DMRN+18: Digital Music Research Network

One-day Workshop 2023

Queen Mary University of London

Tuesday 19th December 2023

Chair: Simon Dixon

Programme

10:00	**Welcome - Simon Dixon**
10:10	**KEYNOTE** Physics-based Audio: Sound Synthesis and Virtual Acoustics, **Stefan Bilbao- (Acoustics and Audio Group - University of Edinburgh)**
11:10	*Break (Coffee break)*
11:30	"In-depth performance analysis of the state-of-the-art algorithm for automatic drum transcription", **Mickaël Zehren (Umea Universitet, Sweden), Marco Alunno (Universidad EAFIT, Colombia) and Paolo Bientinesi (Umea Universitet, Sweden)**
11:50	"Automatic Guitar Transcription with a Composable Audio-to-MIDI-to-Tablature Architecture", **Xavier Riley, Drew Edwards and Simon Dixon (Queen Mary University of London, UK)**
12:10	"The Stories Behind the Sounds: Finding Meaning in Creative Musical Interactions with AI", **Jon Gillick (University of the Arts London, UK)**
12:30	Announcements: "The Cadenza Challenge for Improving Music for People with a Hearing Loss", **Gerardo Roa Dabike and Trevor Cox (Universities of Salford, UK).** "Timbre Tools Hackathon; Timbre Tools for the Digital Instrument Maker", **Charalampos Saitis, (Queen Mary University of London, UK).**
12:45	*Lunch - Poster Session*
14:15	"An automated pipeline for characterizing timing in jazz trios", **Huw Cheston, Ian Cross, and Peter Harrison (University of Cambridge, UK)**
14:35	"Electric Guitar Sound Restoration with Diffusion Models", **Ronald Mo (University of Sunderland, UK)**
14:55	"DedAI: Advanced AI-Driven Music Composition Informed by EEG-Based Emotional Analysis", **Elliott Mitchell (University of Westminster, UK)**
15:15	*Break (Coffee break)*
15:35	"PolyDDSP: A lightweight, Polyphonic Differentiable Digital Signal Processing Library", **Tom Baker, Ke Chen (University of Manchester, UK) and Ricardo Climent (NOVARS Research Institute, University of Manchester, UK)**
15:55	"A Two-Stage Differentiable Critic Model for Symbolic Music", **Yuqiang Li, Shengchen Li (Xi'an Jiaotong-Liverpool University, China) and George Fazekas (Queen Mary University of London, UK)**
16:15	Close - Simon Dixon

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

Keynote Talk

Keynote: By Stefan Bilbao

Tittle: Physics-based Audio: Sound Synthesis and Virtual Acoustics

Abstract: Any acoustically-produced sound produced must be the result of physical laws that describe the dynamics of a given system---always at least partly mechanical, and sometimes with an electronic element as well. One approach to the synthesis of natural acoustic timbres, thus, is through simulation, often referred to in this context as physical modelling, or physics-based audio. In this talk, the principles of physics-based audio, and the various different approaches to simulation are described, followed by a set of examples covering: various musical instrument types; the important related problem of the emulation of room acoustics or "virtual acoustics"; the embedding of instruments in a 3D virtual space; electromechanical effects; and also new modular instrument designs based on physical laws, but without a counterpart in the real world. Some more technical details follow, including the strengths, weaknesses and limitations of such methods, and pointers to some links to data-centred black-box approaches to sound generation and effects processing. The talk concludes with some musical examples and recent work on moving such algorithms to a real-time setting.

Bio: Stefan is a Professor (full) at Reid School of Music, University of Edinburgh, he is the Personal Chair of Acoustics and Audio Signal Processing, Music. He currently works on computational acoustics, for applications in sound synthesis and virtual acoustics. Special topics of interest include: Finite difference time domain methods, distributed nonlinear systems such as strings and plates, architectural acoustics, spatial audio in simulation, multichannel sound synthesis, and hardware and software realizations.

More information on: https://www.acoustics.ed.ac.uk/group-members/dr-stefan-bilbao/

Location

Mason Lecture Theatre, Bancroft building
Queen Mary University of London - Mile End Campus

Posters

1	“Tokenization Informativeness and its Impact on Symbolic MIR Tasks”, **Dinh-Viet-Toan Le (Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France), Louis Bigo (Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France) and Mikaela Keller (Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France)**
2	“Rhythm Guitar Tablature Continuation from Chord Progression and Tablature Prompt”, **Alexandre D'Hooge (Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France), Louis Bigo ((Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France) and Ken Déguernel (Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France)**
3	“Subjective Evaluation of Roughness for Perceptual Audio Coding”, **He Xie, Bruno Fazenda and Duncan Williams (University of Salford, UK)**
4	“Adapting Beat Tracking Models for Salsa Music: Establishing a Baseline with a novel dataset”, **Antonin Rapini and Anna Jordanous (University of Kent, UK)**
5	“Efficient Optimisation Techniques for Large Generative Audio Models”, **Bradley Aldous and Ahmed M. A. Sayed (Queen Mary University of London, UK)**
6	“Towards Melodic Development with Discrete Diffusion Models for Symbolic Music”, **Keshav Bhandari and Simon Colton (Queen Mary University of London, UK)**
7	“Rethinking music representation learning for music and musicians”, **Julien Guinot (Queen Mary University of London, UK), Eliot Quinton (Universal Music group, UK) and George Fazekas (Queen Mary University of London, UK)**
8	“Towards End-to-End Automatic Guitar Transcription via a Multimodal Approach”, **Zixun (Nicolas) Guo and Simon Dixon (Queen Mary University of London, UK)**
9	“Neuro-Symbolic Meta-Composition”, **Adam Z. He (Queen Mary University of London and DAACI, UK) Doon MacDonald (DAACI, UK), Geraint A. Wiggins (Vrije Universiteit Brussel, Belgium and Queen Mary University of London, UK)**
10	“Improving music recommendation and representation using DJ mix tracklists”, **Gregor Meehan and Johan Pauwels (Queen Mary University of London, UK)**
11	“Self-Supervised Music Source-Separation using Vector-Quantized Source Category Estimates”, **Marco Pasini (Queen Mary University of London, UK), Stefan Lattner (Sony CSL) and George Fazekas (Queen Mary University of London, UK)**
12	“Limited-Data Incremental Learning in Music”, **Christos Plachouras, Johan Pauwels and Emmanouil Benetos (Queen Mary University of London, UK)**
13	“Timbre Tools for the Digital Instrument Maker”, **Haokun Tian and Charalampos Saitis (Queen Mary University of London, UK)**
14	“Music-Driven dance generation”, **Qing Wang and Shanxin Yuan (Queen Mary University of London, UK)**
15	“Using AI to Help Render Orchestral Scores to Expressive Mockups”, **Yifan Xie and Mathieu Barthet (Queen Mary University of London, UK)**

16	“Computational auditory scene analysis: what next?”, **Farida Yusuf and Marcus Pearce (Queen Mary University of London, UK)**
17	“Multimodal AI for musical collaboration in immersive environments”, **Qiaoxi Zhang and Mathieu Barthet (Queen Mary University of London, UK)**
18	“Generative Deep Learning for Explainable AI Music-Making: A survey and Taxonomy”, **Shuoyang Zheng (Queen Mary University of London, UK), Anna Xambó (De Monfort University, UK) and Nick Bryan-Kinns (University of the Arts London, UK)**

Announcements

1	“Planning The 2nd Cadenza Challenge for Improving Music for People with a Hearing Loss”, **Gerardo Roa Dabike (University of Salford, UK), Michael A. Akeroyd (University of Nottingham, UK), Scott Bannister (University of Leeds, UK), Jon Barker (University of Sheffield, UK), Trevor J. Cox (University of Salford, UK) Bruno Fazenda (University of Salford, UK), Jennifer Firth(University of Nottingham, UK), Simone Graetzer (University of Salford, UK), Alinka Greasley(University of Leeds, UK), Rebecca R. Vos(University of Salford, UK), William M. Whitmer (University of Nottingham, UK)**
2	“Timbre Tools Hackathon; Timbre Tools for the Digital Instrument Maker”, **Charalampos Saitis, Haokun Tian, Jordie Shier and Bleiz Macsen Del Sette (Queen Mary University of London, UK).**

Organizing Committee

Supported by UKRI AIM CDT

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

AI + MUSIC

Aditya Bhattacharjee
James Bolt
Carey Bunks
Adam Garrow
Yinghao Ma
Tyler McIntosh
Christopher Mitcheltree
Ashley Noel-Hirst
Jordan Shier
David Südholt
Ioannis Vasilakis
Ningzhi Wang
Alexander Williams
Chin-Yun Yu

DMRN+18: Digital Music Research Network
One-day Workshop 2023

Queen Mary University of London
Tue 19 December 2023

In-depth performance analysis of the state-of-the-art algorithm for automatic drum transcription

Mickaël Zehren*[1], Marco Alunno[2] and Paolo Bientinesi[1]

[1]Department of Computing Science, Umeå Universitet, Sweden, mzehren@cs.umu.se
[2]Department of Music, Universidad EAFIT, Colombia

***Abstract*— In this work, we assess the most common sources of errors in a recent drum transcription algorithm.**

***Index Terms*—** Automatic drum transcription

I. Introduction

In music information retrieval, the task of Automatic Music Transcription (AMT) is especially important because the results it produces—i.e., the notes played by the instruments—help estimating many high-level features of a musical track, such as structure, melody, and rhythm. A subtask of AMT is automatic drum transcription in the presence of melodic instruments (DTM), which focuses on the estimation of the notes' onsets and their corresponding drum instrument in multi-instrument tracks.

