DMRN+17: Digital Music Research Network

One-day Workshop 2022

Queen Mary University of London

Tuesday 20th December 2022

Chair: Simon Dixon

DRAFT - 19th Dec

Programme

Location: Arts 2 Theatre – QMUL Mile end campus
Zoom: https://qmul-ac-uk.zoom.us/j/89668766939

10:00	**Welcome – Andrew McPherson**
10:10	**KEYNOTE** "On generative modelling and iterative refinement", **Sander Dieleman- (Research Scientist at DeepMind)**
11:10	*Break (Coffee break)*
11:30	"Improving Chord Sequence Graphs with Transcription Resiliency and a Chord Similarity Metric", **Jeff Miller, Vincenzo Nicosia and Mark Sandler (Queen Mary University of London, UK)**
11:45	"Bringing the concert hall into the living room: digital scholarship of small-scale arrangements of large-scale musical works", **David Lewis and Kevin R. Page (University of Oxford e-Research Centre, UK)**
12:00	"Leveraging Music Domain Knowledge in Symbolic Music Modeling", **Zixun Guo and Dorien Herremans (ISTD, Singapore University of Technology and Design, Singapore)**
12:15	"Large-Scale Pretrained Model for Self-Supervised Music Audio Representation Learning", **Yizhi Li (University of Sheffield, UK), Ruibin Yuan (Beijing Academy of Artificial Intelligence, China, and Carnegie Mellon University, PA, USA) , Ge Zhang (Beijing Academy of Artificial Intelligence, China and University of Michigan Ann Arbor, USA) , Yinghao Ma (Queen Mary University of London, UK), Chenghua Lin (University of Sheffield, UK) , Xingran Chen (University of Michigan Ann Arbor, USA), Anton Ragni (University of Sheffield, UK), Hanzhi Yin (Carnegie Mellon University, PA, USA), Zhijie Hu (HSBC Business School, Peking University, China), Haoyu He (University of Tübingen & MPI-IS, Germany), Emmanouil Benetos (Queen Mary University of London, UK), Norbert Gyenge (University of Sheffield, UK), Ruibo Liu (Dartmouth College, NH, USA) and Jie Fu (Beijing Academy of Artificial Intelligence, China)**
12:30	Announcements
12:45	*Lunch - Poster Session*

12:45	*Lunch - Poster Session*
14:15	"Time-Frequency Scattering in Kymatio", **Cyrus Vahidi (Queen Mary University of London), Vincent Lostanlen, Han Han, Changhong Wang (Queen Mary University of London) and György Fazekas (Queen Mary University of London)**
14:30	"Working for the AI Man: Algorithmic Rents, Accumulation by Dispossession and Alien Power", **Hussein Boon (University of Westminster, UK)**
14:45	"Remarks on a Cultural Investigation of Abstract Percussion Instruments", **Lewis Wolstanholme and Andrew McPherson (Queen Mary University of London, UK)**
15:00	"Beat Byte Bot: a bot-based system architecture for audio cataloguing and proliferation with neural networks and Linked Data", **J. M. Gil Panal (E.T.S.I. Informática, University of Málaga, Spain and Luís Arandas (INESC-TEC, University of Porto, Portugal)**
15:15	*Break*
15:30	"Symmetries and Minima in Differentiable Sinusoidal Models", **Ben Hayes, Charalampos Saitis, György Fazekas (Queen Mary University of London)**
15:45	"Affordances of Generative Models of Raw Audio to Instrumental Practice and Improvisation", **Mark Hanslip (School of Arts and Creative Technologies, University of York)**
16:00	"Practical Text-Conditioned Music Sample Generation", **Scott H. Hawley (Belmont University, USA and Harmonai), Zach Evans, C.J. Carr (Harmonai), and Flavio Schneider (Harmonai).**
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

Sander Dieleman - Research Scientist at DeepMind

Tittle**: On generative modelling and iterative refinement**

Abstract:

The field of generative modelling has seen a significant upheaval in the past few years. In the audiovisual domain, adversarial approaches have been all but replaced by diffusion models, resulting in a step change in quality improvements and even mainstream adoption. In this talk, I will argue that iterative refinement is the key to generative modelling at scale, discuss some other innovations behind recent quality improvements, and consider the implications for audio and music generation.

Bio:

Sander Dieleman is a Research Scientist at DeepMind in London, UK, where he has worked on the development of AlphaGo and WaveNet. He obtained his PhD from Ghent University in 2016, where he conducted research on feature learning and deep learning techniques for learning hierarchical representations of musical audio signals. His current research interests include representation learning and generative modelling of perceptual signals such as speech, music and visual data.

Announcements

1. **AIM CDT 2023-2024 call open:** https://www.aim.qmul.ac.uk/apply/
   Based at the Centre for Digital Music at QMUL the AIM CDT offers 12+ Fully-funded PhD studentships to start September 2023. The call is open to UK Home and International student and covers fees and a stipend for four years. The application deadline is 31 January 2023

2. **COMPEL**- the Computer Music Preservation Electronic Library! http://compel-dev.vtlibraries.net/
   COMPEL is an electronic music database project from Virginia Tech. They collect data about people (composers, performers, and other contributors); compositions; specific performances; and instruments. The database is intended to serve performers, composers and researchers into the field of computer music

Posters

1	"Which car is moving? A listening approach using distributed acoustic sensor systems", **Chia-Yen Chiang and Mona Jaber (Queen Mary University of London)**
2	" YourMT3: a toolkit for training multi-task and multi-track music transcription model for everyone", **Sungkyun Chang, Simon Dixon and Emmanouil Benetos (Queen Mary University of London)**
3	"Supervised Contrastive Learning for Musical Onset Detection", **James Bolt and György Fazekas (Queen Mary University of London)**
4	"Computational Modelling of Expectancy-Based Music Cognition of Timbre Structures", **Adam Garrow and Marcus Pearce (Queen Mary University of London)**
5	"Self-supervised Learning for Music Information Retrieval" **Yinghao Ma and Emmanouil Benetos (Queen Mary University of London)**
6	"Performance Rendering for Automatic Music Generation Pipelines", **Tyler McIntosh and Simon Dixon (Queen Mary University of London)**
7	"Explainability in End-User Creative Artificial Intelligence", **Ashley Noel-Hirst and Nick Bryan-Kinns (Queen Mary University of London)**
8	" Real-time timbre mapping for synthesized percussive performance", **Jordan Shier (Queen Mary University of London), Andrew Robertson (Ableton), Andrew McPherson and Charalampos Saitis (Queen Mary University of London)**
9	"Machine Learning of Physical Models for Voice Synthesis", **David Südholt and Joshua Reiss (Queen Mary University of London)**
10	"Using Signal-informed Source Separation (SISS) principles to improve instrument separation from legacy recordings ", **Louise Thorpe, Emmanouil Benetos and Mark Sandler (Queen Mary University of London)**
11	"Personalised music descriptors: valuing user perspective", **Yannis Vasilakis (Queen Mary University of London), Rachel M Bittner (Spotify), Johan Pauwels (Queen Mary University of London)**
12	"Learning Music Representations using Coordinated based Neural Network", **Ningzhi Wang and Simon Dixon (Queen Mary University of London)**
13	"User-Driven Music Generation in Digital Audio Workstations", **Alexander Williams (Queen Mary University of London), Stefan Lattner (Sony SCL) and Mathieu Barthet (Queen Mary University of London)**
14	"Conditioning in Variational Diffusion Models for Audio Super-Resolution", **Chin-Yun Yu (Queen Mary University of London) Sung-Lin Yeh (University of Edinburgh) György Fazekas (Queen Mary University of London) Hao Tang (University of Edinburgh)**

Organizing Committee

Supported by UKRI AIM CDT

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

Katarzyna Adamska
Sara Cardinale
Franco Caspe
Ruby Crocker
Carlos De La Vega Martin
Bleiz MacSen Del Sette
Rodrigo Mauricio Diaz Fernandez
Andrew Edwards
Oluremi Falowo
Maryam Fayaz Torshizi
Yazhou Li
Jackson Loth
Teresa Pelinski Ramos
Sai Soumya Vanka
Christopher Winnard
Xiaowan Yi
Huan Zhang

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Improving Chord Sequence Graphs with Transcription Resiliency and a Chord Similarity Metric

Jeff Miller*[1] , Vincenzo Nicosia[2] and Mark Sandler[1]

[1]Centre for Digital Music, Queen Mary University of London, United Kingdom, j.k.miller@qmul.ac.uk
[2]School of Mathematical Sciences, Queen Mary University of London, United Kingdom

***Abstract*—**

We present an improved Chord Sequence Graph schema which is more resilient to discrepancies between sources. We also propose a musically-informed metric for measuring chord similarity within a collection.

