Queen Mary
University of London

DMRN+12: Digital Music Research Network Workshop Proceedings 2017

Kudumakis, P; Sandler, M; DMRN+12: Digital Music Research Network Workshop 2017

© 2017 Centre for Digital Music, Queen Mary University of London

For additional information about this publication click this link.
http://qmro.qmul.ac.uk/xmlui/handle/123456789/30583

Information about this research object was correct at the time of download; we occasionally make corrections to records, please therefore check the published record when citing. For more information contact scholarlycommunications@qmul.ac.uk

DMRN+12: Digital Music Research Network

One-day Workshop 2017

Arts One Lecture Theatre

Queen Mary University of London

Tuesday 19th December 2017

Chairs

Panos Kudumakis and Mark Sandler

Programme

10:30	**Registration** opens **Tea/Coffee**
11:00	**Welcome** and opening remarks **Prof. Mark Sandler** (Director, Media and Arts Technology, Queen Mary University of London)
11:10	**KEYNOTE** "Capturing and rendering spatial audio", **Prof. Augusto Sarti (Politecnico di Milano)**
11:50	"FAST forward to the semantics of design for musical performance", **Alan Chamberlain (University of Nottingham), David De Roure (Oxford University), Steve Benford, Chris Greenhalgh and Adrian Hazzard (University of Nottingham), Maria Kallionpää (Aalborg University), David Weigl, Kevin Page and Pip Willcox (Oxford University)**
12:10	"Artist similarity modelling for music discovery", **Alo Allik and Mark Sandler (Queen Mary University of London)**
12:30	"Are you experienced? Dynamic music listening", **Adrian Hazzard and Chris Greenhalgh (University of Nottingham), Florian Thalmann and Gary Bromham (Queen Mary University of London)**
12:50	**Buffet Lunch**, Networking **Posters** will be on display
14:00	"Deep adaptation: How generative music affects engagement and immersion in interactive experiences", **Andrew Elmsley, Ryan Groves and Valerio Velardo (Melodrive, Germany)**
14:20	"Evaluating machine learning for music generation", **Bob L. Sturm (Queen Mary University of London) and Oded Ben-Tal (Kingston University)**
14:40	"Exploration of emotion-based cross-modal mappings for generating music for videos", **Shahar Elisha and Tillman Weyde (City University of London)**
15:00	"An Internet of Musical Things architecture for performers-audience tactile interactions", **Luca Turchet and Mathieu Barthet (Queen Mary University of London)**
15:20	**Tea/Coffee** **Posters** will be on display
15:40	"Assessing the use of metrical information in a LSTM-based polyphonic music sequence transduction", **Adrien Ycart and Emmanouil Benetos (Queen Mary University of London)**
16:00	"Musicians' binaural headphone monitoring for studio recording", **Valentin Bauer (Paris Conservatoire), Hervé Déjardin (Radio France) and Amandine Pras (University of Lethbridge)**
16:20	"Towards bio-responsive control for music", **Duncan Williams and Damian T. Murphy (University of York) and Bruno M. Fazenda (University of Salford)**
16:40	"A statistical-learning model of harmony perception", **Peter M. C. Harrison and Marcus T. Pearce (Queen Mary University of London)**
17:00	**Panel Discussion**
17:30	**Close***

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

Posters

1	"An agent on my shoulder: AI, privacy and the application of human-like computing technologies to music creation", **Alan Chamberlain (University of Nottingham), Alessio Malizia (University of Hertfordshire) and David De Roure (Oxford University)**
2	"Inverting feature representations of machine listening systems", **Saumitra Mishra, Bob L. Sturm and Simon Dixon (Queen Mary University of London)**
3	"Social music machine: Crowdsourcing for composition & creativity", **Alan Chamberlain (University of Nottingham), David De Roure and Pip Willcox (Oxford University)**
4	"Feature design for intelligent control of the dynamic range compressor using audio decomposition", **Di Sheng and György Fazekas (Queen Mary University of London)**
5	"The art and 'science' of opera: Composing, staging & designing new forms of interactive theatrical performance", **Alan Chamberlain (University of Nottingham), Maria Kallionpää (Aalborg University) and Steve Benford (University of Nottingham)**
6	"Linear and logistic models for music classification experiments", **Francisco Rodríguez-Algarra and Bob L. Sturm (Queen Mary University of London)**
7	"Let's jam! An ethnographic study of collaborative music composing", **Juan Pablo Martinez Avila (University of Nottingham)**
8	"Discovering feature relevance in pedalling analyses of piano music", **Beici Liang, György Fazekas and Mark Sandler (Queen Mary University of London)**
9	"The social character of metadata in 'In the Box' music production", **Glenn McGarry (University of Nottingham)**
10	"Hearing the humanities: Sonifying Steele's Shakespeare", **Iain Emsley (Oxford University), Alan Chamberlain (University of Nottingham) and David De Roure (Oxford University)**
11	"A deeper look at the 2017 ASV spoof challenge", **Bhusan Chettri and Bob L. Sturm (Queen Mary University of London)**
12	"Towards performing a personal interactive musical soundtrack", **Laurence Cliffe (University of Nottingham)**

FAST Forward to the Semantics of Design for Musical Performance

Alan Chamberlain[1*] David De Roure[2] Steve Benford[1*] Chris Greenhalgh[1*] Adrian Hazzard[1*] Maria Kallionpää[3] David Weigl[2] Kevin Page[2] Pip Willcox[2]

[1*] MRL, Computer Science, University of Nottingham, UK, Alan.Chamberlain@Nottingham.ac.uk
[2] Oxford e-Research Centre, Oxford University, UK
[3]Department of Communication and Psychology, Aalborg University, Denmark

***Abstract*— This paper presents research relating to the Performance strand of the *FAST project. An emerging focus of the research on this project has related to the Performance (and Creation) of music. In this paper we briefly discuss two of the demonstrator projects that have been developed as part of the research of the project. We discuss a game-like musical piece called 'Climb!', which allowed the composer to integrate musical codes into their composition that could be triggered by the performers, and 'Numbers into Notes' an experimental digital humanities project that developed music creation software into the future based on Ada Lovelace's writings.**

I. SOUNDING SEMANTICS OUT

Contemporary musical performance has taken advantage or a range of new technological innovations that have enabled composers, performances and audiences to engage with music in new and innovative ways. By experimenting with such technologies we are able to unpack and research the use and development of such technologies in the context of musical performance and creation, and by doing understand the nature and use of semantics in this area. This may be meaning as interpreted by the audience or performer, meta-data as used at a system level or the use of a codified system for compositional purposes.

II. "CLIMB!"

"Climb!" is a composition, performed on a Yamaha Disklavier. It uses a software-based system called Muzicodes as a mechanism to trigger different parts of the composition. At intersection points in the composition the performer can play a different code that in turn take the performer and the audience on a musical journey. The musical structure is best imagined as a set of musical branches, at each point where the branch splits the performer can choose to go in variety of different musical journeys, and in order to choose one of though routes they must first play the code correctly. In our earlier work we have examined the use of the Muzicodes system and interacting with systems that need tight or loose adherence to codes in order to work with performance systems of this type. More recently we have been discussing issues pertaining to the archiving of the performances of "Climb!". "Climb!" raises questions about the nature of 'recording' and curating performances of this type, where each performance could be different, made up of a variety of constituent parts that together make up the piece. This obviously has implications for the meta-data that relates to the recording and also to publishing the score of a given performance. A system called MELD - Music Encoding and Linked Data (MELD) was also used in the performance. This framework retrieves, distributes, and processes information addressing semantically distinguishable music elements. The MELD framework and implementation architecture augments and extends MEI structures with semantic Web Annotations capable of addressing musically meaningful score sections.

III. TURNING NUMBERS INTO NOTES

Numbers into Notes describes a series of tools, prototypes, compositions and performances which have arisen out of a historical thought experiment: Ada Lovelace wrote in 1843 that Charles Babbage's proposed Analytical Engine "might compose elaborate and scientific pieces of music of any degree of complexity or extent", and we have used various kinds of digital prototyping to explore what might have happened, informed by the mathematical and musicological context of the time. The original work, presented at the Ada Lovelace Symposium in 2015, used a simulator for the analytical engine. Subsequently we developed a website to "crowdsource" the generation of musical fragments, and we also translated the algorithms onto arduino-based devices. In August 2017 a live composition at the Audio Mostly conference was based on fragments contributed by musicians. Like Climb!, Numbers into Notes raises questions around the role of the machine and the human, in this case through its combination of algorithmic generation and human selection and assembly. The original idea of "numbers into notes" is due to composer Emily Howard, with whom we continue to work to explore the relationship between mathematics and music.

IV. CONCLUSION

The two technologies that we have briefly discussed are different in their conception and development, but share features in respect to their use as compositional that can support performance, which use patterns and codes to provide a structure for the composition of music, at a systemic levels, as a composer and a performer.

* This research was supported through the following EPSRC project: Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (EP/L019981/1).

Artist Similarity Modelling for Music Discovery

Alo Allik and Mark B. Sandler

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

***Abstract*— This presentation explores different ways of representing music artist similarity, comparing concepts and methods based on collaborative filtering, crowd-sourced information, linked data and content-based feature extraction. Practical implementations of the theoretical methods are demonstrated on a web-based artist discovery platform that makes use of community-run public data sources and open source web technologies, which enables users to discover links between artists in novel ways.**

I. OVERVIEW

Artist similarity is one of the most frequently employed measures for music discovery and recommendation in large music libraries. Similarity models typically rely on collaborative filtering, content-based feature extraction or a combination of both. Known limitations of these methods - including limited exposure of a collection or lack of high-level descriptions - can be alleviated by adopting linked data practices and semantic representations of musical [1]. Publicly accessible open linked data stores combined with music related social media platforms offer an increasing number of opportunities for music researchers. This presentation proposes methodologies for enhancing artist similarity modelling taking advantage of linked data best practices that enable unique identification of musical entities and the discovery of valuable but obscure connections between them. The core of this methodology relies on structured representation of music related knowledge bases and datasets using Semantic Web technologies. This facilitates connecting artists by various commonalities such as style, geographical location, instrumentation, record label as well as less intuitive categories, for instance, artists who have received the same award, have shared the same fate, or belonged to the same organization or religion. Another strategy supplementary to semantic linking employs automated feature extraction data and strives to describe artists according to musical concepts derived from their published recordings such as tonality (most frequent musical keys and chords), rhythm (typical tempo and temporal density), and timbre (spectral information indirectly representing instrumentation). While automatic feature extraction has provided an alternative method of similarity modelling, it is still rather challenging to derive high level musical concepts such as mood or genre from such low level representations. One way of addressing these problems has been attempted using crowd-sourced tagging statistics in order to take advantage of user consensus. This enables mood similarity comparisons between songs and, by aggregating song data, can be extrapolated to modelling similarities between artists.

II. MUSICLYNX PLATFORM

The different linking models have been implemented in the MusicLynx (available at https://musiclynx.github.io) web platform that has been publicly deployed using open source and free front-end and server hosting platforms. Any artist recognised by the platform is linked to others by various categories derived from the models described above. While browsing the interface, users can link to several content providers - for example, BBC Music, Deezer, or Youtube - and collate favourite artists or playlists of their songs. Linked data based artist connections are queried from the Dbpedia knowledge graph (http://wiki.dbpedia.org), while content-based models are derived from the AcousticBrainz project (http://acousticbrainz.org). The experimental mood-based linking is provided directly from the MusicLynx API service (https://musiclynx-api.herokuapp.com).

ACKNOWLEDGMENT

This work was supported by EPSRC Grant EP/ L019981/1, "Fusing Audio and Semantic Technologies for Intelligent Music Production and Consumption".

REFERENCES

[1] M. Mora-Mcginity, A. Allik, G. Fazekas, and M. Sandler. *Musicweb: Music discovery with open linked semantic metadata*. In Proceedings of MTSR 2016, Göttingen, Germany, 2016.