Recently, we presented a new DTM algorithm based on large supervised learning from crowdsourced annotations [1]; thanks to the size and diversity of the datasets curated, we found that this algorithm surpasses the accuracy of the previous methods [2]. However, the resulting models are not perfect, as their estimations still contain mistakes.

In this work, we expose the most common sources of errors in the estimations, aiming to help the development of even more accurate models. This was done in three steps, as described in the following.

II. Experiments

First, to identify the most difficult instruments to transcribe, we independently evaluated the performance of the models on the different instrument classes. When trained and evaluated on (a different split of) the crowdsourced datasets, we observed that the typical difficulty of transcribing the instruments that play the least was attenuated. This is likely because now we can count with a much larger amount of training examples than before. However, despite this improvement, our model cannot yet transcribe cymbals as reliably as drums.

Second, to understand why cymbals are problematic, we employed both a new metric and a pseudo confusion matrix. Through the new metric, we identified that the estimations of cymbals are prone to mistakes because of their timbre. Specifically, their long sustain overlapping with subsequent quiet (ghost) notes makes them difficult to transcribe. Through the pseudo-confusion matrix, we showed that different kinds of cymbals are hard to discern, very likely because of both their similar timbres and the presence of other (non-cymbals) overlapping instruments.

Last, we assessed how much the quality of crowdsourced annotations affected the results of the models' evaluation. Due to discrepancies in the labels, some of the correct estimations from the models could have been mistakenly reported as errors. To estimate the accuracy of the ground truth itself, we quantified the agreement among different annotators of the same tracks. Any difference in their annotations indicates that at least one of them made a mistake. We found that, indeed, the annotators made many mistakes similar to those reported from the models. Thus, the discrepancies between the estimations and the ground truth might be caused by errors from either the models or the annotators. Since the best model evaluated achieves a performance close to the agreements between annotators, we believe it has little measurable margin of improvement on the crowdsourced dataset.

III. Conclusions

Through crowdsourcing, we curated large datasets for the supervised training of DTM models and we conducted an in-depth analysis of their performance. The results of this study highlight the limits of the current method and can be used to steer the development of future algorithms.

IV. References

[1] M. Zehren, M. Alunno, and P. Bientinesi, "High-quality and reproducible automatic drum transcription from crowdsourced data," *Signals*, vol. 4, no. 4, pp. 768–787, 2023. [Online]. Available: https://www.mdpi.com/2624-6120/4/4/42

[2] R. Vogl, G. Widmer, and P. Knees, "Towards multi-instrument drum transcription," in *21th International Conference on Digital Audio Effects (DAFx-18)*, June 2018. [Online]. Available: http://arxiv.org/abs/1806.06676

*The computation was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS).

DMRN+18: Digital Music Research Network
One-day Workshop 2023

Queen Mary University of London
Tue 19 December 2023

Automatic Guitar Transcription with a Composable Audio-to-MIDI-to-Tablature Architecture

Xavier Riley[*], Drew Edwards and Simon Dixon

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

***Abstract*— This work-in-progress demonstrates an end-to-end guitar transcription system. The architecture takes as input a solo guitar recording, transcribes the audio to MIDI, and then estimates a tablature for the performance. The audio-to-MIDI transcription exhibits strong generalisability, including state-of-the-art performance on GuitarSet in a zero-shot setting. The tablature estimation is a novel approach applying masked language modeling to per-note string assignment.**

***Index Terms*—** guitar, transcription, tablature, AMT

I. Guitar Multi-pitch Estimation

Automatic transcription of piano has achieved good results in recent years due to the availability of large datasets such as MAESTRO [1]. Several successful architectures have been proposed, however the guitar does not yet have a comparable dataset with which to train these models. Existing guitar datasets tend to be smaller, with less timbral diversity [2]. We address this lack of data by adapting a recent score alignment technique proposed by Maman and Bermano [3]. We use this to produce aligned MIDI for 78 commercially available guitar recordings. These form our new dataset which we then use to fine-tune an existing piano model. In contrast the work by Maman and Bermano, we use a newer high-resolution piano model proposed by Kong et al. [4] which is shown to be more robust to noisy labels. We also use data augmentations on the MAESTRO dataset when training the base piano transcription model. This helps with generalisability when fine-tuned on guitar recordings.

II. Tablature Estimation

Our approach to guitar tablature estimation uses the MIDI as input instead of audio. This loses timbral information but affords certain advantages. First, since the input and output are symbolic, a user can change the string and fret assignment of a particular set of notes and regenerate the estimated tablature. Second, this modular architecture provides a novel solution to arranging for guitar with a MIDI keyboard. A composer or arranger can play MIDI and quickly view how it could be performed on guitar.

We model the task of guitar tablature estimation as a masked language modeling task. Our ground truth data consists of guitar tablature transcriptions (from the 78 performances mentioned in Section I and GuitarSet [5]), in MusicXML or GuitarPro format. These are converted to six-track MIDI files, with one track per string. We use the Structured tokenizer from MidiTok [6]. For each note event N_i, we output the following tokens: $N_i \rightarrow S_i, T_i, P_i, V_i, D_i$, where $S_i \in \{1, 2, 3, 4, 5, 6\}$ is the string, T_i is the relative time shift, P_i is the pitch, V_i is the velocity, and D_i is the duration. During training, we mask and predict the S_i tokens. Our current best model using a BART [7] Transformer architecture achieves approximately 80% accuracy on a held out test set without any post-processing and is still a work-in-progress.

III. References

[1] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *7th International Conference on Learning Representations*, New Orleans, USA, 2019.

[2] Y. Chen, W. Hsiao, T. Hsieh, J. R. Jang, and Y. Yang, "Towards automatic transcription of polyphonic electric guitar music: A new dataset and a multi-loss transformer model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 786–790.

[3] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 14 918–14 934.

[4] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[5] Q. Xi, R. M. Bittner, J. Pauwels, X. Ye, and J. P. Bello, "GuitarSet: A dataset for guitar transcription," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018, pp. 453–460.

[6] N. Fradet, J.-P. Briot, F. Chhel, A. El Fallah Seghrouchni, and N. Gutowski, "MidiTok: A python package for MIDI file tokenization," in *Extended Abstracts for the Late-Breaking Demo Session of the 22nd International Society for Music Information Retrieval Conference*, 2021.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2019.

[*] XR and DE are research students at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music. XR is supported by UK Research and Innovation [grant number EP/S022694/1]; DE is supported by Queen Mary University of London and Yamaha.

DMRN+18: DIGITAL MUSIC RESEARCH NETWORK ONE-DAY WORKSHOP 2023

QUEEN MARY UNIVERSITY OF LONDON
TUE 19 DECEMBER 2023

Digital Music
Research Network

The Stories Behind the Sounds: Finding Meaning in Creative Musical Interactions with AI

Jon Gillick

Creative Computing Institute, University of the Arts London, United Kingdom, j.gillick@arts.ac.uk

Through a series of three studies, this talk explores the big-picture experiences of musicians, producers, composers, and listeners as they attempt to introduce AI into their musical lives in meaningful ways.

The work covered here makes up the last chapter of my dissertation [1], and it was done very much in response to the work I did in the few *preceding* years that focused more on specific ML algorithms, models, and tools. In this earlier work, I (and most others in the field) typically focused on coming up with technology that could do something new and exciting while still trying to fit into existing creative processes. But what happens when we move into the real world? How do real-world interactions between people and AI music systems play out? In this talk, I start by taking a step back from specific musical problems or situations, instead focusing on the messy but practical experiences of musicians or listeners when we try to introduce machine learning in a meaningful way into some area of their musical experience where it hasn't been involved at all before.

My approach toward the studies in this talk acknowledges that every creative process is different. Some people write music with an instrument, some write with a computer, some use pen and paper; sometimes we create to meet a deadline, sometimes we create because we feel inspired or emotional or bored; sometimes we create alone, sometimes we work together. We all change from moment to moment and year to year as we go through different experiences and face different situations. Might our needs and experiences in working with musical AI be similarly individualized? How can we meet individuals where they are in order to overcome the practical barriers that prevent them from trying a new technology?

