

# 1<sup>st</sup> Latin American Music Information Retrieval Workshop

December 9 to 11, 2024

Rio de Janeiro, Brazil



## Proceedings

LAMIR 2024 was organized by the Universidade Federal do Rio de Janeiro and a diverse international committee of organizers.

Website: <https://lamir-workshop.github.io>

*LAMIR 2024 logo design: Amanda Assis (Universidade Federal do Rio de Janeiro)*

*Cover page design: Leonardo Barreto Alves (Universidade Federal do Rio de Janeiro)*

*Edited by:*

Magdalena Fuentes (*New York University*)

Luiz Wagner Pereira Biscainho (*Universidade Federal do Rio de Janeiro*)

Martín Rocamora (*Universitat Pompeu Fabra and Universidad de la República*)

Lucas Simões Maia (*Universidade Federal do Rio de Janeiro*)

Carlos Eduardo Cancino-Chacón (*Johannes Kepler University Linz*)

ISBN: 978-65-01-30797-8

*Title: Proceedings of the 1st Latin American Music Information Retrieval Workshop, Rio de Janeiro, Brazil, Dec. 9-11, 2024*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee, provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page.

© 2025 Latin American Music Information Retrieval Workshop



# Table of Contents

<b>Table of Contents</b>	<b>iii</b>
<b>Sponsors</b>	<b>v</b>
<b>Organizing Team</b>	<b>vii</b>
<b>Preface</b>	<b>ix</b>
<b>Reviewers</b>	<b>xiii</b>
<b>Keynote Talks</b>	<b>1</b>
Industry: Bridging Research and Real-World Applications in MIR	
<i>Igor Gadelha</i> . . . . .	1
Keynote 1: Including Latin American Music in MIR	
<i>Magdalena Fuentes</i> . . . . .	2
Keynote 2: An AI Dress Rehearsal: Exploring Music Performance and Interaction with Computational Models	
<i>Carlos Cancino-Chacón</i> . . . . .	3
Keynote 3: Digital Images and Symbolic Encoding of Guatemalan Polyphonic Choirbooks: Enhancing Preservation and Access for Early Music Sources through Digitization and Music Information Retrieval	
<i>Martha Thomae</i> . . . . .	4
<b>Tutorials</b>	<b>5</b>
<b>Session 1: Music Analysis and MIR Tasks</b>	<b>9</b>
Analyzing Pitch Content In Traditional Ghanaian Seperewa Songs	
<i>Kelvin Walls, Iran R Roman, Kelsey Van Ert, Colter Harper, Leila Adu-Gilmore</i> . . . . .	9
A Corpus Study Of Rhythm In Brazilian Popular Music	
<i>Hugo T Carvalho, Felipe D Martins, Carlos de Lemos Almada</i> . . . . .	14
Beat Tracking For Salsa Music: Adapting And Benchmarking Models Using A Newly Introduced Salsa Dataset	
<i>Antonin PL Rapini, Anna Jordanous</i> . . . . .	19
Music Source Separation In Historical Brazilian Choro Recordings	
<i>Pedro Donadio de Tomaz Júnior, Martín Rocamora, Luiz W P Biscainho</i> . . . . .	24
Tempo Estimation Using Combined Mel-Spectrogram And Mel-Scalogram Inputs	
<i>Luiz Alberto Guimarães Viana, Antonio Carlos Lopes Fernandes Jr, Eduardo Simas</i> . . . . .	29
Skip That Beat: Augmenting Meter Tracking Models For Underrepresented Time Signatures	
<i>Giovana V Moraes, Brian McFee, Magdalena Fuentes</i> . . . . .	34
Improving Music Emotion Recognition By Leveraging Handcrafted And Learned Features	
<i>Pedro L. Louro, Hugo Redinho, Ricardo S Malheiro, Rui P Paiva, Renato Panda</i> . . . . .	40
<b>Session 2: MIR Applications</b>	<b>47</b>

Symbolic Music Style Transfer Via Latent Space Transformations: Model And Evaluation <i>Lucas Somacal, Pablo Riera, Diego Fernandez Slezak, Martin A Miguel</i> . . . . .	47
Surprising Patterns In Musical Influence Networks <i>Flavio Figueiredo, Tales Panoutsos, Nazareno Andrade</i> . . . . .	53
I've Heard This Before: Tiktok's Impact On The Re-Popularization Of Songs <i>Breno S Matos, Francisco Galuppo Azevedo, Rennan Lima, Flavio Figueiredo</i> . . . . .	58
Assessing The Impact Of Sampling, Remixes, And Covers On Original Song Popularity <i>Guilherme S dos Santos, Flavio Figueiredo</i> . . . . .	63
Shallow Neural Network Architectures For Musical Genre Classification <i>Natanael L. de Matos, Hugo T Carvalho, Carlos Tadeu Pagani Zanini</i> . . . . .	68
Aeromamba: An Efficient Architecture For Audio Super-Resolution Using Generative Adversarial Networks And State Space Models <i>Wallace C de Abreu, Luiz Biscainho</i> . . . . .	74
Long-Form Text-To-Music Generation With Adaptive Prompts: A Case Of Study In Tabletop Role- Playing Games Soundtracks <i>Felipe F Marra, Lucas N. Ferreira</i> . . . . .	80
<b>Author Index</b>	<b>85</b>



# Sponsors

## Sponsors



## Institutional Support





# Organizing Team

## Organizing Committee

Magdalena Fuentes (New York University)

Luiz Wagner Pereira Biscainho (Universidade Federal do Rio de Janeiro)

Martín Rocamora (Universitat Pompeu Fabra and Universidad de la República)

## Program Committee

Lucas Simões Maia (Universidade Federal do Rio de Janeiro)

Carlos Eduardo Cancino-Chacón (Johannes Kepler University Linz)

## Logistics Committee

Giovana Morais (New York University)

Richa Namballa (New York University)

Xavier Juanola Molet (Universitat Pompeu Fabra)

Leonardo Barreto Alves (Universidade Federal do Rio de Janeiro)



# Preface

Bem-vindo ao LAMIR!  
¡Bienvenidos a LAMIR!  
Welcome to LAMIR!

LAMIR, the 1st Latin American Music Information Retrieval Workshop, aims to provide opportunities for local students and researchers to cultivate the Latin American community within the International Society for Music Information Retrieval (ISMIR) and the Artificial Intelligence (AI) communities. LAMIR is a satellite event of the **ISMIR 2024** conference and part of the **KHIPUx 2024** events.

The first edition of LAMIR was held at the Universidade Federal do Rio de Janeiro (UFRJ) in Rio de Janeiro, Brazil. LAMIR is part of a broader effort of the MIR community to promote diversity by supporting the development of new communities interested in MIR and connecting existing communities already working on MIR. Our community embraces a wide range of scientific disciplines, experience levels, professional affiliations, and cultural backgrounds. Our organizing team, composed of Latin Americans living and working across South America, the U.S., and Europe, has been dedicated to making this event a success and warmly welcomes you to LAMIR!

## Scientific Program

The LAMIR scientific program comprised 14 papers, three keynote presentations, one industry talk, and a tutorial and demo session. A total of 35 papers were registered in the CMT submission system, of which 24 were submitted as complete papers eligible for review. We received submissions from 8 different countries across three continents. Following ISMIR conference practices, a two-tier double-blind peer-review process was conducted with 41 reviewers and 5 meta-reviewers. Each paper was assigned to one meta-reviewer and three reviewers, with replacements found as needed, and ensuring that at least one of the reviewers was a senior member of the MIR community. Meta-reviewers handled up to 5 papers, while reviewers handled no more than 2. After the initial review, the Scientific Program and Organizing Committees made final decisions on the papers. A total of 15 papers were accepted (one later withdrawn), resulting in an acceptance rate of 62.5% (or 42.9% including incomplete submissions and desk rejections).

The Scientific Program Committee would like to express their thanks to the MIR community of reviewers for their support of this critical aspect of a successful technical program.

Table 1 summarizes the number of submitted and accepted papers in each subject area (selected by the authors during the submission process) together with the corresponding proportion of papers in the program. The accepted papers had a total of 40 unique authors, with an average of 3.1 authors per paper. Of the accepted papers, 9 (60.%) had at least one student author, and 5 (33.3%) focused on applications to Latin American Music. Figure 1 shows the distribution of authors by country.

Accepted papers were presented as posters and divided into two sessions, each consisting of seven posters. The first session focused on Music Analysis and MIR tasks, covering topics such as beat tracking, tempo estimation, pitch and rhythm analysis, music source separation, and music emotion recognition. The second session focused on MIR applications, including symbolic style transfer, text-to-music generation, the impact

Table 1: Papers submitted and accepted by subject area

Primary Subject Area	Total Papers	Accepted Papers	Accepted %
Applications	5	4	26.7%
Computational Musicology	3	1	6.7%
Generative Tasks	1	1	6.7%
MIR Fundamentals and Methodology	4	3	20.0%
MIR Tasks	6	3	20.0%
Musical Features and Properties	5	3	20.0%
<b>Total</b>	<b>24</b>	<b>15</b>	

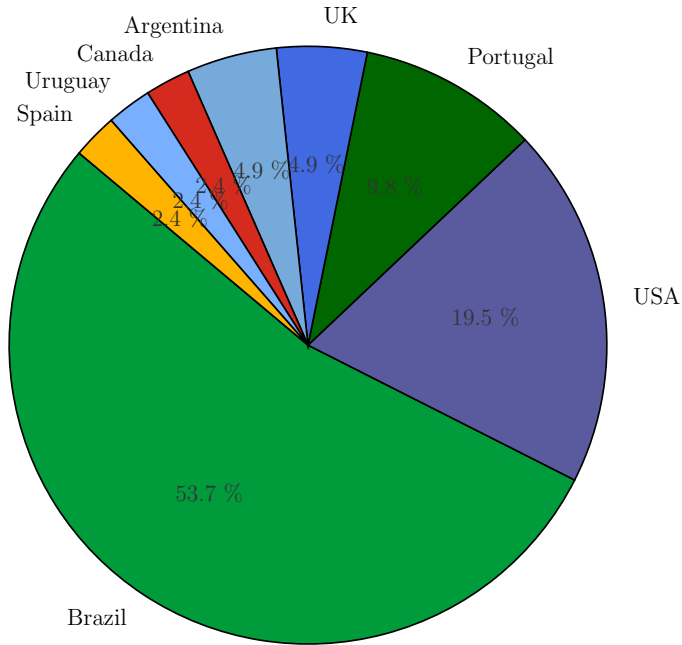


Figure 1: Distribution of Authors by Country

of social media platforms on music popularity, music genre classification, and deep learning models for audio super-resolution.

## Tutorial, Hackathon and Demos

### Adapting Deep Learning Models for Latin American Music Tasks with Little Data

Giovana Morais, Richa Namballa, Xavier Juanola, Martín Rocamora and Magdalena Fuentes

[https://lamir-workshop.github.io/lamir\\_hackathon/intro.html](https://lamir-workshop.github.io/lamir_hackathon/intro.html)

## Keynotes

Industry Talk: **Igor Gadelha**

Head of Machine Learning at Music.Ai

*Bridging Research and Real-World Applications in MIR*

Keynote Talk 1: **Magdalena Fuentes**

Assistant Professor at the Music and Audio Research Lab and the Integrated Design and Media Program at New York University

*Including Latin American Music in MIR*

**Keynote Talk 2: Carlos Cancino-Chacón**

Assistant Professor at Institute of Computational Perception, Johannes Kepler University Linz

*An AI Dress Rehearsal: Exploring Music Performance and Interaction with Computational Models*

**Keynote Talk 3: Martha Thomae**

Postdoctoral Researcher at NOVA University of Lisbon

*Digital Images and Symbolic Encoding of Guatemalan Polyphonic Choirbooks: Enhancing Preservation and Access for Early Music Sources through Digitization and Music Information Retrieval*

## Music Session

The music program included solo performances and ensemble pieces reflecting the diversity of traditional Latin American music.

**Carlos Cancino-Chacón** (Mexico)

Selection of piano works by Manuel Ponce (1882-1948)

**Sara Cohen** (Brazil)

Selection of piano works by Ernesto Nazareth (1863-1934) and Heitor Villa-Lobos (1887-1959)

**Cuarteto Colibriyo** (Uruguay)

Selection of Alberto Mastra (1909-1976) and Alfredo Zitarrosa (1936-1989)

## Travel Awards

The LAMIR workshop aimed to provide opportunities for local students and researchers to connect and foster the MIR/AI community in Latin America. Thanks to the generous support of our sponsors, we have offered financial assistance in the form of travel awards to local Latin American students.

We provided two types of travel awards, each with specific criteria:

1. **Author Awards:** These are intended for student authors of a paper accepted to LAMIR 2024.
2. **Partial Awards (Non-Authors):** A limited number of partial travel awards are available for non-author students who wish to attend and engage in the workshop.

All applicants were required to be enrolled in a degree-granting university or college program for the 2024–25 academic year.

## Acknowledgments

We are happy to present to you the proceedings of LAMIR 2024. The workshop program was made possible thanks to the hard work of many people, including the LAMIR Organizing Committee and Scientific Program Committee, volunteers and the reviewers, who contributed to make the workshop a success.

We would also like to thank our sponsors, whose generous support made this workshop possible possible:

### Sponsors

- CAPES
- Google
- ISMIR

- Khipu
- Music.Ai

### **Institutional Support**

- IEEE Signal Processing Society
- Johannes Kepler University Linz
- New York University
- Universidad de la República
- Universitat Pompeu Fabra
- Universidade Federal do Rio de Janeiro

LAMIR would not have been possible without the outstanding support of our community in response to our call for participation. We extend our deepest gratitude to you—the authors, presenters, and participants—for your invaluable contributions.

Lucas Simões Maia  
Carlos Eduardo Cancino-Chacón

### **Scientific Program Committee**

Magdalena Fuentes  
Luiz Wagner Pereira Biscainho  
Martín Rocamora

### **Organizing Committee**



# Reviewers

## Meta Reviewers

Magdalena Fuentes, New York University  
Luiz Wagner Pereira Biscainho, Universidade Federal do Rio de Janeiro  
Martín Rocamora, Universitat Pompeu Fabra and Universidad de la República  
Lucas Simões Maia, Universidade Federal do Rio de Janeiro  
Carlos Eduardo Cancino-Chacón, Johannes Kepler University Linz

## Reviewers

Jimena Arruti, Universidad de la República	Xavier Juanola-Molet, Universitat Pompeu Fabra
Juan P. Bello, New York University	Jorge David López Ayala, Universitat Pompeu Fabra
Pablo Cancela, Universidad de la República	Lukáš Samuel Marták, Johannes Kepler University Linz
Carlos Eduardo Cancino-Chacón, Johannes Kepler University Linz	Ivan Meresman Higgs, Queen Mary University of London
Estefania Cano, Fraunhofer IDMT	Giovana V. Morais, New York University
Luis Carvalho, Johannes Kepler University	Alia Morsi, Universitat Pompeu Fabra
Hugo T. Carvalho, Universidade Federal do Rio de Janeiro	Richa Namballa, New York University
Joann Ching, Johannes Kepler University	Nestor Napoles Lopez, Avid Technology
Anna-Maria Christodoulou, University of Oslo	Jiyun Park, Korea Advanced Institute of Science and Technology
Alexandre D’Hooge, Université de Lille	Leonardo D. Pepino, Universidad de Buenos Aires
Pablo M. Delgado, Fraunhofer IIS	Silvan Peter, Johannes Kepler University Linz
Sivan Ding, New York University	Marcelo Queiroz, University of São Paulo
Mauricio Do Vale Madeira Da Costa, Osnabrück University	Rafael Ramirez, Universitat Pompeu Fabra
Andres Ferraro, Pandora/SiriusXM	Eduardo A. Silva, Universidade Federal do Rio de Janeiro
Flavio Figueiredo, Universidade Federal de Minas Gerais	Martha E. Thomae Elías, NOVA University of Lisbon
Francesco Foscarin, Johannes Kepler University Linz	Gissel Velarde, International University of Applied Sciences
Esteban A. Gutiérrez, Pontificia Universidad Católica de Chile	Julia Wilkins, New York University
Jan Hajič, jr., Charles University	Pablo Zinemanas, BMAT Music Innovators
Katharina Hoedt, Johannes Kepler University Linz	Wallace C. de Abreu, Universidade Federal do Rio de Janeiro
Patricia Hu, Johannes Kepler University	Yigitcan Özer, International Audio Laboratories Erlangen
Ignacio Irigaray, Universidad de la República	



# Keynote Talks

## Industry Talk

### Bridging Research and Real-World Applications in MIR

Igor Gadelha  
Music.ai

#### Abstract

This keynote delves into the path from Music Information Retrieval (MIR) research to practical applications, highlighting source separation as a pivotal technology. With a background in machine learning and sound engineering, I will discuss the journey of developing and deploying state-of-the-art source separation models, addressing the unique challenges of making these models accurate, efficient, and accessible to a broad user base.

A core focus will be on optimizing source separation models for real-time, low-latency edge deployment. I'll explore the technical intricacies of ensuring these models perform reliably on mobile and constrained devices, where resource limitations challenge both speed and fidelity. Topics will include the trade-offs and design considerations in model compression, latency reduction, and maintaining separation quality. This deep dive will illustrate how we bridge the gap between research and real-world use, making advanced audio separation tools available for musicians, producers, and listeners on-the-go.

Attendees will gain practical insights into the evolving landscape of MIR technology, as well as strategies for overcoming the complexities of real-time deployment, positioning source separation as a powerful, accessible tool in modern music technology.

#### Speaker Bio

As Head of Machine Learning at Music.Ai, Igor Gadelha leads a team dedicated to advancing Music Information Retrieval (MIR) through machine learning. He focuses on developing models for key detection, chord recognition, beat and downbeat tracking, and source separation for audio stems and lyrics transcription. His role includes overseeing data collection projects, aligning research with product requirements, and guiding the technical implementation and deployment of these models. This experience has enabled him to bridge research and practical application, contributing to the development of more precise and accessible music technology tools.

## **Keynote 1**

### **Including Latin American Music in MIR**

Magdalena Fuentes

New York University

#### **Abstract**

In recent years, Music Information Research (MIR) has made remarkable advances in developing computational tools for analyzing, generating, and understanding music. However, there is still a long way to go for MIR tools developed for Latin American music. This talk will examine some examples of the unique challenges that Latin American music presents to MIR research, ranging from data scarcity to complex rhythmic, timbral, and improvisational qualities that defy conventional methods. Specifically, I will highlight the need for culturally sensitive datasets and models, addressing issues like polyphonic textures with overlapping timbres, genre-specific rhythmic nuances, and timing (e.g., rubato and microtiming). Additionally, I will discuss the social importance of adapting MIR tools to underrepresented traditions, including the impact on cultural preservation, educational resources, and creative exploration for Latin American musicians and researchers. I will conclude with a look at open challenges in the field, proposing new directions that include low-data learning approaches, dataset curation, and light-weight models, with the aim to inspire future research for a multicultural MIR landscape that truly reflects the diversity of global music.

#### **Speaker Bio**

Magdalena Fuentes is an Assistant Professor at the Music and Audio Research Lab (MARL) and the Integrated Design and Media (IDM) Program at New York University (NYU), affiliated with both the Steinhardt School of Culture, Education, and Human Development and Tandon School of Engineering. Previously, she was a Provost Postdoctoral Fellow at NYU's Center for Urban Science and Progress (CUSP) and MARL. She completed her Ph.D. in Image and Signal Processing at Université Paris-Saclay, and a B.Eng. in Electrical Engineering from Universidad de la República, Uruguay. Her research focuses on machine listening—a field at the intersection of signal processing and machine learning—where she develops models for understanding natural and everyday sounds, music, and multimodal data. Magdalena is actively involved with the IEEE Audio and Acoustic Signal Processing Technical Committee and regularly serves as an Area Chair/Meta-reviewer for ICASSP and ISMIR. She has held Program Chair roles for DCASE 2021, 2023, and 2025, as well as ISMIR 2025. Her research has been sponsored by NYU, Google and the NIH.

## Keynote 2

### An AI Dress Rehearsal: Exploring Music Performance and Interaction with Computational Models

Carlos Cancino-Chacón

Johannes Kepler University Linz

#### Abstract

The way a piece of music is performed is a very important factor influencing our enjoyment of music. A good performance goes beyond a precise rendering of the score; performers shape aspects like tempo, dynamics, and articulation to convey emotion and engage listeners.

This talk focuses on a specific area of research: computational models of expressive music performance. These models aim to codify hypotheses about expressive performance using mathematical formulas or computer programs, enabling systematic and quantitative analysis. The models serve two main purposes: they allow us to systematically test hypotheses about how music is performed, and they can be used as tools to create automated or semi-automated performances in artistic and educational settings.

In this talk, I will explore two key aspects: data-driven approaches to modeling expressive performance and interdisciplinary collaboration with music cognition to understand how humans interact and develop expressive interpretations. I will illustrate these aspects through three main topics: (1) Basis Function Models, a machine learning framework for generating expressive performances based on musical scores; (2) Studying human interaction in musical performance and insights into the development of a real-time automatic accompaniment system; and (3) The Rach3 Project, an investigation into how pianists learn new music and develop their own expressive interpretations.

#### Speaker Bio

Carlos Cancino-Chacón is an Assistant Professor at the Institute of Computational Perception at Johannes Kepler University Linz (JKU), Austria, and the Principal Investigator of the Rach3 Project, funded by the Austrian Science Fund. The Rach3 Project uses computational, data-driven methods to study long-term music rehearsal, leveraging advances in AI and machine learning. He previously conducted research at the Austrian Research Institute for Artificial Intelligence (OFAI) and was a Guest Researcher at the RITMO Centre, University of Oslo. His research focuses on machine learning models for understanding music performance and listening, with emphasis on three areas: computational modeling of expressive performance, (real-time) human-computer interaction in music, and cognitively plausible machine listening. He holds a PhD in Computer Science from JKU, an MSc in Electrical and Audio Engineering from Graz University of Technology, a Bachelor's in Physics from UNAM, and a Bachelor's in Piano Performance from the National Conservatory of Music of Mexico. He currently serves as a Member-at-Large on the ISMIR Board and as Section Editor for TISMIR.

## Keynote 3

### **Digital Images and Symbolic Encoding of Guatemalan Polyphonic Choirbooks: Enhancing Preservation and Access for Early Music Sources through Digitization and Music Information Retrieval**

Martha Thomae

NOVA University of Lisbon

#### **Abstract**

I will present a pilot project focused on the digitization and encoding of one of six colonial polyphonic choirbooks from the Archivo Histórico Arquidiocesano de Guatemala (AHAG), an archive located next to the Metropolitan Cathedral in Guatemala City. These choirbooks, copied in the 17th and 18th centuries, primarily contain Renaissance European polyphonic music written in mensural notation and provide invaluable insight into Guatemala's colonial-era musical heritage. To preserve and enhance access to this music, I employed a do-it-yourself (DIY) book scanner for high-resolution images, optical music recognition (OMR) software trained for handwritten mensural notation, and an interpreter for mensural notation. Additionally, a music-analysis tool served as an error checker. I will present these tools and their integration into a digitization and music information retrieval (MIR) pipeline to create both digital images and symbolic scores of the choirbook. The symbolic scores are encoded in MEI format, a machine-readable standard that allows for the representation of early music in its original notation. This MIR pipeline can be used to semi-automatically encode other early music sources in mensural notation from both Europe and Latin America. By relying on open, free, and on-line technologies, this pipeline remains accessible to projects with limited resources, furthering the project's mission of "enhancing access" to early music sources.

#### **Speaker Bio**

Martha E. Thomae is a postdoctoral research fellow for the ECHOES project at the NOVA University of Lisbon, where she leads the development of tools to facilitate the search and analysis of chants encoded in the MEI music format. She has a PhD and Master's in Music Technology from McGill University. During her time at McGill, she worked as a research assistant for the Single Interface for Music Score Searching and Analysis (SIMSSA) project, directed by Ichiro Fujinaga.

Most of her research has focused on the preservation and encoding of early music written sources through music information retrieval technologies. Her PhD dissertation focused on digitizing and encoding Guatemalan polyphonic choirbooks from the colonial period, written in mensural notation, using optical music recognition, automatic voice alignment, and computational error detection. She currently serves as a Co-chair of the Mensural MEI IG and has served as a member of the MEI Board.

# Tutorials

During the first day, we will feature a tutorial-hackathon-demo featuring a hands-on session focused on MIR applications in Latin American music. Participants will have the opportunity to work with Latin American datasets and develop AI tools and data loaders that contribute to the MIR and Music-AI community. The results of this work will be shown to the other participants in the demo session.





# **Session 1: Music Analysis and MIR Tasks**



# ANALYZING PITCH CONTENT IN TRADITIONAL GHANAIAN *SEPEREWA* SONGS

Kelvin L. Walls<sup>1</sup>      Iran R. Roman<sup>2</sup>      Kelsey Van Ert<sup>1</sup>  
Colter Harper<sup>3</sup>      Leila Adu-Gilmore<sup>1</sup>

<sup>1</sup> New York University, USA

<sup>2</sup> Queen Mary University of London, UK

<sup>3</sup> University at Buffalo, USA

## ABSTRACT

This study examines the pitch content in traditional Ghanaian *seperewa* (Akan harp-lute) songs, utilizing a unique dataset from field recordings of the mid-twentieth century. We selected 71 songs and used Demucs to isolate vocals from instrumental tracks. We then retrieved the F0 content from these isolated tracks and applied Gaussian Mixture Models (GMM) to approximate musical scales. Comparative F0 analysis between vocals and *seperewa* revealed higher microtonal deviations from equal temperament in vocal tracks. We also note challenges in using MIR tools for musical scale approximation in non-Western music. Our research contributes to the quantitative study of pitch in traditional music of Sub-Saharan Africa.

## 1. INTRODUCTION AND MOTIVATION

This paper examines archived recordings of *seperewa*, a two-course, six or eight string harp-lute that accompanies sung repertoire in Akan languages [1]. The *Akan* are a Ghanaian ethnic and language group that includes the subgroups of *Asante*, *Fante* and *Akuampem* and make up nearly half of the country’s population [2]. European documentation has demonstrated its symbolic importance in the Asante Empire in the 18th century [1]. Due in part to the introduction of guitars from Europe, the *seperewa* nearly disappeared by the mid-20th century, leading to conservation efforts by Ghanaian musicologists Ephraim Amu (1899-1995) and J.H. Kwabena Nketia (1921-2019).

We chose *seperewa* songs because they demonstrate tuning systems other than equal-temperament (though over time it has been tuned to align with equal-tempered instruments like piano or fretted guitar) [1]. The instrument is briefly mentioned in literature addressing West African music; Nketia, however, wrote a comprehensive publication on *seperewa* harmony and melody [3]. A more recent

study by McPherson and Obiri-Yeboah examines Akan language encoding in *seperewa* music [4]. Therefore, due to its cultural significance, studying traditional *seperewa* music could shed light on the original indigenous practices in Africa that made their way to the Americas and shaped the African diasporic musical practice across the globe.

Musicological research has documented and analyzed musical and social aspects of traditional music across the Africa continent, as well as the influences of Western European religious and military music [5]. In particular, there is extensive literature focusing on aspects of rhythm in African music, such as cycle and multidimensionality [6–8], as well as musical connections between Africa and Afro-Latino communities [9, 10]. More recent developments have employed computer analysis to look at micro-timing in African drumming to suggest alternative approaches to meter [11, 12]. However, there is little research on pitch content in African music that explores microtonal variance and tendencies outside the Western equal-tempered tuning system [13]. We seek to redress this balance by investigating Ghanaian music’s unique *scales*. Nketia stated, “*seperewa* music and Akan songs in general are based on the heptatonic scale, though there remains a great deal of variance from Western tuning systems and tonal logic within this framework” [1]. Therefore, we examine *Akan* pitch through the implied scale and the microtonal content of the *seperewa*’s song repertoire. Therefore, our research questions are:

1. Given a known *seperewa* scale, can we use MIR methods (such as source separation, F0 tracking, and probability density function modeling) to detect its presence in the vocal and instrument pitch content of a *seperewa* song?
2. How “equal tempered” are the overall scales we approximate? how microtonally flat or sharp are specific scale degrees?
3. How similar and different are the scales between the *seperewa* and the vocals?

These questions align with our broader goal of decolonizing datasets and studying the effects of music technology’s embedded biases (i.e. the equal-tempered system in synthesis instruments, MIDI protocols, and recording ef-



© KL Walls, IR Roman, K Van Ert, C Harper and L Adu-Gilmore. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** KL Walls, IR Roman, K Van Ert, C Harper and L Adu-Gilmore, “Analyzing Pitch Content in Traditional Ghanaian *Seperewa* Songs”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.

fects like autotune) on traditional and popular musics of the world. In keeping with our decolonizing methods, we obfuscate the original audio material<sup>1</sup> to retain indigenous intellectual property of this archive. Through collaborative efforts with the original authors and their communities, we aim to activate these archives for research in innovative and transformative ways.

## 2. ACTIVATING AN ARCHIVAL GHANAIAAN MUSIC DATASET

This study’s dataset is drawn from field recordings collected in the early 1960s by Ghanaian composer and ethnomusicologist Ephraim Amu (1899-1995). These recordings were part of a larger project by Ghanaian ethnomusicologist, composer, and linguist Dr. J.H. Kwabena Nketia (1921-2019), which began in 1952 and aimed to collect recordings of folklore, music, and poetry in, what was then, the United Kingdom’s Gold Coast Colony.

These recordings are the earliest known of this important endangered music tradition, offering an opportunity to examine pitch content, as well as melodic and harmonic structures in music with pre-colonial origins. The importance of these archival materials are also tied to the locations in which they are housed and the individuals that steward those materials. In working with this data set, we are supporting efforts to not only preserve African cultural heritage but also develop resources and institutions to house the data in the countries from which they originated.

## 3. METHODS

The *seperewa* is traditionally tuned to a heptatonic scale [1]. For this analysis, we paired MIR techniques with the informed analysis of an expert plucked string instrument performer (he/him), ethnomusicologist, and scholar of traditional Ghanaian music. He provided us with the approximate tuning of the *seperewa* for each song. Therefore, for each song we have the *seperewa* heptatonic scale with a “tonic” that corresponds to the frequency of an key in an equal-tempered piano (as identified by the *seperewa* expert), and whether the heptatonic scale had major or minor third and sixth degrees. In this scale, the second was always major, the fourth and the fifth were always perfect, and the seventh was always absent. Note that while the *seperewa* has a known tuning, the vocal parts of these songs can freely sing other notes (such as the tritone or the 7th) or microtonal inflections around all scale degrees. Therefore, our analysis does not impose equal-tempered tuning standards as correct. Rather, we analyze the actual values performed by these traditional Ghanaian musicians.

### 3.1 Scale Approximation Pipeline (SAP).

Our pipeline builds upon other similar ones for F0 vocal analysis [14–16]. Our archive consists of 71 songs. We use Demucs [17] to split our songs into isolated vocal and

instrumental tracks, then CREPE [18] to extract the instantaneous pitch content (F0). We drop F0 values with a confidence score under 0.8, and convert all values from Hz to cents to interpret results on a linear scale. We also quantize F0 estimates to the nearest tenth (i.e. 10 bins per semitone of 100 cents; each bin encompasses 10 cents) to enhance our analysis of densities.

We determine a track’s scale using the method by Roman et al. [14]. This involves training a GMM to identify Gaussian components in the histogram of F0 values for a song. These components signify the notes in a track’s musical scale. We limit our analysis to F0 values within a whole step below and an octave above the known *seperewa* tonic. After GMM modeling we also recursively average components within 50 cents of each other per song. To calculate how aligned a scale’s components are with equal temperament, we use the approach described by Roman et al. obtaining an  $\epsilon_S$  score per song, where a value of zero indicates the highest possible alignment with equal temperament, and 50 complete deviation from it. Furthermore, we assign “scale degree” identities to each component by identifying the tonic based on the known *seperewa* tuning and computing the distance of other components to this tonic anchor. For example, if a component is 930 cents above the tonic, that component is labeled as a major sixth with a sharpness of 30 cents.

## 4. RESULTS

### 4.1 Finding scales in isolated *seperewa* & vocal tracks.

Table 1 quantifies how our SAP identified scale components. First, the column “No. of songs where in *seperewa* tuning” shows the number of times a scale degree (in the specific quality of major or minor, when appropriate) was known to be part of the *seperewa* tuning. This knowledge about the *seperewa* tuning was provided by the expert *seperewa* player that we consulted. Note that the tonic, major second, fourth, and fifth were part of the tuning in all songs. The 3rd and 6th were predominantly “major”, and only a handful of songs featured a “minor” tuning. Note also how no song featured a seventh in the expected *seperewa* tuning, confirming the underlying *seperewa* heptatonic scale according to the *seperewa* expert. Similarly, the minor second and the tritone are never part of the scale. The next column, labeled “Retrieved”, shows that all scale degrees known to be part of the *seperewa* tuning were found in our corpus, although not in all songs. For example, the tonic was found in the *seperewa* in only two thirds of the songs, while in the vocals it was found in almost all songs. The column labeled as “missing” quantifies the number of songs where a scale degree was known to be part of the *seperewa* tuning and was not found by our SAP. We also quantified the number of scale degrees found that were “unexpected” since they were not part of the *seperewa* tuning. Interestingly, in the *seperewa* track (and also in the vocals, although less surprisingly due to the expressive pitch abilities of the voice) our SAP found unexpected minor seconds, minor thirds, and minor sev-

<sup>1</sup> the archive granted permission to share some song examples at [seperewa-pitch-analysis.github.io](https://seperewa-pitch-analysis.github.io)

**Table 1. Given a known *seperewa* tuning, which scale components did our scale approximation pipeline (SAP) find?** The third column indicates the number of songs in our corpus where a scale degree (specified by the first two columns) is part of the known *seperewa* tuning. The other columns denote the number of times our SAP "Retrieved" each degree, "Missed" it, or found it although it was "Unexpected" in the *seperewa* tuning. Results are shown for both *seperewa* and vocals. "—" denotes that the scale degree was not part of the known *seperewa* tuning in any song.

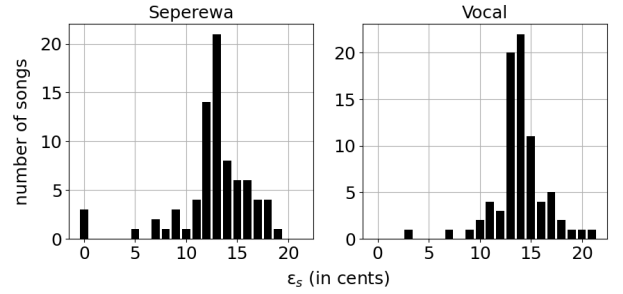
Scale Degree	Quality	No. of songs where in <i>seperewa</i> tuning	Seperewa			Vocals		
			Retrieved	Missing	Unexpected	Retrieved	Missing	Unexpected
Tonic		71	50	21	0	65	6	0
2nd	Minor	—	—	—	19	—	—	14
	Major	71	39	32	0	49	22	0
3rd	Minor	3	2	1	17	1	2	26
	Major	68	39	29	1	44	24	2
4th		71	47	24	0	55	16	0
Tritone		—	—	—	12	—	—	11
5th		71	58	13	0	61	10	0
6th	Minor	3	2	1	6	2	1	8
	Major	68	42	26	1	45	23	1
7th	Minor	—	—	—	25	—	—	25
	Major	—	—	—	44	—	—	42
Avg. (std.)		53.25 ( $\pm$ 29.04)	34.88 ( $\pm$ 19.86)	18.38 ( $\pm$ 11.34)	10.42 ( $\pm$ 13.2)	40.25 ( $\pm$ 23.39)	13.0 ( $\pm$ 8.9)	10.75 ( $\pm$ 13.12)

enths a considerable number of times. It also found unexpected tritones and sixth. Finally, the bottom row in the table summarizes the average performance of SAP on scale degree retrieval from *seperewa* and vocal tracks, also summarizing the average number of missing and "unexpected" components.

Together, these results demonstrate that our SAP can majorly identify the scale degrees in the known *seperewa* tuning in both the isolated *seperewa* and vocal tracks. There are limitations to our approach, however. For instance, the number of missed and "unexpected" components is not trivial and deserves attention. In the case of the SAP applied to the *seperewa*, we hypothesize that the large number of missing components could be caused by the imperfect separation of the instrumental and vocal tracks by Demucs. Through manual inspection we found that Demucs allows for the *seperewa* track to leak into the vocal track, particularly in the lower range of the *seperewa*. This causes a considerable amount of *seperewa* information to be missing in the instrumental track and could be the underlying factor of the relatively low retrieval of some scale degrees in the *seperewa*. In the case of the vocals, the "unexpected" scale degrees are easily explained by the voice's ability to freely show microtonal inflections. The "unexpected" components in the *seperewa* can be explained by voice leakage into the instrumental track. Future work should look into improving Demucs for the type of recordings we used in this study, and better measuring the challenges of using such model to robustly separate vocal and instrumental tracks.

#### 4.2 Deviation of scales from equal temperament.

After approximating each song's scale in the *seperewa* and vocal tracks, we measured how much each scale deviates



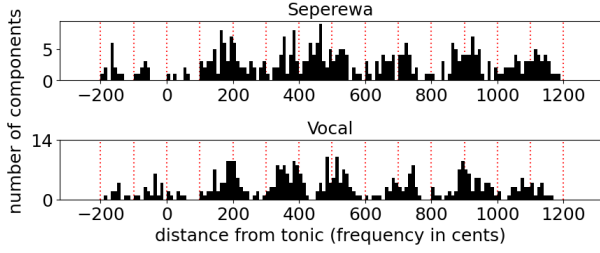
**Figure 1. How equal-tempered are the scales of songs?**

This figure shows the density of the error  $\epsilon_S$  (in cents) between the equal-tempered scale and the scale we approximated with our SAP for each song in the corpus. On the x-axis, an  $\epsilon_S$  at or close to zero would correspond to an equal-tempered scale, while higher  $\epsilon_S$  values denote deviation from equal temperament. Separate histograms show the density of  $\epsilon_S$  for the *seperewa* and the vocals.

from equal temperament by calculating  $\epsilon_S$  [14]. Fig. 1 shows the distribution of  $\epsilon_S$  values across all songs, separately shown for the *seperewa* and the vocals, with a mean of 11.6 (5.07 std.) and 13.67 (4.92 std.), respectively. The larger mean in the vocal  $\epsilon_S$  is expected given the voice's ability to freely represent microtonal inflections that deviate from equal temperament. Also note how the *seperewa* has three songs perfectly aligned with equal temperament ( $\epsilon_S = 0$ ), while the same is not observed for the vocals. In general, these results highlight this music's deviation from equal temperament.

#### 4.3 Microtonal inflections of scale degrees.

We also wanted to understand the microtonality of individual scale degrees. Fig. 2 shows the density of each scale



**Figure 2. How microtonally flat or sharp are individual scale degrees in our corpus with respect to equal temperament?** The density of all scale components found with our SAP on the entire corpus. The vertical red lines correspond to a twelve-tone equal-tempered scale. The x-axis is the frequency distance from the tonic (in cents). The top plot shows the component density in the *seperewa* and the bottom is the corresponding plot for vocals.

degree in terms of distance in cents from the tonic (recall that 100 cents is one semitone), with red vertical lines referencing the equal tempered twelve-tone scale. Results are shown for both the *seperewa* (top) and the vocals (bottom).

Note the density around 200 cents, corresponding to a microtonally flat “major second” in both the *seperewa* and the vocals. Similarly, a density between 300 and 400 cents corresponds to a “third” that is between minor and major, an effect that is more clearly visible in the vocal scales. Next, there are clear densities around 500 and 700, corresponding to the perfect fourth and fifth, respectively. Finally, both *seperewa* and vocals show a clear density corresponding to a major sixth (900 cents). These results give insight into the microtonal tendencies in the individual scale components in this musical corpus.

#### 4.4 Comparing the *seperewa* and vocal scales directly.

It is also interesting to directly compare the scale components that were common between the *seperewa* and the vocals in a given song. This allows to answer whether one is flatter or sharper than the other in general. Table 2 shows the results of this analysis. Here we observe again that major thirds were considerably flatter (negative numbers) in the vocals. All other scale degrees showed microtonal deviations between the *seperewa* and the vocals, most times with the vocals being slightly sharper (positive numbers) than the *seperewa*.

## 5. DISCUSSION

This study analyzed the complex relation between indigenous Ghanaian musical practices and Western tuning influences, revealing systematic and structural deviations and relations to equal temperament. For example, we identified evidence supporting the existence of a heptatonic scale in the tuning systems and practice of the *seperewa* and the vocal singing that it is usually performed with. Our analysis highlights the resilience of traditional tuning practices despite the encroachment of Western musical norms, underscored by the presence of microtonal in-

**Table 2. How flat or sharp is each sung scale degree compared to the *seperewa* scale?** Comparison of scale components between *Seperewa* and Vocals as found in SAP. Negative average distances indicate the component was flatter in Vocals than in *Seperewa*, while positive values indicate it was sharper. The third column specifies the number of songs in our corpus where a given component was found in both the *Seperewa* and the vocals and were used for this analysis. All values are in units of cents.

Scale Degree	Quality	No. of songs with comp. in both	Avg. Distance ( $\pm$ std.)
Tonic		50	-2.03 ( $\pm$ 36.8)
2nd	Minor	7	2.01 ( $\pm$ 15.21)
	Major	30	-1.08 ( $\pm$ 36.92)
3rd	Minor	12	2.2 ( $\pm$ 24.58)
	Major	26	-12.57 ( $\pm$ 42.13)
4th		37	21.42 ( $\pm$ 38.65)
Tritone		4	8.67 ( $\pm$ 30.01)
5th		53	0.67 ( $\pm$ 32.32)
6th	Minor	4	17.92 ( $\pm$ 32.0)
	Major	28	7.08 ( $\pm$ 29.11)
7th	Minor	11	-11.12 ( $\pm$ 45.41)
	Major	24	-21.45 ( $\pm$ 43.59)

flections and the retention of non-equal tempered scales in both *seperewa* and vocals. The findings challenge prevailing assumptions about the universality of Western tuning standards and highlight the importance of context-sensitive musicological analysis. Future studies should enhance this research by incorporating statistical testing of the preliminary observations made here, necessitating a larger dataset and defined, measurable variables to carry out a statistical analysis that determines the significance of the trends we have described.

## 6. CONCLUSION

Our research contributes to a nuanced understanding of Ghanaian musical scales, revealing a rich tapestry of sound that defies simple categorization within the Western equal-tempered system [13, 19]. By documenting the microtonal variances and tuning discrepancies in *seperewa* songs, this study not only aims at preserving a vital aspect of Ghanaian cultural heritage but also fosters a broader appreciation for the musical diversity that defines the African diaspora. This work underscores the necessity of adopting decolonizing methodologies in musicology, advocating for a more inclusive approach that respects and elevates non-Western musical forms. Future research should continue to explore these themes, further bridging the gaps between traditional African music and its diasporic iterations thereby enriching our global musical heritage.

