DMRN+16: Digital Music Research Network

One-day Workshop 2021

Queen Mary University of London

Tuesday 21st December 2021

Chair: Simon Dixon

Programme

10:00	**Welcome, Simon Dixon (Queen Mary University of London)**
10:10	**KEYNOTE** "Sound on the Brain - Insights from Functional Neuroimaging and Neuroanatomy", **Prof Sophie Scott - (Institute of Cognitive Neuroscience - UCL)**
10:50	Building Style-aware Neural MIDI Synthesizers Using Simplified Differentiable DSP Approach, **Sergey Grechin and Ryan Groves (Infinite Album)**
11:00	Completing Audio Drum Loops with Transformer Neural Networks, **Teresa Pelinski (Queen Mary University of London), Behzad Haki and Sergi Jordà (Pompeu Fabra University)**
11:10	Evaluation of GPT-2-based Symbolic Music Generation, **Berker Banar and Simon Colton (Queen Mary University of London)**
11:20	NASH: the Neural Audio Synthesis Hackathon, **Ben Hayes, Cyrus Vahidi and Charalampos Saitis (Queen Mary University of London)**
11:30	**5 min break**
11:35	Designing a Synthesiser to Elicit a Feeling of Perceived Tension, **Connor Welham, Bruno Fazenda, and Duncan Williams (Acoustic Department, University of Salford)**
11:45	Is Automatically Transcribed Data Reliable Enough for Expressive Piano Performance Research?, **Huan Zhang, Simon Dixon (Queen Mary University of London)**
11:55	CAMAT: Computer Assisted Music Analysis Toolkit, **Egor Poliakov (IHMT Leipzig) and Christon R. Nadar (Semantic Music Technologies, Fraunhofer IDMT)**
12:05	An Investigation on Pitch-Based Features on Selected Music Generation Systems, **Yuqiang Li, Shengchen Li (Xi'an Jiaotong-Liverpool University) and George Fazekas (Queen Mary University of London)**
12:15	**Lunch break**
13:15	**KEYNOTE** "Learning Interpretable Music Representations: From Human Stupidity to Artificial Intelligence", **Assistant Prof Gus Xia - (NYU Shanghai)**
13:55	**Announcements and Intro to Gather Town**
14:00	**POSTER SESSION** An open poster session where the participants will be able to view the posters and chat with authors.
16:00	**Close ***

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

Posters

1	Sketching Sounds: Using Sound-shape Associations to Build a Sketch-based Sound Synthesizer, **Sebastian Löbbers and George Fazekas (Queen Mary University of London)**
2	Everyday Sound Recognition with Limited Annotations, **Jinhua Liang, Huy Phan and Emmanouil Benetos (Queen Mary University of London)**
3	Generating Comments from Music and Lyrics, **Yixiao Zhang and Simon Dixon (Queen Mary University of London)**
4	AI-Assisted FM Synthesis, **Franco Caspe, Andrew McPherson and Mark Sandler (Queen Mary University of London)**
5	Algorithmic Music Composition for the Environment, **Rosa Park (San Francisco State University)**
6	The Vienna Philharmonic's New Year's Concert Series: A Corpus for Digital Musicology and Performance Science, **David M. Weigl and Werner Goebl (University of Music and Performing Arts Vienna, Austria)**
7	An Interactive Tool for Visualising Musical Performance Subtleties, **Yucong Jiang (University of Richmond)**
8	A Benchmark Dataset to Study Microphone Mismatch Conditions for Piano Multipitch Estimation on Mobile Devices, **Jakob Abeßer, Franca Bittner, Maike Richter, Marcel Gonzalez and Hanna Lukashevich (Fraunhofer IDMT)**
9	Looking at the Future of Data-Driven Procedural Audio, **Adrián Barahona-Ríos (University of York)**
10	Making Graphical Scores Accessible to Visually Impaired People: A Haptic Interactive Installation**, Christina Karpodini**
11	Acoustic Representations for Perceptual Timbre Similarity**, Cyrus Vahidi, Ben Hayes, Charalampos Saitis and George Fazekas (Queen Mary University of London)**
12	Investigating a Computational Methodology for Quantitive Analysis of Singing Performance Style, **Yukun Li, Polina Proutskova, Zhaoxin Yu and Simon Dixon (Queen Mary University of London)**
13	Variational Auto Encoding and Cycle-Consistent Adversarial Networks for Timbre Transfer, **Russell Sammut Bonnici, Martin Benning and Charalampos Saitis (Queen Mary University of London)**
14	Characterizing Texture for Symbolic Piano Music**, Louis Couturier (Universite de Picardie Jules Verne), Louis Bigo (Universite de Lille) and Florence Leve (Universite de Picardie Jules Verne and Universite de Lille)**
15	Beat-Based Audio-to-Score Transcription for Monophonic Instruments, **Jingyan Xu (Music X Lab, NYU Shanghai)**
16	Predicting Hit Songs: Multimodal and Data-driven Approach, **Katarzyna Adamska, Joshua Reiss (Queen Mary University of London)**
17	Character-based Adaptive Generative Music for Film and Video Games, **Sara Cardinale and Simon Colton (Queen Mary University of London)**
18	Physically-inspired Modelling with Neural Networks, **Carlos De La Vega Martin and Mark Sandler (Queen Mary University of London)**
19	Hearing a Volumetric Drum, **Rodrigo Diaz and Mark Sandler (Queen Mary University of London)**
20	Computational Modelling of Jazz Piano via Large-Scale Automatic Transcription, **Drew Edwards and Simon Dixon (Queen Mary University of London)**

21	Music Emotion Mood Modelling using Graph and Neural Nets, **Maryam Torshizi, George Fazekas, and Charalampos Saitis (Queen Mary University of London)**
22	Virtual Placement of Objects in Acoustic Scenes, **Yazhou Li, Lin Wang and Joshua Reiss (Queen Mary University of London)**
23	Real Time Timbre Transfer with a Smart Acoustic Guitar, **Jack Loth and Mathieu Barthet (Queen Mary University of London)**
24	Music Interestingness in the Brain, **Chris Winnard (Queen Mary University of London), Preben Kidmose (Aarhus University), Kaare Mikkelsen (Aarhus University) and Huy Phan (Queen Mary University of London)**
25	Intelligent music production, **Soumya Vanka (Queen Mary University of London), Jean Baptiste Roland (Steinberg) and George Fazekas (Queen Mary University of London)**
26	Composition-aware music recommendation for music production, **Xiaowan Yi and Mathieu Barthet (Queen Mary University of London)**
27	Dynamic Mood Recognition in Film Music, **Ruby Crocker and George Fazekas**; **(Queen Mary University of London)**
28	The Sound of Care: Researching the Use of Deep Learning and Sonification for the Daily Support of People with Chronic Pain, **Bleiz Del Sette and Charalampos Saitis (Queen Mary University of London)**
29	Embodiment in Intelligent Musical Systems, **Oluremi Falowo and Charalampos Saitis (Queen Mary University of London)**

Keynote Talks

Keynote 1: By **Prof. Sophie Scott** -Director, Institute of Cognitive Neuroscience, UCL.

Title: "Sound on the Brain - Insights from Functional Neuroimaging and Neuroanatomy"

Abstract: In this talk I will use functional imaging and models of primate neuroanatomy to explore how sound is processed in the human brain. I will demonstrate that sound is represented cortically in different parallel streams. I will expand this to show how this can impact on the concept of auditory perception, which arguably incorporates multiple kinds of distinct perceptual processes. I will address the roles that subcortical processes play in this, and also the contributions from hemispheric asymmetries.

Keynote 2: By **Prof. Gus Xia -** Assistant Professor at NYU Shanghai

Title: "Learning Interpretable Music Representations: From Human Stupidity to Artificial Intelligence"

Abstract: Gus has been leading the Music X Lab in developing intelligent systems that help people better compose and learn music. In this talk, he will show us the importance of music representation for both humans and machines, and how to learn better music representations via the design of inductive bias. Once we got interpretable music representations, the potential applications are limitless.

Organizing Committee

Corey Ford
Max Graf
Madeline Hamilton
Benjamin Hayes
Jiawen Huang
Harnick Khera
Yin-Jyun Luo
Luca Marinelli
Xavier Riley
Eleanor Row
Shubhr Singh
Christian Steinmetz
Jingjing Tang
Lewis Wolstanholme
Yixiao Zhang

Supported by UKRI AIM CDT

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

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Building style-aware neural MIDI synthesizers using simplified differentiable DSP approach

Sergey Grechin[1] and Ryan Groves[2]

[1]Infinite Album, grechin.sergey@gmail.com
[2]Infinite Album, ryan@infinitealbum.io

***Abstract*— We explore how simplified differentiable DSP approach can be used to build realistic sounding MIDI-controllable monophonic synthesizers. The simplification involves directly using MIDI data as input to the DDSP decoder rather than continuous F0 and loudness curves. On top of that, we show how incorporating additional style-based and temporal channels can be used to imitate various aspects of performance and improve realism. We demonstrate the results by applying the approach to the task of modelling the sound of electric guitar. The presented results were obtained with a model trained on less than 12 minutes of manually MIDI-annotated audio. The source code is released along with the prepared dataset.**

***Index Terms*—** Deep Learning, DDSP, virtual synthesizers, MIDI

I. MODEL ARCHITECTURE AND INPUTS

The model was built on top of [1] which in turn is a simplified version of original DDSP design [2]. In contrast to the original implementation, in our model audio is not directly used to produce inputs to the decoder. Instead, we use MIDI annotations to generate fundamental frequency curve (F0), loudness, additional timing signals ("distance from onset" , "distance to offset"), proposed in [3], and arbitrary CC (continuous controller) values. CC values are used to capture various performance characteristics such as "openness" - the degree of how muted the guitar string is when plucked. Additional inputs are passed through dedicated stacks of dense layers.

II. DATASET AND RESULTS

For training, 12 minutes of playing chromatic scales on an electric guitar were recorded. MIDI annotations were made manually with CC55 used to represent open (127) and muted (1) sound. For training, velocity values were obtained using the A-weighted loudness-based approach proposed in [3]. At the inference stage velocity was taken from the source MIDI.

We encourage the reader to visit the online supplement page [4] to listen to the generated examples and access additional resources such as the source code and the dataset.

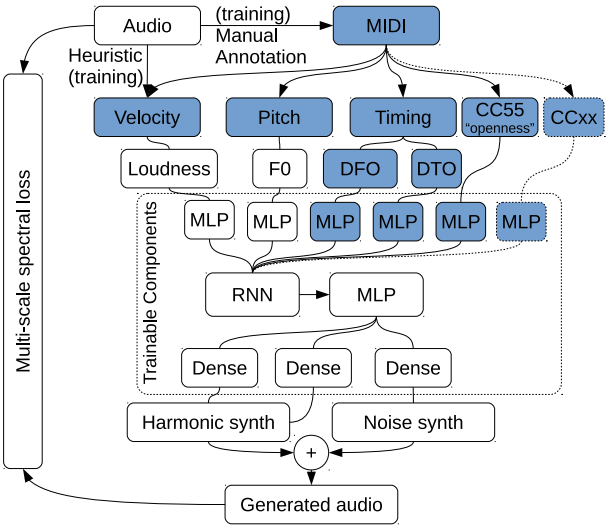

Figure 1: Componets of the model. The white blocks represent the components of original DDSP design [2]. MLP - Multilayer perceptron, DFO - distance from onset, DTO - distance to offset, CC - continuous controller

Figure 2: Waveforms generated on C-5 for two MIDI notes with closed and open sound. This example demonstrates that the model has succeeded in learning temporal characteristics of the guitar sound, including the modelling of initial transient.

III. FUTURE RESEARCH

We plan to research if other aspects of playing style can be captured using this approach, such as playing chords. For that, we plan to add additional harmonic synth stacks with independently learnable parameters.

IV. REFERENCES

[1] DDSP simplified repository. [Online]. Available: https://github.com/raraz15/ddsp_simplified

[2] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "DDSP: Differentiable digital signal processing," in *ICLR*, 2020.

[3] Nicolas Jonason, Bob Sturm, and Carl Thom, "The control-synthesis approach for making expressive and controllable neural music synthesizers," in *Joint Conference on AI Music Creativity*, 2020.

[4] Online supplement. [Online]. Available: https://grechin.org/neural_synthesizers_with_simplified_ddsp.html

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Completing Audio Drum Loops with Transformer Neural Networks

Teresa Pelinski*[1], Behzad Haki[2] and Sergi Jordà[2]

[1]Centre for Digital Music, Queen Mary University of London, UK, t.pelinskiramos@qmul.ac.uk
[2]Music Technology Group, Pompeu Fabra University, Barcelona, Spain

***Abstract*— Infilling drum loops refers to complementing a drum pattern with additional drum events that are stylistically consistent with the loop. We present the Transformer Groove Infilling (TGI), a Transformer based approach to the infilling task. Until now, the infilling of drum beats has been implemented using Recurrent Neural Network (RNN) architectures, in particular, sequence-to-sequence models that employ LSTM cells. However, in such architectures, as a consequence of sequential computation, proximity is emphasised when dealing with dependencies in the input sequence. Furthermore, those models receive the audio loops as symbolic input sequences. In contrast, the TGI is based on the Transformer architecture, which relies entirely on self-attention mechanisms to represent the input sequences, allowing for faster training as a result of parallelisation. In addition, we present a novel direct audio representation that enables the TGI to receive the input drum loops in the audio domain, avoiding their transcription and tokenisation.**

***Index Terms*—** Infilling, Drum generation, Transformer

I. RELATED WORK

An infilling model can be used for computer assisted composition; for instance, the composer can sketch some instrument parts of a drum beat and obtain the system's suggestions for additional parts. So far, the GrooVAE model by Gillick et al. [1] is the only model in the literature. In particular, this system adds the hi-hat part to a drum MIDI performance and is based on LSTMs. The GrooVAE model deals with symbolic musical representations both in the input loop and the output infilled pattern. To our knowledge, the Transformer Groove Infilling model (TGI) is the first model of its kind that tackles the infilling task for drums with an audio input and a Transformer architecture.