In part because of the always-changing contexts in which people create music and art, studying human interactions with AI in situated creative environments is very hard. Controlled studies "in the lab'" might separate creators from their usual processes in ways that color their experiences, making it difficult to isolate the effects of new AI technology [2]. Participants brought in to try out prototypes or learn how to use AI-based creative tools for the first time might find their learning curves to be steep; it can take a long time to start to understand how AI works or how to use it. And finally, participants might not feel very invested in the outcomes of their (often unfamiliar) interactions with AI.

The research in this talk begins with the following question: what are experiences and interactions like for people who *have a reason to be emotionally invested* in music created with AI? I approach this question from different angles in each of the three studies:

In the first study, I use first-person design research methods to probe the experiences of a group of people (not necessarily musicians) *listening* to individually customized music that uses audio samples (and stories about those samples) from meaningful moments in their own lives. I find that when listening to music created personally for them using this material (which could presumably be done in some form through AI), participants cared relatively little about the "quality" of the music; what mattered much more was how well it fit with their individual understanding of the stories behind the samples that were used in the music.

The second study describes my own firsthand experience producing a song together with a group of 4 people that ended up as the winning entry submitted to the 2021 AI Song Contest, an international contest exploring the potential uses of AI for songwriting. Our collaboration built on the findings from my first study through our approach to coming up with a shared narrative starting point for any AI-generated musical material

In the last study, I work with two musicians, one professional and one amateur, using AI to manipulate sound collections that they find meaningful in order to create musical material (new samples, loops, or digital instruments) to compose with. This study points to the potential for productive collaboration between musicians and experts in AI/ML ("AI Music Engineers"), who, much like recording engineers often do with studio gear, might be able to help guide artists through the landscape of available models or methods and help apply appropriate tools for the job in a given situation.

ACKNOWLEDGMENT

The work presented in this talk draws from work done in collaboration with Noura Howell, Wesley Deng, Julia Park, YangYang Yang, Carmine-Emanuele Cella, David Bamman, Matt Sims, Max Savage, and Brodie Jenkins.

REFERENCES

[1] Gillick, Jonathan. *Creating and Collecting Meaningful Musical Material with Machine Learning*. Dissertation. University of California, Berkeley, 2022.

[2] C.Z.A Huang, H.V. Koops, E. Newton-Rex, M. Dinculescu, and C.J Cai, "AI Song Contest: Human-AI Co-creation in Songwriting" in

Proceedings of the 21st International Society for Music Information Retrieval Conference, pp. 708-716, October 2020.

An automated pipeline for characterizing timing in jazz trios

Huw Cheston[1*], Ian Cross[1], and Peter Harrison[1]

[1]Centre for Music & Science, University of Cambridge, United Kingdom

***Abstract*— The jazz rhythm section (piano, bass, drums) has seen little attention in performance studies. We developed a pipeline to automatically extract features relating to swing, rhythmic feel, complexity, and performer interaction from 300 commercial audio recordings. A classification model was able to correctly identify the pianist playing in 52% of these recordings using only the provided rhythmic features (chance accuracy: 10%), with swing and feel proving the strongest predictors. Our results speak to the importance of rhythm in defining a performer's unique improvisational style.**

I. INTRODUCTION

Despite its presence in nearly every jazz ensemble, the rhythm section of piano, bass, and drums has received little scholarly attention. In this paper, we analyze the variation between the different instruments in this unit across a range of rhythmic features. We then consider which of these predictors are the strongest predictors of stylistic variation between rhythm sections led by different musicians.

II. METHOD

Thirty commercial audio recordings by ten bandleaders ($n = 300$) were selected for analysis, based on performer popularity and prolificacy (assessed using real-world listening and discographic data). A pipeline was developed to extract timing onsets automatically from these recordings. The *Spleeter* [1] and *Demucs* [2] source separation models were applied, yielding isolated stems for each instrument. An onset detection algorithm was used to detect the start of notes by each instrument, and the position of the underlying quarter note pulse in the audio mixture. Onsets were matched to their nearest pulse to estimate the meter of each musician. Comparison to a reference set of annotations created for 10% of the dataset indicated a mean $F_1 = 0.86$ for detected onsets.

III. RESULTS

Features relating to a performer's swing (beat-upbeat ratio, BUR [3]), complexity (compression rate of discrete inter-onset intervals), feel (relative position to bandmates), tempo (mean, stability, and slope) and interaction (phase coupling to and from bandmates) were extracted from all recordings. Pianists played the most complex rhythms and marked the pulse 1/64th note after bassists and drummers on average, neither of whom adjusted to match their beat. Bassists and drummers played fewer complex rhythms in tight synchrony (mean asynchrony: < 1/256th note) and coupled strongly to each other. Drummers played with the most swing (mean BUR: 1.65:1) in the ensemble, with pianists playing closer to notated "straight" than "swung" eighths (BUR: 1.21:1). Bassists tended to play notated "straight" eighth note rhythms (mean BUR: 1.04:1).

Features extracted from the piano player in each recording were then entered into a random forest classification model, fitted using stratified k-fold cross-validation ($k = 5$). The model was able to correctly identify the pianist using only rhythmic features in 52% of recordings, i.e., 5x better than chance (accuracy: 10%). Classification accuracy was higher on average for pianists associated with "hard bop" and "soul jazz" genres (e.g., Junior Mance), which typically place great emphasis on rhythmic drive, and lower for "modal" pianists (e.g., Bill Evans) whose innovations were harmonic.

The mean variable importance scores (Fig. 1) computed for each feature category suggested that feel, tempo, and swing were the strongest predictors used in the model. By extension, these categories can be considered the aspects of musical timing that best contributed towards defining a pianist's improvisational style, over and above the complexity of their performance and how they interacted with their bandmates.

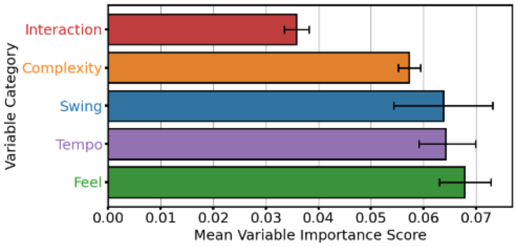


Figure 1. Mean variable importance score across categories. Error bars show 95% confidence intervals via bootstrapping, $n = 10{,}000$ replicates.

IV. DISCUSSION

Our results speak to the importance of rhythm in defining a performer's improvisational style. They also demonstrate the value of applying quantitative methodology to improvised jazz, a subject that often resists empirical study given its lack of notated scores and the freedoms afforded to its performers.

REFERENCES

[1] R. Hennequin, et. al., "Spleeter: a fast and efficient music source separation tool with pre-trained models", *JOSS*, vol. 50/5, April 2020.

[2] S. Rouard, F. Masssa, and A. Défossez, "Hybrid Transformers for Music Source Separation", *arXiv:2211.08553*, Nov. 2022.

[3] F. Benadon, "Slicing the Beat: Jazz Eighth Notes as Expressive Microrhythm", *Ethnomusicology*, vol. 50/1, Winter 2006, pp. 73–98.

H. Cheston, I. Cross, and P. Harrison are all with the Centre for Music & Science, Cambridge, United Kingdom (corresponding author e-mail: hwc31@cam.ac.uk).

DMRN+18: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2023

QUEEN MARY UNIVERSITY OF LONDON
TUE 19 DECEMBER 2023

Electric Guitar Sound Restoration with Diffusion Models

Ronald K. Mo

School of Computer Science, University of Sunderland, United Kingdom
ronald.mo@sunderland.ac.uk

Abstract—**This work aims to investigate the potential of employing Denoising diffusion probabilistic models, commonly referred to as *diffusion models*, to revert a processed electric guitar recording to its original, unaltered form while retaining all the expressive elements of the performance such as dynamics and articulation. Specifically, a parallel dataset is constructed, containing both the unprocessed and processed versions of the guitar recordings, which is used for training a diffusion model. To preserve the expressiveness, the model is *conditioned* on the processed guitar recording when restoring the raw guitar sound. This research has the potential to enhance the accuracy of various music information retrieval tasks, such as automatic music transcription.**

I. BACKGROUND

Denoising diffusion probabilistic models, also known as diffusion models, have showcased their capacity for generating realistic images [1]. In particular, a diffusion model comprises two processes. The *forward process* entails the repetitive addition of Gaussian noise to the input data x, such as images. Conversely, the *reverse process* is responsible for denoising a vector sampled from $p(z)$ (i.e., the latent representation of x in an iterative manner, ultimately restoring the input data to its original state. A well-trained diffusion model excels in learning the data distribution $p(x)$ within a provided set of data, enabling it to create novel data and surpass the performance of traditional Generative AI (GenAI) models.