## 7. REFERENCES

- [1] J. K. Nketia, “Generative processes in seperewa music,” *To the Four Corners: A Festschrift in Honor of Rose Brandel*, vol. 14, no. 1994, p. 117, 1994.
- [2] I. Wilks, *Asante in the Nineteenth Century: the structure and evolution of a political order*. CUP Archive, 1989, vol. 13.
- [3] K. Agawu, “Tonality as a colonizing force in africa,” *Audible empire: Music, global politics, critique*, pp. 334–56, 2016.
- [4] L. McPherson and M. Obiri-Yeboah, “Akan tone encoding across musical modalities,” *Studies in African Linguistics*, vol. 52, no. 1-2, pp. 160–88, 2023.
- [5] K. Agawu, *On African Music: Techniques, Influences, Scholarship*. Oxford University Press, 2023.
- [6] D. Locke, “The metric matrix: Simultaneous multidimensionality in african music,” *Analytical Approaches to World Music*, vol. 1, no. 1, pp. 48–72, 2011.
- [7] V. K. Agawu, *African rhythm: A northern Ewe perspective*. CUP Archive, 1995.
- [8] N. Jacoby, R. Polak, and J. London, “Extreme precision in rhythmic interaction is enabled by role-optimized sensorimotor coupling: analysis and modelling of west african drum ensemble music,” *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1835, p. 20200331, 2021.
- [9] A. Villepastour, “Two heads of the same drum? musical narratives within a transatlantic religion,” *Journal of Transatlantic Studies*, vol. 7, no. 3, pp. 343–362, 2009.
- [10] P. Manuel and O. Fiol, “Mode, melody, and harmony in traditional afro-cuban music: From africa to cuba,” *Black Music Research Journal*, pp. 45–75, 2007.
- [11] R. Polak and J. London, “Timing and meter in mande drumming from mali,” *Music Theory Online*, vol. 20, no. 1, 2014.
- [12] G. Sioros, “Polyrhythmic modelling of non-isochronous and microtiming patterns,” in *Proceedings of the 24th International Society for Music Information Retrieval Conference*, 2023.
- [13] J. Collins, “The early history of west african highlife music,” *Popular Music*, vol. 8, no. 3, pp. 221–230, 1989.
- [14] I. R. Roman, D. Faronbi, I. Burger-Weiser, and L. Adu-Gilmore, “F0 analysis of ghanaian pop singing reveals progressive alignment with equal temperament over the past three decades: A case study,” in *20th Sound and Music Computing Conference, SMC 2023*. Sound and music Computing network, 2023, pp. 27–33.
- [15] E. Georgieva, P. Ripollés, and B. McFee, “Total variation in vocals over time,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2023.
- [16] K. L. Walls, I. R. Roman, B. Steers, and E. Georgieva, “Total variation in popular rap vocals from 2009-2023: Extension of the analysis by georgieva, ripollés & mcfée,” in *Ismir 2023 Hybrid Conference*, 2023.
- [17] S. Rouard, F. Massa, and A. Défossez, “Hybrid transformers for music source separation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [18] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, “Crepe: A convolutional representation for pitch estimation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.
- [19] J. W. Shipley, *Living the Hiplife: celebrity and entrepreneurship in Ghanaian popular music*. Duke University Press, 2013.

# A CORPUS STUDY OF RHYTHM IN BRAZILIAN POPULAR MUSIC

Hugo T. Carvalho<sup>1</sup>

Felipe D. Martins<sup>2</sup>

Carlos de L. Almada<sup>2</sup>

<sup>1</sup> Department of Statistical Methods, Federal University of Rio de Janeiro, Brazil

<sup>2</sup> School of Music, Federal University of Rio de Janeiro, Brazil

## ABSTRACT

The MPB Project (from the Brazilian Portuguese *Música Popular Brasileira*) is a broad corpus study intended to systematically analyze a large group of musical attributes (regarding harmony, melody, and rhythm) in order to map stylistic characteristics of renowned Brazilian composers of popular music. The present work focuses on the rhythmic structure, describing concisely the scope of the project and some of the most important elements, like corpora formation, datasets, methodological tools, analytical criteria, conceptual framework, and some of the results obtained so far, which in accordance with what we believe would be the expectations of someone familiarized with the style.

## 1. INTRODUCTION

It is natural that Brazilian popular music is considered an attractive and rich subject of study. Especially in recent times, numerous academic musicological and/or ethnomusicological works have been published with various approaches, aligned with important contextual, historical, biographical, aesthetic, social, political, racial, and gender-related issues [1]. These studies are shedding increasingly detailed light on this fascinating subject. However, specifically considering the scope of systematic studies in music, the field is still in its infancy compared to other approaches. The present work is part of a larger research, called the MPB Project,<sup>1</sup> which is intended to pursue a rather ambitious goal: to systematically delineate the contours of the aesthetic-musical context informally referred to as Brazilian Popular Music, better known by the acronym MPB (from the Brazilian Portuguese *Música Popular Brasileira*). The main challenge of this task lies in defining what MPB actually is. Given that our focus is primarily technical and structural, we do not intend to delve deeply into the complex issues (especially those related to social, historical, and political aspects) involved in providing a general definition. We get around these obstacles

<sup>1</sup> Webpage in Brazilian Portuguese: <https://projetoMPB.com.br/>. Some publications under the scope of the project are [2–5].

by simply establishing the temporal arc of what we name “our” MPB common-practice period, which corresponds to Tom Jobim’s creative life (roughly from 1953 to 1993), being Jobim widely regarded as one of the most important Brazilian popular composers and a pioneer in shaping the MPB aesthetic. An important and distinctive aspect of our project is that, unlike most other similar approaches, the analyzes consider not only the harmonic structures of the corpora, but also the melodic and, especially, rhythmic structures. This paper is concerned only with the latter, and it is organized as follows: Section 2 describes the corpora under analysis, and presents the model employed to encode rhythmic information; on Section 3 an exploratory data analysis is presented, together with conclusions drawn from a musicological perspective; the paper is briefly summarized in Section 4, where future works are also outlined.

## 2. THE MPB CORPORA

Corpus studies have become one of the most prominent trends in systematic musicological approaches [6–9]. Driven by rapid advances in computational technology and the development of extensive music databases (particularly in the form of scores, audio files and MIDI files), these studies cover a wide range of repertoires and styles. They are generally aligned with working hypotheses, theoretical models, and methodologies specifically developed for the context, and supported by statistical frameworks. Corpus studies in popular music have also emerged with relative frequency in recent years, encompassing different musical genres and styles [10–14], and the MPB Project falls within the general scope of this trend, though somewhat unusually, it focuses not just on a single corpus but on a broad set of them, subdivided into two main groups: (1) primary, the central focus of the analytical attention, comprising corpora of individual composers of identical lengths (50 pieces each), and (2) secondary, the control group, consisting of three genre corpora (Jazz, Samba, and Choro, genres that highly influenced the MPB) each also containing 50 pieces. As mentioned above, a key feature of our project is that, in contrast to many other comparable studies, our analyses take into account not just the harmonic aspects of the corpora, but also melodic and rhythmic structures.

In this first stage of the project, we selected 10 composers to form the primary group to initiate the analytical process. The transcription of musical information fol-





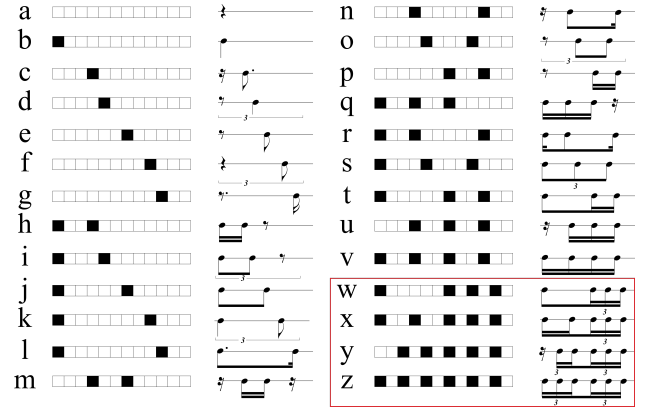
lows some basic criteria: only the main melodic sections are transcribed (i.e., any instrumental introductions, interludes, and *codas* are omitted); repeats and *da capo* repetitions are disregarded. In other words, we are interested in what we call the *nominal form* of the compositions (their essential material, in a certain sense), as opposed to their *realized form* (the final arrangement). The 10 composers selected for the formation of the corpora that make up the primary group in stage one are (listed in the order of analysis): Antonio Carlos (Tom) Jobim (1927–1994), Ivan Lins (1945–), Chico Buarque de Hollanda (1944–), Edu Lobo (1943–), Caetano Veloso (1942–), Djavan (1949–), Milton Nascimento (1942–), João Bosco (1946–), Gilberto Gil (1942–), and Rita Lee (1947–2023). This choice was based on the author’s perception of who are the main composers of the MPB.<sup>2</sup> Regarding the selection of works to be included in the 10 MPB corpora, we favored songbooks of scores available in the Brazilian publishing market [15–24]. All of the scores undergo a review before transcription, during which any errors are corrected.

The main distinction between the control group and the primary group lies in the fact that their repertoires were not selected from the works of specific composers but rather by considering the respective genres.<sup>3</sup> The sources used for the transcriptions were [25,26]. For Samba, the 50 selected pieces were transcribed from their recordings.

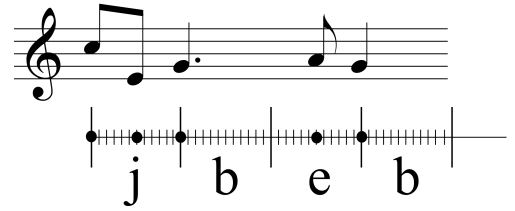
As the present work concerns only the rhythm, we now describe the model used to represent rhythmic information.

## 2.1 Melodic Filtering Model

The foundations for constructing this model, firstly proposed in [2], were established based on three principles: *segmentation*, *abstraction*, and *encoding*. The first principle is grounded in the idea that understanding sequential information requires segmenting it into smaller units, each with relative autonomy (such as short musical “phrases”) [27]. In the model, these groups are referred to as *words*. Abstraction is an essential stage for any analytical process, as it allows different structures to be grouped into equivalence classes based on a specific comparison parameter. In the Melodic Filtering Model (abbreviated as MFM), rhythmic structures are isolated from the melodies under analysis, transforming them into abstract descriptions within this domain. The strategy consists in determining how the onsets of the melodies fall within a grid of 12 equally-spaced sub-divisions of a beat. The onsets in a beat are encoded in an alphabet of *r-letters*, illustrated in Figure 1. In Figure 2 an example of the encoding of a small excerpt is shown. Note that both the dotted quarter note and the quarter note



**Figure 1.** Alphabet of r-letters used in the MFM. The last four letters (“w”, “x”, “y”, and “z”) are suggestions for possible “wildcard letters”, which can be adapted according to the corpus of interest to the analyst.



**Figure 2.** Example of the encoding of a small melodic segment into r-letters, making the r-word “jbeb”.

are encoded by the same r-letter “b”, illustrating that the MFM is sensitive only to attack points.

Note that in Figure 1, the last four letters are highlighted. We observed that the r-letters from “a” to “v” are sufficient to cover almost the entire repertoire of interest. However, MFM (as well as all the frameworks of the entire MPB project) can also be used to analyze other corpora of interest. Thus, the analyst could rely on the letters “a” through “v”, which we believe are sufficient to cover most rhythmic patterns in popular music repertoires, and would have many “wildcard letters” as necessary at their disposal to accommodate less common scenarios. In Figure 1 a suggestion of four wildcard letters is presented.

## 2.2 The dataset

The application of the MFM to the corpora presented at the beginning of Section 2 resulted in a total of 11,119 r-words (groupings of r-letters), stored in a .csv file, whose structure is illustrated in Table 1. The four columns of the file contain, respectively: the name of the corpus, the number of the music, the index of the r-word within the song, and finally, the r-word itself. The relationship between the corpus name and the composer’s name, as well as the song number and its respective title, are provided in two other .csv files containing the corresponding metadata. The dataset and Python scripts to reproduce the results presented in Section 3 are available on a GitHub repository.<sup>4</sup>

<sup>2</sup> It is also important to note that this list comprises only the first stage of the project: overall, the project aims to evaluate a larger group of about 50 composers. Also, the main focus of the project is to analyze *compositional* aesthetics, implicating that artists that were mostly performers, like Gal Costa and Elis Regina for example, will be naturally missing.

<sup>3</sup> In truth, while the pieces in the Jazz and Samba corpora come from different composers, the Choro pieces, by methodological decision, are all composed by Pixinguinha (and occasionally his partner Benedito Lacerda). This decision is justified by the fact that Pixinguinha is widely recognized as the greatest composer of the genre of all time. Therefore, he becomes a kind of ultimate representative of the style.

<sup>4</sup> <https://github.com/ProjetoMPB/LAMIR2024>

corpus	music	position	r-word
BOSCO	1	1	jjb
BOSCO	1	2	fsb
⋮	⋮	⋮	⋮

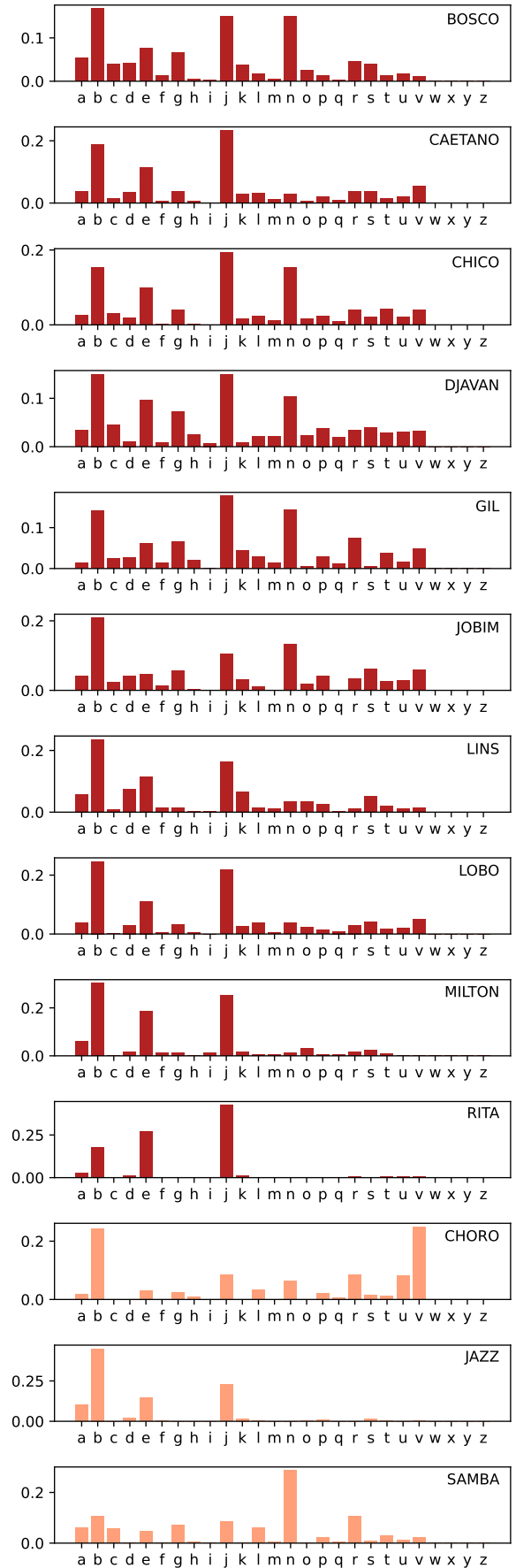
**Table 1.** Header and first two rows of the table corresponding to the .csv file with the r-words related to the corpora analyzed in the first stage of the MPB Project.

### 3. DATA ANALYSIS

Figure 3 illustrates a bar chart for each of the corpora analyzed in the first stage of the MPB Project, showing the occurrences of each r-letter in the respective corpus. Regarding the most common r-letter, the statistics support the practical observation that the r-letter “b” (a single attack, a kind of “tonic degree” rhythmically) is the most prominent in almost all repertoires. Notable are the exceptions that confirm the basic rhythmic cells of the control group genres: r-letters “n” (Samba) and “v” (Choro), as well as the high percentage of “b” in Jazz (about twice as much as in other repertoires). Note also the prominence of “j” in the corpora of Chico Buarque, Caetano Veloso, Milton Nascimento, Gilberto Gil, and Rita Lee (in this last one, the highest percentage of all – 42.5% – almost double the average for this r-letter).

We can also present the information related to the r-letters in another way, through the *metric profile*. This quantity corresponds to the distribution of rhythmic attacks on the micrometric grid of 12 divisions per beat for each corpus, and is presented for the analyzed corpora on Table 2. The metric profile is one of the most important attributes in highlighting a Brazilian “lineage” in MPB, as it deals with the essence of its rhythm. Several aspects deserve commentary: (1) As expected, almost all repertoires have position 1 as the most recurrent point, with notable exceptions being Rita Lee (where point 7 is the most common) and Samba (where the predominant position is 10, the essence of the genre’s syncopation); (2) In this sense, the corpora of Choro, Jobim, Chico Buarque, Djavan, João Bosco, and Gilberto Gil could be considered closest to Samba, given their distributions at this position (more than 1/5 of occurrences). The Jazz corpus (also as expected) shows the lowest percentage at position 10, almost negligible (Milton Nascimento and Rita Lee also have particularly low percentages at this micrometric position); (3) On the other hand, position 7 (which divides the beat in half) shows a high concentration of attacks in the Jazz and Milton Nascimento corpora (both very close and well above the others), and especially in Rita Lee (the highest value of all, 48.3%); (4) The positions that contribute to dividing the beat into three equal parts (5 and 9) show high distribution in Milton Nascimento and Ivan Lins (on average, 12% and 11%, respectively), much higher than in the other repertoires.<sup>5</sup> These positions are rarely occupied

<sup>5</sup> Jobim follows closely after the duo (with an average of 8.6%), which can be attributed primarily to the composer’s early phase, characterized by the strong presence of *samba-canções*.



**Figure 3.** Bar chart showing the proportion of each r-letter in each of the corpora analyzed.

	1	4	5	7	9	10
Bosco	<b>30.5</b>	17.2	6.7	17.9	7.1	20.5
Caetano	<b>37.3</b>	11.2	4.8	27.6	4.6	14.6
Chico	<b>30.2</b>	17.3	3.2	24.5	3.2	21.5
Djavan	<b>29.1</b>	17.7	4.6	23.6	4.5	20.5
Gil	<b>32.2</b>	17.2	1.6	25.8	2.2	21.0
Jobim	<b>33.2</b>	13.7	8.5	15.2	8.7	20.7
Lins	<b>39.6</b>	5.0	11.5	23.4	11.8	8.7
Lobo	<b>39.1</b>	9.6	5.6	25.7	5.6	14.3
Milton	<b>46.5</b>	3.5	6.2	32.8	6.2	4.8
Rita	43.4	2.6	1.1	<b>48.3</b>	1.1	3.4
Choro	<b>44.9</b>	15.2	1.6	14.6	1.6	22.1
Jazz	<b>59.9</b>	0.3	3.1	32.0	3.3	1.4
Samba	24.0	27.8	0.7	12.8	0.7	<b>34.0</b>

**Table 2.** Metric profile by analyzed corpus, in percentage. The columns contain only the attack points present in the corpora. Percentages in bold correspond to the most prominent attack point on each corpus.

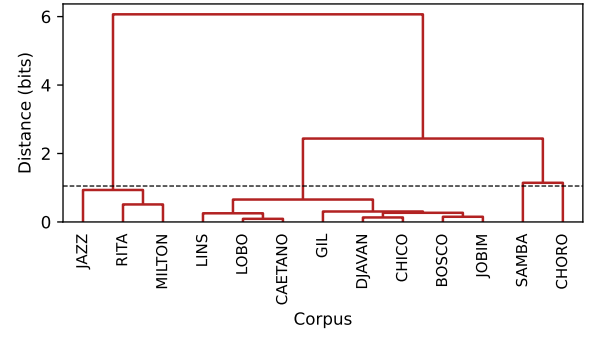
	c	e	g	m	n	p	u
Bosco	4.0	7.7	6.6	0.6	<b>15.1</b>	1.2	1.6
Caetano	1.6	<b>11.5</b>	3.9	1.3	3.0	2.1	2.1
Chico	3.2	10.0	4.1	1.2	<b>15.5</b>	2.4	2.2
Djavan	4.6	9.6	7.3	2.1	<b>10.4</b>	3.8	3.1
Gil	2.7	7.9	6.1	1.9	<b>11.8</b>	3.2	1.5
Jobim	2.5	4.7	5.8	0.1	<b>13.4</b>	4.1	3.0
Lins	1.0	<b>11.4</b>	1.6	1.1	3.5	2.6	1.3
Lobo	0.3	<b>11.1</b>	3.2	0.6	4.0	1.3	2.1
Milton	0.2	<b>18.5</b>	1.3	0.4	1.3	0.5	0.1
Rita	0.3	<b>27.1</b>	0.5	0.1	0.4	0.5	0.8
Choro	0.1	3.1	2.6	0.1	6.3	2.1	<b>8.3</b>
Jazz	–	<b>14.4</b>	0.1	–	–	0.7	–
Samba	5.7	6.9	4.6	0.6	<b>29.0</b>	2.5	1.2

**Table 3.** Distribution of countermetric r-letters by corpus, in percentage. The most prominent countermetric r-letter in each corpus is highlighted in bold.

in Samba (less than 1%) and in Choro (on average, 1.5%).

The affinity relationships suggested above are, in a way, confirmed in Table 3, which details the *countermetric* r-letters.<sup>6</sup> In the corpora where prominence falls on the countermetric r-letter “e”, we might speculate whether Ivan Lins and Edu Lobo exhibit a jazz influence in the aspect of rhythmic organization. On the other hand, Milton Nascimento also shows a strong recurrence of this r-letter, though certainly not only due to jazz influence (which is indeed present), but also likely due to other possible sources (such as folk music, sacred music, or even Beatles songs). In the case of Caetano Veloso and especially Rita Lee, we might attribute the results to the connection their music has with rock and rhythm-and-blues. The corpora of Jobim, Chico Buarque, João Bosco (and, to a slightly lesser extent, Djavan and Gilberto Gil) clearly suggest a closer alignment with Samba, denoted by the prominence of the r-

<sup>6</sup> For the concept of “countermetric rhythms”, see [28].



**Figure 4.** Dendrograms illustrating the hierarchical clustering of the bar charts on Figure 3.

letter “n”, which supports our intuitions about these repertoires. In this context, the Choro corpus stands apart, with a high presence of the r-letter “v”, one of the characteristic rhythmic cells of the genre. Overall, a pronounced use of countermetric r-letters (about one-third of the total configurations) indicates a high degree of rhythmic syncopation.

Finally, Figure 4 illustrates the dendrograms obtained as a result of a hierarchical clustering of the bar charts shown in Figure 3. The symmetrized Kullback-Leibler divergence [29] was computed between pairs of bar charts in Figure 3, producing a distance matrix which is then given to an agglomerative clustering algorithm [30]. Thus, Figure 4 illustrates various possible groupings derived from the collected data (which can be obtained by considering distance thresholds), and highlighting the groupings considering distances smaller than 1.05 bits. This value was chosen for its parsimony in generating few groups and for its ability to distinguish Samba and Choro as two distinct clusters. It should be noted that there is a central grouping containing the corpora of Ivan Lins, Edu Lobo, Caetano Veloso, Gilberto Gil, Djavan, Chico Buarque, João Bosco, and Tom Jobim, as well as a grouping containing Jazz, Rita Lee, and Milton Nascimento. Both are in accordance with the previous considerations.

#### 4. CONCLUSIONS

In this paper we presented preliminary results within the scope of the MPB Project, concerning the rhythmic content of the corpora analyzed so far. The proposed MFM was shown to be capable of capturing idiosyncrasies of the composers considered on the first stage of the project, and also detecting influences of the control group on their rhythmical aesthetic. These conclusions were obtained by a careful and extensive examination of the dataset. However, the application of a hierarchical clusterization technique was able to reveal an information highly correlated with the points previously discussed, which indicates that the proposed framework is capable of assisting in properly understanding the MPB and individual styles. Future work includes a detailed statistical analysis of attributes related to harmony and melodic contour, in order to expand our analytical description of the MPB style.

## 5. ACKNOWLEDGEMENTS

The authors would like to thank the Brazilian funding agencies CAPES (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) for the graduate scholarship awarded to the second author, and CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*) for the academic productivity grant awarded to the third author.

## 6. REFERENCES

- [1] M. T. de Ulhoa, C. Azevedo, and F. Trotta, *Made in Brazil: Studies in Popular Music*. Abingdon: Routledge, 2014.
- [2] C. Almada, *A Melodia de Jobim*. Campinas: Editora da Unicamp, 2023.
- [3] —, *A Harmonia de Jobim*. Campinas: Editora da Unicamp, 2022.
- [4] C. Almada and P. Zisels, “O ritmo de ideias básicas de samba: uma abordagem sistemática,” *Musica Theorica*, vol. 8, no. 1, 2023.
- [5] C. Almada and H. Carvalho, “Entropy, probabilistic harmonic space, and the harmony of antonio carlos jobim,” *Musica Theorica*, vol. 7, no. 1, 2023.
- [6] D. Huron, “On the virtuous and vexatious in the age of big data,” *Music Perception*, vol. 31, no. 3, 2013.
- [7] —, “Error categories, detection, and reduction in a musical database,” *Computers and the Humanities*, vol. 22, no. 1, 1988.
- [8] F. Moss, M. Neuwirth, D. Harasim, and M. Rohrmeier, “Statistical characteristics of tonal harmony: A corpus study of Beethoven’s string quartets,” *PLOS ONE*, vol. 14, no. 6, 2019.
- [9] C. White, *The Music in the Data: Corpus Analysis, Music Analysis, and Tonal Traditions*. New York: Routledge, 2022.
- [10] T. deClercq and D. Temperley, “A corpus analysis of rock harmony,” *Popular Music*, vol. 30, no. 1, 2011.
- [11] M. Mauch, R. MacCallum, M. Levy, and A. Leroi, “The evolution of popular music: USA 1960-2010,” *Royal Society Open Science*, vol. 2, no. 5, 2015.
- [12] J. Serrà, Álvaro Corral, M. Boguñá, M. Haro, and J. L. Arcos, “Measuring the Evolution of Contemporary Western Popular Music,” *Scientific Reports*, vol. 2, no. 521, 2012.
- [13] C. White and I. Quinn, “Chord Context and Harmonic Function in Tonal Music,” *Music Theory Spectrum*, vol. 40, no. 2, 2018.
- [14] F. Moss, W. F. Souza, and M. Rohrmeier, “Harmony and form in Brazilian Choro: A corpus-driven approach to musical style analysis,” *Journal of New Music Research*, vol. 49, no. 5, 2020.
- [15] A. C. Jobim, *Cancioneiro Jobim: Obras Completas (5 vols.)*. Rio de Janeiro: Jobim Music, 2010.
- [16] A. Chediak, *Songbook Ivan Lins (2 vols.)*. Rio de Janeiro: Irmãos Vitale, 2020.
- [17] R. Zappa, *Cancioneiro Chico (2 vols.)*. Rio de Janeiro: Jobim Music, 2018.
- [18] A. Chediak, *Songbook Edu Lobo*. Rio de Janeiro: Lumiar, 1994.
- [19] —, *Songbook Caetano Veloso (2 vols.)*. Rio de Janeiro: Irmãos Vitale, 2020.
- [20] —, *Songbook Djavan (3 vols.)*. Rio de Janeiro: Irmãos Vitale, 2020.
- [21] —, *Songbook João Bosco*. Rio de Janeiro: Irmãos Vitale, 2020.
- [22] W. Lopes and B. Lima, *Songbook Milton Nascimento*. Belo Horizonte: Letramento, 2018.
- [23] A. Chediak, *Songbook Gilberto Gil (2 vols.)*. Rio de Janeiro: Irmãos Vitale, 2020.
- [24] —, *Songbook Rita Lee*. Rio de Janeiro: Irmãos Vitale, 2020.
- [25] Various, *The Real Book*, 6th ed. Milwaukee: Hal Leonard, 2004.
- [26] M. J. Carrasqueira, *O Melhor de Pixinguinha*. Rio de Janeiro: Irmãos Vitale, 2020.
- [27] R. Snyder, *Music and Memory: An Introduction*. Cambridge: The MIT Press, 2001.
- [28] C. Sandroni, *Feitiço decente*. Rio de Janeiro: Zahar/UFRJ, 2001.
- [29] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New Jersey: Wiley, 2005.
- [30] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning with Applications in Python*. Cham: Springer Verlag, 2023.

# BEAT TRACKING FOR SALSA MUSIC: ADAPTING AND BENCHMARKING MODELS USING A NEWLY INTRODUCED SALSA DATASET

**Antonin Rapini**

University of Kent, UK  
apl3@kent.ac.uk

**Anna Jordanous**

University of Kent, UK  
a.k.jordanous@kent.ac.uk

## ABSTRACT

This study addresses the challenge of adapting current beat tracking algorithms, predominantly trained on Western music, to the rhythmic complexities of Salsa, a genre rich in syncopations and polyrhythms. Using training methods that minimise the need for extensive annotated data, we benchmark the adaptability of two established models: BeatNet and BöckTCN, on our newly introduced beat and downbeat annotated Salsa dataset. We find that, on Salsa music, models trained with Salsa largely outperform models trained without any Salsa, nearly matching the accuracy of these models on Western music. This research not only establishes a baseline for beat and downbeat tracking performance in Salsa music but also contributes to the broader goal of developing more adept music information retrieval systems. We also contribute a 40-song Salsa dataset for beat and downbeat tracking research in this genre.

## 1. INTRODUCTION

Beat tracking is the temporal identification of beats—the basic rhythmic units of a song. Although it is a skill that comes naturally for most people [1], automatic beat tracking—the computational identification of beats from audio data—poses significant challenges for computational systems.

Downbeat tracking involves identifying the first beat of each measure in a musical piece and requires a deeper understanding of a song’s musical structure, making it more challenging for computational models. In the current literature, downbeat tracking generally yields lower accuracy than beat tracking.

Current state-of-the-art beat tracking algorithms typically rely on machine learning models trained on large datasets predominantly featuring Western musical styles [2]. These models achieve high accuracy when evaluated on these same genres. However, their performance on genres such as Salsa remains largely unexplored.

Salsa music, known for its rich rhythmic structure characterized by syncopations and poly-rhythms, holds significant cultural importance and enjoys global popularity [3] [4]. The absence of annotated data for Salsa hinders the formal assessment and adaptation of these algorithms to such complex rhythmic patterns.

This study aims to bridge this gap by evaluating the adaptability of two state of the art beat tracking models: BeatNet [5] and BöckTCN [6], on Salsa music. We introduce a new beat-annotated Salsa dataset and explore training methods that minimize the need for extensive annotated data.

## 2. BACKGROUND

Salsa is known for its rich and dynamic rhythmic tapestry, reflecting the genre’s deep roots in Afro-Cuban musical traditions, African rhythms, and the cultural fusion brought about by Latin American communities in New York City [3] [4]. Central to the genre is the clave pattern, a fundamental rhythmic motif that serves as the structural backbone, often alternating between 3-2 and 2-3 patterns within a 4/4 meter. Additionally, Salsa incorporates polyrhythms, where multiple rhythmic patterns are played simultaneously by different percussion instruments such as congas, timbales, and bongos. This layering of diverse rhythms results in off-beat accents and irregular syncopations. Furthermore, the variable tempo and expressive timing variations in Salsa performances add another layer of difficulty, requiring models to adapt to subtle fluctuations and maintain consistent beat detection. These features are not often represented in current beat tracking datasets, which predominantly focus on genres with more straightforward rhythms typically found in many Western music genres.

Annotating music is a time-consuming and arduous process [7]. The temporal nature of music means the manual annotation process takes at least the length of the annotated segment, often requiring multiple listens and minor corrections to achieve accurate beat placement. This labour-intensive process limits the availability of large, annotated datasets, particularly for genres like Salsa that have been overlooked in previous beat tracking research.

Addressing this lack of data for genres such as Salsa is crucial for progress toward music information retrieval systems that are more representative of diverse musical genres.



© A. Rapini and A. Jordanous. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** A. Rapini and A. Jordanous, “Beat Tracking for Salsa Music: Adapting and Benchmarking Models Using a Newly Introduced Salsa Dataset”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.



### 3. RELATED WORK

The adaptation of beat tracking models to non-Western musical genres has been a growing focus in Music Information Retrieval (MIR), as existing models trained primarily on Western music often struggle with different rhythmic structures. This section reviews three relevant studies that address these challenges.

Maia et al. [8] investigated adapting beat tracking models to Latin American music, specifically Samba and Candombe, using minimal annotated data and computational resources. They tested strategies including training from scratch, fine-tuning a pre-trained model, and applying data augmentation with a TCN model. Their approach demonstrates the potential for adapting beat tracking models to under-represented musical genres with limited annotated data. Building upon their strategies, our work focuses on Salsa music, which, like Samba and Candombe, features complex rhythmic structures. We explore similar methods to evaluate their effectiveness in the context of Salsa.

Fiocchi et al. [9] applied transfer learning to beat tracking in Greek folk music by utilizing a deep BLSTM-based RNN originally trained on popular Western music datasets. They collected and manually annotated a dataset of Greek folk music, which includes a variety of rhythms with irregular time signatures and tempo fluctuations. By freezing the lower layers of the pre-trained network and retraining the top layer on the Greek dataset, they achieved significant improvements compared to models trained from scratch on the limited data. Their results demonstrate the effectiveness of transfer learning in adapting beat tracking models to new genres with limited annotated data. This approach underscores the potential for leveraging existing models to handle diverse musical traditions without the need for extensive new datasets.

Pinto et al. [10] proposed a user-driven fine-tuning approach for beat tracking, aiming to enhance the performance of state-of-the-art models on specific challenging musical pieces. Their method involves adapting a pre-trained TCN by fine-tuning it using a small, user-annotated segment of the target piece. This approach allows the model to better handle expressive timing variations and complex rhythms without the need for large annotated datasets. They demonstrated significant improvements in beat tracking accuracy across various datasets. However, their approach is tailored to individual pieces, which limits its scalability and applicability to genre-wide adaptation. This raises questions about its effectiveness for broader applications where generalization across an entire genre is desired.

These studies provide valuable insights into the challenges and potential strategies for adapting beat tracking models to under-represented musical genres. However, limitations remain in achieving generalization across an entire genre with minimal annotated data. Our work extends these efforts by benchmarking multiple models and training conditions specifically for Salsa.

### 4. OBJECTIVES

This study sets out to establish a benchmark for beat tracking accuracy on Salsa music by evaluating two state-of-the-art models, Beatnet and BöckTCN, on a newly created beat-annotated Salsa dataset. This benchmark will allow future research to measure progress in developing more effective beat tracking systems for diverse musical genres.

Leveraging techniques such as transfer learning, we aim to optimise these models for Salsa music despite the scarcity of annotated data.

Additionally, this research introduces a novel beat and downbeat annotated Salsa dataset, with the objective of further improving beat tracking systems for diverse musical genres.

### 5. METHODOLOGY

We assess the accuracy of two prominent beat tracking models: BeatNet and BöckTCN, on an unseen Salsa test dataset created for this study. The models were trained under three distinct conditions:

#### 5.1 Training Data

##### 5.1.1 Other Datasets Used

The non-Salsa music datasets used (referred to as “Others”) include: GTZAN [11], Ballroom [12], SMC [13], Beatles [14] and Rock corpus [15]

We were not able to obtain some of the datasets used in the original training of the two models, such as the Hainsworth dataset, due to accessibility constraints.

##### 5.1.2 Salsa Dataset

The Salsa dataset comprises 40 tracks, which were divided into five folds of eight tracks for training. All tracks were used for evaluation, following the process described below.

#### 5.2 Training Conditions

Three training conditions were used: (1) ‘Others Only’, using non-Salsa datasets; (2) ‘Salsa Only’, training solely on our Salsa dataset; and (3) ‘Fine-Tuning’, training with ‘Others’ and then fine-tuning with Salsa data.

We employed 5-fold cross-validation to measure the average F-measure accuracy of the models on the Salsa dataset. In each fold, the models were trained on 32 songs (with 10% used for validation) and evaluated on 8 unseen songs. This method ensures that every song in the dataset is used for testing exactly once.

#### 5.3 Model Configurations

In all cases except fine-tuning, the models were trained with the original parameters presented in their respective papers or official implementations. For “Fine-Tuning” and “Salsa Only”, we experimented with reduced learning rates ranging from  $1 \times 10^{-3}$  to  $2 \times 10^{-6}$ . We found that a learning rate of  $5 \times 10^{-4}$  resulted in a stable training process for both models, with consistent decreases in validation loss.

Training was conducted for a large initial number of epochs, and we monitored the validation loss throughout. The model checkpoint with the lowest validation loss was selected for evaluation.

The models were implemented using their official repositories when possible to ensure consistency with the original designs [16, 17]. Very minor changes were made to enable training with the datasets at our disposal.

### 5.3.1 Fine-Tuning Details

For fine-tuning, specific layers of each model were trained while others were frozen:

- **BeatNet:** The convolutional layers were frozen, and fine-tuning was applied to the LSTM layers and the final layer.
- **Böck TCN:** The convolutional layers were frozen, allowing fine-tuning of the Temporal Convolutional Network (TCN) layers.

## 5.4 Evaluation Metrics

We used the F-measure [7] as the primary metric to assess beat and downbeat tracking accuracy. For comparison, we include in Table Table 1 the average F-measure accuracy on popular music datasets previously reported for the two prominent beat tracking models we investigate in this study. A standard tolerance window of  $\pm 70$  milliseconds was applied when matching detected beats to the ground truth, accounting for slight timing variations and reflecting human perception of beat alignment.

## 5.5 Creation of the Salsa Dataset

### 5.5.1 Song Selection

The Salsa dataset compiled for this study consists of 40 tracks, selected to capture a diverse range of eras, regional styles, sub-genres, tempos, and instrumentation that characterize Salsa music. These tracks span various origins, with representation from areas such as Puerto Rico, Cuba, and the United States, and include popular sub-genres like Salsa Romántica, Salsa Dura, and Cuban Salsa. Release dates span from the 1970s to the 2020s. Tempos vary from 155 to 246 BPM, with an average around 191 BPM, calculated from the annotated beat intervals.

### 5.5.2 Beat Annotation Process

Beat annotations were created using the *Sonic Visualiser* software [18]. Each beat was manually placed through a combination of visual waveform inspection and auditory analysis to ensure precise timing. Subjective choices were made regarding the inclusion or exclusion of beats in certain sections; for instance, intros and outros with ambiguous or free rhythms were sometimes omitted to maintain annotation consistency.

The Salsa dataset can be accessed publicly via GitHub [github.com/AntoninRap/Salsa-dataset](https://github.com/AntoninRap/Salsa-dataset).

## 6. RESULTS

Our findings reveal that accuracy increases with specialisation and can benefit from the general musical knowledge derived from training on other datasets. This is an unsurprising result to some extent, but it was useful to see that this consistent improvement could be obtained even with a small amount of genre-specific training data.

	BeatNet	BöckTCN
GTZAN	0.806	<b>0.885</b>
Ballroom	N/A	<b>0.962</b>

**Table 1.** Reported average F-measure accuracy on popular music datasets of two prominent beat tracking models. BeatNet did not report any results for the Ballroom dataset.

	BeatNet	BöckTCN
Fine-tuned	0.845	<b>0.771</b>
Salsa only	<b>0.855</b>	0.437
Others (base)	0.560	0.420

**Table 2.** Average beat F-measure accuracy on the Salsa dataset of two prominent beat tracking models under the three training conditions outlined in the Methodology section.

	BeatNet	BöckTCN
Fine-tuned	<b>0.522</b>	<b>0.216</b>
Salsa only	0.516	0.052
Others (base)	0.215	0.042

**Table 3.** Average downbeat F-measure accuracy on the Salsa dataset of two prominent beat tracking models under the three training conditions outlined in the Methodology section.

Tables Table 2 and Table 3 present the average beat and downbeat F-measure accuracies, respectively.

BeatNet achieved its highest F-measure when trained solely on the Salsa dataset (0.855), slightly surpassing its performance when fine-tuned (0.845). The base BeatNet model scored significantly lower (0.560). For BöckTCN, fine-tuning resulted in the highest F-measure (0.771), outperforming the Salsa-only training (0.437) and the base model (0.420). These results suggest that incorporating Salsa data enhances beat tracking performance. BeatNet benefits more from training exclusively on Salsa data, while BöckTCN shows greater improvement through fine-tuning.

Both models exhibited lower F-measure scores for downbeat tracking compared to beat tracking. BeatNet achieved its highest downbeat F-measure when fine-tuned (0.522), closely followed by Salsa-only training (0.516). The base model had a considerably lower score (0.215). BöckTCN’s best downbeat F-measure was 0.216 when fine-tuned; performance decreased with Salsa-only training (0.052) and was minimal for the base model (0.0\*).

Fine-tuning improves downbeat detection, particularly for BeatNet, but overall accuracy remains low. The original BöckTCN Paper does not report any downbeat capabilities or results. These results were obtained using Ben Hayes’s BöckTCN implementation [16]

A closer look into individual beat tracking results reveals that most models obtain higher accuracy on most songs in the dataset after training with Salsa specific data. Specifically, for BeatNet in the "Salsa Only" and "Fine-Tuning" training conditions, a majority of songs achieved higher-than-average F-measure scores compared to the overall average in their respective training conditions. The average accuracy was brought down by a few songs with significantly reduced performance. Interestingly, for both models, these particular songs actually achieved higher results with the base models and present differences in instrumentation compared to the rest of the tracks in the dataset. We explore these findings in more detail in the discussion section below.

## 7. DISCUSSION

This study established a baseline for beat and downbeat tracking in Salsa music using a new, small-scale dataset. Our results show that models trained with Salsa-specific data perform better than those trained on non-Salsa datasets, and, in the case of BeatNet, outperforms its accuracy obtained on Western music genres.

Upon analysing the outlier negative results presented in Table 4, it became apparent that the results were likely due to the difference in instrumentation and lack of strong rhythmic elements rather than the complexity of the rhythms. For instance, in challenging tracks such as “Venenosa”, “Es Tu Mirada” and “Juntando Amores” rhythmic instruments are either less prominent or played more subtly. In “Venenosa”, for example, the rhythm is primarily carried by a soft Tumbao on the conga, while the piano, guitar, bass, and vocals dominate the mix. One thing to note here is that in Salsa, the piano most often functions as a rhythmic instrument, rather than serving a primarily melodic role as it does in many Western genres.

“Es Tu Mirada” is widely enjoyed by Salsa dancers around the world; however, its instrumentation differs from traditional Salsa music and more akin to fusion of Cuban pop with traditional Cuban music elements. In this track, the rhythmic instruments are slightly muted compared to the prominent vocals and bass. Similarly, “Juntando Amores” blends Salsa rhythmic elements with flamenco guitar and is characterized by a very fast tempo. In this track, the rhythmic elements take a backseat to the prominent guitar. These deviations in instrumentation and emphasis on non-traditional elements may have impacted the models’ performance on these three tracks.

The high beat tracking accuracy obtained on most of the dataset suggests that the models effectively specialized in traditional Salsa instrumentation. Training on 32 songs enabled the models to generalize effectively to 8 unseen songs with similar characteristics, and, notably, this generalization occurred consistently across all five folds in our

	Venenosa	Es Tu Mirada	Juntando Amores
BeatNet Fine-tuned	0.378	0.333	0.387
BeatNet Salsa only	0.324	0.249	0.442
BeatNet (others)	0.634	0.660	0.639
BöckTCN Fine-tuned	0.416	0.368	0.514
BöckTCN Salsa only	0.404	0.463	0.465
BöckTCN (others)	<b>0.974</b>	<b>0.662</b>	<b>0.657</b>

**Table 4.** Beat tracking F-measure accuracy on three songs with outlier accuracies for BeatNet and BöckTCN under the three training conditions outlined in the Methodology section.

cross-validation. However, the outlier results highlighted above seem to indicate that the models lost some of their ability to accurately track beats in songs whose instrumentation differed from traditional Salsa arrangements.

For downbeat tracking, the results are more challenging to interpret. The F-measure accuracy varies significantly—from 0 to 1—and it is not clear yet why this variation occurs.

## 8. FUTURE WORK

The promising results obtained during this study with a limited amount of data highlight the need for further research. In following experiments we will focus on leveraging large amount of unannotated data, such as dance videos posted online, through techniques such as self-supervised learning. Through this research, we hope to further improve beat tracking accuracy by exploring multimodal data that integrates visual information from dance movements.

## 9. CONCLUSION

The study demonstrates that fine-tuning beat tracking models with genre-specific data can significantly improve accuracy for Salsa music. It also establishes a baseline for the performance of beat and downbeat tracking on this genre, providing a reference point for the efficacy of more intricate future methodologies. This work contributes to the ongoing efforts to develop beat tracking systems that better account for the rhythmic diversity found in global music genres, and with this objective in mind, introduces a new beat-annotated dataset of Salsa music.



