DMRN+19: Digital Music Research Network

One-day Workshop 2024

Queen Mary University of London

Tuesday 17th December 2024

Chair: Simon Dixon

Programme

9:40	***Registration - Coffee***
10:00	**Welcome - Simon Dixon**
10:10	**KEYNOTE** Models of Musical Signals: Representation, Learning & Generation, **Stefan Lattner (Research Leader at Sony CSL Paris)**
11:10	*Break (Coffee break)*
11:30	"PAGURI: a user experience study of creative interaction with text-to-music models", **Francesca Ronchini, Luca Comanducci, Gabriele Perego and Fabio Antonacci (Politecnico di Milano, Italy)**
11:50	"REBUS: Exploring the space between instrument and controller", **Eleonora Oreggia (Goldsmiths, University of London, UK)**
12:10	"An Improvisation Analysis Method with Machine Learning for Embodied Rhythm Research", **Evan O'Donnell (Goldsmiths, University of London, UK)**
12:30	Creative Apps Hackathon Winners Presentation; **Jean-Baptiste Thiebaut (Music Hackspace, UK) and György Fazekas (Queen Mary University of London, UK)**
12:45	*Lunch - Poster Session*
14:15	"Analysis of MIDI as Input Representations for Guitar Synthesis", **Jackson Loth, Pedro Sarmento, Saurjya Sarkar and Mathieu Barthet (Queen Mary University of London, UK)**
14:35	"Emulating LA-2A Optical Compressor with a Feed-Forward Digital Compressor Using the Newton-Raphson Method", **Chin-Yun Yu and György Fazekas (Queen Mary University of London, UK)**
14:55	"AFX-Research: a repository and website of audio effects research"**, Marco Comunità and Joshua D. Reiss (Queen Mary University of London, UK)**
15:15	*Break (Coffee break)*
15:35	"Characterizing Jazz Improvisation Style Through Explainable Performer Identification Models", **Huw Cheston, Reuben Bance and Peter Harrison (Centre for Music & Science, Cambridge, UK)**
15:55	"Framework for Predicting Eurovision Song Contest Results," **Katarzyna Adamska and Joshua D. Reiss (Queen Mary University of London, UK)**
16:15	"MINDS: Mutual Inclusion through Neurodiversity in Science", **Daniel Gill (Queen Mary University of London, UK)**
16:35	Close – Emmanouil Benetos

* - There will be an opportunity to continue discussions after the Workshop in a nearby Pub/Restaurant for those in London.

Posters

1	"Improving Automatic Guitar Tablature Transcription with LLMs", **Omar Ahmed (University of Oxford and Queen Mary University of London, UK), Pedro Sarmento and Emmanouil Benetos (Queen Mary University of London, UK)**
2	"Towards Detecting Interleaved Voices in Telemann Flute Fantasias", **Patrice Thibaud, Mathieu Giraud and Yann Teytaut (Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France)**
3	"Embodied Movement-Sound Interaction in Latent Terrain Synthesis", **Shuoyang Zheng, Anna Xambó Sedó (Queen Mary University of London, UK) and Nick Bryan-Kinns (University of the Arts London, UK)**
4	"Towards Differentiable Digital Waveguide Synthesis", **Pablo Tablas de Paula and Joshua D. Reiss (Queen Mary University of London, UK)**
5	"Classification of Spontaneous and Scripted Speech for Multilingual Audio", **Shahar Elisha (Spotify Ltd, Queen Mary University of London, UK), Andrew McDowell, Mariano Beguerisse-Díaz, (Spotify Ltd, UK) and Emmanouil Benetos (Queen Mary University of London, UK)**
6	"Personalising equalisation using psychological and contextual factors", **Yorgos Velissaridis, Charalampos Saitis and György Fazekas (Queen Mary University of London, UK)**
7	"Can downbeat trackers predict hypermetre?" **Jose Alejandro Esquivel de Jesus and Jordan B. L. Smith (Queen Mary University of London, UK)**
8	"A Transposition-Invariant Chord Encoder for Bigram Modelling" **Yuqiang Li and György Fazekas (Queen Mary University of London, UK)**
9	"Multimodal techniques for the control of procedural audio", **Xavier Marcello D'Cruz and Joshua D. Reiss (Queen Mary University of London, UK)**
10	"Towards differentiable modular approaches for dynamic sound synthesis: A case study in vehicular sound effects", **Minhui Lu and Joshua D. Reiss (Queen Mary University of London, UK)**
11	"Procedural Music Generation for Games", **Shangxuan Luo and Joshua D. Reiss (Queen Mary University of London, UK)**
12	"Advancing Expressive Performance Rendering in Pop Music Using Computational Models", **Jinwen Zhou and Aidan Hogg (Queen Mary University of London, UK)**

Location

Arts Two Theater
Queen Mary University of London - Mile End Campus

Keynote Talk

Keynote: By Stefan Lattner (Research Leader at Sony CSL Paris)

Tittle: Models of Musical Signals: Representation, Learning & Generation

Abstract: Low-level audio representations and higher-level representation learning are at the heart of music analysis and synthesis. Thus, the talk will dive into some previous works of Sony CSL on audio representations, covering different concepts and use cases. Learning first- and second-order basis functions to obtain desired invariances, investigating the choice of low-level audio representations for generation, self-supervised learning of higher-order representations, and audio codecs. Finally, musical audio synthesis will be discussed, ranging from GANs to latent diffusion to recent advancements in continuous autoregressive models.

Bio: Stefan Lattner serves as a researcher leader at the music team at Sony CSL Paris, where he focuses on generative AI for music production, music information retrieval, and representation learning. He earned his PhD in 2019 from Johannes Kepler University (JKU) in Linz, Austria, following his research at the Austrian Research Institute for Artificial Intelligence in Vienna and the Institute of Computational Perception Linz. His studies centered on the modeling of musical structure, encompassing transformation learning and computational relative pitch perception. His current interests include human-computer interaction in music creation, live staging, and information theory in music. He specializes in latent diffusion, self-supervised learning, generative sequence models, computational short-term memories, and models of human perception.

More information on: **https://csl.sony.fr/member/stefan-lattner-phd/**

Organizing Committee

Supported by UKRI AIM CDT

UK Research and Innovation Centre for Doctoral Training in Artificial Intelligence and Music.

Bradley Aldous
Keshav Bhandari
Louis Bradshaw
Julien Guinot
Zixun (Nicolas) Guo
Adam He
Gregor Meehan
Marco Pasini
Christos Plachouras
Haokun Tian
Qing Wang
Yifan Xie
Farida Yusuf
Qiaoxi Zhang
Shuoyang Zheng

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

PAGURI: a user experience study of creative interaction with text-to-music models

Francesca Ronchini, Luca Comanducci, Gabriele Perego and Fabio Antonacci

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy, francesca.ronchini@polimi.it

***Abstract*— In recent years, text-to-music models have been the biggest breakthrough in automatic music generation. This paper aims to address this question via Prompt Audio Generation User Research Investigation (PAGURI), a user experience study where we leverage recent text-to-music developments to study how musicians and practitioners interact with these systems, evaluating their satisfaction levels. We developed a tool through which users can generate music samples and/or apply recently proposed personalization techniques, based on fine-tuning, to make the text-to-music model generate sounds closer to their needs and preferences. Using questionnaires, we analyzed how participants interacted with the proposed tool, to understand the effectiveness of text-to-music models in enhancing users' creativity. Results show that even if the audio samples generated and their quality may not always meet user expectations, the majority of the participants would incorporate the tool in their creative process. Furthermore, they provided insights into potential enhancements for the system and its integration into their music practice.**

***Index Terms*—** Text-to-Music, Generative Models, Human-AI Interaction, Human-computer co-creativity.

I. INTRODUCTION

The introduction of TTM models has lowered the technical competencies needed to use music generative models, posing the need to consider if and how AI is introducible into music creative practice. Although generative models for music are gaining more and more popularity, there is still a lack of comprehensive research on this topic. We advocate that it is important to conduct this type of research closely with potential final users and music practitioners, to develop tools that not only showcase the impressive capabilities of the technology in the music field but also highlight how these tools can be viewed and perceived as musical instruments for creating music.

This study proposes Prompt Audio Generator User Research Investigation (PAGURI), which aim to analyze if TTM models are ready to be used as tools for music creation and composition and explore their potential integration into the music creation process based on feedback from users and music professionals. Our main focus is determining the specific phases and purposes for which users would apply these tools in their creation processes, and where they are considered most useful.

II. METHOD

We developed an interface through which the user can indicate through a text prompt the audio sample they wish to generate. Additionally, the user can upload up to 5 desired audio samples to personalize the TTM generative model. We use AudioLDM2 [1] as a generative model, and a TTM personalization technique proposed in [2] to let the users fine-tune the model according to the music samples of their choice. We then conducted a user experience study, both online and in presence, where users can interact with PAGURI and we analyzed their experience using questionnaires and open questions. Specifically, we first quantify their background related to music and AI tools, and then we analyze their level of satisfaction after each use of the TTM model, until reaching the desired result. Finally, we let them answer a questionnaire analyzing the whole experience and gathering information related to the perceived usability of the TTM model in music practice.

III. RESULTS AND CONCLUSION

Results show that these models have several application opportunities, not only in the music creative process. Particular attention was given to the opportunities of the model's personalization. At the same time, worries about plagiarism and unauthorized use of generated personalized sounds were raised. Moreover, the majority of participants strongly asserted that these tools have significant growth potential and offer a wide range of applications across various domains. Future works aim to include these results in TTM generative models and develop a better interface to allow further interaction and control for the final users.

IV. REFERENCES

[1] H. Liu, Q. Tian, Y. Yuan, X. Liu, X. Mei, Q. Kong, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *arXiv preprint arXiv:2308.05734*, 2023.

[2] M. Plitsis, T. Kouzelis, G. Paraskevopoulos, V. Katsouros, and Y. Panagakis, "Investigating personalization methods in text to music generation," *arXiv preprint arXiv:2309.11140*, 2023.