Beyond generating images, diffusion models have found applications in various GenAI tasks [3]. To facilitate the conditioning of the generated content, diffusion models often incorporate a *conditioning mechanism*. Conditional diffusion models are designed to learn the conditional distribution of $p(z|y)$ where y is the *conditioning input* such as class labels, text, or audio. Nevertheless, it's worth noting that the application of GenAI models to audio signal processing remains an area that has not been thoroughly explored, to the best of our knowledge.

II. OVERVIEW

This work seeks to explore the potential of utilizing diffusion models to transform a processed electric guitar recording to its raw format (i.e., a clean electric guitar sound), akin to image denoising. To accomplish this, a parallel dataset containing both the clean and processed guitar sounds is constructed. The clean guitar playing is performed and recorded by the author who is a professional studio guitarist. The recorded guitar recording is processed using *Logic Pro*. Due to the preliminary nature of this study, it only considers *distortion* as the processing technique.

A diffusion model is developed and trained using the dataset mentioned above. To reduce the training time, a *latent* diffusion model is used [4]. It first *encodes* the input x (i.e., $\mathcal{E}(x)$) into a lower-dimension representation, carries out both the forward and reverse processes on $\mathcal{E}(x)$, and eventually *decodes* $\mathcal{E}(x)$ (i.e., $\mathcal{D}(\mathcal{E}(x))$) to obtain $\tilde{x}$. To preserve the expressiveness of the guitar playing, the model conditions its output on the *processed guitar recording*. More precisely, the conditional input y is encoded using the Diffusion Magnitude-Autoencoding introduced by Schneider et al. [5] and concatenated with the sampled vector during the reverse process. The generated outputs will be evaluated both objectively and subjectively.

III. CONCLUSION

While the full potential of diffusion models in audio signal processing remains largely unexplored [6], this study introduces a novel approach for the restoration of electric guitar sounds using diffusion models. Given the potential to extend this method to multi-track scenarios, this research has the capacity to enhance the performance of various music information retrieval tasks, including automatic music transcription, music source separation, and beat detection.

REFERENCES

[1] J. Ho et al. “Denoising diffusion probabilistic models,” in *Advances in neural information processing systems 33*, 2020, pp. 6840–6851.

[2] M.W.Y. Lam et al. “BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis,” in 2022-*10th International Conference on Learning Representations*, 2022.

[3] G. Mittal et al. “Symbolic Music Generation with Diffusion Models”, in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021, pp. 468-475

[4] R. Rombach et al. “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684-10695.

[5] F. Schneider et al. “Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion,” in *arXiv preprint arXiv:2301.11757,* (2023).

[6] Kong, Zhifeng, et al. “Diffwave: A versatile diffusion model for audio synthesis. ” in *arXiv preprint arXiv:2009.09761* (2020).

Ronald K. Mo is with the School of Computer Science. University of Sunderland, United Kingdom (corresponding author e-mail: ronald.mo@ sunderland.ac.uk).

DedAI: Advanced AI-Driven Music Composition Informed by EEG-Based Emotional Analysis

Elliott Mitchell

Department of Music Production, University of Westminster, United Kingdom,
w1864126@my.westminster.ac.uk & elliott@iamdedeye.com

ABSTRACT

This presentation introduces DedAI, a state-of-the-art platform leveraging artificial intelligence to create emotionally resonant music compositions. At the core of DedAI's innovation is the integration of advanced EEG (Electroencephalography) technology and sophisticated AI models, including MERT (Music Emotional Recognition Transformer) variants like MERT-v1-330M and MERT-v1-95M. This integration facilitates an unprecedented real-time translation of emotional states, discerned through EEG, into AI-generated musical compositions.

An auxiliary research project enriches DedAI's approach, "Measuring Music's Emotional Impact with EEG: A Study on Musicians and Non-Musicians." This study provides invaluable insights into the neural underpinnings of music-induced emotions, enhancing DedAI's capability to tailor musical compositions based on EEG-derived emotional cues. By processing EEG data with Emotiv's sophisticated emotion detection algorithms, DedAI classifies emotional states, which are then interpreted by MERT models to guide the AI composition process.

This research employs advanced signal processing techniques for EEG data, including spatial filtering and Independent Component Analysis (ICA), followed by extracting specific EEG frequency bands using bandpass filters. These bands are correlated with distinct mental states and emotional responses. Subsequent machine learning algorithms, notably SVMs and neural networks, classify these EEG patterns into discernible emotional states, which are intricately woven into the music composition process. DedAI stands at the forefront of this research, showcasing a pioneering intersection of neuroscience, AI, and musicology. A potential collaboration with AudioSparx is envisioned to enrich DedAI's music database, providing a wide range of royalty-free music for model training and enhancing the system's ability to generate diverse and personalised musical experiences. This project contributes to the field of music psychology and opens new avenues for personalised music therapy and cognitive enhancement applications.

ABBREVIATIONS AND ACRONYMS

EEG: Electroencephalography
MERT: Music Emotional Recognition Transformer

ACKNOWLEDGMENT

I extend sincere thanks to Hussein Boon, Principal Lecturer at the University of Westminster, for his invaluable guidance and support as the supervisor of my Final Major Project. His recommendation to submit this work has been crucial to its development.

Appreciation is also due to Dr Jasmine Taylor, Senior Lecturer at the Westminster School of Arts, University of Westminster, for her insightful supervision of my research project "Measuring Music's Emotional Impact with EEG: A Study on Musicians and Non-Musicians."

Gratitude is extended to Yizhi Li from the University of Sheffield for his significant contributions to the MERT models and assistance in sourcing this model, which has been pivotal in advancing the DedAI project.

REFERENCES

1. AudioSparx. Commercial Music for Video, TV, Film, and Media from AudioSparx.com. Retrieved from https://www.audiosparx.com/.
2. Farnell, A. (2010). Designing sound. MIT Press.
3. Juslin, P. N., Sloboda, J. A. (2010). Handbook of music and emotion: theory, research, and applications. Oxford University Press.
4. Li, Y., Yuan, R., Zhang, G., et al. (2023). MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. arXiv:2306.00107.
5. Lima, M. (2013). Visual complexity: mapping patterns of information. Princeton Architectural Press.
6. Plourde-Kelly, A. D., Saroka, K. S., Dotta, B. T. (2021). The impact of emotionally valenced music on emotional state and EEG profile: Convergence of self-report and quantitative data. Neuroscience Letters, 758, 136009.
7. Ramirez, R., Planas, J., Escude, N., et al. (2018). EEG-Based Analysis of the Emotional Effect of Music Therapy on Palliative Care Cancer Patients. Front Psychol., 9, 254.
8. Sanyal, S., Nag, S., Banerjee, A., et al. (2019). Music of brain and music on brain: a novel EEG sonification approach. Cogn Neurodyn, 13(1), 13–31.
9. Tandle, A. L., Joshi, M. S., Dharmadhikari, A. S., Jaiswal, S. V. (2018). Mental state and emotion detection from musically stimulated EEG. Brain Inform., 5(2).
10. Zhou, T. H., Liang, W. L., Liu, H., et al. (2023). EEG Emotion Recognition Applied to the Effect Analysis of Music on Emotion Changes in Psychological Healthcare. Int J Environ Res Public Health, 20(1), 378.

DMRN+18: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2023

QUEEN MARY UNIVERSITY OF LONDON
TUE 19 DECEMBER 2023

PolyDDSP: A lightweight, Polyphonic, Differential Digital Signal Processing Library

Tom Baker, Ke Chen[1] and Ricardo Climent[2]

[1]Department of Computer Science, University of Manchester, UK, tom.baker@manchester.ac.uk
[2]NOVARS Research Institute, University of Manchester, UK

***Abstract*— In this abstract, we introduce an ongoing extension of the Differentiable Digital Signal Processing (DDSP) [1] framework, expanding its capabilities to accommodate multi-voice (n-voice) processing. Our approach encompasses three key strategies: leveraging parallel operations within a unified decoder pipeline, incorporating a multi-fundamental frequency (multi-F0) pitch detection system, and devising a new, pre-trained, lightweight timbre encoder. This work-in-progress proposes a scalable and efficient technique for managing complex polyphonic audio situations, while preserving the core benefits of the original DDSP model, including interpretable latents, rapid inference and relatively low computational and data demands for training.**

***Index Terms*—** Digital Signal Processing, Machine Learning, Real-time, Polyphony, Timbre Transfer

I. Multi-Pitch

Adapting the Differentiable Digital Signal Processing (DDSP) [1] for polyphonic audio first required addressing the pitch encoder. The original model, developed by Engel et al., utilised the CREPE [2] pitch detection method. This method, notable for its lightweight convolution structure, excelled at identifying the dominant fundamental frequency (F0) in audio. However, its limitation was that it could only detect a single F0. Bittner et al. [3] overcame this by using Harmonically Stacked-Constant Q (HS-CQT) spectrograms, which allowed for the identification of multiple F0 frequencies through a similarly efficient 2D convolution structure. This advancement not only enabled the detection of multiple pitches but also their conversion into discrete note events with individual velocities.