## 10. REFERENCES

- [1] J. Phillips-Silver and L. Trainor, “Feeling the beat: Movement influences infant rhythm perception,” *Science*, vol. 308, no. 5727, p. 1430, 2005.
- [2] S. Böck, Matthew, M. E. P. Davies, and P. Knees, “Multi-task learning of tempo and beat: Learning one to improve the other,” in *The 20th International Society for Music Information Retrieval*, Delft, The Netherlands, 2019, pp. 486–493.
- [3] P. Manuel, “Puerto rican music and cultural identity: Creative appropriation of cuban sources from danza to salsa,” *Ethnomusicology*, vol. 38, no. 2, pp. 249–280, 1994.
- [4] L. Waxer, *The City of Musical Memory: Salsa, Record Grooves, and Popular Culture in Cali, Colombia*. Middletown, CT: Wesleyan University Press, 2002.
- [5] H. Mojtaba, F. Cwitkowitz, and Z. Duan, “Beatnet: A real-time music integrated beat and downbeat tracker,” in *The 22nd International Society for Music Information Retrieval*, Online, 2021, pp. 270–277.
- [6] M. E. P. Davies and S. Böck, “Temporal convolutional networks for musical audio beat tracking,” in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, 2019, pp. 1–5.
- [7] M. Davies, N. DeGara, and M. Plumbley, “Evaluation methods for musical audio beat tracking algorithms,” Centre for Digital Music, Queen Mary University of London, London, UK, Tech. Rep. C4DM-TR-09-06, 2009.
- [8] L. Maia, M. Rocamora, L. Biscainho, and M. Fuentes, “Adapting meter tracking models to latin american music,” in *The 23rd International Society for Music Information Retrieval*, Bengaluru, India, 2022, pp. 3–11.
- [9] D. Fiocchi, F. A. M. Buccoli, M. Zanoni, and A. Sarti, “Beat tracking using recurrent neural network: a transfer learning approach,” in *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018, pp. 1929–1933.
- [10] A. Pinto, J. C. S. Böck, and M. Davies, “User-driven fine-tuning for beat tracking,” *Electronics*, vol. 10, no. 13, pp. 10–13, June 2021.
- [11] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [12] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, “An experimental comparison of audio tempo induction algorithms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1832–1844, 2006.
- [13] A. Holzapfel, M. Davies, J. Zapata, J. Oliveira, and F. Gouyon, “Selective sampling for beat tracking evaluation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2539–2548, 2012.
- [14] M. E. P. Davies and M. Plumbley, “Context-dependent beat tracking of musical audio,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009–1020, 2007.
- [15] T. de Clerq and D. Temperley, “A corpus analysis of rock harmony,” *Popular Music*, vol. 30, no. 1, pp. 47–70, 2011.
- [16] H. Mojtaba. (2021) Beatnet. [Online]. Available: <https://github.com/mjhydri/BeatNet>
- [17] B. Hayes. (2020) Beat tracking tcn. [Online]. Available: <https://github.com/ben-hayes/beat-tracking-tcn>
- [18] C. Cannam, C. Landone, and M. Sandler, “Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files,” in *Proceedings of the ACM Multimedia 2010 International Conference*, 2010, pp. 1467–1468.

# MUSIC SOURCE SEPARATION IN NOISY BRAZILIAN CHORO RECORDINGS

Pedro Donadio<sup>1,2</sup>

Martín Rocamora<sup>3,4</sup>

Luiz Wagner Biscainho<sup>2</sup>

<sup>1</sup> Department of Electronics and Computation, Universidade Federal do Amazonas, Brazil

<sup>2</sup> Universidade Federal do Rio de Janeiro, Brazil

<sup>3</sup> Music Technology Group, Universitat Pompeu Fabra, Spain

<sup>4</sup> Facultad de Ingeniería, Universidad de la República, Uruguay

pedro.donadio@smt.ufrj.br, rocamora@fing.edu.uy, wagner@smt.ufrj.br

## ABSTRACT

Choro music is considered the first musical style to originate in Brazil, dating back to the 1870s. Some historical recordings from the early 20th century include noise inherent to the process of recording and playing shellac records. In this work, we investigate the instrument separation task applied to historical recordings of this Brazilian music genre, using models originally trained on clean tracks. We used a choro dataset composed of modern recordings of songs from the most important composers of this style, and a 78 RPM (rotations per minute) noise dataset to emulate old choro records. Using an available neural network architecture — Hybrid Demucs — trained to separate the characteristic choro musical instruments into the string, wind, and percussion families without background noise, we evaluate the separation result in the presence of different types of 78 RPM noise. Furthermore, we study the impact of the additive noise on separation when the signal-to-noise ratio (SNR) ranges from 10 to 40 dB. The experiments showcase that the model is robust, although the performance depends on the type and level of noise.

## 1. INTRODUCTION

The task of music source separation consists in isolating the sound of each instrument, or a family of instruments, from an audio mixture (i.e., a track containing various instruments playing together) [1–7]. In recent years, deep learning approaches have achieved state-of-the-art performance in music source separation. Various neural network architectures have been explored for this task, as demonstrated in previous works [8–12]. We chose to base the investigation of this work on the Hybrid Demucs model [10], which uses an encoder-decoder architecture, due to its excellent performance in instrument separation and the availability of a model we trained on a dataset of choro music within our current research.

In many contexts, isolating the singer’s voice or even erasing the sound of a specific instrument from a musical excerpt (preserving the rest of the mixture) are practical tasks that may help professional and amateur musicians, music students, audio engineers, musicologists, and researchers. However, most of the works in this area explore separation on modern recordings, obtained in controlled studio environments, such as those found in MUSDB [13] database. The open question of our interest is the effectiveness of these models on historical recordings, i.e., data extracted from discs where noise is an inherent part of the recording/reproduction process.

In this work, we explore the traditional musical genre called *choro*, which, along with samba, is one of the most representative elements of Brazilian and Latin American culture. Historical recordings of choro [14, 15] date back to the beginning of the 20th century and typically involve groups where a rhythmic and harmonic base of 6-string guitars, 7-string guitars, *cavaquinho*, and *pandeiro* accompanies soloists playing flute, mandolin, clarinet, or *cavaquinho*. This music is instrumental in its conception (although some choros have received lyrics later) and is formed through the fusion of various rhythms performed in a characteristic choro way, such as polka, baião (a rhythm from northeast of Brazil), waltz, and maxixe (also known as the Brazilian Tango). It results from a mixture of the African rhythm lundu with European genres. All these aspects introduce different types of challenges. Regarding rhythm, there are various notions of meter, which may vary between 2/4, 3/4, or 4/4. In melodic terms, the improvisational nature of choro allows the soloist freedom to create melodies in certain musical passages. From a technical perspective, the timbral overlap of various instruments increases the difficulty of separation. The main composers of the genre are Ernesto Nazareth (1863-1934), Heitor Villa-Lobos (1887-1959), Pixinguinha (1897-1973), Jacob do Bandolim (1918-1969) and Waldir Azevedo (1923-1980). It is interesting to note that, in most cases, the recordings have the composers themselves performing as soloists. For example, Jacob do Bandolim, who was an excellent mandolinist; Waldir Azevedo, credited with introducing the *cavaquinho* as a soloist instrument in choro groups; and Pixinguinha, who played both the flute and saxophone, and is also considered one of greatest musicians in Brazil.



© P. Donadio, M. Rocamora, and L.W. Biscainho. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** P. Donadio, M. Rocamora, and L.W. Biscainho, “Music source separation in noisy Brazilian choro recordings”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.

This study investigates the impact of background noise present in the original mix on the performance of a musical instrument family separation system. We utilize a Hybrid Demucs model specifically trained on a choro music dataset composed of 10 albums of modern recordings from the most prominent composers of the genre. In its original form, the model is capable of isolating instruments into three families: strings, wind, and percussion. We conduct a series of experiments by adding different levels and types of 78 RPM noise to the mixtures, simulating the characteristics of old shellac recordings. The noisy mixtures are then separated by the model and compared to the corresponding results obtained from clean (non-noisy) mixtures. In both cases, the separation performance is evaluated according to the most usual objective metrics adopted for source separation [16] — signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) — supplemented by informal but careful subjective listening evaluation. The results demonstrate that the model is robust and can effectively separate instrument families under most conditions, although the performance depends on the type and level of noise.

## 2. METHODS

### 2.1 The Model

Hybrid Demucs<sup>1</sup> (v3) is a neural network that utilizes convolutional layers as encoder/decoder schema. Its main highlight is the fact that it receives the input signal simultaneously in the time and frequency domains, subjecting it to 6 encoder layers and 6 decoder layers. The model was trained for 360 epochs with a batch size equal to 4, using Adam optimizer and a learning rate  $Lr = 10^{-4}$ .

The available model has been trained (from scratch) on a choro music dataset that consists of 10 albums for training and 2 albums for validation<sup>2</sup>. Table 1 presents the number of songs per album utilized for training, validation and testing the model.

The data structure for each song includes a mixture (with all instruments playing together) and 4 sources: the string family (6-string guitar, 7-string guitar, *cavaquinho*, mandolin, etc.), the wind family (flute, saxophone, clarinet, etc.), the percussion family (*pandeiro*, *reco-reco*, *tamborim*, snare drums, etc.), and the “others” family (consisting of instruments that do not belong to the other three classes). All tracks have a sampling rate of 48 kHz.

In a future publication, we will provide a detailed presentation of the choro dataset, outlining all stages of track production for each instrument family and their specific characteristics, as well as details of the training process.

### 2.2 The noise dataset

The noise dataset described in [17] comprises various noise segments extracted from recordings of 78 RPM shellac

**Table 1:** List of songbooks in *Choro* dataset.

Number	Songbook	# Songs
1	Altamiro Carrilho	13
2	Benedicto Lacerda	12
3	Chiquinha Gonzaga	12
4	Choro Meets Bach	14
5	Ernesto Nazaré 1	11
6	Ernesto Nazaré 2	11
7	Ernesto Nazaré 3	11
8	Inéditos	14
9	Jacob do Bandolim 1	12
10	Jacob do Bandolim 2	12
11	Pixinguinha	12
12	Roda de Choro	12
13	Severino Araújo	12
14	Waldir Azevedo 1	12
15	Waldir Azevedo 2	12
16	Zequinha da Abreu	12

records. These include electrical circuit noise, ambient noise, noise caused by the turntable, and clicks. From this set, we choose 5 noise samples with different characteristics to simulate historical choro recordings; Figure 1 shows their respective spectrograms. To make the tracks from the noise dataset compatible with the clean mixtures from the choro dataset, the noise tracks have been pre-processed: we took only one of the two original stereo channels, re-sampled it from 44.1 to 48 kHz (to match the sampling rate of the choro dataset), periodically looped the noise in cycles consistent with a 78 RPM recording, and ensured that both the noise track and the musical track were the same length. The original noise track titles (with references to various metadata) in the original dataset are extremely long and were shortened to *cristree*, *vucchella*, *ma-journey*, *springfield*, and *alacarte* for reference. The samples used in this study are available upon request.

## 3. EXPERIMENTS

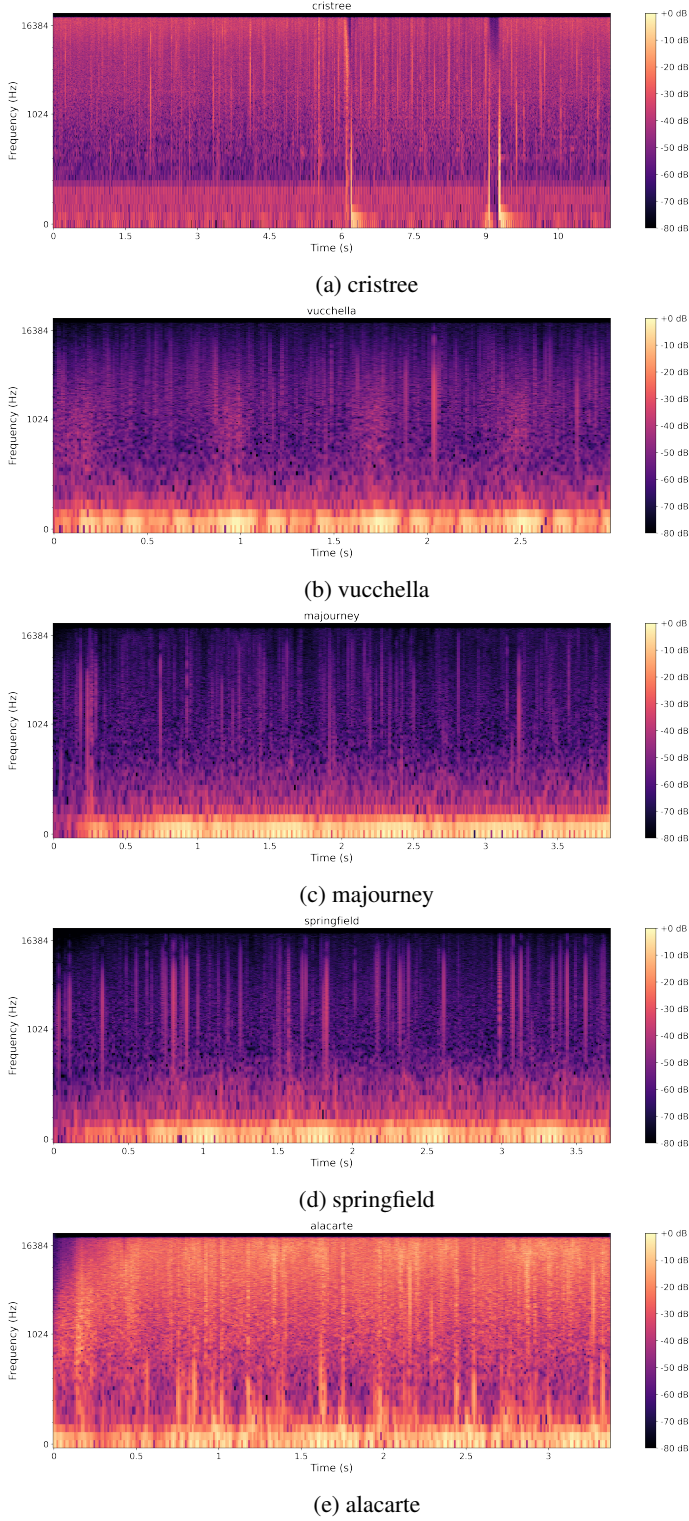
The experiments were conducted to test the separation models on simulated historical choro recordings. To do that, 20 tracks were carefully selected from the test set of the choro dataset to ensure they were free of leakage. One of the typical characteristics of choro is the rich contrapuntal interaction between performers, which is facilitated and spontaneously captured in studio if the musicians play in the same space. This usually results in some sound leakage from one instrument (or family of instruments) into the microphone of another. Of course, this effect would prevent the proper use of the original segregated tracks as reference signals for evaluating system performance.

To simulate historical 78 RPM noisy recordings, we combine the clean tracks of choro test set and the 5 pre-conditioned noise tracks according to

$$C_{\text{noisy}} = \beta \cdot (G \cdot N_{pp} + C_{\text{clean}}), \quad (1)$$

<sup>1</sup> Available at <https://github.com/facebookresearch/demucs/tree/v3>

<sup>2</sup> The albums mentioned are available for purchase at <https://www.choromusic.com/>



**Figure 1:** Spectrogram of the five noise samples selected.

where  $\beta$  represents the global track gain,  $G$  is the gain applied to the pre-processed noise track  $N_{pp}$  to produce the desired SNR, and  $C_{\text{clean}}$  is the track extracted from choro test set. Noise gains were adjusted to induce SNR values of 10, 20, 30, and 40 dB for each track. In total, 400 noisy mixtures were generated according to this scheme, outlined in Figure 2. It should be noticed that the choro dataset albums used for training were not subjected to any

pre-selection regarding instrument leakage.

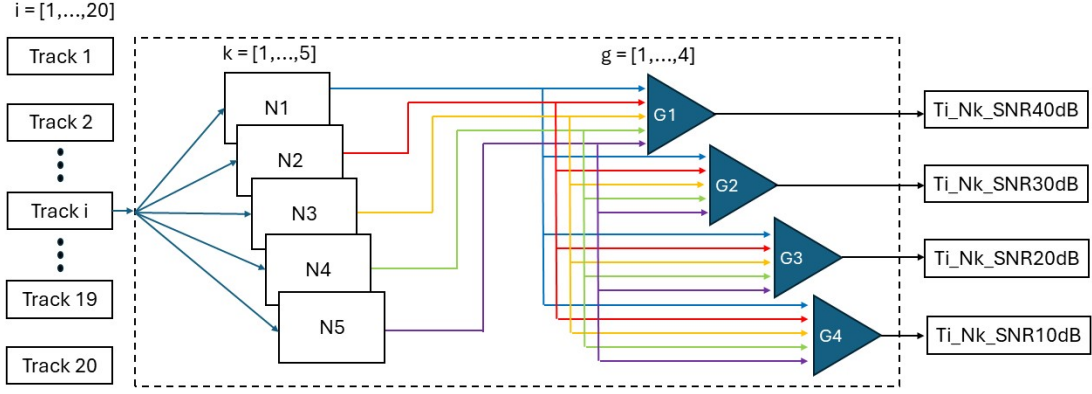
Another key aspect to highlight is the choice to separate simulated historical tracks instead of original recordings from the early 20th century, which was based on two main considerations. First, we aim to test the robustness of the model, i.e., its ability to accurately separate instrument families in the presence of background noise. To achieve this, it is essential to vary the types of noise both in terms of their spectral components and their power levels—an approach that would be impractical with real recordings, given the limited availability of such material. Secondly, to calculate the objective measures commonly used in the source separation community, it is essential to compare the separated track with the ground truth, i.e., the isolated recording of the instrument without noise. Clearly, acquiring such material from original recordings of that era is practically unfeasible, as technology for multichannel recording per instrument was not yet available.

To assess the quality of separation, three objective metrics are widely adopted in the literature: signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR), defined in [16]. Table 2 presents these metrics computed to compare the performances for different levels of SNR. The average value is calculated considering all the test tracks and all selected noises. It can be observed that for SNRs of 20, 30, and 40 dB, the separation results are closely aligned with those obtained from the clean tracks in the last row. This indicates a degree of robustness in the system when processing noisy recordings. Table 3 addresses the results obtained exclusively for the SNR of 10 dB. For this case of worst performance, it is possible to observe the behavior of the separation for each noise individually. As the values of the metrics SDR, SIR, and SAR metrics decrease, respectively, the signal tends to become almost imperceptible, presents a higher amount of artifacts that mask the original signal, and the separation is affected by interference from other families, indicating that the system struggles to differentiate them.

Overall, the separation works well for recordings with noise. From an auditory perspective, the instrument families are separated in a similar way as in the cases of noise-free tracks for almost all noise types and SNR values, with the percussion family being the most critical case for low SNR values.

Lower SNRs tend to impair the separation quality, particularly for the percussion family. We verified auditorily that in certain cases, the nature of the noise is very similar to that of the *pandeiro* (the primary percussive instrument in the database), especially due to the high-frequency sound produced by the *platinelas* (small iron plates). This is predominantly observed in the cristree and alacarte noise signals, which contain a considerable amount of information above 512 Hz. In contrast, the other noise signals do not exhibit this characteristic with the same intensity. Reflecting the low values in SDR and SAR, a similar behavior was found through perceptual evaluation.

As for the string family, separation is quite reasonable



**Figure 2:** Scheme for preparing the noisy tracks of the test set.

**Table 2:** Average value of SDR, SIR and SAR for each SNR (considering the 20 noisy tracks in the test set).

Noise SNR	STRINGS			WIND			PERCUSSION		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
20 dB	14.210	28.248	16.104	14.134	24.596	15.590	5.657	16.918	5.234
30 dB	15.777	28.298	17.878	15.365	25.130	17.112	7.176	19.231	7.380
40 dB	16.128	28.230	18.503	15.582	25.280	17.414	7.453	19.452	8.233
Clean	16.130	28.373	18.399	15.700	25.407	17.576	7.6172	19.340	8.334

**Table 3:** Average value of SDR, SIR and SAR computed for the 20 noisy tracks in test set with an SNR of 10 dB.

Noise category	STRINGS			WIND			PERCUSSION		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
alacarte	12.749	27.398	14.309	14.169	24.699	15.617	-0.016	6.042	-6.685
cristree	11.880	27.291	14.128	14.035	25.083	16.023	0.597	7.475	-6.396
majourney	6.264	27.656	9.166	9.545	21.626	10.154	6.373	17.835	6.850
springfield	5.663	27.465	8.090	9.935	21.005	10.774	6.077	17.126	6.395
vucchella	5.102	26.898	7.091	10.484	20.101	11.447	5.631	16.241	6.122

with an SNR of 10 dB with the alacarte and cristree noises. This occurs mainly because, for these two cases, the model extracts some of the noise from the string family and classifies it as percussion, thus favoring the string metric, at the cost of impairing the percussion metric.

In the case of the wind family, results for SNR of 30 and 40 dB show little variation for SDR, SAR e SIR values, indicating that separation is easier for this family. In many tracks, the system acts as a form of denoiser, partially eliminating noise, particularly during pauses in the melody played by the wind instrument.

#### 4. CONCLUSIONS AND FUTURE WORK

In this work, we investigate the effects of separating musical instruments into families on historical recordings of the traditional Brazilian genre choro. To achieve this, we employ a deep learning approach using an architecture known as Hybrid Demucs, trained on a choro database. We use 20 leakage-free test tracks combined with a database of noise from 78 RPM records to simulate historical record-

ings, creating different levels of SNR (10, 20, 30, and 40 dB). Finally, we evaluate the results using traditional objective metrics for music separation (SDR, SIR, and SAR), in addition to a careful consideration of our own subjective evaluation.

The results obtained are promising, demonstrating that the system is robust when dealing with tracks containing additive noise, even though it has been pre-trained on clean recordings. Some families, such as percussion, face greater challenges in separation at lower SNRs, while others, such as wind instruments, show good results across all SNR levels. Some separation results, as well as the noises used, are available for listening at [https://www02.smt.ufrj.br/~pedro.donadio/index\\_Lamir.html](https://www02.smt.ufrj.br/~pedro.donadio/index_Lamir.html).

Several ideas for future work arise from this approach, with the main one being fine-tuning using tracks with noise. The expectation is that, in addition to separating the families, a form of denoising will occur if the noise is included as training data for the model.



## 5. REFERENCES

- [1] Y. Özer and M. Müller, “Source separation of piano concertos with test-time adaptation,” *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 493–500, June 2020.
- [2] S. Sarkar, E. Benetos, and M. Sandler, “EnsembleSet: A new high quality synthesised dataset for chamber ensemble separation,” *23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.
- [3] F. Stöter, A. Liutkus, and N. Ito, “The 2018 signal separation evaluation campaign,” *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, p. 293–305, 2018.
- [4] G. Fabbro, “The sound demixing challenge 2023 – music demixing track,” *Transactions of the International Society for Music Information Retrieval*, vol. 7, no. 1, p. 63–84, 2024.
- [5] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F. Stöter, “Musical source separation: An introduction,” *IEEE Signal Process. Mag.*, vol. 36, no. 1, p. 31–40, Jan. 2019.
- [6] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, “An overview of lead and accompaniment separation in music,” *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 8, p. 1307–1335, Aug. 2018.
- [7] P. Patel, S. Shah, S. Prasad, A. Gada, K. Bhowmick, and M. Narvekar, “Audio separation and classification of indian classical instruments,” *Eng. Appl. Artif. Intell.*, vol. 133, 2024.
- [8] F. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, “Open-unmix - a reference implementation for music source separation,” *Journal of Open Source Software*, 2019. [Online]. Available: <https://doi.org/10.21105/joss.01667>
- [9] R. Hennequin, A. Khelif, F. Voituret, and M. Moussallam, “Spleeter: a fast and efficient music source separation tool with pre-trained models,” *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 6 2020.
- [10] A. Défossez, “Hybrid spectrogram and waveform source separation,” in *ISMIR 2021, Music Demixing Workshop (DMX)*, 11 2021.
- [11] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 8 2019, pp. 1256–1266.
- [12] —, “Tasnet: Time-domain audio separation network for real-time, single-channel speech separation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, 4 2018, pp. 696–700.
- [13] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “The MUSDB18 corpus for music separation,” december 2017. [Online]. Available: <https://doi.org/10.5281/zenodo.1117372>
- [14] J. B. Siqueira, *Três vultos históricos da música brasileira: Mesquita - Callado - Anacleto*. Rio de Janeiro: FUNARTE, 1970.
- [15] A. Diniz, *Almanaque do Choro*, 3rd ed. Rio de Janeiro: Zahar, 2003.
- [16] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers*, vol. 14, p. 1462–1469, 6 2006.
- [17] E. Moliner and V. Välimäki, “A two-stage u-net for high-fidelity denoising of historical recordings,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 841–845.

# TEMPO ESTIMATION USING COMBINED MEL-SPECTROGRAM AND MEL-SCALOGRAM INPUTS

Luiz A. G. Viana<sup>1</sup>   Antonio C. L. Fernandes Júnior<sup>1</sup>   Eduardo F. de Simas Filho<sup>1</sup>

<sup>1</sup> Departamento de Engenharia Elétrica e de Computação (DEEC), Universidade Federal da Bahia

luiz.guimaraes@ufba.br, antonio.lopes@ufba.br, eduardo.simas@ufba.br

## ABSTRACT

This paper proposes a novel method for musical tempo estimation by combining mel-spectrograms and mel-scalograms in a custom three-dimensional image format. The proposed model consists of a convolutional neural network trained with these images, correlating them with predefined tempo values. The model is trained and evaluated on widely used databases in the literature and it is compared to the state-of-the-art. Data augmentation was iteratively applied during training. The combination of mel-spectrogram and mel-scalogram data resulted in an accuracy 2 improvement. Future work may explore the combination of additional audio signal representation methods.

## 1. INTRODUCTION

Musical tempo can be understood as the speed at which a musical piece is performed, typically measured in BPM (beats per minute), and its estimation is one of the fundamental tasks in Music Information Retrieval (MIR) [1]. Determining the tempo of a musical piece allows for possibilities such as automatic classification and even automation of musical accompaniment for live performances. Since the early research of [2] and [3], the MIR community has been conducting researches on tempo estimation for the past 25 years [4].

Although the works of [1] and [5] have achieved excellent results, the problem of musical tempo estimation is still open [4]. Aspects such as evaluation metrics focused on real-world applications, audio transformations for relevant feature extraction, and novel machine-learning methods are still under investigation. Recent studies have shown significant progress in the field of musical tempo estimation. Notably, works such as [6, 7] have achieved promising results using different neural network architectures, specifically Bidirectional Recurrent Neural Networks and Multi-scale Grouped Attention Networks. Additionally, several studies have explored self-supervised approaches, with works like [8–11] achieving results com-

parable to state-of-the-art methods, despite the challenges posed by the scarcity of labeled examples. More recently, the works [12, 13] have improved results in some datasets, surpassing the state-of-the-art in specific cases by utilizing Temporal Convolutional Networks (TCNs) trained to jointly estimate beat, downbeat, and tempo [9, 14].

In this context, this work aims to contribute to the musical tempo estimation problem by introducing a novel representation of audio signals through the creation of a customized three-dimensional image. This image combines a mel-spectrogram in red tones and a mel-scalogram in green tones, resulting in a unique three-dimensional representation. The employed database is composed of a combination of public music signal datasets with tempo annotations. The generated images are fed into a convolutional neural network (CNN). Given the limited number of examples available in public datasets, an online data augmentation procedure was proposed to modify real examples and generate new synthetic ones. Finally, the 5-fold cross-validation method was used to enhance the statistical reliability of the proposed model. Our results are compatible with state-of-the-art algorithms.

## 2. PROPOSED MODEL

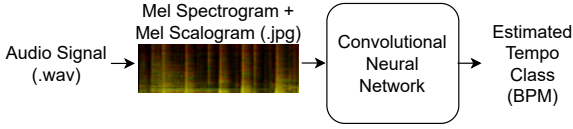
The proposed model for tempo estimation involves generating a custom three-dimensional image from an audio signal of a musical piece with a predefined BPM tempo. This image is created by combining a mel-spectrogram in red tones and a mel-scalogram in green tones. The objective is to perform supervised learning by training a convolutional neural network with the generated images, as illustrated in Figure 1. Intuitively, the tempo estimation problem appears to be a regression problem for an integer value. However, following the approach proposed in [5], this work treats it as a classification problem, assigning a class for each integer BPM value within a specific BPM range.

### 2.1 Used Datasets

For the model to be capable of generalizing the musical tempo estimation problem, it must be trained with datasets that contain examples from various musical styles and different tempo classes. The datasets chosen for this purpose are LMD Tempo (3611 examples), MTG Tempo (1159 examples), and Extended Ballroom (3826 examples). These databases are studied and referenced in [4].



© Luiz Alberto G. Viana, Antonio Carlos L. Fernandes Júnior, Eduardo F. de Simas Filho. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Luiz Alberto G. Viana, Antonio Carlos L. Fernandes Júnior, Eduardo F. de Simas Filho, “Tempo Estimation using Combined Mel-Spectrogram and Mel-Scalogram Inputs”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.



**Figure 1.** Simplified diagram of the proposed model. The audio signal is transformed into a custom three-dimensional image combining a mel-spectrogram and a mel-scalogram. This image is used to train the convolutional neural network, which will estimate the musical tempo value in BPM.

The datasets chosen for model evaluation are widely used in the literature. This allows us to compare the results achieved with other publications. Naturally, examples used in training will not be included in the evaluation sets. The following datasets were utilized: ACM Mirum (1410 examples), Ballroom (698 examples), GiantSteps Tempo (660 examples), GTZAN (999 examples), Hainsworth (222 examples), ISMIR2004 (465 examples), and SMC Mirum (217 examples) [4]. These datasets consist of musical pieces in *wav* files, each with a predefined tempo in BPM.

We chose to focus the training on the range where most examples occur, resulting in classes between 60 and 199 BPM, reducing it to 140 different classes. This range is referred to as the “sweet octave” by some authors [15], as it tends to contain more musical pieces than any other interval.

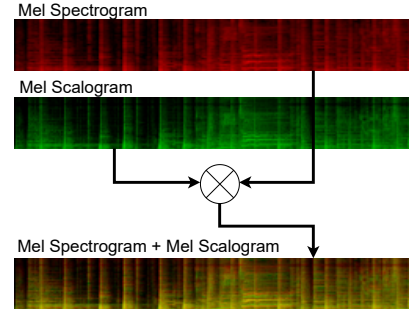
## 2.2 Audio Signal Representation as Image

Whenever deep learning algorithms are used to solve problems involving audio signals, signal representation becomes a crucial aspect to consider. In this work, an offset of 5 seconds was applied to generate the representations, meaning that the first 5 seconds of the audio signal were discarded. This step is necessary because the initial seconds of a musical piece often exhibit different tempo values compared to the overall excerpt. Subsequently, the signals were converted to mono (downmixing) and downsampled to 11025 Hz, which is sufficient to detect tempos up to more than 646 BPM [5]. Given that musical tempo is not an instantaneous feature, the representation must encompass a sufficient time span. Therefore, a duration of 11.888 seconds was selected, ensuring that the length of the audio vector is represented in base 2, optimizing computational efficiency. After downsampling, the resulting audio vector contains 131072 samples.

The final custom image has dimensions of (256, 40, 2), representing two matrices of size 256 by 40, where the first matrix corresponds to the mel-spectrogram and the second to the mel-scalogram, as illustrated in Figure 2.

### 2.2.1 Mel-Spectrogram

The mel-spectrogram combines the Short-Time Fourier Transform (STFT) with a frequency-to-Mel scale conversion, creating a more perceptually relevant representation



**Figure 2.** Diagram illustrating the creation of the custom input image for the CNN. The result is a custom image with dimensions (256, 40, 2) that serves as input to the convolutional neural network.

of audio. This method was chosen for its widespread application in visual audio representation, providing a basis for comparative analysis. The parameters for the mel STFT, such as frame length, hop size, and the use of 128 Mel bands, directly influence the time and frequency resolution of the spectrogram [16]. In this study, a frame length of 2048 samples and a hop size of 512 samples were employed. This configuration ensures a balance between time and frequency resolution, suitable for the task of tempo estimation. Figure 3a illustrates a mel-spectrogram generated from the processed audio signal.

### 2.2.2 Mel-Scalogram

Wavelets have been widely used in previous studies, such as [17], and have proven to be an effective technique for detecting events over time. The wavelet scalogram may be generated from the pre-processed audio signal vector by initially applying the Continuous Wavelet Transform (CWT). Given a signal  $f(t)$ , its CWT is defined as follows:

$$\mathcal{W}_f^\psi(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi^* \left( \frac{t - \tau}{a} \right) dt \quad (1)$$

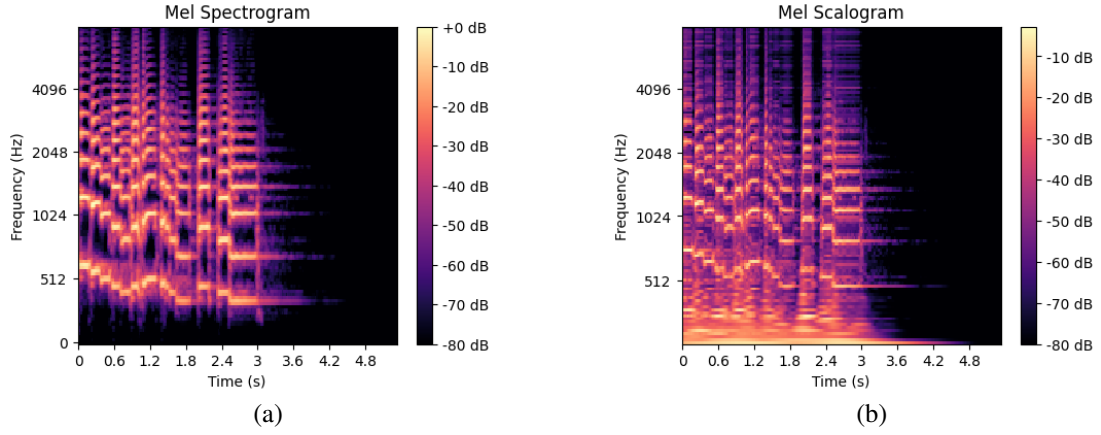
where the parameter  $a$  ( $>0$ ) refers to the scale, and  $\tau$  represents the translation or location of the mother wavelet function  $\psi(t)$ . Both  $a$  and  $\tau \in \mathbb{R}$ . The parameter  $a$  controls the dilation/contraction of the mother wavelet function. The superscript asterisk in  $\psi^*(\cdot)$  denotes the complex conjugate of the function  $\psi(\cdot)$ , and  $\mathcal{W}_f^\psi(a, \tau)$  is known as the wavelet coefficient [18].

Several mother wavelet functions can be used in signal analysis. It is still unknown which continuous wavelet function performs better for the musical tempo estimation problem. In this study, the complex Morlet wavelet was chosen after visually analyzing the mel-scalogram generated by different wavelet functions. The complex Morlet wavelet is defined as:

$$\psi(t) = \frac{1}{\sqrt{\pi B}} e^{-\frac{t^2}{B}} e^{j2\pi C t} \quad (2)$$

with  $B = 6$  and  $C = 6$ , where  $B$  is the bandwidth and  $C$  is the center frequency. The wavelet scale parameter  $a$





**Figure 3.** (a) Mel-Spectrogram and (b) Mel-Scalogram generated from the "trumpet" audio example provided by [19].

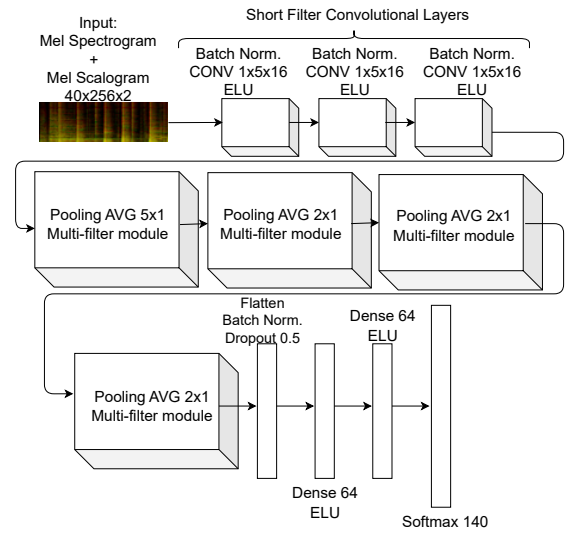
was selected to represent the Mel frequencies, using 128 bands to maintain consistency with the mel-spectrogram generated. A mel-scalogram generated using this approach can be observed in Figure 3b.

### 2.3 Convolutional Neural Network: Proposed Architecture and Training Procedure

The CNN architecture employed by [5] was utilized. This network achieved good results using mel-spectrograms, suggesting it might also perform well when trained with the custom image representation. The distinctive feature of this architecture is that all convolutions are of the “same” type, meaning padding is employed to maintain the image’s dimensions with a stride of one. As the filters possess unit dimensions along the vertical axis, the tensor’s shape remains unchanged along the time axis, which is crucial for tempo detection. The CNN architecture used may be observed in Figure 4. The input is a custom image with dimensions (40,256,2), where the first channel corresponds to the mel-spectrogram and the second to the mel-scalogram. Following the input layer are three sequential convolutional layers with short filters, specifically (1,5,16), and Exponential Linear Unit (ELU) activation functions. These layers are inspired by the traditional approach of creating an Onset Strength Signal (OSS) and then analyzing its periodicity [5]. All layers in the network begin with batch normalization, which is crucial for stabilizing the learning process by normalizing each layer’s inputs.

After the initial convolutional layers, multi-filter modules are employed, whose structure is shown in Figure 5. These modules aim to reduce dimensionality along the scale axis, summarizing the information and combining the signal with a variety of filters capable of detecting temporal dependencies [5].

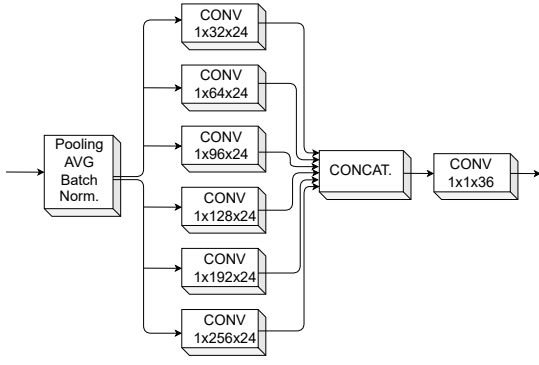
The training datasets are unified and randomly divided into five parts, for use in 5-fold cross-validation. Of these five parts, four are used for training, and one part is split between validation and testing, resulting in 80% for training, 10% for validation, and 10% for testing. All tensors are normalized using the mean and standard deviation before



**Figure 4.** Convolutional Neural Network Architecture diagram. Adapted from [5]. It consists of an input layer followed by three convolutional layers with short filters and four multi-filter modules. Finally, the tensor is flattened and connected to two fully connected layers. The output layer is a softmax with 140 classes.

model training begins, and a data augmentation strategy was employed by iteratively compressing and expanding the custom images along the horizontal axis while keeping the vertical axis unchanged. The modification factor is randomly chosen from a predefined set of values  $F_a \in \{0.8, 0.85, 0.9, 0.95, 1, 1.05, 1.1, 1.15, 1.2\}$ . All data undergo this data augmentation process.

Evaluating the performance of tempo estimation models has a particular challenge. Accuracy 0 is defined as the real accuracy of the model when the convolutional neural network can predict the exact tempo ( $\Gamma$ ) of the musical piece,  $\hat{\Gamma} = \Gamma$ . Accuracy 1 considers values within a 4% precision window,  $\hat{\Gamma} = \Gamma \pm 4\%$ . This criterion accounts for the fact that this minimal difference is imperceptible to the human



**Figure 5.** Multi-Filter Module diagram. Its main feature is parallel convolutions with different filter dimensions. All activation functions are ELU. Adapted from [5].

ear, and even well-trained individuals may estimate tempos within the same margin of error. Finally, Accuracy 2 also considers the submultiples ( $1/2$  and  $1/3$ ) and multiples (2 and 3) of the actual tempo value, within a 4% precision window,  $\hat{\Gamma} = (\Gamma \pm 4\%) M$ , where  $M \in \frac{1}{2}, 1, 2, 3$ . These are known as “octave errors”.

### 3. RESULTS

#### 3.1 Comparison between Mel-Spectrograms and Mel-Scalograms

Using only the training datasets, we applied the 5-fold cross-validation method to evaluate the performance of mel-spectrograms, mel-scalograms, and the custom image that combines both. Table 1 presents the Accuracy 2 values for the validation and test sets, allowing for a clear comparison between the different audio representations.

**Table 1.** Comparison of Accuracy 2 between Mel-Scalograms and Mel-Spectrograms

Experiment	Validation (%)	Test (%)
Mel-Scalogram	$93.5 \pm 0.3$	$90.4 \pm 1.1$
Mel-Spectrogram	$94.1 \pm 0.4$	$91.2 \pm 0.8$
Mel-Spec+Scal	$94.2 \pm 0.6$	$92.0 \pm 0.7$

The results show that the mel-spectrogram alone achieved an accuracy of 94.1% on the validation set and 91.2% on the test set. The mel-scalogram, on the other hand, reached 93.5% on validation and 90.4% on the test set, showing slightly lower performance compared to the mel-spectrogram. However, when both are combined to create a custom image, the model’s performance improves further, achieving an accuracy of 94.2% on validation and 92.0% on the test set.

These results suggest that incorporating the mel-scalogram to complement the mel-spectrogram provides a richer and more detailed representation of the audio signal, contributing to improved tempo estimation. The combination of the two representations appears to better capture relevant temporal and spectral characteristics, resulting in greater accuracy in musical tempo prediction.

#### 3.2 Evaluation of the Proposed Model and Comparison with the State-of-the-art

In 2020, [4] highlighted the works of [5] and [1] as foundational research in musical tempo estimation. More recently, the works [12, 13] have further refined these results, surpassing the state-of-the-art in specific datasets through the use of Temporal Convolutional Networks (TCNs) trained for joint beat, downbeat, and tempo estimation [9, 14].

Table 2 presents a comparison between these state-of-the-art models and the proposed model, which utilizes a custom image combining the mel-spectrogram and mel-scalogram. Note that [13] did not report results for all datasets, so only the published results are included.

From the table, it is evident that the proposed model achieves results very close to the current state-of-the-art, even surpassing some models in specific cases. For instance, while the model excels on the GiantSteps dataset, achieving an accuracy of 93.4%, it slightly underperforms on datasets like Ballroom and GTZAN compared to the most recent state-of-the-art. These results suggest that the incorporation of the mel-scalogram enhances the representation, although there remains room for improvement in certain contexts.

**Table 2.** Comparison with the State-of-the-Art - Accuracy 2 (%)

Evaluation Datasets	Schr [5]	Böck [1]	Böck [13]	Mel-Spec+Scal
ACM Mirum	97,4	97,7	97,7	97,4
Ballroom	98,4	98,7	-	94,8
GiantSteps	89,3	86,4	95,8	93,4
GTZAN	92,6	95,0	93,9	92,2
Hainsworth	84,2	89,2	-	84,2
ISMIR04	92,2	95,0	-	89,2
SMC	50,2	67,3	-	52,5
Combined	92,1	93,6	-	91,4

### 4. CONCLUSION

This study aimed to introduce a new representation layer in the image of the audio signal. We observe that the mel-scalogram, generated from the CWT, improved the model’s performance when evaluating accuracy 2. By simply altering the representation of the audio signal while maintaining the same network architecture, the accuracy 2 values improved compared to those achieved using only the mel-spectrogram. We can conclude that the mel-scalogram captures characteristics that the mel-spectrogram does not, which are crucial for musical tempo estimation.

The proposed model achieved results comparable to the state-of-the-art, surpassing some models in specific databases, such as GiantSteps. Future work could explore other image generation methods incorporated into the model’s input as additional layers. Other possibilities include the use of alternative CNN architectures and improvements in data augmentation methods.

## 5. ACKNOWLEDGMENTS

The authors would like to thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for their support.