II. METHODOLOGY

We trained the TGI[1] for three different tasks: infilling closed hi-hats, jointly infilling kicks and snares, and infilling a pattern without an instrument specification. The datasets were obtained after processing the Groove MIDI Dataset [1]. The TGI model is based on the Transformer Encoder block from the original Vaswani et al.'s [2] Transformer architecture. The model receives a 2 bar audio loop as input and outputs a symbolic representation of the infilled instrument parts. The numerical representation of the input audio corresponds to a novel reduced representation of an onset spectrogram, both in the time and audio domain. The output symbolic representation captures hit, velocity and offsets for each drum kit instrument, which is then easily converted into the MIDI format and synthesised with a given soundfont.

III. EVALUATION

For the task of infilling closed hi-hats, we obtain a moderately good accuracy (75.4%). In order to evaluate the performance of the input audio representation, we train an identical model with a symbolic input, and find that the audio representation only entails a 4.3% decrease in hit prediction accuracy. In the task of infilling kicks and snares, we find that the accuracy improves to 83.4%. We assume that this improvement is caused by two main factors: a larger training dataset (4 times larger) and the benefit of learning to infill two instruments jointly. However, we observed a large gap between validation and training loss, likely caused by an overly specific training dataset. Finally, in the infilling random experiment we infill drum events without attending to the instrument they belong to. We train two different versions of the model, one that removes between 10% and 30% of the initial hits and one that removes between 40% and 70%. Both experiments achieve very high accuracies, 97.7% and 94.7% respectively. These are likely due to the size of the dataset (4 times larger than in the closed hi-hat task). However, since the training datasets are generated by removing random hits of the score, the musical relevance of this task remains questionable.

IV. REFERENCES

[1] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman, "Learning to groove with inverse sequence transformations," in *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, California, USA: PMLR, May 2019, pp. 2269–2279, iSSN: 2640-3498.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 2017-Decem, 2017, pp. 5999–6009, iSSN: 10495258. [Online]. Available: https://arxiv.org/abs/1706.03762

*This research was conducted as T. Pelinski's Sound and Music Computing MSc thesis, under the supervision of B. Haki and S. Jordà at the Pompeu Fabra University.

[1]Standalone code and datasets available at `https://zenodo.org/record/5347908`, with contributions by B. Haki and M. Nieto.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Evaluation of GPT-2-based Symbolic Music Generation

Berker Banar *[1] and Simon Colton[1]

[1]School of EECS, Queen Mary University of London, UK, b.banar@qmul.ac.uk

***Abstract*— In this study, we evaluate the performance of two different GPT-2 models, which are pre-trained on natural language and utilised in symbolic music generation. Our evaluation is based on statistical methods and musical metrics. We also address anomalies that we have found in our initial experiments and suggest some future directions for the analysis.**

***Index Terms*—** Generative Music, Transformers, GPT-2, Music Generation Evaluation, Transfer Learning

I. INTRODUCTION

Transformers [1] are attention-based models originally designed for NLP tasks, which are widely used in AI-based symbolic music generators due to their proven success [2]. One typical generative music practice involves taking a transformer model pre-trained on a natural language corpus and fine-tuning it using a symbolic music dataset [3]. While there are transformer architectures specifically tailored for music generation and trained on a symbolic music dataset from scratch [4], transformer models pre-trained on natural language data, e.g. GPT-2 [5], are still useful for music generation as the fine-tuning process requires less data and lower computational resources. However, using pre-trained models in a cross-domain setting brings its own challenges, as there might be some limitations to transferred knowledge [6]. Evaluation of the generative models provides better insight for practitioners and helps them to choose a method according to their application needs, and is useful for improving the model architecture and training procedure. Transfer learning approaches make the evaluation in target data modality (in this case symbolic music) even more crucial [6].

II. EXPERIMENTS

In this study, we use a GPT-2 small architecture with 124m parameters [5] and evaluate the performance of two models, which are fine-tuned to different loss values; one model is well trained (loss value = 0.17) and the other is poorly trained (loss value = 0.70), where the fine-tuning process starts from the loss value of 1.40, roughly. To fine-tune GPT-2, we use the GiantMIDI-Piano dataset [7], which has 10854 classical piano music pieces, and curate a subset of the dataset with 300 pieces, which follows similar statistics to the whole corpus in terms of our musical metrics. We represent our music data in text format, by using note number, velocity, start time and end time values and process our music in roughly 5 bars format for the fine-tuning corpus and generated material.

We use 6 different musical metrics in the analysis, namely pitch count (number of different note numbers used), pitch range (maximum note number - minimum note number), average note duration, average velocity, chroma metric (ratio of number of least used 5 pitches to number of most used 7 pitches, given that the seven note scales are common in Western classical music), and a chords metric (ratio of number of augmented triad, major, minor, dominant, half-diminished and diminished seventh chords to number of major, minor and diminished triad chords, given the occurrences in Western classical music).

For the evaluation, we generate 750 samples from each of our GPT-2 models and calculate the musical metrics for each of the generated sets and also for the fine-tuning corpus. Then, we generate histograms of the musical metric values per each set and metric. As in [8], we convert these histograms into continuous probability density functions (pdf) using kernel density estimation. Finally, we calculate KL distance and overlapping area between the pdfs of the fine-tuning corpus and each of the generated sets [8].

In an ideal learning setting, generated material from the well-trained model should be statistically closer to the fine-tuning corpus, which means that we expect to see lower KL distance and higher overlapping area in the well trained case for each of the musical metrics.

	Well Trained		Poorly Trained	
	KLD	OA	KLD	OA
Chroma Metric	0.1246	0.1126	0.0798	0.4398
Chords Metric	0.1621	0.5831	0.2479	0.3371
Pitch Count	0.0278	0.2072	0.8252	0.0818
Pitch Range	0.0448	0.0047	0.0485	0.0273
Avg. Duration	0.0097	0.7801	0.0154	0.5841
Avg. Velocity	0.0074	0.8076	0.0046	0.7990

Table 1: Analysis of generated music by well / poorly trained models

As shown in Table 1, for the chords metric, pitch count and average note duration, the results are as expected, but for the chroma metric, we have higher KL distance and lower overlapping area in the well trained case, which is unexpected and an anomaly. Arguably, this might suggest that the model fails to learn chroma/scale properties. For the pitch range and average velocity, the results are inconclusive as KL distance and overlapping area suggest oppositely since the smaller KL distance and overlapping area happen at the same training case, which is an observed phenomenon with these distance measures in [8].

For future work, we would like to verify the validity of these distance measures and experiment with others since it is challenging to determine whether a distance measure is meaningful in the music generation context. Also, to further investigate the effect of training level, we will sample from more models with various loss values. Moreover, data representation, fine-tuning corpus and model architecture play important roles in this learning setting and we would like to deepen our analysis by also introducing these parameters to our experiments.

III. REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[2] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," 2020.

[3] C. Payne, "Musenet," https://openai.com/blog/musenet/, 2019.

[4] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," 2018.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[6] Z. Wu, N. F. Liu, and C. Potts, "Identifying the limits of cross-domain knowledge transfer for pretrained models," 2021.

[7] Q. Kong, B. Li, J. Chen, and Y. Wang, "Giantmidi-piano: A large-scale midi dataset for classical piano music," 2020.

[8] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Computing and Applications*, vol. 32, 05 2020.

*Berker Banar is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

NASH: the Neural Audio Synthesis Hackathon

Ben Hayes*, Cyrus Vahidi, and Charalampos Saitis

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

***Abstract*— The field of neural audio synthesis aims to produce audio using neural networks. A recent surge in its popularity has led to several high profile works achieving impressive feats of speech and music synthesis. The development of broadly accessible neural audio synthesis tools, conversely, has been limited, and creative applications of these technologies are mostly undertaken by those with technical know-how. Research has focused largely on tasks such as realistic speech and musical instrument synthesis, whereas investigations into high-level control, esoteric sound design capabilities, and interpretability have received less attention. To encourage innovative work addressing these gaps, C4DM's Special Interest Group on Neural Audio Synthesis (SIGNAS) propose to host our first Neural Audio Synthesis Hackathon: a two day event, with results to be presented in a session at DMRN+16.**

***Index Terms*—** hackathon, neural audio synthesis

I. INTRODUCTION

In the field of image generation, the creative capabilities of generative models, such as generative adversarial networks (GANs) and vector quantized variational autoencoders (VQ-VAEs), have been extensively explored by an active community of creators, hackers, researchers, and computational artists. Comparatively, the creative capabilities of neural audio synthesis models, which draw on a breadth of deep learning techniques ranging from generative modelling [1] to differentiable rendering [2], have received considerably less attention. The capabilities of these models beyond their performance on well established benchmark tasks are thus poorly understood. Whilst certain prominent community members, such as *Dadabots*[1], *Holly Herndon*[2], and *Hexorcismos*[3], have applied neural audio synthesis models creatively, the technical barrier to entry remains high, and this is compounded by a lack of tools and interfaces for would-be users of neural audio synthesis technology.

II. AIMS

NASH (the Neural Audio Synthesis Hackathon) aims to encourage cross-disciplinary collaboration in neural audio synthesis by encouraging the development of new techniques, tools, and interfaces for neural audio synthesis, with a particular focus on creative musical applications. We thus propose four main topic areas for the hackathon:

1. Interfaces and instruments
2. Novel techniques and models
3. Synthesis control
4. Creative applications

Participants are encouraged to consider these when selecting their project, although this list should not be considered exhaustive — we welcome all hacks that participants believe will be valuable to the neural audio synthesis community.

III. RULES

To be considered, all teams must submit (1) a demonstration video of length two minutes or less, (2) a public source code repository, or steps to replicate the technical portion in the case of creative applications. Submitting an interactive demo is also encouraged where appropriate, although this is not a requirement for entry.

The hackathon will take place over a 24 hour period on the weekend of 18th–19th December UTC, and submissions must be made within this time window. When submissions close, an online voting platform will be made available to hackathon participants and DMRN attendees.

IV. FURTHER INFORMATION

More detailed information, including links to register and join teams, can be found on the hackathon's website[4]. Links to demonstration videos and voting information will also be displayed on this page.

V. REFERENCES

[1] M. Huzaifah and L. Wyse, "Deep generative models for musical audio synthesis," *arXiv:2006.06426 [cs, eess, stat]*, June 2020, arXiv: 2006.06426. [Online]. Available: http://arxiv.org/abs/2006.06426

[2] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," in *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr

*This work was supported by UK Research and Innovation [grant number EP/S022694/1]

[1] https://dadabots.com/

[2] https://www.hollyherndon.com/

[3] https://twitter.com/hexorcismos

[4] https://signas-qmul.github.io/nash/

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Designing a synthesiser to elicit a feeling of perceived tension

C.Welham[1*], B.M.Fazenda and D.Williams[2]

[1*]Acoustic Department, University of Salford, England, C.J.Welham@edu.salford.ac.uk
[2] Acoustic Department, University of Salford, England

***Abstract*— Tension is an emotion which often has a negative connotation. However, tension can also be an important component in the domain of entertainment, for example watching a horror movie. This research describes the design of a synthesiser intended to induce or exaggerate feelings of perceived tension. A study was conducted with a group of participants (n=23), in which several samples created using a synthesiser (n=50) were tested against a reference sound. A Likert-scale was used to rank individual samples based on their relative level of perceived tension. A linear regression and principal component analysis (PCA) were conducted. The PCA demonstrated that several acoustic features correlated with tension.**

I. DEFINITIONS

The APA defines tension as a "feeling of physical and psychological strain accompanied by discomfort, uneasiness, and pressure to seek relief through talk or action".

II. CONTEXT AND PURPOSE

This research aims to better understand perceived tension and how it may be induced via a synthesised musical signal. The goal of the research is to allow for a correlation to be drawn between the acoustic features of a given signal and its level of perceived tension.

III. SYNTHESISER DESIGN

The synthesiser was designed using an FM model [1]. The parameters used were as follows: ADSR for both oscillators, ratio, brightness, and a three-band EQ with adjustable gain.

All samples for testing were generated using random numbers within a given range for each parameter. Each sample used the same note value (F3) and duration.

IV. METHODOLOGY

Participants (N=23) evaluated samples using a Likert scale (1-7) of perceived tension in comparison to a reference. There were three separate tests run, each with the same number of total samples (N=50). The scale ranged from “much less tense” to “much more tense”.