II. Scalability and Generalisation

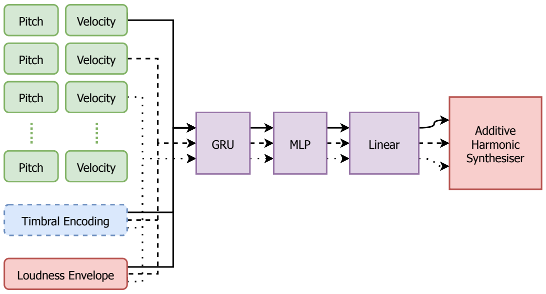


Figure 1: Model structure showing the parallel voices fed batch-wise though the decoder pipeline to generate synthesiser parameters.

With the ability to identify multiple pitches, the next step is to translate these into synthesiser parameters. The initial idea might be to expand the decoder pipeline. However, this approach requires setting a fixed number of voices throughout the model's training and vastly increasing training parameters per voice added. Our solution is a single, generalist decoder pipeline. This decoder, treating each pitch and timbre encoding as a unique 'voice', generates synthesiser parameters for that specific frequency and its harmonic spectrum. This parallel processing approach enables the training of a single decoder that can manage any number of voices, offering flexibility as computational resources vary. When the input pitches exceeds the maximum voices, this operates like a standard polyphonic synthesiser, assigning voices on a first-in-first-out basis.

III. Timbre: Further Work

The next phase involves extracting distinct timbral information for each voice. To achieve this, we plan to utilise the F0 frequency data and the HS-CQT spectrograms from our pitch encoder. By employing a similar convolutional structure, we are developing a model for partial pitch-informed source separation. This model aims to classify individual timbres for each voice. For our goals, a basic timbre classification that can accurately distinguish different timbral characteristics is sufficient and will allow the decoder to learn the corresponding parameterisations. However, the more representative the timbre encoding, the less training the decoder framework requires.

IV. References

[1] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," Jan. 2020, arXiv:2001.04643 [cs, eess, stat].

[2] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A Convolutional Representation for Pitch Estimation," Feb. 2018, arXiv:1802.06182 [cs, eess, stat].

[3] R. M. Bittner, J. J. Bosch, D. Rubinstein, G. Meseguer-Brocal, and S. Ewert, "A Lightweight Instrument-Agnostic Model for Polyphonic Note Transcription and Multipitch Estimation," May 2022, arXiv:2203.09893 [cs, eess].

DCritics: A Two-Stage Differentiable Critic Network for Symbolic Music

Yuqiang Li[1] Shengchen Li[1] and George Fazekas[2]

[1]Music Informatics Group / XJTLU, Suzhou, China, yuqiang.li19@student.xjtlu.edu.cn
[2]C4DM / Queen Mary University of London, UK

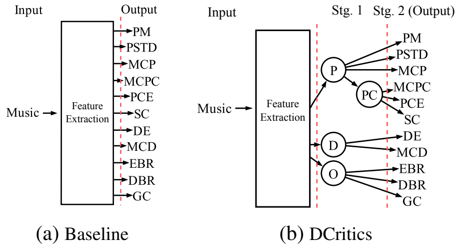
***Abstract*— In symbolic music generation, objective evaluation metrics have been widely employed to examine how close the generated music is to that of the test dataset. However, the computation process of most metrics is non-differentiable, being unable to provide feedback when training generation models. This work proposes *DCritics*, a two-stage differentiable critic model that approximates multiple evaluation metrics of the symbolic music input. Instead of directly modelling the metrics from the hidden space, *DCritics* first estimates the distributions of pitch, pitch class, duration and intra-bar onsets, then model the derived metrics accordingly. It is hypothesized that the two-stage design of *DCritics* can improve the accuracy of estimated metrics and reduce the parameter size. The network is extendable for other metrics according to researchers' needs by applying the same concept.**

***Index Terms*—** Objective Evaluation, Symbolic Music Representation, Pre-Training

I. Methods

Since many evaluation metrics are based on the statistics of specific musical features regarding the music input, *DCritics* would first model the needed distributions then compute or approximate its statistics rather than directly model the statistics. The proposed approach offers two advantages. (1) The statistics are derived from meaningful distributions of relevant musical features instead of from an unexplainable hidden space. (2) When multiple statistics of the same distribution are modeled together, their gradient do not directly affect other distributions that they do *not* depend on.

For the purpose of a concrete demonstration, this work focuses on 12: PM, PSTD, MCP, MCPC, PCE, SC, DE, DBR, EBR and GC[1]. Figure 1 compares a baseline implementation with a *DCritics* implementation. In the DCritics implementation, the distributions of **P**itch, **P**itch **C**lass, **D**uration, and intra-bar **O**nset are estimated before estimating the corresponding statistics. This layout ensures that the metrics learned from meaningful and only-necessary distributions of musical features.



(a) Baseline

(b) DCritics

Figure 1: Two Implementations of Differentiable Critic Networks

II. Training

The entire network can be trained at the same time, but only training some branches is also fine for fine-tuning. For early training, a random music token dataset can be used. As long as the input representation of music X is valid, the ground truth musical feature distributions D (e.g. pitch distributions) and the true metric M can be computed as the labels for the dataset. Since D and M are learnt via supervised learning, the choice of loss function depends on the type of distribution or statistic.

The Wikifonia[2] dataset is selected but the approach applies to other datasets as well. OctupleMIDI[2] representation is used as it provides rich symbolic features for the input. Results would be compared in terms of the similarity between estimated metrics and the true metrics. It is hypothesized that the DCritics implementation would predict closer metrics to the ground truth.

III. Conclusion

The expect conclusion would be that forcing the model to learn musically meaningful features can improve the modelling of objective evaluation metrics accordingly.

IV. References

[1] H.-W. Dong, K. Chen, J. McAuley, and T. Berg-Kirkpatrick, "MusPy: A Toolkit for Symbolic Music Generation," in *The 21st International Society for Music Information Retrieval Conference*, no. arXiv:2008.01951, Montréal, Canada, Aug. 2020.

[2] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, "MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Aug. 2021, pp. 791–800.

[1]They are short for Pitch Mean, Pitch Standard Deviation, Most Common Pitch, Most Common Pitch Class, Pitch-Class Entropy, Scale Consistency [1], Durational Entropy, Down-Beat Rate, Empty-Beat Rate, Groove Consistency [1].

[2]`http://wikifonia.org`

DMRN+18: Digital Music Research Network
One-day Workshop 2023

Queen Mary University of London
Tue 19 December 2023

Tokenization Informativeness and its Impact on Symbolic MIR Tasks

Dinh-Viet-Toan Le[1], Louis Bigo[2] and Mikaela Keller[1]

[1]Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France, dinhviettoan.le@univ-lille.fr
[2]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

***Abstract*— With the breakthrough of Transformer neural networks within the MIR community, there has been an increasing interest in *tokenization*, a representation of music as sequences of elements, for which a variety of methods have been proposed. We explore the impact of two methods aiming at increasing the musical informativeness of token sequences: *intervalization* and *byte-pair encoding*. We find out that improving tokenization informativeness has an impact on downstream task performances, as well as bringing out musical interpretations.**

***Index Terms*—** Tokenization, Symbolic Music, Natural Language Processing (NLP), Machine Learning.

I. Tokenization in Symbolic Music

Tokenization refers to a process of representing a complex content into a sequence of elements. In NLP, tokenization is the task of segmenting a sequence of atomic elements (characters) by grouping them together into informative *tokens* [1] such as subwords or words. In contrast, tokenization for symbolic music can occur at different levels of granularity, thus leading to a variety of processes. The resulting elements, composing the musical sequences, derive from two levels of description: the choice of an initial *alphabet* of atomic elements *encoding* different aspects of music, and the *grouping* of these atomic elements into the more informative elements of a *vocabulary*. It thus exists a variety of tokenization strategies derived from MIDI-performance [2] or MIDI-score [3] data.

II. Tokenization Informativeness

In contrast with words in text, individual musical tokens do not carry much musical information on their own. This has encouraged the elaboration of methods to improve the *informativeness* of an initial tokenization strategy, including the tweaking of *encodings* or *groupings* approaches.

We focus on applying *intervalization* [4], aiming at encoding pitches with intervals rather than absolute MIDI values, and byte-pair encoding (BPE) [5] derived from NLP, that statistically groups atomic characters together into subword units that we refer to as musical *supertokens* in the case of symbolic music.