## 6. REFERENCES

- [1] S. Böck, F. Krebs, G. Widmer. "Accurate Tempo Estimation Based on Recurrent Neural Networks and Resonating Comb Filter". *16th ISMIR Conference*, pp. 486-493, 2015.
- [2] Masataka Goto and Yoichi Muraoka. "A Beat Tracking System for Acoustic Signals of Music". In *Proceedings of the Second ACM International Conference on Multimedia* pp. 365-372, 1994. doi: 10.1145/192593.192700
- [3] Eric D. Scheirer. "Tempo and Beat Analysis of Acoustic Musical Signals". *The Journal of the Acoustical Society of America*, 103(1): 588-601, 1998. doi: 10.1121/1.421129.
- [4] Hendrik Schreiber, Julián Urbano, Meinard Müller. "Music Tempo Estimation: Are We Done Yet?". *Transactions of the ISMIR*, vol. 3, pp. 111, 2020. doi: 10.5334/tismir.43
- [5] Hendrik Schreiber, Meinard Müller. "A Single-step Approach to Musical Tempo Estimation Using a Convolutional Neural Network". *19th ISMIR Conference*, pp. 98-105, 2018.
- [6] Mila Souza, Pedro Moura, and Jean-Pierre Briot. "Tempo Estimation Via Neural Networks - A Comparative Analysis". In *Proceedings of 18th Brazilian Symposium on Computer Music*, 2021. doi: 10.5753/sbcm.2021.19420.
- [7] Xiaoheng Sun, Qiqi He, Yongwei Gao, and Wei Li. "Musical Tempo Estimation Using a Multi-scale Network". *22nd ISMIR Conference*, 2021. doi: 10.48550/ARXIV.2109.01607.
- [8] Elio Quinton. "Equivariant Self-Supervision for Musical Tempo Estimation". *23rd ISMIR Conference*, 2022. doi: 10.48550/arXiv.2209.01478.
- [9] Antonin Gagnere, Slim Essid, Geoffroy Peeters. "Adapting Pitch-Based Self Supervised Learning Models for Tempo Estimation". *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 956-960, doi: 10.1109/ICASSP48485.2024.10447129.
- [10] Matthew C. McCallum, Florian Henkel, Jaehun Kim, Samuel E. Sandberg, Matthew E. P. Davies. "Similar but Faster: Manipulation of Tempo in Music Audio Embeddings for Tempo Prediction and Search". *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 686-690, doi: 10.1109/ICASSP48485.2024.10447877.
- [11] Florian Henkel, Jaehun Kim, Matthew C. McCallum, Samuel E. Sandberg, Matthew E. P. Davies. "Tempo Estimation as Fully Self-Supervised Binary Classification". *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1356-1360, doi: 10.1109/ICASSP48485.2024.10448098.
- [12] M. E. P. Davies and S. Böck, "Temporal convolutional networks for musical audio beat tracking", in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1-5. doi: 10.23919/EUSIPCO.2019.8902578.
- [13] S. Böck, M. Davies, and P. Knees, "Multi-Task Learning of Tempo and Beat: Learning One to Improve the Other", in *Proceedings of the 20th International Society for Music Information Retrieval Conference*, A. Flexer, G. Peeters, J. Urbano, and A. Volk, Eds., 2019, pp. 486-493. Zenodo. doi: 10.5281/zenodo.3527849.
- [14] S. Böck and M. E. P. Davies, "Deconstruct, Analyse, Reconstruct: How to improve Tempo, Beat, and Downbeat Estimation", in *Proceedings of the International Society for Music Information Retrieval Conference*, 2020. [Online]. Available: <https://archives.ismir.net/ismir2020/paper/000223.pdf>.
- [15] Hendrik Schreiber and Meinard Müller. "A Post-Processing Procedure for Improving Music Tempo Estimates Using Supervised Learning". *18th ISMIR Conference*, 2017. doi: 10.5281/zenodo.1415045
- [16] Kah Liang Ong, Chin Poo Lee, Heng Siong Lim, Kian Ming Lim, and Ali Alqahtani. "Mel-MViTv2: Enhanced Speech Emotion Recognition With Mel Spectrogram and Improved Multiscale Vision Transformers". *IEEE Access*, 11, 108571-108579, 2023. doi: 10.1109/ACCESS.2023.3321122
- [17] Tzanetakis, G. and Cook, P. "Musical genre classification of audio signals" *IEEE Transactions on Speech and Audio Processing*, 293-302, 2002. doi: 10.1109/TSA.2002.800560
- [18] M. Domingues, O. Mendes, M. Kaibara, V. Menconi, E. Bernardes. "Exploring the continuous wavelet transform". *Revista Brasileira de Ensino de Física*, 2016. doi: 10.1590/1806-9126-RBEF-2016-0019.
- [19] B. McFee, C. Raffel, D. Liang, et al., "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference (SciPy 2015)*, 2015, pp. 18-24. doi: 10.25080/Majora-7b98e3ed-003.

# SKIP THAT BEAT: AUGMENTING METER TRACKING MODELS FOR UNDERREPRESENTED TIME SIGNATURES

**Giovana Morais**  
MARL  
New York University  
giovana.morais@nyu.edu

**Brian McFee**  
MARL-CDS  
New York University  
brian.mcfree@nyu.edu

**Magdalena Fuentes**  
MARL-IDM  
New York University  
mfuentes@nyu.edu

## ABSTRACT

Beat and downbeat tracking models are predominantly developed using datasets with music in 4/4 meter, which decreases their generalization to repertoires in other time signatures, such as Brazilian samba which is in 2/4. In this work, we propose a simple augmentation technique to increase the representation of time signatures beyond 4/4, namely 2/4 and 3/4. Our augmentation procedure works by removing beat intervals from 4/4 annotated tracks. We show that the augmented data helps to improve downbeat tracking for underrepresented meters while preserving the overall performance of beat tracking in two different models. We also show that this technique helps improve downbeat tracking in an unseen samba dataset.

## 1. INTRODUCTION

Current datasets for beat and downbeat tracking are biased towards music in 4/4 meter, which do not adequately address music with different rhythmic structures, such as jazz, which typically contains 3/4, 12/8, or even 5/4 meters; classical music, typically featuring 2/4, 3/4, 6/8 meters (among others); Turkish Aksak (9/8); Cretan leaping dances (2/4), or Brazilian samba, which is in 2/4.

Recent advancements in deep learning have shifted beat and downbeat tracking from traditional signal processing methods, which use manually crafted features, to data-driven techniques, such as [1–4]. This exacerbated the bias towards music in 4/4 because annotating and creating new datasets for diverse musical meters is both time-consuming and costly. As many culturally specific music genres feature meters other than 4/4, this results in a predominance of mainstream annotated musical data [5]. For example, in [1] from the 2216 training tracks available, 1120 are in 4/4, 882 do not have beat position annotations, and the remaining 214 spans other 4 time signatures. This underrepresentation is not a problem for beat tracking, but it affects downbeat tracking as briefly discussed in [6].

Recent work has explored methods to minimize the number of required annotations while still enabling models to generalize across different music styles [7] and strategies for selecting which tracks to annotate [8]. While this minimizes the number of tracks and time to annotate, it does not remove the need for annotation.

Another way to alleviate the need to annotate new data is through data augmentation. Data augmentation involves modifying existing data to improve model training, such as rotating an image in computer vision tasks or adding noise in audio tasks. In audio applications, [9] proposes a framework to augment annotated data, which supports operations such as pitch shift and time stretch, and evaluate it on instrument recognition. They show that even simple transformations can lead to improvements in the task, but they should be done carefully (e.g. avoid deforming the acoustic audio signal too much). In [1] the augmentation is done by calculating the same STFT with different hop lengths, which improves the tempo estimation on unseen tempo ranges. Finally, [3] proposes augmenting the dataset by applying Harmonic-Percussive Source Separation (HPSS), and training the model with both the original mel-spectrogram, the harmonic component, and the percussive component. Inspired by the fact that data augmentation helps models generalize to out-of-domain situations [1], we explore these ideas in the context of meter tracking.

In particular, we propose a novel augmentation procedure that uses existing annotated data to enhance the representation of underrepresented meters, without the need for new annotations. We evaluate our approach using two models, the Temporal Convolutional Network (TCN) [1] and BayesBeat [10]. We use a test set with unseen meters but seen music genres, and another test set with an unseen music genre (Brazilian samba) and unseen meter. Our results demonstrate that our augmentation technique improves downbeat tracking performance while preserving the overall effectiveness of beat-tracking methods.

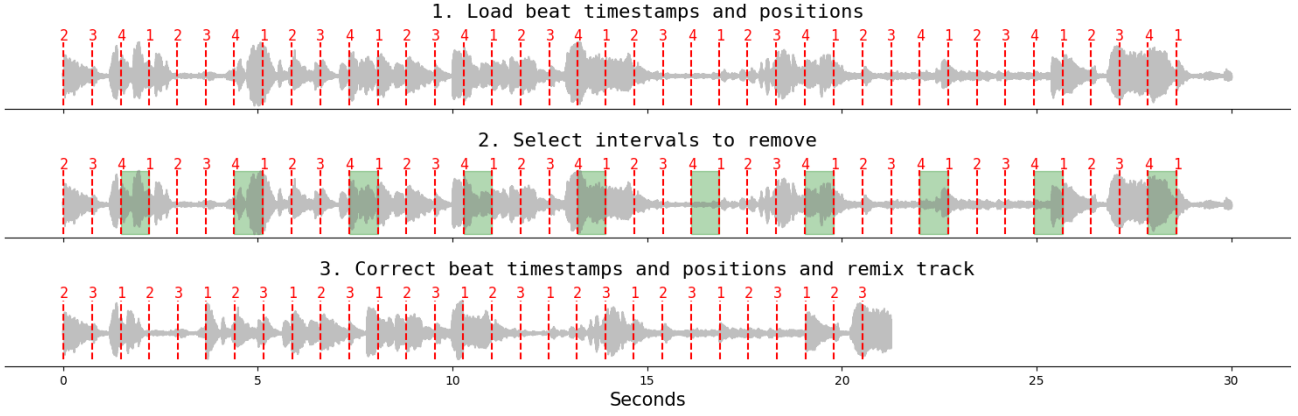
## 2. METHOD

### 2.1 Augmentation Procedure

Given a track that is in 4/4, let  $B = \{b_1, b_2, \dots\}$  be the beat annotation timestamps and  $P = \{p_1, p_2, \dots\}$  be the beat position within the bar, where  $p_i \in \{1, 2, 3, 4\}$ . Let  $IBI$  be the Inter-Beat Interval of  $B$ , i.e.  $IBI = \Delta B$ .



© G. Morais, B. McFee, and M. Fuentes. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).  
**Attribution:** G. Morais, B. McFee, and M. Fuentes, “Skip That Beat: Augmenting Meter Tracking Models for Underrepresented Time Signatures”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.



**Figure 1.** Illustration of the  $4/4 \rightarrow 3/4$  augmentation of the GTZAN track `blues.00009`. On the top plot, we have the original track with the beat timestamps and positions. The middle plot shows the intervals we remove in green. In this case, we remove every interval from beat 4 to beat 1. Then, we correct the annotation labels and timestamps and remix the track. On the bottom plot, we have the augmented track, with corrected time displacements.

We denote  $p^*$  the beat positions we want to keep. For an augmentation  $4/4 \rightarrow 2/4$ ,  $p^* = \{1, 2\}$ , and, similarly,  $4/4 \rightarrow 3/4$ ,  $p^* = \{1, 2, 3\}$ . Therefore, we define an augmentation as keeping positions  $p_i \in p^*$ , that is:

$$\hat{B} = \{b_1, b_2, \dots\} \text{ if } p_i \in p^* \quad (1)$$

$$\hat{P} = \{p_1, p_2, \dots\} \text{ if } p_i \in p^* \quad (2)$$

$$\widehat{IBI} = \{IBI_{i-1} \text{ if } p_i \in p^*\} \quad (3)$$

After we remove unwanted beats, we correct the remaining beat positions and any time displacements of the annotations. We do so by taking the first beat timestamp of  $\hat{B}$  and adding the  $\widehat{IBI}[i]$  for every new position. We denote the corrected beat timestamps as  $\bar{B}$ :

$$\bar{B} = \{\bar{B}[i-1] + \widehat{IBI}[i]\} \quad (4)$$

where  $\bar{B}[0] = \hat{B}[0]$ . Our augmentation procedure is illustrated in Figure 1, and its pseudo-code is provided in Algorithm 1. While we do not use the remaining values of  $\hat{B}$ , we still calculate it because beat timestamps and beat positions do not necessarily start at  $p_i = 1$  so it is incorrect to just do  $\bar{B}[0] = B[0]$  (see Figure 1 for an example).

Once we have the corrected new annotations, we use librosa’s remix function<sup>1</sup> to synthesize the augmented signal. This function receives the original audio and the beat intervals we wish to keep, mapping the interval boundaries to the closest zero-crossing in the signal to avoid discontinuities and concatenating them.

We note that, as of now, the beat position removed is fixed and we do not remove 1st and 2nd beats to keep musical information related to the downbeat, but these can be further explored in the future. It is possible to use the same procedure to augment  $4/4$  tracks to other time signatures, such as  $5/4$ , but instead of removing beats, one would repeat them. For example, we could repeat the third beat. We leave the exploration of these scenarios for future work.

<sup>1</sup> <https://librosa.org/doc/0.10.2/generated/librosa.effects.remix.html>

#### Algorithm 1 Augmentation Procedure

```

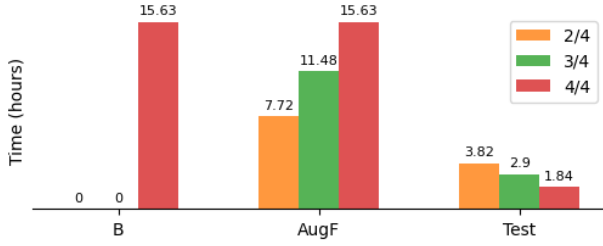
 $B \leftarrow \{b_1, b_2, \dots\}$   $\triangleright$  beat timestamp annotations
 $P \leftarrow \{p_1, p_2, \dots\}$   $\triangleright$  beat positions annotations
 $IBI \leftarrow \{0\}$ 
 $IBI \leftarrow IBI.append(\Delta B)$ 
 $keep \leftarrow \{\}$   $\triangleright$  indices of positions we want to keep
for  $i$  in  $length(B)$  do
    if  $p \in p^*$  then
         $keep \leftarrow keep.append(i)$ 
    end if
end for
 $\widehat{IBI} \leftarrow IBI[keep]$ 
 $\hat{P} \leftarrow P[keep]$ 
 $\bar{B} \leftarrow \{B[keep[0]]\}$ 
for  $i$  in  $range(1, length(\widehat{IBI}))$  do
     $\bar{B} \leftarrow \bar{B}.append(\bar{B}[i-1] + \widehat{IBI}[i])$ 
end for
    
```

## 2.2 Datasets

To create our augmented datasets, we use data from the Beatles [11], GTZAN [12], and the RWC Classical and Jazz [13] datasets. The GTZAN dataset consists of 1000 tracks of 30s each and spans 10 different genres. The rhythmic annotations, such as tempo, swing ratio, and beat/downbeat positions, were provided by [14] with metrical annotations added by [15]. The Beatles dataset consists of 179 songs from the 12 studio albums from The Beatles. RWC Classical and RWC Jazz consist of 50 tracks each with variate lengths. All datasets have beat and downbeat annotations, among others. We discard tracks without beat annotations.

For the four aforementioned datasets, we have tracks with no meter annotations. We infer the meter directly from the beat positions by first calculating the difference between consecutive  $p_i$  and identifying the negative differences, which indicate the transition between bars, e.g.  $[1, 2, 3, 4, 1]$  becomes  $[1, 1, 1, -3]$ . We count how often





**Figure 2.** Training and test data distributions by time signature.

these transitions occur at specific beat positions  $p_i$ . We consider the numerator of the meter the most frequent transition point. The denominator is always 4 for the tracks we are using in this work. This method does not account for tracks with internal time signature changes but gives us a good approximation of the dominant track meter.

While GTZAN and Beatles are more concentrated in 4/4, RWC Classical and Jazz have better representation on 2/4 and 3/4 time signatures. Still, from the 28.41 hours of audio (1283 tracks) from the datasets, 17.48 hours (1096 tracks) are classified as 4/4, 2.9 hours (87 tracks) are classified as 3/4, 3.82 hours (59 tracks) as 2/4 and the other 1.21 hours (41 tracks) are spanned between the other 5 time signatures. In this work, we focus on 2/4, 3/4, and 4/4 meters, leaving the remaining meters for future work.

These datasets give us a reasonable amount of tracks to train and test the models with the different time signatures. We create two different datasets to train the models:

**Baseline (B)** Randomly selected 993 original 4/4 tracks, which stands for 80% of the whole dataset.

**Augmented Full (AugF)** Same tracks as **B** plus their respective augmentations in 2/4 and 3/4. This is the largest dataset between the three, totaling 2979 tracks.

The test split is fixed for the three experiments, has 20% of the total tracks and consists only of original tracks. The split has the 2/4 and 3/4 tracks, plus the remaining 4/4 tracks. As a second and unseen test set, we use the acoustic mixtures from the Brazilian Rhythmic Instrument Dataset, BRID [16], which consists of 93 short tracks of samba (samba-enredo and partido alto). Tempo varies from track to track, but not within the track. All the tracks are in 2/4.

Figure 2 shows the time signature distribution of the data. We note here the TCN methods use 20% of the train split as a validation set, while BayesBeat uses the whole set as training.

### 2.3 Meter Tracking Models

To evaluate our augmentation strategy, we train two different meter-tracking models: the TCN and BayesBeat. In previous works, the two models were shown to be capable of generalizing to different music corpora [7, 17].

The TCN was first proposed to beat tracking in [18], and then expanded to simultaneously estimate beat, downbeat,

and tempo in [1]. We use the open-source implementation provided in [5]. The TCN outputs beat and downbeat activations (i.e. likelihood of beats and downbeats respectively), which are typically further processed using a Dynamic Bayesian Network (DBN) for temporal consistency.

According to [19], the DBN is good at dealing with ambiguous observations and finds the global best state sequence given these observations. [4] argues that the DBN performs well on most pieces commonly used to train and evaluate beat tracking systems because those tracks have common features (e.g., stable tempo, meter in either 3/4 or 4/4) and that the DBN is most likely to mispredict more challenging data. In preliminary experiments, we observed that the DBN would improve overall results by 5% (i.e. similar to the BayesBeat), but as argued in [4] it would obscure the effects of the augmentation. To better analyze the augmentation effect in the learning of the model we replace the DBN post-processing with the adaptive thresholding peak-picking method proposed by [20]<sup>2</sup>. This method has a moving window that compares peaks with the median threshold. We refer to it as *TCN-PP*. Finally, the implementation of the TCN in [5] contains tempo augmentation. We removed this step to make the comparison with the BayesBeat fairer.

The second model we use is BayesBeat [10], a statistical method that tracks beat, downbeat, tempo, meter, and rhythmic patterns. We use the MATLAB implementation available on GitHub<sup>3</sup>. In comparison with the TCN, BayesBeat has fewer parameters and trains faster. For this model, we use the default parameters, i.e. two frequency bands (low and high) and one rhythmic pattern. We refer to this model as *BB*.

## 3. EXPERIMENTS AND RESULTS

### 3.1 Setup

We start by loading all 4/4 tracks from the datasets discussed in Section 2.2. We use *mirdata* [21] to load the tracks' annotations. We discard tracks without annotations. Then, we apply the augmentation procedure described in Section 2.1 to all 4/4 tracks.

For every model, we evaluate the F-measure and continuity metrics CMLt, AMLt, CMLc, and AMLc (Correct and Allowed Metrical Level with and without continuity required, respectively). We use the implementation provided in *madmom*<sup>4</sup> [22].

In our GitHub repository<sup>5</sup>, we have made available the augmentation and experiments code.

### 3.2 Results

Table 1 shows the average performance of the models trained on three different datasets. Overall, the beat met-

<sup>2</sup>We use the implementation provided in [https://www.audiolabs-erlangen.de/resources/MIR/FMP/C6/C6S1\\_PeakPicking.html](https://www.audiolabs-erlangen.de/resources/MIR/FMP/C6/C6S1_PeakPicking.html)

<sup>3</sup><https://github.com/flokadillo/bayesbeat>

<sup>4</sup><https://github.com/CPJKU/madmom>

<sup>5</sup>[https://github.com/giovana-morais/skip\\_that\\_beat](https://github.com/giovana-morais/skip_that_beat)

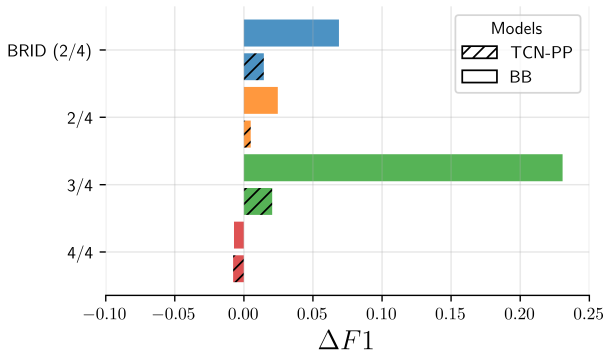
		Beat					Downbeat				
		F1	CMLt	CMLc	AMLt	AMLc	F1	CMLt	CMLc	AMLt	AMLc
BB	B	0.70	0.53	0.46	0.76	0.61	0.41	0.31	0.29	0.49	0.45
	AugF	0.71	0.54	0.47	0.78	0.63	0.49	0.38	0.35	0.60	0.54
TCN-PP	B	0.72	0.42	0.27	0.63	0.36	0.36	0.00	0.00	0.29	0.11
	AugF	0.74	0.47	0.33	0.64	0.39	0.36	0.01	0.00	0.23	0.11

**Table 1.** Average results for the models across all time signatures.

rics remained stable across models and datasets, with a 2% improvement for the TCN-PP model on the **AugF** dataset, mainly due to improvements in the 3/4 and 4/4 time signatures. For downbeat, we see that the BayesBeat model benefited from the **AugF** dataset with an 8% overall improvement. The TCN’s performance does not seem to change on average, but when looking at a breakdown of the results per meter we observe some differences, as discussed below.

### 3.2.1 Results by Time Signature

Figure 3 shows the models’ performance broken down by meter in the test set, for the two augmented training datasets. We see an improvement across models in 2/4 and 3/4, with the biggest improvement seen in downbeat tracking for 3/4 tracks for the BB model. We hypothesize that this is due to the difference between the time signatures. While 2/4 and 4/4 can be considered similar, 3/4 has a more distinct pattern. By inspecting examples, we see that models trained with the augmented datasets misclassified 4/4 tracks as 2/4 and the other way around. Those errors make sense given the perceptual similarity between those meters. Besides the F-measure, **BB** also improves the continuity metrics, as seen in Table 1, showing that once the method understands the correct meter, it propagates it through the track.



**Figure 3.** Relative F-measure for downbeat tracking on test set and BRID. The bars represent the deviation from the baseline F-measure. Hatched bars represent TCN-PP results and non-hatched bars represent BB.

### 3.2.2 Results for BRID

In [7] the authors reported results for a TCN trained only on Western music on the BRID dataset. The model had an F-measure of 0.096 and a CMLt of 5.9, demonstrating the difficulty of generalizing to unseen data. The result had

an increase of 70% when the authors fine-tuned the model with only 0.67 minutes of annotated BRID data. They argued that the difficulty is not only in the meter itself but also in the acoustic properties of the data. The authors also report that BayesBeat can achieve similar downbeat F-measures by being trained from scratch with around 3 minutes of annotated data. We see in Figure 3 that our models show an improvement of about 5% and 15% with respect to the baseline without annotating new data when evaluated on BRID. The improvement for the TCN method was not as big as the BayesBeat, but in absolute values, the TCN had a better performance than BayesBeat. The augmentation was not sufficient to account for the acoustic properties of the track, which are different from training, but it improves downbeat tracking without any extra annotated data. This augmentation could be used to annotate a few minutes of samba to then retrain the models, as a preprocessing step for the approach in [7].

As one of its main characteristics, samba has a strong accent on the second beat and the development of contra-metric structures [7]. In our qualitative analysis, we saw that offbeat mistakes were the most common in all models and training data variations, i.e. the strongest beat was classified as the downbeat, even though it was not.

We provide full results and listening examples in the accompanying website<sup>6</sup>.

## 4. CONCLUSIONS AND FUTURE WORK

In this work, we proposed an augmentation procedure to increase representation in 2/4 and 3/4 time signatures for beat and downbeat tracking. We show that the augmentation approach helps models generalize to unseen meters and datasets but it has limitations for unseen accentuation and rhythmic patterns. The simplicity of the method leaves room for exploration in different directions. For future work, we will explore adding internal meter changes and augmenting different time signatures. We will also experiment with tracks without voice stems, so the augmentation cuts are less perceivable, and explore variations of the removed beat positions, instead of removing, e.g., the 3rd and 4th beats within a bar. Finally, it is worth testing the effect of the augmentations on deep learning models that do not rely on the DBN, such as [4], and evaluate whether the improvement and the generalization to unseen datasets, such as BRID, remains.

<sup>6</sup>[https://giovana-morais.github.io/skip\\_that\\_beat\\_demo/](https://giovana-morais.github.io/skip_that_beat_demo/)

## 5. ACKNOWLEDGMENTS

We thank Martín Rocamora for sharing his BayesBeat training scripts with us. We also thank the reviewers for their comments and suggestions. This work was partially supported through the NYU IT High Performance Computing resources, services, and staff expertise.

## 6. REFERENCES

- [1] S. Böck and M. E. P. Davies, “Deconstruct, Analyse, Reconstruct: How to improve Tempo, Beat, and Downbeat Estimation,” in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, 2020, pp. 574–582.
- [2] J. Zhao, G. Xia, and Y. Wang, “Beat Transformer: Demixed Beat and Downbeat Tracking with Dilated Self-Attention,” in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, P. Rao, H. A. Murthy, A. Srinivasamurthy, R. M. Bittner, R. C. Repetto, M. Goto, X. Serra, and M. Miron, Eds., 2022, pp. 169–177. [Online]. Available: <https://archives.ismir.net/ismir2022/paper/000019.pdf>
- [3] T. Cheng and M. Goto, “Transformer-Based Beat Tracking With Low-Resolution Encoder and High-Resolution Decoder,” in *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023, Milan, Italy, November 5-9, 2023*, 2023, pp. 466–473.
- [4] F. Foscari, J. Schlüter, and G. Widmer, “Beat this! Accurate beat tracking without DBN postprocessing,” *CoRR*, vol. abs/2407.21658, 2024.
- [5] M. E. P. Davies, S. Böck, and M. Fuentes, *Tempo, Beat and Downbeat Estimation*, Nov. 2021. [Online]. Available: <https://tempobeatdownbeat.github.io/tutorial/intro.html>
- [6] F. Krebs, S. Böck, M. Dorfer, and G. Widmer, “Downbeat Tracking Using Beat Synchronous Features with Recurrent Neural Networks,” in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, 2016, pp. 129–135.
- [7] L. S. Maia, M. Rocamora, L. W. P. Biscainho, and M. Fuentes, “Adapting meter tracking models to latin american music,” in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, 2022, pp. 361–368.
- [8] —, “Selective Annotation of Few Data for Beat Tracking of Latin American Music Using Rhythmic Features,” *Trans. Int. Soc. Music. Inf. Retr.*, vol. 7, no. 1, pp. 99–112, 2024.
- [9] B. McFee, E. J. Humphrey, and J. P. Bello, “A Software Framework for Musical Data Augmentation,” pp. 248–254, 2015.
- [10] N. Whiteley, A. T. Cemgil, and S. J. Godsill, “Bayesian Modelling of Temporal Structure in Musical Audio,” in *ISMIR 2006, 7th International Conference on Music Information Retrieval, Victoria, Canada, 8-12 October 2006, Proceedings*, 2006, pp. 29–34.
- [11] M. E. P. Davies, N. Degara, and M. D. Plumbley, “Evaluation Methods for Musical Audio Beat Tracking Algorithms.”
- [12] G. Tzanetakis and P. R. Cook, “Musical genre classification of audio signals,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.
- [13] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical and jazz music databases,” in *ISMIR 2002, 3rd International Conference on Music Information Retrieval, Paris, France, October 13-17, 2002, Proceedings*, 2002.
- [14] U. Marchand, Q. Fresnel, and G. Peeters, “GTZAN-Rhythm: Extending the GTZAN Test-Set with Beat, Downbeat and Swing Annotations,” Oct. 2015.
- [15] E. Quinton, C. Harte, M. Sandler, and C. Shannon, “Extraction of Metrical Structure from Music Recordings,” in *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, Trondheim, Norway, 2015.
- [16] P. Tomaz, L. Maia, M. Fuentes, M. Rocamora, L. Biscainho, M. da Costa, and S. Cohen, “A Novel Dataset of Brazilian Rhythmic Instruments and Some Experiments in Computational Rhythm Analysis,” in *Congreso Latinoamericano de la AES 2018, Montevideo, Uruguay, September 24-26, 2018*, Sep. 2018.
- [17] A. Holzapfel, F. Krebs, and A. Srinivasamurthy, “Tracking the “Odd”: Meter Inference in a Culturally Diverse Music Corpus,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, H. Wang, Y. Yang, and J. H. Lee, Eds., 2014, pp. 425–430.
- [18] E. P. Matthew Davies and S. Böck, “Temporal convolutional networks for musical audio beat tracking,” in *27th European Signal Processing Conference, EU-SIPCO 2019, A Coruña, Spain, September 2-6, 2019*. IEEE, 2019, pp. 1–5.
- [19] S. Böck, F. Krebs, and G. Widmer, “Joint Beat and Downbeat Tracking with Recurrent Neural Networks,” pp. 255–261, 2016.
- [20] O. Nieto and J. P. Bello, “Systematic Exploration of Computational Music Structure Research,” in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York*



*City, United States, August 7-11, 2016*, 2016, pp. 547–553.

- [21] M. Fuentes, R. Bittner, G. Plaja-Roglans, G. Cortès, T. Khandelwal, H. Palan, M. Miron, D. Zasukha, C. Thomé, G. Morais, F. Papaleo, P. Ramoneda, J. Arruti, and M. Rocamora, “Mirdata 0.3.8,” Zenodo, Nov. 2023.
- [22] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, “madmom: a new Python Audio and Music Signal Processing Library,” in *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 10 2016, pp. 1174–1178.

# IMPROVING MUSIC EMOTION RECOGNITION BY LEVERAGING HANDCRAFTED AND LEARNED FEATURES

Pedro Lima Louro<sup>1</sup>

Hugo Redinho<sup>1</sup>

Ricardo Malheiro<sup>1,2</sup>

Rui Pedro Paiva<sup>3</sup>

Renato Panda<sup>1,3</sup>

<sup>1</sup> Centre for Informatics and Systems of the University of Coimbra (CISUC),

Department of Informatics Engineering, LASI, University of Coimbra, Portugal

<sup>2</sup> Polytechnic Institute of Leiria, School of Technology and Management, Portugal

<sup>3</sup> Ci2 — Smart Cities Research Center Polytechnic Institute of Tomar, Portugal

pedrolouro@dei.uc.pt, redinho@student.dei.uc.pt, {rsmal, ruipedro, panda}@dei.uc.pt

## ABSTRACT

Music Emotion Recognition was dominated by classical machine learning, which relies on traditional classifiers and feature engineering (FE). Recently, deep learning approaches have been explored, aiming to remove the need for handcrafted features by automatic feature learning (FL), albeit at the expense of requiring large volumes of data to fully exploit their capabilities. A hybrid approach fusing information from handcrafted and learned features was previously proposed, outperforming separate FE and FL approaches on the 4QAED dataset (900 audio clips). The results suggested that, in smaller datasets, FE and FL could complement each other rather than act as competitors. In the present study, these experiments are extended to the larger MERGE dataset (3554 audio clips) to analyze the impact of the significant increase in data. The best-obtained results, 77.62% F1-score, continue to surpass the standalone FE and FL paradigms, reinforcing the potential of hybrid approaches.

## 1. INTRODUCTION

Recently, several Deep Learning (DL) approaches have been proposed to address research problems in Music Emotion Recognition (MER). These eliminate the necessity of feature engineering efforts since DL architectures can automatically learn relevant features from the input data.

However, as pointed out in a previous study [1], the current state-of-the-art DL approaches are still underperforming compared to classical MER approaches using handcrafted features. The lack of sizeable and quality MER datasets is part of the problem since DL architectures can only reach their full potential with a large, representative set of samples for the problem at hand, which usually takes

hundreds of thousands of samples. Furthermore, the features learned by a neural network depend on the data provided. Emotionally-relevant patterns may be missed if they are rare in the dataset, unlike handcrafted features, which can target specific characteristics even if they appear infrequently, though this might not improve classification performance.

To take advantage of the strengths of both the classical and DL paradigms, a hybrid methodology was previously developed and validated on a set of small datasets (containing 1372 samples), showing promising results. This methodology surpassed all classical and neural network-based baselines [1].

In this article, we further extend the evaluation of this hybrid methodology to two new datasets proposed by Louro et al. [2], containing 3554 samples.

## 2. RELATED WORK

In this section, we briefly review the state-of-the-art approaches relevant to this study. Both classical ML and DL-based methodologies are discussed, concluding with a summary of the advantages and disadvantages of each.

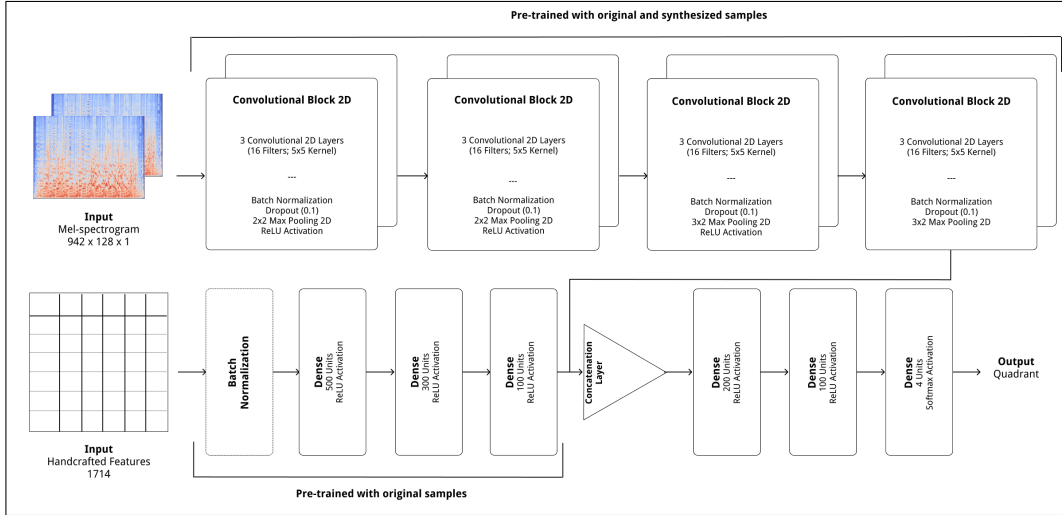
Seminal works in MER follow a common pipeline. First, a set of songs is collected and manually annotated by subjects, followed by the extraction of features relevant to emotion, and finally, training and evaluating a classifier, such as Support Vector Machines (SVM) or Random Forest, to name a few. Feng et al. [3] presents such a pipeline, with a slight difference regarding the song annotations. Instead of manual annotations, these are automatically obtained through predefined intervals of the extracted features, e.g., an excerpt with legato as the predominant articulation and a slow tempo labeled as sad.

Subsequent works improve on the oversights of this approach and explore the problem in other directions. Some of these works include Lu et al. [4], which keep the single label approach but use a Gaussian Mixture Model to classify samples based on intensity, timbre and rhythm features, and Yang et al. [5], defining MER as a regression problem to mitigate the ambiguities inherent to the discrete labels from the previously mentioned approaches.

Recently, Panda et al. [6] proposed a set of new audio



© P. L. Louro, H. Redinho, R. Malheiro, R. P. Paiva, R. Panda. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** P. L. Louro, H. Redinho, R. Malheiro, R. P. Paiva, R. Panda, "Improving Music Emotion Recognition By Leveraging Handcrafted and Learned Features", in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.



**Figure 1.** Hybrid augmented architecture. The architecture can be decomposed into the frontend, or the CNN and DNN portions where features are learned or processed, respectively, and the backend, further processing the concatenated features and outputting the predicted quadrant.

features alongside a small but thoroughly validated balanced dataset of 900 song excerpts and a state-of-the-art methodology focused on classical static MER. There, the 4 Quadrant Audio Emotion Detection (4QAED) dataset was built with the aid of a semi-automatic approach, building on user-generated labels from AllMusic, and manual validation. Features were extracted for each excerpt and ranked, from which only the top 100 were used to train an SVM classifier. The observed results are considerably higher than those previously reported, attaining a 76.4% F1-score.

As discussed previously, neural networks have the ability to automatically learn the most relevant features from the input data. Such an idea is very appealing to any Music Information Retrieval (MIR) problem, considering the hardship of developing and validating features by hand. To our knowledge, the first application of these models to MIR was presented by Choi et al. [7]. Here, the experiments used only convolutional layers, learning and processing the learned features from Mel-spectrogram representations of the considered datasets for validation. Later, the same authors presented a more complex network, consisting of a convolutional portion for feature learning, and a recurrent portion for processing time-related features and performing classification, referred to as Convolutional Recurrent Neural Network (CRNN) [8]. The final system, trained for multi-label classification, attained a 0.86 Area Under the ROC Curve (AUC), outperforming the other proposed architectures.

Several approaches built on this system, iterating mostly on certain aspects of its architecture. Some of these include musically-motivated filters applied to the convolutional portion of the network, focusing on finding timbral and temporal information [9], and end-to-end architectures, which aim to learn the most relevant features directly from the raw audio waveform [10].

Regarding MER specifically, many works build upon

the previously described system’s pipelines, experimenting with different data representations such as chromagram [11] and conchleogram [12], applying transfer learning from related tasks such as speech [13], and experimenting with smaller input sizes [14].

This work builds on Panda et al. [6] and Choi et al. [7] for the FE and FL portions respectively, given their impact in the field.

### 3. MATERIALS AND METHODOLOGY

This section describes the methodology followed in this study, starting with the pre-processing steps, then describing the architecture details, and ending with the optimization strategy.

Dataset	Q1	Q2	Q3	Q4	Total
MERGE Audio C	875	915	808	956	3554
MERGE Audio B	808	808	808	808	3232
MERGE Bimodal C	525	673	500	518	2216
MERGE Bimodal B	500	500	500	500	2000

**Table 1.** Datasets used and their distribution per quadrant.

#### 3.1 Datasets

The methodology was evaluated using two datasets: MERGE Audio and MERGE Bimodal, which includes a complete and a balanced collection of samples for each dataset, detailed in [2]. MERGE Audio contains 3554 and 3232 samples, while the MERGE Bimodal comprises 2216 and 2000 samples. The quadrant distribution of each is detailed in Table 1.

Each dataset entry includes a 30-second audio clip of the most emotionally-representative part of the song. Samples were annotated into one of four emotion quadrants (happy, tense, sad, and relaxed), according to Russell’s Circumplex model [15]. While the MERGE Bimodal dataset

includes the full lyrics for each sample, this research does not explore the lyrical content.

A 70-15-15 train-validate-test (TVT) split was used as our validation strategy, as recommended in [2].

### 3.2 Pre-processing Steps

Initially, the audio samples are converted into WAV format. To obtain Mel-spectrograms, these samples are down-sampled from 22.5 kHz to 16kHz, as per the methodology in [1].

The handcrafted features are extracted from all samples using MIRToolbox [16], Marsyas [17], and PsySound3 [18] audio frameworks, complemented by the novel features proposed by Panda et al. [6]. A final set of 1714 features is obtained after performing feature decorrelation, i.e., eliminating redundant features that would not contribute to increasing the model’s performance.

As for the input data for the CNN portion, the librosa library [19] is used to obtain Mel-spectrogram representations. The library’s default settings for the Fast Fourier Transform window length (2048) and the hop size (512) are used to generate the spectral representations.

Data augmentation is also performed on the train set of each dataset when optimizing the CNN portion. This is done by applying time shifting (shifts the start or the end of the audio clip by a maximum of 5 seconds), pitch shifting (increases or decreases the pitch by a maximum of 2 semitones), time stretching (speeds up or slows down an audio clip by a maximum of 50%), and power shifting (increases or decreases amplitude by a maximum of 10 dB) to each audio clip. Since each transformation is applied individually, the train set essentially increases five-fold.

### 3.3 Architecture Details

The architecture, illustrated in Figure 1, comprises a CNN and DNN portion for feature learning and processing, respectively, and a smaller DNN portion for classification.

The CNN portion, based on the mentioned work by Choi et al. [7], comprises four convolutional blocks, each containing a sequence of Batch Normalization, Dropout, and Max Pooling layers, ending with a ReLU activation layer. The last convolutional block does not contain the Dropout layer. The previously discussed Mel-spectrograms are fed as input to this portion.

The resulting output of both is concatenated at the feature level before being fed to the classifier, a set of three Dense layers, with the last one outputting one of the four quadrants of Russell’s Circumplex model [15]. This way, the classifier could pick the set of patterns that are most relevant to the problem at hand.

The CNN portion’s training phase includes synthesized samples to improve its performance. Therefore, this and the DNN for feature processing are pre-trained separately, freezing their weights before training the classification portion.

Dataset	Best Hyperparameters		
	Batch Size	Optimizer	Learning Rate
MERGE	32	SGD	1e-2
Audio	128	SGD	1e-2
Complete	16	Adam	1e-2
MERGE	32	SGD	1e-2
Audio	128	Adam	1e-4
Balanced	32	SGD	1e-3
MERGE	32	SGD	1e-2
Bimodal	128	SGD	1e-2
Complete	64	SGD	1e-4
MERGE	32	Adam	1e-3
Bimodal	32	SGD	1e-2
Balanced	64	Adam	1e-3

**Table 2.** Optimal hyperparameters per dataset. For each, the optimal values for standalone CNN and DNN portions are shown, followed by the final classifier optimization.

### 3.4 Optimization Strategy

The model optimization was carried out with the Bayesian optimization approach provided by the Keras Tuner library [20]. This technique searches for the optimal combination of hyperparameters within predefined ranges, aiming to maximize or minimize a specific objective function defined by the user.

The tuner’s objective is to maximize the validation set’s accuracy. The optimal values for each hyperparameter, including batch size, optimizer, and the respective learning rate, are detailed in Table 2.

The process involves running ten trials, beginning at the lower end of the specified intervals. For every trial, the model undergoes training up to a maximum of 200 epochs. However, an early stopping mechanism fires if the validation accuracy does not improve for 15 straight epochs or if the training accuracy exceeds 90%. This approach greatly decreases the time required for the optimization by reducing the time spent on hyperparameters that show poor performance, also preventing overfitting.

We used the 70-15-15 train-validate-test (TVT) split defined in [2] as our validation strategy. The resulting models for each trial are backed up for later usage, including the evaluation phase, which is discussed next.

In the TVT strategy, the optimization function uses both training and validation sets to identify the best hyperparameters. Once the model is trained using these, it undergoes evaluation on the test set. This evaluation involves calculating the F1-score, Precision, and Recall by comparing the actual values with the model’s predictions for each category and assessing the model’s overall performance.

## 4. RESULTS AND DISCUSSION

The results and gathered insights are presented in this section. We begin by highlighting the most relevant results from the previously presented metrics, followed by a discussion on the improvements and drawbacks of applying

	Dataset	F1-score	Precision	Recall
	MERGE Audio Complete	68.84%	69.52%	68.80%
	MERGE Audio Balanced	<b>77.62%</b>	<b>78.11%</b>	<b>77.89%</b>
	MERGE Bimodal Complete	73.13%	75.45%	74.40%
	MERGE Bimodal Balanced	70.00%	69.99%	70.33%

**Table 3.** TVT 70-15-15 results for the mentioned datasets

		Predicted			
		Q1	Q2	Q3	Q4
Actual	Q1	<b>70.2%</b>	10.9%	9.4%	9.4%
	Q2	7.1%	<b>92.1%</b>	0.8%	0.0%
	Q3	3.9%	2.6%	<b>55.3%</b>	38.2%
	Q4	18.3%	0.9%	21.7%	<b>59.13%</b>

**Table 4.** Confusion matrix for MERGE Audio Complete

		Predicted			
		Q1	Q2	Q3	Q4
Actual	Q1	<b>77.9%</b>	7.6%	3.8%	10.9%
	Q2	6.0%	<b>93.2%</b>	0.9%	0.0%
	Q3	2.8%	1.4%	<b>68.8%</b>	27.0%
	Q4	8.4%	0.0%	18.9%	<b>72.6%</b>

**Table 5.** Confusion matrix for MERGE Audio Balanced

		Predicted			
		Q1	Q2	Q3	Q4
Actual	Q1	<b>73.9%</b>	5.7%	6.8%	13.6%
	Q2	7.1%	<b>92.9%</b>	0.0%	0.0%
	Q3	5.7%	0.9%	<b>58.5%</b>	34.9%
	Q4	5.1%	0.0%	23.1%	<b>71.8%</b>

**Table 6.** Confusion matrix for MERGE Bimodal Complete

		Predicted			
		Q1	Q2	Q3	Q4
Actual	Q1	<b>74.7%</b>	5.3%	8.0%	12.0%
	Q2	9.1%	<b>90.9%</b>	0.0%	0.0%
	Q3	3.5%	1.2%	<b>58.8%</b>	36.5%
	Q4	14.3%	0.0%	30.2%	<b>55.6%</b>

**Table 7.** Confusion matrix for MERGE Bimodal Balanced

the hybrid methodology to the MERGE Audio and Bimodal datasets.