V. RESULTS

An initial least squares a linear regression model was fitted to establish a correlation between acoustic features in the signal with tension level. However, model fitting failed several key assumptions for the underlying data.

A Principal Component Analysis (PCA) was run to identify several principal components (PCs) representing the acoustic feature space. The PCA identified both frequency-based (Var explained = 51.48%) and time-based (Var explained = 16.14%) features as the dimensions under which most variance of the data can be established. perceived tension scores appear to be almost exclusively correlated to spectral based features in dimension 1, such that tension increases with Brightness, Spectral Centroid, and Roll Off. See figure 1.

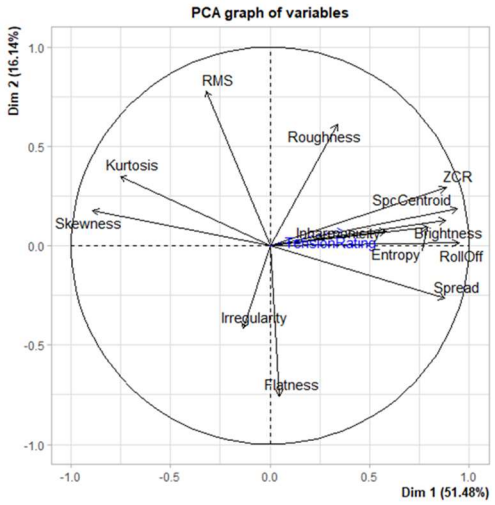

Figure 1. Principal Component Analysis plot, showing associations between the first two principal components.

A further attempt to model a linear regression between the first two PCs and perceived tension was run but failed due to violating underlying assumptions on the acoustic feature data. Primarily the assumption of normality and homoscedasticity.

ACKNOWLEDGMENT

Thanks to the University of Salford for facilitating this study.

REFERENCES

[1] American Psychological Association. (n.d.). Tension. In dictionary.apa.org. Retrieved November 22, 2021, from https://dictionary.apa.org/tension
[2] Chowning, J. M. (1973). Synthesis of Complex Audio Spectra By Means of Frequency Modulation. AES: Journal of the Audio Engineering Society, 21(7), 526–534.

*Research supported by the University Of Salford.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Is Automatically Transcribed Data Reliable Enough for Expressive Piano Performance Research?

Huan Zhang and Simon Dixon*

Center for Digital Music, Queen Mary University of London, UK, huan.zhang@qmul.ac.uk

***Abstract*— In this extended abstract, we discussed a new perspective of obtaining data for piano performance research through deep learning based Automatic Transcription models, and proposed methods for evaluating the reliability of such transcription.**

I. CONTEXT

To study expressive piano performance with computational models, the most widely used format for representing a performance is expressive MIDI, with events carried with control attributes such as velocity and timestamps. Recent advancement of automatic music transcription (AMT) provides new perspective of generating expressive MIDI from audio recordings. With fine-grained transcription as well as symbolic score, we can analyze the piece of piano performance as close as note level. Another motivation to investigate automatic expressive transcription is large-scale dataset curation. As summarized in Table 1, we can observe that the scale of existing datasets is far from enough to distinguish pianistic styles and composition genres.

II. RELATED WORK AND LIMITATIONS

Most recent models of piano transcription involves the high-resolution piano transcription system [1] and the onsets-and-frame system [2]. Their training dataset Maestro, however, is limited to an alignment resolution of 3ms [3]. Both of the systems claimed an F1 score greater than 90% (with 50ms tolerance), but the testing set accuracy is not equivalent to reliability in capturing perceptual expressiveness. Another limitation of existing transcription models is the ability to process historical virtuoso recordings. Given that the training data are obtained from a clean acoustic environment, noisy or live recording of performances can result in poor performance.

name	tot. duration	#. composer	#. pianist
Maestro [3]	∼172h	28	-
Crestmuse [4]	∼50h	6	12
Mazurka [5]	∼140h	1	45

Table 1: Comparison of major datasets in expressive performance

*H. Zhang is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation and Queen Mary University of London.

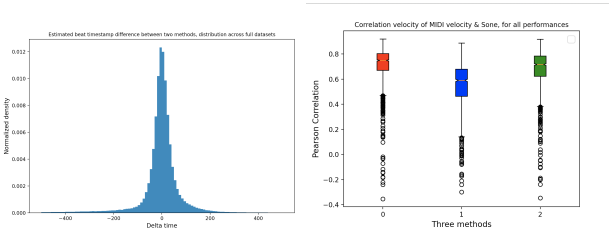

Figure 1: Left: distribution of timestamp differences; Right: Correlation of velocity & dynamics

III. PROPOSED METHODOLOGY

Cross-datasets Comparison We cross-compared the expressive attributes of beat timestamps and velocity with the MazurkaBL [5] dataset. The mazurka pieces were transcribed and their timestamps were compared with the perceptually annotated and audio aligned beat timestamp from MazurkaBL, and Figure 1 left showing that the majority of differences lie within ∼50ms range. Figure 1 right also demonstrates a high correlation between the transcribed velocity and audio loudness in Sone.

Listening Experiment To examine how well expressive transcription works on a perceptual scale, we will first render the transcribed MIDI on a reproducible piano such as a Yamaha Disklavier, and ask participants who know the piece to rate the reproduction deviation from the original performance. Another experimental design involves iterative transcription, where a series of transcription-reproduction-transcription will amplify the inaccuracies.

IV. REFERENCES

[1] Q. Kong, B. Li, X. Song, Y. Wan, Y. Wang, and S. D. Oct, "High-resolution Piano Transcription with Pedals by Regressing Onsets and Offsets Times," *arXiv:2010.01815v2*, 2020.

[2] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," *Proc. of the 19th ISMIR 2018*.

[3] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling Factorized Piano Music Modeling and Generation with the Maestro Dataset," *Proc. ICLR 2019*.

[4] M. Hashida, E. Nakamura, and H. Katayose, "CrestMusePEDB 2nd EDITION: Music Performance Database with Phrase Information," *Sound and Music Computing*, 2018.

[5] K. Kosta, O. F. Bandtlow, and E. Chew, "MazurkaBL: Score-aligned Loudness, Beat, and Expressive Markings Data for 2000 Chopin Mazurka Recordings," *Proc. of the International Conference on Technologies for Music Notation and Representation – TENOR'18*.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

CAMAT: Computer Assisted Music Analysis Toolkit

Egor Poliakov[1] and Christon R. Nadar[2]

[1]HMT Leipzig, Germany, egor.poliakov@hmt-leipzig.de
[2]Semantic Music Technologies, Fraunhofer IDMT, Ilmenau, Germany

***Abstract*— We introduce CAMAT, a python-based sheet-music parsing and analysis tool based on the auditory model. The toolkit aims to provide computer-assisted analytical methods for musicological research, especially for statistical investigations on larger databases with an educational focus.**

***Index Terms*—** sheet music, analysis, python, pandas, auditory model

I. INTRODUCTION

Despite a wide integration of music software in music performance and production, there is still a noticeable gap in the adoption of software tools in various fields of music education. During the one-year scientific and educational project "Computer-assisted Music Analysis" at Hochschule für Musik FRANZ LISZT Weimar (HfM) Weimar, CAMAT was developed, tested, and evaluated several flexibly applicable teaching modules based entirely on open-source software with a goal of providing powerful tools for integration in ongoing musicology and music theory courses. The teaching modules are dedicated, among other things, to the computer-based annotation and visualization of sheet music texts and audio files, the statistical analysis of music corpora, the search for musical patterns (rhythms, melodies, harmony connections, etc.), and the comparison of interpretation. Most of the learning modules are designed as Jupyter Notebooks. In addition, they are available online on a wiki-based resource which also includes a large data bank (over 4800 entries) of MusicXML scores [1].

II. CAMAT

CAMAT (Computer Assisted Music Analysis Tool) was developed as a dedicated tool for parsing and analyzing MusicXML scores. Although there exist various tools like humdrum[2] [1] and music21[3] [2], CAMAT tries to solve some particular design problems that appeared during the conception and realization of certain learning modules. Here is a quick overview of the problem fields we encountered during this process described in the following sections.

II.1. HANDLING OF TIED CHORDS

Analysis of choral and piano music was one of the main topics of developed teaching modules. Therefore, we had to deal with scores that consisted of many tied notes (which is very usual in any polyphonic music). Because the MusicXML is by design a dedicated notation format which is built mainly for engraving purposes, a lot of crucial information about the actual duration of a note is handled as optional information (because it's not an engraved note, but simply an optional character that was added to previous note), that in some cases cannot be easily extracted and allocated. That led to severe problems while trying to parse the exact duration of every single tied note and especially note groups in dense chord and polyphonic structures. In CAMAT, we reconsider the weight of ties and handle them on a top priority level as crucial information besides the pitch and duration. This decision also led to the idea of storing all the parsed information in pandas data-frames to preserve the exact duration of every tied note in a persistent rhythmical grid.

[1]https://analyse.hfm-weimar.de
[2]https://www.humdrum.org/Humdrum/
[3]https://web.mit.edu/music21/

II.2. PARSING OF POLYRHYTHM, POLYMETRIC AND UPBEAT STRUCTURES

The tool can parse polyrhythm, polymetric, and upbeat structures while maintaining the unique metric profiling values. Because of the persistent rhythmical grid structure inside the pandas data-frame, we could now correctly parse and synchronize even very complex polyrhythm and polymetric structures while maintaining the unique metric profiling for every given part even if different time signatures are defined. Parsing of various metric positions of upbeats or repeat signs could also be fully integrated. Overall, the way we designed the storage of notation-based data within the pandas data frame, which includes combining and synchronizing all events along a fixed timeline, is very similar to an auditory model, cause the exact pitch-duration information of the actual perceivable musical events can be extracted.

II.3. APPROPRIATE INFRASTRUCTURE FOR NOTE SHEET BASED CORPUS ANALYSIS

Due to the scalability of pandas dataframes we found an easy solution to parse and store the information of multiple scores, that led to efficient integration of corpus based analysis in the learning modules.

III. CONCLUSION

The goal of the CAMAT is to provide a computer-based analysis tool for music analysis. CAMAT also provides parsing, visualizing musical texts, statistical analysis of music corpora, and searching for musical patterns such as melodies and rhythms.

IV. ACKNOWLEDGMENTS

The project Computergestützte Musikanalyse in der digitalen Hochschullehre (computer-aided music analysis within digital higher education) is funded by the Thuringian Ministry for Economy, Science and Digital Change and by Deutscher Stifterverband.

V. REFERENCES

[1] D. B. Huron, *The humdrum toolkit: Reference manual*. Center for Computer Assisted Research in the Humanities, 1994.

[2] M. Cuthbert and C. Ariza, "Music21: A toolkit for computer-aided musicology and symbolic music data," in *ISMIR*, 2010.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

An Investigation on Pitch-Based Features on Selected Music Generation Systems

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

[1]Xi'an Jiaotong-Liverpool University, yuqiang.li19@student.xjtlu.edu.cn
[2]Queen Mary Univeristy of London

***Abstract*— Pitch-based features are important factors for the evaluation of generated music. This paper investigates the distribution of pitch-based features of melodies generated by selected systems. Based on the CSMT 2020 Data Challenge dataset, three types of pitch-based features are investigated: interval mean (IM), interval standard deviation (ISD) and tonal standard deviation (TSD). Results show that IM tends to be better learned by CNN-GAN systems while ISD and TSD could be better learned by transformer networks than other systems, which suggests that different systems are good at modelling different pitch-based features of music.**

I. INTRODUCTION

Automatic music generation has been a popular research topic in recent years. GAN (Generative Adversarial Network), VAE (Varational Auto-Encoder) and Transformer are considered as commonly used network architectures for music generation. However, the evaluation of generated music still remains a challenging task. This paper investigates a certain range of pitch-related features in generated music.

The proposed pitch-based features are extracted from melodies, named Interval Mean (IM), Interval Standard Deviation (ISD) and Tonal Standard Deviation (TSD).

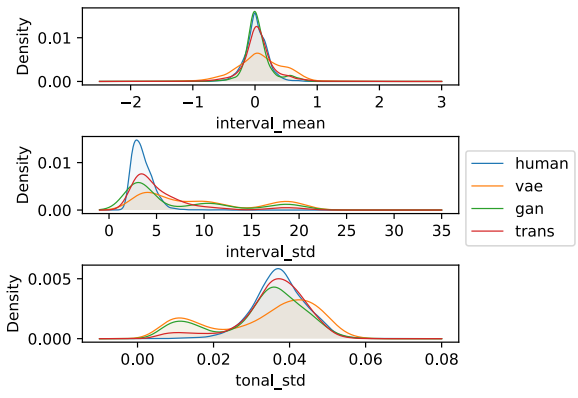

Figure 1: Approximated Probability Density Functions of the 3 Features

Suppose a melody is represented as $X = [p_0, p_1, \dots]$. The intervals of every two neighbour notes, denoted as ΔX, are the first-order (forward) difference of the pitch sequence X. We define $\mathrm{TP}(X)$ as the tonal probability function of X which yields 12 values $[c_0, c_1, \dots, c_{11}]$, where c_i indicates the number of notes in X belonging to the i-th key. The vector is then normalised such that $\sum \mathrm{TP}(X) = 1$. IM, ISD and TSD can be written as the following functions of X,

$$\mathrm{IM}(X) := \mathbb{E}[\Delta X],$$
$$\mathrm{ISD}(X) := \mathrm{SD}[\Delta X],$$
$$\mathrm{TSD}(X) := \mathrm{SD}[\mathrm{TP}(X)].$$

The dataset for CSMT 2020 data challenge[1] is used for this investigation. In this dataset, 2000 melodies are composed by human and the rest 8000 pitch sequences are generated by the selected music generation systems: Music VAE (VAE-based), MidiNet (DCGAN-based) and a Music Transformer (transformer-based).

