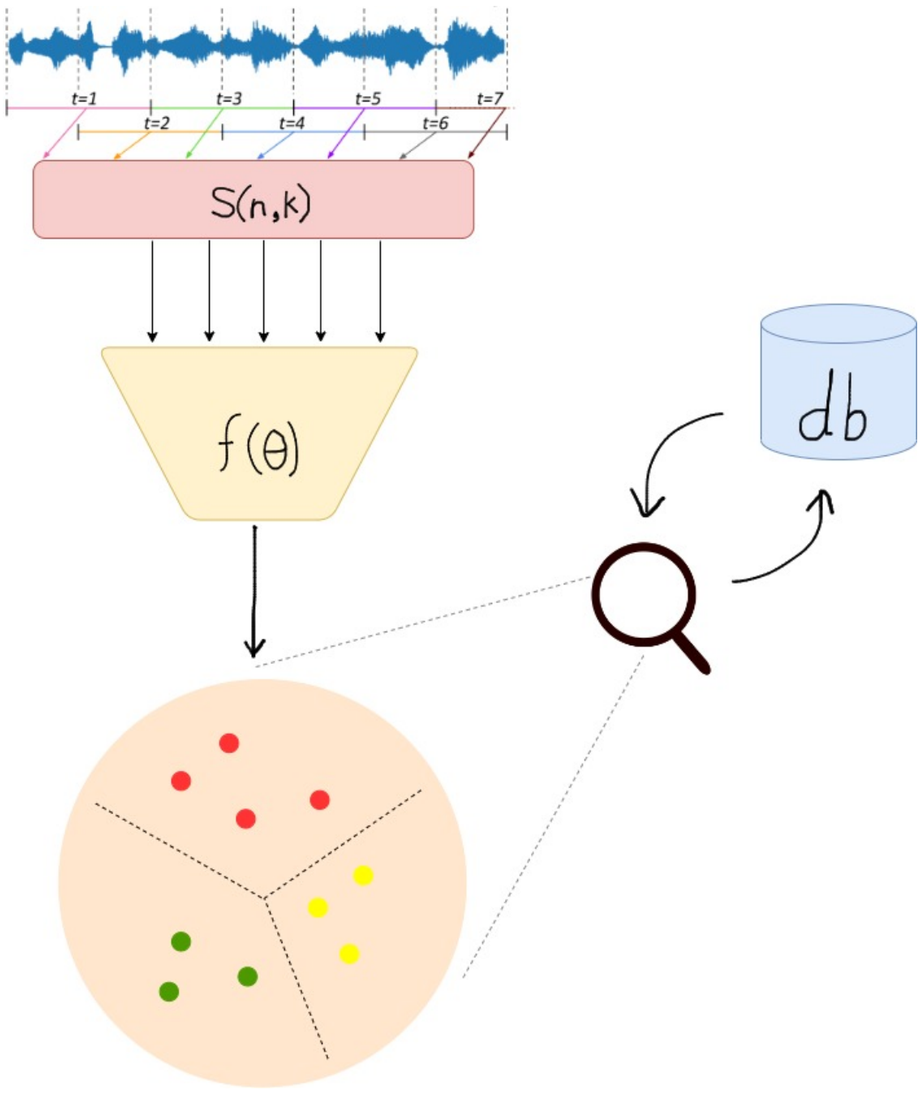
AIM Summer School 2023

Self-supervision in Audio Fingerprinting

Aditya Bhattacharjee

2022



Finding

The state-of-the art audio fingerprinting framework is not robust to pitch-shifting and time-stretching.

Question

How to model the scalability of a search-retrieval task such as audio identification? Real-world performance of audio fingerprinting is measured at a scale which is orders of magnitude bigger than proposed experimental setups in the state-of-the-art. Is there a way to model the "capacity" of an embedding space?

Supervisor(s): Emmanouil Benetos; Joren Six

AIM Summer School 2023

Real-Time Expressive Automatic DJ Mixing of Electronic Dance Music and Digital Audio Workstation Applications

Alexander Williams

2022

Finding

A literature review revealed that methods for automatic evaluation of DJ mixes are few and inconsistent. The length of DJ mixes make listening tests impractical and results are subjective to listener tastes. There are also multiple lenses by which to evaluate a DJ mix at the individual transition and song level and in the sequencing of a complete mix.

Question

Questions remain about how to computationally evaluate and compare DJ mixes in a quantitative way. How do we compute concise and objective measures of the quality and properties of DJ mixes at different temporal resolutions that also reflect subjective differences? Further, what factors affect these qualities and can we control them?

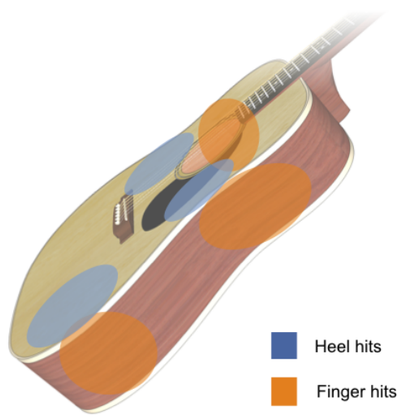
Supervisor(s): Stefan Lattner, Mathieu Barthet

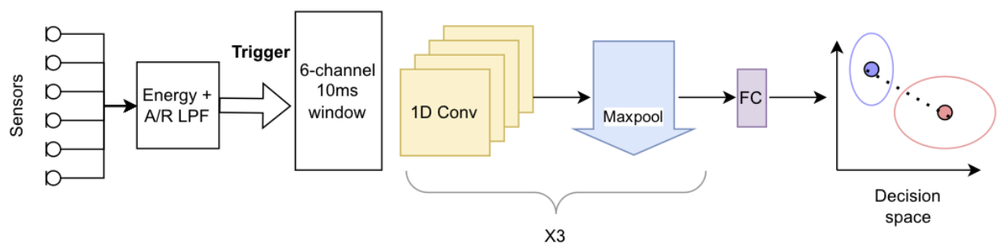
AIM Summer School 2023

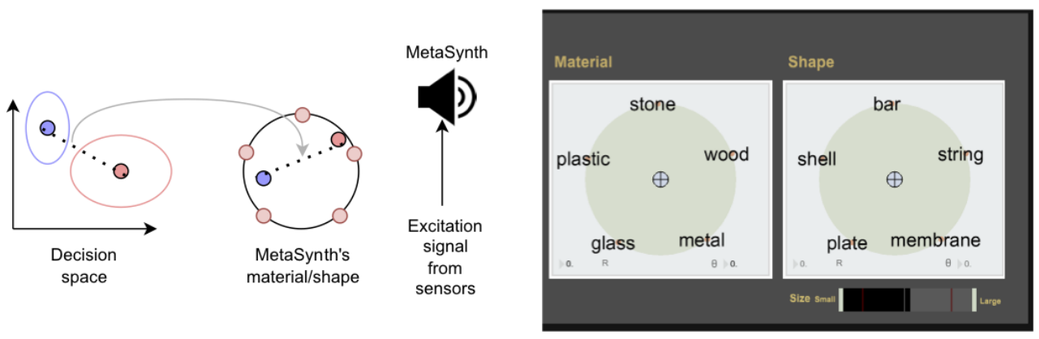
The HITar, an augmented guitar for percussive fingerstyle with DNN body hit description

Andrea Martelloni

2019







Finding

We built an acoustic guitar prototype that uses a soft classifier to control the parameter space of a synthesis engine in real time. The embeddings (last layer before the output layer) are regularised so that they act as a low-dimensional latent space, providing a deeper description than the output layer of the classifier. System latency is ~13 ms.

Question

Usability evaluation: How do players behave with it? Does the re-synthesised sound match the expectation of the player? Does the richer representation carry the subtle differences across hits normally present in expressive playing?

Supervisor(s): Prof Andrew McPherson, Dr Mathieu Barthet

AIM Summer School 2023

Modeling Jazz Piano: Symbolic Music Generation via Large-scale Automatic Transcription

Andrew Edwards

2021

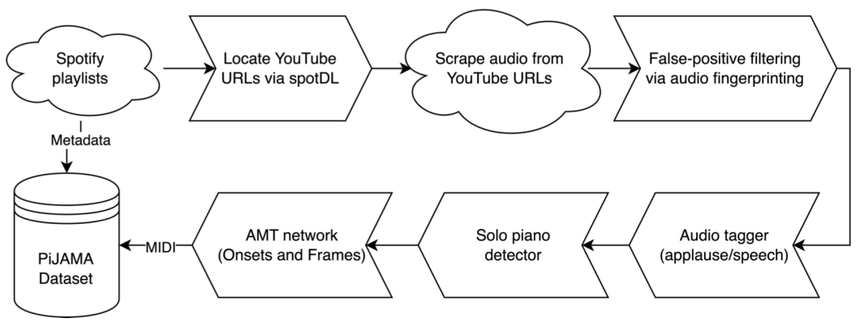


Figure 1: Diagram of the data collection process for the PiJAMA dataset.

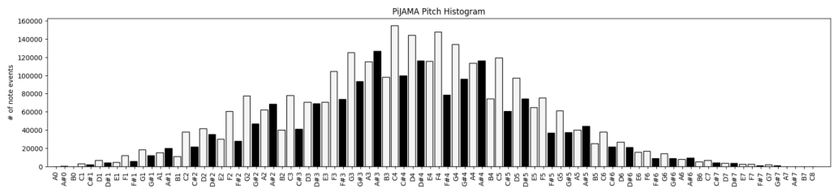


Figure 2: Pitch histogram of all note events in the PiJAMA dataset.

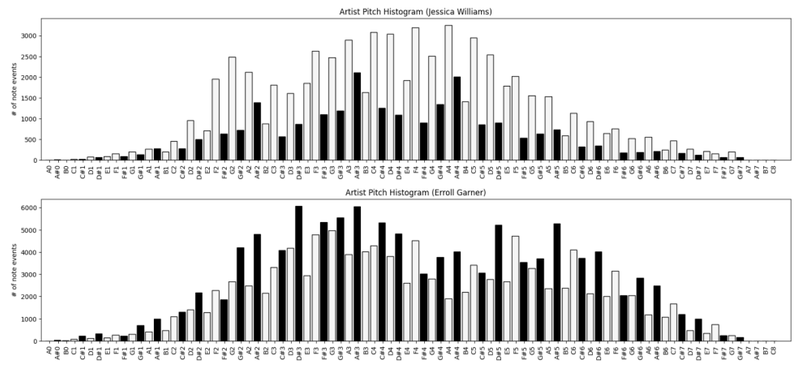


Figure 3: Pitch histograms from pianists Jessica Williams and Erroll Garner.

Finding

In current S.O.T.A. piano transcription work, data augmentation is reported to hurt test set performance. We have preliminary results of matching S.O.T.A. MAESTRO test set results while also setting new S.O.T.A. results on out-of-distribution evaluation using a combination of data augmentation and automatic data collection.

Question

Can we use automatic chord and melody detection on a large dataset of jazz piano MIDI to build a system to perform an automatic arrangement from a lead sheet?

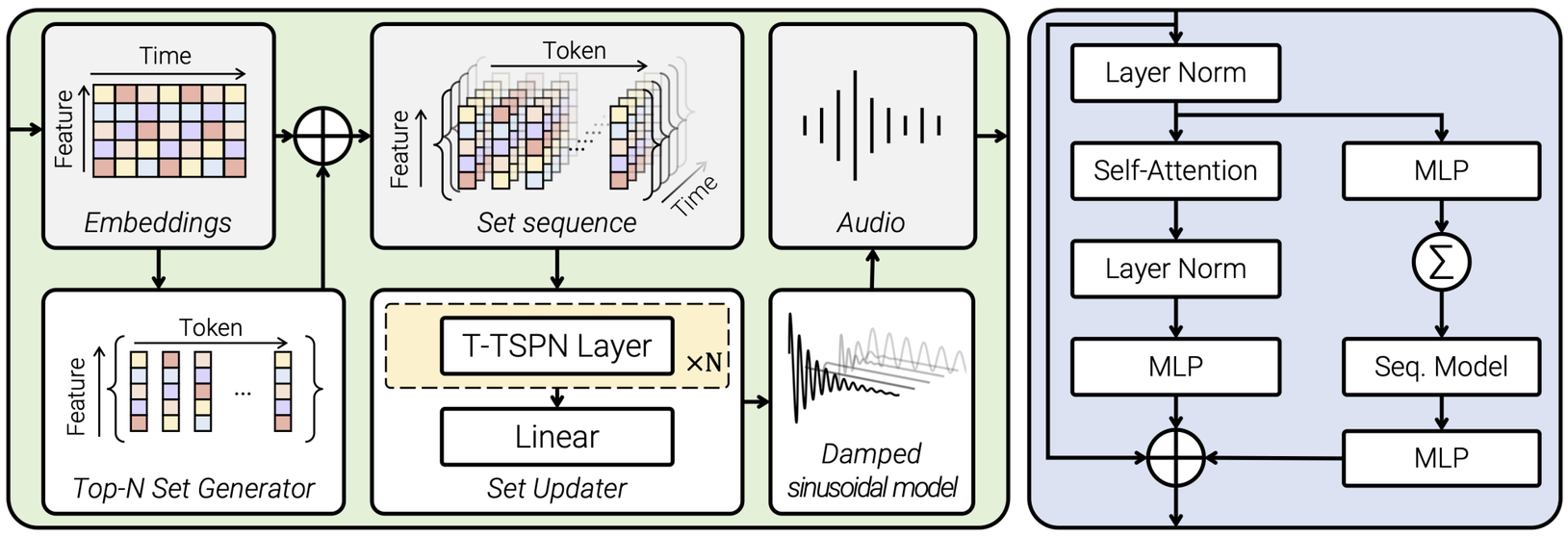
Supervisor(s): Simon Dixon and Emmanouil Benetos

AIM Summer School 2023

Permutations, periodicity, and symmetry: resolving optimisation pathologies in differentiable signal processing

Ben Hayes

2020



(i) Differentiable sinusoidal modelling with T-TSPN

(ii) T-TSPN Layer

Finding

There exists a simple surrogate model that allows direct optimization of sinusoidal parameters.

