DMRN+15: Digital Music Research Network

One-day Workshop 2020

Digital Music
Research Network

Queen Mary University of London

Tuesday 15th December 2020

Chair: Simon Dixon

Queen Mary
University of London

centre for digital music

Programme

10:00	**KEYNOTE** “Creativity at the Era of Artificial Intelligence”, **Prof. Philippe Esling (Institut de Recherche et Coordination Acoustique Musique)**
10:25	**Break**
10:30	“Joint Piano-roll and Score Transcription for Polyphonic Piano Music”, **Lele Liu, Veronica Morfi and Emmanouil Benetos (Queen Mary University of London)**
10:40	“A Modular System for Harmonic Structure Analysis of Music”, **Andrew McLeod and Martin Rohrmeier (École Polytechnique Fédérale de Lausanne)**
10:50	“Choral Music Separation using Time-domain Neural Networks”, **Saurjya Sarkar, Emmanouil Benetos, and Mark Sandler (Queen Mary University of London)**
11:00	“Generating Audio Mosaics with Particle Smoothing”, **Graham Coleman (Oldenburg, Germany)**
11:10	**Break**
11:20	“How to Automatically Calculate Tonal Tension Using AuToTen”, **Germán Ruiz-Marcos, Robin Laney and Alistair Willis (Open University)**
11:30	“Perceptual Similarities in Neural Timbre Embeddings”, **Ben Hayes, Luke Brosnahan, Charalampos Saitis, and George Fazekas (Queen Mary University of London)**
11:40	“Creating and Evaluating an Annotated Corpus Using the Library ms3”, **Johannes Hentschel and Martin Rohrmeier (École Polytechnique Fédérale de Lausanne)**
11:50	“Temporal Classes of User Behaviours on Music Streaming Platforms”, **Dougal Shakespeare and Camille Roth (Centre March Bloch)**
12:00	**Break**
12:05	**KEYNOTE** “Controllable Music Generation: from MorpheuS to Deep Networks”, **Prof. Dorien Herremans (Singapore University of Technology and Design)**
12:30	**Lunch break**
13:00	**Poster session** Posters will be displaced on breakout rooms. Participants will be allowed to freely join the rooms, view the posters, and talk to the authors.
14:30	**Break**
14:35	**KEYNOTE** “Accessibility through sound design and spatialisation: towards more creative and inclusive practices in film and television”, **Dr. Mariana Lopez (University of York)**
15:00	**Close**

Keynote talks

Keynote 1 - by Prof. Philippe Esling- Associate professor and head of the Artificial Creative Intelligence and Data Science (ACIDS) research group at IRCAM

Title: Creativity at the era of artificial intelligence.

Abstract : Creativity is a deeply debated topic, as this concept is arguably quintessential to our humanity. Across different epochs, it has been infused with an extensive variety of meanings relevant to that era. Along these, the evolution of technology have provided a plurality of novel tools for creative purposes. Recently, the advent of Artificial Intelligence (AI), through deep learning approaches, have seen proficient successes across various applications. The use of such technologies for creativity appear in a natural continuity to the artistic trend of this century. However, the aura of a technological artefact labeled as intelligent has unleashed passionate and somewhat unhinged debates on its implication for creative endeavors. In this talk, we aim to provide a new perspective on the question of creativity at the era of AI, by blurring the frontier between social and computational sciences. To do so, we rely on reflections from social science studies of creativity to view how current AI would be considered through this lens. As creativity is a highly context-prone concept, we underline the limits and deficiencies of current AI, requiring to move towards artificial creativity. We exemplify our argument with several very recent research works from our team at IRCAM, called Artificial Creative Intelligence and Data Science (ACIDS).

Keynote 2 - by Prof. Dorien Herremans - Assistant Professor at Singapore University of Technology and Design (SUTD) where she leads the AMAAI lab and is Director of SUTD Game Lab.

Title: Controllable music generation: from MorpheuS to deep networks.

Abstract: In its more than 60 year history, music generation systems have never been more popular than today. In this talk, I will discuss a number of co-creative music generation systems that have been developed over the last few years. These include MorpheuS, a tonal tension-steered music generation system guided by tonal tension and long-term structure. MusicFaderNets, a variational auto encoder model that allows for controllable arousal and rhythmic density of music. Finally, some more recent models by our AMAAI lab which include architectures such as controllable transformers and hierarchical RNN.

Keynote 3 - by Dr. Mariana Lopez - Senior Lecturer in Sound Production and Post Production at the Department of Theatre, Film, Television and Interactive Media at University of York

Title: Accessibility through sound design and spatialisation: towards more creative and inclusive practices in film and television

Abstract: Studies on sound design and spatialisation in the creative arts seldom engage with their potential to create accessible experiences. But these strategies can do much more than just entertain and immerse audiences, they could be put to the service of the creation of more accessible and inclusive experiences. This talk will explore research on the use of creative sound design for the development of accessible film and television experiences for visually impaired audiences. It will do so by exploring the Enhancing Audio Description Methods (EAD Methods), as an alternative to traditional Audio Description practices. The talk will explore the potential of the methods for accessibility practices as well as the creative advantages they hold for sound designers as well as film and television creators. Attendees will be introduced to notions of integrated access, accessible filmmaking and universal design, and how these are key for the creation of creative and accessible film and television productions, in which innovation on sound design and spatialisation is focused on their contribution towards social inclusion.

Posters

1	“Prosociality and Collaborative Playlisting: A Preliminary Study”, **Ilana Harris (Freie Universität Berlin) and Ian Cross (University of Cambridge)**
2	“auraloss: Audio-focused loss functions in PyTorch”, **Christian J. Steinmetz and Joshua D. Reiss (Queen Mary University of London)**
3	“Development of an Audio Quality Dataset Under Uncontrolled Conditions”, **Alessandro Ragano (University College Dublin), Emmanouil Benetos (Queen Mary University of London) and Andrew Hines (University College Dublin)**
4	“Analysis of Chord Progression Networks”, **Lidija Jovanovska and Bojan Evkoski (International Postgraduate School Jozef Stefan)**
5	“Fusion of Hilbert-Huang Transform and Deep Convolutional Neural Network for Predominant Musical Instrument Recognition”, **Xiaoquan Li and Jinchang Ren (University of Strathclyde)**

Organizing Committee

Adan Benito
Berker Banar
Alvaro Bort
Marco Communita
David Foster
Lele Liu
Ilaria Manco
Andrea Martelloni
Mary Pilataki
Saurjya Sarkar
Pedro Sarmento
Elona Shatri
Cyrus Vahidi

Supported by UKRI AIM CDT

Joint Piano-roll and Score Transcription for Polyphonic Piano Music

Lele Liu*,Veronica Morfi, and Emmanouil Benetos

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

Abstract— **We propose a method of joint multi-pitch detection and score transcription for polyphonic piano music. The outputs of our system include both a piano-roll representation (a descriptive transcription) and a symbolic musical notation (a prescriptive transcription). Instead of further converting MIDI transcriptions to scores, we use a multitask model combined with Convolutional Recurrent Neural Networks and Sequence-to-sequence models with attention mechanisms. We propose a reshaped score representation that outperforms a LilyPond representation both in prediction accuracy and time/memory resources, and compare different input audio spectrograms. The joint model outperforms a single task model in score transcription.**

I. INTRODUCTION

A large part of work in Automatic Music Transcription (AMT) falls under the tasks of multi-pitch detection and onset/offset detection. In this work, we discuss the problem of music audio-to-score transcription (A2S). Unlike in [1] which obtains a MIDI output in the beginning and transcribes music audio step by step, we use an end-to-end method that directly converts an audio input to a score format (see some early stage works in [2]).

In this work, we intend to extend the use of end-to-end A2S to a more general application scenario of polyphonic piano music with varying polyphony levels, as well as to support the estimation of music performance characteristics in a piano-roll format. We propose a multitask end-to-end model composed of convolutional layers, recurrent layers and sequence-to-sequence models with an attention mechanism for A2S, which is, to our knowledge, the first holistic model that transcribes polyphonic piano music into both a piano-roll format (corresponding to a descriptive notation of the music audio) and a score in Western staff notation (corresponding to a prescriptive notation of the musical audio). Additionally, we propose a new score representation for modelling polyphonic music that learns and predicts 7 times faster, uses less memory, and performs better than the LilyPond format score representation on this model. We also test the effect of using different input time-frequency representations, and the effect of combining multi-pitch detection and score transcription with a multitask model.