II. RESULTS

Figure 1 shows the Probability Density Functions(PDF)s of the proposed features for each system as well as human composers. PDFs are estimated by a Gaussian kernel with 0.2 bandwidth. Table 1 lists the 1-Wasserstein distances of the feature distributions from systems to human.

Features	VAE-based	GAN-based	Transformer-based
IM	0.184	**0.029**	0.061
ISD	6.008	4.084	**2.218**
TSD (10^{-3})	6.397	5.051	**1.789**

Table 1: Distances of Feature Distributions from Systems to Human

Results show that MidiNet (GAN) approximates IM better than Music VAE and Music Transformer, whereas Music Transformer outperforms the other two systems in ISD and TSD. This reveals that different types of music generation system are good at learning different types of pitch-related features.

A possible explanation of GAN outperforming in IM can be the CNN architecture of MidiNet and the pixel-level pianoroll representation. The IM is close to zero (like human as displayed in the first plot), meaning that the pitch sequence has similar pitches in the beginning and at the end. The long tail shown in the ISD distribution of human composed melodies indicates that most intervals in a melody are within 12 semitones. As to TSD, human show a single-peak distribution while candidate systems learned two peaks instead. In general, the Transformer-based model shows a better approximation of these two characteristics than the VAE-based and GAN-based.

III. REFERENCES

[1] S. Li, Y. Jing, and G. Fazekas, "A Novel Dataset for the Identification of Computer Generated Melodies in the CSMT Challenge," in *Proceedings of the 8th Conference on Sound and Music Technology*. Springer Singapore, 2021, pp. 177–186.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Sketching Sounds: Using sound-shape associations to build a sketch-based sound synthesiser

Sebastian Löbbers[1*] and George Fazekas[1]

[1*]Centre for Digital Music, Queen Mary University of London, U.K., s.lobbers@qmul.ac.uk

***Abstract*— Digital synthesisers are an integral part of modern music, but their complex controls can make it difficult to realise sound ideas in a straightforward way. This extended abstract gives an overview of a research project that aims to harness research into cross-modal associations between musical timbre and shapes to develop an intuitive control interface that can produce sound from a visual sketch input.**

I. BACKGROUND

Cross-modal associations between shapes and sounds have been researched extensively in a theoretical context [1], but only little research has been conducted on how they could be used for sound synthesis or retrieval. Recent advancements in deep learning for sketch recognition, in particular Google's *QuickDraw!* project [2], can inspire new mapping architectures for sketch-driven sound applications as demonstrated by Engeln et al. [3].

II. RESEARCH OVERVIEW

This research is centred around human participant studies and can broadly be divided into two parts that contribute to the development of a sketch-based sound synthesiser. On one side, perceptual studies are conducted to find out the different ways in which humans represent timbre through simple visual sketches and, on the other side, interface design and system usability studies are needed to investigate how this synthesis system could be incorporated into music production.

A first study [4], where twenty-eight participants were asked to sketch their associations with ten different sounds, showed that a mixture of *abstract* (lines, shapes etc.) and *realistic* (objects like musical instruments or scenes like ocean waves) representations can be expected if no restrictions are imposed. Quantitative analysis produced significant correlations between visual and audio features mainly in *abstract* sketches that align with existing sound-shape association research. A second evaluation study showed that participants matched these sketches to their related sound significantly higher than the random baseline, suggesting that at least some sound characterises can be communicated through simple visual representations.

These findings informed a series of three interface design studies with the purpose of finding a setup that guides users towards simple, abstract sketches while maintaining a high level of perceived expressivity. The resulting interface was used to collect sketch representations of a synthesiser dataset by Hayes and Saitis [5] from eighty-eight participants. These sketches were fed into a deep learning (DL) classifier that was pre-trained on abstract sketches[1] from the *QuickDraw!* dataset to distinguish *noisy* from *calm* sketches. The model was then in-cooperated into a first functional prototype seen in Figure 1.

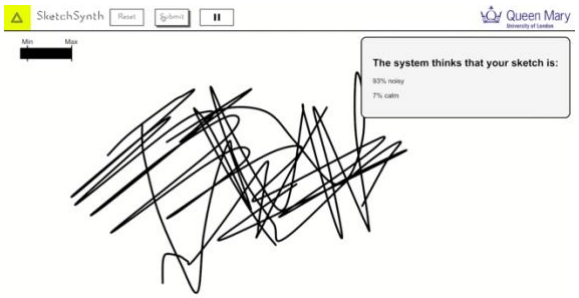

Figure 1. Screenshot of the SketchSynth prototype implemented to run in a browser. See https://youtu.be/ca1LYn8Yy-g for a demonstration video.

III. FUTURE WORK

In the next step, the ability of the prototype to produce appropriate sounds from a sketch input will be evaluated through a user study while continuing to refine and extend the DL architecture and further explore correlations between visual and audio features. A key point of this research is to find out to what extend general sound-shape association can inform the cross-modal mapping and how this system could adapt to individual representational styles to become more robust and nuanced in a music production context.

ACKNOWLEDGMENT

EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

REFERENCES

[1] Adeli M, Rouat J, Molotchnikoff S. *Audiovisual correspondence between musical timbre and visual shapes*. Frontiers in human neuroscience, 2014, p.352.

[2] Ha D, Eck D. *A Neural Representation of Sketch Drawings*. arXiv preprint arXiv:1704.03477, 2017.

[3] Engeln L, Le NL, McGinity M, Groh R. *Similarity Analysis of Visual Sketch-based Search for Sounds*. Audio Mostly, pp. 101-108.

[4] Löbbers S, Barthet M, Fazekas G. *Sketching sounds: an exploratory study on sound-shape associations*. arXiv preprint arXiv:2107.07360. 2021.

[5] Hayes B, Saitis C. There's more to timbre than musical instruments: semantic dimensions of FM sounds.

[1] *Circle, Square, Triangle, Squiggle, Zigzag* and *Line* were chosen

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Everyday Sound Recognition with Limited Annotations

Jinhua Liang, Huy Phan, Emmanouil Benetos

Centre for Digital Music, Queen Mary University of London, United Kingdom
$\{jinhua.liang, h.phan, emmanouil.benetos\}$ @$qmul.ac.uk$

I. BACKGROUND AND MOTIVATION

Everyday sound recognition aims to detect and classify the types of everyday sounds in a recording or online streaming. As a core technology in machine listening, it has many potential applications, including hearing aids, smart devices, and audio retrieval. However, it has been hard to be tackled in the past decades due to the large variety of sound sources, highly different acoustic characteristics, and the complex sound context. Nowadays the quantum leap of machine learning makes it possible to recognise the occurrence of various everyday sounds by mapping raw audio to a latent feature representation. This project thus intends to take everyday sound recognition as the research object.

II. RELATED WORK

Several existing research approaches [1, 2, 3] were dedicated to adopting a series of deep learning methods into everyday-sound tasks. For instance, Kong et al. [1] proposed a stacked convolutional neural network (CNN) to extract spectral information in input spectrograms. Lezhenin et al. [2] designed a long short-term memory network to model long-term temporal dependencies between frames. Gong et al. [3] adopted a self-attention model to address audio classification tasks. Subsequent work attempted to propose a variant or a combination of the existing models. Despite their state-of-the-art results, these techniques usually require large amounts of annotated data, which is especially difficult to get in sound recognition tasks as it takes great effort for annotators to label recordings frame by frame.

III. METHODOLOGIES

The goal of this project is to investigate and propose advanced everyday sound recognition algorithms using limited annotations. The term "limited annotations" refers to (i) the amount of development samples is small, or (ii) the available development samples are not annotated at the frame level. This project intends to address this problem by embracing the progress in self-supervised learning and model design, which are unfolded in the following.

Self-supervised learning is proposed to obtain feature representations that are semantically meaningful via pretext tasks where it is easy to access large amounts of unlabelled training data. There already exist many works related to self-supervised learning in audio related domains [4], but most of them focus on speech- or music- related fields. Since the characteristics of everyday sounds are quite different from the counterpart of speech and music (e.g., everyday sounds do not have as clearly defined units as speech and music do), it is of importance to investigate how to devise appropriate pretext tasks for optimal acoustic representation in sound recognition tasks.

Model design mainly focuses on improving the system performance by modifying its structure. The nature of model design is adding some constraints or inductive biases to existing models. This calls for a priori knowledge in the targeted domain. The duration and the frequency range of everyday sounds are more variable than speech and music. Therefore, it is promising to design a multi-resolution network that captures acoustic features with highly different characteristics. In addition, existing works [5] have indicated the importance of a proper receptive field in CNN design, but few of them tried to investigate the underlying reason. Thus, it is also interesting to investigate how appropriate model design can affect sound recognition performance.

IV. ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/T518086/1].

V. REFERENCES

[1] Q. Kong, I. Turab, X. Yong, W. Wang, and M. D. Plumbley, "DCASE 2018 challenge surrey cross-task convolutional neural network baseline," DCASE2018 Challenge, Tech. Rep., September 2018.

[2] I. Lezhenin, N. Bogach, and E. Pyshkin, "Urban sound classification using long short-term memory neural network," in *2019 federated conference on computer science and information systems (FedCSIS)*. IEEE, 2019, pp. 57–60.

[3] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[4] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2019.

[5] C.-Y. Wang, J.-C. Wang, A. Santoso, C.-C. Chiang, and C.-H. Wu, "Sound event recognition using auditory-receptive-field binary pattern and hierarchical-diving deep belief network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1336–1351, 2017.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Generating Comments from Music and Lyrics

Yixiao Zhang and Simon Dixon*

C4DM, Queen Mary University of London, UK
yixiao.zhang@qmul.ac.uk

***Abstract*— Generating comments based on music is a fairly new topic in the MIR field. In this paper, we propose a deep generative model based on contrastive learning and BART to generate music comments based on music and lyrics.**

***Index Terms*—** Deep Generative Model, BART, Multimodal Representation Learning

I. INTRODUCTION

In recent years, research related to cross-modal generation of music and text has gradually developed, such as generating lyrics from music [1] and generating music from descriptions [2]; [3] uses audio thumbnails to summarise lyrics. However, the work of generating subjective interpretations of music is relatively limited. In order to do so, the model must integrate multimodal and multiclass information to understand music and be able to generate text with high relevance and natural flow. In this paper, we propose a cross-modal deep learning model to generate high quality comments from music and lyrics.

II. METHOD

Our proposed model takes the lyrics and the music spectrogram as input and outputs a comment on this music. The model is divided into two parts, the encoder-decoder module based on BART [4], and the condition module based on multimodal contrastive learning. The whole training process is divided into two stages: in the first stage, the conditional model and BART are pre-trained separately; in the second stage, the two models are trained jointly.

Condition Module. We expect the model can capture global features such as emotion and style from the music. Considering that music metadata can help to interpret music to some extent, we use a multimodal approach based on [5], to incorporate information from music metadata into the music encoding. In the pre-training stage, the music spectrogram and music labels are encoded using FCN and Transformer respectively, and the two representations are made to fuse in a contrastive learning manner; in the joint training stage, the conditional model uses only the music spectrogram as the only input.

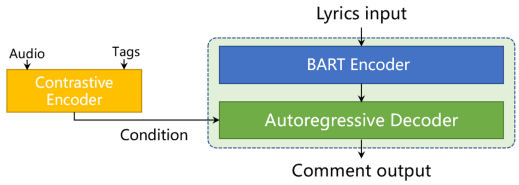

Figure 1: The diagram of our proposed model.

BART Encoder-Decoder. After a preliminary analysis we found that there is a close correspondence between the comments and the lyrics. Therefore we finetune a seq2seq model from lyrics to comments based on a pretrained BART model, making the model understand lyrics and generate interpretations. In the joint training phase, the conditioning is added to the decoder following [6], therefore the decoder can get knowledge both from music and lyrics.

Dataset. We create a new dataset with song titles and metadata from the Music4All Dataset [7] and the corresponding comments for the songs from SongMeanings.com [1]. After processing, we obtain 27,834 songs and the corresponding approximately 490,000 comments.

III. REFERENCES

[1] Z. Sheng, K. Song, X. Tan, Y. Ren, W. Ye, S. Zhang, and T. Qin, "Songmass: Automatic song writing with pre-training and alignment constraint," *arXiv preprint arXiv:2012.05168*, 2020.

[2] Y. Zhang, Z. Wang, D. Wang, and G. Xia, "Butter: A representation learning framework for bi-directional music-sentence retrieval and generation," in *Proceedings of the 1st workshop on NLP for music and audio (nlp4musa)*, 2020, pp. 54–58.

[3] C. I. I. France, Université Côte d'Azur, M. Fell, E. Cabrio, F. Gandon, and A. Giboin, "Song Lyrics Summarization Inspired by Audio Thumbnailing," *Proceedings - Natural Language Processing in a Deep Learning World*, pp. 328–337, 2019.