Are You Experienced? Dynamic Music Listening

Adrian Hazzard[1], Chris Greenhalgh[1], Florian Thalmann[2] and Gary Bromham[2]

[1]School of Computer Science, University of Nottingham, United Kingdom, adrian.hazzard@nottingham.ac.uk
[2]Centre for Digital Music, Queen Mary University of London, United Kingdom

***Abstract*— This document acknowledges how music listening formats have historically evolved inline with technological developments, and then considers the current state of the art, specifically in relation to dynamic renderings of music as driven by user interactions or contextual adaptions. Given the availability of powerful technologies to create such dynamic experiences, we question why their mainstream adoption appears hesitant.**

I. INTRODUCTION

Historically, the development of new technologies has shaped how we listen to recorded music and that relationship appears just as prominent today: from the introduction of the vinyl record in 1948, to the transistor radio in 1954, to the compact disc in 1982 to online streaming in 2008 and projecting forward to further adoption of virtual reality experiences and (possibly) beyond to biometric media formats. Until recently, technological shifts in music formats have not ushered in any significant alternatives to the product consumed, i.e., recorded music rendered as a fixed linear form. Traditionally, listeners have had negligible control over playback, other than volume control, equalization shaping or the 'shuffle' function for playlist re-organisation.

II. DYNAMIC MUSIC

The opportunities presented by mobile, sensing, networked and web technologies over last 10-15 years has seen a flurry of explorations in the dynamic presentation of music which can permit for fluid, non-linear renderings driven by a range of input mechanisms. Some of these endeavors necessitate music specifically produced for proprietary systems, while others enable adaptations to existing stereo mixes. Such approaches can be by categorized by their control mechanism. There are those which leverage user interaction, where a listener also 'performs' the system, directly controlling tailorable features within a musical arrangement, for instance *NI Stems* and *Ninja Jam.* Moving beyond track level interventions, we note examples of dynamic playlists responsive to a listener's mood [1], or constructed upon recommendations based on historical listening habits. Other interactive systems may draw on user input indirectly, such as tempo sensitive playlists that synchronise with a users jogging pace [2]. Alternative dynamic approaches move beyond direct user interaction and look to a listener's context, such as their surroundings, to drive musical adaptations. Location-based soundtracks that unfold in relation to GPS sensing [4], which require bespoke tools to support user-configurable, semi-automated or fully automated control of playback renderings [5], for example.

III. CHALLENGES OF ADOPTION

Musical arrangements that adapt or facilitate user interaction are commonplace within computer game soundtracks. However, we have yet to witness wholesale adoption of alternative dynamic listening formats within mainstream music, where explorations to date have been peripheral; even though some influential artists such as Bjork and Sigur Rós have pushed at these boundaries. Against this backdrop we consider what the challenge points might be that account for this. For instance, the production of commercial music is a multi-faceted and complex process that engages with many stakeholders and relies on many complementary production tools. The production of recorded music has remained a static process, arguably as a result of embedded traditions, the continuing conceptualization of production tools based on traditional studio paradigms and commercial pressures. We speculate that a re-envisaging of the end-to-end production to consumption chain is required to enable artists and composers to recognize the opportunities and explore the new processes required to compose for dynamic music, and a commitment from all stakeholders to deliver and promote such formats at the point of consumption.

REFERENCES

[1] Barthet, Mathieu, György Fazekas, Alo Allik, and Mark Sandler. 2015. Moodplay: an interactive mood-based musical experience. In *Proceedings of the Audio Mostly 2015 on Interaction With Sound*, p. 3. ACM, 2015.

[2] Spotify, 'Spotify Running', 2015. [Online]. Available: https://www.spotify.com/uk/running/. [Accessed: 07-Sep-2015].

[3] Hazzard, A., Benford, S. and Burnett, G., 2015. Sculpting a mobile musical soundtrack. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems* (pp. 387-396). ACM.

[4] Thalmann, F. and Perez Carillo, A., Fazekas, G., Wiggins, G., and Sandler, M., 2016. The semantic music player: A smart mobile player based on ontological structures and analytical feature metadata. In *Proceedings of the 2nd Annual Web Audio Conference*.

* This work is supported by Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (Grant No. EP/L019981/1).

Deep Adaptation: How Generative Music Affects Engagement and Immersion in Interactive Experiences

Andrew Elmsley, Ryan Groves and Valerio Velardo

Melodrive, Germany, andy@melodrive.com

***Abstract*— This paper presents the results of a psychological experiment in which we measured the perceived level of immersion and time spent in the experience in a virtual reality experience under three music conditions: no music, linear music and machine-generated deep adaptive music. We found that deep adaptive music increased both the time spent in the interactive experience and significantly amplified the immersion of participants.**

I. INTRODUCTION

In interactive media, such as virtual reality (VR) experiences and games, the player has the ability to affect the world around them, triggering events and different emotional states at any time. It is often important to the game designer that the player feels immersed and engaged in these experiences. Previous research [1] indicates that music plays an important role in creating immersive experiences.

It is very common for interactive music composers to use adaptive music – music that changes through vertical layering and/or horizontal re-sequencing – to support different player behaviours. However, it's impossible for a human composer to conceive and produce every musical outcome. This means that even an adaptive soundtrack can quickly feel repetitive and break the player's engagement due to listener fatigue.

This is why we built Melodrive: an AI music generation engine that responds to very granular emotional cues in the experience and dynamically composes and produces music in realtime. We call this *deep adaptive music.*

In this paper we give details of a psychological experiment designed to understand whether deep adaptive music increases the level of immersion and engagement of a person exploring an interactive experience.

II. METHOD

We created a simple VR space station scene in which there were two rooms connected by a corridor. Each room had its own emotional feel ('tender' and 'angry'), and there were no interactive objects in the scene.

Each of the 46 participants were randomly assigned to one of 3 music conditions for the experience: no music, linear music and Melodrive-generated music. The linear music condition had a fixed, looping soundtrack. The music by Melodrive was generated in realtime and adapted to the emotional feel in each room while the participants explored the space station. Both the linear music and the deep adaptive music had the same sound design (instrumentation, effects etc.).

Participants had no prior knowledge of what the experiment was designed to test, and were instructed to explore the scene for as long as they liked. They were timed during the experience as an overall measure of engagement, and afterwards were asked to fill out a questionnaire about immersion and music, based off de Oliveira et al.'s metrics [2].

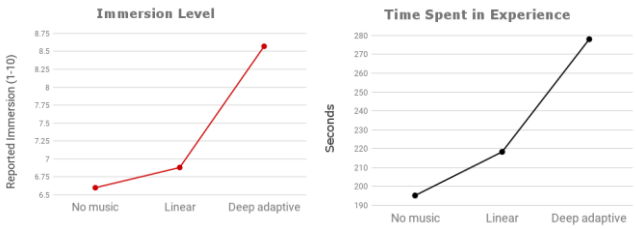

Figure 1. The overall immersion level and time spent in the experience.

III. RESULTS

The time spent in the VR scene with the Melodrive-generated music was 42% more than that for the no-music condition and 27% more than that for linear music. Immersion levels with music generated by Melodrive were 30% higher than those perceived with no music, and 25% higher than linear music. There was no significant difference in time spent and level of immersion between no music and linear music (Figure 1).

90% of participants thought that that the music generated by Melodrive was a very important component that helped them to feel immersed. It was also found that the adaptive music generated by Melodrive fitted the VR scene 49% better than the linear soundtrack.

All the findings were statistically significant ($p < 0.05$).

IV. CONCLUSION

Melodrive and deep adaptive music has the ability to increase both the time spent in an interactive experience and to significantly amplify the immersion of participants.

REFERENCES

[1] Zhang, J., & Fu, X. (2015). The Influence of Background Music of Video Games on Immersion. Journal of Psychology & Psychotherapy, 5(4).

[2] de Oliveira, R. P., de Oliveira, D. C. P., & Tavares, T. F. (2016). Measurement Methods for Phenomena Associated with Immersion, Engagement, Flow, and Presence in Digital Games. Proceedings of SBGames 2016.

Evaluating Machine Learning for Music Generation

Bob L. Sturm[1*] and Oded Ben-Tal[2]

[1*]Centre for Digital Music, Queen Mary University of London, UK, b.sturm@qmul.ac.uk
[2]Department of Performing Arts, Kingston University, UK

***Abstract* — In research applying machine learning to music modelling and generation, it is critical for researchers to engage with user communities to evaluate the usefulness of the resulting applications, and to consider their overall benefits and detriments to the practice domain. This will not only help make the application of machine learning to music more beneficial, but also help focus and improve research. We illustrate several approaches to evaluation in our own research modelling transcriptions of folk music, and applying it to music practice.**

Evaluation is frequently mentioned as a necessary but difficult task in the application of machine learning, or other computational methods, to music modelling and generation. Since aspects of musicality, aesthetic qualities, and creative potential are essential to this kind of research, evaluation that focuses primarily on quantitative results in the service of null hypothesis statistical testing is insufficient, and may even be counter-productive. For this research to be meaningful and relevant, one has to evaluate it in relation to musical practice.

Guidance for making applied machine learning "matter" can come from what we term "The Wagstaff Principles", inspired by [1]:

1. measure the concrete impact of an application of machine learning with practitioners in the originating problem domain;
2. with the results from the first principle, improve the particular application of machine learning, the definition of the problem, and the domain of machine learning in general.

Our research with our music modelling system, *folk-rnn* [2], attempts to follow these two principles by interacting with music practitioners in and out of the folk music tradition on which our models are trained (traditional Irish and English music). We have so far organized three concerts and one workshop featuring material generated by and with models created by *folk-rnn* [3]. This provides a way to evaluate the models from the perspective of the audience and the composer. We are considering several evaluation approaches:

1. **First-order sanity check**: comparing the statistics of generated transcriptions with those of the training transcriptions. How has a model succeeded or failed in capturing the patterns of the training data?
2. **Nefarious testing**: seeking the limits of the musical knowledge encoded by a model by pushing it little by little outside its "comfort zone". How fragile and general is its "music knowledge"?
3. **Music analysis**: examining the ways in which generated transcriptions are and are not successful as compositions. What should be changed to make the piece better?
4. **Performance analysis**: examining the plausibility of generated material through performance. What should be changed to make it perform better?
5. **Assisted composition**: using the models to create music in and out of the conventions of the training data. How well does a model contribute to the music composition "pipeline"? In what ways does the model hinder useful aspects of the composition process?
6. **Cherry picking**: finding useful material in the model output. How hard is it to find something of interest, something good, something really good, in a bunch of material generated by a model?
7. **Ethics analysis:** considering the positive and negative impacts of the application of machine learning on the practice domain. What are the ethical aspects to building and deploying such models?

We are exploring ways to improve *folk-rnn* models and their application via these evaluation approaches. One serious challenge is the lack of a one-to-one correspondence between the outcomes and the quantitative nature of model training. For instance, a composer might want to encourage certain kinds of unpredictable behaviours, but discourage others. Having a human "in-the-loop" could provide a solution. We are currently designing a web-based interface to *folk-rnn* models to facilitate such approaches.

* This research is supported by AHRC Grant No. AH/N504531/1.
BLS is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK (corresponding author).
OBT is with the Department of Performing Arts, Kingston University, UK (e-mail: O.Ben-Tal@kingston.ac.uk).

REFERENCES

[1] Wagstaff, "Machine learning that matters," in Proc. Int. Conf. Machine Learning, pp. 529–536, 2012.
[2] Sturm and Ben-Tal, "Taking the models back to music practice: Evaluating generative transcription models built using deep learning," *J. Creative Music Systems*, vol. 2, Sep. 2017.
[3] Sturm, Ben-Tal, Monaghan, Collins, Herremans, Chew, Hadjeres, Deruty, and Pachet, "Machine Learning in Music Practice: A Case Study", submitted to *J. New Music Research*, 2017.

Exploration of Emotion-based Cross-modal Mappings for Generating Music for Videos

Shahar Elisha and Tillman Weyde

Music Informatics Research Group, Dept. of Computer Science, City, University of London, UK,
t.e.weyde@city.ac.uk

***Abstract*— This work addresses the control of music generation for videos. We control the music generation process with arousal-valence values extract from videos via colour analysis and facial expression recognition. We use two deterministic and two machine-learning models for music generation. Although the generated music is not aesthetically satisfactory, this allows a first exploration of emotion-based film music generation.**

I. INTRODUCTION

Automatic generation of adapted music for videos requires sequential video analysis and a generation process based on that analysis. We describe here an approach to learning cross-modal relationships to generate music from video content based on emotions. Given a video, we extract an emotion vector as an intermediate representation to control the music composition. The music-generating model can then be trained on existing videos with music. We explored the potential of this approach by combining different video analysis and music generation models.

II. METHOD

Using colour analysis and Microsoft's Emotion API (azure.microsoft.com/en-us/services/cognitive-services/emotion), we computed a sequence of arousal and valence (AV) values for a video [1][2]. Two deterministic models that create a variety of modes and registers in different tempi in response to a change in emotion demonstrate the relationship between the video input and the music output [3].