II. EXPERIMENTS

We carry out three experiments: 1) *comparison of time-frequency representations*, including Short-Time Fourier Transform (STFT), Mel Spectrogram, Constant-Q Transform (CQT), Harmonic Constant-Q Transform (HCQT), and Variable-Q Transform (VQT); 2) *comparison of score representations*, including a LilyPond format score representation and a Reshaped score representation (see in Figure 1); 3) *combination of piano-roll and symbolic score* in a multitask model. We use a joint model with shared convolutional layers, and separate recurrent layers/sequence-to-sequence networks for multi-pitch detection and score prediction.

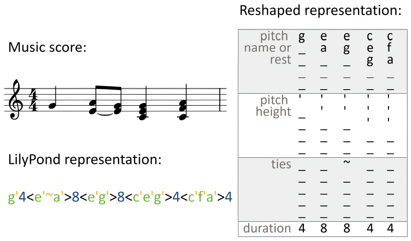

Figure 1. Example music score and corresponding LilyPond and Reshaped representation

We train and evaluate our system in a dataset with scores collected from the MusicScore website and audio recordings synthesized from the scores. Experimental results are shown in Tables 1 and 2. Among the five spectrogram types, VQT shows the best performance. The Reshaped representation runs around 7 times faster, uses around half the memory, and is slightly better than the LilyPond representation in terms of prediction accuracy. Overall, the joint model predicts better scores than a single task model.

Table 1. Benchmark F-measure of piano-roll prediction on different input representations and models.

Input representations/Models	F_f	F_{on}	F_{onoff}
STFT	89.5	81.0	61.7
Mel Spectrogram	89.0	82.1	63.0
CQT	**91.9**	85.4	67.4
HCQT	91.0	84.1	65.3
VQT	**91.9**	**85.7**	**68.5**
Piano-roll only	86.4	**67.6**	52.0
Joint	**88.0**	66.7	**53.6**

Table 2. Word error rates and MV2H [3] results in percentage for different models. LilyPond: Score-only model with LilyPond representation; Reshaped: Score-only model with Reshaped representation; Joint: Joint model with Reshaped representation.

WER	wer_{right}	wer_{left}	wer
LilyPond	38.0	39.0	38.5
Reshaped	37.8	**34.5**	**36.2**
Joint	**37.6**	35.3	36.5

MV2H	F_p	F_{voi}	F_{met}	F_{val}	F_{MV2H}
LilyPond	66.7	90.3	94.8	93.2	86.3
Reshaped	69.6	89.7	94.8	93.7	86.9
Joint	**71.1**	**90.8**	**94.9**	**94.4**	**87.8**

III. REFERENCES

[1] K. Shibata et al., “Non-local musical statistics as guides for audio-to-score piano transcription,” arXiv preprint arXiv:2008.12710, 2020.

[2] M. A. Román et al., “Data representations for autio-to-score monophonic music transcription,” Expert Systems with Applications, vol. 162, pp.113769, 2020.

[3] A. Mcleod and M. Steedman, “Evaluating automatic polyphonic music transcription,” in ISMIR, 2018, pp. 42-49.

*L. Liu 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.

DMRN+15: Digital Music Research Network
One-day Workshop 2020

Queen Mary University of London
Tue 15 December 2020

A Modular System for Harmonic Structure Analysis of Music

Andrew McLeod and Martin Rohrmeier*

Digital and Cognitive Musicology Lab, EPFL, Switzerland, andrew.mcleod@epfl.ch

***Abstract*— Harmonic structure analysis is the task of labeling an input musical piece (be it a score, MIDI, or audio) with chord and local key information. The task involves many interconnected dependencies at various levels of granularity (from low to high: frames, notes, chords, and keys). In this work, we propose a system with a modular design, allowing each component to regard the data at the appropriate level.**

***Index Terms*—** Chord, key, harmonic analysis

I. Task and System

Previous work on full harmonic analysis [1, 2] has treated the input as a sequence of input frames, assigning a label to each with various (sequential and non-sequential) neural network architectures. Our modular system, on the other hand, models each aspect of the analysis at its corresponding level of granularity. We hypothesize that such a design will allow our system to be more interpretable, as well as more adaptable to various use cases.

We use a very large vocabulary of chords and keys taking on a full characterization as used in music theory. Chord roots and key tonics may be any pitch A–G, double-flat to double-sharp (35 total). Chords may be major, minor, augmented (each with no, major, or minor 7th), or diminished (with no, minor, or diminished 7th) (12 total); in any inversion (3 for triads, 4 for 7th chords). This totals 1540 chords and 70 keys (major or minor for each tonic).

Our system is composed of 6 modules, each with a well-defined input and output (see Fig. 1). Its input can be a musical score (notes), a MIDI file (notes), or frames of an audio spectrogram, though we currently use only musical scores, and its output is a list of (absolute or relative) chord symbols and local keys, each corresponding to a range of inputs.

The *Chord Transition Model* (CTM) takes the system's input vectors and outputs the probability of each being the start of a new chord. The *Chord Classification Model* (CCM) takes as input a list of input vectors belonging to the same chord, and outputs a distribution over all chords. The *Chord Sequence Model* takes as input a sequence of chords (whose root pitch is relative to the current key's tonic), and outputs a distribution over the next chord at each step. The *Key Transition Model* takes as input a sequence of relative chords and outputs the probability of each being the start of a new local key. The *Key Sequence Model* takes as input the sequence relative chords from the previous local key section of a piece, plus the first chord symbol of a new key section (still relative to the previous tonic), and outputs a distribution over the next local key (as an interval from the previous tonic plus major or minor). The *Initial Chord Model* outputs a distribution over the first relative chord symbol of a piece given the key.

Table 1: Evaluation results.

Chord			Key		
Root+Triad	+7ths	**+Inv**	Tonic	**+Mode**	**Full**
0.47	0.32	0.26	0.45	0.34	0.15

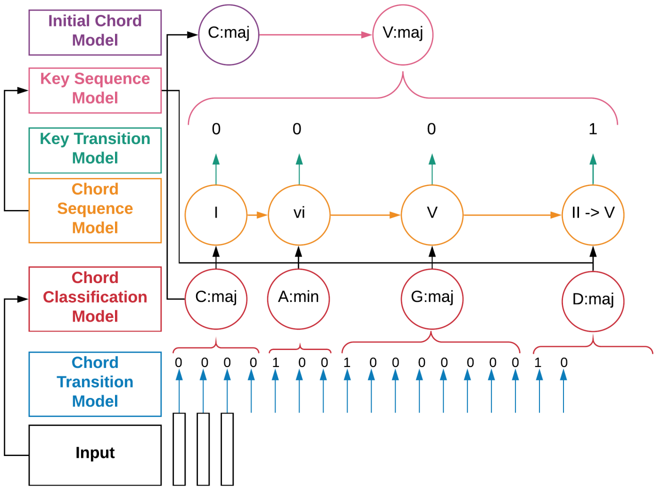

Figure 1: Overview of our fully integrated system. Each component depends on the component directly below it, in addition to the arrows.

II. Results and Discussion

We train and evaluate the system on a private set of annotated musical scores from a variety of composers (including the Annotated Beethoven Corpus [3]). Results are shown in Table 1. Each value is the average proportion of each piece with the correct label. Although each of our modules is currently very simple (most are a single-layer LSTM with a softmax), our results are promising.

Our system's modular allows us to train and improve each component independently, treating each as a black box. Noisy training methods such as scheduled sampling could also be used to make our model more robust to decoding errors. In future work, we plan to adapt the system to different input formats (MIDI and audio—only the CTM and CCM would need to be re-trained), and use the system in a human-in-the-loop way for annotation where a human can force the search process to go through particular manually-input labels.

III. References

[1] T.-P. Chen and L. Su, "Harmony transformer: Incorporating chord segmentation into harmony recognition," in *ISMIR*, nov 2019.

[2] G. Micchi, M. Gotham, and M. Giraud, "Not all roads lead to rome: Pitch representation and model architecture for automatic harmonic analysis," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 42–54, may 2020.

[3] M. Neuwirth, D. Harasim, F. C. Moss, and M. Rohrmeier, "The annotated beethoven corpus (ABC): A dataset of harmonic analyses of all beethoven string quartets," *Frontiers in Digital Humanities*, 2018.

*Research supported through the Swiss National Science Foundation within the project "Distant Listening – The Development of Harmony over Three Centuries (1700–2000)" (Grant no. 182811).