Question

Many signal processors are invariant under the actions of some finite group on their parameters. To what extent do these symmetries hinder the learning of neural network controllers for synthesizers and audio processors? And what strategies can be used to resolve this?

Supervisor(s): Charalampos Saitis & George Fazekas

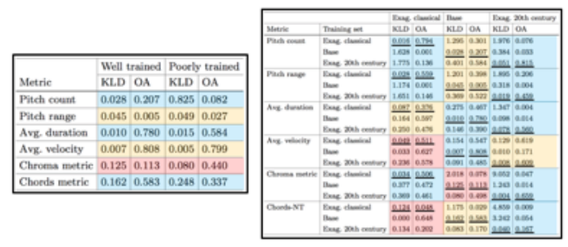
AIM Summer School 2023

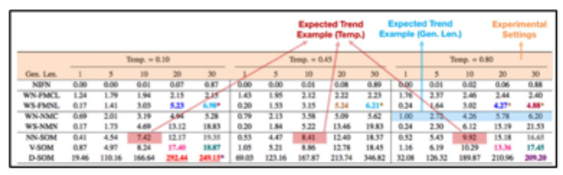
Composing Contemporary Classical Music Using Generative Deep Learning

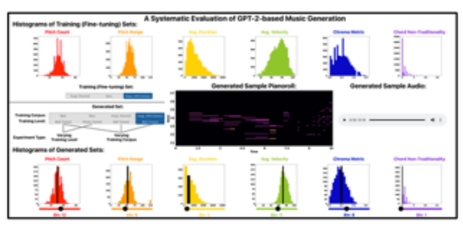
Berker Banar

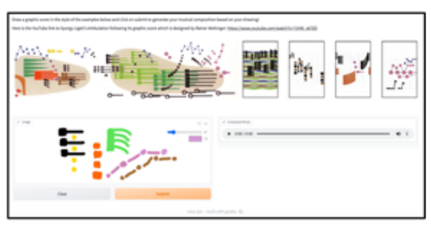
2019

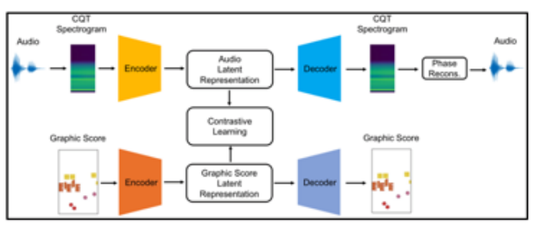
Metric	Well trained		Poorly trained	
	KLD	OA	KLD	OA
Pitch count	0.028	0.207	0.825	0.082
Pitch range	0.045	0.005	0.049	0.027
Avg. duration	0.010	0.780	0.015	0.584
Avg. velocity	0.007	0.808	0.005	0.799
Chroma metric	0.125	0.113	0.080	0.440
Chords metric	0.162	0.583	0.248	0.337

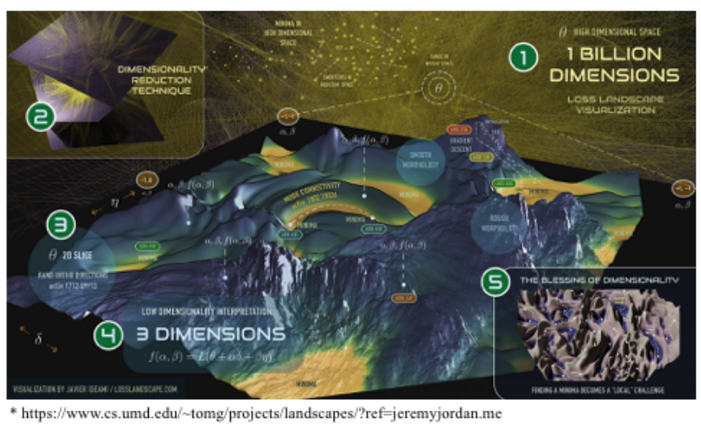




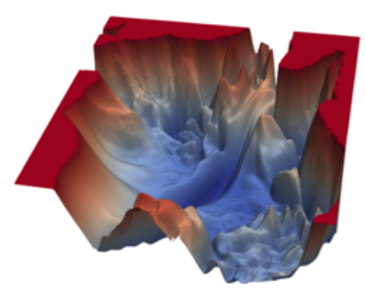


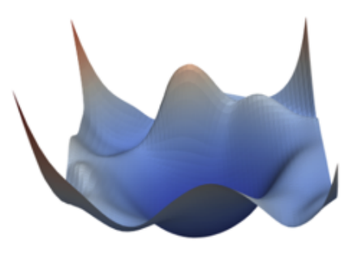






* https://www.cs.umd.edu/~tomg/projects/landscapes/?ref=jeremyjordan.me





Finding

Our methodology for comprehensive and domain-specific assessment of creatively generative models shows the inadequecy of off-the-shelf loss functions in creative contexts, improves the model selection process and helps to harvest novel artefacts while challenging the paradigm of learning for perfection.

Question

While optimising creatively generative models, are there any wormholes in the loss landscape? What about creating black holes? (let's talk more about it!)

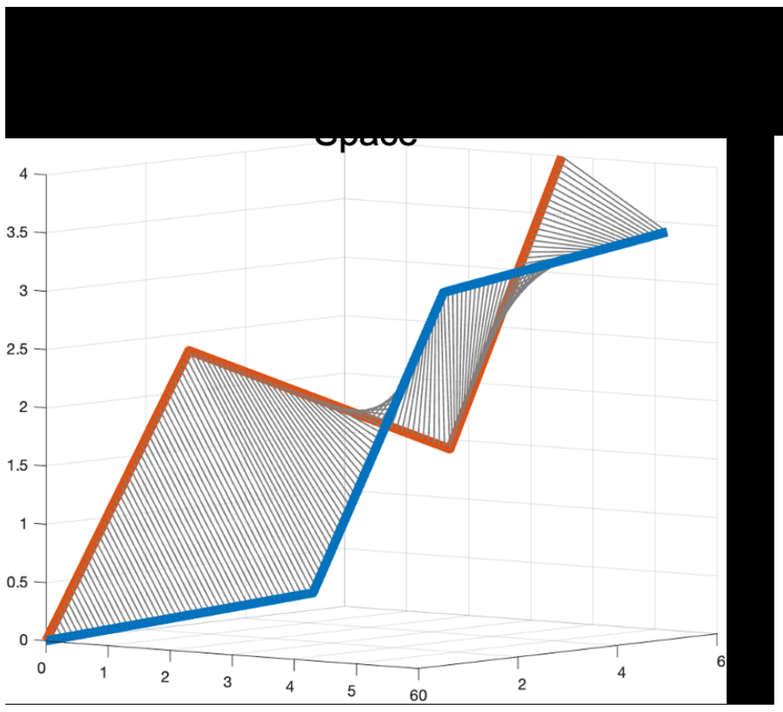
Supervisor(s): Simon Colton

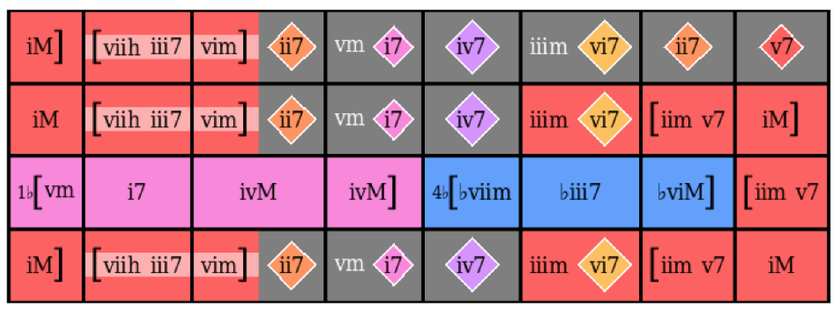
AIM Summer School 2023

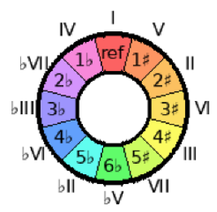
Live-Jazz Cover Song Identification

Carey Bunks

2022







Confirmation
Number of Bars: 32
Time Signature: 4/4
DB Key Signature: F
Ref. Major Scale: F

Finding

Can language models be applied to jazz harmony? I have proposed using co-occurrence vectors to embed jazz chords, and to use them to model chord progressions in the resulting latent space. I have shown that a novel distance metric, the membrane area, is successful at measuring the harmonic similarity between songs, and useful for identifying contrafacts.

Question

As melodic improvisation is a significant component of jazz, it is not a strong identifying feature. However, improvised melodies are predominantly coherent with a song's chord progression. An open question I would like to answer is whether harmony can be used to effectively identify songs played during live jazz performances.

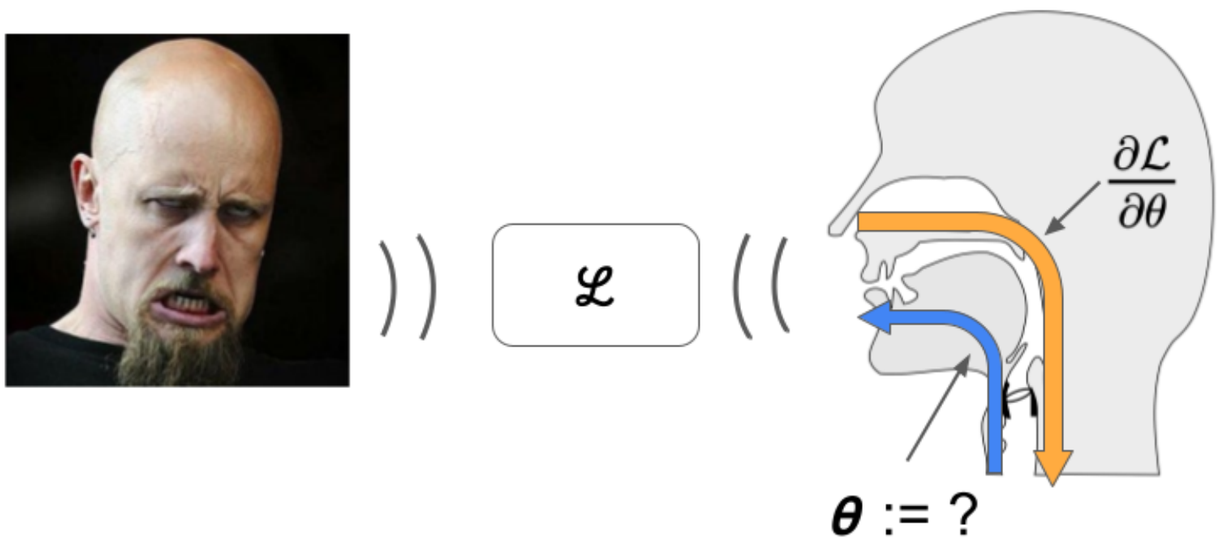
Supervisor(s): Simon Dixon, Bruno Di Giorgi

AIM Summer School 2023

Analysing and controlling extreme vocal expression using differentiable DSP and neural networks

Chin-Yun Yu

2022



Finding

Putting more prior knowledge of vocal production into the design of differentiable operators reduces computational loads and training resources. Moreover, fixing the source oscillator to the shape of glottal pulses makes modelling phase responses possible.