Furthermore, we created two stochastic models using a Restricted Boltzmann Machine (RBM) where 4-bar segments are generated: the feedback model feeds the generated segment as input into the next cycle, whereas the sampling model generates unbiased segments every time.

We trained our models on a dataset extracted from 50 music videos from YouTube, and the equivalent MIDI files from freemidi.org. We chose music videos because they normally have good synchronisation between music and video content.

III. RESULTS

We found that the RBM is capable of learning simplistic artificial relationships between pitches and emotions. Using real music data, we found that the RBM was capable of replicating some statistically significant relations between the MIDI data and the AV values, some of which were counter-intuitive [3]:

TABLE I. EMOTIONS VS MIDI RELATIONS

Relationship	***Regression slope***
Arousal vs. note density	Negative
Valence vs. note density	Positive
Valence vs. average pitch	Negative
Arousal vs. pitch range	Negative
Valence vs. pitch range	Positive

Some relations between emotion values and notes were not transferred from the training to generated data: specifically pitch histograms, interval histograms, and autocorrelation (which was partly transferred).

However, the analysis of note and interval distributions confirms the subjective impression that real music data is too complex for our simple RBM models. Still, the stochastic models, although often random-sounding, produced some music that was subjectively more interesting and well adapted to the video. We found that the feedback model generated subjectively more musical structures than the sampling model by producing variations over 4-bar periods, which was confirmed by autocorrelation analysis.

IV. CONCLUSION

Overall, our results demonstrate that emotion-based music generation for videos is a promising approach to generating music adapted to videos. However, more complex generation models are needed to produce music with an aesthetically satisfying structure.

REFERENCES

[1] Geslin, E., Jegou, L. & Beaudoin, D. (2016). *How color properties can be used to elicit emotions in video games*. International Journal of Computer Games Technology, vol. 2016, pp.1-9.
[2] Russel, J.A. (1980). *A circumplex model of affect*. Journal of Personality and Social Psychology. 1980; 39: 1161-1178
[3] Morreale, F., Masu, R. and Angeli, A. (2013). *Robin: an algorithmic composer for interactive scenarios.* In: Proceedings of the Sound and Music Computing Conference 2013, SMC 2013. Logos Verlag Berlin, pp.207-212.

An Internet of Musical Things architecture for performers-audience tactile interactions

Luca Turchet[1*] and Mathieu Barthet[1*]

[1]Center for Digital Music Queen Mary University of London, UK, luca.turchet@qmul.ac.uk

***Abstract*— This paper presents an architecture supporting novel forms of tactile interactions between live music performers and audience members. Such interactions are enabled by the multidirectional communication between Smart Musical Instruments and Smart Musical Haptic Wearables.**

I. INTRODUCTION

Smart Musical Instruments (SMI) are a novel family of musical instruments characterized by embedded computational intelligence, wireless connectivity, an embedded sound delivery system, and an onboard system for feedback to the player [1]. They offer direct point-to-point communication between each other and other portable sensor-enabled devices connected to local networks and to the Internet. An example of devices that can be connected to SMI are the Smart Musical Haptic Wearables (SMHWs) [2]. This is a novel class of wearable devices for audience members, which encompass haptic stimulation, gesture tracking, and wireless connectivity features.

SMI and SMHWs are components of an ecosystem of interoperable musical devices that has been recently termed as "Internet of Musical Things" (IoMusT) [3]. Such an ecosystem can support novel forms of interactions between live music performers and audience members. This study presents an architecture enabling the multidirectional creative communication between performers playing SMI and audience members using SMHWs.

II. IOMUST ARCHITECTURE AND SUPPORTED INTERACTIONS

We designed a Smart Mandolin and a Smart Cajòn (Fig. 1), which enhance the acoustic instruments with contact microphones, sensors, actuators, the Bela board for low-latency audio processing, an embedded loudspeaker, Wi-Fi, and a lightweight power supply. At software level these instruments run an audio engine written in Pure Data (PD) that processes with effects the sound captured by the microphones, generates sounds from synthesizers and samplers, extract audio features, and maps the interactions of the player with the sensors to sound parameters.

We also designed four prototypes of an armband-based SMHW (Fig. 1), with hardware components similar to the SMIs as well as actuators and push buttons. A dedicated software synthesized tactile stimuli by means of Pulse Width Modulation.

*The work of Luca Turchet is supported by a Marie-Curie Individual fellowship from the European Union's Horizon 2020 research and innovation programme, under grant agreement No. 749561. Mathieu Barthet also acknowledges support from the EU H2020 Audio Commons grant (688382).

SMIs and SMHWs were connected to a local wireless network by means of a router (IEEE 802.11.ac standard over the 5GHz band). Interoperability between devices was achieved with Open Sound Control messages over UDP.

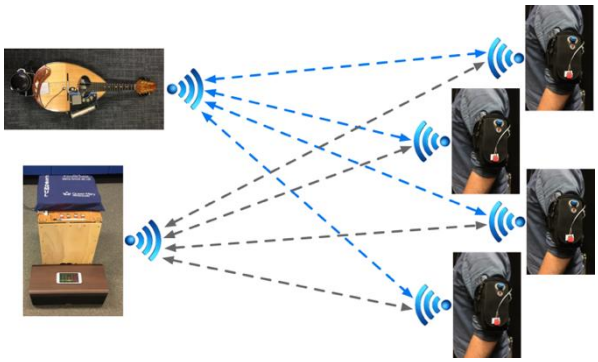

Figure 1. The developed IoMusT architecture.

The onset of hits and strums above an amplitude threshold are extracted in real-time from acoustic signals captured by the microphones. This information was sent to the four SMHWs and mapped to a strong and short vibration so that audience members can experience a tactile stimulation in correspondence of strongest hits and strums. To create participatory interactions, the SMHWs buttons can be used by an audience to deliver to the SMI players tactile vibrations conveying four directions: start/stop playing (continuous vibration of 2 seconds), play faster/slower (intermittent pulses of increasing/decreasing intensity in 5 seconds).

III. CONCLUSIONS

Results of the technical validation of the architecture proved to be stable and reliable in supporting the described interactions. Preliminary perceptual tests show that the latency between sounds and vibrations is perceived as negligible by audience members, as well as that performers correctly react to the tactile directions.

REFERENCES

[1] L. Turchet, A. McPherson, and C. Fischione. *Smart Instruments: Towards an Ecosystem of Interoperable Devices Connecting Performers and Audiences*. In Proceedings of Sound and Music Computing Conference, pp. 498–505, 2016.

[2] L. Turchet and M. Barthet. *Envisioning Smart Musical Haptic Wearables to Enhance Performers' Creative Communication*. In Proceedings of International Symposium on Computer Music Multidisciplinary Research, pp. 538–549, 2017.

[3] L. Turchet, C. Fischione, and M. Barthet. *Towards the Internet of Musical Things*. In Proceedings of Sound and Music Computing Conference, pp. 13–20, 2017.

Assessing the Use of Metrical Information in LSTM-based Polyphonic Music Sequence Transduction

Adrien Ycart[1*] and Emmanouil Benetos[2]

[1*] C4DM, Queen Mary University of London, United Kingdom, a.ycart@qmul.ac.uk
[2]C4DM, Queen Mary University of London, United Kingdom

***Abstract*— Automatic Music Transcription often requires transformation of a sequence of pitch estimations into a binary piano-roll. On this task, we compare LSTMs using time step lengths of 10ms against a sixteenth note. Results indicate that the latter perform better than the former. Further study will confirm to what extent it does.**

I. INTRODUCTION

Automatic Music Transcription (AMT) is one of the most widely discussed Music Information Retrieval tasks. Still, it remains a challenging problem, in particular in the case of polyphonic music.

Most AMT systems use the following workflow. First is extracted from the signal some frame-wise, real-valued pitch estimations in the form of a *posteriogram*, which is then post-processed to obtain a binary *piano-roll*. We focus here on the latter, in the specific case of piano music.

Recurrent neural networks have become increasingly popular for sequence transduction in a variety of domains. For AMT, some quite complex architectures have been developed [1]. We assume that by using musically relevant (*note-based*) time steps, such as a sixteenth note, instead of *time-based* time steps of ten milliseconds, their performance could be greatly improved.

II. DATASET

We use the MAPS [2] dataset. For note-based time steps, the rhythmic grid, *i.e.* the location of each sixteenth note, must be known. In real life we would rely on beat-tracking algorithms; here we consider it given. To determine it, we use symbolic alignment [3] between the MAPS files, and the same piano pieces with quantized durations. We make this rhythm ground truth available for future use[1].

III. MODEL

We use a single-layer Long Short-Term Memory (LSTM) architecture. We aim solely to compare time- and note-based time steps, so we deliberately use a simple architecture. We train it using as inputs posteriograms obtained with [4] and as targets the ground-truth piano-roll.

IV. EXPERIMENTS

We compare two LSTMs (time- and note-based) with two simpler approaches: median filter and thresholding (*Baseline*) and a pitch-wise on-off Hidden Markov Model (*HMM*) method [5]. On the frame level, the LSTM improves the results over both simpler methods. When evaluated on the note-level, the baseline is better than the LSTM: the LSTM over-fragments notes, yielding poor precision. In both cases, using note-based time steps improves the results of the LSTM.

V. DISCUSSION

A disadvantage of sixteenth note time steps is that they misrepresent tuplets and ornaments. Yet, they improve the results in two respects: they quantize the output durations, and they allow to better model dependencies between notes, as suggested in [6]. It is yet unclear which one has the most influence. It is also possible that the LSTM corrects the error of the acoustic model on one specific piano, but doesn't generalize to other piano models. Such questions will be investigated in further study. Future work will also include using note-based time steps with more sophisticated architectures. Finally, we considered the rhythmic ground truth given in this experiment; we will have to make sure that there is still an improvement when using imperfect beat tracking algorithms.

REFERENCES

[1] N. Boulanger-Lewandowski, Y. Bengio and P. Vincent, "High-dimensional Sequence Transduction", in *IEEE ICASSP*, 2013
[2] V. Emiya, R. Badeau, and B. David, "Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle", *IEEE/ACM TASLP*, 2010
[3] E. Nakamura, K. Yoshii, and H. Katayose, "Performance Error Detection and Post-processing for Fast and Accurate Symbolic Music Alignment", in *ISMIR*, 2017
[4] E. Benetos and T. Weyde, "An Efficient Temporally-constrained Probabilistic Model for Multiple-instrument Music Transcription", in *ISMIR*, 2015
[5] G. E. Poliner and D. P. W. Ellis, "A Discriminative Model for Polyphonic Piano Transcription", in *IEEE ICASSP*, 2006
[6] A. Ycart and E. Benetos, "A Study on LSTM Networks for Polyphonic Music Sequence Modelling", in *ISMIR*, 2017

[1] http://www.eecs.qmul.ac.uk/~ay304/code/icassp18.html

*Research supported by Queen Mary University of London

Musicians' Binaural Headphone Monitoring for Studio Recording

Valentin Bauer[1*], Hervé Déjardin[2] and Amandine Pras[1,3]

[1*]Advanced music production program (FSMS), Paris Conservatoire (CNSMDP), France, vbauer@cnsmdp.fr
[2]Quality and innovation department, Radio France, France
[3]Department of music, University of Lethbridge, Canada

***Abstract*— Musicians face various challenges when recording in studio with headphone monitoring. This study investigates the use of binaural technology as a possible monitoring solution in the context of world music, jazz and free improvisation.**

I. INTRODUCTION

Based on the first author's experience as a recording engineer, this study extends previous binaural monitoring solutions. For instance, Soudoplatoff and Pras examined conductors' challenges through an online survey completed by 12 international conductors. They also carried out two comparative tests in situ with binaural rendering and head tracking [1]. One of the tests was conducted in rehearsals of a live film-scoring performance with a symphonic orchestra, which showed positive results in terms of realism, hearing comfort and performance precision.

Our study focuses on musicians' challenges and includes three objectives: to establish musicians' challenges and their need to perceive a realistic auditory scene when monitoring on headphones during recording sessions; to identify the pros and cons of binaural headphone monitoring; and to observe a potential impact of binaural monitoring on musical performance and creativity.

II. METHOD

Twelve international musicians, with an average of 13 years of improvisation practice and 14 years of studio recording experiences, filled out an online survey with 12 semi-directed questions about their experience with headphone monitoring. The first author then conducted two exploratory tests with binaural headphones in a film-scoring recording context, both involving a soloist (one singer and one bass flutist). Eventually, he produced three recording sessions with a world music musician, a free improvisation trio and a jazz trio that allowed for comparative tests between stereo and binaural headphones. These sessions were followed by semi-directed focus groups to investigate in-depth musicians' perceived differences between binaural and stereo.