I. CHORD SEQUENCE GRAPHS

Chord Sequence Graphs (CSGs) are a useful tool for modelling multiple chord sequences as a single, time-aligned, directed graph. [1] Chord sequences, collected from multiple sources in symbolic and/or audio domains, can reveal patterns and provide insights into harmonic practice which might remain undetected at smaller scales. CSGs can be applied to various musicological questions, such as 1) How do different versions of the same song (i.e. covers) compare harmonically? 2) How do different musicians harmonise a piece of music? 3) How does a musician vary their chord choices between performances of a song or piece? Finally - and importantly - 4) How similar or dissimilar are the harmonic sequences of two different pieces of music?

II. TRANSCRIPTION ISSUES AND CSG IMPROVEMENTS

Differences in chord descriptions, vocabularies, and transpositions can make it difficult to compare transcriptions from varied sources. A collection of notes might be described by different chord names, or chord extensions may be ignored and described as simpler chord types. Perhaps most problematically, similar or identical harmonic patterns can appear to be unrelated when played in different keys. [2]

We propose a revised CSG model which mitigates many of these concerns by replacing text-based chord labels with pitch class profile (PCP) vectors. A suitable chord vocabulary is chosen and mapped to PCPs; these become the data elements represented by nodes in the chord sequence graph, facilitating mathematical manipulation of graph data. Directed edges in the graph represent transitions from a chord (N) to the following chord in a sequence (N+1); the weights of nodes and edges represent the prevalence of chords and transitions at each relevant point across the original body of chord sequences.

III. CHORD SET COHERENCE METRIC

The introduction of a musically-informed method for describing the relative similarity within a set of chords would improve the CSG schema and augment the representative and analytic power of the model. Such a metric could describe the degree of similarity within a set of chords but need not describe the qualities of the chords themselves. Rather, it could describe the proximity of the chords relative to one another in a harmonically descriptive space. Each set would be considered in isolation; i.e., the relative position of one set to another within this space would be unimportant. We refer to this metric as the Chord Set Coherence (CSC).

The CSC metric could be applied to a chord sequence graph in several ways. Calculating the CSC for a time frame would produce a value describing all choices of chords at that point in the music. If every musician played the same chord at that point, the coherence would be high, indicating that the various chords chosen sounded similar to one another. A very low coherence would indicate that most musicians chose chords which did not sound like each other.

A profile of successive CSC values could be used to describe the harmonic content of the musical collection at a highly abstract analytical level. Likewise, by applying the CSC metric to a singular chord sequence, a coherence value could be calculated for an individual song. By aggregating these values into a coherence profile for a collection of songs with different chord sequences, one could analyse musicians or genres.

IV. REFERENCES

[1] J. Miller, V. Nicosia, and M. Sandler, "Discovering Common Practice: Using Graph Theory to Compare Harmonic Sequences in Musical Audio Collections," in *8th International Conference on Digital Libraries for Musicology*. New York, NY, USA: ACM, jul 2021, pp. 93–97. [Online]. Available: https://dl.acm.org/doi/10.1145/3469013.3469025

[2] J. Pauwels, K. O'Hanlon, E. Gómez, and M. B. Sandler, in *20th International Society for Music Information Retrieval Conference*, Delft, Holland.

*Research supported by EPSRC grant EP/R512072/1 and the British Broadcasting Corporation through an Industrial CASE studentship in collaboration with the BBC Audio Research Partnership.

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Bringing the concert hall into the living room: digital scholarship of small-scale arrangements of large-scale musical works

David Lewis and Kevin R. Page

University of Oxford e-Research Centre, UK
[david.lewis|kevin.page]@oerc.ox.ac.uk

***Abstract*— We present a study into nineteenth century arrangements of operatic and orchestral works for domestic use, supported by tools designed to support digital musicological research. These tools are built on web standards – Linked Data (particularly Web Annotations), IIIF and MEI – along with a new ontology designed to support the annotation of musical materials that appear in different forms across different resources.**

***Index Terms*—** Digital musicology, Linked Data, Ontologies, IIIF, Music Encoding

Before the twentieth century, and the rise of both affordable concert tickets and technological innovations in music recording and broadcast, people's access to performances of orchestral music and opera was limited by geographical and financial factors. Even for those who lived within reach of concert venues and for whom the cost of tickets was not a barrier, repeated, on-demand listening to any individual work was impossible.

These factors gave rise to a huge market in musical arrangements (or 'translations', as Beethoven called them), where orchestral works were reworked for fewer instruments. These reductions also allowed audience members to prepare for a concert more thoroughly, by playing through the music they would be hearing in advance. More dramatic changes to the source were also common in arrangements, however, including prècis, setting the melody to words, or freely composing a theme and variations or a fantasia around the source.

Despite their significance in how music was heard and understood – and financially in the music industry itself – arrangements are relatively little-studied. The lower status of arranged works has affected musicological discourse and, to some extent, library acquisitions, leaving suitable materials for study widely distributed and they have been seldom collected into scholarly editions.

With the rise of library digitisation and publication through IIIF, it has become practical to support the painstaking task of finding, comparing and analysing these arrangements, using decentralised tools, built on web standards such as Linked Data. Starting with a musicological investigation of musical arrangements in review and edition in *The Harmonicon*, a music periodical of the 1820s and 30s, we illustrate how research of this kind has been supported using an application developed for the Beethoven in the House project. This application allows a scholar to record observations about the musical and practical decisions being made by editors whilst, in the process, creating a reusable Research Object[1] that can be published with any scholarly outcomes. Saving these annotations in Solid Pods[2] allows observations about public resources to be kept privately by a scholar, or published at a time and in a way that matches their needs.

To support musicological investigation of different versions of a piece of music, we have used the Music Annotation Ontology[3], which uses a FRBR-based model to allow comparable passages in different arrangements – evidenced by images, MEI editions or recordings – to be addressed as a single conceptual entity. These can be used simply to indicate where parallel passages occur, for example to support side-by-side browsing, but also to attach annotations at the appropriate level of abstraction to discuss shared musical concepts or instrumentation decisions.

I. ACKNOWLEDGMENTS

This research was undertaken by the project 'Beethoven in the House: Digital Studies of Domestic Music Arrangements', supported in the UK by the Arts and Humanities Research Council (AHRC), project number AH/T01279X/1. We gratefully acknowledge the contributions of our project partners at the University of Paderborn, and the Beethoven-Haus Bonn, and in particular of Johannes Kepper and Mark Saccomano for development of the annotation application.

II. REFERENCES

[1] K. R. Page, B. Fields, D. D. Roure, T. Crawford, and J. S. Downie, "Capturing the workflows of music information retrieval for repeatability and reuse," *Journal Intelligent Information Systems*, vol. 41, no. 3, pp. 435–459, 2013.

[2] D. M. Weigl, W. Goebl, A. Hofmann, T. Crawford, F. Zubani, C. C. S. Liem, and A. Porter, "Read/write digital libraries for musicology," in *7th International Conference on Digital Libraries for Musicology*. New York, USA: Association for Computing Machinery, 2020, p. 48–52.

[3] D. Lewis, E. Shibata, M. Saccomano, L. Rosendahl, J. Kepper, A. Hankinson, C. Siegert, and K. Page, "A model for annotating musical versions and arrangements across multiple documents and media," in *Proceedings of the 9th International Conference on Digital Libraries for Musicology*. New York, USA: Association for Computing Machinery, 2022, p. 10–18.

