DMRN+13: Digital Music Research Network

One-day Workshop 2018

Arts Two Lecture Theatre

Queen Mary University of London

Tuesday 18th December 2018

Chairs

Panos Kudumakis and Simon Dixon

Programme

10:30	Registration opens Tea/Coffee
11:00	Welcome and opening remarks Prof. Simon Dixon (Centre for Digital Music, Queen Mary University of London)
11:10	KEYNOTE “MIR: From Signal Processing to Deep Learning, Learning from Data: Which Data?” by Prof. Geoffroy Peeters (Télécom ParisTech)
12:00	Poster spotlights (5 mins each)
13:00	Buffet Lunch, Networking
14:00	Posters will be on display
15:20	Tea/Coffee
15:40	“The Effect of Simulated Acoustics on Musical Expression”, James Weaver, Mathieu Barthet and Elaine Chew (Queen Mary University of London)
16:00	“Music and Haptic Representation on Mobile Auditory Graph”, Zico P. Putra (Queen Mary University of London)
16:20	“Deep Learning for Audio Event Detection and Tagging on Low-Resource Datasets”, Veronica Morfi and Dan Stowell (Queen Mary University of London)
16:40	“The Audio Commons Initiative”, Alessia Milo, Mathieu Barthet and György Fazekas (Queen Mary University of London)
17:00	Close*

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

Posters

1	“Data-driven Quality Prediction for Digitally Restored Audio Archives”, Alessandro Ragano (University College Dublin), Emmanouil Benetos (Queen Mary University of London) and Andrew Hines (University College Dublin)
2	“Grammar Informed Sound Effect Retrieval for Soundscape Generation”, Emmanouil Theofanis Chourdakis and Joshua D. Reiss (Queen Mary University of London)
3	“A Multi-modal Approach for Learning from Singing”, Helen L. Bear (Technical University Munich), Daniel Stoller, Yukun Li and Emir Demirel (Queen Mary University of London), Wenming Gui (Jinling Institute of Technology, China), Emmanouil Benetos, and Simon Dixon (Queen Mary University of London)
4	“Computation and Visualization of the Differences Between Two Scores”, Francesco Foscarin, Raphael Fournier-S'Niehotta (CNAM, Paris) and Florent Jaquemard (INRIA, Paris)
5	“Eyes-free Music Browsing in a Binaural Auditory Environment”, Rishi Shukla, Rebecca Stewart and Mark Sandler (Queen Mary University of London)
6	“Towards Richer Online Music Public-domain Archives (TROMPA)”, Emilia Gomez (Universidad Pompeu Fabra), Cynthia Liem (Technische Universiteit Delft) and Tim Crawford (Goldsmiths University of London)
7	“Using Triplet Network for the Intelligent Control of Audio Effects”, Di Sheng and György Fazekas (Queen Mary University of London)
8	“Digital Music Objects: Composing on the SOFA”, David De Roure, Graham Klyne, John Pybus, David M. Weigl, Matthew Wilcoxson and Kevin Page (Oxford University)
9	“An Investigation into Automatically Generated Auditory Route Overviews”, Nida Aziz, Rebecca Stewart and Tony Stockman (Queen Mary University of London)
10	“Introducing JADE, a New Digital Instrument”, Sarah Sauvé, Luke Welsh and Andrew Staniland (Memorial University of Newfoundland, Canada)
11	“Characterising Glissando and Flutter-tongue Techniques in Recordings of Chinese Bamboo Flute”, Changhong Wang, Emmanouil Benetos and Elaine Chew (Queen Mary University of London)
12	“Musical Chills: Stimulus Properties, Stylistic Preference & Familiarity”, Rémi de Fleurian (Queen Mary University of London) and Marcus T. Pearce (Aarhus University & Queen Mary University of London)
13	“Tonic and Dastgāh Recognition in Iranian Music”, Peyman Heydarian (Apulse, UK)

The Effect of Simulated Acoustics on Musical Expression

James Weaver^{1*}, Mathieu Barthet¹, and Elaine Chew¹

¹Centre for Digital Music, Queen Mary University of London, UK, j.d.weaver@qmul.ac.uk

Abstract— Quantitative analysis of the impact of reverberation on musical performance is a relatively unexplored area. This paper presents a ‘proof of concept’ methodology looking at global tempo changes for a collocated piano duo and finds an association between tempo and reverberation time.

I. INTRODUCTION

Performing musicians experience varying acoustic conditions which impact their performance. Existing research by Kato, Ueno et al. [1]–[3] investigated a range of objective parameters with soloists playing in simulated acoustics but no work to date has analysed multiple performers working in the same physical space with simulated acoustic environments. This paper outlines a methodology to do this with two digital pianos and convolution reverberation utilising Room Impulse Responses (RIR) with particular analysis on tempo.

II. METHODOLOGY

This study used two piano players playing in convolution reverberation conditions based on RIR.

Two experienced pianists were recruited from the Centre for Digital Music, Queen Mary University of London. The stimulus was Bars 0-33 of *Mozart’s Sonata for 2 Pianos in D Major, K.448 / 375a* as it provided variation and detail for analysis of tempo changes. Participants undertook rehearsal sessions allowing discussion and exploration of the stimulus.

In order to make the study replicable, easily available commercial equipment was sourced. Two Digital Pianos—were used and recordings were made using the Ableton 9 (Suite) Digital Audio Workstation. This DAW was chosen due to the freely available Convolution Pro Reverb Device developed in MAX/MSP. RIR were selected from OpenAir Library. The musicians played the stimulus 40 times across 10 randomised conditions with instructions to play expressively utilising the acoustics. Two methods were used to extract tempo data from the recordings. The first involved manually beat tapping the audio using Sonic Visualiser and computing the tempo information using MATLAB. The second method used the MiningSuite Toolbox to directly analyse the audio file.

III. RESULTS

Across both methods of analysis global tempo in conditions with short (~0.1s Reverberation Time (RT)) or long reverberation (10.22s RT) were overall slower than those of 1-2 seconds where performances were not only faster overall, but also showed more variation beat to beat. This can be explained by the longer RT ‘smearing’ note definition, particularly in faster passages, requiring the performers to play slower to maintain intelligibility. Conversely, very short RT leaves the performer unable to utilise the ‘comfort reverb’ of 1-2s RT which supports their playing. Full analysis of this is available in Weaver, Barthet, Chew [4].

IV. CONCLUSION

The results of this study are in line with that of the previous studies conducted on solo performers and provide a basis for further work which will expand the methodology to look at other areas of expression including: loudness dynamics; notions of ensemble, or togetherness, at boundaries and transitions in the score; and comparison of real acoustic environments with simulations delivered via speaker array.

ACKNOWLEDGMENT

Thank you to Edward Hall and Beici Liang for participating in the pilot that shaped this study.

REFERENCES

- [1] K. Ueno, K. Kato, and K. Kawai, “Effect of room acoustics on musicians’ performance. part I: Experimental investigation with a conceptual model,” *Acta Acust. united with Acust.*, vol. 96, no. 3, pp. 505–515, May 2010.
- [2] K. Kato, K. Ueno, and K. Kawai, “Effect of Room Acoustics on Musicians’ Performance. Part II: Audio Analysis of the Variations in Performed Sound Signals,” *Acta Acust. united with Acust.*, vol. 101, no. 4, pp. 743–759, Jul. 2015.
- [3] K. Kato, K. Ueno, and K. Kawai, “Musicians’ Adjustment of Performance to Room Acoustics, Part III: Understanding the Variations in Musical Expressions,” *Acoust. ’08, Paris*, vol. 123, no. 5, p. 3610, 2008.
- [4] J. Weaver, M. Barthet, and E. Chew, “Analysis of Piano Duo Tempo Changes in Varying Convolution Reverberation Conditions,” in *145th Convention, Audio Engineering Society*, 2018.

* JW is supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1)

Music and Haptic Representation on Mobile Auditory Graph

Zico P. Putra^{1*}

¹Centre for Digital Music, Queen Mary University of London, UK, z.putra@qmul.ac.uk

Abstract— This work addresses the use of music and haptic generation for auditory graph. We control the audio generation process and haptic feedback by varying the frequency of musical notes represents the corresponding Y value and represents the sound of the line graph. We use two interaction modes for audio generation on mobile device by swiping across screen area and single-tapping the graphs. The generated audio feedback allows visually impaired (VI) users to perceive the shape of graphs.