The best F1-score attained was 77.62% on the MERGE Audio Balanced dataset, as seen in Table 3. Again, the model’s performance is shown to be particularly susceptible to quadrant balancing since the lowest result is observed when using the largest but most unbalanced of the validation datasets.

From previous experiments and according to the literature, one of the biggest challenges of audio-only approaches is to accurately differentiate valence when arousal is low, i.e., confusion between the third and fourth quadrants of Russell’s Circumplex model. As observed in Table 4, this is still present in this model when considering MERGE Audio Complete, also with some considerable confusion between the first and fourth quadrants. Using the balanced counterpart, as seen in Table 5, the confusion is reduced considerably in the third quadrant. This improvement is very significant given that it is the quadrant that produces the most confusion, even for human annotators. The fourth quadrant also improves significantly, a consequence of less confusion with the first quadrant.

There are some caveats to consider, such as the overall higher results for MERGE Bimodal Complete against Bimodal Balanced. Although this contradicts the previous idea that quadrant distribution is essential for this model, this could be explained by less disparity between the number of samples of the third and fourth quadrants compared to MERGE Audio Complete. This is further corroborated by the confusion matrices in Tables 6 and 7, as the most significant difference is the performance of the fourth quadrant.

## 5. CONCLUSION AND FUTURE WORK

The Hybrid Augmented methodology is further experimented with in the present study. Due to the promising results of the fusion of handcrafted and learned features, we conducted further experiments on larger datasets, namely the complete and balanced versions of MERGE Audio and MERGE Bimodal. Each portion of the architecture is trained independently, first pre-training the CNN portion for feature learning and the DNN portion for feature processing, with additional synthesized samples added to the optimization phase of the former. The optimal weights for each portion are frozen, finally optimizing the classification portion.

The best result from these datasets is a 77.62% F1-score, attained with MERGE Audio Balanced. This was expected since previously reported results indicate the importance of large datasets with even distribution between quadrants for optimal performance. The confusion matrices for the MERGE Audio datasets further corroborate this conclusion, as low arousal quadrants are more easily distinguished in the balanced version of these. There are some inconsistencies, such as the higher results in the complete version of the MERGE Bimodal datasets, which may be due to a smaller gap between the number of samples of the third and fourth quadrants.

Regarding the methodology, it would be beneficial to analyze further the impact of new data augmentation techniques applied to the CNN portion of the model. It would also be beneficial to experiment with optimizing the DNN portion of the network with the same synthesized data of the CNN counterpart. Finally, the classifier could be further enhanced by including recurrent layers, such as in the CRNN architecture, to process time-related features from

the previously processed information.

## 6. ACKNOWLEDGMENTS

This work is funded by FCT - Foundation for Science and Technology, I.P., within the scope of the projects: MERGE - DOI: 10.54499/PTDC/CCI-COM/3171/2021 financed with national funds (PIDDAC) via the Portuguese State Budget; and project CISUC - UID/CEC/00326/2020 with funds from the European Social Fund, through the Regional Operational Program Centro 2020. Renato Panda was supported by Ci2 - FCT UIDP/05567/2020.

## 7. REFERENCES

- [1] P. L. Louro, H. Redinho, R. Malheiro, R. P. Paiva, and R. Panda, "A Comparison Study of Deep Learning Methodologies for Music Emotion Recognition," *Sensors*, vol. 24, no. 7, p. 2201, 2024.
- [2] P. L. Louro, H. Redinho, R. Santos, R. Malheiro, R. Panda, and R. P. Paiva, "MERGE – A Bimodal Dataset for Static Music Emotion Recognition," Jul. 2024.
- [3] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 375–376.
- [4] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 5–18, 2006.
- [5] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A Regression Approach to Music Emotion Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.
- [6] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, Oct. 2020.
- [7] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 2016, pp. 805–811.
- [8] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proceedings of the 2017 International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2392–2396.
- [9] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018, pp. 637–644.
- [10] J. Lee, J. Park, K. Kim, and J. Nam, "SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification," *Applied Sciences*, vol. 8, no. 1, p. 150, Jan. 2018.
- [11] M. Bilal Er and I. B. Aydılek, "Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features:," *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, p. 1622, 2019.
- [12] M. Russo, L. Kraljević, M. Stella, and M. Sikora, "Cochleogram-based approach for detecting perceived emotions in music," *Information Processing & Management*, vol. 57, no. 5, p. 102270, Sep. 2020.
- [13] J. S. Gomez Canon, E. Cano, P. Herrera, and E. Gomez, "Transfer learning from speech to music: Towards language-sensitive emotion recognition models," in *2020 28th European Signal Processing Conference (EUSIPCO)*. Amsterdam, Netherlands: IEEE, Jan. 2021, pp. 136–140.
- [14] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760–767, Jun. 2021.
- [15] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.
- [16] O. Lartillot, "MIR Toolbox 1.8.1 User's Manual," 2021.
- [17] D. Bogdanov, N. Wack, E. Gomez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: An Audio Analysis Library for Music Information Retrieval," in *14th International Society for Music Information Retrieval Conference*, 2013.
- [18] D. Cabrera, S. Ferguson, and E. Schubert, "'PsySound3': Software for Acoustical and Psychoacoustical Analysis of Sound Recordings," in G. P. Scavone (Ed.), *13th International Conference on Auditory Display*, 2007, pp. 356–363.
- [19] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and Music Signal Analysis in Python," in *Python in Science Conference*, Austin, Texas, 2015, pp. 18–24.
- [20] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi *et al.*, "Keras Tuner," <https://github.com/keras-team/keras-tuner>, 2019.

## **Session 2: MIR Applications**





# SYMBOLIC MUSIC STYLE TRANSFER VIA LATENT SPACE TRANSFORMATIONS: MODEL AND EVALUATION

Lucas Somacal<sup>1</sup>

Pablo Riera<sup>1,2</sup>

Diego Fernandez Slezak<sup>1,2</sup>

Martin Miguel<sup>3</sup>

<sup>1</sup> University of Buenos Aires. Faculty of Natural and Exact Sciences.

Computer Science Department. Buenos Aires, Argentina.

<sup>2</sup> CONICET-University of Buenos Aires. Computer Science Institute (ICC). Buenos Aires, Argentina.

<sup>3</sup> McMaster University, Ontario, Canada

lsomacal@dc.uba.ar

## ABSTRACT

In this work, we addressed the challenge of style transfer in music, aiming to transform a musical fragment from one style to another. We used a Variational Autoencoder (VAE) to encode musical fragments into a latent space, manipulating them using characteristic vectors of each style. We focused on four distinct music styles chosen to present different challenges, comparing two models: one trained on a general music dataset and another fine-tuned on the target styles. We also introduced an automatic evaluation method to assess resemblance to the target style, musicality, and identity preservation. Both models successfully achieved style transfer with stronger manipulations in the latent space enhancing style resemblance but reducing musicality and identity, making this a relevant parametrization. Fine-tuning offered only slight improvements showing that a single model trained on a large dataset can generalize to represent multiple styles.

## 1. INTRODUCTION

The last decade of developments in artificial intelligence has allowed users to generate music with ease, as continuations of an initial melody in a given style [1], create accompaniments to a melody [2] or even generate full audio songs from text prompts [3, 4]. Several of these tools are usually catered for a general audience, allowing them to create music with ease and without expert knowledge. Yet, music creators may desire tools that allow more controlled manipulation of their compositions, such as making a score sound more natural, interpolating between musical fragments or generating drum bases for a given melody [5, 6].

With this in mind, in this work we are interested in the problem of style transfer for symbolic music: change the style of an existing piece of music in symbolic format to mimic a specific music style. This problem has been

approached in different ways. Models like DeepJ [7] or MuseNet [1] generate continuations for an input musical fragment in a specific style. Others models generate style-conditioned chord accompaniments for a melody [2, 8, 9]. In particular, we are interested in models that allow for direct manipulation of a music fragment, generating a new one that maintains the identity of the original while containing features of a target music style.

This task has been approached previously using Generative Adversarial Networks (GANs) [10] and Variational Autoencoders (VAEs) [11]. GANs are models that, given an input seed, attempt to generate content indistinguishable from other examples of a specific domain. In [10], they train generative models that are conditioned on a musical fragment and try to generate a new version similar to the original but resembling the target style. Each model is designed to translate from a specific source to a specific target style. VAEs are models that encode a input to a latent space representation and learn to decode the same input from the latent space. These models have been used to manipulate inputs by decoding modified versions of their latent space encoding. In [11], they use VAEs to perform music style transfer. As in [10], they train specific models that work in pairs of music styles. Moreover, these VAEs are trained to encode the style of the input as a section of the latent space, which is then modified to instruct the decoder to decode the input in the target style.

In this work, we propose using a VAE approach to achieve multi-style transfer with a single model. We propose modifying the fragments by applying characteristic vectors of a target style to the encoding of the fragment prior to decoding. This approach is commonly known as latent space vector arithmetic [12–14]. Here, we show that a single model trained with large dataset can perform style transfer even in styles it was not trained on, as style vectors of a new style are calculated by applying the pre-trained encoder to examples of that style.

Evaluation of generative models for creative tasks is often challenging, as appreciation is aesthetic and therefore hard to capture. Previous work often uses subjective listening tests to rate the preference of generated samples of different models [8, 15, 16]. Other work calculates musical features on generated and original fragments and compare their distributions to assess musicality [7], or measure task



© L. Somacal, P. Riera, D. Fernández Slezak and M. Miguel. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** L. Somacal, P. Riera, D. Fernández Slezak and M. Miguel, “Symbolic music style transfer via latent space transformations: model and evaluation”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.

related features (e.g.: chroma in accompaniment generators [2]). In style transfer, a common approach has been comparing the predictions of style classifiers between original and modified fragments [10, 11, 15].

As part of this work we also present new evaluation methods to incorporate three distinct aspects of evaluation for the style transfer task: whether the generated fragment presents the target musical style, whether the generated fragment is musical, and whether the generated fragment still resembles the input. Through these metrics we can inspect the effects of the transformation process in various and possibly conflicting aspects of style transfer.

## 2. METHODS

### 2.1 Datasets

We used two different datasets for training and testing the model. For evaluation, we wanted to have music styles that would provide different style transfer challenges. In this work, we define a musical style as any set of abstract features that are common in a set of musical fragments (see 2.4 for more detail).

For training the model, we looked for a large dataset of symbolic data. We used the Lakh Midi Dataset [17], an open dataset of MIDIs with metadata, including style tags from different databases. We used the style tags from [musicbrainz.org](https://musicbrainz.org) (`mb_tags`). We then selected MIDIs whose style tags were *classic pop and rock*, *pop*, *folk* or *classical* because there were a significant quantity of each (more than 500 for each tag). The MIDI files in the dataset needed to be adapted to the input format of our model: extracting a melody and bass track and dividing them into fragments of 4 bars of  $\frac{4}{4}$  with the semiquaver as a minimum note (see 2.2 for details). This was done with `MIDI miner` [18]. After processing them we obtained a total of 155,037 music fragments. These were used to train an initial pre-trained model (named *pre*).

For evaluation, we used the symbolic dataset of MIDIs KernScores [19], a library of virtual musical scores in diverse symbolic data formats. In particular, we selected 4 styles that have different rhythmic and melodic complexities. Ordering from least to most complex, our evaluation styles are: Bach’s chorals, Frescobaldi’s canzonis, Mozart piano sonatas and ragtimes. We expected that transferring style between styles of similar complexity would yield better results. After balanced sampling of the styles and processing them, we obtained 2032 fragments that were split in train (80%), test (10%) and validation (10%) sets.

To test the generalizability of the pre-trained model (*pre*), we used the train and test splits of the evaluation dataset (KernScores) to generate a new model by fine-tuning it. We call this model *fine*. We evaluated both models by transferring between the evaluation styles using the musical fragments in the validation set.

### 2.2 Model and music representation

To transform the style of a song, we propose to adapt the model from Guo et al. [20]. In this work, they present a

model for adding or subtracting musical tension to 4-bar musical fragments. Musical tension refers to a aesthetic property where music feels "tense" and requires resolution, achieved by returning to common themes. The model consists of a Variational Autoencoder (VAE) [21] where tension is changed by moving in the latent vector space. The original model takes as input a symbolic representation of the fragment and outputs a fragment in the same representation plus two additional outputs predicting the musical tension of the fragment. The input are matrices of 1s and 0s with 64 rows as time units (semiquavers, spanning 4 bars) and 89 columns indicating pitch and note changes (rhythm). This input is limited to represent two monophonic voices: a melody voice and a bass voice. The 89 columns are divided to represent the two voices: 73 possible pitches for the melody and 12 for the bass, plus 2 columns for each voice indicating silent time units (rests) and note onsets (rhythm). The two rhythm columns are used to indicate when the same note is repeated consecutively.

As a VAE, the model consists of an encoder that transforms the input into a latent space and a decoder that reconstructs the input from this space. The encoder consists of two bidirectional layers made of gated recurrent units (GRU) [22]. The first one has  $64 \times 512$  dimensions and the second one  $1 \times 512$ . That is followed by two separated dense layers (with Linear + RELU activation functions) of  $1 \times 96$ , representing the mean and the standard deviation of the latent space values, respectively. Thus, the latent space has  $1 \times 96$  dimensions.

After sampling a latent space vector (given the mean and standard deviation obtained by the encoder), the decoder has a repeat vector of  $64 \times 96$ , two bi-GRU of  $64 \times 256$  and the output of six dense layers (Linear + RELU) representing melody pitch (with silence), melody rhythm, bass pitch (with silence), bass rhythm and two additional outputs related to musical tension. In this work we use the same model architecture without the two additional outputs. Therefore, the output is equivalent to the input.

### 2.3 Transfer style models

As stated in section 2.1, we trained two style transfer models. The model named *pre* was trained on our subset of the Lakh Midi Dataset following the procedure and training parameters from the original model by Guo et al. [20]. The model named *fine* is the same model as *pre*, fine-tuned to autoencode the train subset of the KernScores dataset.

To perform style transfer, we propose modifying the encoding of the input fragment  $m$  with a style characteristic vector  $v_s$ . To calculate them, given the models, we encoded all the fragments from the validation split of the 4 evaluation styles in the latent space. We calculate the characteristic vector  $v_s$  as the average of the encoding of all the fragments of style  $s$ . Thus, given a fragment  $m$  of style  $s_i$ , if we want to transform it to  $s_j$ , we encode it into the VAE, we add  $v_j$ , subtract  $v_i$  weighted by a value  $\alpha$  between 0 and 1, and finally, we decode this result, i.e.,

$$t_{s_i, s_j}(m) = \text{decode}(\text{encode}(m) + \alpha(v_j - v_i)) \quad (1)$$

In equation 1,  $\alpha$  parameterizes the strength of the transformation. During evaluation, we will explore the effect of different values of the  $\alpha$  parameter on the generated fragments.

## 2.4 Modeling of musical style

We need to characterize the style from a set of pieces in order to measure and determine whether or not they belong to a style. To do so, we consider the work of Rodríguez Zivic et al. [23] in which they distinguish compositions from different western historical periods according to the distribution of successions of 2 melodic intervals. To define the aesthetics of the style we can ask, given a melodic interval  $x$  (with  $x \in \{-12, \dots, 12\}$  semitones), what is the probability that the next interval is  $y$  (with  $y \in \{-12, \dots, 12\}$ ). To compute these probability distributions, we compute for each subset of fragments  $M_s$ , representing a style  $s$ , a  $25 \times 25$  matrix  $\sigma^i(M_s)$  where each element  $\sigma_{xy}^i$  is the number of occurrences of the interval  $y$  occurring after an interval  $x$ . We define the distribution of music intervals in  $M_s$  as  $\Sigma^i(M_s)$ , by normalizing  $\sigma^i(m)$ , i.e.,

$$\Sigma^i(M_s) = \frac{\sigma^i(M_s)}{\sum_{x,y} \sigma_{xy}^i(M_s)} \quad (2)$$

Similarly, we propose to characterise a style in terms of the distribution of bigrams of rhythmic patterns. In particular, we consider the possible patterns that can occur in four time units. If we note with a 1 when a new note starts (same or different pitch) and with a 0 otherwise, we have 16 possible patterns given by the combinations of ones and zeros in 4-unit arrays. Analogously to the case of the intervals, to compute the distributions we consider for each subset of fragments  $M_s$  the matrix  $\sigma^r(M_s)$  with shape  $16 \times 16$  where each element  $\sigma_{xy}^r$  is the number of times that the rhythmic pattern  $x$  is followed by the rhythmic pattern  $y$ . The distribution of rhythmic patterns is notated as  $\Sigma^r(M_s)$ .

## 2.5 Evaluation

We propose three evaluation metrics to assess the quality of the generated fragment. We will evaluate whether the generated music fragment has attributes of the target style, is musical and keeps the identity of the input fragment.

We first look to whether the generated musical fragment belongs to the target style. To do so, we propose calculating interval and rhythmic distributions ( $\Sigma^i$  and  $\Sigma^r$ , respectively) for a single fragment  $m$  (considering  $M = \{m\}$ ). Then, we can measure the distance between a fragment and a style. In particular, given a distance measure  $\delta$  (in this case, we use *optimal transport* [24]), we define the difference  $\Delta$  between fragment and style as:

$$\Delta(m, M_s) = \sum_{k \in \{i, r\}} \delta(\Sigma^k(m), \Sigma^k(M_s)) \quad (3)$$

We will consider a transformation to be successful if the generated fragment  $m'$  is closer to the target style than the original fragment  $m$ , that is:

$$\Delta(m', M_s) < \Delta(m, M_s) \quad (4)$$

We want to assess whether the generated fragment remains musical. The concept of musicality is both culturally universal and varied, making it difficult to formalize [25]. In this work, we assume that there are implicit rules in the use of melodic intervals and rhythmic patterns that make a sequence of notes sound musical. We propose to estimate these rules considering a *universal style* formed by the balanced sum of the fragments of the different styles of a dataset. Thus, we define the *universal style*  $M_u = M_{s_1} \cup M_{s_2} \cup \dots$ , with  $M_i$  the subset of fragments of style  $i$ . Particularly, we took the same number of fragments of the four evaluation styles. We can then calculate the distance between a fragment  $m$  to  $M_u$  as we saw in equation 3 but in this case, we take  $\delta$  as the probability of sampling  $m$  from  $M_u$ . Specifically, we define the sampling probability proportional to

$$\delta(m, M_u) = \sum_{x,y} \log(\Sigma_{xy}^k(M_u)) \sigma_{xy}^k(m) \quad (5)$$

where  $k$  is  $i$  (intervals) or  $r$  (rhythmic patterns). Equation 5 considers that each melodic interval and rhythmic pattern bigram ( $\sigma_{xy}^k$ ) is sampled independently from the musicality distribution  $\Sigma^k(M_u)$ .

To evaluate the musicality of a fragment, we assembled a new set of fragments with permutations of the notes of the original fragment (assuming randomness is detrimental to musicality). In total, 20 new fragments were created for each original fragment. Then, we calculate the distance of each of the fragments with the *universal style*. We expect that a musical fragment will be closer to the *universal style* than its permutations. Thus, we define the musicality of a fragment as the percentage of permutations that are less musical (that are less likely to be sampled from the *universal style* distribution).

The third issue to evaluate is the similarity of the transformed fragment to its original. To do it, we propose to generate a similarity ranking between the original fragment against the set composed of the transformed fragment and all other fragments of the original style. We then consider that the transformed fragment retained characteristics of the original fragment the higher it appears in the ranking. The similarity score is defined as

$$\text{sim}_\sigma(m', m, M) = 1 - \frac{R(\sigma(m, m'), M^*)}{\#M} \quad (6)$$

$$M^* = \{\sigma(m, x) | x \in M \setminus \{m\}\} \quad (7)$$

The similarity score is 1 minus the position normalized by the cardinal of the subset. This score is bound to 0 and 1, being 1 the best value. In equation 6,  $\sigma$  is a measure of similarity between two fragments and  $R(x, M)$  the position of  $x$  in the ranking generated by  $M \cup \{x\}$ . As  $\sigma$ , we count for each time instant how many semitones the notes differ between one fragment and the other. A rest compared with a note is considered as 12.

To evaluate the performance of our models, we generate a transformed version of every fragment from the validation set (10% of the Kern dataset) onto every target style

different than the fragment’s style. Then we calculate these metrics on each generated fragment and summarize them by source-target style pairs.

For the first evaluation, for each pair of source-target styles we calculate the percentage of generated fragments that became closer to the target style. In Figure 1 (top), we present the distribution of these percentages for the 12 style combinations. To evaluate musicality, for each transferred fragment  $m'$ , we calculate the percentage of permutations of the original fragment that are more unlikely to be generated by the *universal* distribution than  $m'$ , and then average these percentages per style pair. Figure 1 (middle), presents the distribution of these percentage values. Similarly, for the similarity, we plot the distribution of the values of the function defined in the equation 6. These evaluations are performed for both models at three different transformation strengths ( $\alpha$  in equation 1).

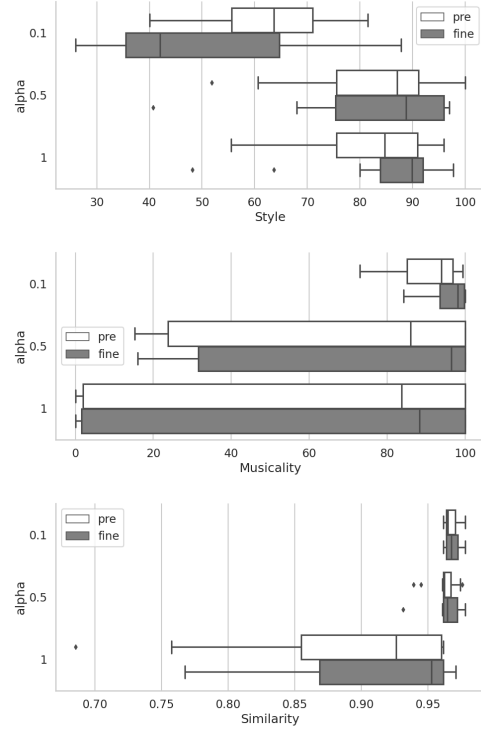
### 3. RESULTS

Comparison between models at different values of  $\alpha$  are presented in Figure 1. In Figure 1 (top) we show the results for style closeness evaluation. We can see how as  $\alpha$  gets larger, more transformed fragments are closer to the target style. We observe no noticeable differences between the *pre* and *fine* models, except with the small  $\alpha$ , where *pre* performs better. In Figure 1 (middle) we present the musicality evaluation. We can observe how musicality degrades as  $\alpha$  increases. A larger  $\alpha$  yields a larger transformation, which may yield less musical results. Nonetheless, most cases are above 80%. Furthermore, the *fine-tuned* model performs slightly better than the *pre* fine-tuning one. Finally, Figure 1 (bottom) displays the similarity evaluation. As in the musicality evaluation, with a large  $\alpha$  value the performance is worse but still good. Likewise, between the two models, the *fine-tuned* one is also slightly better than the other.

### 4. DISCUSSION

In this paper, we addressed the problem of music style transfer with symbolic data. We used a VAE approach where we operated in the latent space using style-characteristic vectors. We tested performing style-transfer between multiple styles with a single model. We further tested generalization of the approach by comparing a model trained on a general music dataset and another one fine-tuned on our 4 evaluation styles. We also evaluated the relevance of the parameter  $\alpha$  that regulates the strength of the transformation. Furthermore, we proposed evaluating the performance of the transfer models on three accounts with three novel metrics: style approach, musicality and similarity with the original fragment.

Our models managed to generate new fragments that remained musical, kept the identity of the original fragment and that were also closer to the target style. This happened for certain values of the transfer strength parameter  $\alpha$  (0.1 and 0.5). As expected, we observed that a greater strength implies more style approach at expense of musicality. This



**Figure 1:** Distributions of our evaluation metric scores for different values of  $\alpha$  and the two models: *pre* (white) and *fine* tuned (grey). (top) represents the percentage of generated rolls that are closer to the target style than the original roll. (middle) represents the distributions of musicality results (percentage of randomized rolls that are less musical than the generated). (bottom) shows the distribution of the fidelity function (1 being the best score).

shows the importance of the hyper-parameter to calibrate the results. Moreover, we noticed that the model trained on a general music dataset was successful even on the distinct set of evaluation styles. Fine-tuning the model on these styles only provided marginal improvements. This leads us to believe that the proposed VAE approach can be used as a single-model solution for multi-style transfer. Yet, when observing the performance between specific source-target style pairs, we noticed performance varied. In particular, the model struggled to transform between Mozart and Ragtime, contrary to our expectation that styles with similar complexity would yield better results.

Future work would benefit from integrating the different approaches used in the literature to address the style transfer task. Firstly, we suggest validating our proposed metrics with listener surveys. Moreover, it would be valuable to compare our metrics with those used in previous work [8]. Regarding modeling, our transformation method could benefit from the latent space disentanglement to represent style as shown in [11], as well as comparing the style-specific vs. general approaches.

From our results, we believe that latent space manipulation can be a powerful tool to develop models that provide fine control for composers to experiment, generalize to multiple styles and are easy to extend.

## 5. REFERENCES

- [1] C. Payne, “Musenet,” [openai.com/blog/musenet](https://openai.com/blog/musenet), 2019, accessed: 2024-10-30.
- [2] O. Cifka, U. Şimşekli, and G. Richard, “Groove2groove: One-shot music style transfer with supervision from synthetic data,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2638–2650, 2020.
- [3] “Udio,” <https://www.udio.com/>, 2024, accessed: 2024-10-30.
- [4] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, “Jukebox: A generative model for music,” *ArXiv*, vol. abs/2005.00341, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:218470180>
- [5] I. Malik and C. H. Ek, “Neural translation of musical style,” *ArXiv*, vol. abs/1708.03535, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:23728768>
- [6] A. Roberts, J. Engel, Y. Mann, J. Gillick, C. Kayacik, S. Nørly, M. Dinculescu, C. Radebaugh, C. Hawthorne, and D. Eck, “Magenta studio: Augmenting creativity with deep learning in ableton live,” in *Proceedings of the International Workshop on Musical Metacreation (MUME)*, 2019. [Online]. Available: [http://musicalmetacreation.org/buddydrive/file/mume\\_2019\\_paper\\_2/](http://musicalmetacreation.org/buddydrive/file/mume_2019_paper_2/)
- [7] H. H. Mao, T. Shin, and G. Cottrell, “Deepj: Style-specific music generation,” in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 2018, pp. 377–382.
- [8] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, “Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/11312>
- [9] C. Lu and S. Dubnov, “Chordgan: Symbolic music style transfer with chroma feature extraction,” in *Proceedings of the 2nd Conference on AI Music Creativity (AIMC)*, Online, 2021, pp. 18–22.
- [10] G. Brunner, Y. Wang, R. Wattenhofer, and S. Zhao, “Symbolic music genre transfer with cyclegan,” in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2018, pp. 786–793.
- [11] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, “MIDI-VAE: modeling dynamics and instrumentation of music with applications to style transfer,” in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 747–754. [Online]. Available: [http://ismir2018.ircam.fr/doc/pdfs/204\\_Paper.pdf](http://ismir2018.ircam.fr/doc/pdfs/204_Paper.pdf)
- [12] X. Hou, L. Shen, K. Sun, and G. Qiu, “Deep feature consistent variational autoencoder,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 1133–1141.
- [13] M. Scarpiniti, E. Massaro, D. Comminiello, and A. Uncini, *Generating New Sounds by Vector Arithmetic in the Latent Space of the MelGAN Architecture*. Singapore: Springer Nature Singapore, 2023, pp. 3–15. [Online]. Available: [https://doi.org/10.1007/978-981-99-3592-5\\_1](https://doi.org/10.1007/978-981-99-3592-5_1)
- [14] Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of gans for semantic face editing,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9240–9249.
- [15] W. T. Lu and L. Su, “Transferring the style of homophonic music using recurrent neural networks and autoregressive model,” in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, 2018, pp. 740–746.
- [16] A. Roberts, J. H. Engel, C. Raffel, C. Hawthorne, and D. Eck, “A hierarchical latent vector model for learning long-term structure in music,” *Proceedings of the 35th International Conference on Machine Learning*, vol. abs/1803.05428, 2018. [Online]. Available: <http://arxiv.org/abs/1803.05428>
- [17] C. Raffel, “Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching,” Ph.D. dissertation, Columbia University, USA, 2016. [Online]. Available: <https://doi.org/10.7916/D8N58MHV>
- [18] R. Guo, D. Herremans, and T. Magnusson, “Midi miner - A python library for tonal tension and track classification,” *Late Breaking Demo of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, vol. abs/1910.02049, 2019. [Online]. Available: <http://arxiv.org/abs/1910.02049>
- [19] “Kernscores,” <https://kern.humdrum.org/>, accessed: 2024-10-30.
- [20] R. Guo, I. Simpson, T. Magnusson, C. Kiefer, and D. Herremans, “A variational autoencoder for music generation controlled by tonal tension,” in *2020 Joint Conference on AI Music Creativity*, 10 2020.
- [21] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: <http://arxiv.org/abs/1312.6114>

- [22] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–decoder approaches,” in Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, D. Wu, M. Carpuat, X. Carreras, and E. M. Vecchi, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: <https://aclanthology.org/W14-4012>
- [23] P. Zivic, F. Shifres, and G. Cecchi, “Perceptual basis of evolving western musical styles,” Proceedings of the National Academy of Sciences of the United States of America, vol. 110, no. 24, pp. 10 034–8, 05 2013.
- [24] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, “Pot: Python optimal transport,” Journal of Machine Learning Research, vol. 22, no. 78, pp. 1–8, 2021. [Online]. Available: <http://jmlr.org/papers/v22/20-451.html>
- [25] S. E. Trehub, J. Becker, and I. Morley, “Cross-cultural perspectives on music and musicality,” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 370, no. 1664, p. 20140096, 2015. [Online]. Available: <https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2014.0096>

# SURPRISING PATTERNS IN MUSICAL INFLUENCE NETWORKS

Flavio Figueiredo  
UFMG

Tales Panoutsos  
UFMG

Nazareno Andrade  
Enveritas

flavio@dcc.ufmg.br tales.panoutsos@dcc.ufmg.br nazareno@gmail.com

## ABSTRACT

Analyzing musical influence networks, such as those formed by artist influence or sampling, has provided valuable insights into contemporary Western music. Here, computational methods like centrality rankings help identify influential artists. However, little attention has been given to how influence changes over time. In this paper, we apply Bayesian Surprise to track the evolution of musical influence networks. Using two networks—one of artist influence and another of covers, remixes, and samples—our results reveal significant periods of change in network structure. Additionally, we demonstrate that Bayesian Surprise is a flexible framework for testing various hypotheses on network evolution with real-world data.

## 1. INTRODUCTION

In recent years, the study of how artists influence [1,2], collaborate [3–10] and sample [11, 12] one another has led to valuable knowledge on the evolution of contemporary western music<sup>1</sup>. A large fraction of these studies rely on social network analysis. Music data is studied via the relationships between songs, albums, or artists in these networks, or *musical influence networks*.

When we look at influence networks evolving (e.g., considering the rise of new artists and changes in network structure), one question arises: considering the evolution of the network, *what was surprising*, as time passed. Take as one example The Beatles. Saying The Beatles are highly influential nowadays is relatively easy. However, understanding The Beatles early in their career may be more complex. Here, we propose both a formal definition of Bayesian Surprise [13–15] for rankings in complex networks. We aim to answer what is (or was) surprising as the musical influence network grew.

To perform our study, we explore two human-curated datasets of music influence. One from the AllMusic Guide<sup>2</sup> and another from WhoSampled<sup>3</sup>. Our study uses

<sup>1</sup> A limitation primarily due to available data. However, some exceptions exist such as: [7,9,10]

<sup>2</sup> <https://www.allmusic.com/>

<sup>3</sup> <https://www.whosampled.com/>



© Flavio Figueiredo, Tales Panoutsos, Nazareno Andrade. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Flavio Figueiredo, Tales Panoutsos, Nazareno Andrade, “Surprising Patterns in Musical Influence Networks”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.

both datasets via temporal views of Pagerank [16] and Disruption [17] centrality scores.

Bayesian Surprise allows the testing of multiple hypotheses over a time-evolving dataset. We are the first to show that it provides a principled approach to combine both centrality scores and multiple hypotheses in a single measure for music influence – or even social – networks. Combining different centralities allows for a more expressive analysis, as each has its interpretation. Pagerank may be seen as an overall influence on an artist. In contrast, even non-influential artists may be disruptive<sup>4</sup> [2]. The framework allows us to combine different centralities and hypotheses to unveil interesting trajectories of influence.

## 2. RELATED WORK

In the past, several authors have looked into musical influence networks. These efforts ranged from looking at Classical music [5], to Jazz [6, 8], Broadway Songs [9], as well as Popular Brazilian Music [7, 10, 18]. With regards to larger and more general datasets, the AllMusic guide and the WhoSampled website have also been the focus of several prior endeavors [1, 2, 19–21].

In these previous efforts, musical influence is captured by a graph, modeling the network. Here, nodes are either artists or songs. Edges encode the influence (e.g., sampled, was influenced by, or collaborated with). With these graphs, social network analysis [22] analysis plays a pivotal role in understanding the importance of music.

Most of these analyses, however, fail to capture temporal aspects of the influence network. With this regard, Andrade and Figueiredo [6] looked into the latent structure of Jazz collaborations via Markovian models. Even though temporal ordering is necessary for Markovian models, the authors do not look into social network properties over time. Here, Shalit, Weinshall, and Chechik [1] consider time in the evaluation for a method to predict interest, but no insights into network evolution are provided. Other authors similarly limit their temporal analyses to overall descriptive statistics per year [2, 11, 18].

This is the first work to adapt Bayesian Surprise as a tool that provides insights into network evolution. The measure can pinpoint each node (artist) point in time where the topology of the network influenced the artist’s centrality. Next, we present the background required to understand our methodology.

<sup>4</sup> Disruption focuses on an artist’s capacity of aggregating influence on itself diverging influence from neighboring nodes.



### 3. NETWORKS AND SURPRISE

We now discuss our notation and ranking scores for the musical influence networks, Section 3.1. We also take the time to explain Bayesian Surprise and how we adapt it for rankings in Section 3.2.

#### 3.1 Influence Networks and Centrality

Let  $\mathcal{G}_t = \{\mathcal{V}_t, \mathcal{E}_t\}$  be a weighted directed graph.  $\mathcal{V}_t$  captures the set of nodes, whereas  $\mathcal{E}_t$  captures edges. For both datasets this graph captures influences across artists. The subscript  $t$  here indicates that our graph is time evolving in discrete steps (decades or years). In other words,  $t$  defines the snapshot time of our graph.  $\mathcal{G}_{2010}$ , is the graph of influences formed up to and including 2010.

We assume nodes are enumerated, i.e.,  $n_i \in \mathcal{V}_t$  determines a node  $n_i$  with each node defining an artist. Edges are similarly defined,  $e_k \in \mathcal{E}_t$  with  $e_k = (n_i, n_j)$  representing an edge from node  $n_i$  to node  $n_j$ . Our graphs are directed, thus i.e.,  $(n_i, n_j) \neq (n_j, n_i)$ . As we are studying musical influence across artists, we impose that  $j \neq i$ , stating that an artist may not influence itself<sup>5</sup>. Both node and edge ids are maintained in subsequent snapshots. We also define edge weights,  $weight(\mathcal{G}_t, e_k)$ , the number of times one artist mentions the other as an influence. Finally, we note that our graphs evolve cumulatively, i.e.,  $weight(\mathcal{G}_{t'}, e_k) \geq weight(\mathcal{G}_t, e_k)$ , when  $t' > t$ . At each time-step  $t$ , nodes are ranked according to Pagerank [16] and Disruption [17] centrality. That is, each node has a score  $s_{t,i} = score(\mathcal{G}_t, n_i)$ .

The first score, Pagerank [16], outputs a probability distribution over the entire set of nodes in the network. If we consider a random walker on the graph, this probability distribution may be interpreted as the chance a random walker “lands” on node  $n_i$  after sufficient steps. The walker jumps from node  $n_i$  to  $n_j$  with a chance proportional to  $weight(\mathcal{G}_t, e_k)$ , where  $e_k = (n_i, n_j)$ .

Disruption [17], focuses on the capability of a node on concentrating influence. That is, considering that artists cite past influences, when some focal artist node,  $n_a$ , is mostly cited by-itself it is deemed are more disruptive. That is, if we consider that for  $n_a$  there are some nodes that reference  $a$ ’s work and at least one of  $a$ ’s own influences. Some other nodes cite only  $a$ . When  $a$  is cited in isolation, disruption increases. When it is not, it decreases. The final score is simply the difference between the fraction of nodes in each group. Positive numbers indicate more disruption.

#### 3.2 Bayesian Surprise

The basis of our analysis is the measure of Bayesian Surprise [13–15], that we now explain. Consider an arbitrary dataset  $\mathcal{D}$  to understand Bayesian Surprise. Now, let us assume that data points come from some probability distribution  $x_i \sim Dist(\theta)$ , with  $\theta$  being the parameters of the probability density function (pdf), i.e.,  $p(x | \theta)$ .

A major part of statistics is focused on finding the best parameters  $\theta$  for an overall model (pdf) and a dataset. Assuming that data points are independent, this comes from maximum likelihood estimation (MLE) is achieved by maximizing:  $\theta = \arg \max_{\theta'} \prod_{i=1}^{|\mathcal{D}|} p(x | \theta')$ . Bayesian Statistics provides an approach to tackle such estimation problem via the usage of prior distributions over parameters:  $p(\theta)$ . By employing Bayes Theorem, we are able to derive another pdf over the parameter space:

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D} | \theta)p(\theta)}{\int_{\theta} p(\mathcal{D} | \theta)p(\theta)d\theta} \quad (1)$$

Here,  $p(\theta | \mathcal{D})$  is called the posterior. It states, after observing the data, what is the probability distribution (which captures uncertainty) over the parameters.

Bayesian Surprise [13–15] is defined as a measure of divergence between prior and posterior distributions. In other words, it captures surprise as the number of bits in this divergence. If prior and posterior are close, we are not surprised after observing the dataset  $\mathcal{D}$ .

Our focus will be on measuring surprise based on a node’s relative position in a ranking. Now, consider a ranking such as  $a_3, a_1, a_2$ . Here, some artist  $a_3$  was ranked above all others, and  $a_2$  was ranked above  $a_1$ . The relative position of  $a_3$  is 1, as all other nodes are below it.  $a_1$  has a relative position of  $\frac{2}{3}$  and  $a_2$  has a relative position of  $\frac{1}{3}$ .

Our dataset  $\mathcal{D}$  will thus be the set:  $\{x_1 = \frac{2}{3}, x_2 = \frac{3}{3}, x_3 = \frac{1}{3}\}$ . In other words,  $x_i = \frac{g_{t,i}}{|\mathcal{V}_t|}$ , where  $g_{t,i}$  is the set of nodes with a score equal to or greater than node  $i$  at time  $t$ , and  $|\mathcal{V}_t|$  is the number of nodes at time  $t$ . Being simple fractions in the  $[0, 1]$  range, each one of these values may be modeled as a Bernoulli distribution, i.e.,  $x_i \in \{0, 1\}$  and  $x_i \sim Bernoulli(\theta)$ . To measure surprise, we can employ the Beta distribution as a prior as it is well known that the Beta distribution is the conjugate prior to the Bernoulli.

Thus, assume a prior such that  $\theta \sim Beta(\alpha_{\theta}, \beta_{\theta})$ , with  $p(\theta)$  being the density function. The prior has parameters, denominated as hyper-parameters  $\alpha_{\theta}$  and  $\beta_{\theta}$ , representing the hypothesis we use for computing surprise. Being the Beta the conjugate prior, the posterior is proven to be  $\theta | \mathcal{D} \sim Beta(\alpha_{\theta|\mathcal{D}}, \beta_{\theta|\mathcal{D}})$ , with  $p(\theta | \mathcal{D})$  being its density. Here,  $\alpha_{\theta|\mathcal{D}} = \alpha_{\theta} + g_{t,i}$ . Also,  $\beta_{\theta|\mathcal{D}} = \beta_{\theta} + (|\mathcal{V}_t| - g_{t,i})$ .

Bayesian Surprise [13–15] thus is computed as the divergence between prior and posterior distributions. A common choice is to use the Kullback-Leibler Divergence:

$$D_{KL}(p(\theta|\mathcal{D}) || p(\theta)) = \int_{\theta} p(\theta|\mathcal{D}) \log_2\left(\frac{p(\theta|\mathcal{D})}{p(\theta)}\right)d\theta \quad (2)$$

Due to the use of conjugate priors, closed-form solutions for the above measure already exists [15, 23, 24]. In particular, ours is the divergence of two Beta distributions [23].

Following the properties of the divergence, surprise will always be positive. Lower values indicate less surprise, as we have better priors (or hypotheses) about the data. It equals zero in the unrealistic setting with a prior perfectly capturing the posterior. This is unrealistic, as the posterior itself transforms the prior after observing some data.

Now, the above definition is valid for a single choice of prior, or hypothesis, as we now call it. We can also

<sup>5</sup> In one of our datasets, the WhoSampled data, artists may sample and cover their own work, and thus we remove this characteristic from graph.



	Weighted	# Artists	# Edges	Range	$\Delta$
AllMusic	False	32,568	119,961	1940-2019	10
WhoSampled	True	166,016	603,487	1940-2019	1

**Table 1:** Overview of the dataset.  $\Delta$  refers to the granularity in years. When  $\Delta = 10$ , each snapshot is a decade.

consider a set of hypotheses:  $\mathcal{H} = \{\theta_h\}$ . Each hypothesis is a prior distribution:  $\theta_h \sim \text{Prior}(\cdot)$ . In the case of the Beta distribution prior, our set will comprise different choices for the hyper-parameters:  $\alpha_{h,\theta}$  and  $\beta_{h,\theta}$ . That is,  $\theta_h \sim \text{Beta}(\alpha_{h,\theta}, \beta_{h,\theta})$ . Given that divergence is additive in nature, the overall surprise for various hypotheses is:

$$\text{Sup}(\mathcal{H}, \mathcal{D}) = \sum_{h=1}^{|\mathcal{H}|} D_{KL}(p(\theta_h|\mathcal{D}) || p(\theta_h)) \quad (3)$$

We currently have the required definitions for computing surprise on music influence networks. Nevertheless, we point out that these definitions are general enough to be adapted to different domains of the musical information retrieval literature. As stated, we shall compute surprise via *rankings* on the network. That is, each artist is ranked according to some centrality score. Each hypothesis is a prior distribution of the position where we expect the node’s relative to be in the ranking. This expected position comes from previous snapshots of the network as we now detail.

In this paper, we consider two hypotheses. The first is called *Past Rank*. In this hypothesis, we define that for each node  $i$ :  $\alpha_{i,\theta} = g_{i,t-1}$  and that  $\beta_{i,\theta} = |\mathcal{V}_{t-1}| - \alpha_{i,\theta}$ . Here, the subscript indicates the hypothesis for node  $i$  on timestamp  $t$  as a prior over parameters  $\theta$ . With these hyperparameters, the expected value of the Beta distribution is equal to the relative position of the node on time  $t - 1$ . This comes from the fact that the expected value of the Beta is:  $\frac{\alpha_{i,\theta}}{\alpha_{i,\theta} + \beta_{i,\theta}}$ . Thus, Past Rank assumes that the node’s position will not change from one snapshot to the next.