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Leveraging Music Domain Knowledge for Symbolic Music Modeling

Zixun Guo*[1] and Dorien Herremans[1]

[1]ISTD, Singapore University of Technology and Design, Singapore, nicolas.guozixun@gmail.com

***Abstract*—Compared to the absolute musical attributes (e.g., pitch), the relative musical attributes (e.g., interval) contribute even more to human's perception of musical motifs and structures. To represent both attributes in a shared embedding space, we propose the Fundamental Music Embedding (FME) for symbolic music based on a bias-adjusted sinusoidal encoding within which the fundamental musical properties are explicitly preserved. Taking advantage of the proposed FME, we further propose a novel attention mechanism based on the relative index, pitch and onset embeddings (RIPO attention) such that the musical domain information can be integrated into symbolic music models. Experimental results show that our proposed model outperforms the state-of-the-art transformers in melody completion and generation tasks both subjectively and objectively.**

I. METHOD

We represent a basic symbolic music sequence with length n using the vector representation of pitch, duration and onset: $P : \{p_1, ..., p_n\}$, $D : \{d_1, ..., d_n\}$, $O : \{o_1, ..., o_n\}$. More generally, these event tokens are defined as fundamental music tokens (FMTs) $F : \{f_1, ..., f_n\}$. The relative attribute of FMT is defined as dFMT: ΔF. The proposed Fundamental Music Embedding (FME) $FME : \mathcal{R}^{n\times 1} \rightarrow \mathcal{R}^{n\times d}$ is shown in Eq 1-3 where B and $[b_{sin_k}, b_{cos_k}]$ represent a base value and a trainable bias vector respectively. The embedding function for dFMT is defined as the Fundamental Music Shift (FMS) and is shown in Eq 4-5. Several fundamental music properties can be observed in the embedding space. For instance, the L2 distance between pitches in the FME conveys the musical interval (but this cannot be guaranteed using one-hot or word embedding).

We further propose a novel attention mechanism in Fig 1 that uses relative index, pitch and onset embeddings (RIPO attention) and incorporates FME. RIPO attention is able to effectively tackle the issue mentioned in [2] that the relative pitch and onset embeddings can not be efficiently utilized beyond the JSB chorale dataset.

$$w_k = B^{-\frac{2k}{d}} \tag{1}$$

$$P_k(f) = [sin(w_k f) + b_{sin_k}, cos(w_k f) + b_{cos_k}] \tag{2}$$

$$FME(f) = [P_0(f), ..., P_k(f), ...P_{\frac{d}{2}-1}(f)] \tag{3}$$

*Zixun Guo is a Senior Research Assistant funded by Singapore's MOE under Grant No. MOE2018-T2-2-161; The full paper of this work [1] has recently been accepted at AAAI 2023.

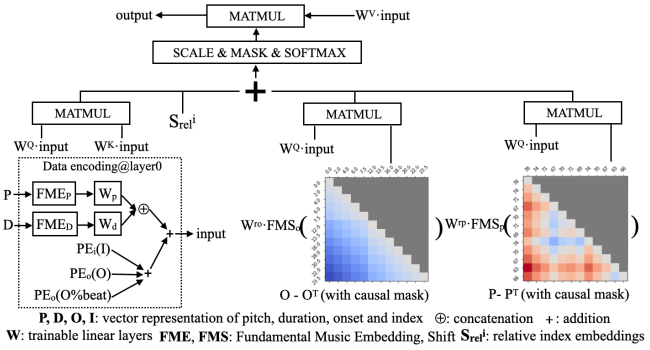

Figure 1: RIPO attention layer.

$$A_k(\Delta f) = [sin(w_k \Delta f), cos(w_k \Delta f)] \tag{4}$$

$$FMS(\Delta f) = [A_0(\Delta f), ..., A_k(\Delta f), ...A_{\frac{d}{2}-1}(\Delta f)] \tag{5}$$

II. RESULTS

We compare our model with the state-of-the-art (SOTA) music models: Music Transformer [2] and Compound Word Transformer [3] in a melody completion and generation task. Table 1 shows that our model outperforms the SOTA models both using analytical metrics as well as in a listening test. Readers are encouraged to check the entire evaluation section in the original paper [1].

Table 1: Model comparison. MT, LT, WE, OH, KL, ISR, AR stand for music transformer and linear transformer, word embedding, one-hot encoding, KL divergence, in-scale ratio, and arpeggio ratio respectively.

		objective evaluation				subjective evaluation
Model	$test\_loss$	KL_p	KL_d	ISR	AR	$overallrating$
MT+OH	2.405	0.015	0.035	0.969	0.038	2.80
LT+WE	2.943	0.026	0.040	0.971	0.032	-
RIPO+FME (ours)	**2.367**	**0.011**	**0.024**	**0.981**	**0.049**	**3.57**

III. REFERENCES

[1] Z. Guo, J. Kang, and D. Herremans, "A domain-knowledge-inspired music embedding space and a novel attention mechanism for symbolic music modeling," in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2023.

[2] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," in *Proc. of the Int. Conf. on Learning Representations*, 2019.

[3] W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh, and Y.-H. Yang, "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proc. of the AAAI Conf. on Artificial Intelligence*, 2021.

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Large-Scale Pretrained Model for Self-Supervised Music Audio Representation Learning

Yizhi Li[1], Ruibin Yuan[2,4*], Ge Zhang[2,5*], Yinghao Ma[3*], Chenghua Lin[1†], Xingran Chen[5], Anton Ragni[1], Hanzhi Yin[4], Zhijie Hu[6], Haoyu He[7], Emmanouil Benetos[3], Norbert Gyenge[1], Ruibo Liu[8], Jie Fu[2†]*

[1]Department of Computer Science, University of Sheffield, UK {yizhi.li, c.lin}@sheffield.ac.uk
[2]Beijing Academy of Artificial Intelligence, China fujie@baai.ac.cn
[3]Centre for Digital Music, Queen Mary University of London, UK yinghao.ma@qmul.ac.uk
[4]School of Music, Carnegie Mellon University, PA, USA
[5]University of Michigan Ann Arbor, USA
[6]HSBC Business School, Peking University, China
[7]University of Tübingen & MPI-IS, Germany
[8]Dartmouth College, NH, USA

***Abstract*— Self-supervised learning technique is an under-explored topic for music audio due to the challenge of designing an appropriate training paradigm. We hence propose MAP-MERT, a large-scale music audio pre-trained model for general music understanding. We achieve performance that is comparable to the state-of-the-art pre-trained model Jukebox using less than 2% of parameters.**

***Index Terms*—** Self-supervised learning, Music representation learning, Music information retrieval

I. INTRODUCTION

Deep learning is undergoing a paradigm shift with the rise of large-scale pre-trained models. In recent years, self-supervised learning (SSL) has achieved significant results in domains like computer vision, natural language processing, and speech processing. SSL leverages large-scale unlabelled data to obtain general representations, which could benefit a wide range of resource-restricted downstream tasks.

Although such a large-scale pre-training paradigm is of potential to improve annotation-limited music information retrieval (MIR) tasks, it is not well-studied in the community. Jukebox, the state-of-the-art SSL model learns music representations by reconstructing the raw audio [1, 2]. But it can barely be fine-tuned or efficiently adapted to more downstream tasks due to the enormous number of 5 billion parameters. To this end, we propose a novel representation learning method for music understanding.

Inspired by HuBERT [4], we obtain discrete pseudo labels by K-Means to conduct the mask prediction pre-training. Apart from only focusing on distinguishing sound textures like HuBERT, we use additional Chroma-based pseudo labels and design a CQT reconstruction task to help Mu-BERT learn the significant pitch information for music tasks. Moreover, we explore and analyse masking strategies and data augmentation techniques appropriate for music audio pre-training. To conclude, the aim and potential innovations of this work include:

1. developing self-supervised methods for music understanding;
2. providing a general music pre-trained model with trainable size; and
3. establishing a user-friendly and extendable MIR benchmark.