Question

What's the efficient way to parameterise the non-harmonic components (e.g., roughness) of voice? Are the resulting parameters (or representation) closely related to how humans perceive and describe screaming vocal?

Supervisor(s): George Fazekas, Emmanouil Benetos

AIM Summer School 2023

Decoding Auditory Attention and Musical Emotions with EarEEG

Chris Winnard

2021

“Traditional” scalp electroencephalography:

cEEGrid around-the-ear EEG:

Finding

We have recently built on previous attention and emotion decoding works (particularly An et al., 2021) by developing calibration tools to ensure that the loudness/spatialisation settings for various instruments are tailored to the participant. This minor part of the experiment will help to ensure consistent participant engagement, and meaningful EEG data.

Question

A practical issue that we have been struggling with is how to implement “oddballs” into an experiment. Broadly, “oddballs” are deviations from what would be expected in stimuli (e.g, a sudden but transient pitch shift in music). We are trying to implement these in polyphonic stimuli so that they can be heard, but with some effort, for attention-based tasks.

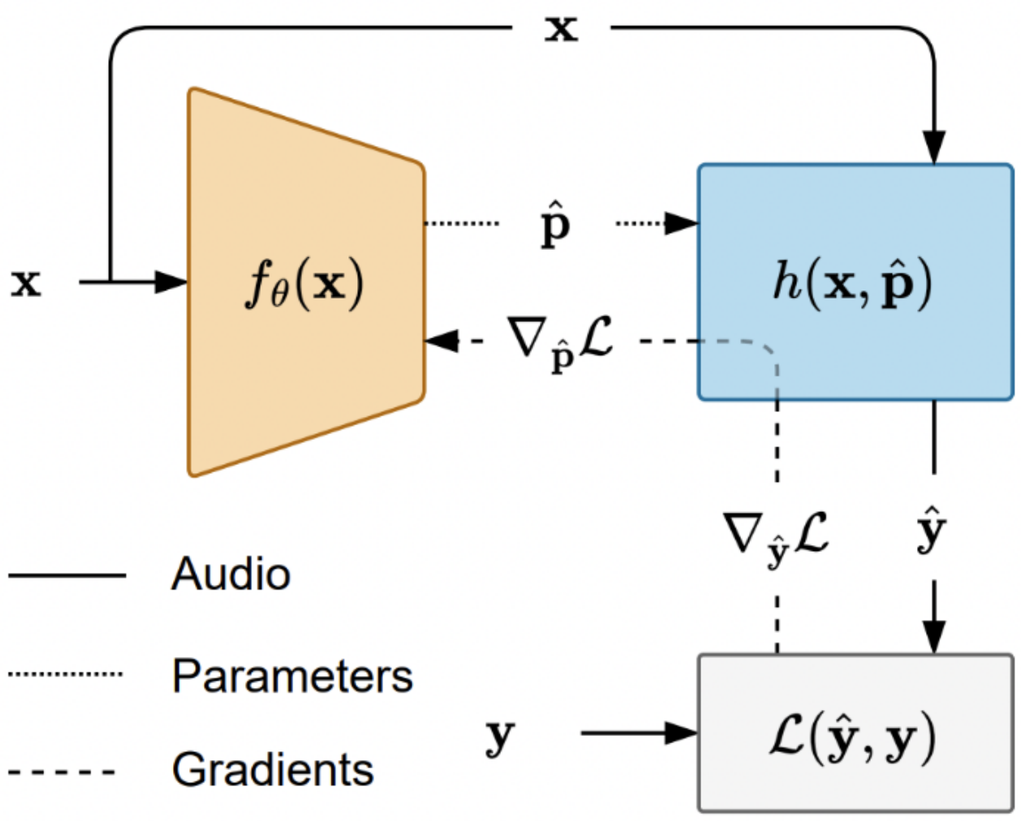
Supervisor(s): Dr. Marcus Pearce, Prof. Preben Kidmose, Prof. Kaare Mikkelsen

AIM Summer School 2023

Designing and Controlling Audio Effects with Machine Learning

Christian Steinmetz

2020



Finding

Automatic differentiation has been shown empirically to perform best in order to train neural network models to learn to control audio effects, however this a white-box approach that limits its applicability in real-world scenarios.

Question

How can we not only learn to control non-differentiable signal processors but also learn to dynamically construct an audio processing graph of these processors for specific tasks.

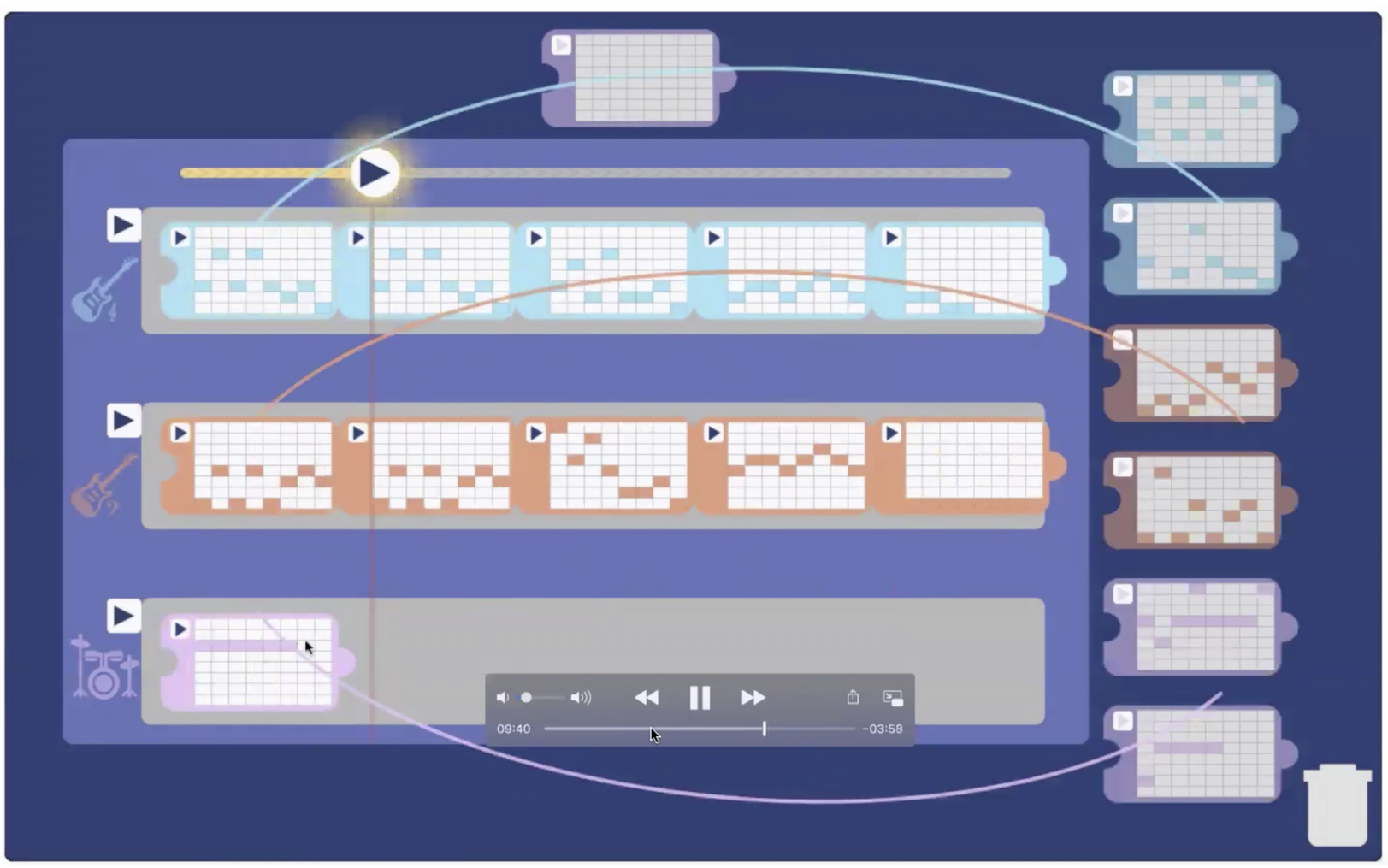
Supervisor(s): Joshua D. Reiss and Shanxin Yuan

AIM Summer School 2023

Exploring Reflection and Engagement in Digital Music Composition with AI

Corey Ford

2020



Finding

Avoiding the Clippy Effect: How to design UIs which encourage reflection, whilst not being detrimental to engagement, for people's music composition with generative AI? Key findings: i) AI perceived as better than user's music ∴ no reflection, ii) ignorable designs helps AI remain unobtrusive, and iii) AI used in lieu of engaging with unfamiliarity.

Question

Should AI designs be purposefully intrusive to make people to people stop and think more deeply (slow thinking)? Is it okay to purposefully interrupt creative flow? What would be an AI's role in doing so? Could we make effective use of LLMs and/or CUIs? How can we design AI in a human-centred way to extend - and not replace - people's creativity?

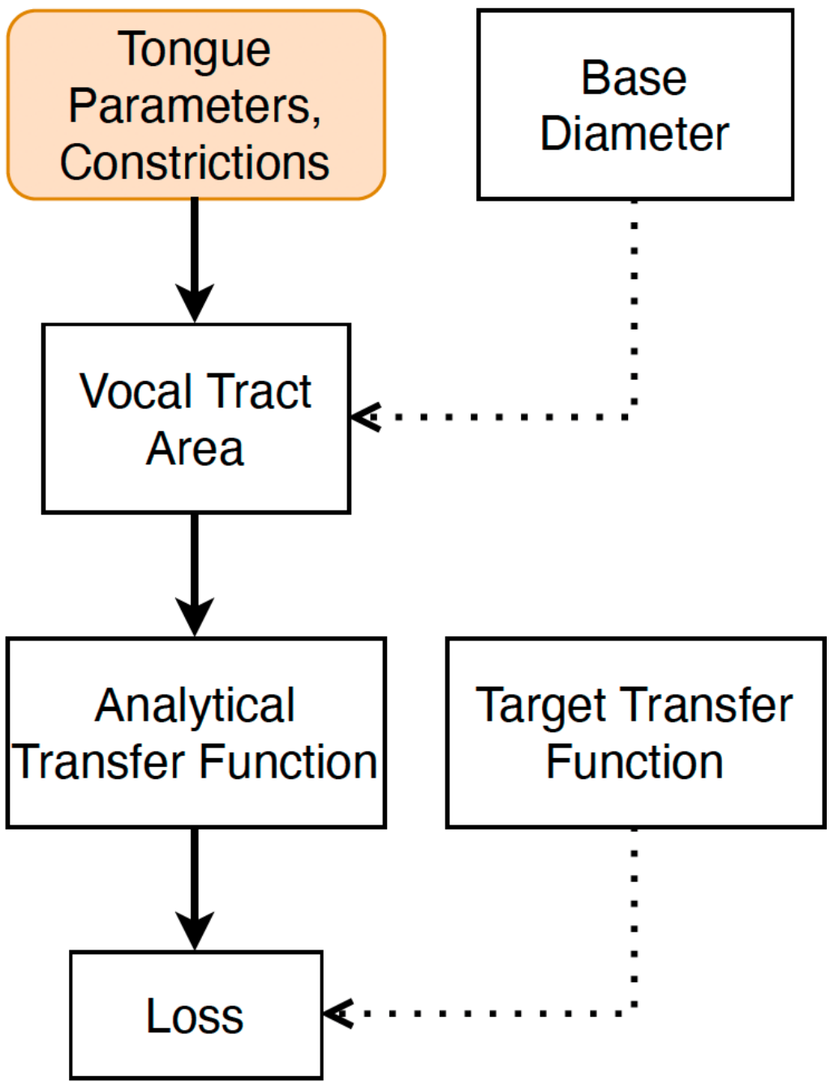
Supervisor(s): Nick Bryan-Kinns

AIM Summer School 2023

Machine learning of physical models for voice synthesis

David Südholt

2022



Finding

In the task of sound matching simple physical models of voice production to real human recordings, white-box optimization with gradient descent and differentiable operations outperforms black-box parameter prediction.

Question

With the goal of matching a digital waveguide to a target pitch, can the length of a fractional delay line be estimated using differentiable allpass filters?