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[5] A. Ferraro, X. Favory, K. Drossos, Y. Kim, and D. Bogdanov, "Enriched music representations with multiple cross-modal contrastive learning," *IEEE Signal Processing Letters*, vol. 28, pp. 733–737, 2021.

[6] C. Li, X. Gao, Y. Li, B. Peng, X. Li, Y. Zhang, and J. Gao, "Optimus: Organizing sentences via pre-trained modeling of a latent space," *arXiv preprint arXiv:2004.04092*, 2020.

[7] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues, *et al.*, "Music4all: A new music database and its applications," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2020, pp. 399–404.

*Yixiao Zhang is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by the China Scholarship Council and Queen Mary University of London, with additional support from Apple.

[1] https://songmeanings.com/

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

AI-Assisted FM Synthesis

Franco Caspe*, Andrew McPherson and Mark Sandler

Centre for Digital Music/Queen Mary University of London, UK, f.s.caspe@qmul.ac.uk

***Abstract*— Frequency Modulation (FM) synthesis is a well-known technique that is used to create interesting timbres at a low computational cost. Recent FM commercial products have seen a resurgence, due to FM's great timbral possibilities, but they still rely on dated and complex sound design paradigms. Scaling up the architecture to improve it seems to be unfeasible due to the increase in complexity it would entail. On the other end of the spectrum, Deep Neural Networks (DNNs), widely employed as classifiers, have been recently used on different generative schemes to classify or to produce musical instrument samples. Moreover, recent works exploit their descriptive power in order to directly control oscillators and filters.**

In our project, we aim to develop a DNN that can describe natural-sounding spectra in terms of the parameters of an FM synthesizer. Obtaining such a decomposition can pave the way to develop novel gestural control strategies or even musical instrument transformations.

***Index Terms*—** Sound matching, FM synthesis, Deep Neural Networks, Instrument Augmentation

I. PROJECT DESCRIPTION

Frequency Modulation (FM) synthesis, firstly presented by John Chowning [1], is an economical mean of generating complex time-varying spectra, by routing simple parametrized signal generators, called *operators*, under different modulator-carrier composition schemes called *algorithms*. Popular throughout the 1980s and '90s, the method fell out of use due to the lack of fine control over the sound and it's sometimes undesirable sonic characteristics [2]. To cope with this, several tools and modifications have been proposed, such as strategies to improve the FM spectra [3] [4], models for gestural control mapping [5] [6], and sound matching techniques [7, 8, 9].

FM has lately regained a fair quote of attention in the music community, with new synthesizers and emulators being released, to name a few: Korg OpSix, Yamaha DX Reface, Elektron Digitone, Korg Volca FM and Dexed. However, their architecture and sound design approach remained similar to that of the Yamaha DX7, a classic six operators architecture from 1983.

Pairing a Deep Neural Network with a synthesizer to describe sounds in terms of its parameters provides a framework for creative control strategies. First, it allows a sound designer to approximate a target sound, being able to continue the workflow with manual fine-tuning if desired [10]. Another interesting possibility is the distillation of mapped meta-controls [11] that can manipulate multiple parameters at the same time in a sonically meaningful way [9]. Finally, we believe that a real-time implementation of such a solution could become an interesting tool for instrument retargeting or intelligent augmentation strategies.

*The author is funded by EPSRC and UKRI under the Centre for Doctoral Training in Artificial Intelligence and Music at Queen Mary University of London (Grant EP/S022694/1).

II. REFERENCES

[1] J. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *Journal of the Audio Engineering Society*, vol. 21:526–534., 1973.

[2] V. Lazzarini, J. Timoney, and T. Lysaght, "The generation of natural-synthetic spectra by means of adaptive frequency modulation," *Computer Music Journal*, vol. 32, pp. 9–22, 06 2008.

[3] B. Schottstaedt, "The simulation of natural instrument tones using frequency modulation with a complex modulating wave," *Computer Music Journal*, vol. 1, no. 4, pp. 46–50, 1977. [Online]. Available: http://www.jstor.org/stable/40731300

[4] V. Lazzarini and J. Timoney, "Theory and practice of modified frequency modulation synthesis," *Journal of the Audio Engineering Society*, vol. 58, pp. 459–471, 06 2010.

[5] M. M. Wanderley and P. Depalle, "Gestural control of sound synthesis," *Proceedings of the IEEE*, vol. 92, no. 4, pp. 632–644, 2004.

[6] E. Miranda and M. Wanderley, "New digital musical instruments: Control and interaction beyond the keyboard (computer music and digital audio series)," 2006.

[7] A. Horner, J. Beauchamp, and L. Haken, "Machine tongues xvi: Genetic algorithms and their application to fm matching synthesis," *Computer Music Journal*, vol. 17, no. 4, pp. 17–29, 1993. [Online]. Available: http://www.jstor.org/stable/3680541

[8] N. Masuda and D. Saito, "SYNTHESIZER SOUND MATCHING WITH DIFFERENTIABLE DSP," *ISMIR 2021*, p. 7, 2021.

[9] G. Le Vaillant, T. Dutoit, and S. Dekeyser, "Improving synthesizer programming from variational autoencoders latent space," *DAFx 2021*, 06 2021.

[10] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos, "Universal audio synthesizer control with normalizing flows," *arXiv:1907.00971 [cs, eess, stat]*, July 2019, arXiv: 1907.00971. [Online]. Available: http://arxiv.org/abs/1907.00971

[11] A. Hunt, Y. Dd, M. Wanderley, and M. Paradis, "The importance of parameter mapping in electronic instrument design," *Journal of New Music Research*, vol. 32, 05 2002.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Algorithmic Music Composition for The Environment

Rosa Park[1*]

[1*] School of Cinema, San Francisco State University, United States, rosapark@sfsu.edu

***Abstract*— "Algorithmic Music Composition for The Environment" is an interactive sound performance that represents scientific data of global warming and climate change. Playing along with the MIDI-equipped interactive interface, "Algorithmic Music Composition for The Environment" aims to reflect the impacts of the climate crisis through sound by representing the alarming records of diverse environmental sectors, such as global land-ocean temperature, Sea Level change, Antarctic Ice mass variation, atmospheric carbon dioxide (CO2) levels, and more. There have been several ongoing collaborative projects among scientists, artists, and musicians in the Bay Area to combat climate change and bring the urgency of this pressing issue to inspire people to take meaningful action through music [1]. Thus, the development of this project is aligned with those endeavors to strengthen collaborative efforts and interdisciplinary solutions, seeking new methods and techniques of experimental music that can raise awareness of environmental challenges.**

I. DESCRIPTION OF THE MUSICAL WORK

The main interface for the music composition of the project has been built in Pure Data (Pd, https://puredata.info/), a data flow programming language for electronic music. The Pd interface of the work is composed of various types of Graphical User Interface (GUI) objects in which the scientific data is stored in the form of tables. These tables contain the information of a growing number of weather-related catastrophes, including Land-Ocean Temperature from 1880 to 2020, Global Mean Sea Level (GMSL) variations data between 1993 and 2021, the monthly records of ocean heat from 1957 to 2020, Antarctic Sea Ice Extent from 1978 to 2020, and CO2 emission trends from 1958 to 2021 measured by five different scientific research organizations that are NASA, NOAA Climate.gov, United States Environmental Protection Agency (US EPA), and the U.S. National Climate Assessment [2][3][4][5].

Values stored in the tables within Pd draw line graphs. Each table expresses unique sound qualities and textural complexities, reflecting regional and seasonal temperature extremes for each year and month since 1880. Figure 1. below shows the examples of the table compositions used in Pd.

II. THE COMPOSITIONAL PROCESS

Sounds generated algorithmically from the table arrays are played through the main Pd interface. The performer controls and improvises on the generated sounds through the GUI modules (Fig. 2) to respond to the trends of the latest climate data, interpreting a sense of urgency about the climate crisis.

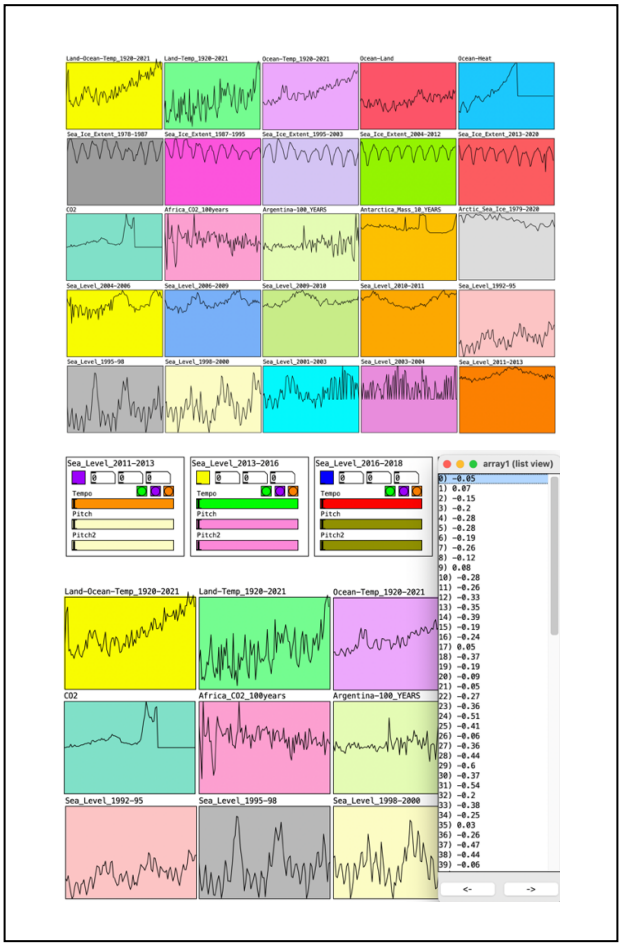

Figure 1. Table compositions and line graphs with the values implemented and the datasheet of Territorial emissions in $MtCO_2$ ($MtCO_2$: 1 million tons of CO2).

The key indicators of the GUI modules affect and change the sonic textures, such as tempo, pitch, note, and octave dramatically to provoke more compelling experiences of the increasing effects of climate change and ultimately portray its catastrophic consequences in the future.

The sonification process allows the performer to add the conceptual domain to the soundscape by enhancing or revealing several notable troubling trends in the data through the main interface system (Fig. 3), which constantly plays sound based on the numbers implemented in the table arrays. By turning data into sound, the project aims to bring a message that climate change is far more than an environmental issue; it is the *cry* of the Earth, and the consequences of climate change are already here.

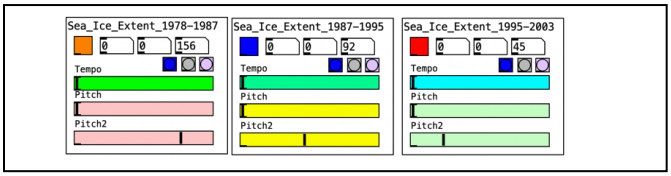

Figure 2. The GUI modules for "Arctic Sea Ice Extent."

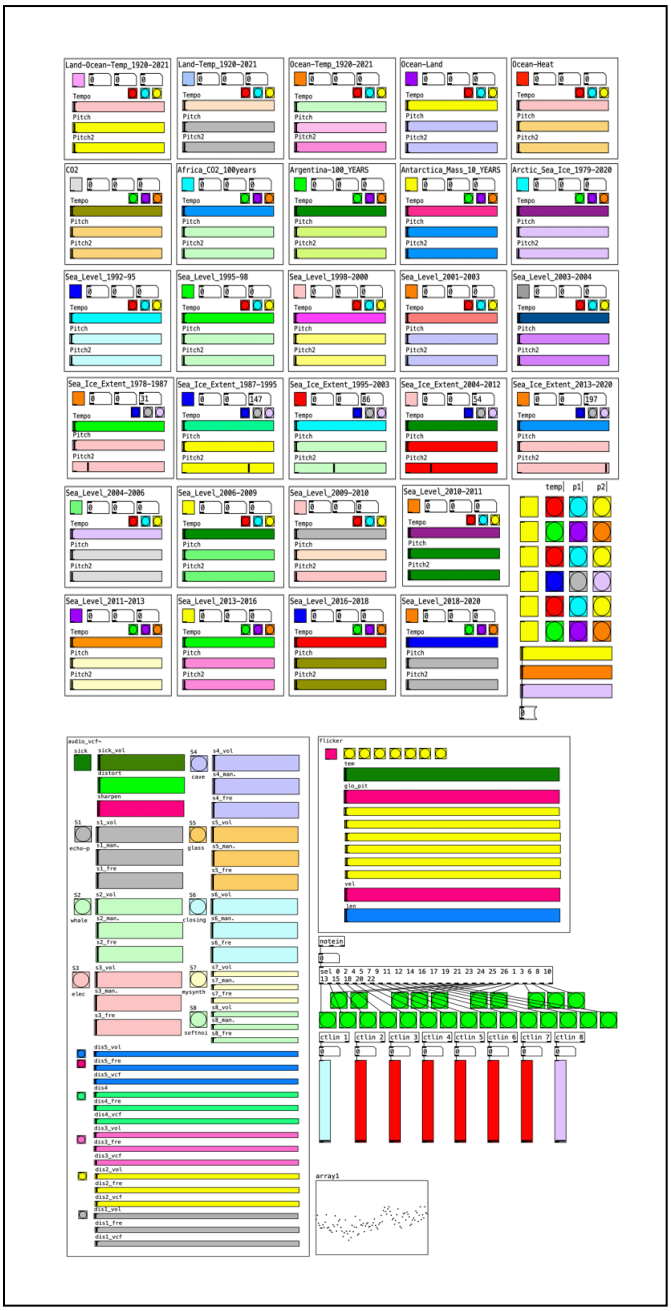

Figure 3. The main Pd interface design for "Algorithmic Music Composition for The Environment."