Approach	MTT		GTZAN	GS	EMO		Average
	AUC	AP	Acc	Score	$R2_{Arousal}$	$R2_{Valence}$	
CHOI	89.7	36.4	75.9	13.1	67.3	43.4	51.9
MUSICNN	90.6	38.3	79.0	12.8	70.3	46.6	53.7
CLMR	89.4	36.1	68.6	14.9	67.8	45.8	50.8
Music2Vec [3]	89.5	35.9	76.6	50.1	69.4	57.4	63.2
Jukebox (5B)	**91.5**	**41.4**	79.7	66.7	**72.1**	**61.7**	**69.9**
MERT (90M)	90.8	38.4	**80.7**	**67.0**	71.2	52.1	66.7

Table 1: Preliminary Results on MIR Tasks. The baseline results (except Music2Vec) and probing protocol are adopted from JukeMIR [2]. All results are produced with probing settings. Our model with 768-D representations under the probing setting achieves performances comparable to the SOTA Jukebox with 4800-D representations on the auto-tagging, genre classification, key detection and emotion regression tasks.

II. REFERENCES

[1] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[2] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," *arXiv preprint arXiv:2107.05677*, 2021.

[3] Y. Li, R. Yuan, G. Zhang, Y. MA, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He, *et al.*, "Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning," in *ISMIR 2022 Hybrid Conference*, 2022.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

* The authors contributed equally to this work.
† Corresponding authors.

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Time–Frequency Scattering in Kymatio

Cyrus Vahidi*[1], Vincent Lostanlen, Han Han, Changhong Wang†[2] and György Fazekas[1]

[1]Centre for Digital Music, Queen Mary University of London, c.vahidi@qmul.ac.uk
[2]LS2N, CNRS, Nantes Université, École Centrale Nantes, France

***Abstract*— We present a differentiable and GPU-enabled implementation of time–frequency scattering in Kymatio, an open-source package for wavelet scattering and deep learning in Python. We outline Kymatio's architecture and the algorithm's key software implementation details and encourage its use in music signal processing.**

I. TIME–FREQUENCY SCATTERING IN KYMATIO

Joint time–frequency scattering (jTFS) is a convolutional operator in the time-frequency domain. It extracts joint spectrotemporal modulations from audio signals at multiple resolutions, serving as a computational surrogate for auditory similarities between sounds [1]. A new Python-based implementation of jTFS was introduced in [2], highlighting its potential for auditory modelling and differentiable computing in MIR-related tasks. We present an extension to the Kymatio software to support jTFS for 1D signals.

Kymatio[1] [3] supports differentiable wavelet scattering, with a focus on portability across modern deep learning frameworks (Torch, TensorFlow, Jax, Keras, Numpy). Each framework has its particular lexis for tensor operations, hence Kymatio makes use of *backends* that implement framework-specific primitive operations. Kymatio's *core* scattering functionality is backend-agnostic and exposed to the user via a consistent *frontend* API.

To implement jTFS, we reuse Kymatio's existing architecture and core routines for filterbank design and scattering path computation. Yet previously, the API was restricted to a single type of filterbank; the specification of a filterbank's parameters and its generation were coupled. Definition of alternative filterbanks was challenging without a major redesign of the core routines. We have redesigned the API to be agnostic to filterbank design. Users can define a filterbank function that yields centre frequencies ξ and bandwidths σ for filterbank construction. Similarly, we yield the computed scattering paths from a generator to allow for conditional computation of scattering paths.

In contrast to time scattering, it is necessary to compute the first-order scattering transform in a breadth-first manner in order to perform frequency scattering. To minimise the memory overhead incurred by storage of intermediate signals, we compute second-order temporal scattering coefficients depth-first prior to subsequent frequency scattering.

Listing I shows an example of the user interface to the core jTFS functionality. We expose necessary parameters for most applications while retaining user-friendliness. We refer the reader to the Kymatio documentation for explanations of the keyword arguments. Our implementation supports both vector (`"time"`) and 3D (`"joint"`) output formats. These are a 2-tuple `(path, time)` and 3-tuple `(path, freq, time)` respectively. The latter allows jTFS to serve as a feature extractor for 2D convolutional neural networks.

```
import torch
from kymatio.torch import TimeFrequencyScattering
jtfs = TimeFrequencyScattering(J=8, J_fr=6, Q=12,
                               shape=8192, Q_fr=1,
                               format="time")
Sx = jtfs(torch.randn(8192))
print(Sx.shape) # (122, 32)
```

II. CONCLUSION

The Kymatio software offers a user-friendly, robust and efficient implementation of the time–frequency scattering transform. It is portable across modern deep learning frameworks, paving the way for usage in music information retrieval and auditory perception research. In future work we will investigate efficiency optimisations and alternative wavelet filterbanks.

III. REFERENCES

[1] V. Lostanlen, C. El-Hajj, M. Rossignol, G. Lafay, J. Andén, and M. Lagrange, "Time–frequency scattering accurately models auditory similarities between instrumental playing techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–21, 2021.

[2] J. Muradeli, C. Vahidi, C. Wang, H. Han, V. Lostanlen, M. Lagrange, and G. Fazekas, "Differentiable time-frequency scattering in Kymatio," *arXiv preprint arXiv:2204.08269*, 2022.

[3] M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, *et al.*, "Kymatio: Scattering transforms in Python." *J. Mach. Learn. Res.*, vol. 21, no. 60, pp. 1–6, 2020.

*Researcher at the UKRI CDT in AI and Music, supported jointly by the UKRI (grant number EP/S022694/1) and Music Tribe
†Atlanstic2020 grant on Trainable Acoustic Sensors (TrAcS)
[1]https://github.com/kymatio/kymatio

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Working for the AI Man: Algorithmic Rents, Accumulation by Dispossession and Alien Power

Hussein Boon*

Music Department, University of Westminster, United Kingdom, h.boon@westminster.ac.uk

***Abstract*— This paper discusses AI, its effects on the production condition of users and conditions of alienation made possible by AI music applications.**

I. INTRODUCTION

This paper places the AI application AIVA under the Marxist lens as a means to stimulate discussion of capitalist practices in AI and the humans behind them. Esling's keynote DMRN+15 [1] described some concerns about AI as somehow "unhinged". Yet Esling ignores how AI changes the conditions of production which is a core aspect of capitalism usually leading to the establishment of 'new' practices, usually for enrichment purposes. Esling's reliance upon slippery terms such as "natural continuity" in the Hobsbawmian sense is an "attempt to establish continuity with a suitable historic past" where claims of "continuity with it [this suitable historic past] is largely factitious." [2]

II. ALIENATION 1: NOT OWNING THE MEANS OF PRODUCTION

Fisher states the process of alienation as "both a pre-condition for exploitation and the result thereof. "[3] The charging of subscriptions (algorithmic rents [4]), removes the ability of users to own the means of production especially where AIVA also restricts users at the free and standard tiers from monetizing their labor efforts. Therefore AIVA changes the production condition at those levels where the only solution for users is higher rents where these restrictive conditions do not apply.

III. ALIENATION 2: CONTROLLING PRODUCTION

AIVA also benefits from human user production activity, at all subscription levels as this work is also ingested by the machine, thus, the machine learns from this unpaid labor. In effect this makes it difficult for any user to obtain a novel compositional or production advantage due to these improvements reflected throughout the machine as a 'benefit' for all users, including the company.

IV. ALIENATION 3: ACCUMULATION BY DISPOSSESSION

Free and standard users also do not own their own copyrights which AIVA assigns to their AI. Therefore, free and standard rate users are dispossessed of their copyrights, which are accumulated for revenue generation opportunities such as synchronization, also revealing AIVA as a competitor in the creative space.

V. ALIENATION 4: COMPOSING WITH INFLUENCE

The final act of dispossession is via AIVA's composing with influence (CWI) option. CWI is AIVA's means to deal with issues observed by Huang et al. [5]. Users at all tiers are encouraged to use CWI which either alienates existing rights holders, or for users to alienate their own rights by using their own piece as influence. The overall effect is that CWI music at the free or standard tiers will be owned by AIVA and all CWI music will also be ingested by the machine for the 'benefit' of all i.e. AIVA.