I. INTRODUCTION

While researchers have performed numerous studies to understand the human interpretation of visual graphs in reading, comprehending and interpreting displayed data; visually impaired (VI) users still face many challenges that prevent them from fully benefiting from these graphs. Consequently, it affects their understanding of data visualisation, and in turn reduces their role in collaborative tasks with their sighted peers in both educational and working environments. Although, the audio graphs have been proposed for these VI users for years, most studies of audio graphs have only been carried out on a non-portable device like personal computer (PC) which does not offer specific modality input like haptic interaction. To fulfil these gaps, we have developed Mobile Auditory Graph application (MAG app) with additional modalities to test the effectiveness of using auditory graphs to perform graph reproduction tasks on mobile device.

II. MOBILE AUDITORY GRAPH DESIGN

In the auditory graph design, the question is centered on how to map the dimensions of sound to the displayed data. The main mapping issue includes whether pitches should be increased or decreased in response to changes in the associated data. Auditory graphs can be considered a class of sonified displays which uses sound to display quantitative information. This means that any changes in quantitative data are mapped to changes in one or more dimensions of sound.

As in the visual graphs, the auditory graph characteristics need to be set up properly so that the listener can understand the meaning of data. While, the properties of the visual graphs (i.e., spatial area, colour, trend, and size) are regularly changed, the audio properties in sonification, such as pitch, pan, rate, volume, and timbre may be modified. These properties describe some mappings to the sound attributes such as loudness (identifies with the sound's amplitude), pitch (a feature that relates to the frequency of sound) and timbre (a

characteristic of a sound that identifies it from the various reference of a similar pitch and volume) [1].

In MAG design, the points were mapped to the Y coordinate following positive polarity. Thus, when the Y value increases, the pitch will also increase. The sound parameter, i.e., the frequency, varies during playback of a data set. The sounds were produced by Bb/A# Piano notes with the 58,270 Hz frequency. The interface is a mobile graph application with multimodal input by allowing gesture interaction and haptic feedback. Using the TalkBack screen reader, VI user can access the menus and the primary navigation around the MAG app interface.

The playback mode is designed as a tool to sonify each data point with a musical note within a specific time interval. The user can interact with this mode by physically touching the playback icon. When the user double taps the play button, a series of tones are played representing the points on a line graph, the pitch of each tone being proportional to represent the Y coordinate of the respective point on the graph. The swipe interaction gesture mechanism is implemented to help VI users to imagine the location of the points on the X-Y coordinates. The swipe is a haptic interaction that works on many smartphones with a sense of touch. Comparing to the playback mode, the swipe mechanism for MAG app is designed to let user perceives the data notes by sequencing the Y-data manually, i.e., by touching each note along the x-axis. The vibration feedback is added to let the VI users identify their finger position on the X-axis of the data point on the screen. Therefore, the sense of haptic sensation comes on the X-axis in a vertical line rather than on a specific <X, Y> coordinate.

III. CONCLUSION

MAG test for the graph reproduction tasks from the playback mode were compared against the swipe mode. We found that there is no significant difference in the correlation means obtained from both modes, shown by p-value = 0.07. Both modes allowed the VI users to better identify the shape of data during the graph reproduction tasks.

REFERENCES

- [1] K. V. Nesbitt and S. Barrass, "Finding trading patterns in stock market data," *IEEE Comput. Graph. Appl.*, vol. 24, no. 5, pp. 45–55, 2004.

*Research supported by LPDP (Indonesia Endowment Fund for Education).

Deep Learning for Audio Event Detection and Tagging on Low-Resource Datasets

Veronica Morfi and Dan Stowell

Machine Listening Lab, Centre for Digital Music, Queen Mary University of London, UK
g.v.morfi@qmul.ac.uk

Abstract— In training a deep learning system to perform audio transcription, two practical problems may arise. Firstly, most datasets are weakly labelled, having only a list of events present in each recording without any temporal information for training. Secondly, deep neural networks need a very large amount of labelled training data to achieve good quality performance, yet in practice it is difficult to collect enough samples for most classes of interest. In this work, we propose factorising the final task of audio transcription into multiple intermediate tasks in order to improve the training performance when dealing with this kind of low-resource datasets.

I. INTRODUCTION

In recent decades, there has been an increase in the demand for transcription predictions for a variety of audio recordings instead of just the tags of a recording. Transcription of audio recordings refers to audio event detection and classification, which provides a list of audio events active in a recording along with temporal information about each of them, i.e., starting time and duration for each event.

Annotating data with strong labels, labels that contain temporal information about the events, to train transcription predictors is a time-consuming process involving a lot of manual labour. On the other hand, collecting weakly labelled data takes much less time, since the annotator only has to mark the active sound event classes and not their exact boundaries. We refer to datasets that only have these types of weak labels, may contain rare events and have limited amounts of training data as low-resource datasets.

We propose a factorisation of the final full transcription task into multiple simpler intermediate tasks of audio event detection and audio tagging in order to predict an intermediate transcription that can be used to boost the performance of the full transcription task. For each intermediate task, we propose a training setup to optimise their performance. Finally, we train the intermediate tasks independently and in two multi-task learning settings and compare their results.

II. METHOD

A full audio transcription task can be described as audio event detection followed by event classification. In order to properly train a full transcription network, we need a large amount of data that is not available in a low-resource dataset.

Since it is very hard to train a network to predict full transcription on a low-resource dataset, we factorise the final task of full transcription into intermediate tasks that can predict an intermediate transcription matrix that can later be used to boost the performance of a full transcription network. Fig. 1 depicts the overall task factorisation into the intermediate tasks and how they interact with the final task of full transcription.

We define a WHEN network that performs audio event detection considering all classes as one general class in other words, it predicts when any event is present without taking into consideration the different event classes. We also define a WHO network that performs audio tagging without predicting any temporal information. By combining the two different predictions from these networks, we create an intermediate transcription that provides us with the events present in a recording and the times where any of these events could be present in a recording. This intermediate transcription is to be used as supplementary information when training the full transcription network in order to improve its performance by focusing its attention to the classes present in a recording and the time frames that may contain them.

III. EXPERIMENTS

We use a multi instance learning (MIL) setting and multi-task learning (MTL) in order to train networks to predict the intermediate transcription. We compare the broadly used joint training with our proposed tied weights training and separate training for each individual task.

Figure 1. Factorisation of the full transcription task.

The Audio Commons Initiative

Alessia Milo¹, Mathieu Barthet¹, and György Fazekas¹

¹Centre for Digital Music, Queen Mary University of London, UK, a.milo@qmul.ac.uk

Abstract— Significant amounts of user-generated audio content, such as sound effects, musical samples and music pieces, are uploaded to online repositories and made available under open licenses. Nevertheless, the creative industries are not yet using much this content in media production. A big share of creative commons content remains unreachable primarily because it is not well organised and annotated. In this paper we present the Audio Commons Initiative, which is aimed at promoting the use of open audio content and at developing technologies to support an ecosystem composed of audio content repositories, production tools and users.

I. INTRODUCTION AND MOTIVATIONS

The Audio Commons Initiative aims at delivering an ecosystem which gathers audio content creators, content providers, consumers, data scientists and the music tech industry. Within the project, Creative Commons is applied to audio as a legal framework to foster sharing practices, creating a distributed infrastructure which enables the reuse and remix of creative artefacts. Creative commons licenses provide different level of permissions: attribution (BY), non-derivative (ND), non-commercial (NC), share-alike (SA). CC-licensed content can be repurposed in a more fluid way in music and media production, synchronisation, games, etc.

II. TECHNOLOGY SOLUTION

In order to integrate CC-licenses in the creative network, our group developed technologies and software tools to find CC material, including the AC Ontology, a data model for audio content [1], the AC API integrating some existing content repositories (Jamendo, Freesound, and Europeana), services using audio content analysis to augment metadata from content providers with musical concepts such as keys, mood, and instrumentation [2]. Browser-based and embedded tools (DAW plugins) have also been developed to support CC-audio retrieval, repurposing and music production. Jam with Jamendo [3], a music discovery tool aiming to provide new practice material to music learners is another example of how this material may benefit users [5]. Playsound [4] provides a collaborative space for music and soundscape composition using Creative Commons audio.

The authors are with C4DM at QMUL. Email: {a.milo, m.barthet, g.fazekas}@qmul.ac.uk.

III. EVALUATION AND DISSEMINATION

Promoting a holistic framework of evaluation, we assessed the integration of the prototypes in the music studio through a user study with music producers and composers, aimed at creating a musical piece in one hour. Participants adopted the prototypes in their workflow by handling possible lack of control and experience with the tools through creative strategies based on serendipity, and providing further suggestions to improve frictions in usability.

The project outcomes are widely disseminated through Audiocommons.org including a blog to report updates and related projects. Creative workshops, a hackathon at Abbey Road and the FAST Industry Day were also used to share Audio Commons outcomes with communities.