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Exploring the space between instrument and controller

Eleonora Oreggia*

Computing Department, Goldsmiths, University of London, UK, e.oreggia@gold.ac.uk

***Abstract*— Electromagnetic sensing systems allow for expressive interaction in electronic music, performance, and time-based media art. REBUS is a machine which emits and receives electromagnetic waves to form a sensing space where human movement and presence can be detected with previously unknown precision. The potential of the system is explored through a workshop and a study targeted to selected groups: whereas the workshop observed the approach of coders and musicians in proposing compositional methods that this electromagnetic field sensing instrument offers, the study inspected the behaviour of classical musicians, mainly trained on strings or keyboard instruments, vis-a-vis that of selected electronic musicians. The terrain between instrument and controller is also explored through a fruitful deception. The paper narrates the conceptual assumptions behind these experiences through an initial analysis and some reflections on the data collected.**

***Index Terms*—** Musical instrument, controller, electromagnetic waves, audiovisual, performance

I. ACKNOWLEDGMENTS

The study described in this paper was made possible by Early Career Research Fund, Goldsmiths University of London. The authors would like to thank Rebecca Superbo, Sebastian Gafencu and Niccolò Perego from Polytechnic of Milan (POLIMI) for their assistance during the workshop and for contributing code and ideas for the study. Last but not least, gratitude to Professor Augusto Sarti and Professor Mark D'Inverno for believing in our work.

*Research supported by Early Career Reaserch Network.

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Evan O'Donnell

Department of Music, Goldsmiths, University of London, United Kingdom,
eodon002@gold.ac.uk

An Improvisation Analysis Method with Machine Learning for Embodied Rhythm Research

Abstract

This project introduces a hybrid music improvisation analysis method incorporating gestural data, allowing for retroactive conversion of qualitative analyses into machine learning (ML) models. It is specifically intended as a tool for investigating associations between gesture and rhythmic phrasing, harnessing practice research methods to uncover questions and correlations for future research relevant to both musical practice and interactive AI development.

This method has roots in a broader inquiry into microtiming in computer music practice, strongly influenced by Vijay Iyer's work examining relationships between embodiment and rhythmic expression [1]. It is also inspired by an interest in "small data" AI development, the issue of cultural diversity and musician involvement in data set creation, and challenges in building interactive AI-based rhythm models. Previous work by Zbyszyński, Reed, Tanaka, and others using gestural interfaces in embodied musical practice has provided invaluable insight into these questions [2, 3], but few of these studies have focused specifically on rhythmic phrasing. Meanwhile, practice researchers such as Ben Spatz have detailed rigorous systems for describing the specifics of embodied technique [4]. Combining these research tools could offer a fresh angle for understanding embodied rhythmic expression, generating new lines of inquiry via creative practice.

The approach presented here adds gestural data and machine learning to components of an improvisation analysis system developed by Rodrigo Constanzo [5]. Users film and record rhythm-based improvisations while wearing gestural sensors. (For clarity, I am currently performing improvisations and recording expressive gestures with separate arms, but other approaches are possible.) Performers then conduct auto-analyses using the video, audio, and gestural data gathered, leaving time-marked comments about specific movements and phrases they feel have interesting or relevant correlations, and classifying these according to meaningful common characteristics. A shared time base for all data and commentary means that these classifications can be used to segment audio and gesture data into relevant pairs for exploratory machine learning at any point in the future, creating models that allow the original musicians or others to test the identified correlations, offering further insight via embodied interaction. The larger goal is an accumulation of gesture-to-phrase correlations which form a basis for more detailed investigations into embodied rhythmic phrasing, firmly rooted in musical practice.

I am currently refining this method within my own practice as part of a larger gesture-based performance and composition project for the following year, and will share some of my initial findings as a basis for further discussion. In the long run I would also like to try this approach with other musicians, and I'm curious how an evolution of this method could help in documenting a more diverse range of embodied relationships to rhythm, in a way that is useful to both HCI researchers and practicing musicians.

REFERENCES

1. Iyer, V. (2002) 'Embodied Mind, Situated Cognition, and Expressive Microtiming in African-American Music,' in *Music Perception: An Interdisciplinary Journal*, 19(3), pp. 387-414.
2. Reed, C.N. *et al.* (2024) 'Sonic Entanglements with Electromyography: Between Bodies, Signals, and Representations,' *Designing Interactive Systems Conference*, pp. 2691–2707.
3. Zbyszyński, M. *et al.* (2021) 'Gesture-Timbre Space: Multidimensional feature mapping using machine learning and concatenative synthesis,' in *Lecture notes in computer science*, pp. 600–622.
4. Spatz, B. (2017) 'Colors Like Knives: Embodied Research and Phenomenotechnique in Rite of the Butcher,' in *Contemporary Theatre Review,* 27(2), pp. 195–215.
5. *Analysis Guide « Rodrigo Constanzo* (no date). https://rodrigoconstanzo.com/analysis-guide/.

Music Technology Hackathon: Build a Creative App in 24 hours

Jean-Baptiste Thiebaut[1], Tiberius Treppner[2], Sebastian Murgul[3], György Fazekas[4]

[1]Music Hack Space, [2]Halbestunde, [3]Klangio [4]Queen Mary University of London

Abstract

I. Music Technology Hackathon: Build a Creative App in 24 hours will be held at QMUL campus on 14th and 15th December 2024.

Join us at the Centre for Digital Music and collaborate with fellow attendees in groups of 3-4 to **build a creative app** from scratch. You can build anything ranging from audio utilities, like a tuner, to music generators or sound effects. Each team gets to decide what they want to build.

A jury will evaluate the best hacks at the end of the hackathon. The presentations will be livestreamed on the YouTube channel of our partner, The Audio Programmer, with an international jury evaluating the presentations. All apps deemed ready may be released on Muse Hub, and the winner will be featured as the Indie app of the month! This is the ideal launchpad for new ideas and future startups.

About the organisers

This event is organised by Muse Hub, in partnership with the Centre for Digital Music, and supported by The Audio Programmer. The event was held on 14 and 15 December 2024 at the Centre for Digital Music, Queen Mary University of London.

Winners' presentation at DMRN+19

Collider: A chaotic, physics-based delay plugin.

We were interested in exploring ways a physical model could be tied to audio - and how that might sound. We were hoping to see how that might introduce an element of chaos or randomness, but also how that could be used musically.

Collider was built in C++ using JUCE and the open source https://github.com/sudara/melatonin_blur library to achieve a degree of glass-morphism.

Music Transcription Studio.

Writing down sheet music by ear is a time consuming and difficult task. Wouldn't it be nice to have a MP3 to Sheet Music feature for your favorite Sheet Music Editor (it's MuseScore)!

We have already built some web apps for music transcription and would love to release them on MuseHub. It is based on the Klangio Transcription API (https://klang.io/api).

Gamified Auditory Training for Hearing Rehabilitation

The app offers gamified auditory training for those recovering from hearing loss, with modules for speech-in-noise, memory, and 3D sound to enhance listening skills.

Rehabilitating hearing capabilities requires engaging, effective, and scientifically backed training methods. Traditional approaches lack interactivity, making it harder for users to stay motivated. Inspired by the need for accessible and enjoyable auditory rehabilitation tools, we designed an app that provides multiple auditory training games for individuals recovering from hearing loss or surgeries

Winner Most creative:

- RelaxSync: Controlling music via user feedback to achieve a goal.
- Collider: A chaotic, physics-based delay plugin
- Music Transcription Studio.
- DAW-XR: Mixed Reality Digital Audio Workstation in a Browser

Winner Ready for Muse Hub:

- Writing down sheet music by ear is a time consuming and difficult task.
- Collider: A chaotic, physics-based delay plugin.
- Ambiful. Granular effect engine for adding stereo, textured ambience to tracks that might need movement

Winner Most fun

- JamSync: Collaborative space for musicioans
- Memoloop: experimental recorder for sound collage
- Wind Chime. Use the wind as a composition tool

More information on Muse Hub Hackathon: https://muse-hub-hackathon.devpost.com/project-gallery

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Analysis of MIDI as Input Representations for Guitar Synthesis

Jackson Loth , Pedro Sarmento, Saurjya Sarkar and Mathieu Barthet

Centre for Digital Music, Queen Mary University of London, United Kingdom, j.j.loth@qmul.ac.uk

***Abstract*— We analyse the effectiveness of MIDI as an input representation in the task of guitar synthesis. Given the results, we argue that more comprehensively annotated datasets are needed for this task.**

***Index Terms*—** Guitar, Synthesis, MIDI, Diffusion;

I. INTRODUCTION

Musical Instrument Digital Interface (MIDI) is a common representation in music instrumental synthesis, as it easily represents note oneset, pitch, duration and velocity. While this might work great for instruments such as piano, instruments such as guitar have significantly more expressive and dynamic control over the sound. To investigate this, we utilise an implementation[1] of a diffusion-based synthesis model [1] which has been trained to synthesise realistic audio from MIDI notes. Examples from guitar performances which have been transcribed to MIDI from the GuitarSet [2] and EGDB [3] datasets are used to generate examples. Thus, we can compare real guitar performances to synthesised versions[2] of the same performance and annotate areas where they differ.

II. ANALYSIS

The timbre of the synthesised audio exhibits some features akin to a real guitar, including some realistic imperfections. However, issues arise with expressive playing techniques like slides, hammer-ons, pull-offs, muted notes, and bends, which are predominantly focused around note transitions and dynamics that are not captured by MIDI. The model struggles to replicate these techniques accurately but can produce transitions resembling a mix of hammer-ons and pull-offs when one MIDI note starts as another ends. This sometimes leads to "phantom" transitions in the synthesised audio (i.e. generated due to adjacent MIDI notes) even if the transition is absent in the original recording.

Furthermore, none of the bends from the original audio appear in the synthesised audio, as the MIDI only records the starting note of the bends. This results in awkward, lifeless audio. A muted string is annotated as a simple MIDI note, leading the model to synthesise a normal note. The model also misses ringing notes from imperfect string muting, as these are absent in the transcription. One synthesised note resembles a slide but sounds more like a violin than a guitar, likely due to the model being trained on multiple instruments, not just guitar.