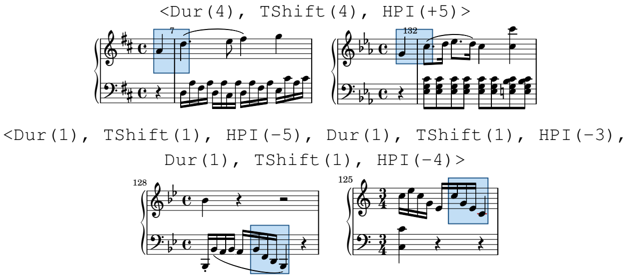

<Dur(4), TShift(4), HPI(+5)>
7
132
<Dur(1), TShift(1), HPI(-5), Dur(1), TShift(1), HPI(-3),
Dur(1), TShift(1), HPI(-4)>
128
125

Figure 1: Most common start-of-phrase (resp. end-of-phrase) supertokens.

III. Tokenization informativeness in MIR tasks

In order to evaluate the impact of tokenization informativeness, we evaluate these tokenization strategies on multiple downstream tasks involving various types of data (monophonic, polyphonic), models (Naive-Bayes, LSTM) and task scopes (composer classification, end-of-phrase detection). Our results show that such customized tokenization induce varying effects on the model's performance depending on the context. In particular, combining BPE and intervalization can lead to a +16% performance increase on a monophonic/Naive-Bayes/classification context. These improvements are higher on performance-based tokenization compared to score-based tokenization.

Supertokens also carry stylistic content. Analyses of the learned supertokens show that common start-of-phrase and end-of-phrase supertokens are matching musicology studies, such as a rising fourth at the beginning of a phrase or arpeggios on the tonic chord as an end-of-phrase motif (Figure 1).

IV. References

[1] S. J. Mielke, *et al.*, “Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP,” *arXiv preprint arXiv:2112.10508*, 2021.

[2] G. Hadjeres *et al.*, “The piano inpainting application,” *arXiv preprint arXiv:2107.05944*, 2021.

[3] Y.-S. Huang *et al.*, “Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[4] M. Kermarec, *et al.*, “Improving tokenization expressiveness with pitch intervals,” in *ISMIR – Late-Breaking Demo Session*, 2022.

[5] N. Fradet, *et al.*, “Byte pair encoding for symbolic music,” *arXiv preprint arXiv:2301.11975*, 2023.

DMRN+18: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2023

QUEEN MARY UNIVERSITY OF LONDON
TUE 19 DECEMBER 2023

Rhythm Guitar Tablature Continuation from Chord Progression and Tablature Prompt

Alexandre D'Hooge*[1], Louis Bigo[2] and Ken Déguernel[1]

[1]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France, alexandre.dhooge@algomus.fr
[2]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

***Abstract*— We propose a new approach to generate rhythm guitar continuation in tablature notation (tab), given a chord progression and a tab prompt. Our approach builds on existing work for suggesting chord positions and generating guitar tabs, and proposes a two-part model that provides user-control and explainability.**

***Index Terms*—** guitar tablature, symbolic music generation, assisted composition

I. Problem Statement

We can call rhythm guitar any guitar track that has an accompaniment role in a song [1]. In this work, we focus on generating continuation of rhythm guitar in the tablature (tab) domain. Such a tool could help artists generate backing tracks from a chosen chord progression, or be used by guitar beginners for assisted composition. An illustration of that objective is proposed Fig. 1.

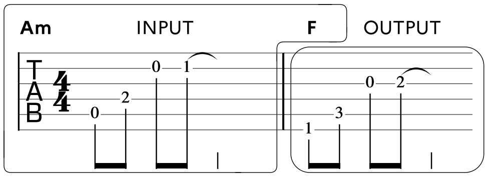


Figure 1: From an input tab prompt with a chord label and the name of the following chord, we want to generate a tab continuation for the next chord.

We call this task *continuation*, as we want the output of our model to guarantee the textural consistency of the input tab prompt. We define symbolic *texture* with two dimensions: *strumming* – which strings are plucked and when; and *position* – the strings and frets used to play the chord on the fretboard. The desired texture for the requested chord progression is obtained from an input prompt. The prompt is a tab notation of how the first chord is played, and we aim at mimicking the underlying texture on the following chords.

II. Proposed Approach

To replicate a provided texture over new chords, we propose the architecture shown Fig. 2.

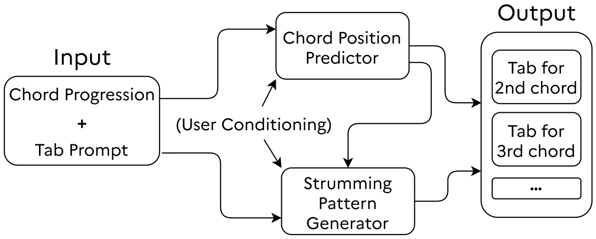


Figure 2: Summary diagram of our proposed approach.

Instead of generating rhythm guitar tabs directly, we suggest splitting the problem into subtasks: first predict an appropriate chord position to preserve textural properties, then generate a strumming pattern matching the pre-existing texture. Both sub-models are open to user conditioning for enhanced control. In particular, we will explore how conditioning signals can be used to reflect the user's preferences, like a favorite style, or preferred chord positions. We aim at comparing this approach with previous state-of-the-art, from a simple machine learning model [2] to more advanced end-to-end transformer models [3, 4]. This evaluation will assess in particular the playability and controllability of the generated tabs. Preliminary experiments were conducted with a recurrent neural network, and comparing the generated content with the original tabs yielded promising results. Some challenges remain, for instance when the expected chord shape has open strings, or when strumming patterns change significantly from one chord to the next, leaving room for improving our approach in the near future.

III. References

[1] D. Régnier, N. Martin, and L. Bigo, "Identification of rhythm guitar sections in symbolic tablatures," in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021.

[2] M. McVicar, S. Fukayama, and M. Goto, "AutoRhythmGuitar: Computer-aided Composition for Rhythm Guitar in the Tab Space," in *Proceedings ICMC|SMC*, 2014.

[3] 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," in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021.

[4] P. Sarmento, A. Kumar, Y.-H. Chen, C. Carr, Z. Zukowski, and M. Barthet, "GTR-CTRL: Instrument and Genre Conditioning for Guitar-Focused Music Generation with Transformers," in *Artificial Intelligence in Music, Sound, Art and Design*, ser. Lecture Notes in Computer Science, C. Johnson, N. Rodríguez-Fernández, and S. M. Rebelo, Eds. Cham: Springer Nature Switzerland, 2023, pp. 260–275.

*Research supported by ANR, project ANR-22-CE38-0001.

DMRN+18: DIGITAL MUSIC RESEARCH NETWORK ONE-DAY WORKSHOP 2023

QUEEN MARY UNIVERSITY OF LONDON
TUE 19 DECEMBER 2023

Subjective Evaluation of Roughness for Perceptual Audio Coding

Xie He, Bruno Fazenda, Duncan Williams[1]
[1*]University of Salford, H.Xie@edu.salford.ac.uk,

***Abstract*— This study investigates correlations between roughness metrics (musical [1], acoustic [2]) and the perceived audio quality of CODEC compressed musical intervals. 16 samples were synthesised using 2 musical instruments across 8 different note intervals and 4 audio conditions, which include uncompressed and 96kbps mp3 compressed audio and spectral manipulations of the harmonic content. Roughness metrics were extracted from the signals. 17 participants evaluated the BAQ of each sample using a MUSHRA listening test. A significant difference with a large effect size was found for audio conditions, whereas a significant but small effect size difference was found for musical intervals and instruments. Further, a correlation was found between perceived quality and roughness metrics, but the explained variance in the model was low.**

I. INTRODUCTION

Perceptual audio coding (CODECs) is known to create artefacts in the coded signal. It is posited that the acoustic metrics of roughness are reasonable, objective representations of these from a subjective standpoint [3]. A MUSHRA-based listening test was designed to draw a correlation between the subjective quality assessment and the roughness of audio samples.

II. METHOD

Two instruments, cello and pipe organ, playing the following note intervals, were synthesised: semitone, tritone, perfect-five, and minor seventh across two octaves, to produce various roughness levels created by the interaction of harmonic content. A digital equaliser was used to implement acoustic manipulation of the spectrum in each audio file. These were fundamental and 2nd harmonic enhancements of 10dB for musical EQ, whereas fundamental and harmonic adjacent at 1kHz enhancement by 10dB for acoustic EQ. Subsequently, samples were compressed using a LAME encoder to 96kbps mp3 format. The listening test was conducted using the WebMUSHRA[4] interface in an ITU listening room. Samples were reproduced from a laptop via a pre-amp through Beyerdynamic DT990 PRO headphones.