IV. CONCLUSIONS

From a technological perspective, the project showed to be successful but the API for content provisioning should still be trialed in the wild (i.e. outside tools built by the project consortium). Future challenges include tracking rights in a content reuse scenario, which might require blockchain and appropriate models such as smart contracts. Broader range of content analysis and repurposing tasks might be better supported with deep learning algorithms.

ACKNOWLEDGMENT

The project is funded by EU H2020 Audio Commons grant (No. 688382).

REFERENCES

- [1] Ceriani, M. & Fazekas, G. (2018) Audio Commons ontology: a data model for an audio content ecosystem. In Proc. of the 17th International Semantic Web Conference (ISWC 2018).
- [2] Pauwels, J., Fazekas, G., & Sandler, M. B. (2018). Recommending songs to music learners based on chord content. In Proc. of the 2018 Joint Workshop on Machine Learning for Music.
- [3] Pauwels, J., Xambó, A., Roma, G., Barthet, M., & Fazekas, G. (2018). Exploring Real-time Visualisations to Support Chord Learning with a Large Music Collection. In Proc. of the Web Audio Conference.
- [4] Stolfi, A., Milo, A., Ceriani M., & Barthet, M. (2018b). Participatory musical improvisations with Playsound.space. In Proc. of the Web Audio Conference.
- [5] Xambó, A., Pauwels, J., Roma, G., Barthet, M., & Fazekas, G. (2018). Jam with jamendo: Querying a large music collection by chords from a learner's perspective. In Proceedings of the 13th International Audio Mostly Conference.

Data-driven Quality Prediction for Digitally Restored Audio Archives

Alessandro Ragano¹, Emmanouil Benetos² and Andrew Hines¹

¹Insight Centre for Data Analytics, University College Dublin, Ireland, alessandro.ragano@ucdconnect.ie

²Centre for Digital Music, Queen Mary University of London, UK

Abstract— Restoring the fidelity of music recordings in audio archives is crucial in order to sustain our musical heritage with a good quality of experience (QoE). We propose to investigate and understand the QoE aspects of audio archives and to develop data-driven approaches for QoE estimation and audio restoration based on deep learning.

I. INTRODUCTION

Digital audio restoration is employed for both preserving the original conditions of the recorded material, that meets the needs of audio archivists and for enhancing the audio quality which aims to improve the audience satisfaction. Even though digital audio restoration has been widely explored from a statistical perspective, tasks such as impulsive disturbances detection and audio inpainting are still far from being solved. In order to advance the current state of the art we propose to develop a new QoE-driven audio enhancement approach based on deep learning.

II. RELATED WORK

Assessment of audio restoration techniques has mainly relied on listening tests and objective audio quality metrics. The former are time consuming and expensive to carry out while the latter were developed for different purposes. Among the most common objective metrics we found Perceptual Evaluation of Audio Quality (PEAQ) [1], mostly used for low audio impairment, Perceptual Evaluation of Speech Quality (PESQ) [2] developed for speech quality assessment and ViSQOLAudio [7] which outperformed PEAQ. Even though these measures establish the current state of the art in audio quality evaluation, they do not appropriately correlate with the subjective quality of highly impaired audio signals [3]. Deep learning has been recently taken into account for audio enhancement tasks: Kuleshov et al. [4] designed a deep Convolutional Neural Network (CNN) with residual connections while Lim et al. [5] proposed a fusion of a time domain network and a frequency domain network. A CNN which exploits the context information across spectral dependency in music signals is found in [6]. However, none of these models integrated perceptual information when training and assessing deep networks, e.g., using a perceptual cost function or evaluating with an ad-hoc audio quality metric.

III. GOAL AND EXPECTATIONS

We propose to investigate new audio enhancement techniques which utilize both deep learning and QoE, more specifically we are interested in: (1) Understanding relevant audio restoration approaches from a QoE-enhancement perspective. (2) Investigating the appropriateness of the objective measures with respect to the perceived quality (3) Defining an audio restoration ontology and creating a dataset for related research tasks such as audio enhancement. (4) Proposing new models that estimate the perceived quality targeting archive related issues. (5) Proposing new QoE-driven enhancement deep learning-based audio restoration methods.

We believe that adhering to the mentioned points will bring benefits to various fields such as QoE, Machine Listening, Audio Archiving, Cultural Heritage and Digital Humanities.

REFERENCES

- [1] Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., & Colomes, C. (2000). PEAQ-The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2), 3-29.
- [2] Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vol. 2, pp. 749-752).
- [3] Creusere, C. D., Kallakuri, K. D., & Vanam, R. (2008). An objective metric of human subjective audio quality optimized for a wide range of audio fidelities. *IEEE Transactions on audio, speech, and language processing*, 16(1), 129-136.
- [4] Kuleshov, V., Enam, S. Z., & Ermon, S. (2017). Audio Super Resolution using Neural Networks. *International Conference on Learning Representation (ICLR)*.
- [5] Lim, T. Y., Yeh, R. A., Xu, Y., Do, M. N., & Hasegawa-Johnson, M. (2018, April). Time-Frequency Networks for Audio Super-Resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 646-650). IEEE.
- [6] Porov, A., Oh, E., Choo, K., Sung, H., Jeong, J., Osipov, K., & Francois, H. (2018, October). Music Enhancement by a Novel CNN Architecture. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
- [7] Sloan, C., Harte, N., Kelly, D., Kokaram, A. C., & Hines, A. (2017). Objective assessment of perceptual audio quality using ViSQOLAudio. *IEEE Transactions on Broadcasting*, 63(4), 693-705.

Grammar Informed Sound Effect Retrieval for Soundscape Generation

Emmanouil Theofanis Chourdakis and Joshua D. Reiss

Centre for Digital Music, Queen Mary University of London, UK, e.t.chourdakis@qmul.ac.uk

Abstract— This paper introduces a simple method for retrieval of sound effects from a sound library, with the goal of creating soundscapes from a sentence in natural language. Grammatical constituents are extracted from the sentence in order to construct search queries relevant to the objects described in the sentence. A comparison against a previous method of retrieving sound effects for soundscape generation is shown.

I. INTRODUCTION

An essential part of soundscape generation from text is retrieving relevant sound effects. Consider sentence (1): *Crows are feeding on rubbish at a garbage dump*. We expect to retrieve sound effects relevant to crows and the noise they make while feeding, and the sounds of a garbage dump. Current methods that generate soundscapes from text do retrieval by constructing text queries from generic sentence features. Such approaches do not take in consideration the role of the words in a sentence and can lead to undesirable results. For example, when using n-grams (e.g. in [1]) word combinations such as *feeding rubbish* will be used to construct queries. This might lead to retrieve *feeding rubbish to a cat* which we expect to be different from *crows feeding on rubbish*. In this paper we aim to constrain searches to sound effects relevant to the *subject* and *object* of the sentence in order to avoid such problems. Main motivation for this paper is previous work on producing sound from story narrative where we needed methods to convey story text as sound [2].

II. APPROACH

We suggest before constructing queries to use a simple rule-based information extraction approach, as in [3], to do sentence simplification and extract the *Subject* (*S*), *Verb* (*V*), *Object* (*O*), and *Adverbials* (*A*) of the simplified sentences. We then combine those parts to construct queries. The main assumption is that a sentence describes one or more events and each event is an action of something (*subject*) possibly interacting with something else (*object*) which happens somewhere (*prepositional adverbials*). Simplification transforms the original sentence to simpler sentences where each describes a single event. An initial query is constructed by concatenating all the parts while omitting auxiliary verbs and articles for each sentence. Since the relevant sound effects might not be available in the library for the complete

query, we gradually omit those parts that are less important. For example, in (1) the sound of crows feeding on rubbish may not be available but the sound of crows feeding on something else might exist and will be more relevant than e.g. *cats* feeding on rubbish. Finally, if the *adverbial* contains a preposition such as *at*, we assume that it tells us where the action happens (*at a garbage dump, on the road*, etc.) Taking sentence (1) as an example, our method first simplifies it to sentences (2a): *Crows are feeding on rubbish* and (2b): *Crows are feeding at a garbage dump* where $S_{2a,2b} = \text{Crows}$, $V_{2a,2b} = \text{Feeding}$, $A_{2a} = \text{On rubbish}$, and $A_{2b} = \text{At a garbage dump}$. It then constructs a query by concatenating the *S*, *V*, and *A* parts together. Additional queries are constructed by removing *A*, then *V*. Since *A* contains a preposition (*at*) the part that follows the preposition (*garbage dump*) is also added. The queries we construct as well as queries constructed using [1] can be seen in Table 1.