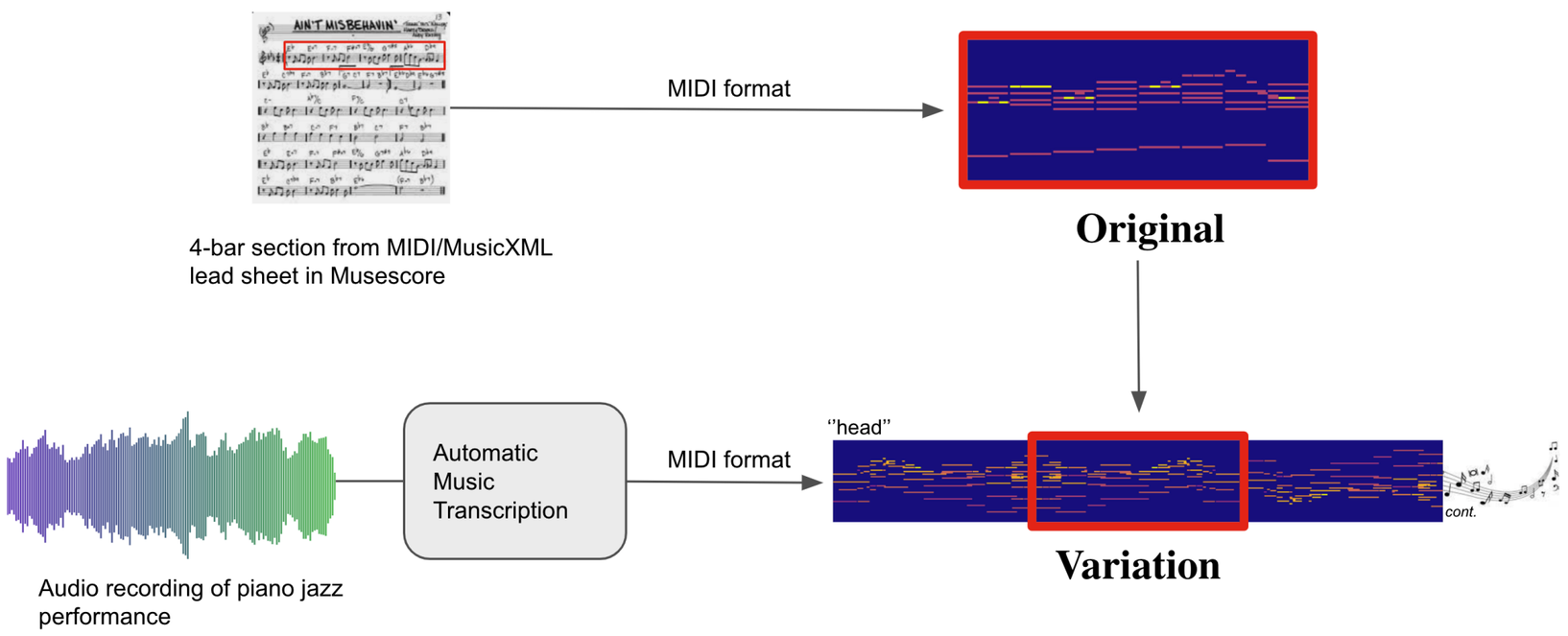
Supervisor(s): Joshua Reiss

AIM Summer School 2023

Music Overpainting: A Generative Music Model and AI Music Composition Tool for Creating Musical Variation

Eleanor Row

2020



Finding

A key finding was that many generative music tasks were not relatable tasks in music composition. I defined a new task, Music Overpainting, which was based on a composer creating variation upon an existing musical idea. I also created a dataset for this, and found that transformer models were better at generalising to the data than more simple models.

Question

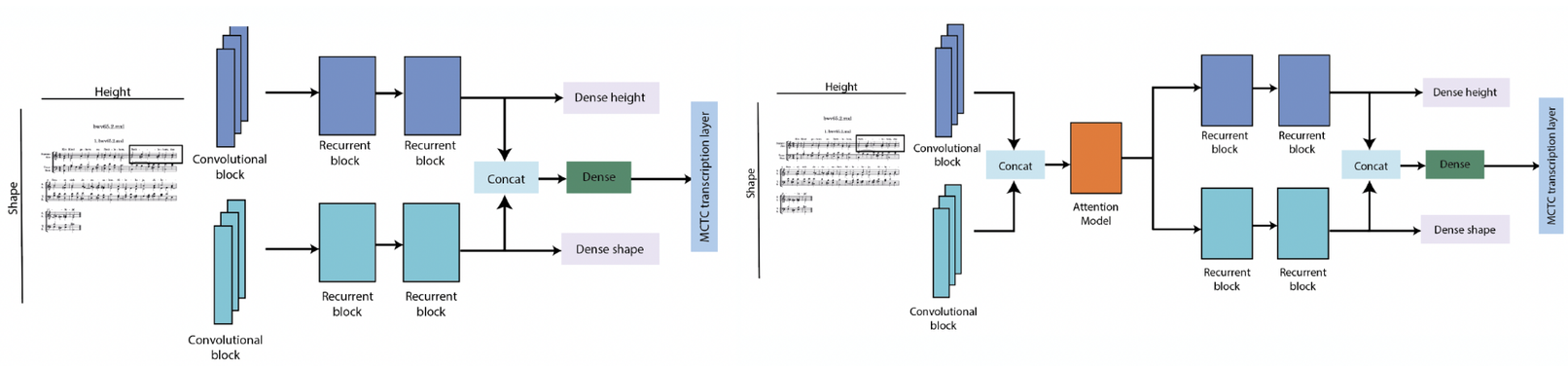
I will be focusing on creating a larger dataset for the task, and training various models using this data. I will also be exploring the use of transfer learning my dataset to models I've trained on larger more well-known datasets to see if this will improve generated outputs.

Supervisor(s): George Fazekas, Nick Bryan-Kinns

AIM Summer School 2023

Optical Music Recognition

Elona Shatri
2019



Finding

We use instance segmentation for entire sheet music images. It provides an accurate representation of musical symbols, especially in dense scores with overlapping symbols, and artefacts. Further we want to explore with RPN anchor ratio, increasing dataset size, and augmenting data with noise and artefacts for real-world applicability.

Question

OMR poses challenges for computer vision, similar to text recognition. In music scores, we propose a unified model with hybrid attention to segment staves and recognise musical symbols. Processes each stave iteratively, using an encoder for feature maps, an attention module for vertical weighted mask generation, and a decoder for symbol recognition.

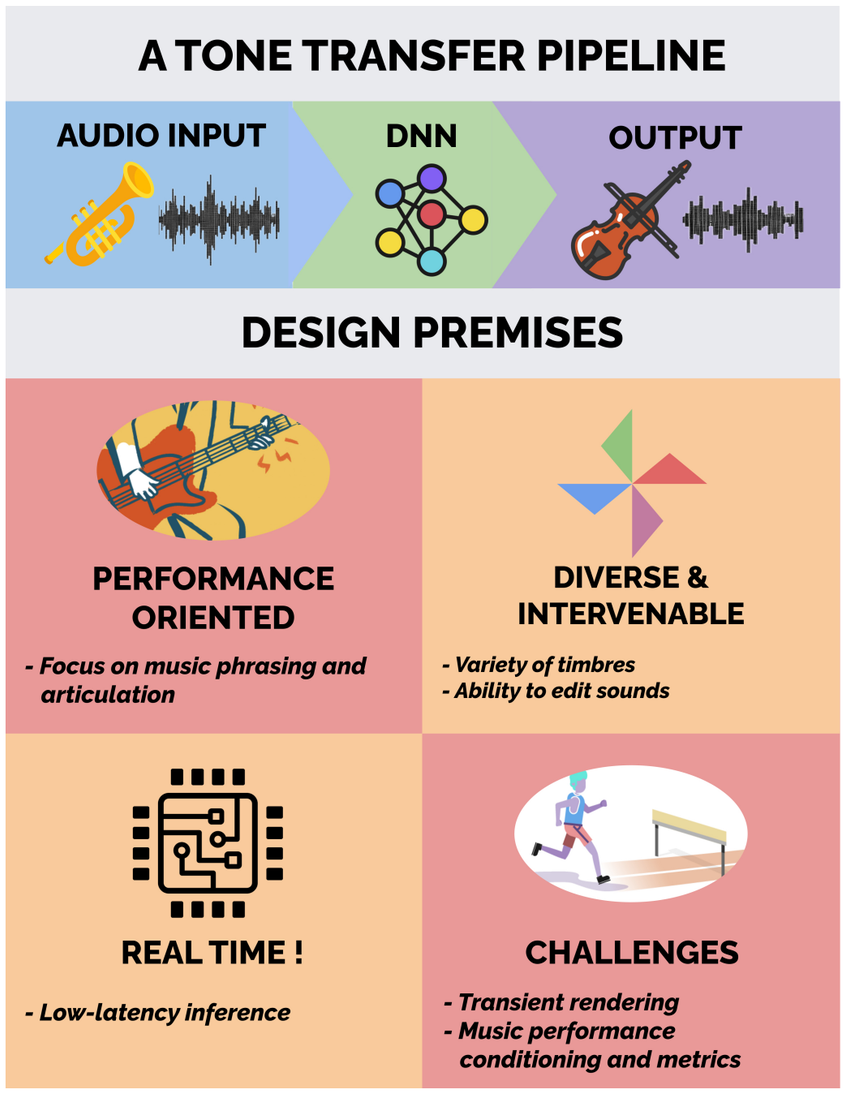
Supervisor(s): George Fazekas

AIM Summer School 2023

Interpretable and Expressive Tone Transfer Algorithms

Franco Caspe

2021



Finding

Can we expose synthesis controls in neural synthesis algorithms that are familiar to sound designers? Yes. We designed an approach for differentiable FM synthesis where a DNN learns to control the envelopes of a compact set of FM oscillators. Now are working on a second approach that leverages the vast amount of FM patches and employs them for Tone Transfer.

Question

How can we improve the musical phrasing capabilities of Tone Transfer algorithms? We may approach this from different angles, analyzing how to disentangle performance from musical form and timbral identity, how to improve audio rendering including transient information in the learning objective, and developing a strategy to quantify phrasing diversity.

Supervisor(s): Andrew McPherson, Mark Sandler

AIM Summer School 2023

Computational Modeling of Expressive Piano Performance

Huan Zhang

2021

Finding

A dataset that documents variety of expressions, a set of performance profile features that characterize expressiveness, and also some attempts in generating performances.

Question

To what extend we can capture the human performance patterns and reproduce them? To what extend we can assess expression from the existing performances?

Supervisor(s): Simon Dixon

AIM Summer School 2023

Bridging audio and language for music understanding

Ilaria Manco

2019

Multimodal deep learning for MIR

Our experience of music is multimodal: we listen to audio, watch music videos, look at album cover art, read and write reviews, etc. Machine music understanding instead often relies on audio and/or metadata in isolation. My PhD research aims at developing ***multimodal representation learning*** *models for* ***music understanding and description****, with a focus on* ***audio and language***

Ilaria Manco

i.manco@qmul.ac.uk
www.ilariamanco.com

Research Interests

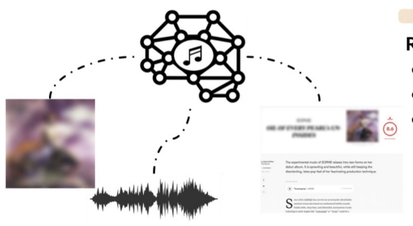
- Music information retrieval
- Multimodal representation learning
- Audio-and-language models

Applications

- Music captioning
- Cross-modal retrieval
- Music search, discovery, recommendation

MusCaps: Generating Captions for Music Audio

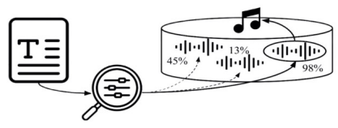
arXiv: 2104.11984
Code: github.com/ilaria-manco/muscaps



Song Describer: a Platform for Collecting Textual Descriptions of Music Recordings

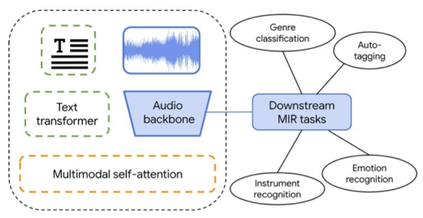
Extended abstract: [link]
Website: song-describer.streamlit.app
Code: github.com/ilaria-manco/sond-describer

Contrastive Audio-Language Learning for Music



arXiv: 2208.12208
Code: github.com/ilaria-manco/muscall

Learning Music Audio Representations via Weak Language Supervision



arXiv: 2112.04214
Code: github.com/ilaria-manco/mulap

Finding

Joint audio-language learning leads to better music representations and enables new music-related tasks.

Question

Are we doing joint multimodal training right?