REFERENCES

[1] G. O. Young, "Synthetic music (Book style with paper title and editor)," in *Music Technology*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1984, pp. 15–64.

[2] W.-K. Smith, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[4] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electronic music studies" *IEEE Transl. J. Mus. Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Music* Japan, 1982, p. 301].

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

The Vienna Philharmonic's New Year's Concert Series: A Corpus for Digital Musicology and Performance Science

David M. Weigl and Werner Goebl*

Dept. of Music Acoustics – Wiener Klangstil, University of Music and Performing Arts Vienna, Austria, {lastname}@mdw.ac.at

***Abstract*— We present *Signature Sound Vienna*,[1] an ongoing project to collect, interrelate, and analyse performance recordings pertaining to the Vienna Philharmonic's New Year's Concert series, combining approaches from historical musicology, performance science, music informatics, and Web science.**

***Index Terms*—** Digital Musicology, Performance Science, Linked Data, Music Encoding

I. VIENNA'S NEW YEAR'S CONCERTS

The Vienna Philharmonic Orchestra (VPO)'s yearly New Year's Concert broadcast provides listening enjoyment to tens of millions in nearly 100 countries. The series has featured a variety of conductors, compositions, and composers, but incorporates the same favourites–most notably, *An der schönen blauen Donau* (Blue Danube Waltz; Johann Strauss II) and *Radetzkymarsch* (Radetzky March; Johann Strauss I)–year after year. The ever-repeating, ever-changing nature of these concerts make them appealing for musicological analysis: How have the performances evolved over time? Are changes explicable using historical factors, e.g., based on the conductor or concert master? Can we find signatures of the VPO, compared to other orchestras' performances of these compositions? How about other Viennese, versus international, orchestras–does Vienna really have a signature sound?

II. REPERTOIRE: COMPOSITIONS AND RECORDINGS

We have restructured data from the VPO concert archive[2] to determine the most frequently performed works in the series (Fig. 1). We are engaging in a media-purchasing campaign incorporating online and physical record stores, second-hand media and auction websites, as well as Vienna's music flea-markets to acquire commercial recordings of the VPO New Year's Concerts, alongside recordings of other orchestras performing (some of) this pertinent material.

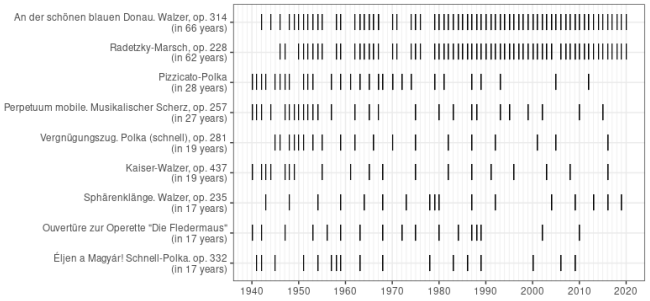

Figure 1: New Year's Concert: Most frequently performed compositions

III. CORPUS HYPERSTRUCTURE AND ANALYSES

We will establish a richly interlinked corpus describing salient aspects of our repertoire, by: *i)* quantifying our performance recordings using audio feature extractors; *ii)* encoding the corresponding music scores, and performing audio-to-score and audio-to-audio alignments to interlink the timelines of our recordings with score elements at the note level; and *iii)* identifying historical data relating to the orchestras, the performers, and the reception of the performances from catalogue sources, blog postings and newspaper archives, in order to contextualise the recordings within their performance events. These diverse data will be represented using established vocabularies, interrelated with online music authorities including MusicBrainz and AcousticBrainz,[3] and published as Linked Open Data, making the data reproducible, reusable, and reinterpretable beyond the immediate scope of our project. We will engage in musicological scholarship to interrogate and characterise this corpus, formulating hypotheses which will drive the development of score- and audio-feature-informed analyses. We will report our findings within hypermedia narratives published as Web apps tying together scholarly arguments with multimedia resources providing their empirical basis. This approach builds on recent trends in Digital Musicology, applied for the first time to an extensive but focused collection of performances.

IV. ACKNOWLEDGMENTS

Project collaborators: Fritz Trümpi, Markus Grassl, Chanda VanderHart, Delilah Rammler, Matthäus Pescoller.

*This research was funded in whole, or in part, by the Austrian Science Fund (FWF) P 34664-G. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

[1] https://iwk.mdw.ac.at/signature-sound-vienna

[2] https://www.wienerphilharmoniker.at/en/konzert-archiv

[3] https://musicbrainz.org

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

An Interactive Tool for Visualising Musical Performance Subtleties

Yucong Jiang

Department of Math & Computer Science, University of Richmond, USA, yjiang3@richmond.edu

***Abstract*— I introduce a tool that visualises key aspects of a performance (such as the tempo), built on Sonic Visualiser and utilizing audio-to-score alignment techniques. I describe an application scenario of this tool in piano education.**

***Index Terms*—** Sonic Visualiser, audio-to-score alignment, piano education, musical performance

I. PERFORMANCE VISUALISATION

One of the most popular tools for studying musical performance recordings is Sonic Visualiser [1], with which users can visualise features of an audio recording. The scope of the features is largely defined by existing audio analysis plugins: e.g., various low-level signal features or fundamental frequency estimation [2]. I introduce a prototype tool built on Sonic Visualiser, with modifications to the layout of the main window and an additional plugin for automatic analysis. This tool aims to visualise subtleties of a musical performance (e.g., tempo control, loudness, or articulation). Given a digital score and a performance recording of this score, this tool first automatically aligns the audio to the score (based on [3]), and then displays key aspects of the performance, while highlighting the corresponding score positions in the sheet music as the user navigates different sections of the recording. Fig. 1 depicts the main window of this tool. The upper pane shows the spectrogram and the onsets of each chord, and the lower pane displays one essential aspect of the performance—in this prototype, the tempo fluctuation of the performance. The next section describes an application scenario of this tool useful in piano education.

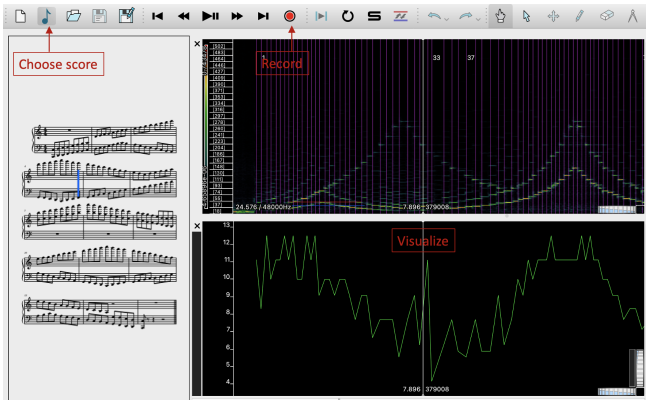

Figure 1: The main window of this tool.

II. AN APPLICATION EXAMPLE: PIANO EDUCATION

Making music is about much more than playing the correct notes. The purpose, as applied here, is to help learners reflect on the quality of their playing, and evaluate it critically by examining their practice recordings in greater detail. Through this process, students should gradually improve their self-reflection skills. There are three steps (Fig. 2) involved. The user first chooses the score to practice, which then appears in the left pane of the main window. The user then clicks the "record" button and starts to play. After each recording, the performance is automatically analysed and visualised in the two panes on the right. At this point, the user may wish to reflect on the past performance by playing back the recording (perhaps multiple times) and observing the visual feedback in the lower pane (in this instance, tempo). The user may keep practicing this score by repeating Step 2 and Step 3, working to improve the quality of play through insights gained from previous takes. After any take, the user may switch to a different score by going back to Step 1.

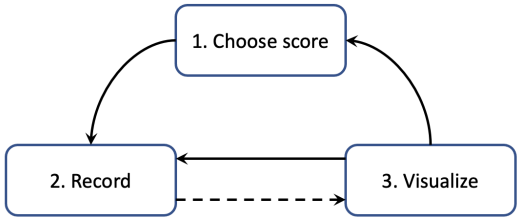

Figure 2: Three steps to this application of the tool. The dashed arrow represents an automatic process.

This tool can also be used to collect and analyse student practice data. The results of this analysis could be used to support a data-driven approach to studying musical learning, and facilitate new pedagogical techniques. In addition, this tool may also contribute to musicology research by making it easier to analyse large-scale performance data sets.

III. REFERENCES

[1] C. Cannam, C. Landone, and M. Sandler, "Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the ACM Multimedia 2010 International Conference*, Firenze, Italy, October 2010, pp. 1467–1468.

[2] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, 2014, in press.

[3] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden Markov models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 4, pp. 360–370, 1999.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

A Benchmark Dataset to Study Microphone Mismatch Conditions for Piano Multipitch Estimation on Mobile Devices

Jakob Abeßer, Franca Bittner, Maike Richter, Marcel Gonzalez, Hanna Lukashevich

Semantic Music Technologies, Fraunhofer IDMT, Ilmenau, Germany, jakob.abesser@idmt.fraunhofer.de

***Abstract*— In this paper, we present the IDMT-PIANO-MM dataset, which allows to evaluate piano transcription algorithms under microphone mismatch conditions. In particular, we discuss specific constraints that these algorithms need to face when being used in music learning applications on mobile devices. Then, we describe the dataset w.r.t. recording locations and devices as well as the recorded music pieces. We intend this dataset to be a public benchmark to evaluate the robustness of AI-based MPE models within realistic microphone-mismatch conditions, which are to be expected with the large number of potential users of music learning applications.**

***Index Terms*—** Multipitch estimation, piano transcription, microphone mismatch, mobile devices

I. INTRODUCTION

In the field of Music Information Retrieval (MIR), the pitch detection of multiple simultaneous tones (multipitch estimation, MPE) is a challenging research task. MPE is commonly approached by recognizing characteristic patterns such as fundamental frequencies and their corresponding overtones in spectrogram representations. Traditional methods use decomposition techniques such as Non-Negative Matrix Factorization (NMF) whereas recent methods solely focus on deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [1].

A particularly interesting application scenario for multipitch estimation algorithms are music learning applications. Here, audio recordings of musical performances need to be transcribed and compared to a given reference notation with near real-time latency in order to assess the user's performance. Furthermore, music learning applications need to run on mobile devices, which have limited computational resources and microphones of very different quality. As a consequence, these constraints limit the complexity of the applied MPE algorithm. Finally, the applied MPE algorithm needs to be robust to acoustic parameters such as reflection times in the users' practice rooms. It has been widely observed that AI-based audio analysis algorithms exhibit a performance drop in domain shift scenarios, which are caused for instance by a microphone mismatch between the initial training data and the test data.

II. DATASET

Established evaluation datasets for piano transcription include audio files from YouTube videos[1], recordings of acoustic grand pianos such as the Yamaha Disklavier[2], or synthesized using professional sample libraries[3]. These datasets mostly lack the required variety of instrument models as well as metadata details about the spatial parameters of the recording locations. As the main contribution of this work, we present the IDMT-PIANO-MM dataset[4], which allows to study microphone mismatch conditions for piano multipitch estimation recorded with mobile phones. The dataset includes a total of 432 piano recordings (around four hours), which cover nine music pieces recorded in eight different rooms using six different recording devices. The pieces cover classical music (B. Bartók, W. A. Mozart, J. Pachelbel, and L. v. Beethoven) as well as jazz (S. Joplin as well as own compositions) and range from simple to medium difficulty. All music pieces are in the public domain. The recording locations range from small rooms to a large lecture hall. Information about the room geometries, piano position within the room, as well as wall materials are documented. The rooms include four different grand pianos, three upright pianos, and one stage-piano. At each location, audio recordings were made with three mobile phones (iPhone 6S Plus, Redmi Note 8, LG G6), two tablets (iPad Air 2, Amazon Fire tablet), and one stereo setup using two high-quality Oktava MK 012 microphones in an AB recording setup. In our presentation, we will show the results of an initial data inspection focusing on properties such as the dynamic range of the recordings. Also, we compare the different microphone characteristics using the spectrum correction method proposed in [2].

III. REFERENCES

[1] 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, 2019.

[2] M. Kośmider, "Calibrating Neural Networks for Secondary Recording Devices," DCASE2019 Challenge, Tech. Rep., 2019.