DMRN+15: Digital Music Research Network
One-day Workshop 2020

Queen Mary University of London
Tue 15 December 2020

Choral Music Separation using Time-domain Neural Networks

Saurjya Sarkar[*1], Emmanouil Benetos[1] and Mark Sandler[1]

[1]Centre for Digital Music, Queen Mary University of London, United Kingdom
saurjya.sarkar@qmul.ac.uk

Abstract— **Polyphonic vocal recordings are an inherently challenging task for source separation due to the melodic structure of the vocal parts and unique timbre of its constituent parts. In this work we utilize a time-domain neural network leveraged for speech separation and modify it to separate 4 acapella vocals (soprano, alto, tenor and bass) at a high sampling rate. To our knowledge this work is the first attempt to use permutation invariant training with time-domain neural networks for this task with audio data only. The results obtained are comparable to the state-of-the-art score-informed separation methods.**

Index Terms— Time Domain Source Separation, Choral Music

I. Architecture

We leverage the Conv-TasNet [1] and Dual-Path RNN [2] architecture with Permutation Invariant Training (PIT) to separate mixtures of 4 source choral polyphonic vocal mixtures. We modify the Conv-TasNet architecture to handle 22.05kHz sampling rate data by increasing the input window to 20 samples and adding 1 dilated convolutional layer to have a receptive field of 1.4 second. We use a permutation invariant loss function for the 4 source mixtures with scale-invariant signal to distortion ratio (SI-SDR) as the loss function. We utilize the Asteroid [3] framework for the experiments presented here.

II. Dataset

We use a combination of 26 Bach Chorales (BC) and 22 Barbershop Quartet (BQ) acapella multitracks from [4] for this experiment. Both the datasets had a combined duration of 104 minutes which was split 8:1:1 between training, test and validation sets. BC has 2 male (tenor and bass) and 2 female vocal sources (soprano and alto) while BQ has all 4 male vocalists. Each file is 10-second at with a sampling rate of 22.05kHz and 16 bits per sample.

III. Training

We trained the network 200 epochs on 10-second-long segments with early stopping (patience of 10 epochs). The initial learning rate is set to $5e^{-4}$ and is subsequently halved if the validation loss does not improve for 3 consecutive epochs. The remaining training parameters are the same as used in the original implementation of Conv-TasNet [1].

IV. Results

We evaluate the performance of our separation using the Asteroid implementation of signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) [5] and SI-SDR. We compare our results with reported non-PIT and score-informed separation results on choral mixtures presented in [6,7].

Table 1. Average SIR, SAR and SDR for choral music separation.

Model	***SIR***	***SAR***	***SDR***
ConvTasNet @ 22.05 kHz	**+12.45 dB**	+7.81 dB	+6.18dB
DPRNN @ 11.025 kHz	+11.80 dB	**+8.14 dB**	+6.24 dB
U-Net without score [6]	+9.30 dB	+5.69 dB	-
Wave-U-Net without score [6]	+7.07 dB	+5.54 dB	-
C-U-Net with score [6]	+12.08 dB	+7.21 dB	-
Wave-U-Net with score [7]	-	-	**+8.1 dB**[#]

Our preliminary results suggest that time-domain separation with permutation invariant training is indeed a suitable tool for this task. Audio examples[†] from our models and our code[^] based on Asteroid is available online.

V. References

[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation." *IEEE/ACM TASLP*, vol. 27, pp. 1256-1266, Aug 2019.

[2] Y. Luo and N. Mesgarani, "Dual-path RNN: Efficient long sequence modelling for time-domain single channel speech separation." *IEEE ICASSP*, pp. 46-50, May 2020.

[3] M. Pariente et al. "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech, ISCA,* 2020.

[4] R. Schramm and E. Benetos. "Automatic transcription of a cappella recordings from multiple singers", *AES International Conference on Semantic Audio*, June 2017.

[5] E. Vincent et al., "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, Jul. 2006.

[6] D. Petermann et al., "Deep Learning based Source Separation Applied to Choir Mixtures", *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montreal, Canada (Virtual), 2020.

[7] M. Gover et al., "Score-informed source separation of choral music," *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montreal, Canada (Virtual), 2020.

[*] S. Sarkar 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. This research utilized Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045 .

- Data not reported in [6, 7].

[#] Median SDR value reported in [7].

[†] http://c4dm.eecs.qmul.ac.uk/ChoralSep/

[^] https://github.com/saurjya/asteroid/tree/ChoralSep

Generating Audio Mosaics with Particle Smoothing

Graham Coleman

Oldenburg, Germany, ravelite@gmail.com

Abstract— **Bayesian sampling techniques, such as particle filters, offer a way to solve state estimation and optimization problems within audio synthesis and transformation. By defining a varispeed random tape that jumps between source audio segments, and a likelihood function expressing harmonic and timbral similarity, particle smoothing was used to generate tape control sequences that imitate a target music segment.**

Index Terms— Particle filters, smoothing, audio mosaic, concatenative synthesis, sampling synthesis.

I. Introduction

Sampling synthesis, for example, when imitating a target music signal, offers alternate views to an audio corpus. As these create complex multi-step decision problems, there is a rich mathematical space available for solving them. Previous systems in these space include [1, 2].

One broad approach, rather than trying to exactly solve an intractable optimization problem, *samples* in the probabilistic sense; that is, it runs a procedure with randomized solutions that tends to produce better solutions as output. [1] is one of the few concatenative synthesis systems from this family of approaches.

More specifically, sequential monte carlo (SMC), or particle filters, offer a principled way to serially decompose the problem of estimating time-varying state. These were previously used for musical tempo tracking tasks [3], as they are admissible even when using complex distributions and/or non-linear state evolutions.

Thus, SMC offers a simple, idiosyncratic, and perhaps overlooked inference method for audio synthesis and transformation.

II. Particle Smoothing

In order to specify the problem, one chooses the prior distributions, that is, the initial state distribution as well as how the state evolves through time. One also chooses a likelihood function, a conditional probability of different states given the observed data, giving a kind of similarity measure. Lastly, one chooses an importance function, which affects which states are favored in the sampling process.

Given a variety of choices for the distributions above, the main algorithm proceeds identically, using the sequential importance resampling (SIR) particle smoother [4]. This produces a distribution of N sampling paths (like the ones of Figure 1) that we can sonify.

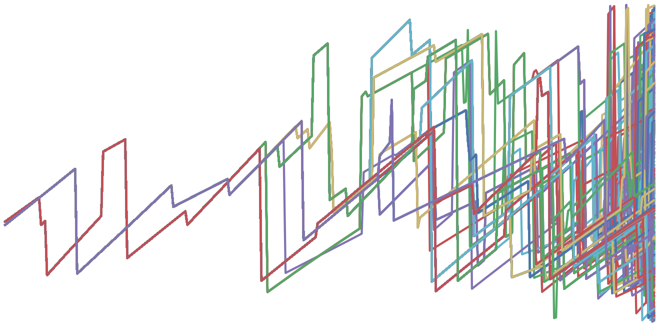

Figure 1: Typical distribution of smoothing paths (sampling position over time) produced by SIR smoothing algorithm. Most paths share common histories.

The author has implemented a prototype framework in python, allowing for configuration of the different distributions and SMC smoothing. Once computed, a representative sampling path is synthesized. Some sound examples will be presented.

III. References

[1] M. D. Hoffman, P. R. Cook, and D. M. Blei, “Bayesian Spectral Matching: Turning Young MC into MC Hammer via MCMC Sampling,” in *Proceedings of ICMC 2009*, 2009.

[2] G. Coleman, “Descriptor Control of Sound Transformations and Mosaicing Synthesis,” Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, 2016. [Online]. Available: http://mtg.upf.edu/node/3449

[3] S. W. Hainsworth and M. D. Macleod, “Particle Filtering Applied to Musical Tempo Tracking,” *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 15, pp. 1–11, Dec. 2004, number: 15 Publisher: SpringerOpen. [Online]. Available: https://asp-eurasipjournals.springeropen.com/articles/10.1155/S1110865704408099

[4] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge University Press, Sept. 2013.