Supervisor(s): Emmanouil Benetos, George Fazekas, Elio Quinton

AIM Summer School 2023

Timbre Transfer For A Smart Acoustic Guitar

Jack Loth

2021

*Table 1: Spearman correlations for MDS dimensions of each playing style with the pruned timbre descriptors (representation noted in parentheses; significant correlations in bold: * p<.05, ** p<.01, *** p<.001).*

Playing style	**Picking**				**Strumming**				**Fingerstyle**			
Dimension	1	2	3	4	1	2	3	4	1	2	3	4
Spectro-temporal variation (ERB)	-0.24	0.22	-0.04	-0.52	0.47	0.35	0.2	0.48	0.24	0.35	0.14	-0.47
Frame Energy (Harmonic)	0.24	-0.03	0.53	0.16	0.08	-0.32	0.02	-0.39	-0.12	0.47	0.08	**0.67***
Harmonic Energy (Harmonic)	**-0.89*****	0.16	-0.12	0.3	0.41	-0.49	**0.73***	-0.04	-0.42	-0.24	0.04	0.16
Noisiness (Harmonic)	**0.92*****	-0.42	0.1	-0.16	-0.39	0.56	**-0.72***	0.03	0.54	0.13	0.25	-0.25
Odd to even harmonic ratio (Harmonic)	-0.18	-0.48	-0.39	0.19	0.36	-0.48	0.59	0.53	-0.36	0.35	0.02	**0.64***
Spectral Centroid (Harmonic)	**0.65***	-0.49	0.15	-0.04	-0.41	0.47	**-0.71***	-0.21	**0.82****	-0.32	0.33	-0.53
Spectral Kurtosis (Harmonic)	**-0.67***	-0.09	-0.02	0.04	-0.01	-0.12	0.62	0.01	-0.52	0.47	-0.02	0.62
Spectral Slope (Harmonic)	0.62	**-0.67***	0.27	0.04	-0.45	0.1	-0.5	-0.32	**0.79****	-0.53	0.47	-0.3
Spectro-temporal variation (Harmonic)	0.32	-0.3	-0.2	**-0.73***	**0.75***	-0.25	0.44	0.12	**0.88*****	-0.38	0.27	-0.5
Spectral Decrease (STFT)	-0.01	-0.31	0.31	0.36	-0.45	-0.38	-0.16	-0.39	-0.35	-0.24	-0.05	0.6
Spectro-temporal variation (STFT)	0.05	-0.24	-0.04	-0.62	0.14	0.37	-0.3	0.47	0.47	0.05	0.1	**-0.78****
Effective Duration (TEE)	0.1	0.27	**-0.73***	0.08	0.54	0.04	0.07	-0.18	0.26	0.1	0.24	0.49

Finding

How are acoustic guitars perceived in relation to one another? We conducted a guitar timbre listening test which revealed a dependence of a steel-string acoustic guitar’s timbral profile on the playing style.

Question

What kinds of model architectures would be best suited for guitar-to-guitar timbre transfer?

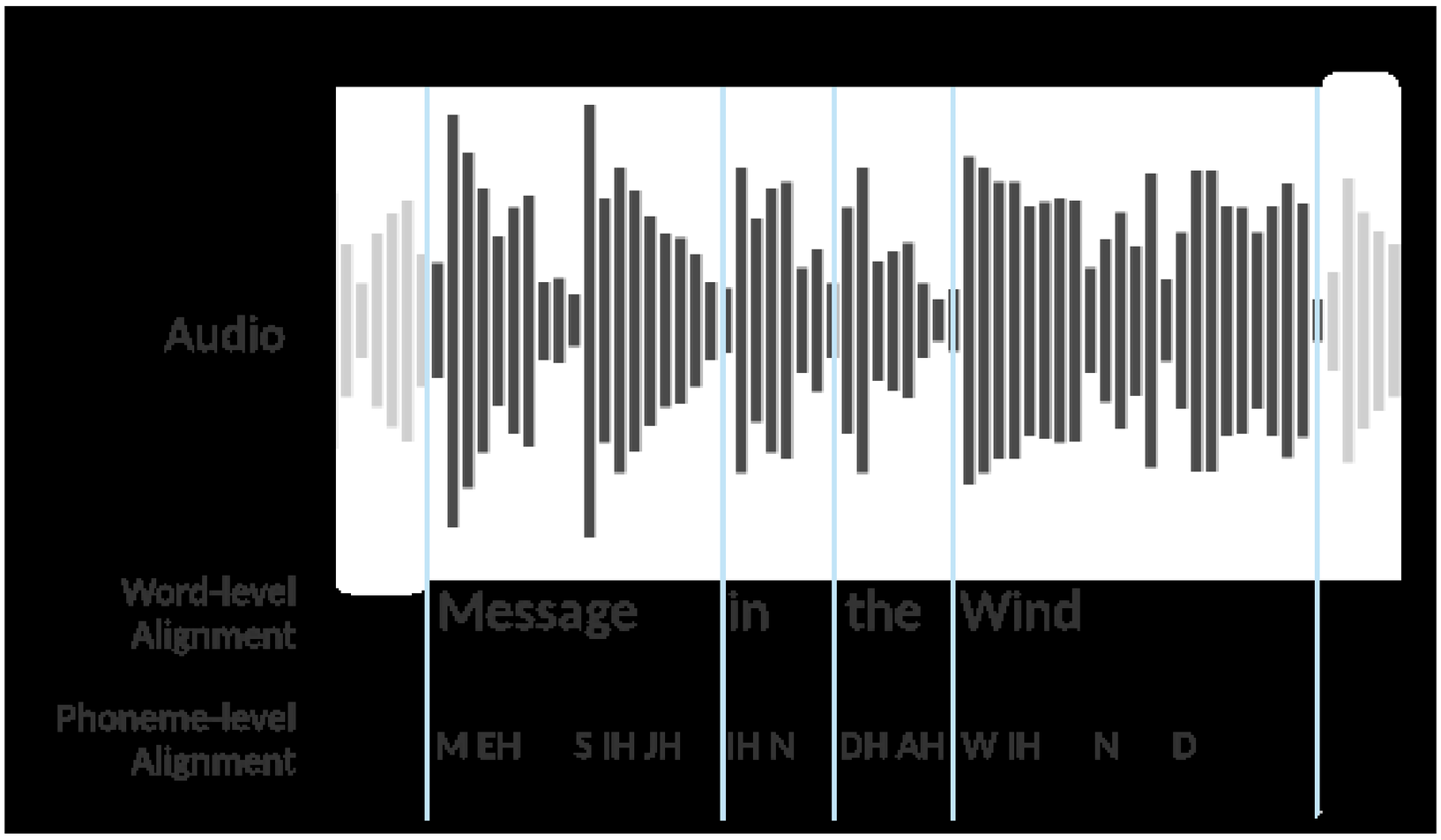
Supervisor(s): Mathieu Barthet

AIM Summer School 2023

Deep Learning for Singing Analysis and Manipulation

Jiawen Huang

2020



Finding

How could pitch and structure information improve the accuracy of lyrics alignment?

Question

How can speech conversion techniques be adapted for manipulating harmonics and rhythm in singing?

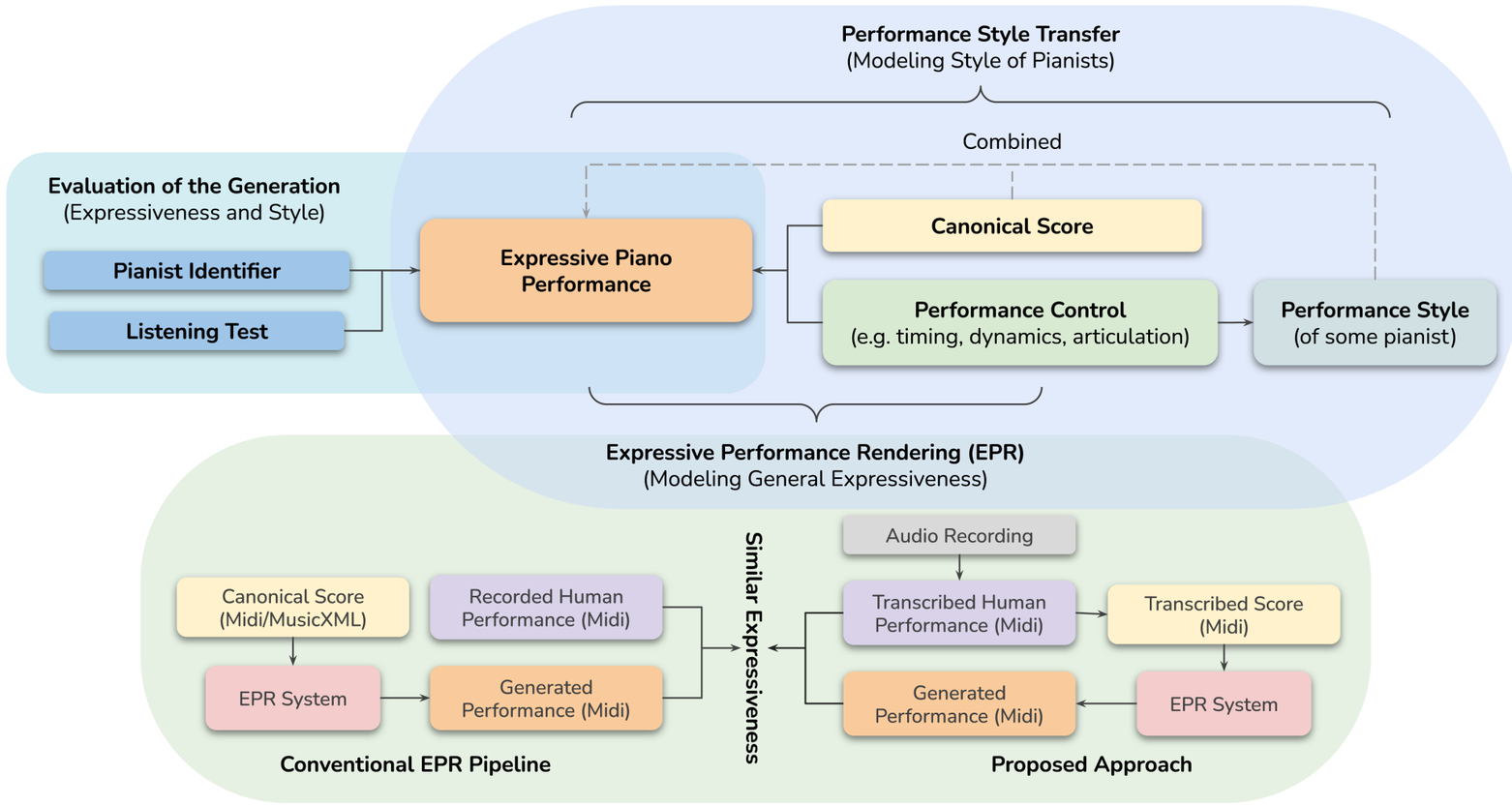
Supervisor(s): Emmanouil Benetos, Sebastian Ewert

AIM Summer School 2023

End-to-End System Design with Deep Learning for Performance Style Transfer and Generation

Jingjing Tang

2020



Finding

Transformers can learn and reconstruct human expressiveness in music performances and the differences of pianists could be learned by convolutional neural networks.

Question

Is that possible to generate performances of a composition in the styles of different pianists through contrastive learning?