III. DISCUSSION

Most issues stem from the fact that existing guitar datasets do not encode information beyond the note pitch and duration, limiting the synthesis results in several ways. The model can only generate transitions, like slides or hammer-ons, by placing MIDI notes directly adjacent. This requires a gap between picked notes, making fast alternate picking difficult to reproduce. Additionally, transitions like slides lack control over the rate of change. The model struggles to handle muted strings and playing imperfections, as MIDI assumes all events have a pitch. These imperfections add character and should be included in re-synthesis.

There are a few possible reasons why existing guitar datasets do not typically encode expressive playing techniques. For one, including additional annotations would likely significantly increase the workload to create a dataset. Secondly, using MIDI as a representation allows the use of models that have been already built with MIDI in mind, as well as benchmarking against previous work more easily. While MIDI does support control channels which could be used to encode these techniques, there is no standardisation for this procedure, thus potentially limiting the use of any dataset which attempts to use this feature.

These results lead to a conclusion that existing guitar datasets are insufficient for detailed, realistic guitar synthesis. A dataset with better annotations for expressive techniques (e.g. tablature-based similar to the DadaGP dataset [4]) would greatly increase the capabilities of guitar synthesis models, and likely help in many other guitar-related tasks as well such as transcription and playing technique recognition.

IV. REFERENCES

[1] C. Hawthorne, I. Simon, A. Roberts, N. Zeghidour, J. Gardner, E. Manilow, and J. Engel, "Multi-Instrument Music Synthesis with Spectrogram Diffusion," *ISMIR*, 2022.

[2] Q. Xi, R. M. Bittner, J. Pauwels, X. Ye, and J. P. Bello, "GuitarSet: A Dataset for Guitar Transcription," *ISMIR*, 2018.

[3] Y.-H. Chen, W.-Y. Hsiao, T.-K. Hsieh, J.-S. R. Jang, and Y.-H. Yang, "Towards Automatic Transcription of Polyphonic Electric Guitar Music: A New Dataset and a Multi-Loss Transformer Model," *ICASSP*, 2022.

[4] P. Sarmento, A. Kumar, C. J. Carr, Z. Zukowski, M. Barthet, and Y.-H. Yang, "DadaGP: A Dataset of Tokenized GuitarPro Songs for Sequence Models," *ISMIR*, July 2021.

Research supported by the EPSRC UKRI Centre for Doctoral Training in Artificial Intelligence and Music (Grant no. EP/S022694/1) and UKRI - Innovate UK (Project no. 10102804).

[1] Available in this repository.

[2] Audio samples available to listen here

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Emulating LA-2A Optical Compressor With a Feed-Forward Digital Compressor Using the Newton-Raphson Method

Chin-Yun Yu* and György Fazekas

Centre for Digital Music, Queen Mary University of London, UK, chin-yun.yu@qmul.ac.uk

***Abstract*— This paper presents a method to emulate the LA-2A optical compressor using a feed-forward digital compressor. The Newton-Raphson algorithm is used to find the optimal parameters that minimise the differences between the actual recordings and the compressor's output. We build an LA-2A emulation plugin using the learnt parameters, serving as a creative tool for music producers.**

***Index Terms*—** Virtual analog modelling, Newton-Raphson, feed-forward compressor, differentiable DSP

I. INTRODUCTION

Compressors control audio dynamic range and are often used in music production and broadcasting. Early analog compressors, such as the popular LA-2A by Universal Audio, have iconic sounds that have attracted considerable interest in replicating their characteristics in the digital world, particularly as an audio plugin inside a DAW.

This paper explores modelling the LA-2A (its peak reduction parameter specifically) by matching its sound using a simple feed-forward digital compressor. Although the expressiveness is constrained by the standard five parameters (threshold, ratio, make-up gain, attack, and release), the learnt mapping ($\mathbb{R} \to \mathbb{R}^5$) can give us a further understanding of this classic unit and enables more creative control than just replicating the sound.

II. METHODOLOGY

Given an input signal $\mathbf{x}$, we aim to match the sound of a feed-forward compressor $\hat{\mathbf{y}} = f_{\mathbf{x}}(\theta)$ to the real recordings $\mathbf{y}$ of LA-2A by finding the optimal parameter values that minimise the L_2 distance between them as $\theta^* = \min_\theta \mathcal{L}(\theta)$ where $\mathcal{L}(\theta) = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})$. Both $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ are from the SignalTrain dataset [1], and $\theta \in \mathbb{R}^5$ are the parameters.

Assuming $\mathcal{L}$ is convex, we can get θ^* using the Newton-Raphson method, which performs the following step repeatedly until convergence:

$$\theta \leftarrow \theta - [\nabla^2 \mathcal{L}(\theta)]^{-1} \nabla \mathcal{L}(\theta). \tag{1}$$

We use the differentiable implementation of $f_{\mathbf{x}}$ by Yu et al. [2] to calculate both the gradients ($\nabla \mathcal{L}(\theta) \in \mathbb{R}^5$) and the Hessian matrix ($\nabla^2 \mathcal{L}(\theta) \in \mathbb{R}^{5\times5}$) in PyTorch.

III. RESULTS

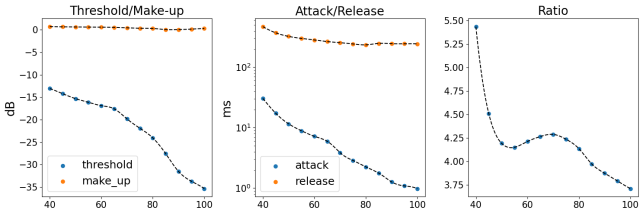

Figure 1: The resulting mapping from LA-2A peak reduction (x-axis) to the compressor's threshold, attack/release time, ratio, and make-up gain.

We show the learnt parameters with peak reduction ranging from 40 to 100 with a spacing of 5 in Fig. 1. The five parameters generally follow some smooth trajectories that can be approximated easily. For example, the threshold is approximately linear, and the attack/release varies exponentially to peak reduction.

We build an LA-2A emulation plugin using APE [3] that runs our feed-forward compressor under the hood based on the above results[1]. We linearly interpolate the parameters for peak values not included in the training set.

IV. CONCLUSION AND FUTURE WORK

This paper demonstrates that the behaviour of LA-2A can be described and controlled by the parameters of a feed-forward compressor using the Newton-Raphson method. We plan to publish the resulting model as a standalone plugin using JUCE and evaluate different interpolation methods for the LA-2A's peak reduction parameter.

V. REFERENCES

[1] B. Colburn and S. Hawley, "SignalTrain LA2A Dataset," May 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3824876

[2] C.-Y. Yu, C. Mitcheltree, A. Carson, S. Bilbao, J. D. Reiss, and G. Fazekas, "Differentiable all-pole filters for time-varying audio systems," in *International Conference on Digital Audio Effects (DAFx)*, Guildford, UK, 2024.

[3] J. L. Thorborg, "Audio programming environment," 2021, https://www.jthorborg.com/ape.html [Accessed: (2024-11-20)].

*Research supported jointly by UKRI (grant number EP/S022694/1) and Queen Mary University of London.

[1] https://github.com/aim-qmul/4a2a

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

AFxResearch: a repository and website of audio effects research

Marco Comunità and Joshua D. Reiss

Centre for Digital Music, Queen Mary University of London, UK, m.comunita@qmul.ac.uk

***Abstract*— We present *AFxResearch*[1], a repository and associated website[2] gathering scientific literature about audio effects. Our database includes topics like: modeling, classification, estimation, removal, style transfer, processing and review papers. Our website contains a detailed table of publications, with search, filter and sort features, while the repository enables users to submit requests for new entries or revise existing ones.**

I. REPOSITORY AND WEBSITE

Research on audio effects has significantly expanded over the last few decades [1], driven by advances in digital signal processing [2], machine learning [3], and auditory perception. As the body of literature on audio effects grows, the need for a centralized repository with tools for easy access and exploration becomes essential. *AFxResearch* addresses this by providing a comprehensive and flexible database of publications, serving as a go-to resource for researchers, educators, and practitioners. *AFxResearch* is hosted on a user-friendly website that facilitates exploration through search, filter, and sort functionalities. It also supports community-driven updates, allowing users to submit new publications or suggest modifications, ensuring the database remains current and reflects the evolving field of audio effects research.

II. DATABASE

Metadata — The database offers detailed metadata for each publication, allowing users to assess its relevance before accessing the full text. At the moment of publication, the metadata includes: title, author(s), paper URL, publication date, main task, paradigm(s), device(s) type(s), device(s) model, method(s), webpage URL, code URL, dataset URL, abstract.

Tasks — Publications are categorized based on the primary task they address, included and not limited to:

- Classification/Identification - studies that classify different types of audio effects (e.g., distortion, phaser, reverb) or identify specific devices (e.g., ProCo Rat distortion, Teletronix LA2A compressor) from audio signals [4].
- Estimation/Regression - works concerned with estimating the controls settings (e.g., gain, cutoff frequency, modulation speed) used to process audio examples or the internal coefficients of processing blocks (e.g., all-pass filter, biquad filter, low-frequency oscillator) [5].
- Modeling - research on developing mathematical or computational models of audio effects [6].
- Removal - research aimed at removing audio effects from processed signals.
- Style Transfer - studies about replicating the sonic characteristics of a reference audio example onto an input example, independent of content, effects, or processing methods used [7].
- Processing - broad category about processing audio signals that includes: automatic audio effects control, automatic mixing, audio processing graph estimation, creative uses of audio effects or derivation of new ones.
- Review - overviews of a specific subtopic or task.

Paradigms — The database includes publications that employ any modeling or emulation paradigms: white-box, gray-box [8], black-box.

Methods — The methods section categorizes publications based on the technical approaches and tools used in the research. Common methods include: differentiable DSP , dynamic convolution, equations solving or approximation, neural networks [3], state-space, wave digital filters, port-hamiltonian, Volterra series, waveshaping, Wiener-Hammerstein.