III. RESULT

The data has 3 independent variables (intervals, instruments, audio conditions) and one dependent variable (BAQ). A 3-way rmANOVA (Table. 1) with post-hoc paired comparisons found significant differences with large effect sizes for *Audio Conditions*. Anchor was judged significantly lower than other conditions. A significant difference was also found in the EQ enhancements (Fig. 1). For the objective metrics of roughness and the BAQ scores, we found a negative correlation for musical roughness and a positive correlation for acoustic roughness. However, both are very weak, with a variance below 4% (Table. 2).

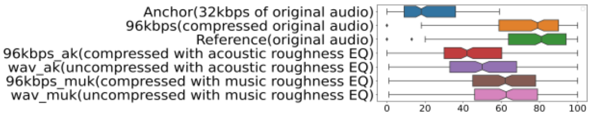


Figure 1. Boxplots of quality rating versus audio conditions collapsed across instruments and intervals of a figure caption.

TABLE I. TABLE TYPE STYLES

Effect Type	*F value*	*p value*	*Effect size/ Measure Scale*
Audio Conditions	31.907	**.000**	**.421(Large)**
Intervals	2.467	**.023**	.008(Small)
Instruments	5.903	**.030**	.004(Small)
Audio Conditions × Intervals	2.139	**.000**	.035(Small)
Audio Conditions ×Instruments	20.481	**.000**	**.073(Medium)**
Intervals × Instruments	0.949	.453	-
Audio Conditions × Instruments × Intervals	0.826	.774	-

TABLE II. CORRELATION OF AUDIO PERCEPTUAL RATING WITH THE ROUGHNESS VALUES (ACOUSTIC AND MUSICAL)

Factor	*R/R^2*	*p-value*
Acoustic roughness	.084(-) /.007	.000
Musical roughness	.190(.036)	.000

IV. CONCLUSION

This study suggests that manipulations of spectral content are perceived as significant alterations of quality, which are independent of musical interval or instrumentation. Further, correlations between roughness metrics and perceptual quality were found to be weak, suggesting that perceptual quality might be more closely related to other signal metrics. Future studies will examine other timbral features (brightness, sharpness, etc.) for possible explanations.

REFERENCES

[1] W. A. Sethares, ‘Local consonance and the relationship between timbre and scale’, J. Acoust. Soc. Am., vol. 94, no. 3, Aug. 1998.

[2] P. DANIEL and R. WEBER, ‘Psychoacoustical Roughness: implementation of an optimised model’, vol. 83, no. 1, 1997.

[3] M. Bosi and R. E. Goldberg, Introduction to Digital Audio Coding and Standards. Springer Science and Business Media, 2002.

[4] Schoeffler, Michael, et al. “WebMUSHRA — a Comprehensive Framework for Web-Based Listening Tests.” Journal of Open Research Software, vol. 6, 2018.

Adapting Beat Tracking Models for Salsa Music: Establishing a Baseline with a novel dataset

Antonin Rapini[1*] and Anna Jordanous[2]

[1*]School of Computing, University of Kent, United Kingdom, aplr3@kent.ac.uk
[2]School of Computing, University of Kent, United Kingdom

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. We benchmark the adaptability of three established models: BeatNet, Wavebeat, and Böck TCN, using our own newly introduced beat-annotated Salsa dataset and focusing on training methods that minimize the need for extensive annotated data. We find that, on Salsa music, models trained with popular datasets and fine-tuned with Salsa generally outperform models trained under other training conditions. This research not only establishes a baseline for beat tracking performance in Salsa music but also contributes to the broader goal of developing more universally adept music information retrieval systems.**

I. BACKGROUND

Beat tracking is the temporal identification of "beats", the basic rhythmic unit of a song. Although it is a skill that comes naturally for many people, automatic beat tracking, the programmatic identification of beats from audio data, is a highly complex task. Current state of the art beat tracking algorithms perform well on Western music [2] but often stumble when encountering the rhythmic intricacies of Salsa. Salsa is, a genre rich with syncopation and polyrhythms [4], features not often found in current available beat tracking datasets. Addressing this discrepancy is crucial for progress towards more universal music information retrieval.

II. OBJECTIVES

This study sets out to benchmark the adaptability of beat tracking models to Salsa music. It makes a point to emphasise methods that circumvent the need for large amounts of annotated data, a rare commodity that requires hours of tedious work and expert knowledge to be created.

III. METHODOLOGY

We assess the accuracy of three prominent beat tracking models: BeatNet [1], Wavebeat [2], and Böck TCN [3], on an unseen Salsa test dataset created for this study. The dataset contains 40 songs for a total of 2h53 of beat-annotated music. It will be made available at github.com/AntoninRap/Salsa-dataset. The models were trained under four distinct conditions: training on "other"* datasets of non-Salsa music, training exclusively on Salsa, an initial training on "other" datasets followed by fine-tuning on Salsa, and simultaneous training on both Salsa and "other" datasets.

IV. RESULTS

Our findings reveal a consistent pattern: models fine-tuned with Salsa music outperformed those trained on more generalised datasets. This trend persisted across models: Models trained on "other" datasets then fine-tuned on Salsa achieved the highest accuracy, followed by those trained on a combination of Salsa and "other" datasets, then models trained only on Salsa, with the least accuracy seen in models trained exclusively on "other" datasets. These results are still generally below what we see on "other" genres.

V. 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 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 for that goal, introduces a new beat-annotated dataset of Salsa music.

Table 1. average f-measure accuracy on unseen Salsa test dataset (10 songs) of three prominent beat tracking models under the four training conditions outlined in the Methodology section.

	F-measure accuracy			
Model	***Fine-tuned***	***Salsa + others***	***Salsa only***	***Others***
BeatNet	**0.844**	0.807	0.710	0.539
TCN	0.789	0.644	0.461	0.422
Wavebeat	0.739*	*	0.789	0.704*

REFERENCES

[1] Heydari, Mojtaba, Frank Cwitkowitz, and Zhiyao Duan. "BeatNet: A real-time music integrated beat and downbeat tracker." International Society for Music Information Retrieval (2021).

[2] Steinmetz, Christian J., and Joshua D. Reiss. "WaveBeat: End-to-end beat and downbeat tracking in the time domain." arXiv preprint arXiv:2110.01436 (2021).

[3] Matthew Davies, E. P., and Sebastian Böck. "Temporal convolutional networks for musical audio beat tracking." 2019 27th European Signal Processing Conference (EUSIPCO). IEEE, 2019.

[4] Laura M. Getz, Scott Barton & Lynn K. Perry (2021) Context-specific Knowledge Is the "Key" to Salsa Music, Auditory Perception & Cognition, 4:1-2, 14-32, DOI: 10.1080/25742442.2021.1964341

*"Other" refers to four datasets of mostly western music popular in the field of beat tracking: GTZAN, Rock, Ballroom and SMC

* Initial results with base model ("other") not trained by us, final results still pending

Planning The 2nd Cadenza Challenge for Improving Music for People With a Hearing Loss

Gerardo Roa Dabike[1*], Michael A. Akeroyd[2], Scott Bannister[3], Jon Barker[4], Trevor J. Cox[1], Bruno Fazenda[1], Jennifer Firth[2], Simone Graetzer[1], Alinka Greasley[3], Rebecca R. Vos[1] and William M. Whitmer[2]

[1]University of Salford , [2]University of Nottingham, [3]University of Leeds and [4]University of Sheffield, UK
[*]cadenzachallengecontact@gmail.com

***Abstract*— Cadenza is an ongoing EPSRC project that aims to improve music quality for those with a hearing loss. The project is running signal processing and machine learning challenges to address different listening issues and scenarios. During the first round, the challenge focused on non-causal music source separation to allow remixing for those with hearing loss. This fed into an ICASSP 2024 challenge, which has cross-talk from loudspeaker reproduction included. There are three potential arms to our upcoming 2024 challenge: causal audio source separation, lyric intelligibility, and loudness/dynamic range control. DMRN is an opportunity for the community to shape these arms.**

***Index Terms*—** challenge, source separation, lyric intelligibility, loudness, machine learning

I. Introduction

The Cadenza signal processing and machine learning challenges[1] are designed to grow a research community that embeds diverse listening when designing music processing algorithms. Figure 1 illustrates the general structure of these challenges. The blue box generates both the music to be enhanced and the reference music signal, while the green oval generates the listeners. In the pink "music enhancer" box, the music is enhanced for the listener. The "evaluation processor" is a fixed module that prepares the "enhanced music" for both objective evaluation and listener panel assessment.