[1] e. g., Giant-MIDI Piano, https://github.com/bytedance/GiantMIDI-Piano

[2] e. g., MAESTRO, https://magenta.tensorflow.org/datasets/maestro, MAPS https://hal.inria.fr/inria-00544155/en

[3] e. g., SMD-Synth https://zenodo.org/record/4637908

[4] https://www.idmt.fraunhofer.de/en/publications/datasets.html

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Looking at the Future of Data-Driven Procedural Audio

Adrián Barahona-Ríos*

Department of Computer Science, University of York, UK, ajbr501@york.ac.uk

***Abstract*— Data-driven methods can be seen as an alternative to pure digital signal processing (DSP) approaches for the sound synthesis of sound effects. We present an overview of some of the advancements in machine learning for this task, hinting at what the future of data-driven procedural audio could sound like.**

***Index Terms*—** Procedural Audio, Sound Synthesis, Game Audio, Sound Design, Neural Synthesis

I. DATA-DRIVEN PROCEDURAL AUDIO

While procedural audio systems are usually built upon DSP-based synthesisers running in real-time [1], creating bespoke procedural audio models in a timely manner with the sound quality required by the video game and interactive media industries still remains a challenge. Data-driven approaches could help to overcome this issue by generating sounds directly, in combination with DSP methods or by helping with aspects of the creative process.

One option can be the neural synthesis of one-shot sound effects, where a generative deep learning model trained on a corpus of sounds directly produces novel assets on demand. The sound synthesis can be driven by categorical descriptors of the training dataset, such as in the case of emotions in knocking sound effects [2] or surfaces in footsteps [3]. It is also possible to condition the synthesis with high-level audio attributes such as timbral features [4]. Training the models on a single sound is an option as well (especially convenient when the category of sound effects is rare), such as in SpecSinGAN [5] for generating arbitrary one-shot variations and Catch-A-Waveform [6] for synthesising longer sequences.

Another approach could be the combination of DSP with deep learning techniques. Among others, differentiable spectral modeling synthesisers have been used for pitch-conditioned sound synthesis [7] or engine sounds conditioned on the revolutions per minute [8]. Other DSP methods such as waveshaping synthesis have been explored in a musical context too [9]. Deep modal synthesis has also been considered, with the possibility of generating real-time impact sounds from arbitrary 3D shapes with different materials [10]. Tangentially, yet another option could be to use deep learning to automatically program a synthesiser to produce a target sound from a provided audio asset [11].

Regarding alternative ways of interacting with neural audio synthesisers, sound designers could discover new sounds by inpainting (masking and automatically reconstructing) spectrograms [12], providing an onomatopoeic word [13] or even a set of frames from a video [14]. High-quality timbre transfer can be performed as well [7][9][15], such as transforming speech into violin sounds.

There are however multiple challenges still to be addressed, such as interfacing, co-creativity, interactivity, efficiency, evaluation metrics or the overall quality and flexibility of the models to name a few.

*Research supported by the EPSRC Centre for Doctoral Training in Intelligent Games & Game Intelligence (IGGI) [EP/L015846/1] and Sony Interactive Entertainment Europe.

II. REFERENCES

[1] A. Farnell, *Designing Sound*. Mit Press, 2010.

[2] A. Barahona-Ríos and S. Pauletto, "Synthesising Knocking Sound Effects Using Conditional WaveGAN," in *17th Sound and Music Computing Conference, Online*, 2020.

[3] M. Comunità, H. Phan, and J. D. Reiss, "Neural Synthesis of Footsteps Sound Effects with Generative Adversarial Networks," *arXiv preprint arXiv:2110.09605*, 2021.

[4] J. Nistal, S. Lattner, and G. Richard, "DrumGAN: Synthesis of Drum Sounds With Timbral Feature Conditioning Using Generative Adversarial Networks," *arXiv preprint arXiv:2008.12073*, 2020.

[5] A. Barahona-Ríos and T. Collins, "SpecSinGAN: Sound Effect Variation Synthesis Using Single-Image GANs," *arXiv preprint arXiv:2110.07311*, 2021.

[6] G. Greshler, T. R. Shaham, and T. Michaeli, "Catch-A-Waveform: Learning to Generate Audio from a Single Short Example," *arXiv preprint arXiv:2106.06426*, 2021.

[7] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," *arXiv preprint arXiv:2001.04643*, 2020.

[8] A. Lundberg, "Data-Driven Procedural Audio: Procedural Engine Sounds Using Neural Audio Synthesis," 2020.

[9] B. Hayes, C. Saitis, and G. Fazekas, "Neural Waveshaping Synthesis," *arXiv preprint arXiv:2107.05050*, 2021.

[10] X. Jin, S. Li, T. Qu, D. Manocha, and G. Wang, "Deep-Modal: Real-Time Impact Sound Synthesis for Arbitrary Shapes," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1171–1179.

[11] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos, "Flow Synthesizer: Universal Audio Synthesizer Control With Normalizing Flows," *Applied Sciences*, vol. 10, no. 1, p. 302, 2020.

[12] T. Bazin, G. Hadjeres, P. Esling, and M. Malt, "Spectrogram Inpainting for Interactive Generation of Instrument Sounds," *arXiv preprint arXiv:2104.07519*, 2021.

[13] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "Onoma-to-Wave: Environmental Sound Synthesis From Onomatopoeic Words," *arXiv preprint arXiv:2102.05872*, 2021.

[14] V. Iashin and E. Rahtu, "Taming Visually Guided Sound Generation," *arXiv preprint arXiv:2110.08791*, 2021.

[15] A. Caillon and P. Esling, "RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

Making graphical scores accessible to visually impaired people: A haptic interactive installation.

Christina Karpodini
c.karpodini@gmail.com

I. ABSTRACT

Graphic scores have always been a liberating way of composing, combining a score that is free from the traditional ways of writing music and while also constituting a visual art piece with independent artistic value. However, Both traditional and experimental notation systems have been unhelpful to people with low vision. The traditional notation system has been made inclusive through the braille system only in the 20th century in contrast to graphic scores that seem in many cases to be failing to be accessible to those people. In this work, I present a project that aims to fill this crucial gap and offer visually impaired participants a new experience of composing and performing with an alternative notation system. With the development of an interactive installation I am proposing the idea of making alternative notation tactile to enable this accessibility. Physical objects can act both as a notation system and as a tactile transformation of the artfulness of a graphic score. With the help of computer vision technology these new objects/notation will be interpreted into sound through an algorithmic process that will map their position in space to score properties. This tool could be used in the context of the work of specific composers that were pioneers in the field of graphical scoring such as Anestis Logothetis. Previous research has provided significant amount of analysis of his work that could be used in relation to this project. [1],[2] Using characteristic graphical patterns as individual notation we can create a series of 3D printed objects that can be used to make his work tactile and interactive thus accessible to visually impaired people.

A primary stage installation design of this project, Block's sound, has been exhibited in London and feedback forms selected by participants with focus on the effect of haptic notation and interactive composition to their perception of sound and interactivity. These feedback forms constitute the first evaluation of the tactile interactive scoring installation idea and give the floor to further development.

II. REFERENCES

[1] Michael McInerney, 2015. New Notational Strategies for New Interpretative Paradigms: Revisiting the Scores of Anestis Logothetis (1921-1994). *Perspectives of New Music*, 53(1), pp.99-120.

[2] Baveli, M. and Georgaki, A., 2008. Towards a decodification of the graphical scores of Anestis Logothetis (1921-1994) . The graphical space of Odysee(1963). In: *SMC-5th Sound and Music Computing Conference, 2008*.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Acoustic Representations for Perceptual Timbre Similarity

Cyrus Vahidi, Ben Hayes, Charalampos Saitis, George Fazekas*

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

***Abstract*— In this work, we outline initial steps towards modelling perceptual timbre dissimilarity. We use stimuli from 17 distinct subjective timbre studies and compute pairwise distances in the spaces of MFCCs, joint time-frequency scattering coefficients and Open-L3 embeddings. We analyze agreement of distances in these spaces with human dissimilarity ratings and highlight challenges of this task.**

***Index Terms*—** timbre, acoustic representations, psychoacoustics

I. METHOD

We used 17 timbre dissimilarity datasets that were compiled in a previous meta-analysis publication [1]. We share an open-source repository containing 17 dissimilarity matrices and corresponding audio sampled at 44.1 kHz[1].

We extracted temporally averaged mel-frequency cepstral coefficients (MFCCs), joint time-frequency scattering coefficients (jTFS) [2] and OpenL3 embeddings [3] for 1000ms of audio of each stimulus. We consider jTFS as it characterises *spectrotemporal modulations*, analogously to the model used in [1]. We used a window length of 25ms for MFCCs with 40 coefficients. jTFS coefficients were computed using Kymatio[2] with maximum scale $J = 8$, $Q = 12$ filters per octave, temporal averaging of $T = 1000ms$ and frequential averaging of $F = 1$ octave, yielding 869 coefficients. 512-dimensional OpenL3 embeddings were extracted using an open-source Python package[3].

Pairwise euclidean distances of the form in Eqn. (1) were computed between all embeddings within each dataset.

$$D_e(x_i, x_j) = \sqrt{(Ix_i - Ix_j)^T (Ix_i - Ix_j)} \qquad (1)$$

II. RESULTS

We collected all triplets (a, i, j) from a dissimilarity matrix, where i and j belong to the k-nearest neighborhood of an anchor a and satisfy the triplet inequality $D(a, i) < D(a, j)$.

Table 1: Mean triplet agreement using a $k = 5$ nearest neighborhood

Dataset	*MFCC*	*OpenL3*	*jTFS*
Barthet2010	0.71	0.77	0.88
Grey1977	0.57	0.64	0.61
Grey1978	0.41	0.48	0.45
Iverson1993_Onset	0.59	0.59	0.56
Iverson1993_Remainder	0.57	0.54	0.54
Iverson1993_Whole	0.59	0.66	0.64
Lakatos2000_Comb	0.55	0.53	0.55
Lakatos2000_Harm	0.64	0.73	0.61
Lakatos2000_Perc	0.53	0.55	0.48
McAdams1995	0.62	0.63	0.58
Patil2012_A3	0.65	0.65	0.65
Patil2012_DX4	0.48	0.6	0.54
Patil2012_GD4	0.58	0.64	0.45
Siedenburg2015_e2set1	0.73	0.71	0.65
Siedenburg2015_e2set2	0.68	0.69	0.61
Siedenburg2015_e3	0.58	0.56	0.5

Triplet agreement is the average number of triplets that satisfy $D_e(a, i) < D_e(a, j)$, i.e the distance ranking is respected in acoustic feature space e. Table 1 shows the mean triplet agreements per dataset using 5 nearest neighbors.

III. CONCLUSION

Initial experiments indicate that acoustic features alone are not sufficient to match perceptual distances. We highlight that the only dataset containing a homogeneous category for all stimuli, *Barthet2010*, produces a considerably higher figure than other datasets, which may suggest that its timbre space only encodes acoustical cues. Otherwise, we observe no clear differences between the representations. Further experiments will aim to learn a unified metric to approximate timbre space distances across datasets, considering specificity and categorical cues. This may give a clearer indication of the suitability of the proposed representations.

IV. REFERENCES

[1] E. Thoret, B. Caramiaux, P. Depalle, and S. Mcadams, "Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre," *Nature Human Behaviour*, vol. 5, no. 3, pp. 369–377, 2021.

[2] J. Andén, V. Lostanlen, and S. Mallat, "Joint time–frequency scattering," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3704–3718, 2019.

[3] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

*This work was supported by UK Research and Innovation [grant number EP/S022694/1]. Cyrus Vahidi is supported jointly by the UKRI and Music Tribe.

[1] https://github.com/ben-hayes/timbre-dissimilarity-metrics/

[2] https://github.com/kymatio/kymatio

[3] https://openl3.readthedocs.io/en/latest/index.html

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Investigating a computational methodology for quantitive analysis of singing performance style

Yukun Li[1] *, *Polina Proutskova*[1], *Zhaoxin Yu*[2] † *and Simon Dixon*[1]

[1]Centre for Digital Music, Queen Mary University of London, UK, yukun.li@qmul.ac.uk
[2]Shandong College Of Arts, China

***Abstract*— As an aspect of high-level cognition, style analysis is subjective and tends to be performed qualitatively. Quantitative computational analysis of singing performance can provide an objective view of style. The question is how to implement a systematic analysis, especially on a large scale of data. This talk proposes a computational methodology and lists several research questions we seek to answer: what computational framework is appropriate for the task; how to extract pitch from the audio and segment it; how to formalise the characteristics of a performance style; how the musical content affects the performance style; how to model and measure musical features such as pitch contour or dynamics. The exploration so far is summarised into some assumptions, possible solutions and preliminary experimental results, which we explain in this talk.**

***Index Terms*—** Computational musicology, Singing performance style, Note segmentation, Pitch modelling

Our research involves the analysis of recordings containing mixtures of vocals and instruments. We utilize automatic source separation to extract the monophonic vocal track. Then we extract features such as f0, amplitude, spectral flux and phonemes. An HMM note tracker is applied to output note segments, which consist of sub-regions labelled with their state (attack, sustain, release or transition). Based on the estimated notes, the following aspects of performance style can be investigated: how the singer begins or ends a note, how subsequent notes are connected, and how notes are sustained by the singer.

Note is a very fundamental conception to help us understand music. However, note segmentation is difficult because the vocal pitch trace is continuous and unstable [1], which leads to disagreements among human judgements. Different purposes or musical backgrounds of annotators result in different segmentations. Currently, automatic vocal transcription systems are much less robust to different styles and less precise than human annotations [2]. We propose ideas to build a controllable model which could potentially improve the performance of note segmentation in terms of generality and accuracy.