The second hypothesis is called *Regular Growth*. Here, we state that:  $\alpha_{i,\theta} = \frac{g_{i,t-1}^2}{g_{i,t-2}}$ . And again,  $\beta_{i,\theta} = |\mathcal{V}_{t-1}| - \alpha_{i,\theta}$ . Here, we aim at capturing regularities in changes. To understand it, observe that  $\frac{g_{i,t-1}}{g_{i,t-2}}$  captures the rate of change considering the last two snapshots. Consider, for example, that this rate of change is equal to two. When we multiply the rate by  $g_{i,t-1}$ , we state that we expect the node to rise two places. When the rate is below one, the node will fall in ranking. Thus, this hypothesis places the expected value of the Beta distribution in the position where we expect the node to be given the past two timestamps of the network.

## 4. RESULTS

Our results focus on influence networks. The AllMusic network is publicly available (see [2])<sup>6</sup>. Using a large seed of 73,000 thousand AllMusic URLs present in MusicBrainz<sup>7</sup>, AllMusic’s influence network was captured

<sup>6</sup><https://github.com/flaviovdv/allmusic-disruption>

<sup>7</sup><https://musicbrainz.org/>

	Kendall	Spearman
AllMusic	0.026	0.036
WhoSampled	0.203	0.290

**Table 2:** Correlation coefficients between PageRank and Disruption on the Last Snapshot of the two datasets.

via snowball sampling. After filtering nodes with at least one edge, the network comprises of 32,568 artists connected by 119,961 edges. Edges capture that one artist influenced another and were defined by the editors of the website. In this dataset, each snapshot comprises a decade.

The WhoSampled dataset was provided after an agreement with *WhoSampled.com*. WhoSampled lists for several songs the other songs that either sampled, remixed or covered it. The provided dataset contained a total 1,250,246 songs. For our study, we aggregate from the song level to the artist level. That is, we define an edge between artists that sampled, remixed or covered other artists. The weight of this edge is the total number of interactions (samples + remixes + covers). After filtering the data to consider the same temporal range as AllMusic, we are left with 166,016 nodes and 603,487 edges.

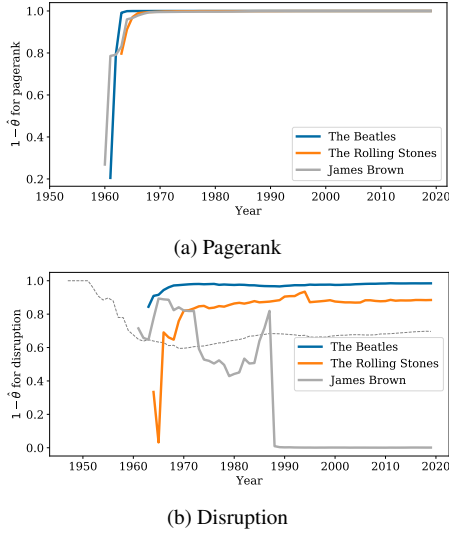
### 4.1 Motivating Surprise

With these datasets, we present some motivating examples of Bayesian Surprise. Our goal here is to show that a single ranking does not tell the whole story of an artist. By combining several rankings in a principled manner, Bayesian Surprise mitigates this issue.

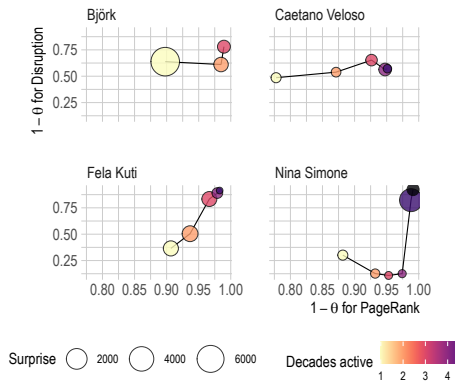
Thus, our first results show that Pagerank and Disruption play different roles when ranking nodes. In Table 2, we present both Spearman [25] and Kendall [26] correlations of the two centrality scores. Scores were measured on the last snapshot for each dataset. Correlations for both datasets are low, thus showing evidence that Pagerank and Disruption scores capture complementary aspects of the network topology. Such a result motivates the need to combine both scores. Bayesian Surprise provides a principled way to achieve this goal. Next, we further explore individual trajectories.

In Figure 1, we show three artists their position in the ranking for WhoSampled. These are three of the most influential artists both scores. Positions are shown as  $1 - \theta$ , thus higher values indicate higher positions. The figure shows that The Beatles quickly rises to influence in terms of Pagerank (Figure 1-a). When we consider Disruption (Figure 1-b), the band takes some time for the band to rise to the top ranks of the scores. The Rolling Stones follows a similar pattern, but with more “jumps”.

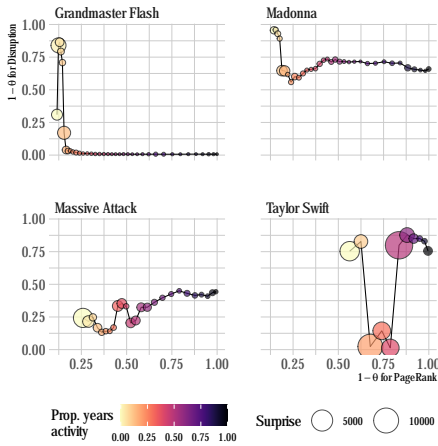
It is also interesting how James Brown both rises and falls in Disruption over time. Eventually, the artist declines in Disruption but remains stable in Pagerank. Looking at only one of the scores would only tell part of this story. Bayesian Surprise can highlight the changes for a single artist, such as James Brown’s trajectory, by adding several hypotheses. We further explore this next.



**Figure 1:** Comparing the position in rankings. The dashed line indicates the position where Disruption changes sign.



**Figure 2:** Trajectories of four artists in AllMusic as a function of their Pagerank, Disruption and surprise.



**Figure 3:** Trajectories of four artists in WhoSampled.

#### 4.2 Making use of Surprise

Next, in Figures 2 and 3, we show the trajectories of four artists in each dataset as connected scatter plots. Here, the surprises of the two hypotheses were summed up. To provide a concise view, we also sum the hypotheses in the Pagerank and Disruption section. The x-axis shows the

position in Pagerank, while the y-axis shows the position in Disruption. Surprise is shown by the size of each point.

Starting from Figure 2, some artists such as Bjork are surprising early in their careers (notice that larger points are the brighter ones). That is, the artist quickly rises to top positions in each centrality score. Further movements from the artist are less surprising, but still highlighted (when we compare with the size of other artists) as Bjork continues to rise to the top of PageRank and Disruption rankings. Artists like Caetano Veloso have smoother trajectories, being thus less surprising. To understand this, recall both of our hypotheses are functions of the previous snapshots. Hypothesis *Reg. Growth*, in particular, aims at capturing these smoother transitions. Nina Simone shows an example of abrupt transitions being highlighted due to a surprisingly higher disruption later in her career.

In WhoSampled (Fig. 3), we are able to observe more granular trajectories. The Hip Hop pioneer Grandmaster Flash provides an example of an initially highly disruptive career that afterwards gains increasing relevance, but acts as a consolidator. Additionally, his peak and sudden decrease of disruption are surprising, marking rapid creation of a large and consolidated stream of artists. Massive Attack has a somewhat similar trajectory, with most surprise in its early years, but maintaining some disruption, likely being sampled by artists that do not sample its influences. Madonna and Taylor Swift are examples of artists whose trajectories have peak surprise points after their initial years in the dataset. Finally, for these examples changes in Disruption are more highlighted than changes in Pagerank. Given that these are popular artists, we can notice that their evolution in Pagerank scores (x-axis) is quite smooth over time. Thus, they are less surprising.

This is most evident with Taylor Swift, whose trajectory has multiple surprising changes in disruption, denoting points where there is first a wave of artists who are influenced by her and by her influences (drop in disruption), and afterward, a wave of artists who are influenced by her but do not share her influences (increase in disruption on 6th point in her trajectory). When we look into Taylor Swift’s disruption values, they are reduced tenfold from 2011 to 2013 (from 0.1 to 0.01). Later in the dataset, the artist regains her position.

Our previous results motivate Bayesian Surprise over alternatives. The trajectories we presented here exemplify how Surprise can guide the analysis in identifying relevant phenomena in the evolution of dynamic networks.

## 5. CONCLUSIONS

In this paper, we derived a Bayesian Surprise measure based on rankings that was validated on real-world datasets. Our is the first work to provide this measure for social networks and for the music domain. Secondly, our analysis on the temporal nature (via artist trajectories) of music influence networks provider novel insights into how influence evolves based on two centrality scores. In particular, our results help understand music datasets from a historical perspective.

## 6. REFERENCES

- [1] U. Shalit, D. Weinshall, and G. Chechik, “Modeling musical influence with topic models,” in *ICML*, 2013.
- [2] F. Figueiredo and N. Andrade, “Quantifying disruptive influence in the allmusic guide,” in *ISMIR*, 2019.
- [3] J. Park, O. Celma, M. Koppenberger, P. Cano, and J. M. Buldú, “The social network of contemporary popular musicians,” *International Journal of Bifurcation and Chaos*, vol. 17, no. 07, pp. 2281–2288, 2007.
- [4] T. Teitelbaum, P. Balenzuela, P. Cano, and J. M. Buldú, “Community structures and role detection in music networks,” *Chaos: An Interdisciplinary Journal of Non-linear Science*, vol. 18, no. 4, p. 043105, 2008.
- [5] A. Bae, D. Park, Y.-Y. Ahn, and J. Park, “The multi-scale network landscape of collaboration,” *PloS one*, vol. 11, no. 3, p. e0151784, 2016.
- [6] N. Andrade and F. Figueiredo, “Exploring the latent structure of collaborations in music recordings: A case study in jazz,” in *ISMIR*, 2016.
- [7] C. Gunaratna, E. Stoner, and R. Menezes, “Using network sciences to rank musicians and composers in brazilian popular music,” in *Proc. ISMIR*, 2011.
- [8] P. M. Gleiser and L. Danon, “Community structure in jazz,” *Advances in complex systems*, vol. 6, no. 04, pp. 565–573, 2003.
- [9] B. Uzzi and J. Spiro, “Collaboration and creativity: The small world problem,” *American journal of sociology*, vol. 111, no. 2, pp. 447–504, 2005.
- [10] D. d. L. Silva, M. M. Soares, M. Henriques, M. S. Alves, S. de Aguiar, T. de Carvalho, G. Corso, and L. Lucena, “The complex network of the Brazilian Popular Music,” *Physica A: Statistical Mechanics and its Applications*, vol. 332, 2004.
- [11] N. J. Bryan and G. Wang, “Musical influence network analysis and rank of sample-based music,” in *ISMIR*, 2011.
- [12] J. Van Balen, J. Serrà, and M. Haro, “Sample identification in hip hop music,” in *CMMR*, 2012.
- [13] L. Itti and P. Baldi, “Bayesian surprise attracts human attention,” *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2009.
- [14] —, “A principled approach to detecting surprising events in video,” in *CVPR*, 2005.
- [15] P. Baldi and L. Itti, “Of bits and wows: A bayesian theory of surprise with applications to attention,” *Neural Networks*, vol. 23, no. 5, pp. 649–666, 2010.
- [16] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.” Stanford InfoLab, Tech. Rep., 1999.
- [17] R. J. Funk and J. Owen-Smith, “A dynamic network measure of technological change,” *Management Science*, vol. 63, no. 3, pp. 791–817, 2016.
- [18] F. Falcão, N. Andrade, F. Figueiredo, D. Silva, and F. Morais, “Measuring disruption in song similarity networks,” in *ISMIR*, 2020.
- [19] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. L. Arcos, “Measuring the evolution of contemporary western popular music,” *Scientific reports*, vol. 2, 05 2012.
- [20] B. G. Morton and Y. E. Kim, “Acoustic features for recognizing musical artist influence,” in *ICMLA*, 2015.
- [21] J. Atherton and B. Kaneshiro, “I said it first: Topological analysis of lyrical influence networks,” in *ISMIR*, 2016.
- [22] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.
- [23] W. Penny, “Kl-divergences of normal, gamma, dirichlet, and wishart densities,” [www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps](http://www.fil.ion.ucl.ac.uk/~wpenny/publications/densities.ps), 2001.
- [24] J. Soch and C. Allefeld, “Kullback-leibler divergence for the normal-gamma distribution,” *arXiv preprint arXiv:1611.01437*, 2016.
- [25] C. Spearman, “The proof and measurement of association between two things,” *The American journal of psychology*, vol. 100, no. 3/4, 1904.
- [26] M. G. Kendall, “A new measure of rank correlation,” *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

# I’VE HEARD THIS BEFORE: INITIAL RESULTS ON TIKTOK’S IMPACT ON THE RE-POPULARIZATION OF SONGS.

Breno Matos<sup>1,2</sup>

Francisco Galuppo<sup>1,2</sup>

Rennan Cordeiro<sup>1,2</sup>

Flavio Figueiredo<sup>1</sup>

<sup>1</sup> Universidade Federal de Minas Gerais

<sup>2</sup> Kunumi

{brenomatos, franciscogaluppo, rennancordeiro, flavioovdf} @dcc.ufmg.br

## ABSTRACT

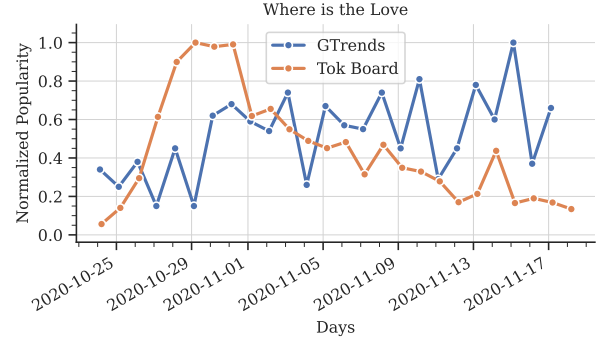
With over a billion active users, TikTok’s video-sharing service is currently one of the largest social media websites. This rise in TikTok’s popularity has made the website a central platform for music discovery. In this paper, we analyze how TikTok helps to revitalize older songs. To do so, we use both the popularity of songs shared on TikTok and how the platform allows songs to propagate to other places on the Web. We analyze data from TokBoard, a website measuring such popularity over time, and Google Trends, which captures songs’ overall Web search interest. Our analysis initially focuses on whether TokBoard can cause (Granger Causality) popularity on Google Trends. Next, we examine whether TikTok and Google Trends share the same virality patterns (via a Bass Model). To our knowledge, we are one of the first works to study song re-popularization via TikTok.

## 1. INTRODUCTION

With the rise of novel social media websites, online music discovery becomes an ever changing ecosystem [1]. While major players such as Spotify and YouTube Music dominate the market, short video-sharing websites like TikTok, the focus of this work, also promote music to end users.

As stated, TikTok is mostly a video-sharing service. After a user uploads a short video, other users can use these snippets to create their videos. In TikTok, songs are present in accompanying video memes or music videos. This phenomenon is a breeding ground for audio to become viral (and thus popular). Nevertheless, many audio snippets on TikTok are from major artists and record labels. With the website becoming a platform for discovering songs from such artists, a question arises: *How do short video-sharing websites impact the revival of older songs?*

To tackle our research goal of understanding TikTok’s impact on the re-popularization of songs, we perform an analysis of songs that were (1) released on or before September 2016 (TikTok’s release date), and (2) trending



**Figure 1:** Popularity trend for the song ”Where is the Love?” from group ”Black Eyed Peas” on both Tok Board and Google Trends. The dates include the peak popularity for both platofms

on TikTok at some point in time. To define this second factor, we use data from TokBoard<sup>1</sup>.

To motivate our work, we cite examples such as ”Dreams”<sup>2</sup> by *Fleetwood Mac* which was released in 1977, and ”Where is the Love?” by *The Black Eyed Peas*<sup>3</sup>. This second example is shown in Figure 1. As reported by news outlets, the viral trend of both songs on TikTok affected the song’s overall popularity online and offline. This is exemplified in the figure when we see the Tok Board curve in a rise-and-fall viral trend, whereas the GTrends curve increased after the song went viral. Motivated by such examples, our study will mainly focus on answering the two research questions described below:

**RQ1:** *Is it possible to predict Web search popularity based on TikTok popularity?* Here, we shall employ the Granger Causality Test [2], in which we evaluate if the TokBoard popularity curve can predict the Google Trends curve. If this is so, we have evidence that TikTok is causing Web search popularity. While this question sheds light on causality and prediction, we still need to understand the viral patterns of both websites. This is why we complement this question with the next one.

**RQ2:** *Do the viral trends of TikTok transfer to Google Trends?* In this question, we shall fit TokBoard and Google



© B. Matos, F. Galuppo, R. Cordeiro and F. Figueiredo. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** B. Matos, F. Galuppo, R. Cordeiro and F. Figueiredo, ”I’ve heard this before: Initial Results on TikTok’s impact on the re-popularization of songs.”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.

<sup>1</sup> tokboard.com – A website aggregating popularity data from TikTok

<sup>2</sup> <https://pitchfork.com/news/watch-mick-fleetwood-recreate-viral-fleetwood-mac-dreams-tiktok/>

<sup>3</sup> <https://newsroom.tiktok.com/en-us/year-on-tiktok-music-2020>

Trend curves with the Bass Model [3]. The Bass Model is a simple differential equation that captures how products get adopted by a population. It has two interpretable parameters to understand virality: (1) Innovation, or who are the adopters that consume a song without influence, and (2) Immitation, who are the adopters influenced by others. After our causal analysis, we aim to understand whether TikTok trends are reflected in Web searches.

Overall, and to the best of our knowledge, ours is the first work to look into the popularity of both TikTok and Web Search based on these two aspects. We also release all code and data for this paper to foster reproducibility.<sup>4</sup> Before discussing our dataset and results, we take the time to summarize related work in the next section.

## 2. RELATED WORK

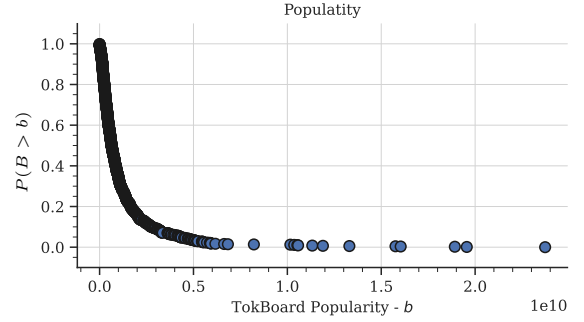
Several authors have looked into popularity patterns for on-line music streaming services. As argued by Greenberg et al. [4], in current times, there is an unprecedented opportunity to investigate the musical listening habits of millions to increase our knowledge of how we consume music.

Narrowing our focus on the impact of social media on music listening, datasets such as the million music tweets and #nowplaying attempted to bridge Twitter and music listening habits [5,6]. As with other cultural products, music is homophilic [7]. The authors of [7] argue that the usage of social media platforms can exacerbate the spread of musical content and reduce its impact. Another interesting case study was done by [8]. The authors show evidence of a growth in nostalgic music consumption during the COVID-19 pandemic [8]. Again, this shows how social media data is a powerful tool for understanding the re-popularization of songs. Complementary to the above efforts, several authors have employed social network analysis in music. Efforts ranged from looking at Classical music [9], to Jazz [10, 11], Broadway Songs [12], as well as Popular Brazilian Music [13–15].

With TikTok being a rather new social media platform, relatively few papers have explored its data and associated phenomena. We now summarize some of these. In [16], the authors perform a characterization of a thousand videos from TikTok. Overall, the authors focus on understanding the types of content shared. Complementary, [17] analyzes TikTok’s impact on collective action. Here, the authors present a case study of the #TulsaFlop phenomenon. The hashtag was associated with a collective action focused on reducing support for a presidential candidate. Other authors propose qualitative [18–22] and quantitative [23] analyses of TikTok.

## 3. THE TOKBOARD DATASET

Since we aim to understand re-popularization, we focused on a smaller subset of the dataset, selected to reduce confounding influences present in the full dataset. Thus, we aimed to remove spurious patterns such as (1) several peaks of popularity, (2) songs not released before TikTok



**Figure 2:** Popularity of TokBoard Songs

was created, and (3) songs where we have evidence that TikTok was the driving aspect of popularity.

Our dataset comes from data collected from TokBoard via a web scraper. Our research group gathered information from all 1,989 songs available in the dataset until May 19, 2022. We then queried each song’s name for Google Trends data displayed on the TokBoard website. This step captures the song’s popularity online and returns results for 1,167 songs we further filtered. To ensure comparable time-series data, we use linear interpolation on TokBoard curves to filter out any missing data. While interpolation may add minor artifacts, it maintains continuity without significant distortion.

In Figure 2, we show the Complementary Cumulative Distribution Function (CCDF) of the popularity of TokBoard songs so far. The  $y$ -axis on this plot captures the fraction of songs with popularity higher than the  $x$ -axis. From the figure, we can see that TokBoard only monitors highly popular content (the scale of the plot is tens of billions). Indeed, the least popular song has a popularity of 11,705 streams. The most popular reaches 23,728,600,741 (over twenty three billion streams). The first, second, and third quantiles were 298,574,069, 637,920,946, and 1,380,260,724, showing that we expect popularity in the hundreds of millions or even billions of streams.

It is essential to point out that even though we collect hundreds of highly popular songs from TokBoard, not all of them will be related to our goals. For instance, not all TikTok songs will be previous hits that went viral in recent years. Moreover, songs may become popular repeatedly, exhibiting several viral-like patterns over time [24, 25].

To focus on the most prominent period of a song, we initially filtered TokBoard popularity time series based on the peak (most popular) day. Thus, we considered points before and after the peak until the popularity of a given date was below 5% of the total popularity. This leads to curves like the ones shown in the Introduction (see Figure 1), and also exemplified in Figure 3. In this second figure, we show the time series for “How Bizarre” by OMC (released in 1995), which was also re-popularized, and “PYRO” by Chester Young & Castion (released in 2019). Although “PYRO,” is popular around July 29th, 2022 (it peaks on TikTok before Google Trends), we must remove it because it was released after TikTok. Thus, it is not a proper example of re-popularization.

<sup>4</sup> <https://github.com/brenomatos/tiktok-lamir>



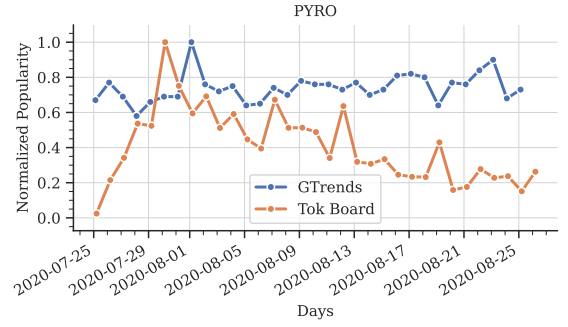
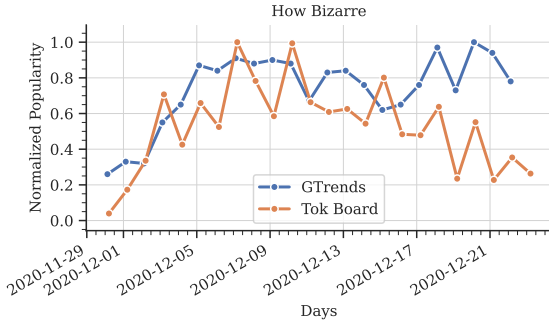


Figure 3: Examples of Peak Focused Time Series on TokBoard

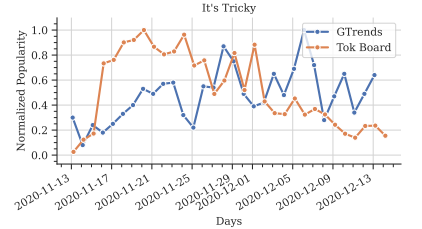
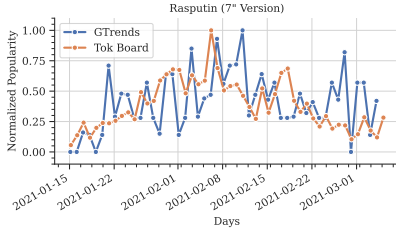
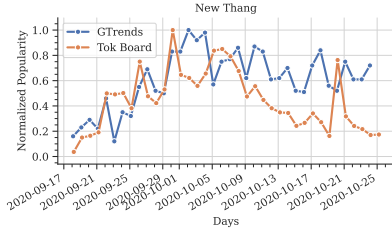
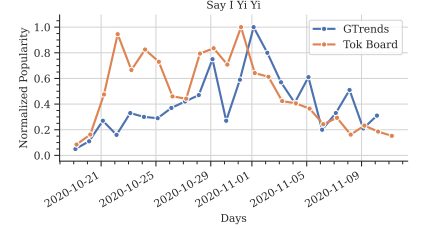
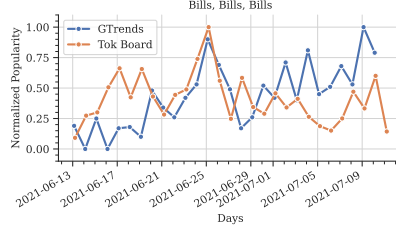
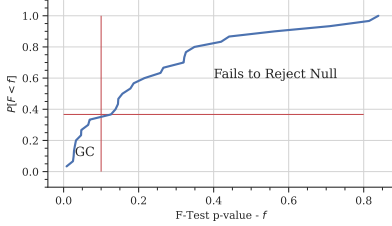


Figure 4: P-values for Granger Causality and Some Examples of Granger Causality

For the above reason, selected songs were released on or before September 2016, when TikTok was released worldwide. This was done in a two-fold manner. Initially, we queried MusicBrainz<sup>5</sup> for the song's title as a *Release Group*. We only kept songs released as **Singles**; this is evidence that the artists deem it good enough to publicize. Then, we removed singles released after TikTok's birth. This left us with 85 songs. Given that MusicBrainz results may return false positives, using a Fuzzy String Matching approach<sup>6</sup>, we matched the MusicBrainz title and artist to the one title shown TokBoard (that combines both strings in one, e.g. "PYRO by Chester Young & Castion"). If **both** matches were above 50% (in a fuzzy match, 50% of the smaller string is on the largest), we kept the song. In the end, this left us with 38 songs. We manually added songs to this list subsequently, as we now detail.

To increase the above list, we manually searched for information on each song to extract evidence that TikTok was a significant influencer on that song's popularity. To do so, we inspected results like *TikTok's this year on music* as well as news outlets<sup>7</sup> after adding songs from this manual procedure, reaching 51 songs. Finally, we examined only songs with at least 20 data points for both time series, leaving us with 30 songs. This was done to have a minimum number of data points to execute statistical tests.

## 4. UNDERSTANDING THE IMPACT OF TIKTOK

In this section, we showcase our results. We begin by unveiling causality using the Granger Causal test. Next, we look into whether the viral dynamics of TikTok transfer to Web searchers via a Bass model. The Granger model allows us to identify temporal correlations, while the Bass model provides insight into the diffusion and adoption patterns of songs within the digital market.

### 4.1 Granger Causality

To assess causality between time series, we employ the Granger Causality Test [2]. To understand Granger causality, let  $b_s(t)$  be the popularity of a song on TokBoard. Moreover, let  $g_s(t)$  be the popularity of the music on Google Trends. Now, let us create two models for the popularity of Google Trends. The first will be based on data from Google Trends only, whereas the second will explore both websites. These are as follows:

$$g_s(t) = \sum_{l=1}^L \gamma_{s,l} \cdot g_s(t-l) + \epsilon \quad (1)$$

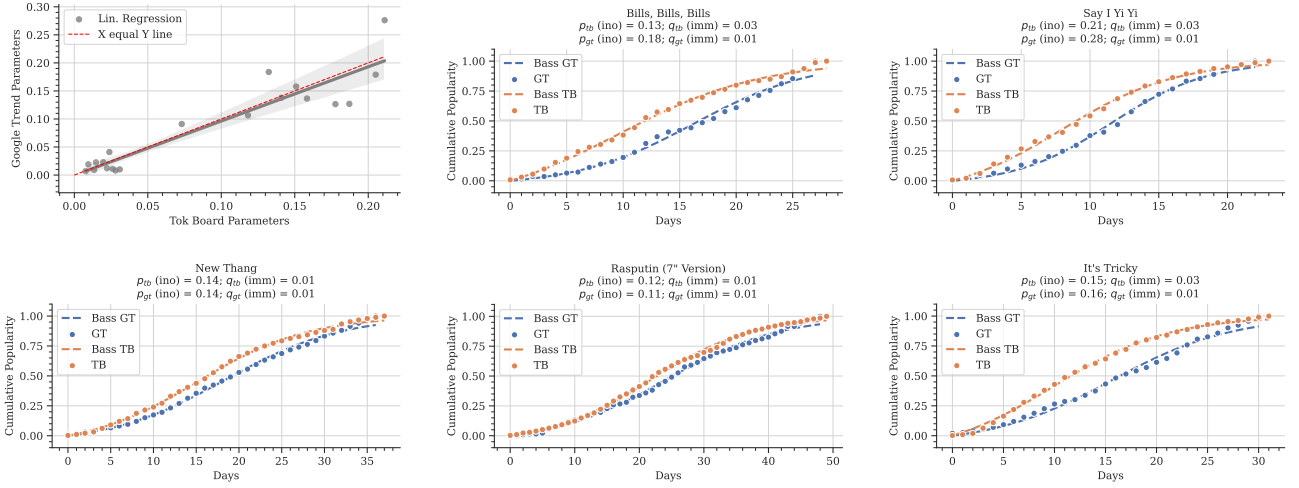
$$g_s(t) = \sum_{l=1}^L \gamma_{s,l} \cdot g_s(t-l) + \sum_{l=1}^L \beta_{s,l} \cdot b_s(t-l) + \epsilon \quad (2)$$

Here,  $L$  is a fixed parameter capturing how far into the past we look. Our analysis tests several lag values for a single pair of time series. We explore  $L \in [1, 5]$  in one to five days.  $\gamma_{s,l}$  indicates the regression weights for the past

<sup>5</sup> <https://musicbrainz.org/>

<sup>6</sup> <https://github.com/rapidfuzz/RapidFuzz>

<sup>7</sup> <https://tonedeaf.thebrag.com/5-old-songs-that-tiktok-weirdly-made-famous-again/>



**Figure 5:** Scatter Plot of Bass Parameters for both websites and Some Bass Model estimation examples.

values of Google Trends, whereas  $\beta_{s,l}$  are the past values for TokBoard. Finally,  $\epsilon$  is Gaussian noise, indicating that this is an ordinary linear regression model.

If, statistically speaking, model (2) is more accurate than model (1), we have evidence that the *past* of TikTok helps to predict the Web search popularity from Google Trends. This indicates Granger causality. We determine that this occurs based on an F-test of the sum of the squares of residuals in the models. The null hypothesis indicates a lack of evidence for causality. Thus, when rejecting it, we have evidence of causality. Figure 4 displays  $p$ -value distribution for the 30 considered songs. Note that around 33% of these  $p$ -values are less than 0.1. This denotes good statistical relevance and indicates some evidence of causality for exactly ten songs. The figure also shows two examples of such songs. Overall, this result suggests that Granger Causality is less present than expected (i.e., on our dataset, TikTok is able to revitalize 33% of songs).

## 4.2 Bass Model

Next, we employ the Bass Diffusion Model, proposed by Frank Bass [26], to understand whether the viral aspects of songs with Granger Causality transfer across platforms. The Bass Model is defined as follows:

$$\frac{b_s(t)}{1 - B_s(t)} = p_s + q_s \cdot B_s(t) \quad (3)$$

Here,  $b_s(t)$  is the popularity of the song at  $t$ . Whereas,  $B_s(t)$  is the popularity up until, or before, time  $t$ . While we use  $b_s(t)$  and  $B_s(t)$  for TokBoard, a similar equation is defined for Google Trends. Finally,  $p_s$  and  $q_s$  are estimated parameters of the model.

Parameter  $p_s$  measures the innovation. In this case, this is the effect of new adopters. In contrast,  $q_s$  measures how imitators follow the new adopters' behaviors, i.e., the influence caused by the initial adopters. Notice from the equation that  $q_s$  multiplies the previous adopters; this is why it captures imitation.  $p_s$  does not depend on past adopters.

We estimated the Bass Model on the ten songs with Granger Causality evidence. This was done via a Least Squares estimation of the parameters on the cumulative popularity curve. Examples are shown in Figure 5. Note that we fit independent models for TokBoard and Google Trends. Finally, we normalized data by dividing by the sum of popularity. In this case, the cumulative curve captures the fraction of adopters or listeners. This is a common practice to fit Bass Models to data. Given that we only have 20 curve fits (two for each song), we visually inspected them. Overall, all the estimations are similar to the ones shown in Figure 5. This result indicates that the Bass Model is a suitable approach to analyzing the adoption of TikTok songs. Overall, the results indicate that a significant portion of viral trends observed on TikTok caused an increase in web search activity, as demonstrated by the statistical findings from the Granger Causality test. In other words, the evidence suggests that TikTok plays an important role in the re-popularization of songs.

## 5. CONCLUSIONS

In this work, we present a study on TikTok's impact on popularizing songs, namely, its effect on bringing older songs back to the mainstream, employing both Granger Causality and Bass Diffusion models to examine TikTok's impact on music popularity across platforms. Our findings illustrate that TikTok significantly contributes to the re-popularization of songs by amplifying their visibility and engagement, as evidenced by the causality and virality patterns that often translate into increased web search activity. Our results showcase a robust methodology alongside a first-of-its-kind analysis, offering insights into how digital platforms can extend the lifecycle of music beyond traditional release cycles.

For future work, a potential direction is modeling the economic impact of TikTok trends on streaming and sales data. Also, we would like to explore the demographic and geographic variations in TikTok-driven music trends, which may provide deeper insights into user engagement.

## 6. REFERENCES

- [1] J. Salo, M. Lankinen, and M. Mäntymäki, “The use of social media for artist marketing: Music industry perspectives and consumer motivations,” *International Journal on Media Management*, vol. 15, no. 1, pp. 23–41, 2013.
- [2] C. W. J. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969. [Online]. Available: <http://www.jstor.org/stable/1912791>
- [3] F. M. Bass, T. V. Krishnan, and D. C. Jain, “Why the bass model fits without decision variables,” *Marketing science*, vol. 13, no. 3, pp. 203–223, 1994.
- [4] D. M. Greenberg and P. J. Rentfrow, “Music and big data: a new frontier,” *Current opinion in behavioral sciences*, vol. 18, pp. 50–56, 2017.
- [5] E. Zangerle, M. Pichl, W. Gassler, and G. Specht, “#nowplaying music dataset: Extracting listening behavior from twitter,” in *Proceedings of the first international workshop on internet-scale multimedia management*, 2014, pp. 21–26.
- [6] D. Hauger, M. Schedl, A. Košir, and M. Tkalcic, “The million musical tweets dataset: What can we learn from microblogs,” in *Proc. ISMIR*, 2013, pp. 189–194.
- [7] Z. Zhou, K. Xu, and J. Zhao, “Homophily of music listening in online social networks of china,” *Social Networks*, vol. 55, pp. 160–169, 2018.
- [8] T. Y.-C. Yeung, “Did the covid-19 pandemic trigger nostalgia? evidence of music consumption on spotify,” *Evidence of Music Consumption on Spotify (August 21, 2020)*, 2020.
- [9] A. Bae, D. Park, Y.-Y. Ahn, and J. Park, “The multi-scale network landscape of collaboration,” *PloS one*, vol. 11, no. 3, p. e0151784, 2016.
- [10] N. Andrade and F. Figueiredo, “Exploring the latent structure of collaborations in music recordings: A case study in jazz,” in *ISMIR*, 2016.
- [11] P. M. Gleiser and L. Danon, “Community structure in jazz,” *Advances in complex systems*, vol. 6, no. 04, pp. 565–573, 2003.
- [12] B. Uzzi and J. Spiro, “Collaboration and creativity: The small world problem,” *American journal of sociology*, vol. 111, no. 2, pp. 447–504, 2005.
- [13] D. d. L. Silva, M. M. Soares, M. Henriques, M. S. Alves, S. de Aguiar, T. de Carvalho, G. Corso, and L. Lucena, “The complex network of the Brazilian Popular Music,” *Physica A: Statistical Mechanics and its Applications*, vol. 332, 2004.
- [14] C. Gunaratna, E. Stoner, and R. Menezes, “Using network sciences to rank musicians and composers in brazilian popular music,” in *Proc. ISMIR*, 2011.
- [15] F. Falcão, N. Andrade, F. Figueiredo, D. Silva, and F. Morais, “Measuring disruption in song similarity networks,” in *ISMIR*, 2020.
- [16] A. Shutsko, “User-generated short video content in social media. a case study of tiktok,” in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 108–125.
- [17] J. Bandy and N. Diakopoulos, “# tulsaflop: A case study of algorithmically-influenced collective action on tiktok,” *arXiv preprint arXiv:2012.07716*, 2020.
- [18] D. Klug, Y. Qin, M. Evans, and G. Kaufman, “Trick and please. a mixed-method study on user assumptions about the tiktok algorithm,” in *13th ACM Web Science Conference 2021*, 2021, pp. 84–92.
- [19] K. Barta and N. Andalibi, “Constructing authenticity on tiktok: Social norms and social support on the” fun” platform,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–29, 2021.
- [20] D. Le Compte and D. Klug, ““it’s viral!”-a study of the behaviors, practices, and motivations of tiktok users and social activism,” in *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, 2021, pp. 108–111.
- [21] A. Vizcaíno-Verdú, P. De-Casas-Moreno, S. Tirocchi *et al.*, “Online prosumer convergence: Listening, creating and sharing music on youtube and tiktok,” *Communication & Society*, vol. 36, no. 1, pp. 151–166, 2023.
- [22] A. Vizcaíno-Verdú and C. Abidin, “Music challenge memes on tiktok: Understanding in-group storytelling videos,” *International Journal of Communication*, vol. 16, p. 26, 2022.
- [23] A. Vizcaíno-Verdú and I. Aguaded, “# thisismechallenge and music for empowerment of marginalized groups on tiktok,” *Media and Communication*, vol. 10, no. 1, pp. 157–172, 2022.
- [24] F. Figueiredo, J. M. Almeida, Y. Matsubara, B. Ribeiro, and C. Faloutsos, “Revisit behavior in social media: The phoenix-r model and discoveries,” in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*. Springer, 2014, pp. 386–401.
- [25] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec, “Do cascades recur?” in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 671–681.
- [26] F. M. Bass, “A new product growth for model consumer durables,” *Management Science*, vol. 15, no. 5, pp. 215–227, 1969. [Online]. Available: <https://doi.org/10.1287/mnsc.15.5.215>



# ASSESSING THE IMPACT OF SAMPLING, REMIXES, AND COVERS ON ORIGINAL SONG POPULARITY

Guilherme Soares S. dos Santos      Flavio Figueiredo

Universidade Federal de Minas Gerais

{guilhermesoares, flaviovd} @dcc.ufmg.br

## ABSTRACT

Music digitalization has introduced new forms of composition known as “musical borrowings”, where composers use elements of existing songs—such as melodies, lyrics, or beats—to create new songs. Using Who Sampled data and Google Trends, we examine how the popularity of a borrowing song affects the original. Employing Regression Discontinuity Design (RDD) for short-term effects and Granger Causality for long-term impacts, we find evidence of causal popularity boosts in some cases. *Borrowee* songs can revive interest in older tracks, underscoring economic dynamics that may support fairer compensation in the music industry.

## 1. INTRODUCTION

Digitization has drastically transformed the production, distribution, and consumption of music. Advances in data storage, transmission, and processing technologies have brought about significant changes in the music industry, making digital music predominant. This shift has led to new forms of composition, known as “musical borrowings”, where composers and producers draw on existing musical influences to create new works, thereby expanding the boundaries of musical genres.

An example of musical borrowing is Beyoncé’s 2003 hit “Crazy in Love” one of her most iconic songs. This work calls this song a *borrowee*. The *borrowee*’s introduction, marked by triumphant brass, drew significant attention. The brass section was sampled from The Chi-Lites’ 1970s song “Are You My Woman”, here called a *borrowed* song. Created in a pre-digital era, the *borrowed*’s influence on “Crazy in Love” is evident, although its precise contribution to the song’s success remains difficult to quantify.

We explore three primary forms of musical borrowing: sampling, remixes, and covers. Sampling, a technique popularized in the late 1970s, involves using fragments of pre-existing songs. Widely used in hip-hop, electronic, and experimental music, sampling allows for creative expression through the manipulation and combination of different

tracks. On the other hand, remixes involve transforming an existing song by altering its structure, rhythm, or effects, often revitalizing it for new audiences. Finally, covers involve re-recordings that usually offer a fresh interpretation and sometimes even surpass the original in popularity.

Musical borrowings raise significant legal issues, particularly regarding the copyright laws that govern these practices – debates around fair use often center on whether the borrowing negatively impacts the market for the original work. A study titled “Sampling Increases Music Sales: An Empirical Copyright Study” highlights that, in the U.S., the Supreme Court considers the market impact on the original song as a key factor in determining fair use [1]. The research suggests that digital sampling can boost sales of the original songs, indicating that musical borrowings may, in some cases, qualify as fair use. Understanding the consequences of using samples in original songs can contribute to fairer legal discussions and more informed decisions by musicians and managers.

Our research aligns with the above discussion, but it also presents counter-arguments, as we now detail. Specifically, we focus on how the popularity of a *borrowee* song (cover, sample, or remix) influences the popularity of the *borrowed* song. Data from WhoSampled<sup>1 2</sup>, a crowd-sourced website cataloging samples, remixes, and covers, and Google Trends<sup>3</sup>, which measures relative search interest, provide the basis for this analysis.

To effectively measure an effect, we employ two techniques. First, we explore Regression Discontinuity Design [2] (RDD) to measure the immediate causal impact of the *borrowee*’s release. Next, we make use of Granger Causality [3] as a means to understand the lasting impact of a *borrower* after release. As an example of a *borrowee* that impacts a *borrowed* song, we show in Figure 1 the song “Somebody” by Natalie La Rose (released in late 2014) that sampled “Shots” from LMFAO. Here, we can see that “Somebody” renews interest in the LMFAO song.

In general, it is important to note that establishing causality is challenging. Among the approximately 884 instances of musical borrowings analyzed, we identified 182 cases (20%) showing any causal evidence. From these, our results show that only a fraction of the *borrowees* (45% or 82 out of 182) have an immediate causal impact via our RDD study. Out of these, 64% (or 117 out of 182) of these

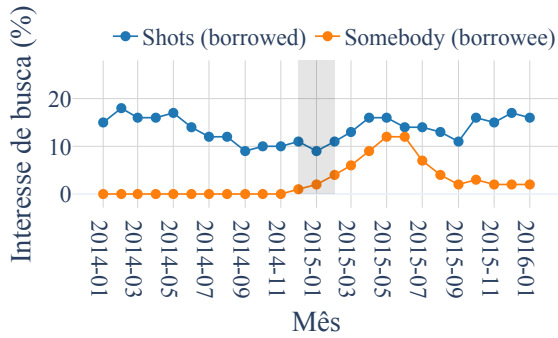


© G. dos Santos and F. Figueiredo. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** G. dos Santos and F. Figueiredo, “Assessing the Impact of Sampling, Remixes, and Covers on Original Song Popularity”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.

<sup>1</sup> <https://www.whosampled.com>.

<sup>2</sup> Our dataset is available at: [github.com/uai-ufmg/whosampleddata](https://github.com/uai-ufmg/whosampleddata)

<sup>3</sup> <https://trends.google.com>



**Figure 1:** Search Interest Over Time for “Shots” (LMFAO) and “Somebody” (Natalie La Rose). The release month is the shaded region.

cases show a lasting effect via Granger Causality. It should be emphasized that this analysis focuses on web search interest, rather than sales metrics. Nonetheless, even with a relatively limited dataset, we observed some influence of *borrowees* on the originals. Thus, we hope that our research sheds light on the actual implications of sampling.