How to automatically calculate tonal tension using AuToTen

Germán Ruiz-Marcos, Robin Laney and Alistair Willis[1]

[1]School of Computing and Communications, Open University, UK, german.ruiz-marcos@open.ac.uk

Abstract— **AuToTen, as in Automatic Tonal Tension, is a Python-based system which automatically calculates the contributions to tonal tension of a piece of music according to Lerdahl's model of tonal tension. We will present a demo to illustrate how to use AuToTen. We believe many research projects could benefit from AuToTen's capabilities.**

Index Terms— Tonal tension, Automation, GTTM, TPS

I. Introduction

In music, the sense of tension created by melodic and harmonic motion is often referred to as tonal tension [1]. From the existing models of tonal tension, Lerdahl's [1] has shown strong correlations against human judgements of tension. Lerdahl's Model of Tonal Tension (MTT) provides us with a method to estimate the degrees of tonal tension a Westerner may perceive when listening to a piece of tonal music. However, the application of MTT needs to be done manually. In order to automate its application, we have developed AuToTen.

II. Lerdahl's model of tonal tension

MTT relies on the Generative Theory of Tonal Music (GTTM) [2], which consists of a collection of rules to extract the metrical components of a piece of music, its inner groups and patterns, and its hierarchical relations. From GTTM's outputs, MTT calculates two components. First, a value of harmonic tension, which concerns the vertical arrangement of chords and the cognitive distances between them within the Tonal Pitch Space (TPS) [3]. Second, a value of attraction, which concerns the horizontal arrangement of chords and the voice-leading paths between chords.

III. AuToTen

What is AuToTen?

AuToTen [4], as in Automatic Tonal Tension, is a publicly available system[1] capable of automatically calculating the degrees of tonal tension, of a given piece of music, according to Lerdahl's MTT. As input, AuToTen needs to be fed with a piece of music and its GTTM representations, all in MusicXML format. The latter can be calculated using the Interactive GTTM Analyser [5] .

How has AuToTen been implemented?

AuToTen consists of five sub-systems: (1) the *metre analyser*, which produces a list of the input piece's offsets from the input GTTM representations; (2) the *matrix calculator*, which calculates a representation of the input piece's hierarchical relations according to the input GTTM representations; (3) the *harmonic analyser*, which calculates the most suitable key and chord labels of the input piece of music; (4) the *parameter calculator*, which calculates the parameters needed to apply the rules in Lerdahl's MTT; and (5) the *tension calculator*, which calculates the input piece's values of harmonic tension and attraction according to Lerdahl's MTT.

How to use AuToTen?

AuToTen includes the file `run.py`, which automatically calls all five AuToTen's sub-systems. When running this file, the user will be asked to select a piece of music and its GTTM representations, all in MusicXML format. The user will also be asked to select a location to save AuToTen's outputs. These will consist of two CSV files which include the quantitative values of the piece's harmonic tension and attraction according to Lerdahl's MTT.

What else can AuToTen be used for?

AuToTen's built-in functions can also be called independently and may be useful in other projects which do not concern musical tension. It is worth mentioning two of these functions. First, `distance()`, which can be used to calculate the distance between two chords within TPS. Second, `generator()`, which can be used to transform a piece's GTTM hierarchical representation into a more readable representation in the form of a matrix.

IV. References

[1] F. Lerdahl and C. L. Krumhansl, "Modeling tonal tension," *Music perception*, vol. 24, no. 4, pp. 329–366, 2007.

[2] F. Lerdahl, R. S. Jackendoff, and R. Jackendoff, *A generative theory of tonal music*. MIT press, 1983.

[3] F. Lerdahl, *Tonal pitch space*. Oxford University Press, 2004.

[4] G. Ruiz-Marcos, A. Willis, and R. Laney, "Automatically calculating tonal tension," in *The 2020 Joint Conference on AI Music Creativity*, B. Sturm and A. Elmsley, Eds., 2020. [Online]. Available: http://oro.open.ac.uk/72732/

[5] M. Hamanaka and S. Tojo, "Interactive gttm analyzer." in *ISMIR*, 2009, pp. 291–296.

[1] `https://doi.org/10.21954/ou.rd.13026578.v1`

DMRN+15: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2020

QUEEN MARY UNIVERSITY OF LONDON
TUE 15 DECEMBER 2020

Digital Music
Research Network

Perceptual Similarities in Neural Timbre Embeddings

Ben Hayes, Luke Brosnahan, Charalampos Saitis, and George Fazekas

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

Abstract— **Many neural audio synthesis models learn a representational space which can be used for control or exploration of the sounds generated. It is unclear what relationship exists between this space and human perception of these sounds. In this work, we compute configurational similarity metrics between an embedding space learned by a neural audio synthesis model and conventional perceptual and semantic timbre spaces. These spaces are computed using abstract synthesised sounds. We find significant similarities between these spaces, suggesting a shared organisational influence.**

Index Terms— Neural audio synthesis, psychoacoustics, timbre, representation learning

I. Introduction

Many neural audio synthesis models use representation learning techniques to enable interpretable control. For example, Kim *et al* learned an instrument embedding when training their *Mel2Mel* model, in a manner that required only reconstruction loss [1]. In this work, we compare the organisation of a *Mel2Mel* embedding space with perceptual and semantic timbre spaces computed from human ratings.

II. Method

We use a set of twelve sounds created with frequency modulation (FM) synthesis in a previous study [2]. Participants ($n = 30$) provided pairwise dissimilarity ratings on these stimuli, and an English speaking subset ($n = 24$) provided semantic ratings along 30 adjective scales. Adjectives were sourced by text-mining a corpus from a popular modular synthesis forum.* A 3D timbre space was constructed by performing multidimensional scaling (MDS) on the dissimilarity scores, and a 2D semantic space was computed with exploratory factor analysis (EFA) on the semantic ratings.

The organisation of *Mel2Mel*'s embedding space is guided by the network's overall reconstruction objective [1]. Two versions of the model were trained, with 2D and 3D embedding spaces.

III. Results

The two semantic factors showed strong loadings for terms associated with mass and texture, respectively. The mass factor correlated strongly with the first dimension of the 3D timbre space.

Table 1: Configurational Similarity Metrics

Space	Embed.	*T.C.C.*	m^2	*RVmod*
EFA	2D	0.884	0.439*	0.683
MDS	3D	0.923	0.721	0.325

*PROTEST significance $p < 0.001$

To compare the perceptual and neural spaces, three configurational similarity metrics were used. Tucker's congruence coefficient (TCC) is related to the cosine similarity between factors. A TCC of $0.83 - 0.95$ is considered significant, and > 0.95 nearly identical [3]. m^2 is the minimisation objective of Procrustes rotation. The modified RV coefficient is an extension of Pearson's r to matrices. Table 1 shows these metrics for each timbre space and the embedding space of corresponding dimensionality. We see strong similarity across all metrics in the semantic EFA space, and very strong similarity in only TCC in the MDS space.

IV. Conclusion

The similarities between the timbre spaces and the *Mel2Mel* embedding spaces suggest that both systems rely on similar attributes to discriminate timbres. Whilst not conclusive, our results warrant further investigation. This will include inquiry into whether these results generalise to other sonic domains and NAS architectures, including those with different representational spaces. The finer structure of these spaces can also be studied by observing the positioning of latent space interpolations in perceptual and semantic timbre spaces.

V. References

[1] J. W. Kim, R. Bittner, A. Kumar, and J. P. Bello, “Neural Music Synthesis for Flexible Timbre Control,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, United Kingdom, 2019, pp. 176–180.

[2] B. Hayes and C. Saitis, “There's more to timbre than musical instruments: Semantic dimensions of FM sounds,” in *Proceedings of the 2nd International Conference on Timbre*, Thessaloniki, Greece (Online), 2020.

[3] U. Lorenzo-Seva and J. M. F. ten Berge, “Tucker's congruence coefficient as a meaningful index of factor similarity,” *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, vol. 2, no. 2, pp. 57–64, 2006.