TABLE 1. QUERIES BY PROPOSED APPROACH AND [1] FOR SENTENCE (1)

Proposed Approach	Thorogood et al [1].
Crows feeding on rubbish at garbage dump, Crows feeding on rubbish, Crows feeding at garbage dump, Crows feeding, Crows, Rubbish, Garbage dump	Crows feeding rubbish garbage dump, crows feeding rubbish garbage, feeding rubbish garbage dump, crows feeding rubbish, ..., crows, feeding, rubbish, garbage, dump

III. CONCLUSION

We presented a method that given a source sentence can construct queries to retrieve sound effects relevant to the objects described in that sentence. While this method only works on sentences that have grammatical structure, it can retrieve relevant sound effects for sentences describing events, such as ones found in stories, supplementing our work done in [2] as a radio-play generation system from story text.

REFERENCES

- [1] Thorogood, M., Pasquier, P., Eigenfeldt, A., “Audio Metaphor: Audio information retrieval for SoundScape Composition”, *Sound and Music Computing*, 2012.
- [2] Chourdakis, E.T., Reiss J.D., “From my pen to your ears: Automatic production of radio plays from unstructured story text”, *Sound and Music Computing*, 2018.
- [3] Corro, L., Gemulla, R., “ClausIE: Clause-Based Open Information Extraction”, *Intl. Conf. World Wide Web*, 2013.

A Multi-modal Approach for Learning from Singing

Helen L. Bear^{1,2}, Daniel Stoller¹, Yukun Li¹, Emir Demirel¹, Wenming Gui^{1,3}, Emmanouil Benetos¹, and Simon Dixon¹

¹Queen Mary University of London, UK h.bear@qmul.ac.uk

²Technical University Munich, Germany

³Jinling Institute of Technology, China

Abstract— We present a proposed project with the goal to use visual information with audio vocals to accurately detect lyrics. Lyric recognition is multi-faceted and includes a number of challenges which we plan to address: detecting lyric timing, that is matching on- and off-set timestamps of lyrics to the audio; lyric transcription, that is recognising the lyric being sung at each time point; and vocal separation, that is isolating the singer from accompaniment and other noises.

I. INTRODUCTION

Recognition of lyrics from a singer’s vocals has many applications such as karaoke, subtitling, and redubbing. Recognition includes the classification of lyrics and the detection of start and stop positions (also referred to as onsets and offsets), the latter of which is referred to as lyrics alignment. With the maturity of speech recognition technologies one approach for this task is to apply existing speech systems to music data. However, classification of singing differs from speech classification, to such an extent that this approach failed to achieve robust results. Thus we seek to develop singing specific recognition methods that make use of additional input from music videos by estimating the singer’s lip movements to aid detection of lyrics and timbre.

II. LYRIC RECOGNITION

Lyric recognition is the classification of the lyrics as sung. One reason this is different to speech is because lyrics may include words that one can not locate in a dictionary, e.g. long ‘aaahs’ and ‘ooohs’ of harmonies are common in popular music as are ‘screaches’ in heavy metal, and drum sounds in beat-boxing. In addition, the duration of phonemes follows an entirely different distribution to that found in speech. Lyrics, or words to a song, are singularly a sentence or sequence of words. Lyrics do not necessarily follow a grammatical language model like and any probable syntax of lyrics would vary by genre. This means that lyric transcription accuracy does not benefit from language models to the same effect as speech recognition systems do.

III. ALIGNMENT

In addition to lyric recognition, we also detect when the vocal are active or not. This means first to separate music into segments contain singing or speaking or non-vocal only. Finding the onsets and offsets of lyrics (vocal activity detection) is the first part of aligning transcriptions to music.

The next step, identifying word boundaries, is more difficult and we propose that video input may be beneficial, eventually enabling more accurate the subtitling systems and timing for karaoke machines referenced in Section I.

IV. CHALLENGES

Transcription and alignment are made all the more difficult in multi-lingual situations such as redubbed videos, vocals in a second language for the speaker, and classification of lyrics which change language. To address these we seek to experiment with methods to enable language recognition from video inputs, vocalist first language identification from video and furthermore, for spotting cover songs (to enforce copyright protection).

These tasks require a pre-processing stage that is itself a research challenge. Most music recordings are sold only as a final mixture of the individual instrument tracks. For this task, multi-track sources would be needed to train a separation model to isolate lead vocals from accompanying instruments or backing vocals.

To utilise video data with audio for this work, tasks like face detection and tracking are essential. In one approach we will develop systems for lip-reading singing. We anticipate adaptation challenges going from speech to singing, including having to deal with the additional acting and expression element of singing performance compared to natural speech.

V. DATA

For all tasks we have difficulties in sourcing data that is accompanied by video streams of the singer(s). Some options available to us include:

1. Collecting, recording, and annotating a new dataset from scratch; this requires decisions regarding genre and lyrics selection to avoid biases;
2. Using existing datasets with different ground truth units (solfege, words, phonemes, lyrics);
3. Curating a dataset from music available online. The difficulty here is isolating background accompaniment from the singer, or to obtain aligned lyrics, as ground truth annotation for the extraction/transcription task, respectively;
4. Collaborating with the music industry to gain access to popular music data that has associated video and fine-grained lyrics annotations, or individual instrument tracks.

Computation and Visualization of the Differences Between Two Scores

Francesco Foscarin^{1*}, Raphael Fournier-S'Niehotta¹ and Florent Jacquemard²

¹Vertigo, CNAM, Paris, France, francesco.foscarin@cnam.fr

²INRIA, Paris, France

Abstract— The comparison of music scores is a crucial step for many tasks, such as collaborative score creation, version control systems and evaluation of music transcription. We propose two measures of score difference aiming at comparing the scores at semantic and syntactic level.

I. INTRODUCTION

The Unix *diff* utility is a line-based comparison of two text files, useful in particular for source code files with few differences.

As recently observed in [1] it's not meaningful to apply the Unix *diff* utility to XML scores because the line structure of a XML file does not reflect the its musical structure. They propose a hierarchic *diff* algorithm for XML MEI files based on a tree edit distance applied to xml trees. We follow here a different approach, proposing a comparison on two different objectives. The first level is semantic: we compare the musical content of the score (seen as a sequence of events characterized by their pitches and durations), with techniques similar to melodic similarity [2][3]. The second level is syntactic, *i.e.* the graphical content of the score (beamings, bar positions, dots, ties, etc.). For both cases, we compute a measure of the difference and a list of modifications necessary to transform one score into the other.

II. SEMANTIC DIFFERENCE

For this case we use an intermediate lossy representation of the scores called timeline (similar to MIDI files) that contains the explicit note pitches and temporal positions.

Considering first the monophonic case, we build a sequence of couples (pitch, duration) from each timeline. We then use an edit distance with the usual operations (insert, delete and update) [4] to compute a similarity index and to retrieve the list of transformations of one score to the other (Fig.1).

Figure 1. Example of the differences computation and visualization on two monophonic scores. The edit operations are notated by different figures: square (insert), triangle (update), rhombus (delete).

This difference is feature-based: we can compute it for pitch only, for durations only or for both features, and we can have different cost/weight values for the edit operations, depending on the needings.

We may generalize to a polyphonic score, computing a similarity index between the voices of the two scores and consider the pairs that give the maximum similarity.

III. SYNTACTIC DIFFERENCE

For the syntactic difference we model the graphical representation of the score with notation trees [5], building the trees from each voice in each measure. Considering first the monophonic case, the computation of the differences reduces to the problem of finding the longest common subsequences (LCSs)[4] for the sequences of tree hash of the two scores. This is similar to compute an edit distance with only insertion and deletion operations.

The generalization to a polyphonic score is done similarly to the semantic difference.

IV. CONCLUSIONS

We propose a tool that allow the computation of two kinds of differences between two scores. It is partially implemented in the digital music library NEUMA (<http://neuma.huma-num.fr/>) to be easy to access and use. Our future goals are to upgrade the visualization tool and to perform a finer tree distance computation for the notation trees, that will allow to compute and visualize a fine grain comparison of the measures (*e.g.*, beaming breaks or continuations and different tuplet groupings).

REFERENCES

- [1] Antila, C., & Weaver, G. (2017). A hierarchic diff algorithm for collaborative music document editing. In *Third International Conference on Technologies for Music Notation and Representation (TENOR)*.
- [2] Lemstrom, K. (2000). String matching techniques for music retrieval. PhD thesis, University of Helsinki.
- [3] Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, 24(3), 161-175.
- [4] Hunt, J. W., & MacIlroy, M. D. (1976). *An algorithm for differential file comparison*. Murray Hill: Bell Laboratories.
- [5] Foscarin, F., Fournier-S'Niehotta, R., Jacquemard, F., & Rigaux, P. (2018). Évaluation de la correction rythmique des partitions numérisées. In *Journées d'Informatique Musicale (JIM 2018)*.

*Research supported by ANR France-Quebec MUNIR Project.