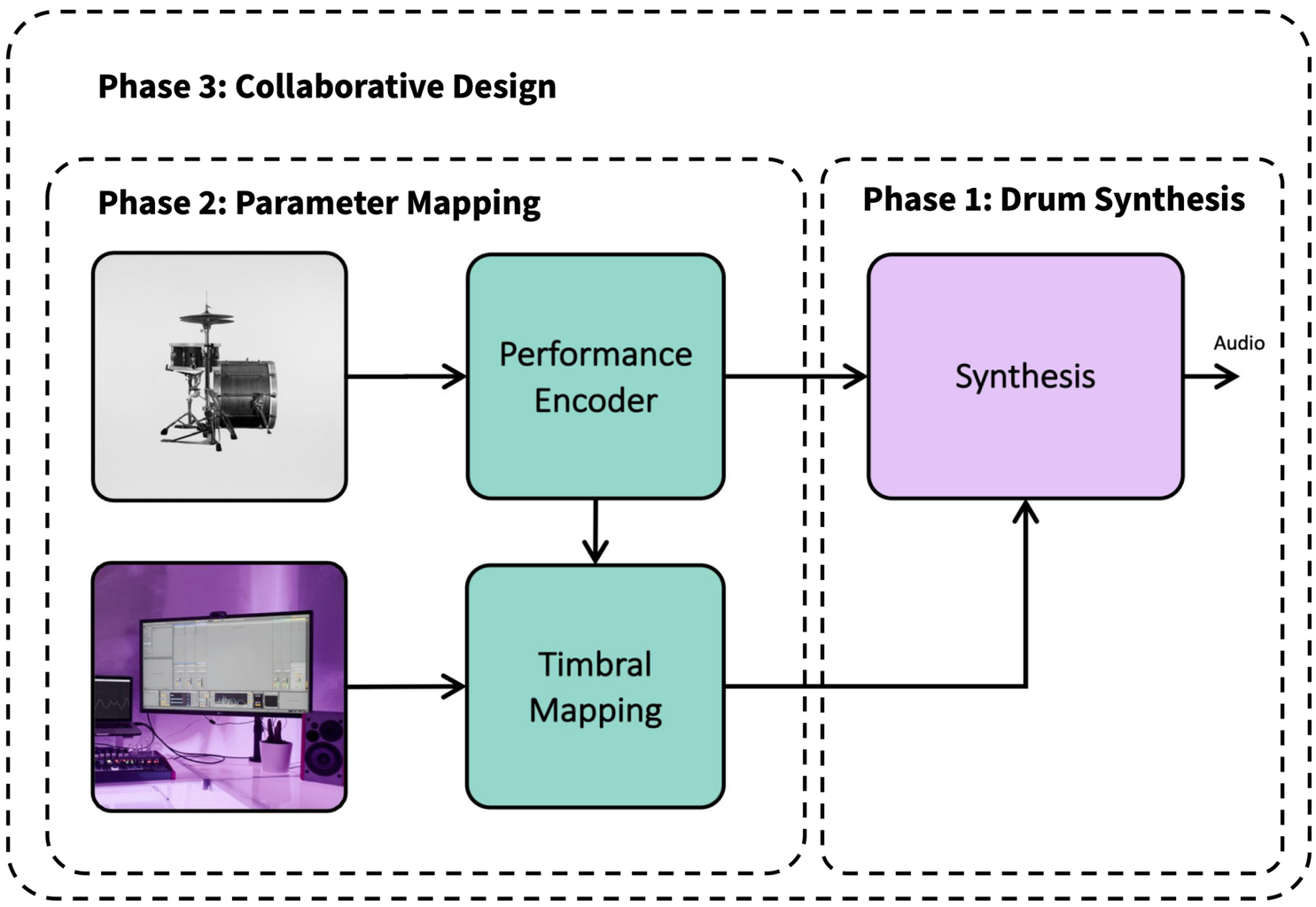
Supervisor(s): George Fazekas, Geraint Wiggins

AIM Summer School 2023

Real-time Timbral Mapping for Synthesized Percussive Performance

Jordie Shier

2022



Finding

Q: How can differentiable digital signal processing (DDSP) be extended to support timbre transfer of non-harmonic percussive instrument tones? A: Two problems with DDSP for non-harmonic percussive tones are: frequency estimation using gradient descent and transients. We address these using hybrid architectures that augment DDSP with neural audio synthesis.

Question

What methodologies can we employ to extract salient performance features from percussive audio in real-time and dynamically map them to controls of a synthesizer to enable nuanced control over timbre?

Supervisor(s): Andrew McPherson, Charalampos Saitis, and Andrew Robertson

AIM Summer School 2023

Predicting hit songs: multimodal and data-driven approach

Kasia Adamska

2021

Finding

Audio properties and attributes cannot explain music appeal on their own. So far, I've discovered that including indicators of public interest in a song or an artist can help predict the success of a music track. I recently used Google Trends data in conjunction with audio features to forecast the outcomes of the Eurovision Song Contest.

Question

A song must capture the attention of the audience in order to become a hit with the public. What qualities does a song need to have to accomplish this? I am interested in developing a deeper understanding of musical and lyrics features that are more descriptive of conveyed sentiments, evoked emotions and 'catchiness'.

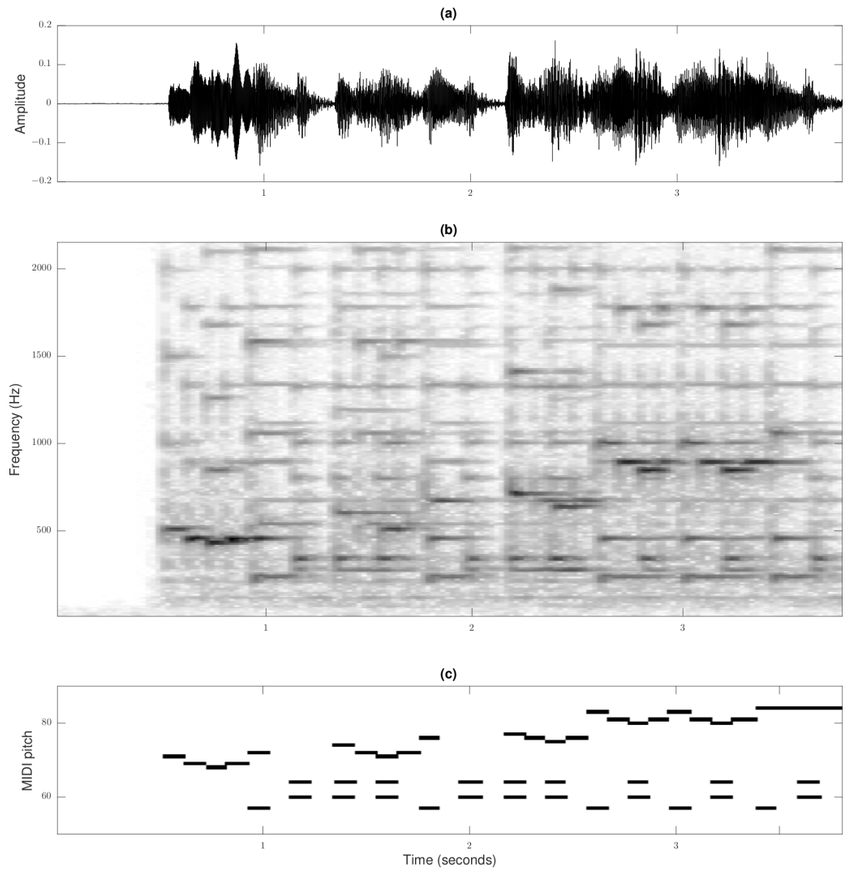
Supervisor(s): Joshua Reiss

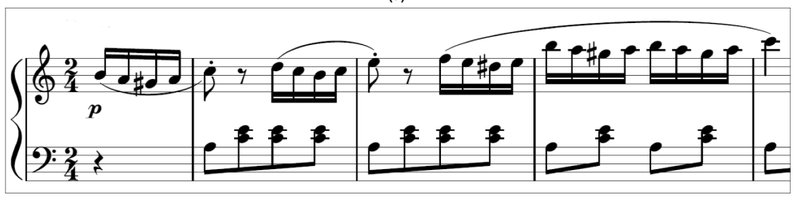
AIM Summer School 2023

Automatic Audio-to-Score Piano Transcription with Deep Neural Networks

Lele Liu

2019





Finding

A2S Transcription converts a music recording into a readable musical score format. In my research, I explored deep learning methods including the use of sequence-to-sequence models, the attention mechanism, and convolutional-recurrent neural networks applied for holistic and pipeline-based piano transcription.

Question

How do we develop better systems for cross-instrument music transcription? How can we avoid using large datasets in transcription systems?

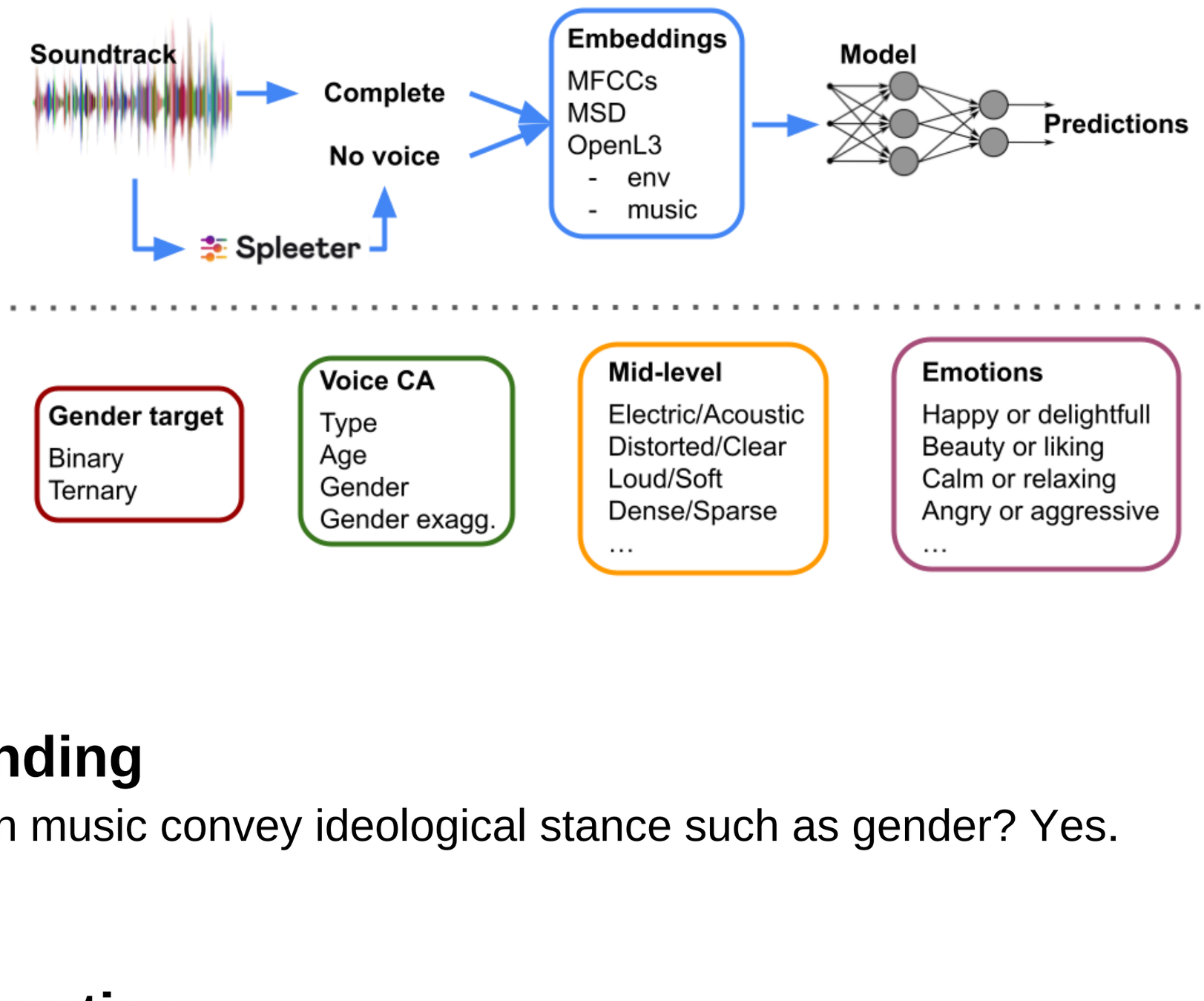
Supervisor(s): Emmanouil Benetos, Veronica Morfi and Simon Dixon

AIM Summer School 2023

Gender-coded sound: A multimodal analysis of gender encoding strategies in music for advertising

Luca Marinelli

2020



Finding

Can music convey ideological stance such as gender? Yes.