VI. ALIEN POWER: CONCLUSION

This paper's examples satisfy Braverman's definition of alien power "the machines must not be the property of the producer, nor of the associated producers, but of an alien power." [6] There is sufficient evidence that these changes to production conditions are significant, neither neutral nor "unhinged".

REFERENCES

[1] C4DM - Centre for Digital Music. (2020). 'Creativity in the Era of Artificial Intelligence - DMRN+15 2020. Prof. Philippe Esling', *YouTube*, 27 January. URL: https://www.youtube.com/watch?v=6jzd9-SN6uc.

[2] E. Hobsbawm, "Introduction: Inventing Traditions," in *The Invention of Tradition*, E. Hobsbawm and T. Ranger, Eds. Cambridge: Cambridge University Press (Canto Classics), 2012, pp. 1–14.

[3] E. Fisher, "How Less Alienation Creates More Exploitation? Audience Labour on Social Network Sites," in *Marx in the Age of Digital Capitalism*, C. Fuchs and V. Mosco, Eds. Leiden: Brill, 2016, 178–203.

[4] H. Boon, "Cyborg Composers: AI as Collaborative Assistant, as Creator and as Competitor." *Library Music in Audiovisual Media - RMA.* University of Leeds, 15 - 16 Sep 2022.

[5] C.-Z. A. Huang, H. V. Koops, E. Newton-Rex, M. Dinculescu, and C. Cai, 'AI Song Contest: Human-AI Co-Creation in Songwriting', *Magenta*, 13 October, 2020.

[6] H. Braverman, "Technology and capitalist control" in *The Social Shaping of Technology* 2nd ed. D. Mackenzie, and J. Wajcman, Eds. Buckingham: Open University Press, 1999, pp. 158–60.

*Hussein Boon is with the University of Westminster and is a member of the Black Music Research Unit (BMRU).

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Remarks on a Cultural Investigation of Abstract Percussion Instruments

Lewis Wolstanholme* and Andrew McPherson

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

***Abstract*— As a part of my larger body of work, I have been developing numerous digital tools to aid in the creative exploration and simulation of abstract percussion instruments. To further understand the creative implications of these abstract percussion instruments, I have curated a study which encourages participants to exhibit their practical intuition towards both familiar and unfamiliar percussive sonorities. The aim of this approach is to ascertain a deeper understanding of the cultural and semantic distance between abstract digital musical instruments and the more traditional musical instruments that they represent.**

***Index Terms*—** digital musical instruments, arbitrarily shaped drums, creative research, practice based research

I. BACKGROUND & MOTIVATION

As both an artist and a researcher, I am often working towards the fruition of my own creative ideas, utilising and developing upon a reflexive [1] and process [2] oriented understanding. This self driven approach seems to underpin a majority of the current developments and explorations of digital musical instruments [3], and is interwoven amongst many contemporary artistic narratives [4]. The rational behind the study presented here is to extend beyond this viewpoint, whilst examining the sentiments and idioms of the digital musical instruments I have created from outside of my own cultural perspective. This project aims towards generating *research through art* and *research for art* [5] which serves to recontextualise and redefine my own epistemic directives in design and creativity.

II. METHODOLOGY

For this study, participants were approached by invite only to compose a piece of music using a sample library of 2000 arbitrarily shaped drums [6]. The drums were all of varying shapes and size, and were generated using a two-dimensional physical model, with each drum being sampled in five unique strike locations. Participants were encouraged to compose freely, and in a way that felt familiar and natural to them, using whatever means they felt most comfortable with. The only limitation imposed upon them were the five briefs shown below, which served to focus their attention primarily on composing music with percussion instruments.

1. Compose only with the sounds contained in the sample pack, without employing audio manipulation or effects.
2. Compose only with the sounds contained in the sample pack, allowing for audio manipulation or effects.
3. Compose with the sounds contained in the sample pack, without the use of audio manipulation or effects, but with the inclusion of any other live/acoustic percussion performance.
4. Compose with the sounds contained in the sample pack, where audio manipulation and effects are allowed, with the inclusion of a live/acoustic percussion performance.
5. Compose with the sounds contained in the sample pack and any other percussive sounds or effects that the practitioner desires.

Once each participant had finished their composition, they were interviewed for approximately one hour. The first half of the interview was dedicated to the participant's cultural impression of percussion instruments in general, questioning their semantic, idiomatic and functional role in an open musical context. The second half of the interview featured a critical examination of the percussion instruments used as part of this study, in comparison with the formal characteristics previously outlined. The interview would then conclude with an analysis of each participant's composition, assessing its material and form alongside the practicalities and affordances used to create it.

*Research supported by UK Research and Innovation, EPSRC grant EP/S022694/1.

III. REFERENCES

[1] H. Borgdorff, *The Conflict of the Faculties: Perspectives on Artistic Research and Academia*. Leiden, Netherlands: Leiden University Press, 2012.

[2] A. N. Whitehead, *Process and Reality*, D. R. Griffin and D. W. Sherburne, Eds. New York, NY: The Free Press, 1978.

[3] F. Morreale and A. Mcpherson, "Design for longevity: Ongoing use of instruments from NIME 2010-14," in *17th International Conference on New Interfaces for Musical Expression (NIME)*, Copenhagen, Denmark, 2017.

[4] C. Molitor and T. Magnusson, "Curating experience: Composition as cultural technology – a conversation," *Journal of New Music Research*, vol. 50, no. 2, pp. 184–189, 2021.

[5] C. Frayling, "Research in art and design," *Royal College of Art Research Papers*, vol. 1, no. 1, pp. 1–5, 1994.

[6] L. Wolstanholme, "kac_drumset: A dataset generator for arbitrarily shaped drums," Zenodo, 2022.

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Beat Byte Bot: a bot-based system architecture for audio cataloguing and proliferation with neural networks and Linked Data

J. M. Gil Panal[1] and Luís Arandas[2]

[1]E.T.S.I. Informática, University of Málaga, Spain, josgilpan@uma.es
[2]INESC-TEC, University of Porto, Portugal

***Abstract*— This article presents a preliminary study on the usefulness of Web 3.0 accessible information regarding musical instrument recognition (MIR) and digital audio analysis through neural networks (NN). We present a bot-based modular architecture prototype inspired by previous research [1], where we develop a complementary module composed of a two-part process: 1) a deep learning (DL) sound processing algorithm; and 2) a service for audio tagging that can be used as a platform to study model usage. We integrate all of the code through a Virtuoso Universal Server instance to deal with SPARQL queries and contextualise positive or negative results of (e.g.) prediction.**

***Index Terms*—** Bot, Telegram, Semantic Web, Deep Learning

I. INTRODUCTION

One of the biggest challenges in Web 3.0 is to keep up with accurate and truthful behaviour in terms of information [2, 3]. Millions of data points need to be ordered and structured for better understanding, of both humans and divergent algorithms [4]. This research tries to assess quality, validity and reliability of information obtained through deep NN on audio signals [5]. For this we built on the already proposed architecture (Fig. 1) Beat Byte Bot [1] and developed it into an open source platform[1] that can be used to study deep NN accuracy [6, 7]. The architecture uses Telegram Bot API [8, 9] and allows for a bot to search and analyse audio files on group chats. With this new research, we are now able to run those files through *Tensorflow.js* [10] audio models and link their usage and generated data to a semantic knowledge graph [11]. We also document experiences with *mtt-musicnn* and *mtt-vgg* [12] where we integrate with Virtuoso [13] for Linked Data semantic queries using a SPARQL endpoint [14].

II. METHODOLOGY

Our methodology starts by running an uploaded audio file into the system where inference occurs and leads to storing the information generated on a graph with custom objects to deal with the ontology [15, 16]. Both of the used models are compatible with the library *Essentia.js* [17] and trained with the MagnaTagATune dataset, the one used for our testing [18]. Different subsets of samples within the MagnaTagATune audio dataset have been validated for content and comparing the output of the classifiers with the manual annotations made on the original dataset. Each file processed by the system has been translated into a new entry in the knowledge graph, so that each audio track has an associated prediction, a series of labels and parameters that provide information of the inference process [19].