“Eyes-free” Music Browsing in a Binaural Auditory Environment

Rishi Shukla*, Rebecca Stewart and Mark Sandler

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

Abstract— We present an interactive, proof-of-concept system that deploys binaural auditory display for music content search and discovery. This “eyes-free” approach to audio content navigation is motivated by developments in mobile smart device technology and the growth of music streaming services.

I. CONTEXT

Much of the leading-edge in consumer mobile device development is currently focused on wearables. Smartwatches released in the last four years, such as the Apple Watch and Samsung Gear series, present a mechanism for remote control of audio playback from smartphones via Bluetooth. On-board WiFi functionality or – in the case of Apple Watch 3 or 4 – 4G networking even provide the capability for direct connectivity to streaming services. Smart headphones, such as Apple AirPods and Bragi Dash Pro, enable speech or gestural interaction to control paired phones or watches. As the world’s first 3G connected headphones, the Vinci 1.5 takes a further step in widening possibilities for “hearables”, providing standalone access to online music with voice activated control.[‡] These options for interacting with on-demand audio services on smartphones, watches or headphones present scenarios where there is in each case (respectively) restricted, little or no visual real estate for providing a graphical user interface. This presents an additional layer of complexity to the problem of interaction design for libraries with near limitless music tracks.

II. MOTIVATION

Spatial auditory display has previously been explored as a means of tackling the dual challenge of shrinking visual interfaces and increasing ubiquity of large-scale music collections. Limitations of earlier approaches have been attributed to two broad considerations. Firstly, a design paradigm based on virtual 2D or 3D mapping of libraries tended to predominate. Although this was frequently combined with a number of inventive solutions to aid user orientation and rapid aural scanning of surrounding tracks, such models were inherently inefficient for navigating large and diverse volumes of content. Furthermore, whilst arranging items according to measures of relative similarity assists one particular mode of browsing, it does not lend itself so well to defined retrieval tasks (i.e. known item search), socially driven discovery, or playlist construction. Secondly, even five years ago, the computational constraints of devices and paucity of

available binaural rendering implementations limited the possibilities and efficacy of 3D auditory display techniques. Implementation of head-tracking and reverberation — both of which improve accuracy of perceived localisation and, therefore, the scope and utility of the auditory display design — were not technically viable [1, 2]. Combined with the commercial and consumer trends outlined above, recent advances in binaural technology provide a clear motivation for revisiting auditory display for music content navigation.

III. “EYES FREE” BROWSING PROTOTYPE

The prototype demonstrates how a 3D head-tracked binaural environment might be deployed to display and enable navigation of music content. The design is implemented using Bela — a commercially-available, open-source, embedded Linux platform for low-latency audio and sensor processing — and the Bosch BNO055 nine degrees of freedom inertial measurement unit sensor. The comparatively restricted processing capability of Bela (a 1GHz ARM Cortex-A8 processor) offers a testbed for software development that is comparable to other low-powered, portable systems — i.e. embedded computing, wearable technology or mainstream mobile platforms where only a defined proportion of CPU resource might be available for audio handling.

The system will be used in future research to explore how binaural auditory display can be best applied to facilitate “eyes-free” music search and discovery. One line of investigation aims at understanding how concurrent playback of spatialised sound sources can be exploited to enable rapid aural browsing of content from search results. This principle has been evaluated previously in the context of instrumental sounds (where five concurrent sources was shown to be viable) but not using multi-part, polyphonic audio tracks [3].

REFERENCES

- [1] R. Stewart & M. Sandler, “An Auditory Display in Playlist Generation” in *IEEE Signal Processing Magazine* 28.4, 2011, pp. 14–23.
- [2] R. Stewart & M. Sandler, “Spatial Auditory Display in Music Search and Browsing Applications” in *J. Audio Eng. Soc.* 60.11, 2012, pp. 936–946.
- [3] G. Lorho, J. Marila & J. Hiipakka, “Feasibility of Multiple Non-Speech Sounds Presentation Using Headphones” in *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, 2001, pp. 32–37.

* Research supported by EPSRC and AHRC under the grant EP/L01632X/1 (Centre for Doctoral Training in Media and Arts Technology).

[‡] See www.apple.com/uk/watch/; www.samsung.com/uk/wearables/smart-watch/; www.apple.com/uk/shop/product/MMEF2ZM/A/airpods; www.bragi.com/thedashpro/; en.vinci.im [accessed 16/11/2018].

Towards Richer Online Music Public-domain Archives (TROMPA)

Emilia Gomez¹, Cynthia Liem² and Tim Crawford^{3*}

¹ Information and Communication Technologies, Universidad Pompeu Fabra, Barcelona, Spain

² Intelligent Systems - Multimedia Computing, Technische Universiteit Delft, The Netherlands

^{3*} Department of Computing, Goldsmiths College, UK, t.crawford@gold.ac.uk

Abstract— We propose to present a DMRN poster describing and illustrating the main features of TROMPA, a Research and Innovation Action project in the Horizon 2020 Topic, Cult-Coop-09-2017. TROMPA stands for Towards Richer Online Music Public-Domain Archives, and continues work done in the earlier PHENICX project, which involved several of the same partners.

Classical music is one of the greatest treasures of Europe's cultural heritage. Although a historical genre, it is continually (re)interpreted and revitalised through musical performance. Today, most of the classical repertoire is in the public domain; massive numbers of scores and recordings are now available in online repositories actively used by scholars and musicians. Technology offers ways to enrich and contextualise this repertoire, so users might better understand and appreciate it. However, this does not happen automatically. Amidst a deluge of data, links across repositories and modalities (e.g. from scores to recordings) still must be made manually, while insights by previous users are not explicitly stored for future users to learn from. It is thus impossible to get comprehensive insight into the full wealth of our musical cultural heritage.

TROMPA changes this by massively enriching and democratising our public-domain musical heritage. For analysing and linking music data at scale, the project will employ and improve state-of-the-art technology. Music-loving citizens will work with the technology, giving feedback on results, and annotating the data according to their personal expertise.

Following an open innovation philosophy, all knowledge derived will be released back to the community in reusable ways. This enables many uses in applications which directly benefit crowd contributors and further audiences.

The principal goals of TROMPA are:

- to enrich public archives of classical music in ways that enhance the public's experience
- to allow and encourage both professionals and music-loving amateurs to actively engage with the music and its performance
- to combine human and machine intelligence, bringing them to bear on real-world use cases serving target communities

Through its human-in-the-loop paradigm, TROMPA will seek an optimal balance between human assessment of meaningful and valuable information, and algorithmic information processing and selection, considering massive amounts of musical cultural heritage.

TROMPA explores five main use cases with various target audiences.

- **Music scholars.** We support and innovate musicology by offering richer digitisation pipelines, better annotation of music material, and user-friendly ways to automatically search and analyse musical data and link related resources across modalities and collections.
- **Content owners.** Content owners, such as orchestras, play an important role in developing commercial exploitation and growing new audiences. We will build a digitisation, annotation and sharing infrastructure for privately-governed archive material from orchestras.
- **Instrument players.** We will offer ways for instrumentalists (professionals, students and amateurs) to explore archives of scores and their performances in ways beyond metadata search, learn about various interpretations, estimate how hard a new piece will be to learn, and inform and improve individual performance based on existing recordings.
- **Choir singers.** We will leverage knowledge of existing performance interpretations and repertoire to offer novel interactive feedback mechanisms surrounding rehearsals. We will employ (multi-lingual) singing voice synthesis mechanisms to automatically generate study and feedback solutions for amateur vocal performance (e.g. virtual choirs with flexible toggling of parts).
- **Music enthusiasts.** We will propose novel and playful interaction mechanisms for musical cultural heritage content aimed at people without formal musical knowledge, but interested in learning more about music. This audience will be recruited at cultural and science festivals.

TROMPA Participants: Universitat Pompeu Fabra (Spain), Technische Universiteit Delft (Netherlands), Goldsmiths College (UK), Universität für Musik und darstellende Kunst Wien (Austria), Video Dock BV (Germany), Peachnote GmbH (Germany), Voctro Labs SL (Netherlands), Stichting Koninklijk Concertgebouworkest (Netherlands), Stichting Centrale Discotheek (Netherlands)

TROMPA Associated Partners: British Library (UK), Escola Superior de Música de Catalunya (Spain), Barcelona Town Hall/Institute of Culture (Spain), Wikidata (Germany), European Choral Association – Europa Cantat (Germany), IMSLP/Petrucci Music Library (USA)

Using Triplet Network for the Intelligent Control of Audio Effects

Di Sheng^{1*} and György Fazekas¹

^{1*}Center for Digital Music, Queen Mary University of London, UK, d.sheng@qmul.ac.uk

Abstract— Audio effects influence different perceptual attributes of sound due to linear and non-linear processing. They are typically applied to fulfil technical or aesthetic goals. Although audio effects are essential and widely used in music production, their use requires expert knowledge amateurs and hobbyists don't necessarily have. To reduce time and labour requirements, we designed an intelligent control system for a specific audio effect: dynamic range compressor (DRC). In previous research, we have established efficient feature sets for each individual DRC parameter. In this research, we are aiming to build a DNN model to extract features that are suitable to predict multiple features simultaneously given a sound example.