[1] https://github.com/mcomunita/afx-research
[2] https://mcomunita.github.io/afx-research

III. REFERENCES

[1] T. Wilmering, D. Moffat, A. Milo, and M. Sandler, "A history of audio effects," *Applied Sciences*, vol. 10, no. 3, p. 791, 2020.

[2] U. Zölzer, X. Serra, M. Sandler, *et al.*, "Digital audio effects," pp. 1–2, 2011.

[3] T. Vanhatalo, P. Legrand, M. Desainte-Catherine, P. Hanna, A. Brusco, G. Pille, and Y. Bayle, "A review of neural network-based emulation of guitar amplifiers," *Applied Sciences*, vol. 12, no. 12, p. 5894, 2022.

[4] M. Comunità, D. Stowell, and J. Reiss, "Guitar effects recognition and parameter estimation with convolutional neural networks," *Journal of the Audio Engineering Society*, vol. 69, no. 7/8, pp. 594–604, 2021.

[5] C. Mitcheltree, C. Steinmetz, M. Comunità, and J. Reiss, "Modulation extraction for lfo-driven audio effects," *arXiv preprint arXiv:2305.13262*, 2023.

[6] M. Comunità, C. Steinmetz, H. Phan, and J. Reiss, "Modelling black-box audio effects with time-varying feature modulation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[7] C. Steinmetz, S. Singh, M. Comunità, I. Ibnyahya, S. Yuan, E. Benetos, and J. Reiss, "St-ito: Controlling audio effects for style transfer with inference-time optimization," in *International Society for Music Information Retrieval (ISMIR) Conference, 2024*, 2024.

[8] J. Colonel, M. Comunità, and J. Reiss, "Reverse engineering memoryless distortion effects with differentiable waveshapers," in *Audio Engineering Society Convention 153*. Audio Engineering Society, 2022.

Characterizing Jazz Improvisation Style Through Explainable Performer Identification Models

Huw Cheston, Reuben Bance, Peter Harrison

Centre for Music & Science, University of Cambridge, United Kingdom, hwc31@cam.ac.uk

Abstract

Famous music performers tend to have distinctive personal styles. In jazz improvisation, these personal styles affect not only "expressive" musical parameters (timing, dynamics) but also the musical content itself (melody, harmony, rhythm). Machine learning models can be trained to recognise the style of individual performers with a high degree of accuracy [1, 2]. In turn, these models can help reveal the contributions of different musical features towards performance style [1]. To do so, however, their decision-making processes must be explained in a "musicologist friendly" manner [3]. In this paper, we contribute improvements and extensions to previous explainability work in MIR to better understand the decisions made by performer identification models and how these may relate to jazz improvisation style.

We consider a variety of supervised learning architectures and techniques for explaining them, only several of which are reported here. The dataset used to train each model consists of automatically transcribed piano improvisations in MIDI format taken from both Jazz Trio Database (JTD, piano solos, with double bass and drums accompaniment [4]) and Piano Jazz with Automatic MIDI Annotations (PiJAMA, unaccompanied piano [2]). We select performers who appear in both datasets and have at least 80 minutes of material, leaving twenty pianists and a total of 1,629 unique performances (84 hours). These performances are then split into training, validation, and test subsets in the ratio 8:1:1, stratified by source database (JTD or PiJAMA).

Architecture	*Accuracy (track-level)*
Logistic Regression (*n*-grams & voicings)	0.509
CRNN [5]	0.769
ResNet-50 [6]	0.875
This work	**0.913**

Table 1: Held-out test split accuracy for the twenty-class jazz pianist identification task considered in this work.

We begin with testing several "white-box" supervised learning methods, with results for a multinomial logistic regression reported in Table 1. These models are trained on handcrafted feature sets of melodic *n*-grams and chord voicings; *n*-grams are obtained by applying the "skyline" algorithm to the transcription and selecting groups of 2 or 3 contiguous pitch-class intervals, with voicings extracted by grouping near-simultaneous pitches according to onset time and taking triads and tetrads. All models achieved between 40–50% test accuracy using the 16,197 unique *n*-grams and 3,010 voicings obtained from the dataset.

One way in which these models can be explained is by permuting all melody or harmony features and computing the loss in test accuracy. The decrease was ~3x greater when permuting all *n*-grams, which may suggest that melodic patterns are more predictive of jazz performer identity than chord voicings. Additionally, we ascertain individual voicings or *n*-grams that are predictive of particular performers by ranking the odds ratios obtained from binary logistic regression models fitted to predict each class in a one-vs-rest fashion. Many of these features reflect prevailing pedagogical understanding, such as Oscar Peterson being associated with "bluesy" major-minor third shifts.

A second set of experiments tests a novel architecture that learns separately from four fundamental musical domains – melody, harmony, rhythm, and dynamics. Each domain is represented as a piano roll with the same shape as the original transcription, but with all other domains ablated. For example, the melody roll preserves "skyline" pitches and discards timing, while the rhythm roll preserves onset/offset times and randomizes pitch/velocity. Data augmentation (transposition, time dilation, etc.) is applied jointly to all rolls during training. Each domain is modeled separately using a small convolutional network, whose outputs are combined using global average pooling. Random masking is applied to the outputs of up to three domains during training for regularization. The model is trained using thirty-second excerpts from each performance and class probabilities are aggregated to produce track-level predictions.

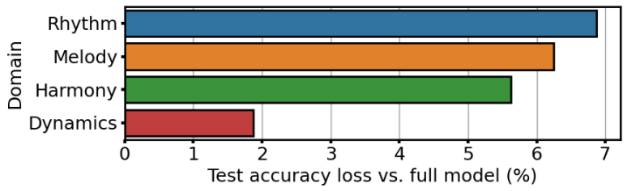

Figure 1: Loss in test accuracy when masking different domains.

Our proposed approach achieves greater classification accuracy than existing music composer identification models [5–6], despite not using the full piano roll as an input (Table 1). During evaluation, we calculate the importance of each musical domain to our model by masking the corresponding embedding prior to pooling and computing the loss in test accuracy. As shown in Figure 1, masking either the melody, harmony, or rhythm domain leads to a similar drop in accuracy (with the greatest loss for rhythm), with a significantly smaller loss when masking dynamics. This reflects previous work emphasizing the importance of microtiming in defining jazz performance style [1].

Taken together, our work extends existing approaches to explainable supervised learning to the task of understanding style in jazz improvisation. Further work could consider, e.g., whether the features considered important for supervised learning models map onto human judgements of jazz improvisation style.

REFERENCES

[1] Cheston, H., et al. (2024). "Rhythmic Qualities of Jazz Improvisation Predict Performer Identity and Style in Source-Separated Audio Recordings", *R. Soc. Open Sci.* 11 (240920).

[2] Edwards, D., et al. (2023). "PiJAMA: piano jazz with automatic midi annotations". *Trans. Int. Soc. Music Inf. Retriv*. 6 (1).

[3] Foscarin, F., et al. (2022). "Concept-Based Techniques for 'Musicologist-Friendly' Explanations in a Deep Music Classifier", in *Proc. 23rd Int. Soc. Mus. Inf. Retriv.*, Bengaluru, India.

[4] Cheston, H., et al. (2024). "Jazz Trio Database: Automated Annotation of Jazz Piano Trio Recordings Processed Using Audio Source Separation", *Trans. Int. Soc. Music Inf. Retriv.* 7 (1).

[5] Kong, Q., et al. (2020). "Large-Scale MIDI-based Composer Classification". http://arxiv.org/abs/2010.14805

[6] Kim, S., et al. (2020). "Deep Composer Classification Using Symbolic Representation", in *ISMIR Late Breaking and Demo Papers*. Montréal, Canada.

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Framework for Predicting Eurovision Song Contest Results

Katarzyna Adamska[1] and Joshua Reiss[2]

[1]C4DM, Queen Mary University of London, United Kingdom, k.m.adamska@qmul.ac.uk
[2]C4DM, Queen Mary University of London, United Kingdom

Abstract

Prior studies on hit song prediction have mostly concentrated on how to predict a song's success in the music charts, overlooking the investigation of song contests such as Eurovision[1], an annual international music competition where artists from primarily European countries perform original songs, and viewers, along with national juries, vote to determine the winner.

The fundamental concept of 'Hit Song Science' assumes that successful songs share a set of musical, contextual, and cultural factors that make them appealing to a general audience [1, 2]. Based on this premise, we propose a framework for predicting the 2024 Eurovision result rankings, focusing first on predicting which songs qualify from the semi-finals to the grand final and then determining the most successful songs in the final.

This study accounts for the contest format, as all participating songs do not compete together in one event. Therefore, the dataset was adjusted into two versions: the semi-finals dataset and the grand final dataset, which included only songs that qualified from the semi-finals and those that prequalified. The semi-finals dataset included 523 songs, while the grand final dataset included 386 songs, from contests between 2008 and 2023.

The methodology centres on a multi-modal perspective, proposing four feature sets: (1) the first with intrinsic characteristics of songs represented as audio and lyrics features, (2) the second adding YouTube daily views as a measure of public appeal, (3) the third incorporating the running order of performances and previous year's country voting results and vote reciprocation, and (4) the fourth excluding audio and lyrics features. Therefore, we try to account for some of the arbitrary aspects of Eurovision results and find the best indicators of competition success. The predictions are evaluated against the actual results and the bookmakers' betting odds [3, 4] as well as through regression error metrics.

Trained using audio, lyrics, and YouTube daily views features, the XGBoost regression model produced rankings for the second semi-final and the grand final that achieved significant ordinal correlation with the actual rankings, with Spearman (ρ) correlation coefficients of 0.64 and 0.57, and Kendall (τ) correlation coefficients of 0.45 and 0.41, respectively. This model also explained 30% of the variance in the Eurovision grand final results and placed the actual top three songs in the 2nd, 3rd, and 7th positions, respectively.