Question

Multimodal fusion techniques in the context of transfer learning

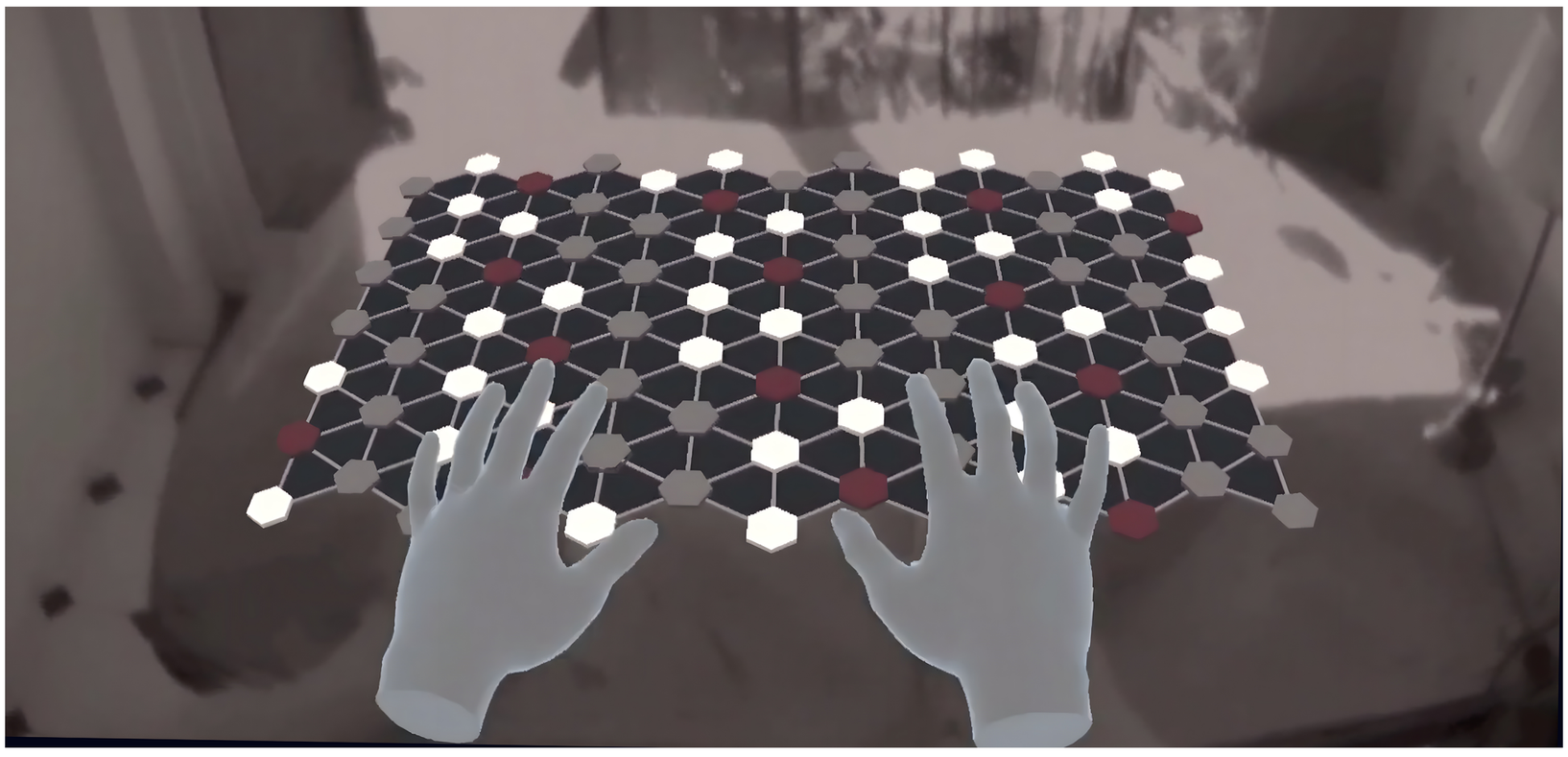
Supervisor(s): Charalampos Saitis

AIM Summer School 2023

Towards a Surface-Based Extended Reality Musical Instrument for Keyboardists

Max Graf

2020



Finding

Hand tracking is one of several key factors for embodied performance with an extended reality musical instrument (XRMI). Current vision-based hand tracking systems that are integrated with head-mounted XR devices are error-prone and have several failure cases in the context of XRMI performance.

Question

How can we overcome the limitations of vision-based tracking approaches? We have investigated the use of surface electromyography (sEMG) wearables + deep learning for articulated finger tracking without relying on cameras. The first results are promising, however, more work is needed for fully articulated hand tracking based solely on sEMG sensors.

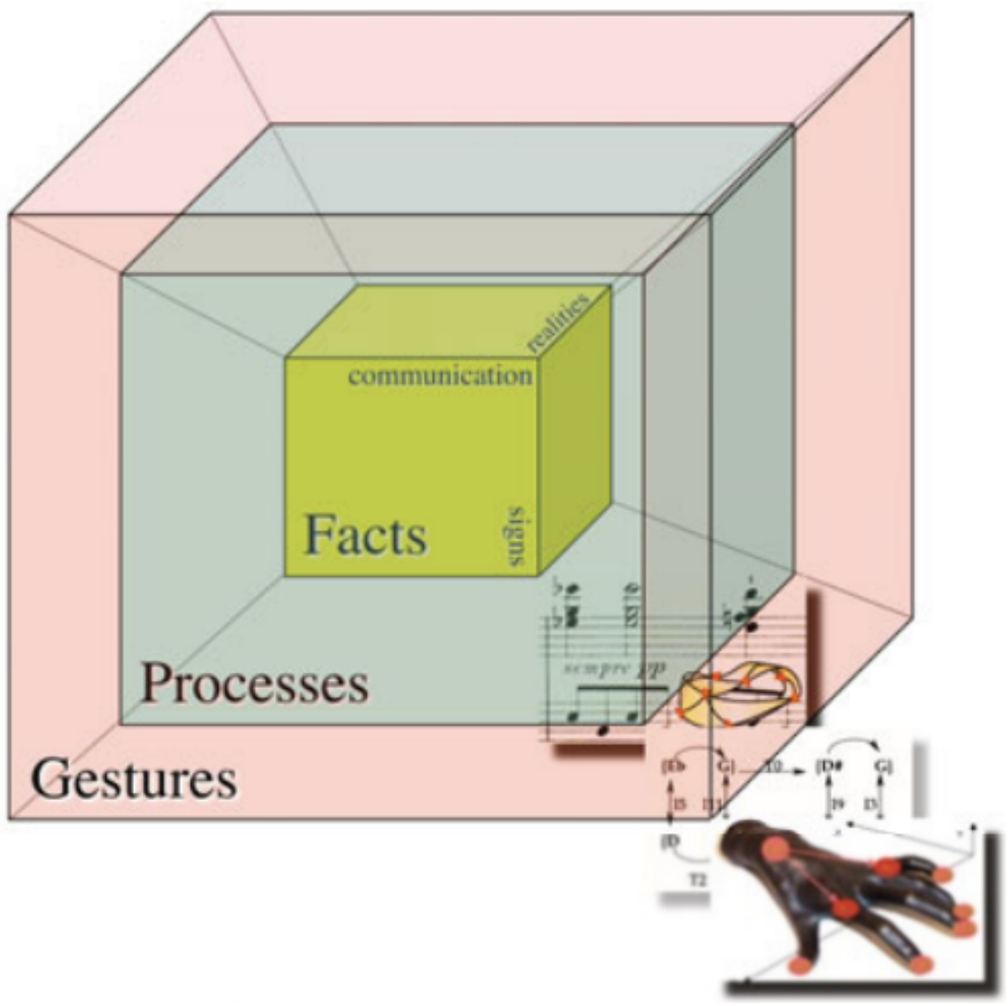
Supervisor(s): Mathieu Barthet, Andrew McPherson

AIM Summer School 2023

Embodied cognition for intellignent musical systems

Oluremi Falowo

2021



Finding

Can corporeal data improve the emotion representation in MER systems? Due to there not being a suitable dataset, my initial experiments have focused on the task of congruency detection between music audio features and dance features (labanotation). It was found that even with relatively small networks the model could detect congruency with %79 precision

Question

Can Graph NNs provide better feature extraction for skeletal postioning data as opposed to hand crafted features (labanotation)?

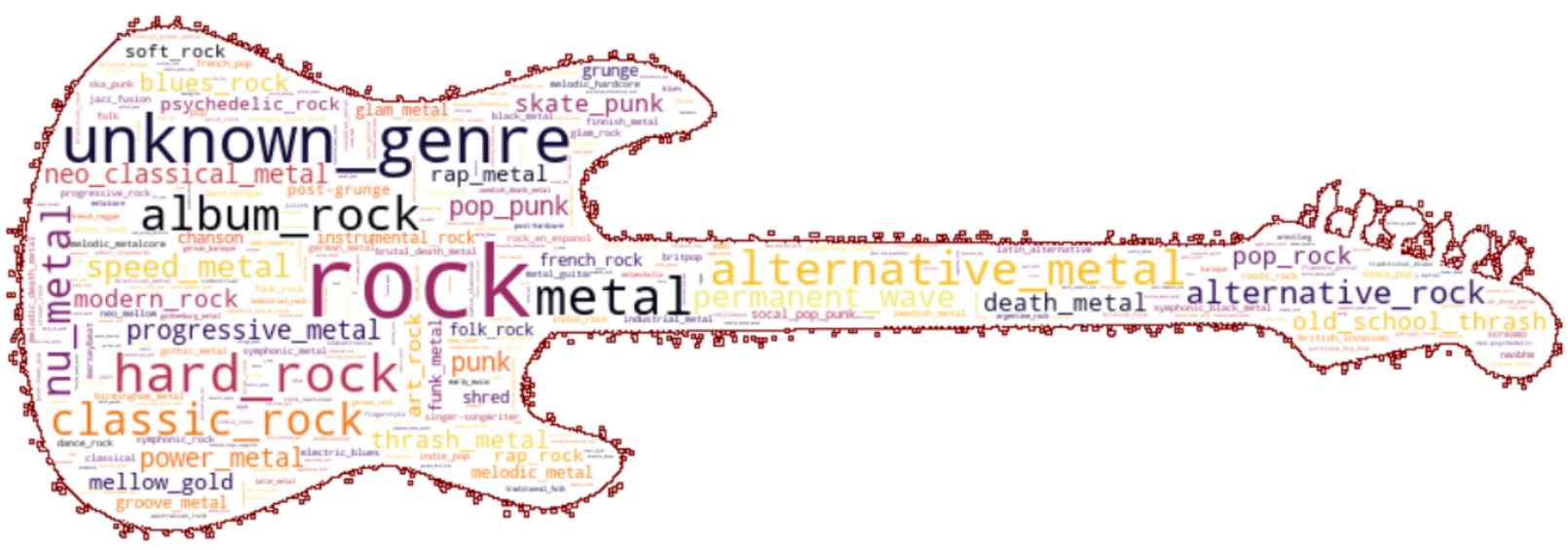
Supervisor(s): charalampos saitis

AIM Summer School 2023

Guitar Tablature Generation with Deep Learning

Pedro Pereira Sarmento

2019



Finding

We've showed that it is possible to automatically generate symbolic music that encompasses prescriptive information useful for guitar players, namely how and where to play specific musical phrases on the instrument.

Question

Although not entirely linked to the findings in my PhD, I'd like to tackle automatic guitar transcription, by leveraging the token format proposed in our DadaGP dataset.

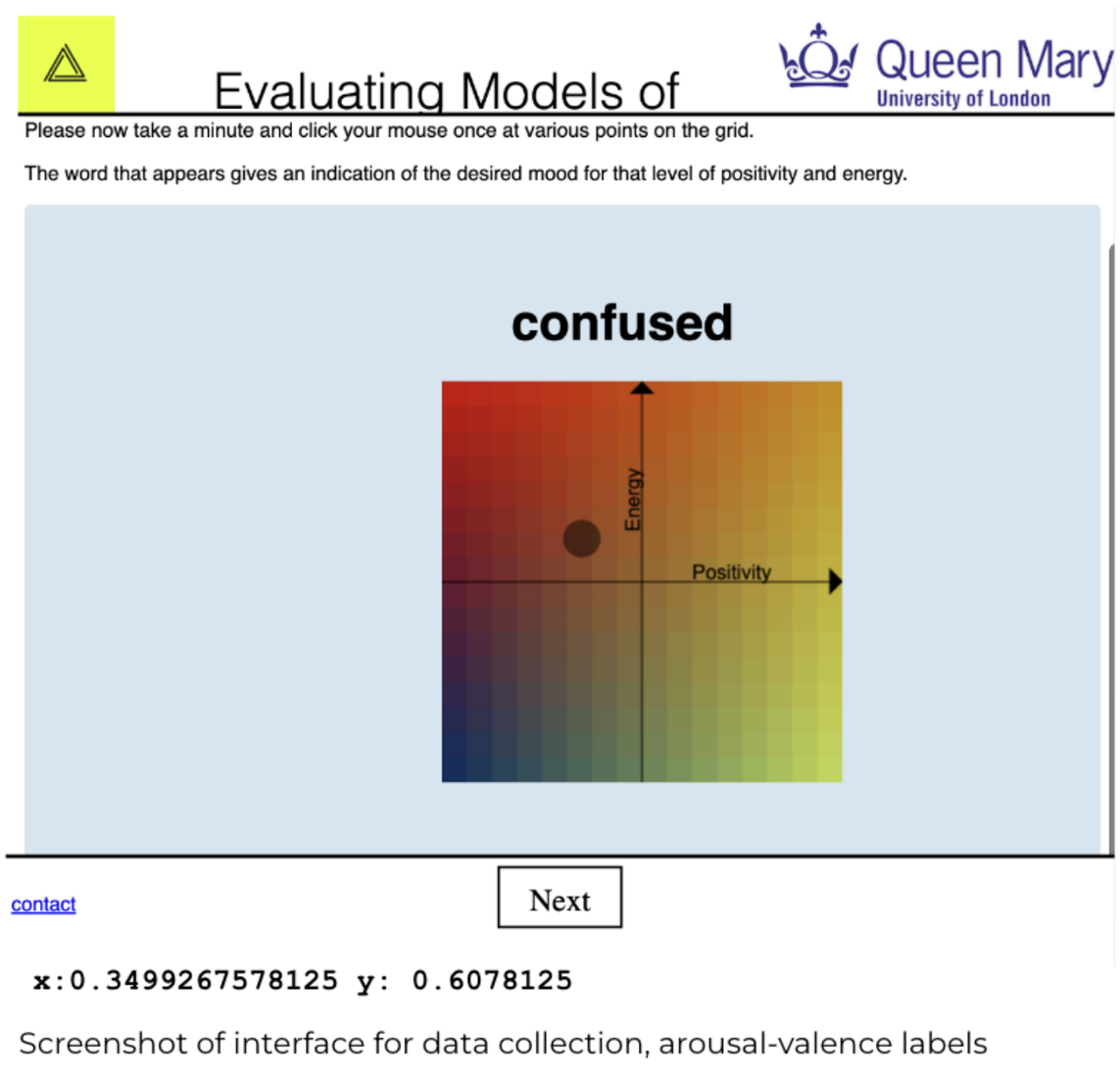
Supervisor(s): Dr. Dorien Herremans, Dr. Mathieu Barthet

AIM Summer School 2023

Continuous Mood Recognition in Film Music

Ruby Crocker

2021



Screenshot of interface for data collection, arousal-valence labels

Finding

Using CNN-type models to predict arousal and valence values accurately for a music and emotion regression task. VGG-model performed well, Arousal was easier to predict than valence, and currently working on creating a modern film music dataset with arousal valence ratings.