III. RESULTS AND DISCUSSION

From the content analysis of the online survey emerged headphone monitoring challenges such as coping with a situation that is perceived as unusual by performers; a lack of realism or sound quality of the auditory scene, the need for individual headphone monitoring and technical issues of the devices or systems. Furthermore, musical consequences due to headphone monitoring are only mentioned in a negative way, which implies that headphone monitoring could only worsen the musical performance.

In-situation film-scoring recording sessions showed that binaural technology allows for sound immersion and a pleasant sound quality, though a lack of definition was perceived by the bass flutist. Thanks to the *sound unmasking* potential of binaural technology, the comparative tests highlighted that binaural monitoring enhances musicians' comfort and pleasure, as well as the general sound quality of the mix.

It also emerged that musicians do not need to monitor an auditory scene on headphones that respects the real acoustic sound spatialization of the studio (musicians positioning). On the contrary, the use of binaural space for headphone spatialization allow musicians to improve their perception of the instruments, and/or to renew with a known musical performance situation (rehearsals or concerts). Lastly, results showed that a more detailed auditory scene enables a better musical performance and more creativity and freedom in the studio, with binaural monitoring sometimes even described as a creative support for musical performance.

IV. CONCLUSION

This research contributes to improve sound quality and realism of auditory scenes on headphones during recording sessions, and thus musicians' comfort. Our positive results regarding binaural headphone monitoring as a significant support for creativity call for further research including the development of a methodology that would allow musicians to describe these creative aspects more precisely.

REFERENCES

[1] Soudoplatoff, D., & Pras, A. (2017). Augmented reality to improve orchestra director's headphone monitoring. In the proceedings of the 142nd Audio Engineering Society Convention in Berlin, Germany.

Towards Bio-responsive Control for Music

Duncan Williams[1], Damian T. Murphy[1] and Bruno M. Fazenda[2]

[1]Digital Creativity Labs, University of York, UK, duncan.williams@york.ac.uk
[2]Acoustics Research Centre, Sch. of Computing Science and Engineering, University of Salford, UK

***Abstract*— Music and audio applications are well suited to tactile control [1]. In sound and music computing there can be a disconnect between design of human-computer interfacing and application congruent design. A categorical approach is proposed, considering active and passive control methods. This work has implications for the design of adaptive or 'on-the-fly' recalibration of music and sound in various contexts, including health and wellbeing, video game soundtracking, and perceptual evaluation of auditory stimulus (e.g., noise annoyance, concentration and attention, relaxation and mindfulness). Due to a lack of agreement on suitable evaluation strategies, a multi-criteria decision aid strategy adopted from the auditory display community is suggested.**

I. INTRODUCTION

Bio-physiological interfacing (for example heart rate, skin response, or brain activity as measured via electroencephalography) is beginning to offer tools which might realistically be adapted to sound and music computing, facilitating new ways of interacting with music, and widening participation to maximise the health and wellbeing benefits which music can provide the listener. Music has been shown to improve athletic performance, reduce stress, increase mindfullness, and aid concentration. There is a large potential audience of individuals who might otherwise be unable to take part in music making via traditional means (either due to lack of training, or physical disability), who might benefit from biophysiologically-informed computer aided interaction with music. Additionally, the delivery of adaptive music benefits from listener-state information which can be gathered via biosensors. In order to be useful, the mapping between biophysiological cue and audio parameter must be intuitive and useful to a neophyte audience.

II. CATEGORICAL APPROACH

We propose three categories of system: conscious, unconscious, or hybrid. Various HCI systems for interacting with music have been developed which can be placed in these categories: an emotion-driven music generator under the control of galvanic skin response (GSR) [2] would fall under the unconscious category. An audio mixer using alpha and beta waves measured by EEG to adjust fader gains is an example of a conscious control [3]. Hybrid systems would make use of both active and passive control. In an end user/consumer context, the use of audio mappings could give the listener a new way to select, create, or manipulate emotionally-congruent music (e.g., biophysiologically informed playlist generation) to enhance a mood or emotional state – perhaps relaxation or concentration. This requires a system to respond adaptively and intuitively without direct user input, in order to avoid distracting the listener from the intended emotional state. The user responses could then be utilised to train a machine learning algorithm, adjusting the mapping on-the-fly according to biophysiological response for optimal performance. This would allow significant progress in developing individual and adaptive systems.

III. FURTHER WORK

Significant further work involving careful mapping between the categorical, context-mapping, and adaptive loop remains. Evaluating the success of such systems is difficult partially due to the infancy of the field and the lack of agreement regarding appropriate strategies. We propose borrowing from the world of auditory display where multi-criteria decision aid analysis has been shown to be useful [4]. Criteria should be selected in line with the end use goals; e.g., utility of control, congruence of mapping to audio feature. The challenge is interdisciplinary and requires collaboration end-user populations, computer scientists (particularly in the training stage of the feedback response), and specialists in biophysiological measurement.

ACKNOWLEDGMENT

A portion of this work was conducted in the Digital Creativity Labs (www.digitalcreativity.ac.uk), jointly funded by EPSRC/AHRC/InnovateUK under grant no EP/M023265/1.

REFERENCES

[1] S. Merchel, M. E. Altinsoy, and M. Stamm, "Touch the sound: audio-driven tactile feedback for audio mixing applications," J. Audio Eng. Soc., vol. 60, no. 1/2, pp. 47–53, 2012.

[2] I. Daly et al., "Towards human-computer music interaction: Evaluation of an affectively-driven music generator via galvanic skin response measures," 2015, pp. 87–92.

[3] E. R. Miranda, "Plymouth brain-computer music interfacing project: from EEG audio mixers to composition informed by cognitive neuroscience," Int. J. Arts Technol., vol. 3, no. 2, pp. 154–176, 2010.

[4] K. Vogt, "A quantitative evaluation approach to sonifications," in Proceedings of the 17th International Conference on Auditory Display (ICAD 2011). Budapest, Hungary. CD-ROM, 2011.

A Statistical-Learning Model of Harmony Perception

Peter M. C. Harrison[1*] and Marcus T. Pearce[1]

[1]Centre for Digital Music, Queen Mary University of London, UK, p.m.c.harrison@qmul.ac.uk

***Abstract*— This work presents and validates a new computational model that explains harmony perception in terms of unsupervised statistical learning.**

I. INTRODUCTION

A core methodology in cognitive science is the construction and evaluation of computational models of human cognition. Here we present a new computational model that explains harmony perception in terms of statistical learning. According to this model, listeners learn statistical regularities of harmonic styles through unsupervised learning, and make probabilistic predictions about upcoming harmonic events on the basis of these learned statistics.

II. MODEL

Many statistical learning models could be used to model human listeners. We guide our model choice through two assumptions. First, we treat humans as optimal learners. This means that, given a set of candidate cognitive models, we prefer the one with the highest predictive accuracy. Second, we assume that listeners have access to a wide range of harmony representations, typically of physiological or psychological origin, that they can use when learning.

The proposed model employs a multiple-viewpoint architecture (Conklin & Witten, 1995; Pearce, 2005) to model sequences of chord symbols. Under this approach, separate predictive models are trained for each representation, and the predictive distributions of these models are combined through ensemble averaging. The resulting model predicts the probability of successive chords conditional upon their previous context.

We make two extensions to the multiple viewpoint framework. The first is the introduction of a continuous ensemble weighting scheme, whereby the contribution of each representation (or viewpoint) is optimized to maximize the predictive accuracy of the ensemble. The second is the introduction of continuous representations, which are modeled by means of *k*-means quantization.

III. STUDY 1: REPRESENTATIONS

An initial computational study investigated which representations best facilitate the prediction of harmonic structure in different musical genres. A large set of candidate representations was constructed from music theory and music psychology, embodying concepts such as dissonance, tonality, relative pitch, and voice-leading distance. These representations were evaluated on datasets (c. 1000 compositions each) of chord sequences from three musical genres: classical, popular, and jazz. Results indicated the following: a) the best representation depends on training-set size; b) simple representations (in particular dissonance) are useful for small training sets; c) complex representations (in particular expressing the chord relative to the local tonic, or relative to the bass note of the previous chord) are useful for large datasets; d) using these derived representations achieves substantial performance improvements (10-30% reduction in cross-entropy) for all genres and training-set sizes tested.

IV. STUDY 2: PSYCHOLOGICAL VALIDATION

A subsequent psychological study investigated how well the new model explains human perception of chord sequences. Forty-four participants were played 300 8-chord sequences sampled randomly from the Billboard popular music corpus, and rated the sixth chord in each sequence for surprisingness. Participant ratings were standardized to *z*-scores and averaged to producing one surprisal rating for each sequence. The new statistical-learning model was then compared with three prominent models of harmony perception from the psychological literature in terms of the ability to predict surprisal ratings. Results indicate that the new model predicts much of the variation in surprisal ratings ($r(298) = .62$, $p < .001$), and substantially outperforms the next best-performing model from the literature ($r(298) = -.17$, $p = .004$).

V. CONCLUSION

The new statistical-learning model provides a fairly good account of harmony perception. However, there is still significant room for improvement. We are continuing to work on the model, using two main strategies: optimizing the model's ability to predict harmonic sequences (a machine-learning task) and incorporating suitable perceptual biases into the model (a psychological task).

ACKNOWLEDGMENT

The authors thank Emmanouil Benetos and Matthew Purver for helpful discussions concerning this project.

REFERENCES

[1] Conklin, D., and Witten, I. H. (1995), "Multiple viewpoint systems for music prediction". J. New Music Res, vol. 24, pp. 51–73.

[2] Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* (PhD thesis). City University, London, UK.

*P. M. C. H. is supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

An Agent on my Shoulder: AI, Privacy and the Application of Human-Like Computing Technologies to Music Creation

Alan Chamberlain[1*] Alessio Malizia[2] & David De Roure[3]

[1*] MRL, Computer Science, University of Nottingham, UK, Alan.Chamberlain@Nottingham.ac.uk
[2]School of Creative Arts, University of Hertfordshire, UK
[3]Oxford e-Research Centre, University of Oxford, UK

***Abstract* — Human-Like Computing technologies are intelligent systems that interact with people in human-like way. By bringing together the disciplines of Artificial Intelligence, Ethnography and Interaction Design, and applying them in a real world context we are able to understand some of the ways that such technologies can be applied. This work in progress poster applies such technologies to the music creation and develops a design that is based on the notion of an 'Intelligent' Agent that is able to support in the music creation process.**

I. MUSIC, IN PRACTICE

For anyone involved in the creative industry, the world appears to be full of concerns and challenges when it comes to dealing with digital rights, ownership, originality and re-use. The music industry is a prime example of this. With the development of different ways to create, distribute, compose, consume and perform the music industry is becoming more complex. The rise in digital technologies for creating and distributing music has meant that billions of audio files (and related content) are served to the public, and with many services offering multiple channels and methods of consuming music it is difficult to see what the most appropriate ways are for musicians to sell and keep a record of what is being served. If we then think about the way that one can create music using Digital Audio Workstations, songs can and are made out of multiple stems, samples, sets and in many respects it is difficult to know how and when to protect and keep this work private, or when to and how to release developing work in order to generate publicity. This abstract discusses an initial description of a conceptual system that is looking at the provision and design of Human-Like intelligent agents that can support musicians as they create and release music, in respect to the challenges that we have outlined.

II. AN "AGENT" ON MY SHOULDER

We use the metaphor of an "*Angel on my shoulder*" in order to represent the Intelligent Agent that acts in a human-like way to give advice and support, this extends our earlier work relating to robots and music [1]. This isn't advice that relates to composition practice, but as we earlier outlined information and advice that relates specifically to the rights, releasing, re-use and distribution of music – offering advice on the implications of such actions. In our early research we found that people can have a playful attitude to new musical technologies [2]. Based on this, and a need to train the system, we propose a game-like approach that can both support the user in terms of getting used to the system and for the system to learn more about the user and to 'learn' to react in a more human-like way. Using a gaming approach enables the user to get used to the agent-based system in more 'natural' way and also learn about the issues and legalities of the music industry.