Then vocal style can be investigated based on notes. Up to now, we have found two examples of vocal style which we plan to analyse on a larger scale. The first concerns the timing of note boundaries. In the pop music from the NUS-48E dataset [3] there are many intra-vowel glides and pitch suspensions beyond the syllable boundaries. As a result, pitch change boundaries and phoneme change boundaries, which usually jointly indicate note boundaries, do not align (they can differ by more than 50ms). We hypothesise that the discrepancy between them are larger for some vocal styles than for others. For example, we investigated Children's songs included in Molina et al.'s dataset. To verify the hypothesis, we plan to implement more analysis of those changes. The other example is concerned with note transitions in Zhangqiu Bangzi. Bangzi is a Chinese folk music genre and Zhangqiu is a region where a style of Bangzi is formed. Zhangqiu Bangzi is characterized by large pitch jumps of more than an octave between pairs of notes. Two features were observed: loudness is increased almost synchronously with pitch; pitch overshoots the target pitch. The hypothesise is that the first feature is shared across styles of Bangzi, while the second feature is specified by the local accent. To demonstrate that this is the case, comparison across different Bangzi styles will be applied by automatically analysing the note transitions for these features.

I. REFERENCES

[1] M. Mauch, K. Frieler, and S. Dixon, "Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 401–411, 2014.

[2] Y. Ozaki, J. McBride, E. Benetos, P. Pfordresher, J. Six, A. Tierney, P. Proutskova, E. Sakai, H. Kondo, H. Fukatsu, *et al.*, "Agreement among human and automated transcriptions of global songs," 2021.

[3] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–9.

*YL is supported by a China Scholarship Council and Queen Mary University of London joint Ph.D. Scholarship. and the 2021 Chinese National Social Science Funding in Art Project No. 21BD061.

†ZY is supported by the 2021 Chinese National Social Science Funding in Art Project No. 21BD061.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Variational Auto Encoding and Cycle-Consistent Adversarial Networks for Timbre Transfer

Russell Sammut Bonnici*, Martin Benning, & Charalampos Saitis

Queen Mary University of London, United Kingdom, r.sammutbonnici@gmail.com

***Abstract*— The combination of Variational Autoencoders (VAE) with Generative Adversarial Networks (GAN) motivates meaningful representations of audio in the context of timbre transfer. This was applied to different datasets for transferring vocal timbre between speakers and musical timbre between instruments. Variations of the approach were trained and generalised performance was compared using the Structural Similarity Index and Frechét Audio Distance. Many-to-many style transfer was found to improve reconstructive performance over one-to-one style transfer.**

***Index Terms*—** Deep learning, Audio, Generative Adversarial Networks, Auto-encoders, Style Transfer, Timbre

I. INTRODUCTION

Timbre transfer is a task concerned with modifying audio signals such that their timbre is reformed while their semantic content is persisted. Through this, utterances of a speaker can be changed such that they sound like they were spoken by another speaker. Recordings of a source instrument can be manipulated in a similar way such that they sound like another target instrument played them. The challenge in making the modification take place first lies in how exactly timbral features can be captured.

II. METHOD

The approach adopted follows a UNIT inspired architecture that was initially proposed for voice conversion [1]. It uses a VAE for motivating content persistence that is embedded in a GAN for motivating timbre transfer. By applying this to the URMP dataset [2] for musical instruments, the generalisibility of the approach was challenged. An ablation study was also carried out on URMP and the Flickr 8k Audio dataset [3] for insight on what makes the architecture effective. Variations of the model included; a version with no Kullback–Leibler divergence (KLD) cyclic component for the VAE, a version where bottleneck residual blocks were used in place of basic residual blocks, and a version where the same model was trained for multiple style transfers at once (many-to-many) rather than one transfer (one-to-one).

III. RESULTS

Table 1: Structural Similarity Index of Cyclic Reconstructions

Target	Initial	No KLD Cyclic	Bottleneck Residual	Many to many
Female 1	0.73	0.74	0.73	**0.77**
Male 1	0.80	0.78	0.68	**0.82**
Trumpet	0.83	0.83	0.78	**0.89**
Violin	0.81	0.81	0.78	**0.88**

Table 2: Frechét Audio Distance (General Vocoding)

Target	Initial	No KLD Cyclic	Bottleneck Residual	Many to many
Female 1	2.96	**2.77**	9.10	4.31
Male 1	1.65	2.48	6.97	**1.40**
Trumpet	**5.26**	5.52	6.06	5.85
Violin	**4.50**	5.52	12.68	4.99

The VAE-GAN approach was found general enough for applicability to instrument timbre transfer [4]. Basic residual blocks superseded bottleneck residual blocks around the latent space of the VAE for enriching content information. The presence of KLD for the cyclic loss component did not significantly impact performance. The many-to-many extension outperformed the initial one-to-one version in terms of reconstructive capabilities due to the increased variation of data passed through the universal encoder, yet improvements on the adversarial translation aspect were inconclusive. More clarity may be produced by training the utilised vocoder further.

IV. REFERENCES

[1] E. A. AlBadawy and S. Lyu, "Voice Conversion Using Speech-to-Speech Neuro-Style Transfer," in *Proc. Interspeech 2020*, 2020, pp. 4726–4730.

[2] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2019.

[3] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 237–244.

[4] R. S. Bonnici, C. Saitis, and M. Benning, "Timbre transfer with variational auto encoding and cycle-consistent adversarial networks," 2021.

*Research supported by ENDEAVOUR Scholarships Scheme (Malta)

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Characterizing Texture for Symbolic Piano Music

Louis Couturier[1], Louis Bigo[2] and Florence Levé[1,2]

[1]MIS, Université de Picardie Jules Verne, France, louis.couturier@u-picardie.fr
[2]CRIStAL, UMR 9189 CNRS, Université de Lille, France

***Abstract*— Musical texture describes how sounding components are organized, individually or with each other. In this work, we propose a new syntax to characterize symbolic texture in classical piano music and thus allow its automated analysis. We annotated the score of 9 movements of W.A. Mozart's sonatas (totalling 1164 bars), on which tested binary classification models show promising results for automatic texture retrieval.**

***Index Terms*—** Texture, symbolic music, piano, modeling, machine learning

I. MUSICAL SYMBOLIC TEXTURE

Symbolic texture is a high-level feature of music which strongly relates to musical style and form. Huron [1] links the term *texture* with the density, the diversity or the overall quality of sound. In this work, we set aside the physical timbre to focus on what Hérold [2] calls the "textural factors of timbre" [1]. Ultimately, a textural configuration can be defined by grouping *threads* (voices or instrumental parts) into several *layers* (textural units) [3]. In 2014, Giraud et al. [4] proposed a first syntax of texture annotation for classical string quartets. We aim at broadening this definition and apply it to piano music. Improving the global understanding of musical texture can also open new perspectives in style analysis or texture control in music generation.

II. MODELING PIANISTIC TEXTURE

The number of threads and their mutual relationships can vary during a piece of music. Piano scores do not explicitly separate musical threads, in contrast to ensemble music. This challenge needs to be addressed by proposing a more general model of texture.

Hence, we model the separation of vertical layers, their thickness (number of threads in each) and functions (melodic, harmonic and/or static support). Other attributes were added to characterize the content of certain layers (sustained or repeated notes, arpeggios etc.) or link them together (like homorhythmy or parallel motions). This model was defined syntactically and semantically, and resulted in an object-oriented implementation in *Python*.

[1]In the original paper: "les facteurs texturaux du timbre".

III. ANNOTATED DATASET

In order to evaluate this syntax through supervised machine learning tasks, we built a minimal dataset from the 9 movements of Mozart's piano sonatas 1, 2 and 5 (K. 278/189^{d}, 279/189^{e}, 283/189^{h}), for a total of 1164 annotated measures. Selected movements all share sonata form but feature various time signatures and tonalities. Scores are taken from the *Mozart Annotated Sonatas* [5], based on the *Neue Mozart-Ausgabe*. Furthermore, 62 high-level descriptors of symbolic music were implemented and computed on each measure of the corpus, adding expert knowledge in the process. For each bar, we retrieve 15 *textural elements* from the annotation. They indicate the presence of certain functions or attributes in the textural layers (melodic layer, homorhythmy, arpeggios...).

IV. PREDICTING TEXTURAL ELEMENTS

A Logistic Regression model is trained to predict the presence of each of the 15 *textural elements* in a bar given its associated 62 computed high-level features. The model is cross-validated with a *leave-one-piece-out* strategy: the prediction on the bars in one movement is performed by a model trained on the other 8 movements in the dataset. Other models including Support-Vector Machine and Decision Trees showed no significant improvements. Although the identification of melodic layers or homorhythmy showed reasonable results – with F1-scores of respectively 96% and 85% –, scale motives or non-melodic functions remain challenging elements to predict.

V. REFERENCES

[1] D. Huron, "Characterizing musical textures," in *Int. Computer Music Conf. (ICMC)*, 1989, p. 131–134.

[2] N. Hérold, "Timbre et forme : La dimension timbrique de la forme dans la musique pour piano de la première moitié du dix-neuvième siècle," Ph.D. dissertation, Université de Strasbourg, 2011.

[3] D. Moreira de Sousa, "Textural Design: A Compositional Theory for the Organization of Musical Texture," Ph.D. dissertation, Universidade Federal Do Rio de Janeiro, 2019.

[4] M. Giraud, F. Levé, F. Mercier, M. Rigaudière, and D. Thorez, "Towards Modeling Texture in Symbolic Data," *International Society of Music Information Retrieval (ISMIR)*, p. 59–64, 2014.

[5] J. Hentschel, M. Neuwirth, and M. Rohrmeier, "The Annotated Mozart Sonatas: Score, Harmony, and Cadence," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, p. 67–80, 2021.

DMRN+16: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2021

QUEEN MARY UNIVERSITY OF LONDON
TUE 21 DECEMBER 2021

Beat-Based Audio-to-Score Transcription for Monophonic Instruments

Jingyan Xu

Music X Lab, NYU Shanghai, joy_xjy @ sjtu.edu.cn

***Abstract*— We propose a model to generate readable scores from audios for monophonic instruments in classical music. Firstly, we obtain the beats from the transcribed MIDI. Secondly, we analyze the most likely tone combinations and pitches according to the beats. Thirdly, we do the recreation to refine the potential mistakes in pitches and rhythms to make the musical semantics more reasonable. The generated scores are subjected to the performers' intentions and are meaningful in musicology.**

***Index Terms*—** audio-to-score, beat tracking, music generation

I. INTRODUCTION

Audio-to-score is to estimate the human-readable score from the input audio signal. For instruments with stable rhythms and fixed pitches like piano, there exist decent methods to obtain accurate scores [1]. However, for monophonic instruments, the unstable rhythms and the constantly changing pitches make the score difficult to obtain. As the performers add their own improvised recreations in real performances, the original scores and the recreated scores might be different. We want to stress this problem in our work.

We try to recover the performers' intentions into human-readable scores. To acquire the performance scores, people first estimate the tones, beats, rhythms, and pitches by repeatedly listening to the recording. The ambiguity makes it impossible to obtain the rhythms and pitches accurately. In face of this problem, people may add their own recreations to make the score be reasonable in musicology. Our model does a similar job by stretching MIDI notes based on the extracted beats. After that, the model performs recreation based on music semantics.

Our model can generate scores as long as the transcribed information meets the minimum requirement [2]. With the common information in the part scores, a further recreation could lead to a readable full score for a band or an orchestra in the future.

II. METHODOLOGY

We primarily consider 8-measure music segments in 4/4 time signature. The tones are transposed to C major or A minor , and there are no off-key notes.

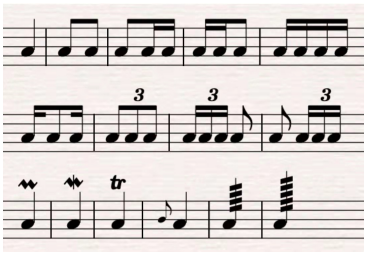

Figure 1: The possible combinations in one quarter note

Step 1: We extract beats from the MIDI, and MIDI are acquired from a transcription model.

Step 2: We jointly estimate the most possible tone value combinations, pitches, onsets and offsets.

We suppose that one beat is a quarter note. Therefore, the possible tone combinations for a quarter note are finite as Figure1 illustrates.

To distinguish between different note durations, we also need to label the onsets and offsets of the notes.

Step 3: We jointly refine the potential mistakes in pitches according to a music language model and fix the notes.

We use a public score editing software MuseScore 3 for score typesetting and generate the readable scores in the MusicXML format. For baselines, scores are generated by the MIDI data from Step 1 or Step 2, but not both. For evaluation, we use the mean of the 5 error rates in [3] to evaluate the quality of our generated scores. As there are some recreations in our results, the subjective evaluations are also indispensable.

III. REFERENCES

[1] Y. Hiramatsu, E. Nakamura, and K. Yoshii, "Joint estimation of note values and voices for audio-to-score piano transcription," in *Proceedings of the 22th International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[2] L. Lin, Q. Kong, J. Jiang, and G. Xia, "A unified model for zero-shot music source separation, transcription and synthesis," in *Proceedings of the 22th International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[3] E. Nakamura, E. Benetos, K. Yoshii, and S. Dixon, "Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 101–105.