The YouTube daily views feature was the strongest predictor of Eurovision success, while audio and lyrics features alone were insufficient. Features representing previous voting results and vote reciprocation favoured countries with a history of success, accounting for strategic voting but introducing bias that diminishes the natural unpredictability of Eurovision.

[1]Link to the official Eurovision Song Contest website: https://eurovision.tv/

I. ACKNOWLEDGMENTS

This study made use of the MIRoVision dataset [5, 6], which contains metadata, lyrics, contest ranking, and voting data of songs from the Eurovision song contests. The author is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1].

II. REFERENCES

[1] F. Pachet and P. Roy, "Hit song science is not yet a science," in *Proc. of the 9th Int. Society for Music Information Retrieval Conf.*, Philadelphia, USA, 2008, pp. 335–360.

[2] D. B. Seufitelli, G. P. Oliveira, M. O. Silva, C. Scofield, and M. M. Moro, "Hit song science: A comprehensive survey and research directions," *Journal of New Music Research*, vol. 52, no. 1, p. 41–72, 01 2023.

[3] D. Demergis, "Predicting eurovision song contest results by interpreting the tweets of eurovision fans," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Granada, Spain, 2019, pp. 521–28. [Online]. Available: https://doi.org/10.1109/SNAMS.2019.8931875.

[4] I. Kumpulainen, E. Praks, T. Korhonen, A. Ni, V. Rissanen, and J. Vankka, "Predicting eurovision song contest results using sentiment analysis," in *Artificial Intelligence and Natural Language*, ser. Communications in Computer and Information Science. Cham: Springer International Publishing, 2020, vol. 1292, pp. 87–108. [Online]. Available: https://doi.org/10.1007/978-3-030-59082-6_7

[5] J. Spijkervet, "The Eurovision Dataset," https://zenodo.org/badge/latestdoi/214236225, mar 2020.

[6] J. A. Burgoyne, J. Spijkervet, and D. J. Baker, "Measuring the Eurovision Song Contest: A living dataset for real-world MIR," in *Proceedings of the 24th International Society for Music Information Retrieval Conference*, Milan, Italy, 2023. [Online]. Available: https://archives.ismir.net/ismir2023/paper/000097.pdf

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

MINDS - Mutual Inclusion through Neurodiversity in Science

Daniel Gill

School of Electronic Engineering and Computer Science, Queen Mary University of London, UK, d.a.gill@qmul.ac.uk

I. ABSTRACT

Through this submission, I hope to create a beacon and prompt for an enhanced discussion on the importance and value of the thoughtful inclusion of neurodivergent people in research in digital music and other fields. Whether in the form of a poster during networking sessions or an interactive talk, I would like to use this opportunity to bring together researchers interested in working with neurodivergent people to share their own experiences and best practices. As a neurodivergent PhD student researching in human-centred design, it was a priority of mine to set up the MI**ND**S (Mutual Inclusion through Neurodiversity in Science) network[1] to connect neurodivergent people, researchers, and neurodivergent researchers, which some DMRN members may be interested in.

Through digital music, autistic people can be provided with a therapeutic technique to manage wellbeing and social interaction [1, 2]; individuals with ADHD may benefit from timing perception and regulation training through rhythm-based music integration in serious video games [3]; and those with Tourette's syndrome may experience a reduction in tics [4]. But, as noted eloquently by Dobesh *et al.* [5], one of the main goals of such research is "to reduce stereotyped behavior". Dobesh *et al.* were talking about music therapy for autistic people, but the sentiment holds for many of the papers discussed in the reviews above.

While there is little doubt that many neurodivergent people would appreciate help with managing their experiences, many technological solutions take the approach of trying to fix or cure a symptom [6] - particularly putting the emphasis on the neurodivergent individual to change. However, when neurodivergent people are involved in the process from the beginning, in line with Le's [7] tenets, the final result can be greatly improved.

MI**ND**S hopes to give practical support to researchers to include neurodivergent people in their research, as well as communicating relevant work to this group (an important first step in rebuilding trust in research practices [8]). Through a talk/poster, I hope to share the importance of the inclusion of neurodivergent people in research that affects them, as well as to hear how colleagues have put this in practice in their own work for others to learn from.

[1]More information about MI**ND**S is available on our website: https://mindsintech.org.uk/

II. ACKNOWLEDGMENT

The author would like to thank Ekaterina Ivanova and Tony Stockman for their support.

III. REFERENCES

[1] G. Ragone, J. Good, and K. Howland, "How Technology Applied to Music-Therapy and Sound-Based Activities Addresses Motor and Social Skills in Autistic Children," *Multimodal Technologies and Interaction*, vol. 5, no. 3, p. 11, Mar. 2021, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2414-4088/5/3/11

[2] D. Johnston, H. Egermann, and G. Kearney, "Innovative computer technology in music-based interventions for individuals with autism moving beyond traditional interactive music therapy techniques," *Cogent Psychology*, vol. 5, no. 1, p. 1554773, Dec. 2018, publisher: Cogent OA. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/23311908.2018.1554773

[3] M. Martin-Moratinos, M. Bella-Fernández, and H. Blasco-Fontecilla, "Effects of Music on Attention-Deficit/Hyperactivity Disorder (ADHD) and Potential Application in Serious Video Games: Systematic Review," *Journal of Medical Internet Research*, vol. 25, no. 1, p. e37742, May 2023, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://www.jmir.org/2023/1/e37742

[4] S. Scataglini, G. Andreoni, M. Fusca, and M. Porta, "Effect of Rhythmic Music Auditory Stimulation On Tics Modulation in Tourette Syndrome: A Case Study," *Open access Journal of Neurology & Neurosurgery*, vol. 5, Aug. 2017.

[5] S. Dobesh, J. Albert, S. Ahmed, and M. Sharmin, "Moving Towards an Accessible Approach to Music Therapy for Autistic People: A Systematic Review," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, June 2023, pp. 472–480, iSSN: 0730-3157. [Online]. Available: https://ieeexplore.ieee.org/document/10196942

[6] K. Spiel and K. Gerling, "The Purpose of Play: How HCI Games Research Fails Neurodivergent Populations," *ACM Transactions on Computer-Human Interaction*, vol. 28, no. 2, pp. 1–40, Apr. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3432245

[7] L. Le, ""I Am Human, Just Like You": What Intersectional, Neurodivergent Lived Experiences Bring to Accessibility Research," Aug. 2024, arXiv:2408.04500 [cs]. [Online]. Available: http://arxiv.org/abs/2408.04500

[8] S. Fletcher-Watson, K. Brook, S. Hallett, F. Murray, and C. J. Crompton, "Inclusive Practices for Neurodevelopmental Research," *Current Developmental Disorders Reports*, vol. 8, no. 2, pp. 88–97, June 2021. [Online]. Available: https://doi.org/10.1007/s40474-021-00227-z

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Improving Automatic Guitar Tablature Transcription with LLMs

Omar Ahmed[1,2], Pedro Sarmento[2] and Emmanouil Benetos[2]
[1]University of Oxford, UK, omar.ahmed@pmb.ox.ac.uk
[2]Centre for Digital Music, Queen Mary University of London, UK, {p.sarmento, emmanouil.benetos}@qmul.ac.uk

***Abstract*— This work-in-progress demonstrates the usability of large language models (LLMs) in correcting the outputs of guitar tablature transcription models. In our methodology, we first convert the output of existing automatic guitar tablature transcription models into text tokens, leveraging the format proposed in the DadaGP dataset, and prompt a fine-tuned LLM to adjust erroneous tokens with respect to fingerings typically used by guitarists. We showcase the feasibility of our approach with examples.**

***Index Terms*—** Guitar, Tablature, LLM, Automatic Music Transcription;

I. RELATED WORK

Guitar tablature is the dominant form of notation for guitarists. The layout of the guitar means that notes of the same pitch can be played in multiple locations, so transcription into tablature format requires not only pitch estimation but also fingering arrangement. As a result, one song may have multiple valid transcriptions, but accurately predicting the ground truth is a much more difficult task. A common model for baseline testing is TabCNN, introduced by Wiggins and Kim [1]. One of the key recurring errors made by this model is that of a *miss-fretting*, in which the model fails to choose the correct string based on timbral features, whilst still identifying the correct note. We posit that this type of error can be tackled by prompting an LLM to maintain fingering consistency. Typically, guitarists tend to prefer fingerings that minimise the distance between a set of notes. TabCNN does not take this information into account, rather approaching the problem by viewing each frame of audio independently. We postulate that a rule-based approach is less versatile due to fringe circumstances that may require unique fingerings that can be learned by an LLM.

II. PROPOSED METHODOLOGY

The proposed methodology is inspired by the findings from LyricWhiz [2]. Here, current lyrics transcription technology is improved by first using Whisper as an 'ear' to transcribe audio followed by GPT-4 as a 'brain' that can syntactically correct the Whisper output. Similarly, we categorise TabCNN as an 'ear' and use GPT-4o as a 'brain' that can capture the thought process employed by guitarists when playing. This aims to take into account the fact that guitarists choose easier fingering arrangements, therefore addressing the miss-fretting error, which results in impractical hand positions. However, in the case of guitar transcription, the model outputs must first be expressed as text before being inputted into a GPT-4o model. We use the text token notation introduced in the DadaGP dataset [3], mostly used for generative purposes [4].

III. EXAMPLE USE

The feasibility of this methodology is expressed through an example use. An artificial miss-fretting error is created by editing a snippet from DadaGP, as shown in Figure 1.

LLM input (w/ missfretting)	**LLM output (corrected)**
artist:unknown_artist	artist:unknown_artist
downtune:0	downtune:0
tempo:120	tempo:120
start	start
new_measure	new_measure
clean0:note:s6:f5	clean0:note:s6:f5
nfx:tie	nfx:tie
wait:240	wait:240
clean0:note:s3:f5	clean0:note:s3:f5
nfx:let_ring	nfx:let_ring
wait:240	wait:240
clean0:note:s2:f2	clean0:note:s3:f6
nfx:let_ring	nfx:let_ring
wait:240	wait:240

Figure 1: The LLM input, with the missfretting in bold, and the LLM corrected output.