III. IOT, DATA & MUSICAL EQUIPMENT

With the development of IoT-based musical equipment instruments are possible to capture data throughout the creative process, and this data may be used in a variety of ways that could support the system in learning about user practices, but could also be used in other ways to support the design of future systems, provenance composition and even support learning. So although our initial framework is based in managing digital rights a Human-Like Computing approach could work in a variety of contexts within the music domain, which could lead to the development of intelligent plugins. Using intelligent plugins that can support and inform musical practices across software, hardware, platforms and channels offer new and interesting challenges for the development of an *Ecology of Audio Technologies* that has yet to be realized or fully understood.

IV. CONCLUSION

Human-Like Computing technologies and the Internet of Things offer new opportunities for the music industry, but also create new challenges. In order to fully appreciate the ramifications of using such technologies there still needs to be more understanding of the practices of people making music in the 'wild' [3]. It is hoped that this poster gives some insight into this area and generates some interest in this emerging area of research.

REFERENCES

[1] Alan Chamberlain (2017) "Are the Robots Coming? Designing for Autonomy & Control in Musical Creativity & Performance." Audio Mostly 2017: 23-26 August. QMUL, ACM.
[2] Andrew McPherson, et al (2016) "Designing for Exploratory Play with a Hackable Digital Musical Instrument", Proceedings of Designing Interactive Systems, DIS'16, June 4 - 8, ACM.
[3] Crabtree, A. et al. (eds.) (2013) Special issue on "The Turn to the Wild" with authored introduction", ACM Transactions on Computer-Human Interaction - ToCHI 20(3), 13:1-13:4.

* This research was supported through the following EPSRC project: Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (EP/L019981/1).

Inverting Feature Representations of Machine Listening Systems

Saumitra Mishra, Bob L. Sturm and Simon Dixon

Centre for Digital Music, Queen Mary University of London, U.K., saumitra.mishra@qmul.ac.uk

***Abstract*— We extend supervised generative model-based inversion of feature representations [1] to machine listening, and demonstrate it for singing voice detection [2]. Inverting the generated representations highlights the information preserved at each layer during discriminative learning, thus assisting to analyse the behaviour of the model.**

I. MOTIVATION

Convolutional neural network (CNN) models are state-of-the-art for several pattern classification problems in vision, text, audio etc. But, they suffer from uninterpretable predictions and hard-to-analyse behaviour. Recent research in machine learning introduces several techniques to understand the behaviour of these black-box models (e.g., instance-based saliency maps, activation-maximisation). One such method [1] aims to understand the *information* (input properties) preserved by a CNN, while it learns discriminatively (throw-off irrelevant information) from data. To gain such an insight about a CNN model, we invert the feature representations captured at each layer of hierarchy. Feature inversion-based reconstructions will highlight the features and the invariances (amount of blur) captured at each layer of CNN, providing insight into the model behaviour.

II. SUPERVISED GENERATIVE MODEL-BASED FEATURE INVERSION FOR MACHINE LISTENING

We extend the method proposed in [1] to machine listening systems. It inverts a feature representation (λ) by training a supervised generative model (e.g., CNN as a decoder) to predict the *expected input* (weighted average of all inputs that could be mapped to λ). This method is particularly useful as it learns an audio prior automatically from the data, thus avoiding the difficult task of creating hand-crafted regularizers. Formally, given an input audio excerpt X_i, and its feature representation λ_i, the method solves (1) to learn the weights w of an up-convolutional network $\Upsilon\ (\lambda, w)$.

$$\arg min_w \sum_i ||X_i - \Upsilon(\lambda_i, w)||_2^2 \qquad (1)$$

We apply the proposed method to invert the feature representations captured by a CNN-based singing voice detection system [2]. This model attains state-of-the-art results on standard datasets used for vocal detection task. We train two generative models, one for each fully-connected layer (FC7 and FC8) in the network. The architecture of the model trained to invert last fully connected layer (FC8) with dimensionality 64 is shown in Table 1. Filters used in the upconvolutional and convolutional layers are of size 4 x 4 and 3 x 3, respectively. Fig. 1 depicts the reconstructed Mel-spectrograms when for a randomly selected excerpt, the generative models invert the feature vectors captured at the layers FC8 and FC7, respectively.

TABLE I. GENRATIVE MODEL ARCHITECTURE TO INVERT FC8 FEATURE VECTOR. LAST DIMENSION OF UCONV AND CONV LAYER INPUTS CORRESPONDS TO THE NUMBER OF CHANNELS.

Layer	*Input*	*Neurons/Filters*	*Padding, Stride*
FC1	64 x 1	64	-
FC2	64 x 1	256	-
Reshape	256 x 1	-	-
Uconv1	4 x 4 x 16	16	1, 2
Conv1	8 x 8 x 16	16	1, 1
Uconv2	8 x 8 x 16	8	1, 2
Conv2	16 x 16 x 8	8	1, 1
Uconv3	16 x 16 x 8	4	1, 2
Conv3	32 x 32 x 4	4	1, 1
Uconv4	32 x 32 x 4	2	1, 2
Conv4	64 x 64 x 2	2	1, 1
Uconv5	64 x 64 x 2	1	1, 2

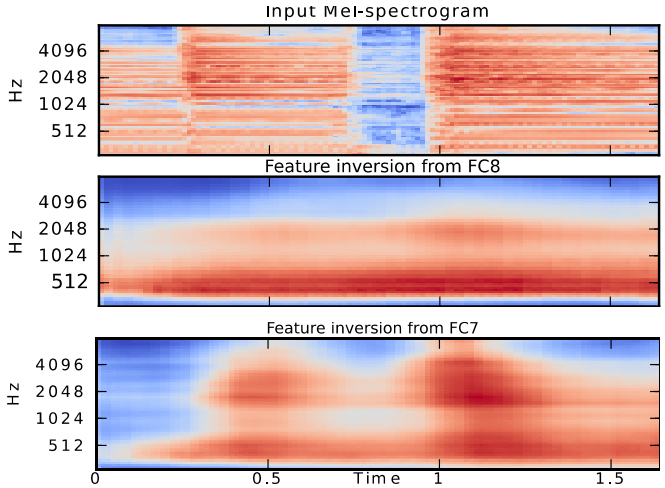

Figure 1. Feature inversion-based reconstructed Mel-spectrograms

Above visualizations suggest that the high frequency content of the input excerpt is not preserved in the FC8 layer. Moreover, it also loses track of the onset positions. FC7 layer on the other hand preserves the high frequency content and the approximate shape of the spectrum. The blurriness in the reconstruction corresponds to the amount of invariance of the feature representation.

REFERENCES

[1] A. Dosovitskiy and T. Brox, "Inverting Visual Representations with Convolutional Networks," in *Proc.* CVPR, 2016.

[2] J. Schluter and T. Grill, "Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks," in *Proc. ISMIR*, 2015.

Social Music Machine: Crowdsourcing for Composition & Creativity

Alan Chamberlain[1*], David De Roure[2] and Pip Willcox[2]

[1*] MRL, Computer Science, University of Nottingham, UK, Alan.Chamberlain@Nottingham.ac.uk
[2] Oxford e-Research Centre, Oxford University, UK

***Abstract*— This poster describes a compositional technique that used crowd-sourced midi clips in order to develop a piece of music, which was later performed. This work in progress highlighted some of the issues facing the designers of systems that enable the 'crowd' to compose.**

I. INTRODUCTION

Can the crowd get creative? And what sort of tools might be used to support this? These are the sorts of questions that we thought about when we initially started to think about these problems. Using software originally developed as part of an Experimental Digital Humanities [1] project, we started to wonder about how such software - "*Numbers into Notes*" [2] might work in the real world if multiple people used it in creative way, and what lessons might we learn from carrying out such an intervention.

II. CROWDS MAKE MUSIC - TOOLS AND EXPERIMENTS

People were asked to use the "*Numbers into Notes*" software and make a sequence (of notes). They then gifted the sequence to the 'composer' who used the sequence as part of a larger piece of music. Overall five sequences where gifted and used in the performance. The reasoning behind the intervention was to see if this was a viable compositional technique, how a performer / composer might use the sequences, and in order to explicate and unpack the issues and practices that might emerge from engaging in such an 'experiment'.

III. THE COMPOSITION AND PERFORMANCE

Each algorithm that was generated was a simple sequence of notes. 5 of these were *gifted* by people and these used in the piece. – *The performance* [3] – used Ableton Live as a platform to play the loops (sequences) and to bring the loops 'in' and 'out' of the mix. The interface was laid out in a way that enabled the performer to follow the performance workflow/order. A *Monome* (Walnut 64) was used with a *Max for Live* patch, which was set to semi-random, this controlled a Grand Piano sound and vocal samples (created by the performer) simultaneously. The performer was able to control this in order to avoid blandness and too much repetition. The gifted algorithms where brought in and out of the mix throughout the piece (the sounds and effects were developed prior to the performance), and the piece was brought to an end with fading in of some field recordings of church bells and a vocal recording of some related readings. Beats where used in the piece towards the movement into the field recording. This was purposefully done for the live performance to keep the audience interested. As the performer *Alan* felt it was important to understand the structure of the piece and its constituent parts, but practicing the piece would have led to an uninteresting performance, so parts of the performance are purposefully random, but controllable. It may appear fairly obvious, but a key part of performing and composing the work related to putting the pieces together in a way that worked, wasn't bland, overly repetitive and kept the audience interested. This is a key issue for systems that have autonomous elements [4] and can inform the design of Human-Like Computing systems for creative applications such as creating music.

IV. CONCLUSION

Crowdsourcing musical composition appears simple in many respects, but to really understand compositional practice and performance, one really has to '*do it'*, and become part of the social machine. Using autoethnographic methods [5] would be a way to further 'unpack' such systems and inform design. This initial experiment has helped us to understand and think about a whole range of issues that can impact upon designing creative crowdsourcing systems.

REFERENCES

[1] David De Roure, et al. (2016) Experimental Digital Humanities: Creative interventions in algorithmic composition on a hypothetical mechanical computer. In: DMRN+11: Digital Music Research Network, December 2016.

[2] Alan Chamberlain, et al. (2017) "Audio Technology and Mobile Human Computer Interaction: From Space and Place, to Social Media, Music, Composition and Creation", In the *International Journal of Mobile Human Computer Interaction* (IJMHCI) V9/4, pp. 25 – 40

[3] Alan Chamberlain, and David De Roure, (2017) The gift of the algorithm: beyond autonomy and control. Performed at Oxford House, London 25th.

[4] Alan Chamberlain (2017) "Are the Robots Coming? Designing for Autonomy & Control in Musical Creativity & Performance." Audio Mostly 2017, QMUL, London. ACM.

[5] A. Chamberlain, M. Bødker & K. Papangelis (2017) "Mapping Media and Meaning: Autoethnography as an approach to designing personal heritage soundscapes" Audio Mostly 2017, QMUL, London. ACM.

* This research was supported through the following EPSRC project: Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (EP/L019981/1). Thanks to Geert De Wilde, Maria Kallionpää, Matthew Yee-King & Dafydd Roberts.

Feature Design for Intelligent Control of the Dynamic Range Compressor Using Audio Decomposition

Di Sheng and György Fazekas

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

***Abstract*— We propose a method for the intelligent control of the dynamic range compressor targeting mono-timbral loops. Initial research using random forest regression has been shown to work in the context of isolated notes [1]. Since audio loops have become the important in many production scenarios, this paper addresses this problem by decomposing loops into appropriate inputs for the initial system. We explore three types of audio decomposition approaches, onset event detection, NMF, and audio transient/stationary separation using ISTA, and extract features correspondingly. Results show a convincing trend that using features extracted in the decomposition domain to train the regression model improves the performance both numerically and perceptually.**

I. INTRODUCTION

Our previous work for the intelligent control of dynamic range compressor (DRC) proposed in [1] uses a regression model to map audio features to compressor parameters. The model is trained using standard audio features as well as features designed specifically to detect acoustic qualities related to DRC in the context of mono-timbral notes. In this paper we adapt this method to mono-timbral loops by applying audio decomposition. We choose three approaches to decompose loops. The most straightforward method is onset event detection. In [2], the authors suggested guidelines for choosing the appropriate onset event detection function. Time domain methods are normally adequate for percussive signals, while spectral methods based on phase distributions or spectral difference are suitable for pitched transients. Complex-domain spectral difference methods work well for many cases but the computational cost is larger. Since mono-timbral loops do not exhibit complex music structure, we opt for the High Frequency Content detection function as a strating point. The second approach utilises source separation using Non-negative Matrix Factorisation (NMF) to decompose complex audio into activation patterns [3]. Finally we assess transient/stationary audio separation. We apply a recently proposed algorithm called Iterative Shrinkage Threshold Algorithm (ISTA) [4] for our purpose.