I. INTRODUCTION

An audio similarity model is designed to operate in an intelligent control system for the dynamic range compressor (DRC). The objective is to learn a feature embedding that relates to the DRC parameters. To train an efficient model, we use a dataset of input audio files, the corresponding compressed audio and the known compression parameters: $\theta = \{\tau_a, \tau_r, \text{Ratio}, \text{Thd}\}$.

Our problem focuses on detecting the minor differences between data from the same audio source (e.g. guitar loop) compressed with different settings, including no compression. Under this situation, a triplet net commonly applied in computer vision [1] is suitable. We designed a model with the structure shown in Figure 1. In this research, we use mono-instrument loops from AppleLoop. The anchor is the original audio; the positive sample is a heavily compressed audio file. For the negative sample, we consider to use a light compressed audio file of the same source. This way, the model is assumed to learn features that characterise "heavy compression".

Figure 1. System Overview

We formulate a deep triplet ranking model with a ranking loss. The loss is defined using the max-margin framework.

For a given triplet $t = (s, p^+, p^-)$, where s : anchor, p^+ : positive sample, p^- : negative sample, its loss is defined as:

$$L_\theta(t) = \max(0, \Delta + D(f_\theta(s), f_\theta(p^+)) - D(f_\theta(s), f_\theta(p^-)))$$

II. EXPERIMENTS AND RESULTS

In our experiment, all the audio loops are compressed with two settings: 1) *threshold 15dB, ratio 2*; 2) *threshold 30dB, ratio 15*. The former is considered the negative sample, the latter is the positive sample and the uncompressed files are the anchors. The input is transformed with a window size of 512 samples and MFCC features are computed with 40 coefficients. The model hyperparameters are: batch size of 8, learning rate is 0.001, kernel regulariser is L2 regulariser with 0.0005 weight decay. We used 1632 drum loops as training data and generated 1632 triplets. 90% is used as training data, 10% validation data, check point on validation error rate. To simplified the experiment, we feed the feature embedding to an SVM classifier.

The initial design is $K1 = 2, K2 = 1, N1 = \{25, 20\}, M2 = 5, N2 = \{1000\}$. The last dense layer with size of 1000 is used as feature extractor. The 3-way classification accuracy is 44%. Since the initial design is a rather shallow model. It is reasonable that it is not able to learn an efficient feature embedding. We increase $K1, K2$ gradually and when $K1 = 3, N1 = \{20, 15, 10\}$, and $K2 = 2, N2 = \{1000, 200\}$, the 3-way classification accuracy reaches 99%. We then change the classifier to a regressor, and used the features to predict compression parameters. The results did not provide a better accuracy than using hand-picked features. For example, the model may produce a threshold prediction error rate of 1.45dB, while the hand-pick features provide error at 0.903dB.

III. CONCLUSION

In this experiment, the input audio is compressed with only 2 different parameter settings. We also restricted the model to help solving a classification problem. While good results were achieved to this end, the learnt features are not yet efficient in predicting a full range of DRC parameter settings. We have to consider a more suitable input representation, and tune the model to a regression problem.

REFERENCES

- [1] Yu, Qian, et al. "Sketch me that shoe." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Digital Music Objects: Composing on the SOFA

David De Roure, Graham Klyne, John Pybus, David M. Weigl, Matthew Wilcoxson and Kevin Page

*Oxford e-Research Centre, University of Oxford, UK, david.deroure@oerc.ox.ac.uk

Abstract— We describe a semantic music system which illustrates the assembly of a music composition using semantically annotated music fragments. The system, which we call SOFA (SOFA Ontological Fragment Assembler), demonstrates the concept of Digital Musical Objects (DMOs), and in particular DMO processing and recomposition. The prototype builds upon two existing tools developed by the authors: Music Encoding and Linked Data (MELD), which augments and extends MEI structures with semantic Web Annotations capable of addressing musically meaningful score sections; and Numbers Into Notes, an algorithmic composition tool that acts as a ‘Semantic Signal Generator’ (SSG) to drive the tool chain.

I. SOFA

The SOFA Ontological Fragment Assembler (SOFA)[1] is an application which takes a set of musical fragments produced by the Semantic Signal Generator, along with annotations informed by the semantic context and expressed as Linked Data; it provides a remixer interface offering a guided environment for the user to experiment with assembling the fragments to create a new musical composition. In order to provide a source of fragments for the SOFA interface, we use the existing Numbers Into Notes software[2] which exports fragments in multiple representations, including MIDI, W3C PROV, and musical score encoded as MEI, using it as a form of Semantic Signal Generator, providing source material to the rest of the system.

II. MUSIC FLOWS OF DMO

These two components define the start and end of a music flow; at mid-points in the flow each fragment is associated with additional semantic context, realised through annotations, which encode the properties by which remixing in the SOFA can proceed.

For the SOFA scenario, annotations must be added within a music flow to indicate a fragment’s key, length, and the instruments for which the fragment is within performable range.

Annotation agents analyse the fragments within these specific semantic contexts – e.g. key, instrumentation – and indicate compatibility through the addition of annotations. At a later stage in the flow, matching agents use the semantic annotations to indicate groups of compatible fragments which

are candidates for remixing e.g. of the same or related key, playable on the same instrument.

Figure 1. SOFA music flows

There is a pipeline of increasingly enriched fragments moving from the SSG to the SOFA remixer – a music flow of fragments, gathering semantic context according to the annotation and matching agents. At the point an agent interacts with one, we conceptualise a fragment and its associated scoped semantic context (drawn from the Web of Linked Data) as a DMO, with a series of related and derivative DMOs the artefacts of the music flow (illustrated in the lower half of Figure 1).

These DMOs express accumulating annotations which provide the semantic context to the marked up musical content. They accord with the MELD framework[3], being implemented as web annotations stored in LDP containers.

We can thus consider our architecture at a high level through two lenses: as a music flow with fragments moving from agent to agent towards a particular goal; and as a Web of DMOs which are individually crystallised into existence when an agent requires a semantically scoped fragment upon which to operate.

REFERENCES

- [1] D. De Roure, G. Klyne, J. Pybus, D.M. Weigl, and K. Page, Music SOFA: An Architecture for semantically informed recomposition of Digital Music Objects. SAAM’18, in conjunction with ISWC 2018.
- [2] D. De Roure and P. Willcox. 2017. Experimental Humanities: An Adventure with Lovelace and Babbage. In 13th International Conference on e-Science. IEEE, 194–201. DOI:<http://dx.doi.org/10.1109/eScience.2017.32>
- [3] Weigl, David M., and Kevin R. Page. “A Framework for Distributed Semantic Annotation of Musical Score: ‘Take It to the Bridge!’.” ISMIR, 2017, pp. 221–228.

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

An Investigation into Automatically Generated Auditory Route Overviews

Nida Aziz, Rebecca Stewart and Tony Stockman

School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
n.aziz@qmul.ac.uk

Abstract— This project explores effective methods of generating auditory route overviews for providing pre-navigation route information to blind and visually impaired individuals (BVI). It also looks into the usefulness and usability of auditory overviews for the general population. An auditory route overview provides a summary of route information in audio format. This includes functional route information, such as distance and directions, as well as landscape information, e.g. schools or parks, etc. The display will be interactive, providing the users with the options of controlling various aspects of it, such as volume, speed and level of detail.

I. INTRODUCTION

This project explores the usefulness and usability of auditory route overviews for unseen navigation. This would include both blind and visually impaired (BVI) pedestrians, as well as cyclists and others who may wish preview routes eyes-free before navigating them. The system is designed to be used for planning a journey, as a pre-navigational tool. Even though there are now quite a number of apps widely available to assist with independent, unseen travel, there remain gaps in provision, and the apps that do exist have seen variable levels of take-up within the BVI community. This shows a lack of input to the design process from the stake holders. Therefore, this project will focus on exploring ways to have a more participatory design approach for increasing usefulness of the system. The system will provide functional route information such as distance and directions as an auditory display (AD). The display will comprise a combination of auditory cues: earcons and auditory icons as well as speech. The display will be interactive, providing the users with the options of controlling various aspects of it, such as volume, speed and level of detail. Research will be done on determining suitable ways to incorporate customisability to this auditory interface, and the effects of adding it to the usability and usefulness of the system.