Note that the edited transcription is still valid, playing the exact same pitch whilst straying from the true transcription. Even without fine-tuning, GPT-4o was able to identify the error and correct the transcription; additionally, GPT-4o cited the "desire for easy transitions" as the reason behind the correction. We provide more examples[1] for further inspection.

IV. REFERENCES

[1] Wiggins, A. and Kim, Y., "Guitar Tablature Estimation with a Convolutional Neural Network," *ISMIR*, 2019.

[2] Zhuo, L., Yuan, R., Pan, J., Ma, Y., Li, Y., Zhang, G., Liu, S., Dannenberg, R., Fu, J., Lin, C., Benetos, E., Xue, W. and Guo, Y., "Lyricwhiz: Robust Multilingual Zero-Shot Lyrics Transcription by Whispering to ChatGPT," *ISMIR*, 2023.

[3] Sarmento, P., Kumar, A., Carr, CJ, Zukowski, Z., Barthet, M. and Yang, Y., "DadaGP: A Dataset of Tokenized GuitarPro Songs for Sequence Models," *ISMIR*, 2021.

[4] Sarmento, P., Kumar, A., Xie, D., Carr, CJ, Zukowski, Z. and Barthet, M., "ShredGP: Guitarist Style-Conditioned Tablature Generation," *CMMR*, 2023.

[1]Available at: `https://shorturl.at/6kX4c`

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Towards Detecting Interleaved Voices in Telemann Flute Fantasias

Patrice Thibaud, Mathieu Giraud and Yann Teytaut

Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France, patrice.thibaud@univ-lille.fr

***Abstract*— The *12 fantasias* by Telemann are one a the most representative works in the solo flute repertoire. We gathered musicological annotations and labeled the harmonies and the structures. We propose an implementation of a perceptual model able to detect changes of voice within a monophonic line. The algorithm detects interleaved voices with an F_1-score of 70%.**

***Index Terms*—** symbolic music representation, implied polyphony, voices detection, flute

I. MONOPHONIC, REALLY?

The *12 fantasias for solo flute* by Georg Philipp Telemann (*TWV 40:2-13*) are quintessential examples of the Baroque era. Each fantasia is a free-form composition, characterized by a sequence of themes evoking improvisation, with some passages featuring implied polyphony through interweaving of two or more melodic lines. Telemann achieves what W. Piston describes as a "compound melody"[1]. Such voices may be notated by a beamed group of notes where some stems point downward and others upward [2] (Fig. 1).

Figure 1: The *Vivace* from Fantasia n°2 (TWV 40:3) starts with two interleaved voices, here being two themes of a double fugue [2]. Telemann sometimes underlines these voices using stems pointing in opposite directions, as here in the bars 2 and 3.

II. THE CORPUS

We started from MusicXML files from *IMSLP* (International Music Score Library Project)[1], and amended these transcriptions by carefully checking the only authentificated copy of the fantasias (Royal Library of Brussels, T 5823). Collecting upon 8 musicological sources such as [2] or [3], we gathered annotations about local keys, movements names (missing from the original), implied chords and voices annotations on musical sequences containing interleaved voices – they account for 29.7% of the notes of the score. We labeled these elements with the Dezrann open-source platform [4].

[1] https://imslp.org/wiki/Special:ReverseLookup/236786 by Árpád Zoltán Szabó (Creative Commons Attribution-ShareAlike 4.0 License)

III. INTERLEAVED VOICES

Whereas voice separation in polyphonic voices is well studied [5, 6], it is not the case for implied polyphony. Stacey Davis [7] proposed a model based on music perception elements to identify where a voice change occurs. Three criteria are applied to analyze all intervals in a piece: (*i*) the size of the diatonic interval; (*ii*) the study of the contour changes (ascending and descending voices around the interval); and (*iii*) the movement of joint notes before and after.

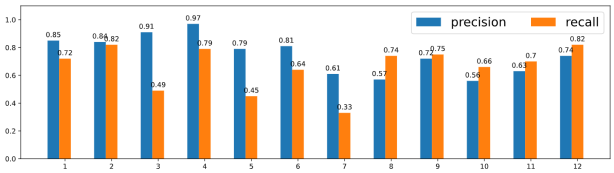

Figure 2: Precision and recall on voice prediction on all the 12 fantasias

We implemented this model and evaluated its output against the existing annotations used as reference on all the passages with interleaved voices (Fig. 2). Most voice changes are correctly predicted (global precision: 78%, recall: 64%, F_1-score: 70%). Note that the most challenging cases are when voices are close.

IV. PERSPECTIVES

These results, with a pure deterministic algorithm, could be improved by taking account of the pedal passages not always detected due to the size of intervals or, of the repetitions of certain patterns within a same voice. Further work could also study how performers may highlight these changes.

V. REFERENCES

[1] W. Piston, "Counterpoint", 1947

[2] P. da Silva, A. Carlos, "A performance guide to three of Telemann's 12 fantasias for flute", Ph.D. Dissertation, University of Alabama, 2012

[3] S. Eppinger, "Georg Philipp Telemann : 12 Fantasien für Flöte solo, Teil I", *TIBIA*, vol. 2, pp. 86–99, 1984

[4] M. Giraud, R. Groult, and E. Leguy, "Dezrann, a Web Framework to Share Music Analysis", TENOR 2018

[5] E. Chew and X. Wu, "Separating Voices in Polyphonic Music: A Contig Mapping Approach", CMMR 2004

[6] E. Cambouropoulos, "Voice And Stream: Perceptual And Computational Modeling Of Voice Separation", *Music Perception*, 2008

[7] S. Davis, "Implied Polyphony in the Solo String Works of J. S. Bach: A Case for the Perceptual Relevance of Structural Expression", *Music Perception*, 2006

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Embodied Movement-Sound Interaction in Latent Terrain Synthesis

Shuoyang Zheng*[1], Anna Xambó Sedó[2] and Nick Bryan-Kinns[3]

[1]Centre for Digital Music, Queen Mary University of London, UK, shuoyang.zheng@qmul.ac.uk
[2]Centre for Digital Music, Queen Mary University of London, UK
[3]Creative Computing Institute, University of the Arts London, UK

***Abstract*— We present *Latent Terrain Synthesis*, an algorithmic strategy for unfolding a neural audio synthesis model's high-dimensional latent space into a mountainous two-dimensional plane, as a research prototype to investigate how musicians perceive and respond to affordances in movement-based interaction with neural synthesis, and how they bundle affordances into embodied performance techniques.**

***Index Terms*—** Neural Audio Synthesis, Human-Computer Interaction, Movement-Sound Interaction

I. EMBODIED PERSPECTIVE ON NEURAL AUDIO SYNTHESIS

Various attempts have been made to investigate musical affordances provided by neural audio synthesis [1], and various ways of harnessing these affordances into musical practices have been found [2]. However, how neural audio synthesis can be appropriated as an embodied musical instrument is still underexplored. Existing literature lacks details on how musicians perceive affordances in movement-based interaction, and how they develop embodied movement-based strategies and techniques for performance.

To approach this gap, we designed a digital musical instrument with a stylus and tablet interface as a research prototype. It embeds *Latent Terrain*, a sonic material adapted from a neural audio synthesis model that one can use any line-drawing movement to interact with by a stylus, inspired by wave terrain synthesis [3]. We present a workshop in which 12 participants explored two latent terrains and actively tested out affordances based on their capabilities and curiosities. We studied how participants' embodied capacities shaped their perceived affordances and how these affordances are bundled into diverse performance techniques.

II. LATENT TERRAIN SYNTHESIS

We present *Latent Terrain Synthesis*, an algorithmic strategy for unfolding a latent variable neural audio synthesis model's latent space into a mountainous two-dimensional plane. The unfolded structure is a streamable "terrain", which can be used to access latent vectors of the synthesis model given coordinates on the 2D plane. Different from a typical XY pad surface, with a mountainous and steeper latent terrain, one can create sound pieces with greater spectral complexity by tracing a linear path through it.

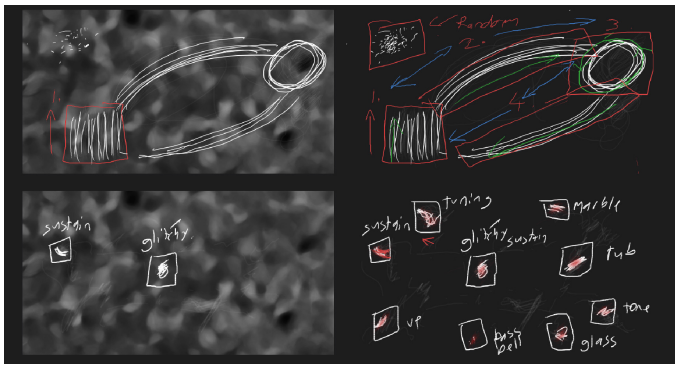

Figure 1: Scores created by musicians, superimposed on the latent terrains.

Various types of controls can be used to navigate a latent terrain. As a probe for our research question, we embedded it into a simple movement-based interface: a tablet and a stylus. We used the Realtime Audio Variational autoEncoder (RAVE) [4] due to its capacity for real-time continuous responses. In addition, we offer our MaxMSP external *nn_stylus* [5] that works together with IRCAM's *nn_tilde* [6] to generate latent terrains for pre-trained RAVE models and allow users to navigate the terrain.

III. REFERENCES

[1] M. Yee-King, "Latent Spaces: A Creative Approach," in *The Language of Creative AI: Practices, Aesthetics and Structures*. Cham: Springer International Publishing, 2022, pp. 137–154.

[2] S. Zheng, A. Xambó, and N. Bryan-Kinns, "A Mapping Strategy for Interacting with Latent Audio Synthesis Using Artistic Materials," in *Proceedings of The second international workshop on eXplainable AI for the Arts (XAIxArts)*, 2024.