This work was supported by UKRI [grant number EP/S022694/1]

*MuffWiggler: https://www.muffwiggler.com/forum/

DMRN+15: Digital Music Research Network
One-day Workshop 2020

Queen Mary University of London
Tue 15 December 2020

Creating and Evaluating an Annotated Corpus Using the Library *ms3*

Johannes Hentschel and Martin Rohrmeier*

Digital and Cognitive Musicology Lab, École Polytechnique Fédérale de Lausanne, Switzerland, johannes.hentschel@epfl.ch

***Abstract*— This contribution focuses on the production of an annotated dataset, supported by the Python library *ms3*[1]. The process is centered around the open-source notation software MuseScore 3 and resulted in the first digital edition of W.A. Mozart's 18 piano sonatas according to the *Neue Mozart Ausgabe* [1]. The 54 MuseScore files are annotated with harmony, phrase, and cadence labels, and the first section focuses on how *ms3* was exploited to extract, manipulate, and add annotations. The second section presents an example of how the extracted data may be combined and evaluated in order to map out the harmonic make-up of the roughly 1,100 cadences contained in the dataset.**

***Index Terms*—** corpus research, corpus creation, dataset, Viennese Classic, solo sonatas, music theory, music annotation, expert analyses, harmony, cadence, data validation

I. Overview

Within Digital Musicology, the computer-aided analysis of large annotated corpora is one of the prevailing methods for gaining music theoretical insights into the musical language of a particular composer and/or of a particular style (for an example, see [2]). The structural aspects of a musical language are often considered as emerging from an interplay between harmony (vertical relationships) and voice-leading (horizontal relationships). Harmonic analyses encoded by human experts therefore make up the majority of annotated datasets in this domain (for an approach to annotating voice-leading, cf. [3]). There are, however, only few datasets where more high-level, formal analyses are encoded which account for phrase structure, modulation plans, or formal patterns (for a suggested form annotation standard, see [4]). The *Annotated Mozart Sonatas* [5] described in this contribution address this issue by including analytical labels for Roman numerals, phrase boundaries, *and* cadences. The annotated corpus currently represents one of the largest datasets allowing for the investigation of the Classical cadence [6] and other morphogenetic features such as phrases and their relation to a piece's tonal hierarchy.

II. Corpus Creation

The 54 MuseScore files were partly downloaded and converted from online sources and the missing movements were typeset in MuseScore. All scores have been checked for accordance with the *Neue Mozart Ausgabe*. The harmony and phrase annotations were added to and reviewed in the respective scores, and extracted as tabular files (TSV format) with *ms3*. The cadence labels were manually created in a tabular format and can be automatically added to the MuseScore files using the Python library's interface. Furthermore, the data has been verified using a novel data triangulation procedure based on expert consensus.

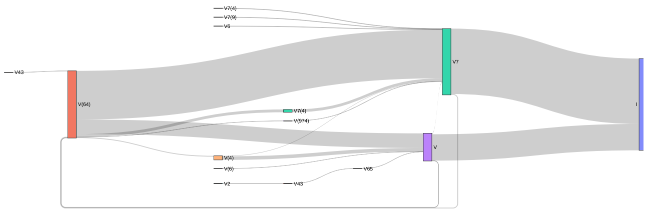

Figure 1: Flow chart aggregating the progressions of dominant chords in Perfect Authentic Cadences. The height of a coloured node expresses the chord's relative frequency. Any flow leaving on the right side shows the proportion of its progressions to other chords.

III. Evaluation

In order to investigate the harmonic make-up of the five annotated cadence types, *ms3* was used for merging the TSV files representing the annotation sets. The harmony labels preceding each cadence label were grouped by chordal roots and aggregated. For example, Figure 1 shows, for all 517 Perfect Authentic Cadences (PACs), the progressions between different dominant chords (chordal root `V`) that precede the tonic ultima. The plot reveals that in this repertoire the typical dominant progressions of a PAC starts on a cadential six-four chord `V(64)` and proceeds to only one other chord, namely `V7` or `V`. In contrast, dominants featuring a fourth suspension (e.g. `V(4)` or `V7(4)`) are rare events. Comparing equivalent plots for other cadence types and chordal roots may shed light on the question whether their harmonic make-up differs.

IV. References

1 Plath, W., & Rehm, W. (Eds.). (1986). *Klaviersonaten* (Neue Mozart-Ausgabe IX/25, Vol. 1-2). Bärenreiter.

2 Moss, F. C., Neuwirth, M., Harasim, D., & Rohrmeier, M. (2019). Statistical Characteristics of Tonal Harmony: A Corpus Study of Beethoven's String Quartets. *PLOS ONE*, *14*(6). https://doi.org/10.1371/journal.pone.0217242

3 Ericson, P., & Rohrmeier, M. (2020). Hierarchical Annotation of MEI-encoded Sheet Music. *Extended Abstracts for the Late-Breaking Demo Session of the 21st International Society for Music Information Retrieval Conference, Montréal*.

4 Gotham, M., & Ireland, M. T. (2019). Taking Form: A Representation Standard, Conversion Code, and Example Corpus for Recording, Visualizing and Studying Analyses of Musical Form. *20th International Society for Music Information Retrieval Conference, Delft*, 633–699.

5 Hentschel, J., Neuwirth, M., & Rohrmeier, M. (2020). *The Annotated Mozart Sonatas: Score, Harmony, and Cadence*. Manuscript submitted for publication.

6 Caplin, W. E. (2004). The Classical Cadence: Conceptions and Misconceptions. *Journal of the American Musicological Society*, *57*(1), 51–118. https://doi.org/10.1525/jams.2004.57.1.51

*Research generously funded by the Swiss National Science Foundation and Claude Latour.

[1] https://pypi.org/project/ms3/

DMRN+15: Digital Music Research Network
One-day Workshop 2020

Queen Mary University of London
Tue 15 December 2020

Temporal Classes of User Behaviours on Music Streaming Platforms

Dougal Shakespeare*[1] and Camille Roth[1]

[1]Computational Social Science Team, Centre March Bloch, Berlin, Germany, dougal.shakespeare@cmb.hu-berlin.de

***Abstract*— Music Recommender Systems and the algorithms they encapsulate have become central to elicit efficient navigation of the vast informational landscapes of popular music streaming platforms. Questions have been raised as to the extent to which recommender algorithms may function as tools or rather act to distort ones initial input preference and subsequent behaviour. Recent multi-disciplinary endeavours have appraised this issue by drawing comparison to an *organic* reference point i.e., absent of algorithmic influence. Whilst the focus of such debates often revolves around distributions of underlying features of recommended items –most often, some measure of their diversity– this work takes a novel approach of exploring temporal variations in organic, algorithmic and editorial content access through a user-side analysis. By classifying users based upon aggregated item access type histories, we characterise the inherent properties of users within each set to trace factors connected to temporal changes in access types.**

***Index Terms*—** Music Recommender System, human and algorithmic curation, user behaviour, ROM-COM

I. Context

Music Recommendation algorithms are designed to elicit personalisation thereby alleviating *choice overload* - a product of the vast collections of music now at the disposal of modern music streaming platforms. Whilst this definement implies recommendation algorithms to act as cognitive helpers, recent years have given rise to substantial literature critiquing recommendation algorithms for distorting a user's *organic* (absent of algorithmic influence) preference. Extending this line of work, we perform a user-side assessment of temporal changes in item access modes on the popular music streaming platform - Deezer. On most music streaming platforms, users are indeed able to access songs by three main modes: *organic* (e.g. manual search, plays from personal library), *editorial* (e.g. curated playlists) and *algorithmic* (e.g. Flow or Daily playlists). Utilising such content access histories, our work defines distinct user taxonomies which capture evolving behavioural dynamics – to our knowledge a novel approach in this domain and a key prior step to disentangle the joint influence of organic and recommendation-based usage on the formation of taste.

II. Relevant Work

Beuscart et al. [1] study the impact of algorithmic recommendation on user autonomy. Their findings show the influence of algorithmic recommendation to be minimal supporting the theory that algorithms act as tools to be utilised. Nonetheless, Anderson et al. [2] find users of Spotify become more diverse in their listening by shifting away from *algorithmic* and towards *organic* music consumption suggesting the impact of algorithmic influence is not be understated. Notwithstanding, Munson & Resnick [3] show users to variously seek diversity. In this sense, the "average user" does not exist and a more fine-grained approach must be deployed to capture a certain number of families of user behaviour, especially in terms of recommendation usage and its temporal evolution.

III. Proposed Methodology

We work with a snapshot of user activity on Deezer, focusing on users who registered in September 2017 and remained active over a two year observation period. This yields approximately $17K$ users, a substantial user base we deem meaningful to perform a temporal assessment of changes in user access types.

We perform a temporal clustering to define a small number of typical user types which exhibit markedly distinct dynamics in the composition of their access modes in terms of *organic*, *editorial* and *algorithmic* content. For each user type we compute gender, age (binned) and activity level to characterise defining attributes which may be implicit of evolving user behaviours.

IV. References

[1] J. S. Beuscart, S. Coavoux, and S. Maillard, *Les algorithmes de recommandation musicale et l'autonomie de l'Auditeur: Analyse des écoutes d'un panel d'utilisateurs de streaming*, 2019, vol. 213, no. 1.