III. RESULTS AND CONCLUSIONS

Through script testing we verified the direct influence of manual annotations on the dataset and how certain errors in sound identification can translate into inaccuracies in the models [20]. Those errors can cause one instrument classification failure, e.g., between the cello and the violin. We found that the accuracy of both models tends to be equivalent, albeit inaccuracy of classification between on the harp (40%) and the guitar. We also emphasise that the *semantic gap* phenomenon also happens when classifying human voice in this context. Trying to tackle issues of accuracy we propose to use within our architecture Linked Data methodologies to derive a more meaningful and semantic relationship with the audio source. By previous results, we believe this can lead to more precise outcomes [21].

IV. FUTURE RESEARCH

We plan to amplify the current bot-based ecosystem to include cross validation through Linked Open Data databases to complement NN misleading results [22]. That can be tamed by providing more information in a graph, as more elaborated analysis and vocabularies are required. In this regard we build on the philosophy of reliable entities such as MusicBrainz or DBPedia [23] [24], and the Solid "Social Linked Data" project that can be taken into account for future experiments in the same line [25].

[1]https://github.com/gilpanal/b8b_virtuoso

V. REFERENCES

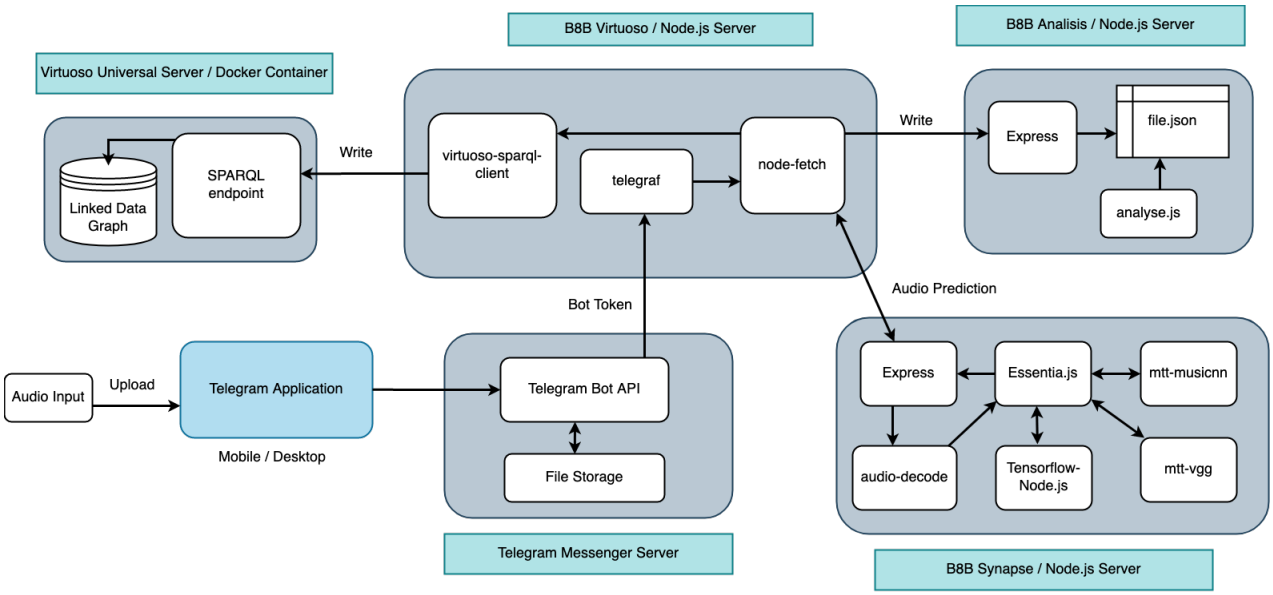

Figure 1: Block diagram to represent the system architecture, built on top of previous research. In grey are the main modules of the system, including the Telegram server, in blue the user connection point -mobile or desktop- and in green the infrastructure for each module.

[1] G. Panal and L. Arandas, "Beat byte bot: A chatbot architecture for web-based audio management," in *Proceedings of the 11th Workshop on Ubiquitous Music (UbiMus 2021)*. g-ubimus, 2021, pp. 72–82.

[2] A. Rettinger, U. Lösch, V. Tresp, C. d'Amato, and N. Fanizzi, "Mining the semantic web," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 613–662, 2012.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web. a new form of web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 5, 2001.

[4] V. Tresp, M. Bundschus, A. Rettinger, and Y. Huang, "Towards machine learning on the semantic web," in *Uncertainty reasoning for the Semantic Web I*. Springer, 2006, pp. 282–314.

[5] P. Kasnesis, N.-A. Tatlas, S. A. Mitilineos, C. Z. Patrikakis, and S. M. Potirakis, "Acoustic sensor data flow for cultural heritage monitoring and safeguarding," *Sensors*, vol. 19, no. 7, p. 1629, 2019.

[6] A. Solanki and S. Pandey, "Music instrument recognition using deep convolutional neural networks," *International Journal of Information Technology*, pp. 1–10, 2019.

[7] V. S. Kadandale, "Musical instrument recognition in multi-instrument audio contexts," Ph.D. dissertation, MSc thesis, Universitat Pompeu Fabra, 2018.

[8] T. Khaund, M. N. Hussain, M. Shaik, and N. Agarwal, "Telegram: Data collection, opportunities and challenges," in *Annual International Conference on Information Management and Big Data*. Springer, 2020, pp. 513–526.

[9] D. Korotaeva, M. Khlopotov, A. Makarenko, E. Chikshova, N. Startseva, and A. Chemysheva, "Botanicum: a telegram bot for tree classification," in *2018 22nd Conference of Open Innovations Association (FRUCT)*. IEEE, 2018, pp. 88–93.

[10] D. Smilkov, N. Thorat, Y. Assogba, C. Nicholson, N. Kreeger, P. Yu, S. Cai, E. Nielsen, D. Soegel, S. Bileschi, *et al.*, "Tensorflow. js: Machine learning for the web and beyond," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 309–321, 2019.

[11] Š. Čebirić, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika, "Summarizing semantic graphs: a survey," *The VLDB journal*, vol. 28, no. 3, pp. 295–327, 2019.

[12] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," *arXiv preprint arXiv:1909.06654*, 2019.

[13] O. Erling and I. Mikhailov, "Virtuoso: Rdf support in a native rdbms," in *Semantic web information management*. Springer, 2010, pp. 501–519.

[14] S. Ferré, "Sparklis: a sparql endpoint explorer for expressive question answering," in *ISWC posters & demonstrations track*, 2014.

[15] F. N. Al-Aswadi, H. Y. Chan, and K. H. Gan, "Automatic ontology construction from text: a review from shallow to deep learning trend," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 3901–3928, 2020.

[16] Ş. Kolozali, M. Barthet, G. Fazekas, and M. Sandler, "Automatic ontology generation for musical instruments based on audio analysis," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 10, pp. 2207–2220, 2013.

[17] P. Alonso-Jiménez, D. Bogdanov, J. Pons, and X. Serra, "Tensorflow audio models in essentia," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 266–270.

[18] D. Wolff and T. Weyde, "Adapting similarity on the magnatagatune database: effects of model and feature choices," in *Proceedings of the 21st international conference on world wide web*, 2012, pp. 931–936.

[19] S. Mohammed, P. Shi, and J. Lin, "Strong baselines for simple question answering over knowledge graphs with and without neural networks," *arXiv preprint arXiv:1712.01969*, 2017.

[20] S. Dieleman and B. Schrauwen, "Learning content-based metrics for music similarity," in *5th International Workshop on Machine Learning and Music (MML-2012)*, 2012, pp. 13–14.

[21] G. Futia, A. Vetrò, and J. C. De Martin, "Semi: A semantic modeling machine to build knowledge graphs with graph neural networks," *SoftwareX*, vol. 12, p. 100516, 2020.