[3] Y. Mitsuhashi, "Audio Signal Synthesis by Functions of Two Variables," *Journal of the Audio Engineering Society*, vol. 30, no. 10, pp. 701–706, Oct. 1982. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=3815

[4] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," Dec. 2021, arXiv:2111.05011 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2111.05011

[5] https://github.com/jasper-zheng/nn_stylus.

[6] https://github.com/acids-ircam/nn_tilde.

*SZ is a research students at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1]

Towards Differentiable Digital Waveguide Synthesis

Pablo Tablas de Paula and Joshua D. Reiss

Centre for Digital Music (C4DM), Queen Mary University Of London, United Kingdom,
p.tablasdepaula@qmul.ac.uk

Abstract

This presentation explores the integration of Differentiable Digital Signal Processing (DDSP) techniques into digital waveguide synthesis (DWS), proposing a novel approach termed Differentiable Digital Waveguide Synthesis (DDWS). DWS revolutionized physical modeling by efficiently producing rich string-like timbres using a delay line and a filter, as first demonstrated in the Karplus-Strong algorithm [1]. Over the past 40 years, DWS has developed a rich legacy with highly specialized waveguide networks for specific musical instruments [2], reverberation [3], and speech processing [4].

Recent advancements in neural audio synthesis, particularly the development of DDSP techniques, have significantly improved the sound quality, interpretability, and training efficiency of synthesis models by incorporating established principles from digital signal processing (DSP). Initially, DDSP integrated basic DSP components—such as oscillators tuned to produce harmonically related frequencies—into neural networks. This allowed models to optimize signal processing parameters through gradient-based learning, effectively combining traditional DSP with modern machine learning techniques [5]. Since then, numerous DSP methods have been incorporated into DDSP frameworks. By applying DDSP to digital waveguide synthesis, we can further enhance this integration by leveraging the highly specialized waveguide networks optimized for specific instruments. This approach builds upon decades of research, utilizing the extensive prior knowledge embedded in highly-specialized DWS models to achieve more accurate and expressive sound synthesis with smaller datasets and training time.

Notably, this technique has already been successfully applied to articulatory synthesis using the Kelly-Lochbaum model, as demonstrated by Südholt et al. [6]. This work exemplifies how highly biased models, such as digital waveguides, can be easily optimized using DDSP techniques. This integration presents opportunities to further develop applications in artificial reverberation by optimizing waveguide meshes [7], waveguide webs [3], and scattering delay networks [8], where DDWS could develop into interactive reverberation modeling for procedural audio settings. Additionally, applying DDWS to instrument modeling could enhance expressivity and significantly reduce memory expenses in sample libraries; by using the sample library as the training dataset, the physical accuracy and efficiency of DWS can create dynamic and expressive instruments without relying on large sets of recordings with often, limited expressivity.

ABBREVIATIONS AND ACRONYMS

DDSP: Differentiable Digital Signal Processing
DWS: Digital Waveguide Synthesis
DDWS: Differentiable Digital Waveguide Synthesis

ACKNOWLEDGMENT

I extend sincere thanks to my supervisor at C4DM, Joshua Reiss, Queen Mary University of London, for invaluable guidance and support in this research. Appreciation is also due to colleagues who contributed to discussions on digital waveguide synthesis and integrating physical modeling techniques into differentiable digital signal processing. Gratitude is extended to the research community whose foundational work has paved the way for this project.

REFERENCES

1. Karplus, K. and Strong, A., 1983. Digital synthesis of plucked-string and drum timbres. Computer Music Journal, 7(2), pp.43-55.
2. Välimäki, V., Laurson, M. and Erkut, C., 2003. Commuted waveguide synthesis of the clavichord. *Computer Music Journal*, *27*(1), pp.71-82.
3. Stevens, F., Murphy, D.T., Savioja, L. and Välimäki, V., 2017. Modeling sparsely reflecting outdoor acoustic scenes using the waveguide web. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(8), pp.1566-1578.
4. Speed, M., Murphy, D. and Howard, D., 2013. Modeling the vocal tract transfer function using a 3D digital waveguide mesh. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(2), pp.453-464.
5. Engel, J., Gu, C. and Roberts, A., 2020, DDSP: Differentiable Digital Signal Processing. In International Conference on Learning Representations.
6. Südholt, D., Cámara, M., Xu, Z., and Reiss, J.D. (2023) 'Vocal tract area estimation by gradient descent', Proceedings of the 26th International Conference on Digital Audio Effects (DAFx23), Copenhagen, Denmark, 4–7 September 2023.
7. Murphy, D., Kelloniemi, A., Mullen, J. and Shelley, S., 2007. Acoustic modeling using the digital waveguide mesh. IEEE Signal Processing Magazine, 24(2), pp.55-66.
8. De Sena, E., Hacıhabiboğlu, H., Cvetković, Z. and Smith, J.O., 2015. Efficient synthesis of room acoustics via scattering delay networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(9), pp.1478-1492.

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Classification of Spontaneous and Scripted Speech for Multilingual Audio

Shahar Elisha[1,2], Andrew McDowell[1], Mariano Beguerisse-Díaz[1], Emmanouil Benetos[2]

[1]Spotify Ltd
[2]Centre for Digital Music, Queen Mary University of London

***Abstract*— Distinguishing scripted from spontaneous speech is an essential tool for better understanding how speech styles influence speech processing research. It can also improve recommendation systems and discovery experiences for media users through better segmentation of large recorded speech catalogues. This paper addresses the challenge of building a classifier that generalises well across different formats and languages. We systematically evaluate models ranging from traditional, handcrafted acoustic and prosodic features to advanced audio transformers, utilising a large, multilingual proprietary podcast dataset for training and validation. We break down the performance of each model across 11 language groups to evaluate cross-lingual biases. Our experimental analysis extends to publicly available datasets to assess the models' generalisability to non-podcast domains. Our results indicate that transformer-based models consistently outperform traditional feature-based techniques, achieving state-of-the-art performance in distinguishing between scripted and spontaneous speech across various languages.**

***Index Terms*—** spontaneous/scripted speech classification, speech paralinguistics, audio processing, podcasts, multilingual

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Personalising equalisation using psychological and contextual factors

Yorgos Velissaridis, Charalampos Saitis and György Fazekas

EECS/C4DM, Queen Mary University of London, UK, g.velissaridis@qmul.ac.uk

***Abstract*— This research investigates how EQ settings interact with contextual and individual factors, such as personality and mood, to shape listener experiences. Through psychological experiments and crowdsourcing for semantic descriptors, a dataset will be created for training an AI model to predict optimal EQ settings. The overall project aim is to establish a unified framework for EQ personalization, leading to the development of an adaptive, real-time EQ system that benefits listeners.**

***Index Terms*—** Personalisation, Audio Equalisation, Music Psychology, Contextual Factors, Listener Experience

I. BACKGROUND

Audio equalization (EQ) plays an important role in shaping listeners' emotional responses, particularly in genres like electronic music. McCown et al. [1] have demonstrated a link between extraversion and psychoticism and a preference for bass-heavy music. More recently, Dourou et al. [2] found that mood affects EQ preferences, with low-arousal listeners favouring less high-frequency boosting compared to high-arousal listeners.

II. RESEARCH MOTIVATION AND GOALS

No research currently examines the combined influence of personality, mood, and psychoacoustic features of audio on EQ preferences. This research aims to address this gap by investigating the hypothesis that EQ settings interact with a range of factors—musical, contextual, and individual—to shape the listener's experience.

These research questions guide the investigation:

> *RQ1: What is the relative contribution of each factor to the listener's experience?*
> *RQ2: How can adjustments to EQ settings enhance the enjoyment of audio across media?*

Research aim: The research aims to establish a unified framework to explain and predict the relationship between EQ settings and listener enjoyment, leveraging this framework to design and refine a real-time adaptive EQ system.

III. METHODS

Psychological experiments: Psychological experiments will support theory development and model training. Participants will report their personalities, moods, and musical preferences, then select preferred EQ settings for music excerpts, as in Dourou and colleagues [2], with mood monitored throughout.

Crowdsourcing: To gather extensive data for model training and to inform interface design, semantic descriptors for EQ settings will be crowdsourced, inspired by Cartwright and Pardo [3], as well as using gamification techniques.

Model Development: The model will fuse the semantic descriptors, processed by a pre-trained multimodal model, with personality, mood, contextual data and music embeddings into a unified model. Other possibilities include utilising interpretable knowledge graphs implemented with graph neural networks and cognitive models such as MicroPsi [4] to dynamically adapt EQ settings based on mood changes.

IV. EXTENSIONS

Future studies could expand the proposed system to enable users not only to match their emotional state but also to explore a wider range of emotions through customized EQ adjustments.

V. ACKNOWLEDGMENTS

The first author is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Yamaha.

VI. REFERENCES

[1] W. McCown, R. Keiser, S. Mulhearn, and D. Williamson, "The role of personality and gender in preference for exaggerated bass in music," *Personality and Individual Differences*, vol. 23, no. 4, pp. 543–547, Oct. 1997.

[2] N. Dourou, V. Bruschi, S. Spinsante, and S. Cecchi, "The Influence of Listeners' Mood on Equalization-Based Listening Experience," *Acoustics*, vol. 4, no. 3, pp. 746–763, Sept. 2022.

[3] M. Cartwright and B. Pardo, "Social-eq: Crowdsourcing an equalization descriptor map." in *ISMIR*, 2013, pp. 395–400.

[4] J. Bach, "A Framework for Emergent Emotions, Based on Motivation and Cognitive Modulators," *International Journal of Synthetic Emotions*, vol. 3, no. 1, pp. 43–63, Jan. 2012.

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

Can downbeat trackers predict hypermetre?