[2] A. Anderson, L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas, "Algorithmic Effects on the Diversity of Consumption on Spotify," vol. 2, pp. 2155–2165, 2020.

[3] S. A. Munson and P. Resnick, "Presenting diverse political opinions: How and how much," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1457–1466. [Online]. Available: https://doi.org/10.1145/1753326.1753543

*Research supported by the ANR-funded RECORDS project, grant agreement number ANR-2019-CE38-0013

Prosociality and Collaborative Playlisting: A Preliminary Study

Ilana Harris[1†] and Ian Cross[2]

[1]Center for Cognitive Neuroscience Berlin, Freie Universität Berlin, DE, il.harris@fu-berlin.de
[2]Centre for Music and Science, Faculty of Music, University of Cambridge, Cambridge, UK

Abstract— **Joint music-making has been found to promote prosocial tendencies (i.e. empathy) across various populations. However, experimental study of prosociality resulting from everyday musical engagement is lacking. We conducted an online experiment to investigate whether mere perceived presence of a partner during playlist-making activated core social processes implicated in empathy. Preliminary results suggest that in younger individuals, some of the social processes involved in joint music-making and implicated in empathy are likely to be elicited.**

Index Terms— Prosociality, collaborative playlisting

I. BACKGROUND

Joint music-making has been empirically shown to activate core social processes and result in prosocial transfer effects [1]. Similar empirical study of everyday musical behaviors is lacking. Collaborative playlisting, a growing site of everyday musical engagement [2], likely elicits *some* of the social processes involved in joint music-making and may also shed light on *technologically mediated* musical interaction.

II. AIMS AND METHOD

We designed an online experiment using PsychoPy [3] to investigate whether perceived presence of a partner during playlist-making is sufficient in eliciting social processes and prosocial consequences known to occur with face-to-face joint music-making, and how these effects may differentially hinge on music and demographic background.

Participants were asked to answer questionnaire items assessing demographic and musical backgrounds and assigned to either an algorithm (ALG) or a fake partner (FP) condition. Participants were then told to create 3 fixed-length playlists with song clips provided by the experimenter, and that either another participant (FP) or a song recommendation algorithm (ALG) would add additional clips to each playlist; in reality, clip additions were random. Participants were played back each resultant playlist (shuffled). A recognition task subsequently assessed participants' memory of who selected each clip provided in the previous sessions ('Q1' items = added by participant, 'Q2' items = added by FP/ALG, 'Q3'items = added by neither). Finally, participants answered self-report items assessing inclusion of other in self (IOS) and trait empathy (interpersonal reactivity index; IRI) [4, 5].

III. RESULTS AND CONCLUSIONS

Participants in the FP condition showed decreased memory sensitivity for recognition of their own clip selections in comparison with those in the ALG condition. Further, a significant main effect of age (>=25 or <25) on IRI and IOS scores was found. We conclude that for younger individuals perceived presence of a partner may increase activation of social processes during, and promote prosocial tendencies resulting from, online everyday musical engagement.

Table 1. Memory sensitivity scores for Recognition Task (25 trials total)

Recognition Task: Sensitivity Scores			
	ALL (n=90)	***ALG (n=44)***	***FP (n=46)***
Q1	1.31	1.42	1.21
Q2	.55	.54	.56
Q3	.98	.92	1.03

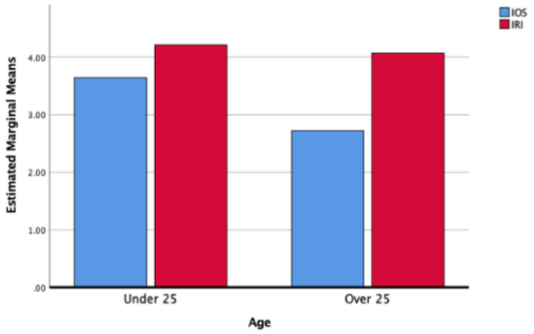

Figure 1. Mixed Repeated Measures ANOVA (n=90) showed main between-subjects effect of age on IOS and IRI scores (α= .02).

IV. REFERENCES

[1] T.-C. Rabinowitch, I. Cross, and P. Burnard, "Long-term musical group interaction has a positive influence on empathy in children," *Psychology of Music*, vol. 41, no. 4, pp. 484–498, Apr. 2012, doi: 10.1177/0305735612440609.

[2] S. Y. Park and B. Kaneshiro, "An Analysis of User Behavior in Co-curation of Music Through Collaborative Playlists," National University of Singapore Research Institute, 2017.

[3] J. W. Peirce, "PsychoPy—Psychophysics software in Python," *J Neurosci Methods*, vol. 162, no. 1–2, pp. 8–13, May 2007, doi: 10.1016/j.jneumeth.2006.11.017.

[4] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of Other in the Self Scale and the structure of interpersonal closeness," *Journal of Personality and Social Psychology*, vol. 63, no. 4, pp. 596–612, 1992, doi: 10.1037/0022-3514.63.4.596.

[5] M. Davis, "A Multidimensional Approach to Individual Differences in Empathy," *JSAS Catalog Sel. Doc. Psychol.*, vol. 10, Jan. 1980.

[†]Thanks to Herchel Smith Scholarship, Emmanuel College Cambridge, and University of Cambridge Faculty of Music for funding.

DMRN+15: Digital Music Research Network
One-day Workshop 2020

Queen Mary University of London
Tue 15 December 2020

auraloss: Audio-focused loss functions in PyTorch

Christian J. Steinmetz and Joshua D. Reiss

Centre for Digital Music, Queen Mary University of London, c.j.steinmetz@qmul.ac.uk

***Abstract*— We present `auraloss`[1], a PyTorch package that implements time and frequency domain loss functions designed for audio generation tasks. The package provides a straightforward interface, as well as multichannel support. We demonstrate its application by using each loss function to train a model on the task of emulating an analog dynamic range compressor.**

I. Loss functions

Error-to-signal ratio — The error-to-signal ratio (ESR) [1] is equivalent to the squared error between the input $\hat{y}$ and target y, both N samples in length, normalized by the energy of the target.

$$\ell_{\text{ESR}}(\hat{y}, y) = \frac{\sum_{i=0}^{N-1} |\hat{y}_i - y_i|^2}{\sum_{i=0}^{N-1} |y_i|^2} \tag{1}$$

Following [2], we also provide perceptually motivated pre-emphasis filters. These include an FIR first-order highpass filter, folded differentiator, as well as an approximation of the A-weighting filter.

Log hyperbolic cosine — The log hyperbolic cosine (log-cosh) [3] aims to strike a balance between the L_1 and L_2. It is similar to the L_2 for small values, providing a level of smoothness, and similar to the L_1 for large values, providing robustness. It is defined in Eq. 2, where a is a hyperparameter that controls the overall smoothness.

$$\ell_{\text{log-cosh}}(\hat{y}, y) = \frac{1}{a} \sum_{i=0}^{N-1} \log(\cosh(a(\hat{y}_i - y_i))) \tag{2}$$

Short-time Fourier transform — The Short-time Fourier transform (STFT) loss is composed of the spectral convergence (Eq. 3), and spectral log-magnitude (Eq. 4), where $|| \cdot ||_{\text{F}}$ is the Frobenius norm, $|| \cdot ||_1$ is the L_1 norm, and N is the number of STFT frames. The overall STFT loss is defined as the sum of these two terms [4].

$$\ell_{\text{SC}}(\hat{y}, y) = \frac{|| \, |\text{STFT}(y)| - |\text{STFT}(\hat{y})| \, ||_{\text{F}}}{|| \, |\text{STFT}(y)| \, ||_{\text{F}}} \tag{3}$$

$$\ell_{\text{SM}}(\hat{y}, y) = \frac{1}{N} \left\| \log\left(|\text{STFT}(y)|\right) - \log\left(|\text{STFT}(\hat{y})|\right) \right\|_1 \tag{4}$$

Multi-resolution STFT — The STFT loss can be extended by computing the loss at multiple different resolutions [5]. This improves robustness and avoids potential bias arising from the STFT parameters. The multi-resolution STFT (MR) loss is defined in Eq. 5 as the average of the error at each of the M resolutions.

$$\ell_{\text{MR}}(\hat{y}, y) = \frac{1}{M} \sum_{m=1}^{M} \left(\ell_{\text{SC}}(\hat{y}, y) + \ell_{\text{SM}}(\hat{y}, y) \right). \tag{5}$$

For optimal performance, the appropriate frame size, window type, and hop size must be selected. Often there is no clear choice. To address this we introduce the random-resolution STFT (RR), which randomly selects these parameters each time the loss is computed, ensuring the model is not biased by a fixed set of parameters.