II. INITIAL STUDIES

For the initial study, we asked some audio and design experts to design auditory overviews of routes. These designs were then evaluated in the next study by 8 visually impaired end-users. Analysis of these studies gave an insight into the perspectives of the designers and the end-users highlighting the similarities and contrasts in them. Several design

guidelines were extracted from this procedure. A summary of these guidelines is given in the next section as charts.

III. DESIGN REQUIREMENTS

Figure 1. Components of an AD for route overview

Figure 2. Subcomponents of the AD: components of the integrated audio

Figure 3. Customizable components of the entire AD as well as its individual stems

IV. FUTURE PLANS

The project aims to provide an interactive and customizable auditory route overview to BVI individuals. This system would be zoomable to increase or decrease the amount of information as required by the user.

REFERENCES

- [1] Loeliger, Esther & Stockman, Tony. (2013). Wayfinding without Visual Cues: Evaluation of an Interactive Audio Map System. *Interacting with Computers*. 26. 403-416. 10.1093/iwc/iwt042.
- [2] Csapó, Á., Wersényi, G., Nagy, H. et al. *J Multimodal User Interfaces* (2015) 9: 275.
- [3] Heuten, Wilko & Henze, Niels & Boll, Susanne. (2007). Interactive exploration of city maps with auditory torches. 1959-1964. 10.1145/1240866.1240932.

Introducing Jade, a New Digital Instrument

Sarah Sauvé^{1*}, Luke Welsh^{2**} and Andrew Staniland²

¹Cognitive Aging and Auditory Neuroscience Laboratory, Faculty of Medicine, Memorial University of Newfoundland, Canada, sarah.sauve@mun.ca

²Memorial Electroacoustic Research Laboratory, School of Music, Memorial University of Newfoundland, Canada

Abstract— In this paper we introduce the new digital instrument Jade. Jade is a digital instrument that makes music in three modes called ambient mode, creator mode and mindfulness mode. Together these modes integrate bio-signals and physical interaction. Furthermore, as Jade creates music using Ableton Live, the performer can manipulate the mappings between the instrument's input and its output. Playing with Jade requires no previous musical training or knowledge and is highly flexible and intuitive.

I. INTRODUCTION

Jade was created to fulfill an interest in novel electronic musical instruments that are inclusive and that allow the audience to see the interaction between performer and sound. The prototype is shown in Figure 1. The central processing of the instrument is an Arduino Mega 2560, which reads input from all sensors and buttons and processes the signals for transmission. Jade's three performance modes and its software are detailed below, followed by a list of possible research directions for the instrument.

II. MODES

Ambient mode collects environmental information via a BME280 atmospheric sensor and an RGB light sensor to detect temperature (°C), humidity (%), pressure (kPa) and altitude (m), and ambient light as well as red, blue and green light components respectively.

Creator mode collects gestural information via four ribbon potentiometers, which detect touch over their 100mm range, four ZX sensors, which detect vertical and horizontal placement at a range of 30cm and 15cm respectively, and eight RGB on/off control buttons which control the SoftPot and ZX sensors.

Mindfulness mode collects bio-signals using a custom-built external pulse and body temperature sensor and a Muse EEG headband. The pulse sensor detects reflected light from an LED to measure blood flow (volts) and heart rate (volts) while body temperature (°C) is measured using an MLX90614 IR thermometer. The Muse EEG headband sends

Figure 1. Jade instrument prototype.

data to the accompanying MuseDirect app, which is connected to Jade's software.

III. SOFTWARE

All Jade modes feed data to an Ableton Live patch, where input sources are connected to sound samples and Max/MSP patches. Performers can creatively manipulate the input/output mapping and mix the output as they wish.

IV. FUTURE RESEARCH

Jade's future research potential is vast; these are a few research directions we see for Jade. First, we would like to explore preference for a physical instrument versus an app-based instrument, shedding light on how people of all musical backgrounds interact with music and instruments. Second, we would like to see Jade used as an instrument where physical accessibility would otherwise be an issue (i.e. paralysis or uncontrolled movement). Third, we would like to explore the therapeutic potential of Jade as a mindfulness training guide for anxiety, depression, mania or dementia to name a few conditions that could benefit from this type of intervention.

ACKNOWLEDGMENT

Authors acknowledge Professor Andy Fisher, Mark Bennett and Alycia Leonard for the development of the Jade prototype and thank Suncor Energy and Memorial University VP Research for funding the Jade project via the Terra Nova Young Innovators Award.

*Research supported by Canada Research Chair to Dr. Benjamin Zendel.

**Research supported by a SSHRC grant awarded to the International Institute for Critical Studies in Improvisation.

S. Sauvé is with the Cognitive Aging and Auditory Neuroscience Lab, Memorial University of Newfoundland, Canada and has recently joined the Jade team to develop future research (e-mail: sarah.sauve@mun.ca).

L. Welsh is with the Memorial Electroacoustic Research Laboratory, Memorial University of Newfoundland, Canada (e-mail: lgw74@mun.ca).

A. Staniland is faculty at Memorial University of Newfoundland's School of Music and heads the Memorial Electroacoustic Research Laboratory, Canada (corresponding author e-mail: astaniland@mun.ca).

Characterising Glissando and Flutter-tongue Techniques in Recordings of Chinese Bamboo Flute

Changhong Wang, Emmanouil Benetos, Elaine Chew

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

Abstract— Focusing on the Chinese bamboo flute, we propose descriptors for two playing techniques frequently employed in performance: the glissando and the flutter-tongue. Glissandi are represented by their duration, range, and peaks of consecutive note changes, and flutter-tongue segments by their periodicity, timbre, and regularity. The glissando descriptors are used in a hidden Markov model-based detection system and a framework for flutter-tongue analysis is proposed. The study uses a newly created dataset of Chinese bamboo flute recordings that includes samples of the playing techniques in isolation and in performance context.

I. INTRODUCTION

Playing techniques such as ornamentations and articulation effects constitute important aspects of music performance. However, their computational analysis is still in its early stages due to a lack of instrument diversity, established methods and informative data. Most existing work on computational analysis of playing techniques focuses on Western instruments while playing techniques in non-Western instruments are often overlooked. Take for example, one of the world's most ancient instruments, the Chinese bamboo flute (CBF): many listeners are most often captivated by its unique timbre, which belies the twenty or more playing techniques invoked when performing on the instrument. Playing technique detection methods adopted in the literature are typically frame-wise classifiers based on high dimensional feature inputs [1], with little explanation of why the methods work. Using the CBF as our instrument of choice, we start with two playing techniques—glissando and flutter-tongue, rarely explored audio gestures in the literature, aiming to build a systematic methodology for automatically analysing CBF playing techniques. For ecological validity, we collect the playing techniques in isolation and in performance context.

II. DATASET

The dataset, recorded in a professional studio, comprises recordings by ten expert CBF players from the China Conservatory of Music. Each player performs both isolated playing techniques covering all notes and two full-length pieces. Players are grouped by flute type and each player uses their own flute. The fundamental frequency of each recording is estimated using the pYIN algorithm, all errors manually

corrected by the first author using Sonic Visualiser, and playing techniques are annotated and verified by players.

III. GLISSANDO DETECTION

The main characteristic of glissando is the consecutive note change, which we claim can be captured by latent states of a hidden Markov model (HMM) [2]. HMMs enable the decoding of note evolution while smoothing outlier variations within performed glissandi. A rule-based segmentation process first extracts glissando candidates that are consecutive note changes in the same direction. Glissandi are then identified by two HMMs (glissando and non-glissando). Besides traditional descriptors of pitch and intensity, three long-term descriptors—duration, range, and peaks of consecutive note changes are added to the input of HMMs. The results, based on both frame- and segment-based evaluation, confirm the feasibility of the proposed method. Better detection performance of ascending glissandi over descending ones is obtained due to their more regular patterns. Inaccurate pitch estimation forms a main obstacle for successful fully-automated glissando detection.

IV. FLUTTER-TONGUE ANALYSIS

Flutter-tongue is another frequently performed technique in CBF playing. The acting of a rapid rolling of the tongue tip, in interaction with a thin membrane glued to the resonance hole, creates richer harmonics for flutter-tongue than other CBF playing techniques. Different from vibrato, which exhibits sinusoidal patterns for both amplitude and frequency modulations, a sawtooth-like frequency modulation and a sinusoidal amplitude modulation are observed for flutter-tongue. Based on a modulation-scale analysis [3], a modulation rate, ranging from 30 to 60Hz, is found for CBF flutter-tongue. Descriptors consist of periodicity, timbre, and regularity are currently under development.