In the following sections, we begin with a review of related work (Section 2), followed by a description of our datasets and methodology (Section 3) and our results (Section 4). Finally, we conclude with a discussion of our findings and their broader implications (Section 5).

## 2. RELATED WORK

The term “musical borrowings” defined by Burkholder [4], refers to the techniques composers use to create music based on pre-existing melodies or structures. This practice has been studied not only in music but also in legal contexts, especially regarding copyright. Scholars have highlighted discrepancies between copyright law and the realities of musical composition. Arewa [5] critiques copyright law’s focus on “romanticized authorship”. This assumes originality and autonomy, disregarding musical borrowing, particularly in genres like hip-hop, where sampling has raised legal concerns.

Musical borrowing also intersects with social inequalities. Hesmondhalgh [6] explores how white musicians have historically appropriated works by Black artists, reinforcing cultural disparities. This issue extends to non-hegemonic markets, such as Latin America and Africa, where musical borrowing can have broader implications.

Analyzing musical influence presents challenges due to its subjective nature. Shalit et al. [7] tackled this by applying Dynamic Topic Models to a dataset of over 24,000 songs from 1922–2010, revealing that musical influence and innovation are not monotonically correlated. The study identified critical periods of high innovation in the early 1970s and mid-1990s.

Turning our attention to authors exploring a dataset similar to ours, Bryan et al. [8] also explore WhoSampled data to analyze to characterize networks. The study examined centrality in a network of 42,447 samples, applying complex network techniques and power-law distributions

to measure the influence of songs and artists. Katz Centrality was introduced as a new method to quantify musical influence, though the author’s study did not directly address how sampling affects popularity.

Ortega [9] analyzed cover versions using a network of over 106,000 artists and 855,000 covers from Second-Hand Songs, focusing on the impact of covers on popular music history. Key findings showed genre and language as primary factors in collaboration and influence.

## 3. METHODOLOGY AND DATA COLLECTION

Our work focuses on analyzing the popularity of songs involved in musical borrowings over time to measure the impact of the release of a new version on the original version. To do so, we used a dataset obtained from Who Sampled containing about 700,000 songs, including samples, covers, and remixes. This dataset was crawled in 2019, and we asked WhoSampled for permission to crawl those data. The website catalogs musical borrowings for more than 975,000 songs from 300,000 artists, contributed by a community of about 28,000 members and verified by moderators or staff<sup>4</sup>. In this sense, our original dataset comprises most of these songs.

The obtained data is structured as a directed graph, where nodes represent songs and edges represent musical borrowings, each labeled to indicate the type of musical borrowing. Since Who Sampled does not provide popularity data or external identifiers, we supplemented the data with Wikidata<sup>5</sup> and Google Trends information.

Initially, we resorted to Wikidata to gather unique identifiers for each song. The song “Clube da Esquina vol. 2” by Milton Nascimento has a Wikidata identifier of Q20053386. More importantly, on Wikidata, this song also has its Freebase<sup>6</sup> identifier listed: /m/0zjw3z... This Freebase identifier was later used for Google Trends queries. Before describing how we collected time series, we discuss how we gathered the Freebase id of our songs.

To find song identifiers, we used Wikidata with three search queries: one with the song title, one with the artist name, and one combining both. For each, we retrieved the top ten entities, removed duplicates, and processed entities in batches of 50 to extract relevant properties.

Several steps were taken to gather the correct IDs, as now detailed. Initially, it is essential to consider that some entities returned may not strictly represent a musical work during the textual search using the song and artist name. To address this issue, we filtered entities using the “**instance of (P31)**” property in Wikidata, selecting those classified as “**musical work (Q2188189)**” or subclasses like “**song (Q7366)**” or “**composition (Q204370)**”. This approach allows us to filter out entities not directly related to musical works.

From the original 699,123 songs, this filtering reduced the list to 111,238 entities, of which 66,373 (59.6%) lacked

<sup>4</sup> <https://www.whosampled.com/about>

<sup>5</sup> [wikidata.org](https://www.wikidata.org)

<sup>6</sup> A discontinued ontology that is still used internally by Google Trends – [https://en.wikipedia.org/wiki/Freebase\\_\(database\)](https://en.wikipedia.org/wiki/Freebase_(database))

knowledge base identifiers. Focusing on entries with Freebase IDs, we finalized a set of 44,857 songs.

To ensure collected entities matched songs in our dataset, we refined them based on the normalized string similarity between the song and artist names, calculated separately with the `diffliib` library<sup>7</sup> in Python. We retained only entries with a string distance above 0.55 for both names, reducing our dataset from 579,322 to 25,830 songs, representing 4,477 musical borrowings.

With the Freebase identifiers, we accessed Google Trends using each identifier as a URL parameter<sup>8</sup>. Our queries were configured to retrieve the global search interest over time for each song, starting from the inception of Google Trends in 2004. We successfully retrieved data for 4,360 songs, each represented as a monthly time series indicating relative popularity, normalized to peak at 100.

Estimating the causal effects of temporal data is challenging [10]. Initially, we can point out that several events may impact popularity (e.g., a song may be sampled twice). Furthermore, other seasonal effects (e.g., a good year for an artist) may be in place. To reduce this impact, all of our time series were filtered to have 24 points only, one per month. Twelve of these occurred before the release and Twelve after. Thus, we shall look at the impact on monthly interest. Moreover, we observed that some time series contained zero values in all time intervals, indicating the absence of search interest data for these musical entities throughout the analyzed period. After this last filter, we were left with 884 borrowings.

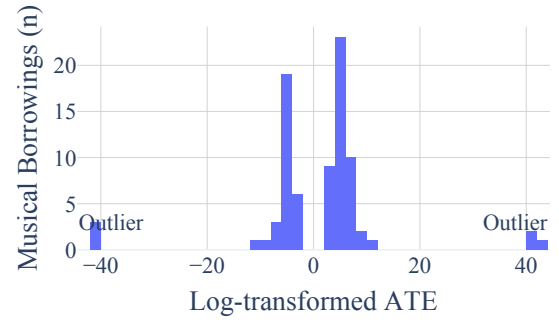
After this pre-processing step, we calculated the impact of the launch on the original in the short and medium term using RDD and Granger Causality.

#### 4. EFFECT OF MUSICAL BORROWINGS ON INTEREST IN A SONG

To assess the impact of samples and re-recordings on audience interest in original music, we explored the hypothesis that these elements introduce discontinuity in nature. Considering that the original music already existed before and so did audience interest, we compared interest in the original music before and after the re-recording. This allows us to estimate the impact of the borrowing on audience interest in the original music.

##### 4.1 Short-term impact

We used the Regression Discontinuity Design (RDD) method to measure the impact of a sample on the original music. RDD is a statistical method that allows us to estimate the causal effect of a treatment on an outcome variable when the probability of treatment changes discontinuously around a cutoff point. In this case, the treatment is the release date of the *borrowee* song. The outcome variable is the search interest in the original song, the *borrowed*. The cutoff point separates the observations into



**Figure 2:** Distribution of the log-transformed average treatment effect (ATE) for the 82 statistically significant musical borrowings.

control and treatment groups. In detail, our RDD was as implemented as follows:

$$googletrends(t) \sim 1 + t + \mathbb{I}_{t>0} + t \cdot \mathbb{I}_{t>0} \quad (1)$$

In this context,  $t$  represents our time variable, shifted so that the release time of the *borrowee* occurs at time zero. This transformation is commonly used in RDD as it simplifies the interpretation of model parameters. The term  $\mathbb{I}_{t>0}$  is an indicator function that equals one when  $t > 0$  (after the release) and zero otherwise. Therefore, at  $t = 0$ , this term’s weight reflects the difference in intercepts, known as the Average Treatment Effect (ATE), which serves as a causal estimate of the release’s impact. We express this measurement in relative terms, where 100% indicates a doubling of the intercept.

Most importantly, we analyzed the ATE estimates only for regressions that achieved statistical significance at a level below 0.05. Our final dataset included 884 musical borrowings, of which 82 (approximately 9.3%) yielded statistically significant  $p$ -values. Figure 2 illustrates the ATE distribution for these 82 significant borrowings, which we will discuss next.

We initially point out borrowings with very high or very low ATEs from the figure. In some cases, this occurred because the original song had no prior search activity. However, after the sample’s release, the original song began to receive increased search interest. Figure 3a shows an example of this phenomenon. The song with the highest ATE in our set is “Something’s Got a Hold on Me” by Etta James, released in 1962, sampled in 2011 by American rapper Flo-Rida in his song “Good Feeling”. We manually inspected these extreme events; all were outliers like this one and we disregarded them.

We thus decided to focus on the non outlier cases, shown in the center of Figure 2. We begin with an interesting case of a negative ATE. Thus, Figure 3b presents the RDD analysis of Lana Del Rey’s “Summertime Sadness,” sampled in “Body Electric,” released five months later on the EP *Paraíso*. The ATE of -152% indicates a negative impact on interest in the original song following the sample release. We believe that the EP’s marketing and the similarity between the two tracks may have influenced the original song’s performance during that period.

<sup>7</sup> <https://docs.python.org/3/library/diffliib.html>

<sup>8</sup> [https://trends.google.com/trends/explore?q=/m/0zjw3z\\_](https://trends.google.com/trends/explore?q=/m/0zjw3z_)

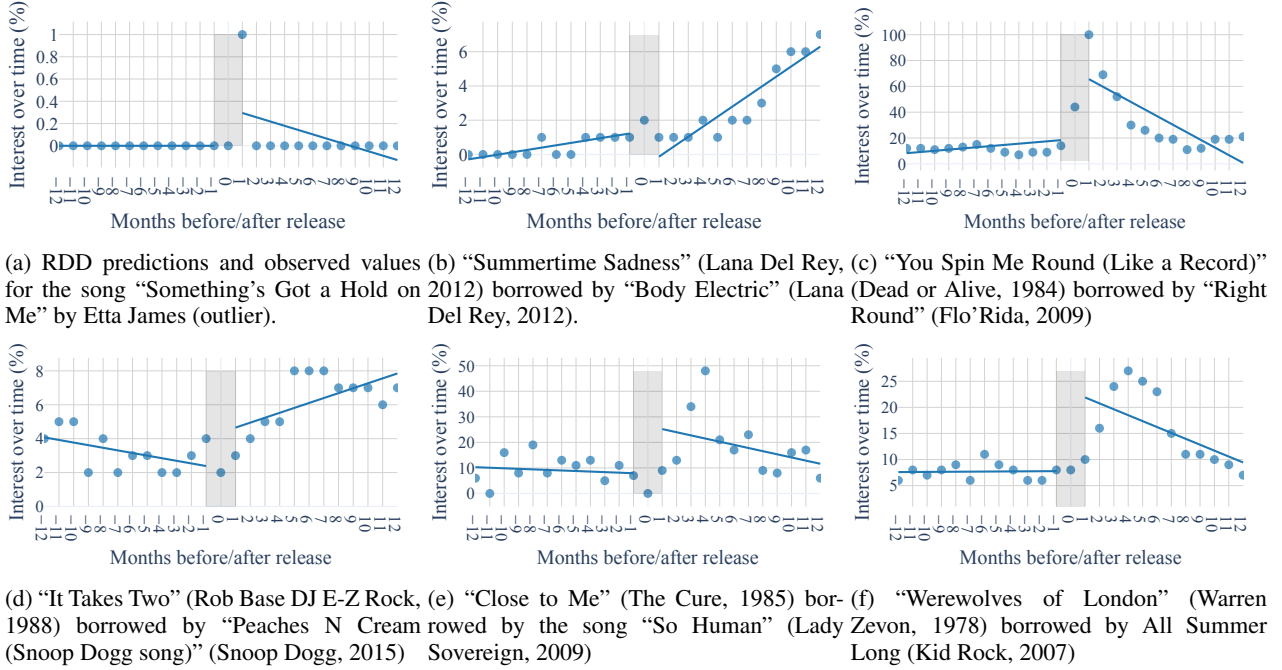


Figure 3: ATE Examples

Figures 3c, 3d, 3e and 3f all show examples of positive ATE. It is quite interesting that several of these examples show samples reviving interest in songs released in the late ’70s to late ’80s. This shows evidence that, in some cases, a *borrowee* may revive interest even in older songs. Nevertheless, we still need to check if this revival is long-lasting. Figures 3c and 3f show a sharp decay a few months after the *borrowee*’s release. This is why we complement this analysis with our Granger Causality analysis next.

## 4.2 Long-term impact

Granger causality is a statistical technique that investigates whether the past of one time series contains valuable information to predict the future of another. Thus, it considers the lags between observations.

We use the *statsmodels* [11] library to perform the Granger causality test, which offers robust statistical tools for time series analysis. The test involves formulating autoregressive (AR) models and comparing them to see if including lagged terms from one time series improves the prediction of another time series.

When applying the Granger test, it is necessary to specify the *max\_lag* parameter, determining the maximum number of lags considered. We set the value of *max\_lag* to 10, assuming that the causal relationship between the original song and the musical borrowing can be captured within 10 months (from one month to almost a year of maximum lag). The test is performed by combining two time series into a complete data frame and running the statistical procedure. The result of the Granger test is a *p*-value associated with the F-test of the sum of squares of the residuals (SSR F-test) for each lag up to the specified *max\_lag*.

A low *p*-value (usually less than 0.05) indicates that the musical borrowing time series contains significant predic-

tive information about the original song time series. When this occurs, we have evidence of a Granger Causal relationship for the time series pair.

After the analysis, we observe that 64% (117) of the musical borrowings have a *p*-value below 0.05, establishing that, in most cases, the time series of the new version (cover, sample, or remix) contains significant predictive information about the time series of the original song. This suggests that *borrowees* generally influence *borrowers* over more extended periods of time.

## 5. CONCLUSIONS

This paper estimated causal evidence that musical borrowings impact the popularity of *borrowed* songs. Our motivation was delving deeper into the argument that borrowing (samples, remixes, and covers) increases the popularity of *borrowed* songs.

Our findings indicate some causal relationships between musical borrowings and the popularity of *borrowed* songs, though effects vary in strength. When present, a *borrowee* song can revitalize interest in older tracks, as shown by RDD analysis. Notably, sustained impacts assessed via Granger Causality suggest borrowings can have a lasting influence on original song popularity.

Our work has performed extensive care in collecting an accurate dataset that links several databases (Freebase, MusicBrainz, and Google Trends). Nevertheless, this does not come with some drawbacks. Out of approximately 700,000 songs analyzed, only 884 instances of musical borrowing were identified. Future work could refine data.

Finally, our results shed light on the complex economics of borrowings, with implications for fairer compensation practices [12].

## 6. REFERENCES

- [1] M. Schuster, D. Mitchell, and K. Brown, "Sampling increases music sales: An empirical copyright study," *American Business Law Journal*, vol. 56, no. 1, pp. 177–229, 2019. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ablj.12137>
- [2] J. Hahn, P. Todd, and W. Van der Klaauw, "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, vol. 69, no. 1, pp. 201–209, 2001.
- [3] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969. [Online]. Available: <http://www.jstor.org/stable/1912791>
- [4] J. P. Burkholder, "The uses of existing music: Musical borrowing as a field," *Notes*, vol. 50, no. 3, pp. 851–870, 1994.
- [5] O. B. Arewa, "From jc bach to hip hop: Musical borrowing, copyright and cultural context," *NCL Rev.*, vol. 84, p. 547, 2005.
- [6] D. Hesmondhalgh, "Digital sampling and cultural inequality," *Social & legal studies*, vol. 15, no. 1, pp. 53–75, 2006.
- [7] U. Shalit, D. Weinshall, and G. Chechik, "Modeling musical influence with topic models," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 2. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 244–252. [Online]. Available: <https://proceedings.mlr.press/v28/shalit13.html>
- [8] N. J. Bryan and G. Wang, "Musical influence network analysis and rank of sample-based music." in *ISMIR*, 2011, pp. 329–334.
- [9] J. L. Ortega, "Cover versions as an impact indicator in popular music: A quantitative network analysis," *Plos one*, vol. 16, no. 4, p. e0250212, 2021.
- [10] C. Hausman and D. S. Rapson, "Regression discontinuity in time: Considerations for empirical applications," *Annual Review of Resource Economics*, vol. 10, no. 1, pp. 533–552, 2018.
- [11] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.
- [12] S. Claflin, "How to get away with copyright infringement: Music sampling as fair use," *BUJ Sci. & Tech. L.*, vol. 26, p. 159, 2020.



# “SHALLOW” NEURAL NETWORK ARCHITECTURES FOR MUSICAL GENRE CLASSIFICATION

Natanael L. de Matos<sup>1</sup>   Hugo T. Carvalho<sup>2</sup>   Carlos T. P. Zanini<sup>2</sup>

<sup>1</sup> Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Brazil

<sup>2</sup> Department of Statistical Methods, Federal University of Rio de Janeiro, Brazil

## ABSTRACT

Deep neural networks are the state-of-the-art in several signal processing tasks, including music genre classification. However, these networks usually require plenty of training data and a lot of computational power to be properly trained. We present four “shallow” neural network architectures, for music genre classification tasks. We also propose a technique of normalization of perceptrons in fully connected layers that, combined with the intentional choice of convolutional kernels tailored to leverage the data structure, appear to facilitate the network extraction of relevant information. Among the results obtained, the accuracy of the normalized models stands out, which in the worst studied cases performed similarly to architectures without normalization. We also demonstrate the promising performance of the proposed architectures, which, although simple, achieved high prediction accuracy metrics.

## 1. INTRODUCTION

Musical classification is a collection of problems within the area of Musical Information Retrieval, which include, among other tasks, estimating the musical genre from observing a musical representation (score, audio, or spectrogram, for example) [1]. Several approaches have been proposed to address the issue of musical genre classification, including Support Vector Machines [2–4], Multi-Layer Perceptron [5, 6], Convolutional Neural Networks (CNN) [1, 7, 8], Recurrent Neural Networks [1, 9, 10] and, more recently, Transformers [11]. In this work, we will demonstrate the feasibility of using neural networks with fewer parameters (compared to state-of-the-art networks [11–14]) in conjunction with data augmentation techniques to solve problems of musical genre classification.

We present four “shallow” neural network architectures based on Convolutional Recurrent Neural Network (CRNN) [15, 16] and Convolutional Neural Network. We

also propose a technique called  $L_2$  normalization of perceptrons in fully connected layers of the network. These, combined with the intentional choice of convolutional layer kernels tailored to leverage data structure, seem to facilitate the network extraction of relevant information concerning the problem of musical genre classification. The audio signals are presented to the neural networks in their raw waveform representation, its mel-spectrogram [17] being computed as a non-trainable layer by properly scaling the magnitudes of the signal’s Short-Time Fourier Transform (STFT) spectra. After this processing, the audio signal can be interpreted as an image with time on the horizontal axis and frequency on the vertical one. In this representation, the use of vertical “one-dimensional” kernels ( $n \times m$  with  $n \gg m$ , where  $n$  is the number of rows and  $m$  the number of columns of the kernel) in the convolutional layers will be explored. The rationale for this choice of kernels is based on the fact that, by analyzing the spectrograms vertically, we jointly observe the frequencies present at each small interval of time. The chosen database to train the proposed architectures was GTZAN [18].

The paper is organized as follows: Section 2 presents the research methodology by describing the proposed architectures and  $L_2$  normalization technique. A discussion about the dataset is also presented, indicating the data augmentation process, and displaying the experimental framework employed. Section 3 discusses the results obtained on the computational experiments. Conclusions are drawn on Section 4, together with indications of future works.

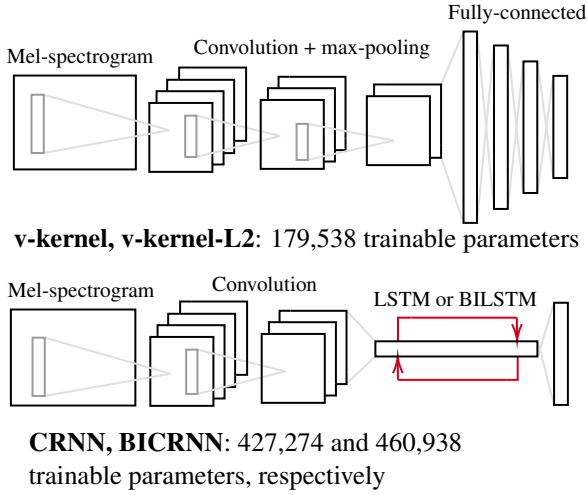
## 2. METHODOLOGY

### 2.1 Architectures

We propose four novel architectures, two CNNs and two CRNNs. The CNN architectures consist on the classic convolutional model with 3 convolutional layers with vertical kernels (each followed by max-pooling and batch normalization), finishing with 3 fully connected layers. These architectures will be referred to as **v-kernel** and **v-kernel-L2**, for the models without and with  $L_2$  normalization, respectively. The CRNN architectures contain 2 convolutional layers with vertical kernels (each followed by max-pooling and batch normalization), finishing with either one recurrent layer (Long Short-Term Memory – LSTM)



© N. Matos, H. Carvalho, and C. Zanini. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).  
**Attribution:** N. Matos, H. Carvalho, and C. Zanini, ““Shallow” Neural Network Architectures for Musical Genre Classification”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.



**Figure 1.** Diagrams of the proposed architectures and number of trainable parameters.

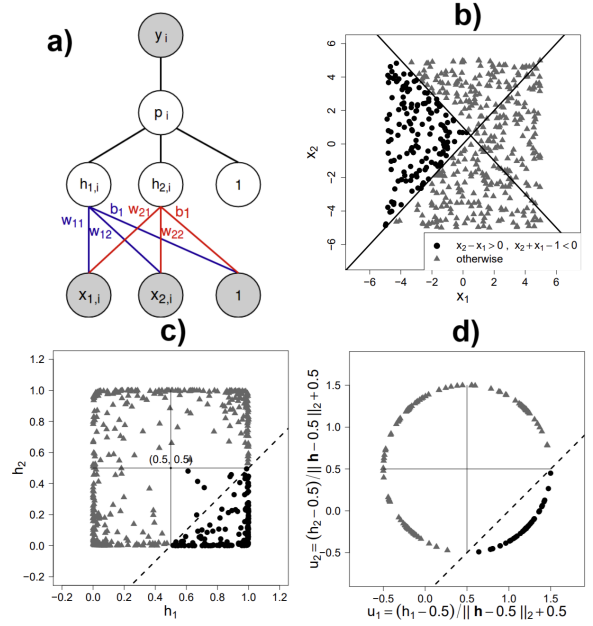
or bidirectional recurrent layer (BiLSTM) [19], followed by one fully connected layer. These architectures will be called **CRNN** and **BICRNN**, respectively (Figure 1). Since there is only one fully connected layer on these architectures, the  $L_2$  normalization is not applied.

## 2.2 $L_2$ Normalization

During the training process of a neural network for classification, each hidden layer is trained to find the best non-linear transformation to a new coordinate system that can approximately separate each class by a hyperplane, meaning that the observations in the training set are learned to be mapped to (approximately) linearly separable regions in the transformed latent space of the last hidden layer with respect to their classes (Figure 2). In this context, the purpose of  $L_2$  normalization of fully connected layers is to assist the neural network in achieving a coordinate change in which the classes, in these coordinates, are easier to separate, as illustrated in Figure 2. The coordinate change performed by  $L_2$  normalization is the projection of the outputs of the fully connected hidden layers onto the surface of a hypersphere with radius 1. Thus,  $L_2$  normalization maps the neurons far away from the center to a region that is easier to be linearly separable at the last network layer which may help to ensure that some data points are not misclassified (Figure 2 panel (d)).

## 2.3 Dataset

The dataset employed to train and validate the proposed architectures was the GTZAN [18], consisting of 1,000 audio signals, each 30 seconds long, mapped into 10 genres: rock, jazz, pop, classical, blues, disco, country, hip hop, metal and reggae. The choice of a small-sized database is strategic: it justifies the adoption of data augmentation techniques and also opens the possibility of extrapolating the strategies employed in this study to other contexts where only limited databases are available (for example, regarding under-represented musical genres), or even



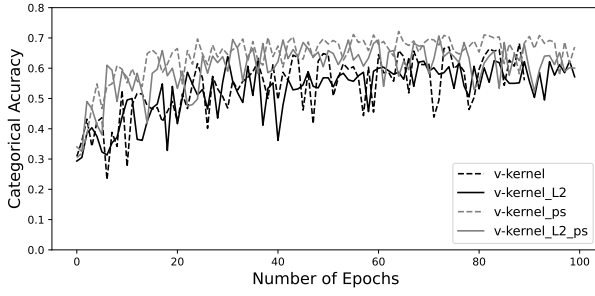
**Figure 2.** Toy example illustrating the coordinate change performed by  $L_2$  normalization. a) Neural network with the inputs ( $x$ 's) and trainable parameters ( $w$ 's and  $b$ 's). b) Training set and its two classes (black and gray points). c) Training set after passing through the hidden layer of the neural network with sigmoid activation function. d) Training set after passing through the hidden layer of the neural network with sigmoid activation function and  $L_2$  normalization, translating and projecting the data onto the unit circumference centered at  $(0.5, 0.5)$ .

to larger databases to introduce robustness into the model predictions, a point to be further investigated.

However, [20] shows that GTZAN has some problems, such as duplicate tracks and some labeling errors. Such problems are very common in classification datasets, especially in more widely explored contexts such as image classification [21]. As the dataset is relatively small, these inconsistencies can negatively affect the performance of models trained on it. Despite this, the proposed architectures achieved satisfactory results in the process of musical genre classification, as reported in Section 3.

## 2.4 Data Augmentation

Since the audio tracks in the dataset are 30 seconds long, we partition the audio signal into three contiguous 10-second signals, assuming that this smaller duration is sufficient to identify the genre of a song. This set will be referred to as  $C$ . Subsequently, we apply a pitch shift to  $C$ , increasing or decreasing the pitch of each segment by a semitone, generating the set  $C_{ps}$ , such that  $C \subset C_{ps}$ . Thus, we have two datasets in total:  $C$  with 3,000 samples and  $C_{ps}$  with 9,000 samples. To ensure a fair comparison between models, the validation set is only taken from  $C$ .



**Figure 3.** Evolution of categorical accuracy (computed on the validation set) during the training of the v-kernel and v-kernel-L2 networks on the  $C$  and  $C_{ps}$  datasets.

## 2.5 Experiments

Each architecture was trained on a subset of each of the datasets  $C$  and  $C_{ps}$ , and a suffix **ps** was added to the names of the architectures trained on the latter. We implemented the proposed architectures on Keras with Tensorflow as backend, and all architectures were trained in the virtual environment of Google<sup>TM</sup> Colab Pro with 28GB of RAM and an NVIDIA<sup>TM</sup> V100 GPU. The training time ranged from 1.5 hours to 2 hours depending on the architecture with inference times of less than 2 seconds for signals up to 4 minutes in the free environment of Google<sup>TM</sup> Colab.

During training, care was taken to ensure that the test set for the networks trained in both  $C$  and  $C_{ps}$  was the same and that there were no excerpts from the same song in the training and test sets to avoid adding correlation between the datasets. Mel-spectrogram hyperparameters and models configurations can be seen on a GitHub repository.<sup>1</sup>

## 3. RESULTS

In this section we analyze performance metrics for the proposed architectures, discuss the effects of  $L_2$  normalization on the fully connected layers of the CNN architectures, and compare the performance of the models when trained on datasets  $C$  and  $C_{ps}$ .

### 3.1 Effects of $L_2$ Normalization

Figure 3 compares the evolution of categorical validation accuracy obtained during training of the CNN models with and without the proposed normalization. Figure 3 shows that the  $L_2$  normalization does not negatively impact the accuracy of the models, since the solid and dashed lines show similar behavior. This result indicates that the  $L_2$  normalization can be, at most, “harmless”, at least in the present scenario. More experiments involving  $L_2$  normalization need to be conducted to better understand its general effects on classification models, especially involving audio data. A finer comparison between the performance of the architectures is drawn in Section 3.2.

<sup>1</sup> <https://github.com/Natanael-Luciano/LAMIR-2024>.

### 3.2 Comparing the architectures

A comparison of the maximum accuracy achieved by each model can be seen in Table 1. By comparing the performance of the architectures trained on the dataset with and without pitch shift, an increase in accuracy can be observed on the former case.

Architecture	Accuracy	Architecture	Accuracy
BICRNN	0.67	v-kernel	0.68
<b>BICRNN-ps</b>	<b>0.77</b>	<b>v-kernel-ps</b>	<b>0.72</b>
CRNN	0.63	v-kernel-l2	0.63
CRNN-ps	0.76	v-kernel-l2-ps	0.70

**Table 1.** Comparison of the maximum accuracy achieved by each proposed architecture. The highest values on each case are marked in bold.

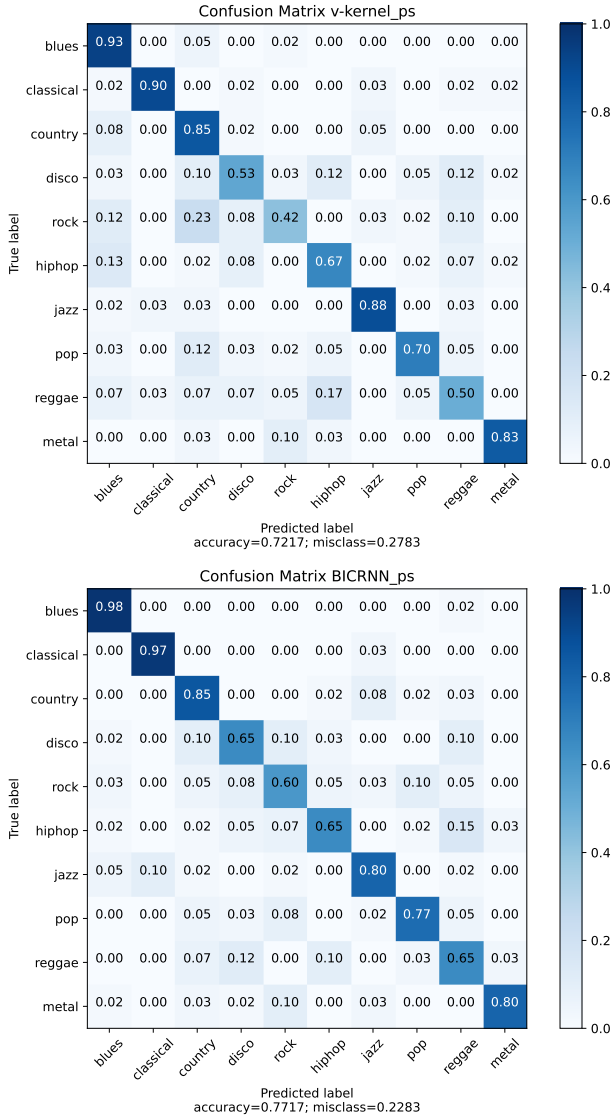
From Table 1 we conclude that the best proposed CNN and CRNN are **v-kernel-ps** and **BICRNN-ps**, respectively. The confusion matrices of these networks are illustrated in Figure 4. The architectures frequently confuse rock with country and blues, and disco with hip hop and reggae. However, the **BICRNN** architecture better classifies these genres, maybe because the **v-kernel** architectures works with local features, while the **BICRNN** takes into account the evolution of features. Therefore, from a network viewpoint, locally rock, blues, and country may seem similar, but globally they are quite distinct.

Inputs of different sizes can be given to the **BICRNN** architecture. To illustrate its behavior on longer (and different) tracks than those present in the dataset, three scenarios were analyzed: the first movement of Beethoven’s *Moonlight Sonata*; first 10 seconds and the whole music *The Only Thing They Fear is You* (soundtrack from the game Doom Eternal). Even with the significantly longer duration of the audio tracks (ranging from five to seven minutes), the network takes on average two seconds to perform inference on the machine available for free on Google<sup>TM</sup> Colab, which allows for real-time classification and also for the evaluation of the “evolution” of its estimated musical genre. Table 2 shows the three most likely estimated genres for each of these tracks. Although the classification for the first movement of Moonlight Sonata is as expected, the classification for Doom Eternal’s soundtrack is not. One possible reason for this is the fact that the guitar sound in this specific track is created via the modulation of a chainsaw sound, together with plenty of electronic samples. Despite recognizing the track as metal in an informal hearing, the intense use of samples is quite common in pop songs, which may explain this misclassification.

### 3.3 Comparison with other studies

When compared to some state-of-the-art models [11, 13, 14], our better-performing architectures have a lower categorical accuracy: [11, 13, 14] reports 0.96, 0.89, and 0.9, respectively, and our better performing architecture is the **BICRNN-ps**, with 0.77 (see Table 1). However, when





**Figure 4.** Comparison of the confusion matrices of the best architectures of each type, CNN and CRNN.

considering the complexity of the models under comparison, our proposal are way simpler, and therefore, faster to train. For example, [11] has about 2,800,000 parameters, and [13, 14] performs respectively 7 and 5 different feature extractions, while our largest architecture, **BICRNN**, has 460,938 trainable parameters, with the advantage of receiving as input a raw audio signal and accepting inputs with any length. Therefore, the importance of using well-tailored architectures becomes evident, enabling simpler architectures to perform well in the musical genre classification problem.

Regarding the metrics used in recent works [22–25] on large databases such as MSD [26] and FMA [27], there is a trend towards using the ROC-AUC metric as a validation measure. The issue with this metric is that it can be high, but the model may perform poorly on some classes, and this information can be lost. This is the case in our work, which achieved an average ROC-AUC (along classes) of 96% with the **BICRNN-ps** architecture, but had classes

Moonlight sonata, 1st. movement		
<b>Classical</b>	Jazz	Country
<b>0.88</b>	0.10	0.006
The Only Thing They Fear is You (first 10 seconds)		
<b>Pop</b>	Metal	Rock
<b>0.82</b>	0.15	0.01
The Only Thing They Fear is You (whole track)		
<b>Pop</b>	Jazz	Blues
<b>0.92</b>	0.04	0.02

**Table 2.** Three genres with the highest probability of allocation for the first movement of *Moonlight Sonata*, first 10 seconds and whole track *The Only Thing They Fear is You*. Highest values and respective classes are marked in bold.

with low accuracy. In comparison, [22–25] report a ROC-AUC between 0.85 and 0.98, which shows that our best-performing model is comparable to theirs in this regard. However, using only the ROC-AUC does not provide information on which classes the model struggles with.

#### 4. CONCLUSION

The problem of musical genre classification is inherently challenging due to the high similarity between some genres, for example, rock and metal, and blues and rock, which can create difficulties even for manual classification. A classic example in the context of Brazilian music is differentiating *pagode* from *samba*.

In this work, we explored the use of “shallow” convolutional neural networks (with and without a proposed technique called  $L_2$  normalization of perceptrons in fully-connected hidden layers), and convolutional recurrent neural networks. Among the results obtained, the observed accuracy of convolutional models when  $L_2$  normalization is applied stands out (which in the worst case performed similarly to architectures without  $L_2$  normalization) and an improvement in the accuracy of the networks when trained on the dataset with pitch shifts, as illustrated in Table 1. However, this improvement was unexpected due to the high similarity of frequencies present in the signals after the pitch shift, and shows that the proposed architectures exhibit a degree of robustness to tonal variations, contrary to the results observed in [28]. Notably, the promising performance of the recurrent architectures stands out, particularly our proposed **BICRNN** architecture, which, although simple, achieved high accuracy in evaluation metrics such as categorical accuracy and was able to better discern similar genres, as demonstrated in Figure 4.

Future works include comparing the proposed models against the state-of-the-art in larger datasets such as MSD and FMA, checking if the use of data augmentation via pitch shift during training have the same effect when working with more genres, and verifying the effects of using our pre-trained model as a starting point when training in larger datasets.

## 5. ACKNOWLEDGMENT

This work was partly funded by the Brazilian funding agency *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES).

## 6. REFERENCES

- [1] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, “Deep Learning for Audio-Based Music Classification and Tagging,” *IEEE Signal Processing Magazine*, pp. 41–51, 2019.
- [2] C. Xu, N. Maddage, X. Shao, F. Cao, and Q. Tian, “Musical genre classification using support vector machines,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 5, 2003, pp. V–429.
- [3] D. R. Ignatius Moses Setiadi, D. Satriya Rahardwika, E. H. Rachmawanto, C. Atika Sari, C. Irawan, D. P. Kusumaningrum, Nuri, and S. L. Trusthi, “Comparison of SVM, KNN, and NB Classifier for Genre Music Classification based on Metadata,” in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2020, pp. 12–16.
- [4] N. Narkhede, S. Mathur, and A. Bhaskar, “Automatic Classification of Music Genre Using SVM,” in *Computer Networks and Inventive Communication Technologies*, S. Smys, R. Bestak, R. Palanisamy, and I. Kotuliak, Eds. Singapore: Springer Nature Singapore, 2022, pp. 439–449.
- [5] J. Ramírez and M. J. Flores, “Machine learning for music genre: multifaceted review and experimentation with audioset,” *Journal of Intelligent Information Systems*, vol. 55, no. 3, pp. 469–499, 2020.
- [6] D. S. Lau and R. Ajoodha, “Music genre classification: A comparative study between deep learning and traditional machine learning approaches,” in *Proceedings of Sixth International Congress on Information and Communication Technology*, X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds. Singapore: Springer Singapore, 2022, pp. 239–247.
- [7] L. Feng, S. Liu, and J. Yao, “Music genre classification with paralleling recurrent convolutional neural network,” 2017. [Online]. Available: <https://arxiv.org/abs/1712.08370>
- [8] Y.-N. Hung, C.-H. H. Yang, P.-Y. Chen, and A. Lerch, “Low-resource music genre classification with cross-modal neural model reprogramming,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [9] J. Zhang, “Music Genre Classification with ResNet and Bi-GRU Using Visual Spectrograms,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.10773>
- [10] W. Wu, F. Han, G. Song, and Z. Wang, “Music genre classification using independent recurrent neural network,” in *2018 Chinese Automation Congress (CAC)*, 2018, pp. 192–195.
- [11] C. Xie, H. Song, H. Zhu, K. Mi, Z. Li, Y. Zhang, J. Cheng, H. Zhou, R. Li, and H. Cai, “Music genre classification based on res-gated CNN and attention mechanism,” *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 13 527–13 542, 2024.
- [12] D. Aditya, S. Murthy, and K. Shashidhar, “Music Genre Classification with Convolutional Neural Networks and Comparison with F, Q, and Mel Spectrogram-Based Images,” in *Advances in Speech and Music Technology*, A. Biswas, E. Wennekes, T.-P. Hong, and A. Wiczorkowska, Eds. Singapore: Springer Singapore, 2021, pp. 235–248.
- [13] P. Fulzele, R. Singh, N. Kaushik, and K. Pandey, “A Hybrid Model for Music Genre Classification Using LSTM and SVM,” in *2018 Eleventh International Conference on Contemporary Computing (IC3)*, 2018, pp. 1–3.
- [14] J. S. Wycliffe, R. M. P. Karthik, K. P. Kanth, and J. Prasanna, “Music Genre Classification Using LSTM and CNN,” in *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, 2023, pp. 205–209.
- [15] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen, “Convolutional Recurrent Neural Networks: Learning Spatial Dependencies for Image Representation,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Singapore: IEEE, 2015, pp. 18–26.
- [16] B. Shi, X. Bai, and C. Yao, “An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition,” Wuhan, China, 2015. [Online]. Available: <https://arxiv.org/abs/1507.05717>
- [17] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Berlin/Heidelberg: Springer, 2015.
- [18] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [19] R. M. Schmidt, “Recurrent Neural Networks (RNNs): A Gentle Introduction and Overview,” Tübingen, Germany, 2019. [Online]. Available: <https://arxiv.org/abs/1912.05911>
- [20] B. L. Sturm, “The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval,” *Journal of New Music Research*, vol. 43, no. 2, pp. 147–172, Apr. 2014. [Online]. Available: <http://dx.doi.org/10.1080/09298215.2014.894533>

- [21] N. M. Muller and K. Markert, “Identifying Mislabeled Instances in Classification Datasets,” in *2019 International Joint Conference on Neural Networks (IJCNN)*. Garching, Germany: IEEE, Jul. 2019, pp. 1–8.
- [22] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, “Deep Learning for Audio-Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from Bach,” *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 41–51, 2019.
- [23] Dieleman, Sander and Schrauwen, Benjamin, “End-to-end learning for music audio,” in *International Conference on Acoustics Speech and Signal Processing ICASSP*. IEEE, 2014, pp. 6964–6968. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6854950>
- [24] G. Song, Z. Wang, F. Han, S. Ding, and M. A. Iqbal, “Music auto-tagging using deep recurrent neural networks,” *Neurocomputing*, vol. 292, pp. 104–110, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231218302431>
- [25] T. Kim, J. Lee, and J. Nam, “Sample-level CNN architectures for music auto-tagging using raw waveforms,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 366–370.
- [26] T. Bertin-Mahieux, D. P. W. Ellis, B. Whiteman, and P. Lamere, “Million Song Dataset,” New York, NY, USA, 2011. [Online]. Available: <https://academiccommons.columbia.edu/doi/10.7916/D8NZ8J07>
- [27] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A Dataset For Music Analysis,” Lausanne, Switzerland; Singapore, 2017. [Online]. Available: <https://academiccommons.columbia.edu/doi/10.7916/1612.01840>
- [28] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, “Evaluation of CNN-based Automatic Music Tagging Models,” Barcelona, Spain, 2020. [Online]. Available: <https://arxiv.org/abs/2006.00751>

# AEROMAMBA: AN EFFICIENT ARCHITECTURE FOR AUDIO SUPER-RESOLUTION USING GENERATIVE ADVERSARIAL NETWORKS AND STATE SPACE MODELS

Wallace Abreu, Luiz W. P. Biscainho

Federal University of Rio de Janeiro

{wallace.abreu, wagner}@smt.ufrj.br

## ABSTRACT

Audio super-resolution aims to enhance low-resolution signals by creating high-frequency content. In this work, we modify the architecture of AERO (a state-of-the-art system for this task) for music super-resolution. Specifically, we replace its original Attention and LSTM layers with Mamba, a State Space Model (SSM), across all network layers. Mamba is capable of effectively substituting the mentioned modules, as it offers a mechanism similar to that of Attention while also functioning as a recurrent network. With the proposed AEROMamba, training requires 2-4x less GPU memory, since Mamba exploits the convolutional formulation and leverages GPU memory hierarchy. Additionally, during inference, Mamba operates in constant memory due to recurrence, avoiding memory growth associated with Attention. This results in a 14x speed improvement using 5x less GPU. Subjective listening tests (0 to 100 scale) show that the proposed model surpasses the AERO model. In the MUSDB dataset, degraded signals scored 38.22, while AERO and AERO-Mamba scored 60.03 and 66.74, respectively. For the PianoEval dataset, scores were 72.92 for degraded signals, 76.89 for AERO, and 84.41 for AEROMamba.

## 1. INTRODUCTION

Audio super-resolution is a technique analogous to what is known in signal processing literature as bandwidth extension [1], whose goal is to reconstruct the upper spectral content of a low-resolution signal. Since a bandlimited audio signal usually sounds muffled, higher-resolution audio yields a better listening experience, in general [1].

Since the 19th-century invention of sound recording devices [2], audio signals have been widely used in communications and entertainment. Technology has evolved to meet specific application requirements, with telephony prioritizing intelligibility and general audio devices focusing on fidelity [3]. High-fidelity systems must cover at least

the human auditory range of 20 Hz to 20 kHz for tones [4], though analog audio may face limitations and media degradation affecting this content [5]. In digital audio, compression lowers transmission and storage costs. Decimation, which discards samples, needs low-pass filtering to prevent aliasing and reduce the signal’s maximum frequency. Lossy audio coding, such as MP3 [6], modifies frequency content based on a psychoacoustic model, mostly affecting high frequencies and sometimes reducing bandwidth. Audio super-resolution is useful in scenarios requiring mitigation of these issues.

Signal processing-based bandwidth extension methods included techniques such as nonlinear devices with linear filtering [1], source-filter modeling [7], codebook mapping [8], and spectral-band replication [9]. In the past few years, solutions based on deep neural networks (DNNs) became the state of the art in audio super-resolution, ranging from pure feedforward networks that operate on raw waveforms [10] or in the spectral domain [11] to generative solutions using Generative Adversarial Networks (GANs) [12, 13] and, more recently, Diffusion Models (DMs) [14–17].