Sum and difference loss — A loss function for stereo music was proposed in [6], which achieves left-right invariance by computing the sum and difference signals (Eq. 6) before applying the MR loss (Eq. 7), instead of directly operating on the left and right channels.

$$y_{\text{sum}} = y_{\text{left}} + y_{\text{right}} \qquad y_{\text{diff}} = y_{\text{left}} - y_{\text{right}} \tag{6}$$

$$\ell_{S/D}(\hat{y}, y) = \ell_{\text{MR}}(\hat{y}_{\text{sum}}, y_{\text{sum}}) + \ell_{\text{MR}}(\hat{y}_{\text{diff}}, y_{\text{diff}}) \tag{7}$$

II. Evaluation

To demonstrate the package, we train the same model each time using a different loss function. We employ a conditional temporal convolutional network (TCN) based on [7] for the task of modeling an analog dynamic range compressor [8]. The model is composed of 10 layers, each with kernel size 15, 32 channels, and exponentially increasing dilation factors for a receptive field of 324 ms at 44.1 kHz. We use Adam with a learning rate of $1 \cdot 10^{-3}$ and a batch size of 128, training each model for 20 epochs. We evaluate on the test set using all of the losses as error metrics as shown in Table 1.

Interestingly, we find that the lowest error for a given metric is not always achieved by optimizing that metric. It appears that training with a time domain loss leads to better performance on time domain metrics, with comparatively worse performance on frequency domain metrics, and vice versa. No formal conclusions can be made from this experiment, as differences in scaling of the losses during training may make comparisons challenging. We present this only as a demonstration of the package. Further work will examine these losses, and others, across more diverse audio generation tasks.

Model	**Test error**					
	L_1	ESR	Logcosh	STFT	MR	RR
L_1	**4.87e-3**	**0.0085**	**2.78e-5**	0.824	0.797	0.558
ESR	5.56e-3	0.0099	3.23e-5	0.806	0.779	0.549
Logcosh	5.30e-3	0.0093	3.03e-5	0.831	0.805	0.566
STFT	9.00e-3	0.0542	1.76e-4	0.451	0.432	0.339
MR	8.98e-3	0.0553	1.80e-4	**0.440**	**0.420**	**0.331**
RR	1.55e-2	0.2187	7.05e-3	0.525	0.504	0.392

Table 1: Test error across a model trained with different loss functions.

III. References

[1] A. Wright, E.-P. Damskägg, V. Välimäki *et al.*, "Real-time black-box modelling with recurrent neural networks," in *DAFx*, 2019.

[2] A. Wright and V. Välimäki, "Perceptual loss function for neural modeling of audio systems," in *IEEE ICASSP*, 2020, pp. 251–255.

[3] P. Chen, G. Chen, and S. Zhang, "Log hyperbolic cosine loss improves variational auto-encoder," 2018.

[4] S. Ö. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 94–98, 2018.

[5] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE ICASSP*, 2020, pp. 6199–6203.

[6] C. J. Steinmetz *et al.*, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," *arXiv:2010.10291*, 2020.

[7] C. J. Steinmetz, "Learning to mix with neural audio effects in the waveform domain," Master's thesis, Universitat Pompeu Fabra, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4091203

[8] S. Hawley, B. Colburn, and S. I. Mimilakis, "Profiling audio compressors with deep neural networks," in *AES*, 2019.

[1] `https://github.com/csteinmetz1/auraloss`

Development of an Audio Quality Dataset Under Uncontrolled Conditions

Alessandro Ragano[1], Emmanouil Benetos[2] and Andrew Hines[1]

[1]School of Computer Science, University College Dublin, Ireland, alessandro.ragano@ucdconnect.ie
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

***Abstract*— A curated dataset for assessing the perceived audio quality of sound archives has not yet been compiled or reported in the literature. In this study, we present the ongoing development of a perceived audio quality dataset using real-world recordings from the NASA Apollo mission audio.**

***Index Terms*—** Audio quality, corpus, Apollo missions.

I. Introduction

Computer technologies provide a more interactive access to historical audio archives and improve the exploration of mankind's historical moments. One of the factors that has been poorly investigated is the sound quality of digitised and restored sound archives. These operations are generally conducted by expert staff as the existing computer-based solutions are few and inefficient. This approach shows two main problems: 1) quality assessment is biased by subjective judgments of staff member experts; 2) given that sound archives are vast, a careful sound quality assessment cannot be conducted on every recording [1]. As a first step, we identify suitable real-world recordings that can be curated for building a dataset. In this work, we show how to build a dataset using data from the audio archive of the Apollo missions. This corpus documents one of mankind's greatest achievements and shows unique signal characteristics, constituted by field recordings.

II. Dataset Development

The Apollo audio archive has been curated for several deep learning-based speech processing tasks [2] but not for speech and audio quality assessment. To create a quality dataset, some issues were identified before annotating the data with quality scores. Extracting random clips from Apollo recordings might cause the presence of repetitive data i.e., data that will cause unbalanced regression. Therefore, a mechanism to control the distribution of the final dataset has been proposed. First, it has been studied whether existing non-intrusive metrics can predict quality in the Apollo recordings. A pilot study with 32 participants and using speech intelligibility as a proxy for quality has been conducted [1]. Results found no correlation between objective and subjective intelligibility [3] except for the word-error-rate (WER) computed on transcriptions of the Google speech-to-text API, which has led to 0.630 and 0.679 for the Pearson and Spearman respectively. Results are shown in Table 1. Therefore, repetitive data can be avoided by controlling the distribution through the Google STT WER. An example of repetitive data is shown in Figure 1, where more than 100 clips show similar objective quality.

Table 1: Inferential statistical tests for assessing the correlation between subjective WER and objective metrics [3].

Metric	Pearson coeff.	Pearson P-value	Spearman coeff.	Spearman P-value
Google STT WER	0.630	0.0001	0.679	1.93e-5
SRMR	0.081	0.658	0.112	0.538
ITU-T P563	-0.246	0.173	-0.295	0.100
MOSNet	-0.073	0.691	-0.163	0.371

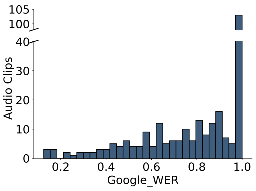

Figure 1: Google WER distribution on Apollo recordings [3].

An unsupervised cluster exploration has been also conducted to control audio stimuli during the preparation of the listening test used for labelling the final dataset and to split the dataset in a stratified fashion to avoid bias towards particular features when training the model. 253 audio features clustered with HDBSCAN are shown in Figure 2.

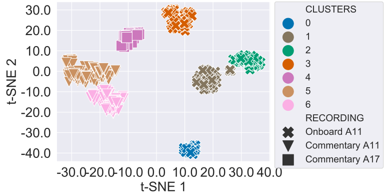

Figure 2: Google WER distribution on Apollo recordings [3].

III. Conclusions and Future Work

In this study, we have shown that simple techniques can be used for preventing the development of an unbalanced dataset. In the future, we will study the relationship between overall quality and intelligibility and we will use the tools shown in this study to curate the extended dataset. The curated dataset will be annotated with subjective quality ratings.

IV. References

[1] A. Ragano, E. Benetos, and A. Hines, "Adapting the quality of experience framework for audio archive evaluation," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019.

[2] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio." in *INTERSPEECH*, 2019, pp. 1851–1855.

[3] A. Ragano, E. Benetos, and A. Hines, "Development of a speech quality database under uncontrolled conditions." in *INTERSPEECH*, 2020, pp. 4616–4620.