REFERENCES

- [1] J. Abeßer, H. Lukashevich, and G. Schuller, "Feature-based Extraction of Plucking and Expression Styles of the Electric Bass Guitar," in *ICASSP*, pp. 2290–2293, 2010.
- [2] C. Wang, E. Benetos, X. Meng, and E. Chew, "HMM-based Glissando Detection for Recordings of Chinese Bamboo Flute," submitted to *ICASSP*, 2019.
- [3] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale Analysis for Content Identification," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 3023–3035, 2004.

C. Wang is funded by the China Scholarship Council (CSC). E. Benetos is supported by a UK RAEng Research Fellowship (RF/128).

Musical Chills: Stimulus Properties, Stylistic Preference & Familiarity

Rémi de Fleurian^{1*} and Marcus T. Pearce^{1,2}

^{1*}Music Cognition Lab, Queen Mary University of London, UK, r.defleurian@qmul.ac.uk

²Center for Music in the Brain, Aarhus University, Denmark

Abstract— Little is known about the effects of stimulus-driven properties, stylistic preference, and familiarity on the occurrence of musical chills. In the present study, participants listened to 12 unfamiliar songs in liked and disliked musical genres. Half were taken from a dataset of songs previously reported as causing chills, while the other half were matched with these songs by artist and popularity. Objective measurements of piloerection and continuous self-reports of the occurrence of chills and intensely pleasurable moments were taken in two lab sessions, separated by a two-week longitudinal phase during which participants listened to the full set of songs another eight times. Preliminary results taken from the first lab session are discussed, in terms of occurrence of chills and intensely pleasurable moments across all conditions.

I. BACKGROUND

Musical chills consist of a pleasurable tingling sensation, sometimes accompanied by piloerection, and represent an emotionally intense physiological reaction to music [1]. They give a convenient insight into what makes music pleasurable because they are widespread, memorable, and observable.

Changes in dynamics, texture, melody, harmony, rhythm, and instrumentation have been linked to chills [2–4], but few studies have looked at the causal influence of such factors [5]. More specifically, it is unclear whether chills can be felt when listening to any piece of music, or whether they require a specific combination of stimulus-driven properties.

Chills are likely due to an interaction of top-down (e.g. development of expectations) and bottom-up (e.g. fulfillment of these expectations) processes [6]. This would suggest potential effects of stylistic preference and familiarity, but such effects have been sparsely explored as of yet.

The present work therefore aimed to investigate the causal influence of stimulus-driven properties, stylistic preference, and familiarity on the occurrence of musical chills, in a longitudinal study using real music.

II. METHODS

A subset of 93 songs was extracted from a previous survey study in which 221 participants reported 671 songs during which they often experience chills. Inclusion criteria were song duration (less than 5 min.), and absence of sung lyrics

within 10 s. from each reported instance of chills. Each song was then matched with three similarly popular songs (as assessed by number of plays on Spotify), from the same artist.

Participants in the present study took an online test in which they listened to randomly selected 15 s. excerpts for 40 songs and their associated matches, and rated them on liking for the genre of each excerpt and familiarity, resulting in an individual set of 12 unfamiliar songs for each participant, containing 3 songs for each combination of song provenance (survey or matched) and liking for the genre (liked or disliked).

Participants listened to the 12 songs in two lab sessions, separated by a two-week longitudinal phase away from the lab, during which participants listened to the full set of songs another eight times. In each lab session, piloerection was measured using a wearable optical device [7], and participants continuously reported the occurrence of chills and of intensely pleasurable moments using button presses.

III. RESULTS

The proposed poster will examine data obtained in the first lab session, with particular attention given to the effects of stimulus-driven properties and stylistic preference on the occurrence of musical chills.

REFERENCES

- [1] D. Huron, *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press, 2006.
- [2] O. Grewe, F. Nagel, R. Kopiez, and E. Altenmüller, “Listening to music as a re-creative process: Physiological, psychological, and psychoacoustical correlates of chills and strong emotions,” in *Music Perception*, 24(3), 2007, pp. 297–314.
- [3] M. Guhn, A. Hamm, and M. Zentner, “Physiological and musico-acoustic correlates of the chill response,” in *Music Perception*, 24(5), 2007, pp. 473–483.
- [4] J. A. Sloboda, “Music structure and emotional response: Some empirical findings,” in *Psychology of Music*, 19(2), 1991, pp. 110–120.
- [5] S. Bannister, and T. Eerola, “Suppressing the chills: Effects of musical manipulation on the chills response,” in *Frontiers in Psychology*, 9, 2018, p. 2046.
- [6] V. N. Salimpoor, M. Benovoy, G. Longo, J. R. Cooperstock, and R. J. Zatorre, “The rewarding aspects of music listening are related to degree of emotional arousal,” in *PLoS ONE*, 4(10), 2009, p. e7487.
- [7] M. Benedek, B. Wilfling, R. Lukas-Wolfbauer, B. H. Katzur, and C. Kaernbach, “Objective and continuous measurement of piloerection,” in *Psychophysiology*, 47(5), 2010, pp. 989–993.

*Research supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

R. de Fleurian: Queen Mary University of London, UK (corresponding author e-mail: r.defleurian@qmul.ac.uk).

M. T. Pearce: Aarhus University, Denmark, and Queen Mary University of London, UK (e-mail: marcus.pearce@qmul.ac.uk).

Tonic and Dastgāh Recognition in Iranian Music

Peyman Heydarian

Apulse, London, UK, moonlight_culture@yahoo.com

Abstract— The dastgāh, the underlying modal system of Iranian classical music is a phenomenon similar and closely connected to maqām in Turkish and Arabic music. It represents the scale and tonic, and is to some extent an indication of the mood of a piece. This paper demonstrates the dastgāh system and presents an algorithm for computational identification of the tonic and mode in Persian audio musical signals.

I. BACKGROUND

Iranian music is based on a modal system, consisting of seven main modes (shur, homāyun, segāh, chāhārgāh, mähur, rāst-panjgāh, navā) and their five derivatives (abu'atā, bayāt-é tork, afshāri, dashti, bayāt-é esfahān) [1]. They fall into five different scales. They can be played with 13 different notes: 7 diatonic notes, 3 semitones and 3 quartertones [2]:

E F $\sharp F$ $\sharp F$ G $\sharp G$ A $\flat B$ B C $\sharp C$ $\sharp C$ D
Sori ($\sharp$) and koron ($\flat$) symbols show half-sharp and half-flat quartertones respectively.

II. DATABASE

A database of 5706 seconds of music in five modes (91 pieces) was constructed, including scale notes, opening sections, melodies, and random sequences of notes played on a santur by a skilled musician (studio recordings) [3].

III. METHOD

Harmonic Pitch Class Profiles (HPCP), also known as chroma are used as the feature set throughout this research. Chroma representation codes by Harte et al are used here. As seen in Fig. 1, finding the scale, involves a tonic detection stage and pitch/chroma alignment. Having identified the scale and tonic, the dastgāh (mode) is recognized, except in rare cases where specific melodic motifs distinguish between similar modes. The bidirectional arrows between mode and melody show that melody recognition reveals the mode and the mode can be used to improve melody recognition systems. A full dastgāh performance can be recognized by tracking the modulations and the respective changing tonics.

Different tone resolutions (6-TET, 12-TET, 24-TET and 48-TET) were compared. 24-TET produced the best results and is used thereafter. The sparse chroma is made by setting the least 11 elements to zero and keeping the 13 remaining elements.

Dot-product, bit-masking, and Manhattan distance are used as the classifiers. Dot-product and bit-masking are computationally more efficient than Manhattan distance. The

classifiers compare the chroma of a given audio sample with either theoretical or data-driven templates. Theoretical-based templates are constructed either based on scale intervals or by making note histograms of existing pieces. Data-driven templates are made by chroma average of the audio samples.

Initially, a shift and multiplication process is performed, the shifted HPCP (Transposed HPCP) or THPCP is obtained, and the test samples are aligned with the training templates (theoretical-based or data-driven templates). Subsequently, the minimum distance between a test sample and the templates for each class determines the mode.

IV. RESULTS

The recognition rates for Manhattan distance and data-driven samples, dot-product and data-driven samples, and theoretic scale intervals and bit-masking are 83.33%, 72.73% and 51.52% respectively (training samples were excluded).

Figure 1: Dastgāh identification flowchart

REFERENCES

- [1] Farhat, H., *The Dastgāh Concept in Persian Music*, Cambridge: Cambridge University Press, 1990.
- [2] Heydarian, P., Automatic Recognition of Persian musical modes in audio musical signals, PhD thesis, 2016.
- [3] www.thesantur.com [accessed 20 November 2018]
- [4] Harte, C.A., Sandler, M.B. "Automatic chord identification using a quantised chromagram", *Proceedings of the 118th Convention of the Audio Engineering Society*, Barcelona, Spain, May 2005.