The choice of using DMs instead of GANs is generally justified by training instabilities, suboptimal mode coverage, and lack of explainability of GANs, while DMs are modeled by statistical physics [18]. Even though efficient DMs [19] constitute an active area of research, their Markov Chain-based sampling requires sequential inference, which makes parallelization difficult and sample generation slower compared to GANs. Additionally, narrow mode coverage is only problematic when diverse data generation is needed, which may not necessarily be the case for audio enhancement.

In this context, this work proposes improvements to the state-of-the-art AERO [13] architecture for super-resolution by replacing its Attention and LSTM layers with Mamba [20], a State Space Model (SSM) created for efficient sequence modeling which has shown promising results when used in speech enhancement [21].

The advantages of this approach are significant: training requires 2x to 4x less GPU memory, and inference runs with a 14x speed gain using 5x less GPU, all with an increase in audio quality, as demonstrated by the listening tests performed.

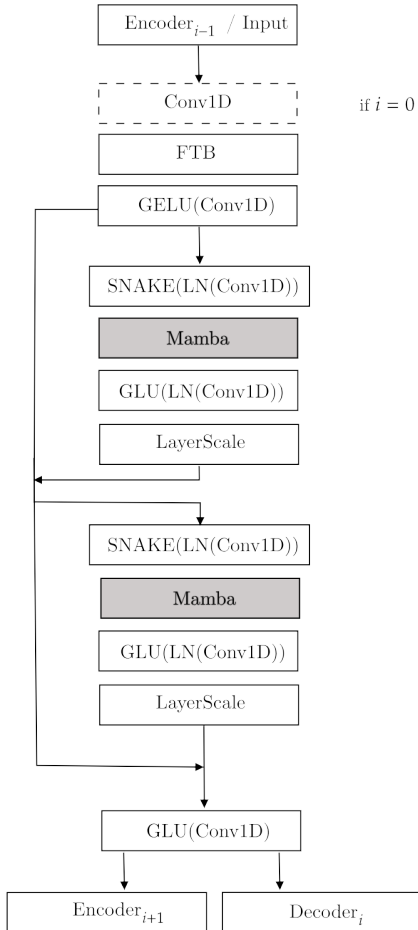


© W. Abreu and L. W. P. Biscainho. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** W. Abreu and L. W. P. Biscainho, “AEROMamba: An efficient architecture for audio super-resolution using generative adversarial networks and state space models”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.

## 2. METHOD

AERO is an audio super-resolution GAN that processes music and speech signals, inspired by Demucs [22]. Its architecture is composed of a generator based on the U-Net, with Attention and BiLSTMs on its last two encoder blocks, and a MelGAN [23] multi-scale discriminator.

In this work, we propose **AEROMamba**, in which we replace all the Local Attention and BiLSTM layers in AERO by Mamba layers, including them in all encoder blocks, not only at specific depths. This modification is motivated by our hypothesis that Mamba can handle both tasks effectively, since it offers a selection mechanism similar to Attention while being also a recurrent network, as a generalized SSM. Specifically, this selection mechanism works through a parameterization of the SSM matrices in relation to the input, different from the linear time-invariant formulation. Additionally, since Mamba exploits the GPU memory hierarchy to perform computations efficiently, we aim to yield high-quality samples while also reducing the use of computational resources in training and inference. The resulting architecture is shown in Figure 1.



**Figure 1.** Detail of the AEROMamba architecture: AERO with BiLSTMs and Local Attentions replaced by Mamba layers in all Encoder blocks.

Consistently with the original model, we optimize the

generator loss

$$L_G = L_{adv} + L_{rec} + \lambda L_{fmap}, \quad (1)$$

composed of the adversarial loss  $L_{adv}$ , the reconstruction loss  $L_{rec}$ , and the features loss  $L_{fmap}$  with  $\lambda = 100$ . In addition, the training objective  $L_D$  for the discriminator is the MelGAN hinge loss [23].

## 3. EXPERIMENTS

### 3.1 Datasets

The PianoEval data set<sup>1</sup> was compiled in two segments: the first segment (training and validation sets) comprises 45 recordings of the Chopin’s 24 Preludes, Op. 28, played by 33 different pianists (totaling 22 hours), while the second segment (test set) incorporates excerpts taken from piano pieces by Ligeti, Schumann and Barber, performed respectively by 3 different performers (amounting to 3.5 hours). We also tested our model on MUSDB18 [24], which contains 150 songs (10 hours) of musical mixtures with their isolated stems, using only the mixture tracks. All files are in WAV format, stereo, sampled at 44.1 kHz.

### 3.2 Training procedure

Originally, AERO was trained in the upsampling configuration from 11.025 kHz to 44.1 kHz using a specific window size  $W = 512$  along with three distinct hop lengths:  $H = 64$ ,  $H = 128$ , and  $H = 256$ . Our model also works in the 11.025-44.1 kHz setting but, since  $H = 64$  or  $H = 128$  settings took impractical training time, our development concentrated on the  $H = 256$  version.

For training on MUSDB18, we started with a pre-trained AERO model (already trained for 696 epochs on the same dataset) and extended its training for an additional 100 epochs, selecting the optimal model based on the Perceptual Audio Quality Measure (PAQM) [25] score. To ensure consistency, AEROMamba was trained for an equivalent number of epochs. The original train/test partition of MUSDB18 was retained.

In the case of PianoEval training set, we created two groups: PianoEval-GQ and PianoEval-HQ (with a difference of 1.5 hours between them) — the first containing all recordings, and the second contained only tracks recorded after the 1960s. The motivation for this was to evaluate whether the greater presence of noise, intrinsic to older recordings, would be beneficial, adding robustness to the models, or would harm the quality of the predicted results. Both AERO and AEROMamba were trained for approximately 800 epochs, also saving the best model.

All other hyperparameters were configured according to the default settings of AERO, such as the learning rate of  $3e-4$ . Both models underwent training on an NVIDIA GeForce RTX 3090 GPU with the maximum batch size allowed for each model.

<sup>1</sup>The metadata of each recording and code repository are available in the accompanying website <https://aeromamba-super-resolution.github.io/>.

### 3.3 Evaluation

For the objective evaluation, we employed the Log-Spectral Distance (LSD) [13] and two distinct perceptual quality metrics: PAQM and the Virtual Speech Quality Objective Listener (ViSQOL) [26]. Furthermore, we conducted subjective listening tests to provide qualitative comparisons between AERO and AEROMamba on the MUSDB18 and PianoEval datasets.

PAQM was used as a validation metric to select the optimal model during a training run, with fixed seed and hyperparameters. The choice of PAQM as a validation metric was motivated by the availability of a vectorized implementation, which is much faster than ViSQOL, and also to its perceptual motivation, taking into account masking effects and other auditory modeling aspects [25]. In addition, ViSQOL (in audio mode) was employed to assess the quality of the processed signals on a scale from 1 to 5, in relation to the test sets.

For the listening tests, the opinions of 20 subjects were compiled regarding the overall similarity of three test signals, one corresponding to each model and one anchor, in relation to a known reference. The score was given through a sliding bar without explicit numerical values, with the words ‘Exactly the same’ on the right end and increasing difference indicated by a left arrow. Although the numerical value was not explicit to the subjects, the scale was defined from 0 to 100, with a 1-point resolution. A total of 12 tracks from the PianoEval dataset and 12 tracks from the MUSDB18-HQ test set were employed, with each track being evaluated by 10 subjects. The order of questions was randomized to mitigate bias.

Test signals were selected to ensure maximum variety in terms of dynamic range, tempo, spectral content, and sound intensity. In the case of PianoEval, this was achieved by selecting four different excerpts from each composer. For MUSDB18HQ, two pieces were selected for each genre, namely rock, pop, electronic, hip hop, world music, and other (an additional pair of songs with no specified genre). The duration of the signals was set between 15 and 20 seconds, according to the standards [27] and providing sufficient listening content for the subjects, with fade-in and fade-out effects to smooth abrupt starts or ends when needed.

## 4. RESULTS

Results and computational performance metrics are summarized in Tables 1, 2, and 3, which make it clear that AEROMamba (with its corresponding HQ version) achieved superior performance to AERO objectively and subjectively, using almost 6x-9x less GPU in inference and running almost 15x as fast (in addition to training at least 2x faster). These performance improvements are due to the efficiency provided by the Mamba layer and also to its ability in sequence modeling tasks. As seen in Table 3, we were able to build a larger architecture, in the sense of parameters, that uses fewer computational resources, thus theoretically being a more powerful model.

Model	MUSDB18		
	ViSQOL $\uparrow$	LSD $\downarrow$	Score $\uparrow$
Low-Resolution	1.82	3.98	38.22
AERO	2.90	1.34	60.03
AEROMamba	<b>2.93</b>	<b>1.23</b>	<b>66.47</b>

**Table 1.** ViSQOL, LSD, and subjective scores for AERO and AEROMamba on the MUSDB18 dataset.

Model	PianoEval		
	ViSQOL $\uparrow$	LSD $\downarrow$	Score $\uparrow$
Low-Resolution	4.36	1.09	72.92
AERO	<b>4.38</b>	<b>0.99</b>	76.89
AERO-HQ	4.34	1.04	-
AEROMamba	4.43	0.98	-
AEROMamba-HQ	<b>4.38</b>	1.00	<b>84.41</b>

**Table 2.** ViSQOL, LSD, and subjective scores for AERO and AEROMamba on the PianoEval dataset. Models labeled with ‘-HQ’ were trained on PianoEval-HQ.

According to Mann-Whitney [28] tests on PianoEval ViSQOL results, all models when paired with Low-Resolution scores distribution achieved statistical significance with  $p$ -value  $< 0.05$ , except for AeroMamba. Therefore, we decided to evaluate AEROMamba-HQ in subjective tests, together with AERO, due to its higher ViSQOL score compared to its ‘-HQ’ version. For subjective scores, all pairs were considered statistically different. Considering that PianoEval-HQ is 1.5 hours shorter in duration than PianoEval, we demonstrate a scenario where AERO-Mamba architecture can attain scores superior to AERO’s even with a reduced amount of data.

Regarding the results of MUSDB18-HQ, the same Mann-Whitney tests reported statistical significance with  $p$ -value  $< 0.05$  between all distribution pairs, both for ViSQOL and subjective listening tests scores.

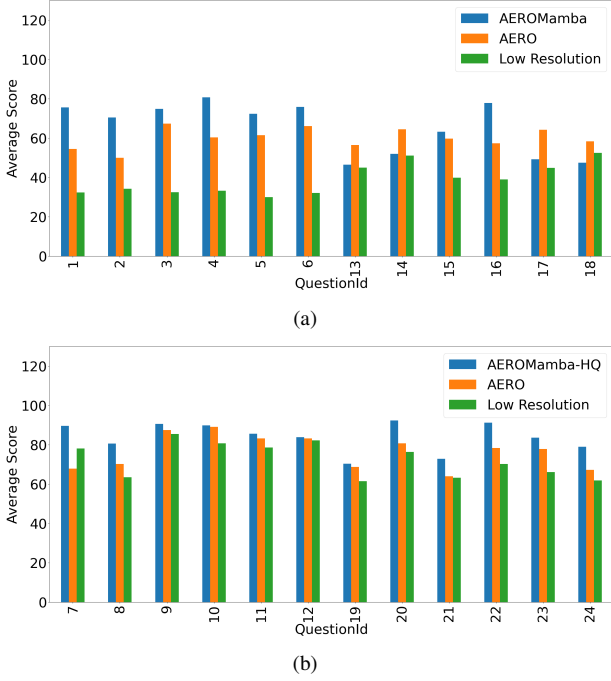
For a detailed visualization, we show in Figure 2 the scores for each track included in the subjective evaluation procedure, identified by their Id number on the testing interface. It is clear that the negative effect of downsampling is more severe on the MUSDB18 tracks, while for PianoEval tracks the improvement of super-resolution is less pronounced. This is explained by the nature of the two datasets: while PianoEval content is limited to a single acoustic instrument played with varied dynamics and agogics, MUSDB18-HQ tracks contain a wide range of electronic sounds and percussion, usually in a heavily saturated mix and high tempo.

As a qualitative illustration, we compare in Figure 3 the frequency spectra of the enhanced signals for AERO and AEROMamba-HQ, using Track 22 of PianoEval as a reference. Evidently, most of the signal’s energy is below 5 kHz, which explains why its low-resolution version scored above 70. However, there is also clear differences between the results of two models. AEROMamba produces a



Method	NVIDIA RTX 3090		NVIDIA RTX 2080 Ti		Parameters
	GPU Usage (MB)	Time (s)	GPU Usage (MB)	Time (s)	
AERO	17091	1.246	16420*	–	19,432,958
AEROMamba	3000	0.087	1914	0.063	20,964,190

**Table 3.** Comparison of GPU usage, inference times for 10 second segments, and number of parameters for AERO and AEROMamba. \*AERO exceeded the GPU memory limit by 5.16 GB.



**Figure 2.** Subjective scores per track id obtained (a) on MUSDB18; (b) on PianoEval.

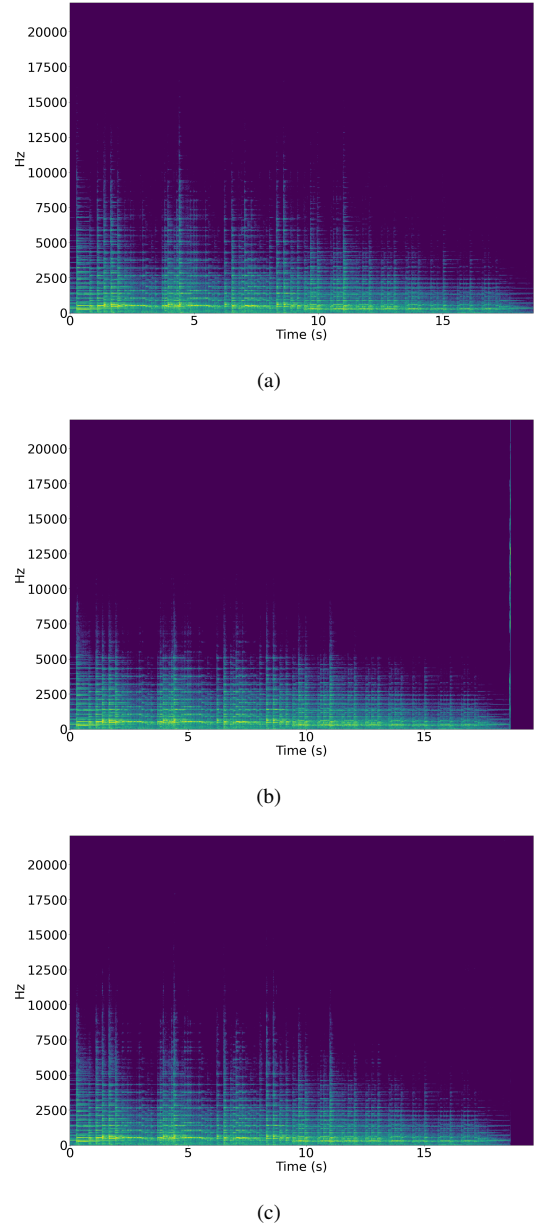
greater amount of high-frequency content than AERO. No visible artifacts are introduced, except for an impulse introduced by AERO at the end of the track due to the abrupt silence induced by zero-padding (easily avoidable).

Finally, in the case of PianoEval, listeners reported that AEROMamba was capable of creating more high-frequency content than AERO, generating a brighter sound that seemed closer to the original, as discussed above. However, this same behaviour resulted in the perception of unexpected artifacts in MUSDB18 evaluation, which led AERO to be indicated as the best model in four tracks.

## 5. CONCLUSION

In this work we proposed an efficient neural network architecture based on a state-of-the-art solution for audio super-resolution. Our method significantly reduces GPU usage and offers faster inference without compromising audio quality. We confirm the superiority of our model to the baseline both through objective metrics and by evaluating the subjective quality of our model via listening tests.

For future works, one of the main modifications that can benefit the usability of the model is to implement a flexible initial sampling frequency, instead of just 11.025 kHz. In



**Figure 3.** Frequency spectra of Track 22 of PianoEval: (a) original; (b) processed by AERO. (c) processed by AEROMamba-HQ.

addition, we would like to evaluate whether the models are invariant to similar instruments, such as the piano and the harpsichord, to the point of achieving good performance on one when trained on the other.

## 6. ACKNOWLEDGMENTS

The authors thank CNPq, FAPERJ, and CAPES for sponsoring this research.

## 7. REFERENCES

- [1] E. Larsen and R. Aarts, *Audio Bandwidth Extension*. Hoboken, USA: John Wiley & Sons, Ltd, 2004.
- [2] D. Schüller and A. Häfner, “Handling and storage of audio and video carriers,” International Association of Sound and Audiovisual Archives, London, UK, Tech. Rep., 2014.
- [3] L. W. P. Biscainho and L. O. Nunes, *Automatic Evaluation of Acoustically Degraded Full-Band Speech*. Boca Raton, USA: CRC Press, Jan. 2017, pp. 181–208.
- [4] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*. Norwell, USA: Kluwer Academic Publishers, 2002.
- [5] P. Copeland, *Manual of Analogue Sound Restoration Techniques*. London, UK: British Library, 2008.
- [6] K. Brandenburg, “Mp3 and aac explained,” in *17th AES International Conference on High Quality Audio Coding*. Florence, Italy: Audio Engineering Society, set. 1999, pp. 99–110.
- [7] V. Berisha and A. Spanias, “Bandwidth extension of audio based on partial loudness criteria,” in *2006 IEEE Workshop on Multimedia Signal Processing*, Victoria, Canada, Oct. 2006, p. 146–149.
- [8] B. Iser and G. Schmidt, *Bandwidth Extension of Telephony Speech*. Berlin, Germany: Springer, 2008, p. 135–184.
- [9] P. Ekstrand, “Bandwidth extension of audio signals by spectral band replication,” in *1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*. Leuven, Belgium: IEEE, nov. 2002, pp. 73–79.
- [10] V. Kuleshov, S. Z. Enam, and S. Ermon, “Audio super resolution using neural networks,” Available in <http://arxiv.org/abs/1708.00853> (09/03/2023), 2017.
- [11] M. Lagrange and F. Gontier, “Bandwidth extension of musical audio signals with no side information using dilated convolutional neural networks,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, mai. 2020, pp. 801–805.
- [12] E. Moliner and V. Välimäki, “Behm-gan: Bandwidth extension of historical music using generative adversarial networks,” Available in <https://arxiv.org/abs/2204.06478> (09/03/2023), 2022.
- [13] M. Mandel, O. Tal, and Y. Adi, “Aero: Audio super resolution in the spectral domain,” Available in <https://arxiv.org/abs/2211.12232> (09/03/2023), 2022.
- [14] E. Moliner, J. Lehtinen, and V. Välimäki, “Solving audio inverse problems with a diffusion model,” Available in <https://arxiv.org/abs/2210.15228> (09/03/2023), 2022.
- [15] H. Liu, K. Chen, Q. Tian, W. Wang, and M. D. Plumbley, “Audiosr: Versatile audio super-resolution at scale,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, p. 1076–1080.
- [16] E. Moliner, F. Elvander, and V. Välimäki, “Blind audio bandwidth extension: A diffusion-based zero-shot approach,” Available in <http://arxiv.org/abs/2306.01433> (14/09/2024), Jan. 2024.
- [17] E. Moliner, M. Turunen, F. Elvander, and V. Välimäki, “A diffusion-based generative equalizer for music restoration,” Available in <http://arxiv.org/abs/2403.18636> (14/09/2024), Mar. 2024.
- [18] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Virtual Event, 12 2020.
- [19] A. Ulhaq and N. Akhtar, “Efficient diffusion models for vision: A survey,” Available in <http://arxiv.org/abs/2210.09292> (14/09/2024), Mar. 2024.
- [20] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” Available in <https://arxiv.org/abs/2312.00752> (12/09/2024), 2024.
- [21] R. Chao, W.-H. Cheng, M. L. Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, “An investigation of incorporating mamba for speech enhancement,” Available in <https://arxiv.org/abs/2405.06573> (15/09/2024), 2024.
- [22] A. Défossez, “Hybrid spectrogram and waveform source separation,” in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Virtual Event, 2021.
- [23] K. Kumar, R. Kumar, T. de Boissiere, L. Geste, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” Available in <https://arxiv.org/abs/1910.06711> (09/03/2023), 2019.
- [24] Z. Rafi, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “Musdb18-hq - an uncompressed version of musdb18,” Available in <https://doi.org/10.5281/zenodo.3338373> (12/09/2024), Aug. 2019.



- [25] J. G. Beerends and J. A. Stemerdink, “A perceptual audio quality measure based on a psychoacoustic sound representation,” *Journal of the Audio Engineering Society*, vol. 40, pp. 963–978, december 1992.
- [26] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, “Visqol: an objective speech quality model,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015 (13), pp. 1–18, 2015.
- [27] International Telecommunications Union (ITU-R), “Methods for the subjective assessment of small impairments in audio systems,” Geneva, Switzerland, Tech. Rep., 2015.
- [28] J. Gibbons, *Nonparametric Methods for Quantitative Analysis*, ser. American series in mathematical and management sciences. Syracuse, NY: American Sciences Press, 1985.

# LONG-FORM TEXT-TO-MUSIC GENERATION WITH ADAPTIVE PROMPTS: A CASE OF STUDY IN TABLETOP ROLE-PLAYING GAMES SOUNDTRACKS

Felipe Marra

Universidade Federal de Viçosa  
Departamento de Informática  
Viçosa, Minas Gerais, Brazil

Lucas N. Ferreira

Universidade Federal de Viçosa  
Departamento de Informática  
Viçosa, Minas Gerais, Brazil

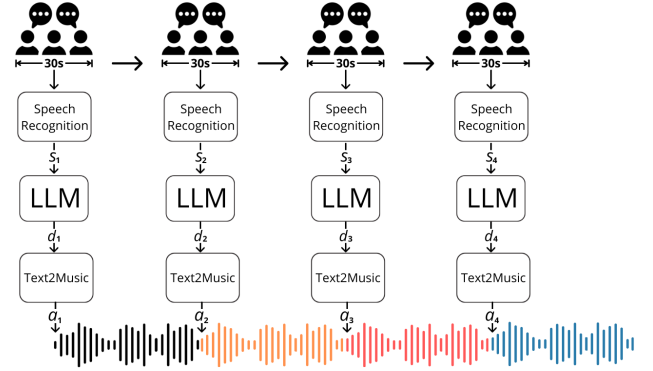
## ABSTRACT

This paper investigates the capabilities of text-to-audio music generation models in producing long-form music with prompts that change over time, focusing on soundtrack generation for Tabletop Role-Playing Games (TRPGs). We introduce Babel Bardo, a system that uses Large Language Models (LLMs) to transform speech transcriptions into music descriptions for controlling a text-to-music model. Four versions of Babel Bardo were compared in two TRPG campaigns: a baseline using direct speech transcriptions, and three LLM-based versions with varying approaches to music description generation. Evaluations considered audio quality, story alignment, and transition smoothness. Results indicate that detailed music descriptions improve audio quality while maintaining consistency across consecutive descriptions enhances story alignment and transition smoothness.

## 1. INTRODUCTION

Recent text-to-audio music generation models such as MusicLM [1] and MusicGen [2] are capable of producing high-quality music in the audio domain that aligns with a given textual description. These models typically generate music autoregressively by predicting the next token from a context window, which limits the size of the signal they can model. While the context size is limited, these models can generate longer signals by sliding a context window through time. Regardless of this capability, they have mainly been evaluated with a fixed prompt and for relatively short music durations. For instance, MusicGen [2] was evaluated considering 30-second music pieces, each generated from a single music description. In this paper, we are interested in evaluating whether text-to-music models can maintain music quality while generating long music pieces, where music descriptions change over time.

It is important to evaluate text-to-music models consid-



**Figure 1.** At every 30 seconds of gameplay, Babel Bardo transcribes the players’ speeches into a text  $s_i$  using a Speech Recognition system and uses a Large Language Model (LLM) to map  $s_i$  into a music description  $d_i$  that matches the scene described by the players. This music description is given to a Text-to-Music system that generates a 30-second piece  $a_i$  directly in the audio domain.

ering long music pieces (greater than 30 seconds, for example) because many music production scenarios involve music durations longer than one can generate with a single short audio context window (e.g., pop music composition, jazz improvisation, soundtrack generation). One key problem of generating long sequences from a small context is that a model has to split the generation into multiple parts, ensuring that the independent parts are smoothly connected in the final composition. Moreover, one might change the initial prompt at any time step, steering the composition in a different direction, and the model must consider both the previous audio context and the new prompt.

In this paper, we investigate long generation with text-to-audio models in the context of Tabletop Role-Playing Games (TRPGs). In this scenario, a music generator takes speech as input and must generate music that matches the story being told by the players. We chose this problem because it inherently poses the challenge of long music generation, where prompts have to change over time to adjust for different story scenes. We also use TRPGs as a research object because TRPG players often enhance their gaming experience by manually selecting songs to play as background music [3], which allows us to compare the results of a generator against a human baseline.



© F. Marra, L. Ferreira. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** F. Marra, L. Ferreira, “Long-Form Text-to-Music Generation with Adaptive Prompts: A Case of Study in Tabletop Role-Playing Games Soundtracks”, in *Proc. of the 1st Latin American Music Information Retrieval Workshop*, Rio de Janeiro, Brazil, 2024.

To investigate the capabilities of current text-to-music models in generating background music for TRPG stories, we’ve built a system called Babel Bardo, which is inspired by Bardo Composer [4], a system that generates symbolic music by transcribing players’ speeches into text and conditioning an autoregressive model with the emotional tone of this text, as given by an emotion classifier. Different than Bardo Composer, Babel Bardo composes music directly in the audio domain by leveraging a Large Language Model (LLM) to transform the speech transcriptions into music descriptions every 30 seconds of gameplay. These descriptions are then given to a text-to-music model to generate a piece of music for that current moment of the story. Figure 1 shows an overview of our system. Babel Bardo is inspired by Herrmann1 [5], which uses LLMs and text-to-music models to generate soundtracks for films.

We compared four different versions of Babel Bardo in two TRPG campaigns played on YouTube: Call of the Wild (in English) and *O Segredo na Ilha* (in Brazilian Portuguese). The first version is our baseline and uses the speech transcriptions directly as prompts for a text-to-music model. All other versions use an LLM to transform the transcriptions into music descriptions. The second one follows the Bardo Composer approach and applies an LLM as an emotion classifier. The music description follows a template that is adjusted based on the emotion given by the LLM. The remaining two versions use the LLM to produce a complete music description; however, one generates a new description for every transcript, while the other can just continue the previously generated segment if the scene hasn’t changed.

We evaluated our models according to audio quality, alignment with the story, and transition smoothness between transcriptions. Results suggest that while detailed music descriptions contribute to improved audio quality, maintaining consistency across consecutive descriptions helps achieve smoother transitions between musical segments. Furthermore, our findings indicate that emotion serves as an effective signal for aligning generated music with TRPG narratives.

## 2. RELATED WORKS

This section reviews audio-based text-to-music models and previous soundtrack generation approaches, focusing on background music generation for TRPGs.

### 2.1 Text-to-Music Models

Text-to-music is the task of generating music pieces from music descriptions in textual format, e.g., “70s punk rock song with fast tempo”. In recent years, various neural models have been proposed to solve this problem in the audio domain [1, 2, 6]. For example, MusicGen [2] is an autoregressive transformer that operates on quantized audio units produced by the EnCodec [7] audio tokenizer. It can be conditioned on textual descriptions using various text encoding methods (e.g., T5 [8], FLAN-T5 [9], and

CLAP [10]), or on melodic structures through an unsupervised approach utilizing chromagram information.

MusicLM [1] is another text-to-music model that extends AudioLM’s [11] multi-stage autoregressive modeling approach by incorporating text conditioning, which is achieved by leveraging MuLan [12] to project music and textual descriptions into a shared embedding space. Moûsai [6] employs a two-stage cascading diffusion approach, where the first stage utilizes a novel diffusion magnitude-autoencoding (DMAE) technique to train a music encoder that compresses audio into a reduced representation. In the second stage, Moûsai implements text-conditioned latent diffusion (TCLD) to generate this reduced representation while conditioning on textual descriptions, enabling the model to produce music that corresponds to given text inputs. Other examples include commercial models such as Suno [13], Mubert [14], and Riffusion [15].

### 2.2 Soundtrack Generation

Soundtrack generation has been investigated in different mediums such as films [5], video games [16], stories [17], and others. This problem has been mainly studied in the symbolic domain. For example, Bardo Composer [4] generates background symbolic music for TRPGs by transcribing the players’ speeches into text at every time step  $t$  and feeding these transcriptions to a music classifier. The emotion given by this classifier is used to condition an autoregressive model that generates music by decoding a sequence using a variation of Stochastic Beam Search. Babel Bardo is similar to Bardo Composer because it also uses transcriptions of players’ speeches to condition the music generation. However, it generates music in audio format by conditioning a text-to-music model with a music description at every time step  $t$ , instead of using a music emotion classifier to condition an autoregressive symbolic music model.

Another important related work is Herrmann-1 [5], which combines an LLM and a text-to-music model to generate background music for movie scenes. Herrmann-1 uses BLIP2 [18] and CLIP [19] to extract a textual description and the affective characteristics of the video, respectively. These characteristics are then provided as input to GPT-4 [20], which generates a description of an appropriate music with the given characteristics. Finally, the description generated by GPT-4 is passed to MusicGen [2], which produces the background music in audio format. Babel Bardo is similar to Herrmann-1, because it also employs an LLM to generate music descriptions for a text-to-music, however, it takes text as input instead of videos. Moreover, in Babel Bardo, the music descriptions change over time, whereas Herrmann-1 uses a single description for each video.

## 3. BABEL BARDO

In TRPGs, players collaboratively construct a narrative through iterative cycles of scene descriptions, decision-making, and action resolutions. The game master presents

scenarios and environmental details, to which players respond by declaring their characters’ intended actions. These actions are then adjudicated using the game’s rule system, often involving probabilistic elements, with the outcomes shaping the evolving storyline and informing subsequent player choices. This process creates an emergent and interactive storytelling experience.

Babel Bardo generates music for a TRPG story by iteratively transcribing the players’ speeches into text and leveraging an LLM to produce a music description that is given to a text-to-music model, which in turn generates a piece of music. Formally, a story can be viewed as a sequence  $S = \{s_1, s_2, \dots, s_n\}$  of transcriptions, where each transcription  $s_i$  is a string generated after  $t$  seconds of gameplay. At each time step, Babel Bardo produces a music description  $d_i$  by asking an LLM to generate one that aligns with the transcript  $s_i$ . The music description  $d_i$  is then given as input to a text-to-music model that realizes the described music in audio format  $a_i$ . The text-to-music model also receives the previously generated audio  $a_{i-1}$  as a conditional input, so it has to generate a music piece based on the description  $d_i$  while continuing  $a_{i-1}$ . It is important to highlight that since we are using an LLM to produce music descriptions, Babel Bardo supports transcripts  $s_i$  in any language.

To illustrate our proposed generative pipeline, consider the transcription  $s_i = \text{"You see a dragon in front of you. A battle will start!"}$ . Babel Bardo could produce a description  $d_i = \text{"A grand orchestral arrangement with thunderous percussion, epic brass fanfares, and soaring strings."}$ , which would be used by the text-to-music model together with the previously generated audio  $a_{i-1}$  to produce the current audio segment  $a_i$ .

We evaluate four different approaches to combine an LLM and a text-to-music model to generate background music for TRPGs. All of them start with an initial prompt to condition the LLM with the generation task: *"You are going to receive a series of Role-playing Game (RPG) video transcript excerpts from players’ dialogues playing a campaign"*. After this initial setup, each model receives a sequence of transcripts  $s_i$ , and each approach employs the LLM in a different way to generate associated music descriptions  $d_i = LLM(s_i)$ .

**Babel Bardo - Baseline (B).** In this first version, Babel Bardo does not use the LLM to generate a music description. Instead, it uses the transcription  $s_i$  directly as a prompt to the text-to-music model ( $d_i = s_i$ ).

**Babel Bardo - Emotion (E).** This version behaves similarly to Bardo Composer and uses the LLM only as an emotion classifier. It processes the transcription  $s_i$  with the following prompt: *"Classify each dialogue into one of the following emotions: Happy, Calm, Agitated, or Suspenseful."* The LLM returns an emotion  $e_i$ , which is used to adjust the following pre-composed prompt  $d_i = \text{"Background music for a Role-playing Game (RPG) dialogue, with the following emotion: } e_i \text{"}$ . We used these four emotions because Bardo Composer [4] also used them.

TRPG	↓ FAD score				
	Babel Bardo				Human
	B	E	D	DC	
COTW	9.66	5.99	6.25	<b>5.82</b>	3.00
OSNI	9.55	6.11	5.63	<b>5.13</b>	4.18

**Table 1.** FAD for each Babel Bardo version in COTW and OSNI in contrast with Human music, i.e., the original soundtracks used by the player in both these TRPGs.

**Babel Bardo - Description (D).** In this third version, Babel Bardo employs the LLM to generate a description  $d_i = LLM(s_i)$  with the following prompt: *"For each transcript excerpt, describe a piece of background music that matches that excerpt."*

**Babel Bardo - Description Continuation (DC).** This last version is similar to the previous one; however, it allows the LLM to keep the same description  $d_i = d_{i-1}$  if the transcription  $s_i$  is part of the same scene as  $s_{i-1}$ . This is achieved by producing a description  $d_i = LLM(s_i)$  with the following prompt to the LLM: *"Determine whether this dialogue is from the same scene as the previous dialogue and based on this determination, either return the previous music description or generate a new one"*. This version is intended to help Babel Bardo keep a consistent soundtrack across a given scene by continuing  $a_{i-1}$  without changing the description.

#### 4. EXPERIMENTS AND RESULTS

We evaluate Babel Bardo<sup>1</sup> in the task of soundtrack generation for two different TRPG campaigns: Call of the Wild (COTW) and *O Segredo da Ilha* (OSNI). The former is a Dungeons & Dragons campaign played in American English and the latter is in Brazilian Portuguese. Both of them were played on YouTube. We used COTW because it was also used to evaluate Bardo Composer [4]. We’ve also included OSNI to evaluate Babel Bardo’s performance with a Latin American language. COTW is composed of 11 episodes, with a total of 6 hours and 37 minutes of gameplay—each episode is approximately 33 minutes long. OSNI is composed of 6 episodes, with a total of 26 hours and 22 minutes of gameplay—each episode is approximately 4 hours and 20 minutes long.

We measured the performance of Babel Bardo with respect to three objective metrics: audio quality, alignment with the story, and transition smoothness between transcriptions. Audio quality was measured using Fréchet Audio Distance (FAD) [21], which compares statistics computed on a set of reconstructed music clips to background statistics computed on a large set of studio-recorded music. Following the approach of Hermann1 [5], we collected 32 hours of high-quality cinematic soundtracks as reference studio-recorded music. Alignment with the story was cal-

<sup>1</sup> [github.com/FelipeMarra/babel-bardo](https://github.com/FelipeMarra/babel-bardo)

TRPG	↓ Mean KL-Divergence			
	Babel Bardo			
	B	E	D	DC
COTW	4.84±2.98	<b>3.34±1.89</b>	4.26±2.65	4.23±2.51
OSNI	5.65±3.23	<b>4.16±2.12</b>	4.85±2.62	4.96±2.86

**Table 2.** Mean/Standard Deviation of KLD for each Babel Bardo version in both COTW and OSNI.

culated with the Kullback-Leibler Divergence (KLD) with respect to the original background music of the campaigns. The transition smoothness was also computed with KLD, but by comparing the 10 seconds before and after a transition  $t_i$ , as shown in Figure 2.

In our experimental setup, we employed the YouTube API as our Speech Recognition system, extracting the transcriptions from the TRPG campaigns. We used Ollama 3.1, with 70B parameters, as Babel Bardo’s LLM for generating the music descriptions. As our text-to-music model, we used MusicGen [2] large, with 1.3B parameters. For every transcription  $s_i$ , we generate a 30-second long audio signal  $a_i$ , which is the maximum length MusicGen supports. We used the VGGish model for computing FAD scores. We computed the KL-Divergence with the PaSST [22] classifier, which was pre-trained on the AudioSet dataset [23]. We fine-tuned PaSST on the MTG-Jamendo dataset [24] to have a model more semantically suited for soundtrack classification (e.g., mood, genre, instrumentation). We fine-tuned PaSST for a single epoch with a learning rate of  $10^{-4}$ .

Table 1 presents the FAD audio quality metric for each Babel Bardo version and the original music from COTW and OSNI (Human). For each method, the FAD score is computed by retrieving a 30-minute window starting at the same random moment in both the reference background music and the generated one. This window is then split into 30-second segments. The FAD score is calculated between all generated samples of a method against the set of reference high-quality soundtracks. Since the lower the FAD value, the better, Babel Bardo- DC outperformed the other methods in both COTW and OSNI. These results suggest that conditioning MusicGen with more detailed music descriptions results in higher audio quality, both in English and Brazilian Portuguese stories. It is important to highlight that even though Babel Bardo- DC had a higher audio quality than the other versions, it is still not as good as professional human musical productions.

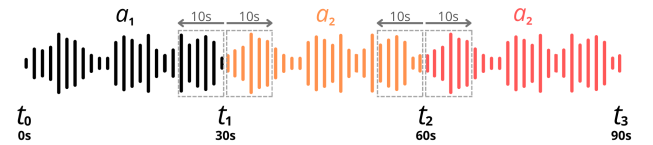
Table 2 shows the mean and standard deviation of the KLD story alignment metric for each Babel Bardo version in both COTW and OSNI. The means were calculated similarly to the FAD scores, but with slices of 10-second segments, since PaSST is limited to this context size. Moreover, each segment in the original background music had a respective segment in the generated piece. Babel Bardo - Emotion outperformed all other methods in both COTW and OSNI. These results indicate that the emotion of the

TRPG	↓ Mean Transition KLDs			
	Babel Bardo			
	B	E	D	DC
COTW	2.33±2.1	<b>1.33±1.19</b>	2.41±2.27	2.19±1.93
OSNI	2.11±1.65	<b>1.37±1.05</b>	1.88±1.47	2.09±1.71

**Table 3.** Mean/Standard Deviation of transition KLD for each Babel Bardo version in both both COTW and OSNI.

story is a strong signal for aligning music with TRPG stories. The lower performance of Babel Bardo - C and Babel Bardo - DC is probably because Ollama can generate descriptions that trigger high-quality audio but do not necessarily align well with the story.

Table 3 shows the mean and standard deviation of the KLD transition smoothness metric for each Babel Bardo version in both COTW and OSNI. These means were computed as in the previous metric, but comparing 10-second segments before and after a transition point  $t_i$ . Babel Bardo - Emotion outperforms all other methods in both COTW and OSNI. These results suggest that keeping a consistent music description  $d_i$  with very little change over time (i.e.,  $d_i \approx d_{i-1}$ ) helps MusicGen create smooth transitions between generated audio clips  $a_i$ . One reason for these results might be that when the music description  $d_i$  is similar to  $d_{i-1}$ , MusicGen focuses more on conditioning the new audio sample  $a_i$  to the previous audio  $a_{i-1}$  than to the new description  $d_i$ .



**Figure 2.** The transition KLD is computed between the 10 seconds before and after every transition moment  $t_i$ .

## 5. CONCLUSION AND FUTURE WORK

This paper presented Babel Bardo, a system that combines an LLM and a text-to-music model to generate background music for tabletop role-playing games. Our goal with Babel Bardo was to evaluate the performance of text-to-music models in long-generation tasks. We’ve presented four different versions of the system and evaluated them in two TRPG campaigns, one in English and another in Brazilian Portuguese. Results showed that while detailed music descriptions help improve audio quality, it is important to maintain consistency across consecutive descriptions to have smoother transitions. Moreover, emotion is a strong signal for generating soundtracks for TRPGs.

As future work, we plan to investigate how to maintain the consistency of the generated music over time while still using detailed music descriptions. Moreover, we will conduct user studies to evaluate the quality of the generated music with subjective metrics.

## 6. REFERENCES

- [1] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, “Musiclm: Generating music from text,” *arXiv preprint arXiv:2301.11325*, 2023.
- [2] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, “Simple and controllable music generation,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [3] K. Bergström and S. Björk, “The case for computer-augmented games,” *Transactions of the Digital Games Research Association*, vol. 1, no. 3, 2014.
- [4] L. Ferreira, L. Lelis, and J. Whitehead, “Computer-generated music for tabletop role-playing games,” in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 16, no. 1, 2020, pp. 59–65.
- [5] M. T. Haseeb, A. Hammoudeh, and G. Xia, “Gpt-4 driven cinematic music generation through text processing,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6995–6999.
- [6] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, “Moûsai: Text-to-music generation with long-context latent diffusion,” *arXiv preprint arXiv:2301.11757*, 2023.
- [7] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *Transactions on Machine Learning Research (TMLR)*, 2023.
- [8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research (JMLR)*, vol. 21, no. 140, pp. 1–67, 2020.
- [9] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, “Scaling instruction-finetuned language models,” *Journal of Machine Learning Research (JMLR)*, vol. 25, no. 70, pp. 1–53, 2024.
- [10] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [11] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, “Audiolm: a language modeling approach to audio generation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 31, pp. 2523–2533, 2023.
- [12] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, “Mulan: A joint embedding of music audio and natural language,” *International Society for Music Information Retrieval (ISMIR)*, 2022.
- [13] Suno-Inc, “Suno,” <https://www.suno.com>, 2022.
- [14] Mubert-Inc, “Mubert,” <https://mubert.com/>, 2022, <https://github.com/MubertAI/Mubert-Text-to-Music>.
- [15] Riffusion-Inc, “Riffusion,” <https://www.riffusion.com>, 2022, <https://github.com/riffusion/riffusion-hobby>.
- [16] I. Cardoso, R. O. Moraes, and L. N. Ferreira, “The nes video-music database: A dataset of symbolic video game music paired with gameplay videos,” in *Proceedings of the 19th International Conference on the Foundations of Digital Games*, 2024, pp. 1–6.
- [17] H. Davis and S. M. Mohammad, “Generating music from literature,” *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL)*, pp. 1–10, 2014.
- [18] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 19 730–19 742.
- [19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.
- [20] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [21] D. Roblek, K. Kilgour, M. Sharifi, and M. Zuluaga, “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms,” in *Proc. Interspeech*, 2019, pp. 2350–2354.
- [22] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, “Efficient training of audio transformers with patchout,” *Interspeech*, 2022.
- [23] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.
- [24] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, “The mtg-jamendo dataset for automatic music tagging,” in *International Conference on Machine Learning (ICML)*, 2019.

# Author Index

Adu-Gilmore, Leila: 15  
Almada, Carlos de Lemos: 20  
Andrade, Nazareno: 25  
Biscainho, Luiz W P: 57, 63  
Carvalho, Hugo T: 20, 46  
de Abreu, Wallace C: 57  
de Matos, Natanael L.: 46  
de Tomaz Júnior, Pedro Donadio: 63  
dos Santos, Guilherme S: 41  
Fernandes Jr, Antonio Carlos Lopes: 78  
Fernandez Slezak, Diego: 9  
Figueiredo, Flavio: 25, 30, 41  
Fuentes, Magdalena: 35  
Galuppo Azevedo, Francisco: 30  
Harper, Colter: 15  
Jordanous, Anna: 68  
Lima, Rennan: 30  
Louro, Pedro L.: 52  
Malheiro, Ricardo S: 52  
Marra, Felipe F: 73

Martins, Felipe D: 20  
Matos, Breno S: 30  
McFee, Brian: 35  
Miguel, Martin A: 9  
Morais, Giovana V: 35  
N. Ferreira, Lucas: 73  
Paiva, Rui P: 52  
Panda, Renato: 52  
Panoutsos, Tales: 25  
Rapini, Antonin PL: 68  
Redinho, Hugo: 52  
Riera, Pablo: 9  
Rocamora, Martín: 63  
Roman, Iran R: 15  
Simas, Eduardo: 78  
Somacal, Lucas: 9  
Van Ert, Kelsey: 15  
Viana, Luiz Alberto Guimarães: 78  
Walls, Kelvin: 15  
Zanini, Carlos Tadeu Pagani: 46