[1] Dataset available at `10.5281/zenodo.3969507`

Analysis of Chord Progression Networks

Lidija Jovanovska, Bojan Evkoski

International Postgraduate School Jozef Stefan, Slovenia
lidija.jovanovska@ijs.si, bojan.evkoski@ijs.si

***Abstract*— We create a chord progression network using guitar chord tabs. By representing chords as nodes and transitions between chords as edges, we obtain a network suitable for the plethora of analysis methods that have been developed in the field of network science. We use the network to analyze communities, identify influential chords and compare differences between chord progressions across genre and decade. Finally, we apply stochastic walks to generate new chord progressions.**

***Index Terms*—** chord progression, network science, symbolic music generation

I. Chord Progression Networks

A large number of chord progression annotation datasets are available online. However, the alignment of data from multiple sources is not trivial. The work of Bien et al. is one of the few efforts in using network science methodology to analyze musical data [1]. However, the dataset used in the study contains only 360 songs and only two genres. To address this limitation, we collect data from the largest guitarist community website - Ultimate Guitar.[1] To create a representative and balanced dataset, we scrape the most popular songs by decade (from the 1960s to 2010s) and by genre. After filtering out noisy data, approximately 12,000 songs remain.

We define a chord progression network as a directed, weighted graph $G = (V, E)$. Each node $v \in V$ represents a chord, while an edge $e \in E$ from node u to node v indicates a transition between the chords. Additionally, edge weights correspond to the number of times the transition occurred in the dataset. This means that a chord progression is simply a path in the network.

Once the chord progression network was generated, we applied computational methods unique to graphs: community detection [2] and node influence metrics [3]. The result is presented in Figure 1. By analyzing multiple networks created from specific genres or decades, we can detect the differences between styles of music.

II. Generating Chord Progressions

To address the challenge of symbolic music generation, we experiment with graph traversal methods. We either specify the starting chord of the progression or choose randomly from a probability distribution derived from centrality measures, e.g. PageRank [3]. Next, we traverse the network by choosing the subsequent chord randomly, while taking into account the weights of the transitions and the community assigned to the previous chord. Thereby, we guarantee that the output progression is not always the most probable one in the network, while the weights ensure it sounds reasonable. The network and the implementation of the approach can be accessed online.[2]

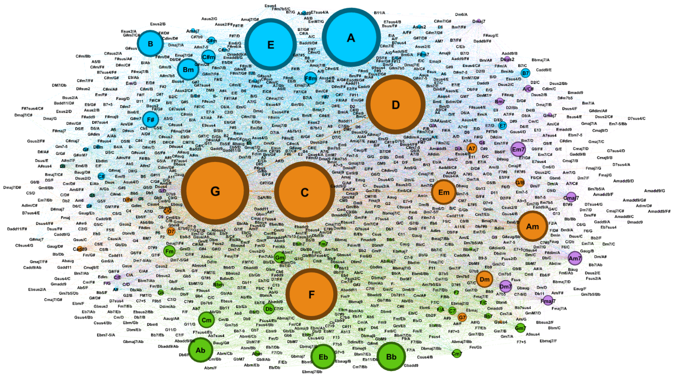

Figure 1: Chord progression network where nodes are chords and edges are transitions between chords. Node size corresponds to the PageRank. Node color corresponds to the communities.

To build upon this, we intend to utilize several neural network architectures, such as LSTMs, attention models, and graph neural networks. To harvest the power of these models, we plan on including all available data from Ultimate Guitar and building a bigger feature set based on knowledge hosted in open-source music databases.

III. References

[1] C. H. . R. N. Bien, N., "Network analysis of chord progressions in rock and jazz music," http://snap.stanford.edu/class/cs224w-2018/reports/CS224W-2018-94.pdf, 2018.

[2] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[3] . G. A. Xing, W., "Weighted pagerank algorithm," in *In Proceedings. Second Annual Conference on Communication Networks and Services Research*, 2004, pp. 305–314.

[1] Ultimate Guitar: https://www.ultimate-guitar.com

[2] Github: https://github.com/lidija-jovanovska/HAMR-2020

DMRN+15: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2020

QUEEN MARY UNIVERSITY OF LONDON
TUE 15 DECEMBER 2020

Fusion of Hilbert-Huang Transform and Deep Convolutional Neural Network for Predominant Musical Instruments Recognition

Xiaoquan Li[1] and Jinchang Ren[1*]

[1] Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK, xiaoquan.li@strath.ac.uk

***Abstract*— As a subset of music information retrieval (MIR), predominant musical instruments recognition (PMIR) has attracted substantial interest in recent years due to its uniqueness and high commercial value in key areas of music analysis research such as music retrieval and automatic music transcription, etc. With the attention paid to deep learning technology and artificial intelligence technology, they have been more and more widely applied in the field of MIR, thus making breakthroughs in some sub-fields that have been stuck in the bottleneck. In this paper, the Hilbert-Huang Transform (HHT) is employed to map one-dimensional audio data into two-dimensional matrix format and then a deep convolutional neural network is developed to learn affluent and effective features for PMIR. In the experiment, 6705 audio pieces including 11 musical instruments are used to validate the efficacy of our proposed approach. The results are compared to four benchmarking methods and show significant improvements in terms of precision, recall and F measures.**

***Index Terms*— Predominant musical instrument recognition, Convolutional neural network, Hilbert-Huang Transform.**

I. METHODOLOGY

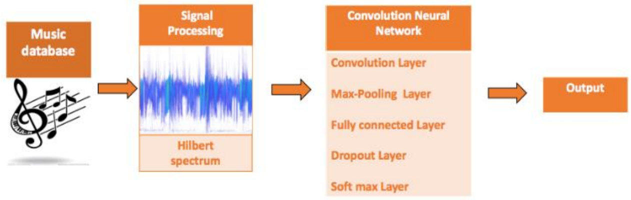

Figure 1. The flowchart of the proposed PMIR system.

In the proposed framework (**Error! Reference source not found.**), we use the HHT [1] to generate the Hilbert spectrum for each instrument in the polyphonic music pieces. Then we build a deep convolutional neural network (DCNN) to take the Hilbert spectrum as input and produce the classification label as the output.

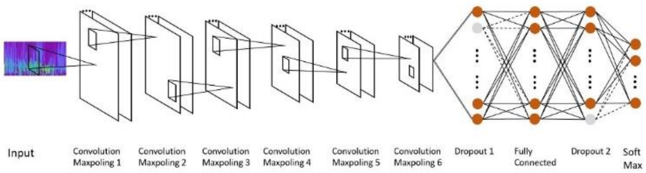

Figure 2. Flowchart of the proposed DCNN.

Our DCNN was inspired by the VGG-16 model[2], which contains 16 hidden layers (13 convolutional and 3 fully connected). The polling size is always set as 2 × 2 and the filter size is set as 3 × 3, and the VGG-16 shows that when deepening the network layers can improve performance.

II. EXPERIMENTS

To further evaluate the effectiveness of the proposed PMIR framework, three conventional approaches are used to benchmark in terms of precision, recall and F1-measurement. Three conventional frameworks are based on Audio Content Analysis (ACA) system [3, 4] and three machine learning model (i.e. random forest (RF)[5], SVM[6] and shallow neural network (SNN)[7]).

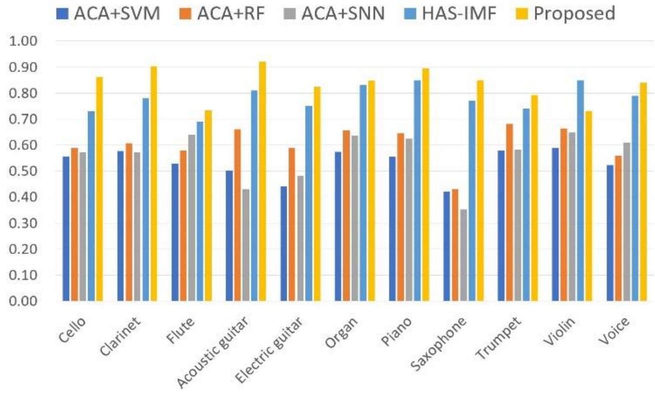

Figure 3. F1-measurement of each instruments of five methods.

III. REFERENCES

[1] S. Sandoval, et al., "Hilbert spectral analysis of vowels using intrinsic mode functions," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 569-575, 2015.

[2] K. Simonyan, et al., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, 2014.

[3] A. Lerch, *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press, 2012.

[4] G. Peeters, "A large set of audio features for sound description (similarity and classification)," *CUIDADO project Ircam technical report*, 2004.

[5] A. Liaw et al., "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18-22, 2002.

[6] C.-C. Chang, et al., "LIBSVM: A library for support vector machines," ACM trans. on int. sys. and tech. (TIST), vol. 2, no. 3, p. 27, 2011.

[7] R. Battiti, "First-and second-order methods for learning: between steepest descent and Newton's method," *Neural computation*, vol. 4, no. 2, pp. 141-166, 1992.