II. METHODOLOGY AND EVALUATION

The features designed for attack and release time require the most attention. Eqn.1-3 correspond to the length, average energy and ascending speed of the attack phase. *NstartA* and *NendA* represent the start and end positions of the attack, which is calculated using the RMS curve through a fixed threshold method. Release features are calculated in a similar manner.

$$\mathsf{T_A} = (N_{endA} - N_{startA})/Fs, \quad (1)$$

$$\mathsf{A1_{att}} = \frac{1}{N_{endA} - N_{startA}} \sum_{n=N_{startA}}^{N_{endA}} rms_curve(n), \quad (2)$$

$$\mathsf{A2_{att}} = rms_curve(N_{endA}), \quad (3)$$

Since most loops contain heavily overlapped notes, using onset event detection alone cannot guarantee each note has a clearly detectable attack/release part. We select notes that contain clear attack/release by the condition of *goodness of fit*. NMF is able to provide note-like active pattern which can be used for attack/release feature extraction. Finally, ISTA can directly provide the position of attack/release, which means we can apply Eqn. 1-3 directly. The mean of the features for each individual notes are used as the feature for the loop. Numerical test results and similarity test results are given in the following tables. NMF stands out for both test, but using three features together delivers the best performance.

MAE(ms)	*Std*	*Onset*	*NMF*	*T/S*	*All*
$Guitar, \tau_a$	0.934	0.897	**0.845**	0.863	**0.807**
$Bass, \tau_a$	1.449	1.196	**1.071**	1.244	**0.995**
$Drum, \tau_a$	1.384	1.361	**1.194**	1.274	**1.134**
$Guitar, \tau_r$	12.115	10.604	**10.442**	11.802	**9.981**
$Bass, \tau_r$	11.701	11.143	**10.733**	10.886	**9.381**
$Drum, \tau_r$	16.327	14.946	**12.714**	13.315	**12.043**

TABLE I. PREDICTED MEAN ABSOLUTE ERROR(MAE) USING DIFFERENT FEATURE SETS FOR LOOPS OF THREE INSTRUMENTS.

	$D2_{std}$	$D2_{onset}$	$D2_{nmf}$	$D2_{t/s}$	$D2_{all}$
$Guitar, \tau_a$	0.918	0.916	**0.914**	0.916	**0.916**
$Bass, \tau_a$	0.384	0.375	**0.371**	0.383	**0.362**
$Drum, \tau_a$	**0.251**	0.252	**0.251**	0.257	0.252
$Guitar, \tau_r$	0.934	0.936	0.940	**0.919**	**0.917**
$Bass, \tau_r$	0.738	0.732	**0.726**	0.733	**0.729**
$Drum, \tau_r$	0.583	0.589	**0.580**	**0.582**	0.584

TABLE II. D1 AND D2 COMPARISON USING DIFFERENT FEATURE SETS, WHEN D() IS THE AUDIO PERCEPTUAL SIMILARITY.

REFERENCES

[1] Di Sheng and György Fazekas, "Automatic control of the dynamic range compressor using a regression model and a reference sound," in Proceedings of the 20th International Conference on Digital Audio Effects (DAFx- 17), 2017.

[2] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler, "A tutorial on onset detection in music signals," IEEE Transactions on speech and audio processing, vol. 13, no. 5, pp. 1035–1047, 2005.

[3] Nancy Bertin, Roland Badeau, and Gael Richard, "Blind signal decompositions for automatic transcription of polyphonic music: Nmf and k-svd on the benchmark," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 1, pp. I–65.

[4] Kai Siedenburg and Simon Doclo, "Iterative structured shrinkage algorithms applied to stationary/transient sep- aration," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, 2017.

The Art and 'Science' of Opera: Composing, Staging & Designing New Forms of Interactive Theatrical Performance

Alan Chamberlain[1*] Maria Kallionpää [2] Steve Benford[1*]

[1*]MRL, Computer Science, University of Nottingham, UK, Alan.Chamberlain@Nottingham.ac.uk
[2]Department of Communication and Psychology, Aalborg University, Denmark

***Abstract*— New technologies, such as Virtual Reality (VR), Robotics and Artificial Intelligence (AI) are steadily having an impact upon the world of opera. The evolving use of performance-based software such as Ableton Live and Max/MSP has created new and exciting compositional techniques that intertwine theatrical and musical performance. This poster presents some initial work on the development of an opera using such technologies that is being composed by Kallionpää and Chamberlain.**

I. BEGINNINGS

Opera as an art form is interesting because it is both musical and theatrical, and as such it is an ideal new-media research platform that can allow researchers to develop new technologies and interactional techniques that at a high-level explore the interplay between audience, performance, composition and staging. In many respects opera lends itself to being explored though methods developed in Participatory Design, Human-Computer Interaction and Computer-Supported Cooperative Work. We offer a brief insight into some of the initial work and discussions that have started to emerge in regard to the development of an opera called '*Spirits of the Land, Lake and Sea'*, which aims to explore the use of Virtual Reality/Mixed Reality, non-linear performance and narratives, autonomous compositional techniques and mulit-sited/distributed performance.

II. THE OPERA

In this section we quickly discuss our vision for the opera, which is currently in its early phases, we briefly discuss three constituent parts of the opera that will be of interest to the conference audience. '*Spirits of the Land, Lake and Sea'* is envisaged as being a participatory experience. Communities in rural Finland and the UK will form the core inspiration for the narrative of the performance, which will be based on their experiences of living with the land, lakes and sea. In past projects we have used participatory approaches to great effect, as they engender the involvement of the community, which lead to a sense of ownership and a greater commitment to, and involvement in the project [1].

Virtual and Mixed Reality – We intend use VR and Mixed Reality in order to share different environments that will consist of different audio environments and differing VR experiences relating to the different sites where the opera is performed. Mixed Reality techniques will bring the virtual and physical together by the blending of audio, location and performance into a narrative structure.

Non-Linear Narratives & Performance – Developing non-linear compositional and performance techniques is a key area exploration. Tools such as Ableton live mean that non-linear performance and 'triggering' is possible, however the integration of computational performance and acoustic instruments is more complex – relating this to the narrative of the opera will be challenging, using modular compositional approaches may offer a solution.

Autonomous Composition & Streamed Content – We aim to use compositional techniques that will offer us the opportunity to engage with, understand and develop a theoretical framework for composing and performing with autonomous music. As part of the opera we aim to build on our existing research in this area [2] and develop the integration of streamed and sonified content [3] relating to the sites explored.

III. CONCLUSION

Opera is an intriguing space to work when one wants to explore the design and development of new technologies that might impact up the composition, performance and staging of such work. Working in such spaces are complex, but there is value in understanding the way that this art form can offer exciting and new possibilities to further understand the way that new technologies relating to autonomous systems for compositional practice, Virtual Reality spaces for performing and non-linear narrative/performance structures can be developed and applied in the real world.

REFERENCES

[1] Alan Chamberlain, Andy Crabtree and Mark Davies (2013) "Community Engagement for Research: contextual design in rural CSCW system development", The 6th International Conference on Communities and Technology 2013, C&T 2013, Germany. ACM

[2] Alan Chamberlain (2017) "Are the Robots Coming? Designing for Autonomy & Control in Musical Creativity & Performance." Audio Mostly 2017: Augmented and Participatory Sound/Music Experiences, 23-26 August. QMUL (London, UK) ACM

[3] Iain Emsley, David De Roure & Alan Chamberlain (2017) "A Network of Noise: Designing with a Decade of Data to Sonify JANET." Audio Mostly 2017: Augmented and Participatory Sound/Music Experiences, August. QMUL (London, UK) ACM

* This research was supported through the following EPSRC project: Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (EP/L019981/1).

Linear and Logistic Models for Music Classification Experiments

Francisco Rodríguez-Algarra and Bob L. Sturm

Centre for Digital Music, Queen Mary University of London, f.rodriguezalgarra@qmul.ac.uk

***Abstract*— Benchmark classification experiments are pervasive in Music Content Analysis (MCA) evaluation. We here review statistical tools to decompose measurements from such experiments into their main contributions. This is an intermediate step towards the integration of targeted interventions in a formal experimental framework for MCA.**

I. INTRODUCTION

Music Content Analysis (MCA) research often relies on classification experiments to assess and compare methods for the construction of systems and their components. Studies rarely, if ever, include analyses about the contribution of each component to the measurements, though, relying on summary performance metrics instead. Well-established statistical tools could provide greater insights about the suitability of proposed solutions. We here discuss two different approaches: linear mixed-effects models, proposed for learning algorithms evaluation in [1], and mixed-effects logistic models, which, to the best of our knowledge, have not been considered before.

II. MODELLING MEASUREMENTS

Consider a hypothetical MCA study to compare a novel feature extraction algorithm (`b`) against the state of the art (`a`), that reports the mean classification accuracies (± standard deviation), in percentage, in Table I for a series of distinct learning algorithms in 4-fold Cross Validation (CV). What is the contribution of the novel feature extractor?

TABLE I. EXAMPLE RESULTS TABLE

	Learning Algorithm		
Feature Extractor	**I**	**II**	**III**
`a`	70.38 ± 2.45	76.83 ± 0.96	69.13 ± 1.92
`b`	69.90 ± 0.48	79.13 ± 1.38	74.81 ± 1.62

Identifying the inherent structure in the *treatments* of the study – the conditions we aim to compare – and in the *units* – where we obtain measurements from – allows us to isolate the effects of interest from nuisance ones. In MCA, this closely resembles what [1] proposes for the evaluation of learning algorithms. Treatments consist on the system-construction methods, which we can decompose into feature extraction and learning algorithms; units, in general, relate with the collection subsamples on which trained systems perform predictions and from which we calculate performance metrics.

Given the set of feature extractors {`a`, `b`}, and learning algorithms {I, II, III}, as well as the test subsamples (the 4 folds), we can define categorical variables (*factors*) whose *levels* describe each measurement y_i with respect to its corresponding treatment and unit characteristics, which we denote `X`, `L` and `K`, respectively. We can relate the effects by these factors with the sequence of measurements **y** through *structural models*, such as the *linear mixed-effects* model:

$$\mathbf{y} = \boldsymbol{\tau}_{\mathtt{X}} + \boldsymbol{\tau}_{\mathtt{L}} + \boldsymbol{\tau}_{\mathtt{XL}} + \boldsymbol{\beta}_{\mathtt{K}} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{\tau}_{\mathtt{X}}$ and $\boldsymbol{\tau}_{\mathtt{L}}$ denote the *fixed effects* of `X` and `L`, $\boldsymbol{\tau}_{\mathtt{XL}}$ their *interaction*, $\boldsymbol{\beta}_{\mathtt{K}}$ the *random effect* of `K`, and $\boldsymbol{\varepsilon}$ the *residual* – the variability not explained by any other parameter in the model. We can then estimate their corresponding effect and/or infer about whether such effects differ across the distinct levels. Fitting this model with the `lme` function of `R`'s package `nlme`, we find that, e.g., the mean effect of `b` is a *decrease* of 0.48 percentage points in accuracy.

We could instead treat the loss of individual predictions (here, {0, 1}) as observations. The binary responses in this case, however, make the linear model unsuitable. To this end, *mixed-effects logistic* models can appropriately capture the structure underlying the measurements and provide estimates of the effects of the parameters we consider in the models we pose. This approach involves models such as:

$$\text{logit}(\mathbf{z}) = \boldsymbol{\tau}_{\mathtt{X}} + \boldsymbol{\tau}_{\mathtt{L}} + \boldsymbol{\tau}_{\mathtt{XL}} + \boldsymbol{\beta}_{\mathtt{K}} + \boldsymbol{\beta}_{\mathtt{A}} \qquad (2)$$

where **z** is the sequence of individual losses, logit is the natural log of the odds (p / (1-p)), and $\boldsymbol{\beta}_{\mathtt{A}}$ is the random effect of the classes (the annotations). Fitting this model using the `glmer` function of `R`'s package `lme4` also yields a negative effect for `b` (-0.0246). Moreover, we see the estimated standard deviation of $\boldsymbol{\beta}_{\mathtt{A}}$ is relatively high (0.7961).

Using logistic modelling we can include and estimate parameters for, eg., the class, the artist, or the album, which not possible in the linear case, allowing a more fine-grained characterization. The interpretation of these estimates, however, may not be obvious. In any case, no sophisticated statistical technique by itself will reveal whether a solution actually addresses the problem at hand. Our proposal here contributes to a broader formal framework that integrates targeted interventions for MCA evaluation.