[22] M. Hausenblas and M. Karnstedt, "Understanding linked open data as a web-scale database," in *2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications*. IEEE, 2010, pp. 56–61.

[23] A. Swartz, "Musicbrainz: A semantic web service," *IEEE Intelligent Systems*, vol. 17, no. 1, pp. 76–77, 2002.

[24] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[25] Z. Zhao, C. Ma, H. Yuan, and Z. Wang, "A dtp and solid based service for multi-source semantically-heterogeneous personal data management," in *2022 International Conference on Service Science (ICSS)*. IEEE, 2022, pp. 255–262.

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Symmetries and Minima in Differentiable Sinusoidal Models

Ben Hayes*, Charalampos Saitis, György Fazekas

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

***Abstract*— Recent work has enabled direct optimisation of unconstrained sinusoidal model frequencies by gradient descent, with applications in audio analysis and synthesis via differentiable digital signal processing. However, two challenges continue to impede further progress. First, we observe that the solution space is symmetric under permutation, leading to a phenomenon known as the responsibility problem when training neural network controllers. Second, in the case where relative model frequencies are constrained, we observe that an inability to exploit this symmetry induces a class of local minimum that can prevent convergence. In this abstract we describe these symmetries and minima. In our talk, we will discuss possible approaches to circumventing them and present early experimental results.**

***Index Terms*—** differentiable signal processing, machine learning, parameter estimation

I. SYMMETRY

The surrogate model described in [1] allows the frequency parameters of a sinusoidal model to be optimized by gradient descent by extending its parameters to the complex domain:

$$\mathfrak{s}_n(z) = \mathfrak{Re}\left(z^n\right) = |z|^n \cos n\angle z. \tag{1}$$

We consider the mean-squared error loss of our model under parameters $\mathbf{z} \in \mathbb{R}^K$ and some target signal $\mathbf{y} \in \mathbb{R}^N$:

$$\mathcal{L}(\mathbf{y}, \mathbf{z}) = \sum_{n=1}^{N} \left(y_n - \sum_{k=1}^{K} \mathfrak{s}_n(z_k) \right)^2 \tag{2}$$

It is clear that due to the summation over $\mathbf{z}$, the function $\mathcal{L}$ is symmetric under permutations $\{\pi_1, \ldots, \pi_{K!}\}$ of $\mathbf{z}$. When $\mathbf{z}$ is the output of a neural network, this leads to the *responsibility problem*, as described by Zhang et al. [2]. Briefly summarised, the ordered nature of a neural network's outputs causes a discontinuous partitioning of its output space when trained with an orderless objective. This generally prevents convergence in such problems.

II. MINIMA IN CONSTRAINED MODELS

We define a constrained sinusoidal model as:

$$x_n = \sum_{k=1}^{K} \alpha_k \cos c(\boldsymbol{\omega}, k) n, \tag{3}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ are, respectively, the independent amplitude and frequency parameters. The function c defines a relationship between the independent frequency parameter and the true component frequencies. It is straightforward to show that signals produced by a harmonic model or through amplitude and frequency modulation can be expressed in this way.

We observe that, despite the summation across values of k, such a model is *not* symmetric under permutations of $\boldsymbol{\omega}$. Further, the dependence of true component frequencies on k imposes an effective ordering on frequency components. Thus, when optimsing such a model using the surrogate from Eqn 1, local minima occur for values of $\boldsymbol{\omega}$ where there exists a k such that $c(\boldsymbol{\omega}, k) \in \{\hat{\omega}_j \mid j = 1, \ldots, J\}$ where $\hat{\boldsymbol{\omega}}$ is the vector of ground truth component frequencies.

In other words, a gradient based optimiser will become stuck when any of the model's true frequency components match with any of the frequency components of the target signal, even if other components remain unmatched. Whilst an unconstrained model would exploit the symmetry of the objective space to match the remaining components with its free parameters, the ordering imposed by constraint c prevents this in the constrained model.

III. CONCLUSION

Building on our recent work on sinusoidal frequency optimisation, we present two related challenges in optimising differentiable sinusoidal models by gradient descent. Resolving these issues would be of particular benefit in differentiable digital signal processing, with direct applications in neural music and speech synthesis. In our talk, we will present our work to date on candidate solutions and discuss directions for future research.

IV. REFERENCES

[1] B. Hayes, C. Saitis, and G. Fazekas, "Sinusoidal Frequency Estimation by Gradient Descent," Oct. 2022, arXiv:2210.14476 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2210.14476

[2] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Deep Set Prediction Networks," Apr. 2020, arXiv:1906.06565 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1906.06565

*Ben Hayes is supported by UK Research and Innovation [grant number EP/S022694/1].

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022
QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Affordances of Generative Models of Raw Audio to Instrumental Practice and Improvisation

Mark Hanslip
mwh512@york.ac.uk

Contemporary Music Research Centre, School of Arts and Creative Technologies,
University of York, UK

Abstract

Neural network architectures for generative modelling of raw audio are fast evolving, but their creative applications remain the preserve of very few artists. Part of my research* as a PhD candidate has been to investigate their usefulness to my long-standing practice as an improvising saxophonist with a view to generating wider knowledge of their potential utility to others. I will showcase outputs of my practice-based research, including raw audio datasets spawned from various aspects of my practice and samples generated from models of these data. I will then discuss the musical value of these outputs. My research finds generative models of raw audio to be of particular value as assistive technologies for the practice of improvisation, as agents for musical human-computer interaction and naturally as engines for sample-based musics. This work motivates future studies on their usefulness to other practitioners of improvised music, researcher-practitioners interested in interactivity and musicians working with sampling.

MODEL ARCHITECTURES USED

While recent advancements in generative modelling of raw audio such as RAVE [1] show impressive output fidelity, priorities of my work include practicality of training and likelihood of engagement from musicians. I have therefore opted for two longer-established methods, SampleRNN [2] and WaveGAN [3]. Both are trainable within a small number of hours on a single GPU, are well-documented with stable, up-to-date implementations [4][5], are straightforward to understand and train, and show distinct characteristics of outputs owing to their divergent processes. I expect the findings of my work to be applicable to alternative model architectures anyway.

REFERENCES

[1] A. Caillon, P. Esling, RAVE: A variational auto-encoder for fast and high-quality neural audio synthesis *https://github.com/acids-ircam/rave*
[2] Mehri S., Kumar K., Gulrajani I., Kumar R., Jain S., Sotelo J., Courville A., Bengio Y. (2017) SampleRNN: An Unconditional End-to-End Neural Audio Generation Model, ICLR 2017 paper. arXiv:1612.07838 [cs.SD]
[3] Donahue C., McAuley J. & Puckette M. (2018) Adversarial Audio Synthesis, *ICLR 2019.* arXiv: 1802.04208 [cs:SD]
[4] *https://github.com/rncm-prism/prism-samplernn*
[5] *https://github.com/mattjwarren/wavegan*

*Research supported by a UKRI / AHRC / WRoCAH grant.

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Practical Text-Conditioned Music Sample Generation

Scott H. Hawley[1,21] and Zach Evans, C.J. Carr, and Flavio Schneider[2]

[1]Chemistry and Physics Department, Belmont University, USA, scott.hawley@belmont.edu
[2]Harmonai

***Abstract*— We present a system whereby producers and composers can generate new short musical audio sound samples and foley effects by fine-tuning a text-conditioned generative diffusion model on their own sound libraries. Inference proceeds by users typing a description of the sound they would like to hear by supplying a "fake" full file path of a conceivable sound as it might appear in a library of sound samples. This model is capable of generating stereo or mono at high sample rates (e.g. 48 kHz), can run locally on small GPUs or Apple Silicon, and offers high-quality examples. This system is intended as a practical tool for music creators to be able to generate new sounds, while avoiding copyright infringement or other IP issues. We present an overview of such a system currently in operation, which will be released soon.**

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

Which car is moving? A 'listening' approach using distributed acoustic sensor systems

Chia-Yen Chiang[1] and Mona Jaber[2]

[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK, c.chiang@qmul.ac.uk
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

***Abstract*— Identifying the type, size, and occupancy level of moving vehicles is a crucial problem in intelligent transportation systems (ITS) as it enables the modeling of motorised traffic and, therefore, the conception and optimisation of sustainable mobility solutions. Computer vision is the most promising solution as it captures all the required details for ITS precise modeling. However, the method suffers from occlusion, adverse weather and visibility conditions and, more importantly, privacy intrusion. This project examines distributed acoustic sensor (DAS) systems as an alternative data source in which moving traffic is 'listened to' instead of watched.**

I. DAS TECHNOLOGY AND DATASET

A DAS system contains a interrogator and an optical fibre. The interrogator locates signals by travel time within fiber (similar to RADAR) and collects back-scattered light reflected by an optical fibre when it is deformed by acoustic vibrations (see Figure 1). Each point on the optical fibre can be used as a sensing unit to achieve continuous detection of acoustic event along the length of the fibre (up to 50km).