The first challenge (CAD1) [1] addressed a non-causal demix-remix problem, focusing on rebalancing different components of the music to enhance the listening experience. An ICASSP 2024 SP challenge [2] followed the CAD1 problem. In this case, a cross-talk scenario was introduced, and the challenge welcomed both causal and non-causal submissions. Currently, the project is in the planning phase of the second challenge (CAD2), which involves three potential arms: causal audio source separation, lyric intelligibility and loudness/dynamic range control.

II. Potential arms of CAD2

Causal source separation. Impressive performances have been achieved in non-causal music source separation with deep neural networks (DNN) for separating pop/rock into vocals, bass and drums [3]. However, many listening scenarios involve live music that requires causal and low-latency approaches. Furthermore, hearing aid processing power is limited, which is challenging for very large DNNs. The plan is to run a causal, low-latency, cross-talk music-remixing scenario. The objective is to rebalance the live music captured by the hearing aid microphones for an enhanced music-listening experience. Current databases (e.g., [4, 5]) are not conducive to the listening habits of older adults with hearing loss [6]; to make a relevant challenge, datasets involving more appropriate genres are needed.

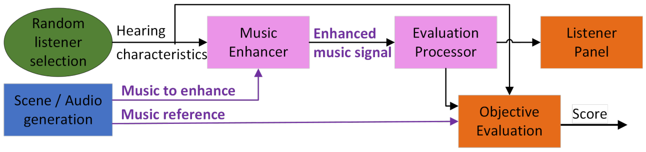


Figure 1: General structure of challenges in Cadenza Project.

The lack of *lyric intelligibility* is one of the more commonly cited issues for people with a hearing loss [6]. This challenge would aim to enhance the clarity of the lyrics while still improving the overall music quality. One of the difficulties in setting up this challenge is what objective metric to use. A pre-trained Automatic Lyric Transcription (ALT) system can be used to evaluate music that has been through a hearing aid and hearing loss processor. But what is the desired intelligibility? There are many factors to consider: (1) reducing the level of the other instruments might increase intelligibility but reduce enjoyment; (2) maintaining the timbre of the voice may be more important than intelligibility; (3) any enhancement needs to consider the gain being applied to compensate for the hearing loss, and (4) there is likely to be a large variation in the desired intelligibility depending on several factors, including the style of music (e.g., ballad vs metal).

Hearing aid users often struggle with music either being too quiet, too loud or not having enough dynamic variation [6]. In the *loudness/dynamic range* arm, the challenge will be to compress the dynamic range of the music to match the listeners' dynamic range of hearing without introducing distortions that impact their listening enjoyment. The training and evaluation dataset could be constructed by synthesizing from MIDI files. This method enables the creation of reference signals with distortion-free dynamic control simply by adjusting the velocity values of the instruments. That allows an intrusive objective metric to be used. The evaluation set for the listening panel will consist of real-world stereo recordings.

III. References

[1] G. Roa Dabike, S. Bannister, J. Firth, S. Graetzer, R. Vos, M. A. Akeroyd, J. Barker, T. J. Cox, B. Fazenda, A. Greasley, and W. Whit-

[*]Research supported by EPSRC [grant: EP/W019434/1].
Thanks to BBC, Google, Logitech, RNID, Sonova, Universität Oldenburg.

[1] https://www.cadenzachallenge.org

mer, “The first cadenza signal processing challenge: Improving music for those with a hearing loss,” in *Workshop on Human-Centric Music Information Retrieval*, 2023.

[2] G. Roa Dabike, M. A. Akeroyd, S. Bannister, J. Barker, T. J. Cox, B. Fazenda, J. Firth, S. Graetzer, A. Greasley, R. Vos, and W. Whitmer, “The Cadenza ICASSP 2024 Grand Challenge,” *arXiv preprint arXiv:2310.03480*, 2023.

[3] Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Défossez, M. Kim, W. Choi, C.-Y. Yu, and K.-W. Cheuk, “Music demixing challenge 2021,” *Frontiers in Signal Processing*, vol. 1, p. 808395, 2022.

[4] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “Musdb18-hq - an uncompressed version of musdb18,” 2019.

[5] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, “Moisesdb: A dataset for source separation beyond 4-stems,” 2023.

[6] A. Greasley, H. Crook, and R. Fulford, “Music listening and hearing aids: perspectives from audiologists and their patients,” *International Journal of Audiology*, vol. 59, no. 9, pp. 694–706, 2020.

Timbre Tools Hackathon; Timbre Tools for the Digital Instrument Maker

Charalampos Saitis, Haokun Tian, Jordie Shier and Bleiz Macsen Del Sette
Queen Mary University of London, UK haokun.tian@qmul.ac.uk

***Abstract*— This project will investigate how timbre can play an active role in designing sound synthesis and AI tools which empower everyone to partake in digital music instrument making and use. A research-through-design approach is adopted, based initially on a hackathon design activity.**

I. INTRODUCTION

Timbre is among the most evocative yet elusive attributes of music. Musicians can express emotions through timbre by manipulating the physical responses of their acoustic instrument. Yet timbre is conspicuously absent from the digital luthier's toolbox [1]. Synthesiser design is still primarily based on concepts from early analog and digital synthesis, or emulation of it, using more recent techniques. Furthermore, an audio engineer's workbench is still based mainly on classical tools like oscilloscopes and signal generators. Such technologies, whether commercial or open-source, value technical knowledge for *producing sounds* (e.g., pitch, rhythm) over perceptual knowledge for *designing timbres*. This effectively marginalises sonic cultures where timbre-based practice predominates or is equally important (e.g., didgeridoo, tabla, techno) from partaking in the music maker movement.

Timbre Tools proposes a techno-cognitive [2], timbre-first approach to digital musical instrument (DMI) design, leveraging the latest advances in music artificial intelligence (AI) to restore timbre to the same level of accessibility as pitch and rhythm. The premise of this project is to promote a learn-by-making approach: through creating digital instruments using flexible, open-ended AI powered tools for control and analysis of timbre, artists and makers without formal training can learn more about sound synthesis and AI, become more aware of timbre phenomena, and so create their own highly expressive bespoke instruments, widening participation in computing and AI and enriching the cultural landscape.

II. A HACKATHON APPROACH

We propose an exploratory design activity for ideating and prototyping timbre tools based on a 48-hour hackathon with audio developers, researchers, music technologists, and interaction designers [3]. Here the user is a blurring between instrument maker, composer, producer, and performer, as these roles tend to merge in music interaction design [2].

Hackathons are time-bounded, low-pressure collaborative events that present themselves as *observatories* of design thinking [4]. We will prompt participants to consider the role of timbre/ interacting with timbre in the development of a DMI. Drawing on the notion of problem and solution spaces, which form a general model of the design thinking process [4], our exploratory research questions are:

- *Exploring the problem space*: How do participants think about the concept of timbre in the design of tools for makers? What (collaborative) strategies do they use to conceptualize their design?
- *Exploring the solution space*: What tools are required by our participants to realize their concepts? How do they use the tools currently available to them to develop their concepts?

We will borrow from methods of rapid ethnography [5] (e.g., self-reports, workbooks) to observe the design thinking process of participants, using the answers to the above questions to inform our future work.

III. PREPARATORY WORK

We will use the Ethically Aligned Stakeholder Elicitation (EASE) framework [6] to identify all project stakeholders, including users, considering their level of Power and Interest in the project, with particular attention to those who may be inadvertently or marginally damaged by it.

Subsequently, we will interview users about their practice, the tools they use, and their needs as makers. We will aim to better understand how they think about the concept of timbre and what current practices and tools of DMI design constitute *timbre tools*. Interviews will inform a follow-up workshop, aiming to produce prompts for the hackathon.

REFERENCES

[1] C. Saitis, M. F. Torshizi, V. Preniqi, B. M. Del Sette, G. Fazekas, "When NIME and ISMIR Talk Timbre," in *Proceedings of Timbre 2023, 3rd International Conference on Timbre*, 2023, pp. 125–129.

[2] A. R. Jensenius, *Sound Actions: Conceptualizing Musical Instruments.* MIT Press, 2022.

[3] N. N. Correia and A. Tanaka, "Prototyping audiovisual performance tools: a hackathon approach," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2015, pp. 319–321.

[4] K. Gama, G. Valença, P. Alessio, R. Formiga, A. Neves, and N. Lacerda, "The Developers' Design Thinking Toolbox in Hackathons: A Study on the Recurring Design Methods in Software Development Marathons," *International Journal of Human–Computer Interaction*, vol. 39, 2023, pp. 2269-2291.

[5] D. R. Millen, "Rapid ethnography: time deepening strategies for HCI field research," in *Proceedings of the 3rd Conference on Designing Interactive Systems (DIS)*, 2000, pp. 280–286.

[6] A.-K. Kaila, P. Jääskeläinen, and A. Holzapfel, "Ethically Aligned Stakeholder Elicitation (EASE): Case Study in Music-AI," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2023.