ACKNOWLEDGMENT

The authors would like to thank Dr. Hugo Maruri-Aguilar.

REFERENCES

[1] Eugster, M. J. A. "Benchmark Experiments. A Tool for Analyzing Statistical Learning Algorithms", 2011, PhD Thesis, Ludwig-Maximilians-Universität München.

Let's Jam! An Ethnographic Study of Collaborative Music Composing

Juan Pablo Martinez Avila*

Mixed Reality Laboratory, The University of Nottingham, United Kingdom
psxjpma@nottingham.ac.uk

Abstract— In this paper we describe the underlying workflow of collaborative music composing uncovered by a series of ethnographic encounters with a group of musicians from Nottingham. These observations raise design challenges for computer-supported cooperative work applications for musicians, such as version control for compositions that depends on collective agreement of song structures, evaluation of new musical arrangements as well as peer tutoring.

I. INTRODUCTION

For most contemporary music groups, getting together to improvise with their instruments (also known as "jamming sessions") is a crucial part of their composition process. For this ethnographic study, a series of observations were made as a member [1, 2] of a rock band from Nottingham. This band frequently gathered to play at an empty establishment owned by an acquaintance of the bass player. In particular, there were two types of sessions: (a) unstructured jamming sessions, in which the band would freely improvise new musical material and record it with mobile devices; and (b) collaborative composition sessions, in which the raw material was glossed out and restructured by the members of the band in order to compose original music. The workflow underlying collaborative composition is the main focus of this study.

II. WORKFLOW

In this particular band work began by playing recordings of previous improvisations and reviewing them collaboratively until a particular fragment was selected to be further developed. In order to build upon a fragment the musicians needed to construct a shared knowledge about its musical features, e.g. the sequence of notes and rhythms. In some cases this required the musicians to engage in a tutoring process, in which one or more members would demonstrate a sequence of notes clearly and slowly for others to learn, and would also provide feedback on correctness. Conversely, when a new musical arrangement was introduced, the members of the band first had to evaluate its adequacy within the song [4], before deciding to learn it. Nonetheless, this learning process was required by all in order to test a fragment's adequacy for the composition as a whole, even if it was subsequently discarded. The workflow would then alternate between the evaluation and subsequent integration of new arrangements into the composition until a relatively finished version was reached. To wrap up the session, the composition was documented on sheets of paper and text files in a laptop, which would also contain lyrics, chord progressions and other annotations (e.g. context and motives of the song). These files were usually shared through a cloud-based file storage, which was administered and compiled by the leader of the band. These documents were organized mainly in three categories: (1) Raw improvisational music material, (2) Songs in progress, and (3) Consolidated songs.

III. DESIGN IMPLICATIONS

The workflow described in Fig. 1 evidences the importance of constructing shared knowledge and the role of negotiation in collaborative music making environments [3]. This raises new challenges for the design of collaborative music composition software that can support the end-to-end process. Moreover, this study leads the way to future in depth ethnographic research on peer tutoring and the processes embedded in the act of learning an instrument, and its computer-supported cooperative work (CSCW) implications for collaborative music composing.

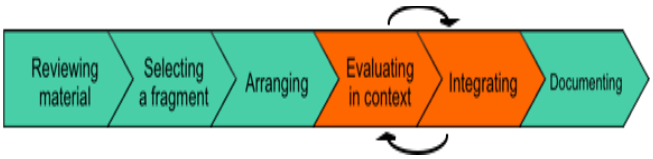

Figure 1. Collaborative composition workflow

IV. REFERENCES

[1] Crabtree, A., Rouncefield, M., and Tolmie, P. Doing design ethnography. (2012).
[2] Garfinkel, H. Studies in ethnomethodology. Prentice-Hall, Englewood Cliffs, N.J, (1967).
[3] McGrath, S., Hazzard, A., Chamberlain, A. and Benford, S. (2016) An ethnographic exploration of studio production practice. In: 2nd AES Workshop on Intelligent Music Production. (2016).
[4] Tolmie, Peter, Steve Benford, and Mark Rouncefield. "Playing In Irish Music Sessions". In Ethnomethodology At Play. (2013). 227-256.

*PhD Student supported by the National Council of Science and Technology (CONACyT) of Mexico and the Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption grant (EP/L019981/1).

Discovering Feature Relevance in Pedalling Analyses of Piano Music

Beici Liang[1*], György Fazekas[1], Mark Sandler[1]

[1]C4DM, Queen Mary University of London, UK, beici.liang@qmul.ac.uk

***Abstract* — Notations of piano pedalling technique in the music score are usually lacking in detail: they provide boundary locations of pedalling techniques, but do not indicate what musical attribute prompts the pedalling change. Understanding this relationship would be useful for musicology and piano pedagogy. We propose to model how musically-motivated features correlate with pedalling transitions. Our aim is to employ this model as prior information for the detection of pedal onsets and offsets from audio recordings.**

I. INTRODUCTION

Pianists typically receive abundant advice on ways to improve their techniques. However, nearly all of this advice refers to the use of the keys, little is usually taught about the use of the pedals. Given that pedal markings are not always notated in the score, acquiring good pedalling techniques merely by guessing and experimenting can be burdensome and difficult. Yet, we believe that the pedal is very important as an invaluable aid in expressive performance and has a decisive influence on the production of piano sound.

The sustain pedal is the most commonly used pedal. It lifts all dampers and sets all strings into vibration due to sympathetic coupling. Pianists frequently exploit the sustain pedal in the playing of Chopin's music for instance. Many declare that pedalling is mainly changed with the harmony of music. Here, "harmony" could mean different things with respect to melody, chord, rhythm, or any combination of these or other musical attributes. Therefore we select midi files and their corresponding audio recordings of Chopin's piano pieces to interpret which musical features form the basis for pedalling transitions.

II. RELATED WORK

The effects of the sustain pedal on the sound has been analysed using isolated notes in [1]. By using the features extracted from harmonics and residuals, it is possible to detect whether notes in the middle region of the piano were played normally or with any of the pedalling techniques employed [2]. However, in the context of polyphonic piano music, the presence of overlapping partials still poses significant challenges to automatic detection of pedalling. In this case, techniques commonly used in source separation or automatic music transcription, such as Non-negative Matrix Factorisation may become necessary for improved feature extraction. Moreover, Bayesian models have been successfully applied to several MIR tasks, such as meter analysis [3]. In this paper, inspired by the technique for automatic interpretation of music structure analysis in [4], we investigate if there are interrelationships between pedalling transition and musical attributes. This could facilitate the development of a Bayesian model for the detection of pedal onsets and offsets using acoustical and musical features simultaneously.

*Research supported by Centre for Doctoral Training in Media and Arts Technology (EPSRC and AHRC Grant EP/L01632X/1), EPSRC Grant EP/L019981/1 "*Fusing Audio and Semantic Technologies for Intelligent Music Production and Consumption (FAST-IMPACt)*" and the European Commission H2020 research and innovation grant AudioCommons (688382). Beici Liang is funded by the China Scholarship Council (CSC).

III. HYPOTHESES AND DEVELOPMENTS

We select four musical attributes. The extracted features should match each of these attributes independently as listed in Table 1. Information about the sustain pedal can be obtained from midi data.

TABLE I. LIST OF FEATURES CHOSEN

Attribute	***Plugins used to obtain feature***
Chord	Chord estimation from *Chordino*
Bass note	Bass chroma from *NNLS Chroma*
Melody	Melody extraction from *MELODIA*
Rhythm	Beat from *DBNBeatTracker*

The goal of this present work is to determine to what extent pedalling transitions can be influenced by the changes in different musically-motivated features. We estimate feature relevance using Pearson correlation coefficient between feature-derived and pedal-derived data matrices. Although we restrict the music features in this study, it will be possible to incorporate more musical parameters in future work.

REFERENCES

[1] H. M. Lehtonen, et al. "Analysis and modeling of piano sustain-pedal effects." *The Journal of the Acoustical Society of America*, 122(3), 2007, pp. 1787-1797.

[2] B. Liang, G. Fazekas, and M. Sandler, "Detection of piano pedalling techniques on the sustain pedal." in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[3] A. Srinivasamurthy, A. Holzapfel, and X. Serra. "Informed automatic meter analysis of music recordings." in *Proceedings of the 18th International Society for Music Information Retrieval (ISMIR) Conference*, 2017.

[4] J. B. Smith, and E. Chew, "Automatic interpretation of music structure analyses: a validated technique for post-hoc estimation of the rationale for an annotation." in *Proceedings of the 18th International Society for Music Information Retrieval (ISMIR) Conference*, 2017.

The Social Character of Metadata in 'In the Box' Music Production

Glenn McGarry

Mixed Reality Laboratory, Department of Computer Science, University of Nottingham, UK,
glenn.mcgarry@nottingham.ac.uk

***Abstract*—Digital technologies have vastly broadened possibilities for music creation in and beyond traditional sites of music production, from audio capture, processing and effects; to new performance interfaces, and possibilities for remote collaboration. Further potential lies in musical metadata, which is progressively being exploited in novel applications to create new value prospects in music production and consumption. The studies presented here build upon ongoing research that aims to understand the *social* character of metadata in contemporary music production, to inform these developments further.**

I. BACKGROUND

In the contemporary musical landscape the ease of access to digital marketplaces coupled with affordable music production tools, such as the Digital Audio Workstation (DAW), has enabled a broader range of people to engage in producing and distributing music [1]. Amateurs and professionals alike can make music in ever diverse locations and the concept of a 'studio' now includes almost any space that can be configured with a DAW and other musical equipment. With less budget, and fewer resources and formal process management available in such production sites, members' methods of bringing about order and organisation to the work of making music is an area of interest in this research.

II. PRODUCTION METADATA

Another focus of the research is the social role and character of metadata in music making practice. Metadata is commonly defined as 'data about data', but more correctly applies to *organisation* of that descriptive data in standards, formats, and conventions [4]. Although aimed toward technical concepts of metadata, these definitions are also useful in identifying the formal and informal *social* uses of metadata in the field data, i.e. human readable description and organisation of music objects used in the accomplishment of music production work.

III. THE STUDIES

The studies presented here build on previous empirical research of contemporary music making practices. [2,3]. The **first study** observes a professional producer working in a home studio alone to create music only using a DAW. The process includes several iterative activities including: setting up the DAW to prioritise musical elements in the mix e.g. grouping drums and bass before everything else; using tools to create and apply tempo and key signature metadata to sound files; drawing on these metadata to search and browse sound files; layering sound files and Virtual Instruments (VI) in the DAW to create musical parts; and exporting the musical parts to replace component layers and streamline the composition. Metadata from file names and VI presets propagate unpredictably through the process as DAW track names, which also are not always meaningful. However, this is of little concern as the producer prefers to adapt the display order of layers and organise sound at the group level to coordinate the process locally.

Not attending to these metadata however, creates problems where musical data is exchanged between people and locations. In the **second study** a producer in a project studio receives demo compositions from a collaborating artist in which, like the first study, tracks are unlabeled, but also containing unorganized sound and by-products of layering retained. The producer then must apply significant effort to unpick and understand the composition before doing the intended work of mixing the song. This further highlights the essential role of metadata in *coordinating* music-making, as concluded in previous work [3].

IV. CONCLUSION

In this abstract, contemporary practices of 'in the box' computer-based music making beyond the recording studio are briefly explored. Uniquely, the role of metadata in the organisation, coordination and exchange of music data in these practices are also accounted for to identify metadata troubles and inform system design.

ACKNOWLEDGMENT

I would like to acknowledge the grant Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption - EP/L019981/1

REFERENCES

[1] Michaela Hoare, Steve Benford, Rachel Jones, and Natasa Milic-Frayling. 2014. Coming in from the Margins : Amateur Musicians in the Online Age. *Chi 2014*, February: 1295–1304. https://doi.org/10.1145/2556288.2557298

[2] Glenn McGarry. 2016. Understanding the Social Character of Metadata in Music Production. In *DMRN+11: Digital Music Research Network Workshop Proccedings 2016, Queen Mary University of London*. Retrieved from http://dx.doi.org/10.26494/DMRN.2016.19345

[3] Glenn McGarry, Peter Tolmie, Steve Benford, Chris Greenhalgh, and Alan Chamberlain. 2017. "They're all going out to something weird." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 995–1008. https://doi.org/10.1145/2998181.2998325

[4] Richard Wright. 2004. CORE METADATA : BACKGROUND AND SIGNIFICANCE. In *AES 25th International Conference, London, United Kingdom, 2004 June 17–19*, 1–5.