In a controlled field trial, DAS data was collected on a 4.8km road stretch equipped with a DAS system. Five different cars were driven in a predefined order in both directions of the road. In a two-dimensional signal-displacement map (see Figure 2(Top-left)), the x-axis indicates the time in fibre *shots*, where one shot s is equivalent to $1/1000.04$ seconds. The y-axis shows the position along the fibre in *bins* where one bin b is equivalent to 0.68 metres. The color intensity of each pixel, at bin b and shot s, shows the strength of displacement in radians $[rad]$.

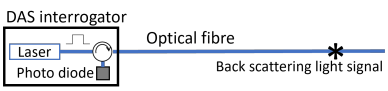

Figure 1: DAS system

II. PROPOSED METHODOLOGY AND RESULTS

We posit that each passing vehicle generates a unique DAS signature (sequence of displacements) that is repeated in different experiments with various speeds, occupancy, and other conditions. To this end, we propose a one dimensional convolution neural network (1D-CNN) in Figure 2(Bottom) to solve two classification problems: (1) identify one of five cars and (2) estimate the car size as *Large* and *Small* [1].

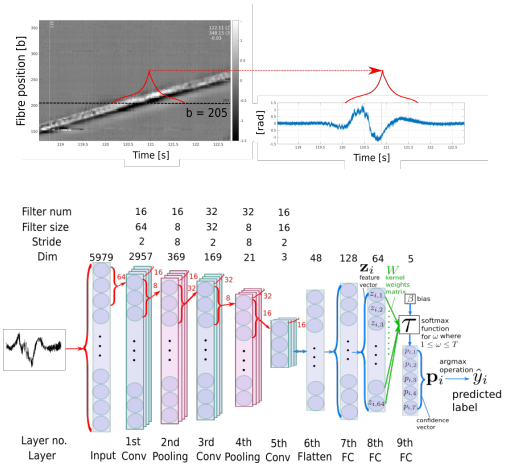

Figure 2: (Top-left) A two-dimensional recording of DAS signals in which the signal captured at b=205 is shown as displacement in $[rad]$ over time in the top-right figure. (Bottom) 1D-CNN architecture used for DAS signal classification. Number of filters, filter size, stride and dimension for each layer are presented.

Results of problem (1) and (2) are shown in Table 1. Two conclusions are drawn. Firstly, each car has a unique signature that is not affected by the speed, as seen from the results of (1). Secondly, the DAS signal contains information about the car size (related to weight) in addition to that of the car type, as seen from the results of (2).

	speed/ Accuracy	30	40	50	60	70	JS
(1)	mean	0.91	0.89	0.94	0.96	0.93	0.913
	SD	0.01	0.02	0.004	0.01	0.009	0.008
(2)	mean	0.97	0.89	0.93	0.97	0.96	0.926
	SD	0.009	0.01	0.001	0.007	0.009	0.007

Table 1: The accuracy results of problems (1) and (2) for specific speeds (in km/h) and for joint speeds JS where samples of all five speeds are mixed.

III. REFERENCES

[1] C.-Y. Chiang, M. Jaber, and P. Hayward, "A distributed acoustic sensor system for intelligent transportation using deep learning," 2022. [Online]. Available: https://arxiv.org/abs/2209.05978

DMRN+17: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2022

QUEEN MARY UNIVERSITY OF LONDON
TUE 20 DECEMBER 2022

YourMT3: a toolkit for training multi-task and multi-track music transcription model for everyone

Sungkyun Chang Simon Dixon and Emmanouil Benetos

Centre for Digital Music, Queen Mary University of London, UK

***Abstract*— This work-in-progress presents a compact toolkit for reproducing the training of `T5`-based multi-instruments and multi-track transcription models on multiple MIR tasks. We plan to release our code in early 2023 at `https://3.ly/amamR`.**

***Index Terms*—** automatic music transcription, music information retrieval, transformers, multi-task learning

I. INTRODUCTION

In recent years, multi-task learning approach [1, 2] has shown significant performance gains in multi-track and multi-instrument automatic music transcription (AMT). However, the training was not easy to reproduce due to the more complex structure required for data processing and task management compared to previous works for single-task and single-instrument AMT [3] or symbolic music generation [4]. In this work, our goal is to present an easy toolkit for training multi-task music transcription models [1, 2].

II. YOURMT3

The proposed toolkit in Figure 1 consists of two main components: task and trainer. Listed below are some design considerations to simplify training in the context of multi-task learning on audio and symbolic music data.

Defining an MIR Task: A task is simply definable with a set of MIDI tokenizer, vocabulary, and an audio processor. Vocabulary interacts with tokenizer, and together with audio processor it configures the data-stream for mixing subtracks.

Builder: We provide a task builder that can import various datasets through the `mirdata` library. All data are preprocessed and written only once into our format.

Data Streaming I/O: Our requirement is to have instant access to partial segments at specific timings from a large number of multi-track audio and MIDI files. For this we record note events in small separate segments and merge them on load. For piano rolls, we pre-load *compressed sparse matrix*.

Trainer: Our models are primarily based on T5 from the `huggingface` library, which has an active community among Transformer practitioners. Powered by `Pytorch Lightning` [1] and `DeepSpeed` [2], a highly efficient training is expected on a single GPU, as well as multi GPUs.

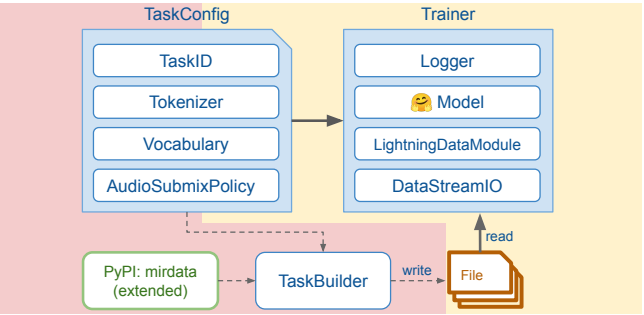

Figure 1: An overview of how to prepare a single MIR task (pink block), and train (yellow block) a model on the task. The dotted line represents preprocessing that runs only once during the task build, and the solid line represents streaming of data during training.

III. LIMITATIONS

Compared to the generic data I/O [3] adopted by MT3 [1], our toolkit is more simplified with only music data in mind. This design focus, however, can limit flexibility for non-music data. Unlike previous MT3, this toolkit includes a training code and on-the-fly audio processor with full randomness. Despite these advantages, the inability to support TPUs due to the CPU workloads can be another limitation.

IV. ACKNOWLEDGMENTS

This work is supported by Huawei Technologies.

V. REFERENCES

[1] J. P. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "Mt3: Multi-task multitrack music transcription," in *ICLR*, 2021.

[2] K. W. Cheuk, D. Herremans, and L. Su, "Reconvat: A semi-supervised automatic music transcription framework for low-resource real-world data," in *ACM Multimedia*, 2021.

[3] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.

[4] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *ACM Multimedia*, 2020.

[1] `https://www.pytorchlightning.ai/`
[2] `https://github.com/microsoft/DeepSpeed`
[3] `https://github.com/google/seqio`