Jose Alejandro Esquivel de Jesus [1] and Jordan B. L. Smith [2]

[1]Queen Mary University of London, UK, j.esquiveldejesus@qmul.ac.uk
[2]Queen Mary University of London, UK

***Abstract*— Hypermetre is the metric structure between the timescales of downbeats and sections in a piece of music. Hypermetre tracking systems, in principle, face the same difficulties posed by beat and downbeat detection. In this work, we evaluate the success of piggybacking downbeat tracking methods to perform hypermetre tracking. We aim to demonstrate their limitations and the need for further research.**

I. INTRODUCTION

If downbeats are metrically accented beats, then hyperdownbeats are metrically accented downbeats, and each one defines the start of a hypermeasure [1]. Beat and downbeat tracking have been widely researched. Machine learning methods have been used to perform beat and downbeat tracking [2, 3], as well as to predict large-scale structure in songs [4]. Some methods address both timescales in the same model (e.g., [5]), but we are not aware of any that models the in-between timescale of hypermeasures. However, doing it seems worthwhile because it would provide us with improved regularisation of structure analysis that would benefit MIR tasks (e.g., segmentation, AI music generation, and recommendation systems).

II. DATA

We are not aware of any datasets of hypermetre annotations. However, in the McGill Billboard dataset [6] of chord labels, the line breaks in the annotation files appear to indicate the hypermetre. We manually verified this for a set of 26 songs. We introduce the term *Hyper-time-signature* as a way of defining the number of measures to a hypermeasure, similar to the time signature. One common assumption of downbeat trackers is that the time signature remains unaltered in the song [7]. From the annotations in the Billboard dataset, 0.7% of 890 songs exhibit a change in time signature. This value contrasts with 44% of the songs that appear to have a change in hyper-time signature, suggesting a need for further research on hypermetre tracking systems that account for hypermeasure metric changes, given they seem to occur at much higher frequency than measure metric changes.

III. METHOD AND RESULTS

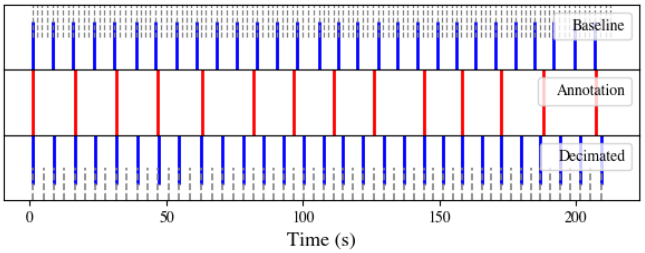

Figure 1: "The Rose" by Bette Middler (McGill Billboard song 0006), as annotated and as predicted by two methods.

We performed an experiment using two methods: the *baseline* and the *decimated*. The *baseline* method serves as a simple baseline for performance floor. We assume 4/4 for the time signature and the hypermeasure-time-signature. This method estimates the beats using spectral flux. Then, downbeats and hyperdownbeats are estimated, assuming they occur every four beats and four downbeats, respectively. The *decimated* method is a first attempt to use an existing beat tracker (madmom) to perform hypermetre tracking. The method uses the madmom downbeattracker to estimate the downbeats locations; then the downbeats are extracted to a new audio file by concatenation and then madmom is re-run on this decimated version to estimate the hyperdownbeats.

In Fig. 1, we can see the results of these methods applied to a pop song. The middle red lines are the human annotations, the blue lines are the hyper-downbeats estimates, and the gray lines are the downbeat estimates for each method. For the *baseline* method, we can observe a misalignment between estimated and annotated times; estimated occur a beat before annotated for the first four red intervals. The fifth red interval is wider than the previous ones because it has a hypermeasure of five measures instead of four. For the *decimated* method, we can see alignment between estimated and annotated hyperdownbeats for the first four red intervals, but this is lost due to the wider fifth red interval.

From these observations, we conclude that hypermetre tracking cannot be reliably performed by only considering beat and downbeat information and that to improve hypermetre tracking, it is necessary to approach hypermetre by accounting for other music characteristics (e.g., harmonic and instrumentation content).

IV. References

[1] H. Krebs, "Hypermeter and Hypermetric Irregularity in the Songs of Josephine Lang." *Engaging Music: Essays in Music Analysis*, 2005.

[2] J. Zhao, G. Xia, and Y. Wang, "Beat Transformer: Demixed Beat and Downbeat Tracking with Dilated Self-Attention," 9 2022. [Online]. Available: http://arxiv.org/abs/2209.07140

[3] T. Kim and J. Nam, "All-in-One Metrical and Functional Structure Analysis with Neighborhood Attentions on Demixed Audio," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, vol. 2023-October. Institute of Electrical and Electronics Engineers Inc., 2023.

[4] O. Nieto, G. J. Mysore, C. I. Wang, J. B. Smith, J. Schluter, T. Grill, and B. McFee, "Audio-Based Music Structure Analysis: Current Trends, Open Challenges, and Applications," pp. 246–263, 2020.

[5] M. Fuentes, B. McFee, H. C. Crayencour, S. Essid, and J. P. Bello, "A Music Structure Informed Downbeat Tracking System Using Skip-chain Conditional Random Fields and Deep Learning," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5 2019, pp. 481–485. [Online]. Available: https://ieeexplore.ieee.org/document/8682870/

[6] J. A. Burgoyne, J. Wild, and I. Fujinaga, "An expert ground truth set for audio chord recognition and music analysis," in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Miami, FL, 2012, pp. 633–638.

[7] F. Foscarin, J. Schlüter, and G. Widmer, "Beat this! Accurate beat tracking without DBN postprocessing," 7 2024. [Online]. Available: http://arxiv.org/abs/2407.21658

DMRN+19: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2024

QUEEN MARY UNIVERSITY OF LONDON
TUE 17 DECEMBER 2024

A Transposition-Invariant Chord Encoder for Bigram Modelling

Yuqiang Li[1] and György Fazekas[2]

[1]EECS, Queen Mary University of London, UK, ec24093@qmul.ac.uk
[2]EECS, Queen Mary University of London, UK

***Abstract*— For deep learning models, the learnt representations for chords often generalise poorly due to sensitivity to the absolute pitch distribution, and over-simplified chord labels in the training data. Without guaranteed transpositional invariance to the input, the harmonic movement derived from the latent space can also be inconsistent. This study proposes a transposition-invariant neural network encoder for chord bigrams, compared against two baseline models, in the task of reconstructing transposed versions of the input chord bigram. The proposed method, which discards the absolute pitches in the input chord bigrams, showed significantly improved reconstruction accuracy and training speed.**

***Index Terms*—** Symbolic Music Representation, Chord Representation, Harmony Modelling

I. METHOD

We represent a chord $\mathbf{c}$ as a 24-d vector, a concatenation of the pitch class vectors $\mathbf{p} \in \{0,1\}^{12}$ and the one-hot bass note vector $\mathbf{b} \in \{0,1\}^{12}$. The bigram is represented by two chords concatenated together, $[\mathbf{c}_1 \oplus \mathbf{c}_2]$, of 48 dimensions.

The *rotation* operation $R_k(\mathbf{c})$ is defined as a cyclic permutation of the chord vector $\mathbf{c}$ by shifting k elements (semitones) to the right: $R_k(\mathbf{c}) := [R_k(\mathbf{p}) \oplus R_k(\mathbf{b})]$. A rotation applied to a chord bigram is of course chord-wise.

Task definition: For the input chord bigram $\mathbf{c}_1$ and $\mathbf{c}_2$, the model should encode the harmonic movement from $\mathbf{c}_1$ to $\mathbf{c}_2$ into a latent variable z, and then to predict the randomly rotated $R_k(\mathbf{c}_2)$ from z, hinted by the query $R_k(\mathbf{c}_1)$. The loss function is the sum of the binary cross-entropy for the $\mathbf{p}$ (pitch class vector) and a cross-entropy loss for the $\mathbf{b}$ (bass note). Our proposed method to address this task is compared to two simple baselines.

Baseline 1: The model encodes the input chord bigram $[\mathbf{c}_1 \oplus \mathbf{c}_2]$ into a latent z, and decode the target chord $R_k(\mathbf{c}_2)$ from the concatenation of the query chord $R_k(\mathbf{c}_1)$ and the latent z. For data augmentation, the input chord bigram is randomly rotated at runtime.

Baseline 2 (Regularization): To encourage the model to encode different rotations of the same input into similar vectors, a regularization term is added to Baseline 1, implemented as the variance of all the encoded vectors $\mathrm{Var}_{k\in\mathbb{Z}_{12}}\left[z^{(k)}\right]$ of all the 12 rotations of the input bigram.

Proposed method: The proposed method encodes all the 12 rotations of the input chord bigram $[R_k(\mathbf{c}_1, \mathbf{c}_2)]$ into 12 latent vectors $\{z_0, z_1, \ldots, z_{11}\}$ and takes the sum as the final latent vector z. This ensures that the latent representation is invariant to transpositions of the input. The (input) augmentation used in the two baseline systems is not needed here.

II. RESULTS

The three models have the exact same trainable parameters: encoder and decoder dimension being 30, and latent dimension being 5. They were trained using 80% of the 71,086 chord bigrams of the iReal Pro jazz playlist, optimised using Adam with a learning rate of 0.003, batch size of 256, until the loss converged. No overfitting was observed. See Table 1.

Model	**Epochs**	**Loss$_{\text{Pitch}}$**	**Loss$_{\text{Bass}}$**
Baseline 1	400	0.082	0.010
Baseline 2	240	0.090	0.005
Proposed	**100**	**0.056**	**0.004**

Table 1: Test Losses on the Pitch Class and Bass Prediction

III. DISCUSSION AND CONCLUSION

The proposed method eliminated the absolute pitch classes, and forces a consistent latent representation for the same harmonic movement, which noticeably sped up the training process. The regularization, despite encouraging consistent representations, showed limited improvement.

For the latent space of chord bigram, we anticipate its use as a prior, e.g. to approximate perceived harmonic similarity, to make corrections on chord recognition algorithms, to model second-order harmonic trajectory, or to extract harmonic features for musical style analysis [1].

Given the fairly small model size, in future work, it may be possible to derive the closed form for the encoder (e.g. circular convolution), or find the exact manifold of the latent space using topological data analysis methods.

IV. REFERENCES

[1] M. Müller, C. Weiß, and F. Brand, "Mid-Level Chord Transition Features for Musical Style Analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 2019.