Hearing the Humanities: Sonifying Steele's Shakespeare

Iain Emsley[1*], Alan Chamberlain[2] and David De Roure[1]

[1*]Oxford e-Research Centre, Oxford University, UK, iain.emsley@oerc.ox.ac.uk
[2]Department of Computer Science, University of Nottingham, UK

***Abstract*— We present initial work that explores the use of sonification to represent Joshua Steele's symbolic notation. This provides a manner of overhearing a previous performance and testing the method's reproducibility and uncertainties within it.**

I. INTRODUCTION

Joshua Steele [1] used a symbolic notation to mark-up performances of Shakespeare. His methodology is briefly discussed before presenting an example of a Web Audio sonification achieved using a fragment of Steele's work before discussing issues arising from such a historical notation and reproducibility.

II. JOSHUA STEELE

An active member and elected a Fellow of the Royal Society of the Arts, Joshua Steele (c1700-1796) wrote the "*Essay towards establishing the melody and measure of speech*" in 1775 [1], revised in 1779 [2]. The books provide a record of his interpretation of the performance in symbolic form using musical notation augmented with extra symbols. Using melody, the accent or force, and rhythmus, the quantity and emphasis, Steele also captured the volume and the way the voice slides upwards or downwards, as can be seen in Fig. 1.

III. METHODOLOGY

A model was created for the Web Audio sonification. The bass clef, which we assume is linked to the actor's perceived pitch, creates a frame map the notes to MIDI numbers by hand. Indicators for related time and loudness are stored with the notes. The start and endpoints for sliding pitches, shown by the accents, defined the way in which the pitches given slide when spoken. The MIDI note was converted into frequencies in the creation of a JSON file. Web Audio is used to present the data as sound. The bars and notes provide clues as to the timing of the syllables and words but no 'actual' time scheme is provided. [3] represented the notes as MIDI in the ELVIS project as a way of analyzing the notation. This project uses sonification to provide an impression of a performance of the text, viewing the text as notes that could be used. Although we cannot reproduce the performance, a glimpse of it is achieved that may support other approaches, such as theatre studies. Volume and time buttons allow for two of the uncertain variables to be altered.

*This research was supported through the following EPSRC project: Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption (EP/L019981/1). With special thanks to Pip Willcox.

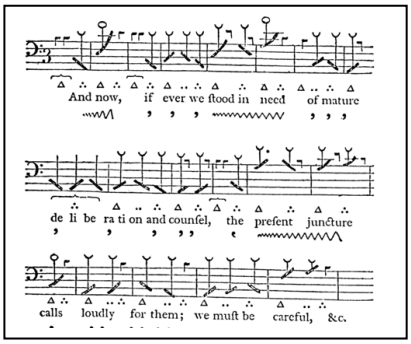

Figure 1. Example Joshua Steele's symbolic notation.

IV. DISCUSSION

Future work will address the challenges of timing, as relating to the assumptions made and Steele's solution of using various times.

V. CONCLUSION

This experimental Digital Humanities [4][5] work demonstrates that the notation might be used to understand the text from the dominant linguistic approach and shows that sonification could be a valid Digital Humanities technique.

REFERENCES

[1] Joshua Steele. *An Essay Towards Establishing the Melody and Measure of Speech to be Expressed and Perpetuated by Certain Symbols.* 1775

[2] Joshua Steele. *Prosodia Rationalis*, J. Nichols: and sold by T. Payne and Son; B. White; and H. Payne. 1779

[3] R. M. Winters and J. E. Cumming, "Sonification symbolic music in the ELVIS project," in *Proceedings of the 20th International Conference on Auditory Display*, New York, NY, June 2014.

[4] Alan Chamberlain, et al. (2017) "Audio Technology and Mobile Human Computer Interaction: From Space and Place, to Social Media, Music, Composition and Creation", In the *International Journal of Mobile Human Computer Interaction* (IJMHCI) V9/4, pp. 25 - 40

[5] David De Roure, et al. (2016) Experimental Digital Humanities: Creative interventions in algorithmic composition on a hypothetical mechanical computer. In: DMRN+11: Digital Music Research Network Proceedings, Queen Mary University of London, Dec. 2016. http://dx.doi.org/10.26494/DMRN.2016.19345

A Deeper Look at the 2017 ASV Spoof Challenge

Bhusan Chettri and Bob L. Sturm

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

***Abstract*— Replaying pre-recorded speech of an enrolled speaker is the simplest spoofing approach to bypass an automatic speaker verification system. The 2017 ASVspoof challenge focused on "replay attack". In this paper, we describe our post-evaluation work. From our analysis of Gaussian Mixture Model (GMM) anti-spoofing systems we find how class-dependent cues in the dataset can significantly affect the results. We show how we can fool the system predictions using such cues. For example, the equal error rate (EER) of Spectral Centroid Magnitude Coefficient (SCMC) system dramatically rises to 44.4 from 14.82 on the evaluation data after we add the "signature" to the test files. This leads us to propose a means of mitigating this problem and improving the internal validity of the dataset.**

I. INTRODUCTION

The 2017 ASVspoof challenge focuses on text-dependent replay attack detection "in the wild" with varying acoustic conditions [1]. Given a recorded speech utterance *s*, the main goal of the challenge is to build an anti-spoofing system that determines if *s* is genuine speech. Fig. 1 illustrates the difference between genuine and replayed speech. The text-dependent ASVspoof 2017 database is based on RedDots corpus and its replayed version, RedDots-Replayed. The latter was created by replaying RedDots signals through various recording and replay devices. The database is divided into training, development and evaluation subsets as described in [1]. Performance is measured using equal error rate (EER) on the evaluation dataset, which is an operating point in detection error tradeoff (DET) curve where false acceptance and miss rejection rate are equal.

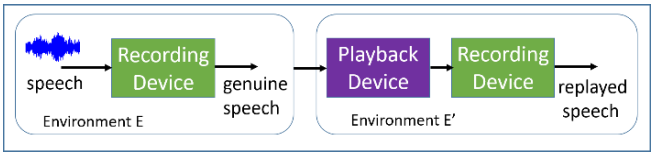

Figure 1. Difference between a genuine and a replayed speech.

II. SYSTEM DESCRIPTION

We use 40 dimensional dynamic coefficients based on inverted-mel frequency cepstral coefficients (IMFCC), spectral centroid magnitude coefficients (SCMC) and constant q-cepstral coefficient (CQCC) to train frame-based GMM systems. No normalization and voice-activity-detection is performed. The results of our anti-spoofing system along with baseline and the best performing system on evaluation data is shown in Table 1.

TABLE I. PERFORMANCE (EER %) ON EVALUATION DATA. ALL THE SYSTEMS EXCEPT [2] ARE TRAINED ON POOLED (TRAIN+DEV) DATA.

Baseline	Top1 [2]	Best-single [2]	IMFCC	SCMC	CQCC
24.6	6.73	7.34	17.4	14.8	17.7
Intervention-I.			34.4	44.4	18.7
Retraining + Intervention-I.			19.1	17.7	19.3

The baseline system provided by organizers is a GMM system trained on 90 dimensional CQCC features. Top1, the best ranking system, use score level fusion of three sub-systems. The best-single system is a GMM based sub-system of Top1 trained on 32 dimensional deep CNN features. They use normalized log power magnitude spectrogram to train the CNN. The challenge results are summarized in [1].

III. DISCUSSION AND CONCLUSION

In our work [3], SCMC-feature based GMM system show the best performance on evaluation data. Our deeper analysis of this system shows the presence of recording artefacts in some genuine examples that is missing from the replayed version. As a consequence, spoofed models assign a very low likelihood, which in-turn influence the classifier prediction. We conduct Intervention-I experiment to prove this hypothesis. We take the first 60ms audio samples from the most confidently and correctly classified genuine audio file in the development set and append it across all the test files and evaluate the system performance. We see a dramatic rise in the EER for IMFCC and SCMC while CQCC show a robust performance. We propose a very simple approach to mitigate against such manipulation attack using intervention. Here, we remove the first 60ms samples from all the genuine audio in the training set and re-train the genuine GMM model. We run Intervention-I experiment again and observe that the system has now become quite robust and resilient against such manipulation.

REFERENCES

[1] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K.A. Lee, *"The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection*," in Proc. Interspeech, 2017, pp. 2-6.

[2] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev and V. Shchemelinin, *"Audio replay attack detection with deep learning frameworks*," in Proc. Interspeech, 2017, pp. 82-86.

[3] B. Chettri and B. Sturm, "*A Deeper Look at Gaussian Mixture Model Based Anti-Spoofing System*," submitted in ICASSP 2018.

Towards Performing a Personal Interactive Musical Soundtrack

Laurence Cliffe

Horizon Centre for Doctoral Training, School of Computer Sciences, University of Nottingham, UK,
laurence.cliffe@nottingham.ac.uk

Abstract— This paper presents on-going research into new methods and understanding to support listener's in self-generating and performing their own interactive, personal musical soundtracks.

I. INTRODUCTION

The availability of locative and biophysical sensors and the personal data they capture offer new opportunities for personal soundtrack generation. Existing approaches to generative and interactive musical composition often rely on pre-composed segments of music [1], and can be plagued with issues regarding musical mechanization, repetitiveness and a lack of creative integrity [1]. This paper presents on-going research into new methods and understanding to support listener's in self-generating and performing their own interactive, personal musical soundtracks. It will attempt to detail how an amalgam of different approaches may present an opportunity for realising meaningful and personal musical soundtracks with creative integrity, expression, and location and listener sensitive generative content for the aural augmentation of daily activities. These approaches include the sonification of personal biophysical and locative data [4, 5] seamful system design [2], and the application of contemporary avant-garde compositional approaches [3] that embrace the inclusion of ambient artifacts, chance and indeterminacy within musical compositions. It also identifies currently available solutions in the area of generative sound tracking and their limitations, such as Weavrun[1], along with the possible applications, societal benefits and possible future research into such an approach.

II. LETTING THE OUTSIDE IN

This paper outlines a seamful approach to location-based musical generation and an open approach to the inclusion of external ambience within the generative composition, as proposed by Cage [3], letting the outside in. This is suggested as a means of complimenting a data-driven approach in an attempt to generate a more spontaneous, personal and location-sensitive composition. This approach could go beyond the appropriation and exploitation of identified seams within a system [2], by utilising a system that creates seams by design, or can dynamically control the size of seams and their impact on the generative composition at given points.

This approach could also further address issues outlined by Berndt et al. [1], who recognise that there is 'an existential danger' involved in the generation of music for interactive purposes which stems from the unknown length of specific scenarios, which leads to soundtracks with repetitive and mechanical characteristics and a loss of musical integrity. Berndt et al. [1] identifies various compositional approaches which are used to tackle these issues, including structural diffusion, sequential variance, polyphonic variance, orchestrational variance and reharmonisation. An examination and evaluation of the open and seamful generative compositional approach outlined here could lead to a potential addition to the interactive music composer's toolkit.

III. FUTURE RESEARCH

Future research includes an examination into the possible curatorial role of the user, and the role that musicians may play within this curatorial role, as a means of specifying genre and instrumentation. Weavrun, a real-time tempo adaptive mobile music application, presents an interesting concept regarding this, where musicians can create adaptive, tempo variant tracks using the WeavMixer[2] software for specific use in the Weavrun application.

ACKNOWLEDGMENT

Laurence Cliffe is supported by the Horizon Centre for Doctoral Training at the University of Nottingham (RCUK Grant No. EP/L015463/1) and the grant Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption - EP/L019981/1

REFERENCES

[1] Berndt, A., Dachselt, R., & Groh, R. 2012. A survey of variation techniques for repetitive games music. *Proceedings of the 7th Audio Mostly Conference*, pp. 61-67.

[2] Broll, G., & Benford, S. 2005. Seamful design for location-based mobile games. *Entertainment Computing - Icec 2005, 3711*, pp. 155-166.

[3] Cage, John. "Silence : lectures and writings." Middletown, Conn. : Wesleyan University Press, 1961. pp. 7-8

[4] Chaparro, I., & Duenas, R. 2015. Psychogeographical Sound-drift. *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pp. 187-188.

[5] Chen, S., Bowers, J., & Durrant, A. 2015. 'Ambient walk': A mobile application for mindful walking with sonification of biophysical data. *Proceedings of the 2015 British HCI Conference*, p. 315.

[1] http://run.weav.io
[2] http://www.weav.io/#/how-it-works.