Question

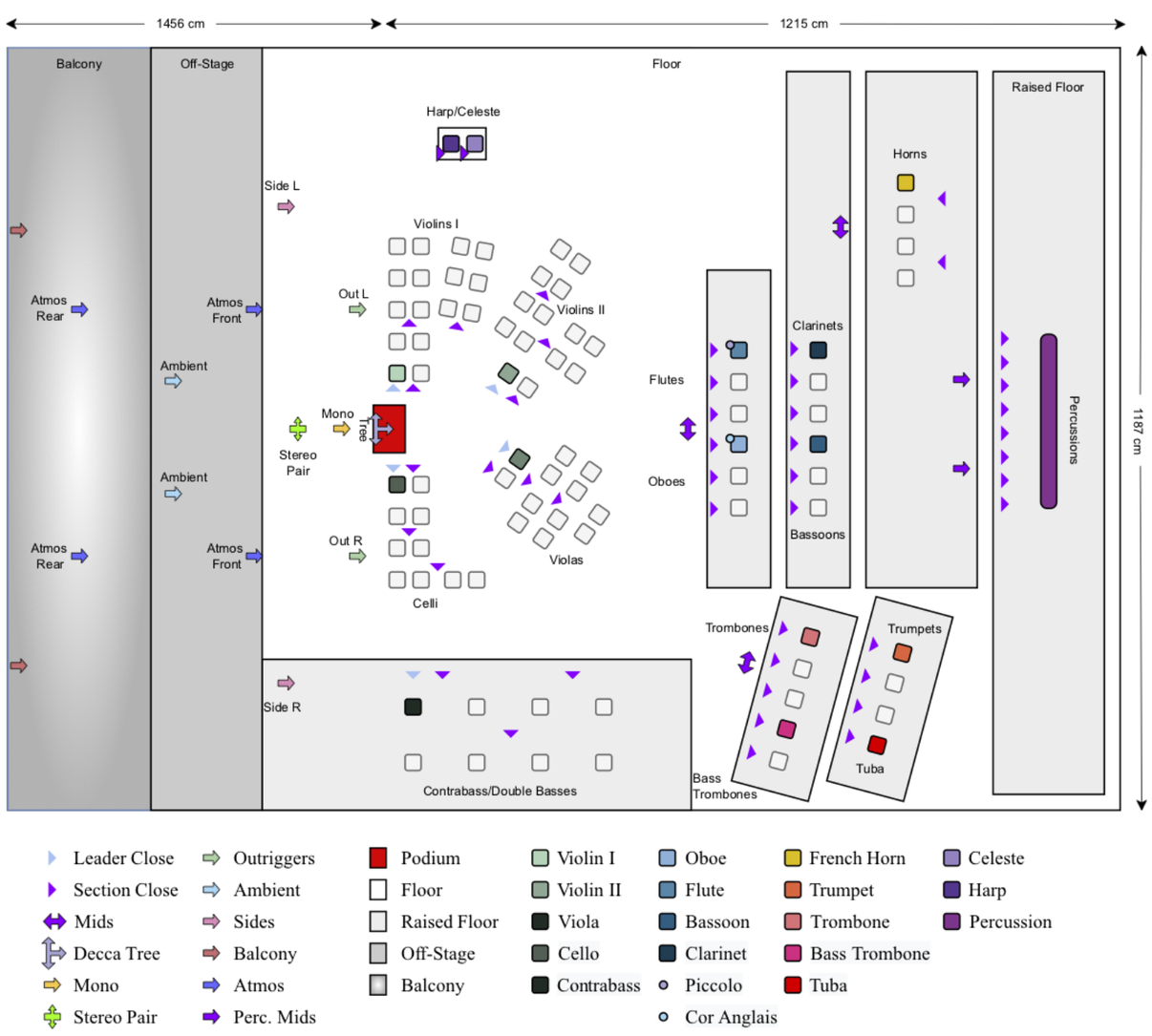
How do composers manipulate emotion? What features/techniques are used to interpret specific mood/emotion in film? How does film music (score) compare with mainstream music in evoking emotion perceptually? Creating the film music dataset

Supervisor(s): George Fazekas

AIM Summer School 2023

Music Ensemble Separation

Saurjya Sarkar
2019



Finding

Permutation invariant training not only enables separating mixtures of identical instruments in a label agnostic fashion, but also allows separation of unseen instruments with the assumption that each source is monophonic.

Question

How do you define source separation? Ideally you should be able to separate any sonic event in an audio recording. If we assume each monophonic event as a source, could we ideally have universal source separation? How do we train a model with variable number of output nodes that are dynamically activated based on the level of polyphony in the mixture?

Supervisor(s): Emmanouil Benetos, Mark Sandler

AIM Summer School 2023

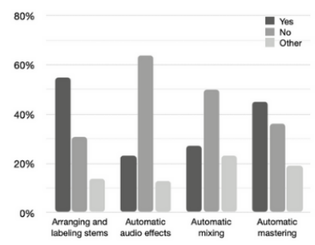
Music Mixing Style Transfer - Informed by Professional Practice

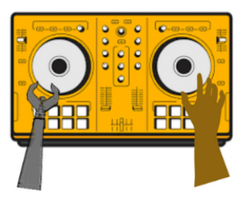
Soumya Sai Vanka

2021

Music Mixing
- Technical and creative considerations
- Highly context-dependent

Context: Intention for the song's sonic sound and emotive quality



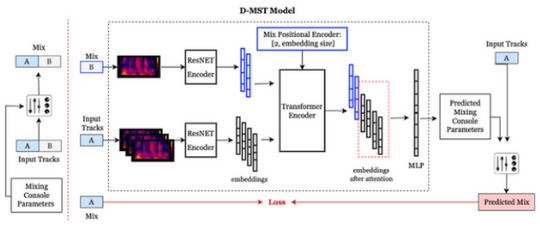


Professionals want co-creative systems that promote holistic interaction

Inspire design of smart mixing systems based on the collaborative practices in studios

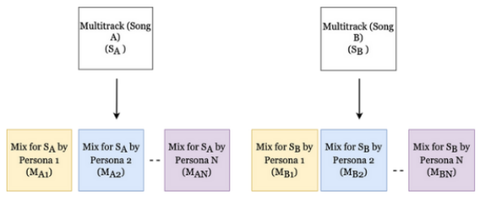
Mixing Style Transfer using Reference Song

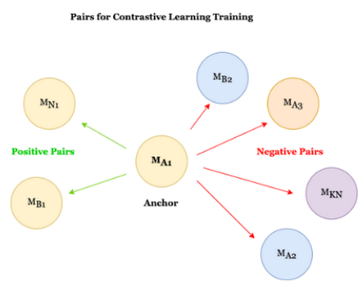
Current Work: Differentiable Mixing Style Transfer

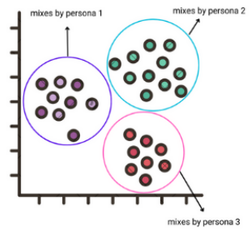


Predicts mixing console parameters for generating a mix given raw tracks and reference mix

Future Work: Self-Supervised learning of Mixing Style Latent Space







- Dataset: Multiple mixes of the same song in the style of different mixing engineers.
- Self-supervised learning to learn mixing styles: Cluster different songs with similar mixing styles together.

Finding

The targeted user group influences automatic mixing system design. Professionals prefer collaborative, controllable systems. Mixing style transfer can be inspired by collaborative practices in studio. Transformer encoder-based transfer learns gain and pan settings from reference song, but advanced audio effects require further investigation.

Question

Reference songs are often used as a medium of tacit agreement for the vision of the mix (mixing style for the song) between the artist and the mixing engineer. In the upcoming work, we want to investigate if it would be possible to disassociate mixing styles from the content of the song. We also want to ask what comprises the definition of a mixing style.

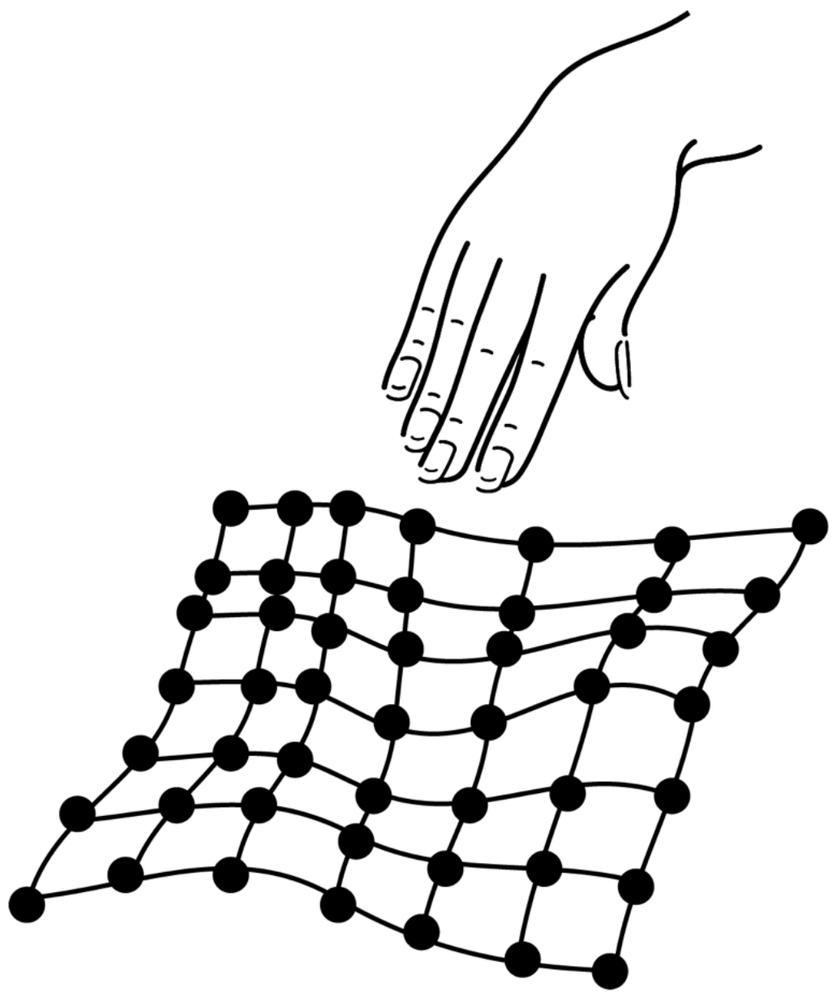
Supervisor(s): George Fazekas, Jean-Baptiste Rolland, Mathieu Barthet

AIM Summer School 2023

Embedding neural networks in low-powered devices for musical practice

Teresa Pelinski

2021



Finding

How can we incorporate embedded deep learning workflows (dataset collection, model training, inference) into existent workflows with embedded systems? (in progress) --> answer: https://github.com/pelinski/bela-dl-pipeline

Question

How can we incorporate embedded deep learning workflows (dataset collection, model training, inference) into existent workflows with embedded systems? What makes a tool (How does a tool become) suitable for prototyping and experimenting?How do the computational limitations of low-powered embedded systems impact musical practice with deep learning models?

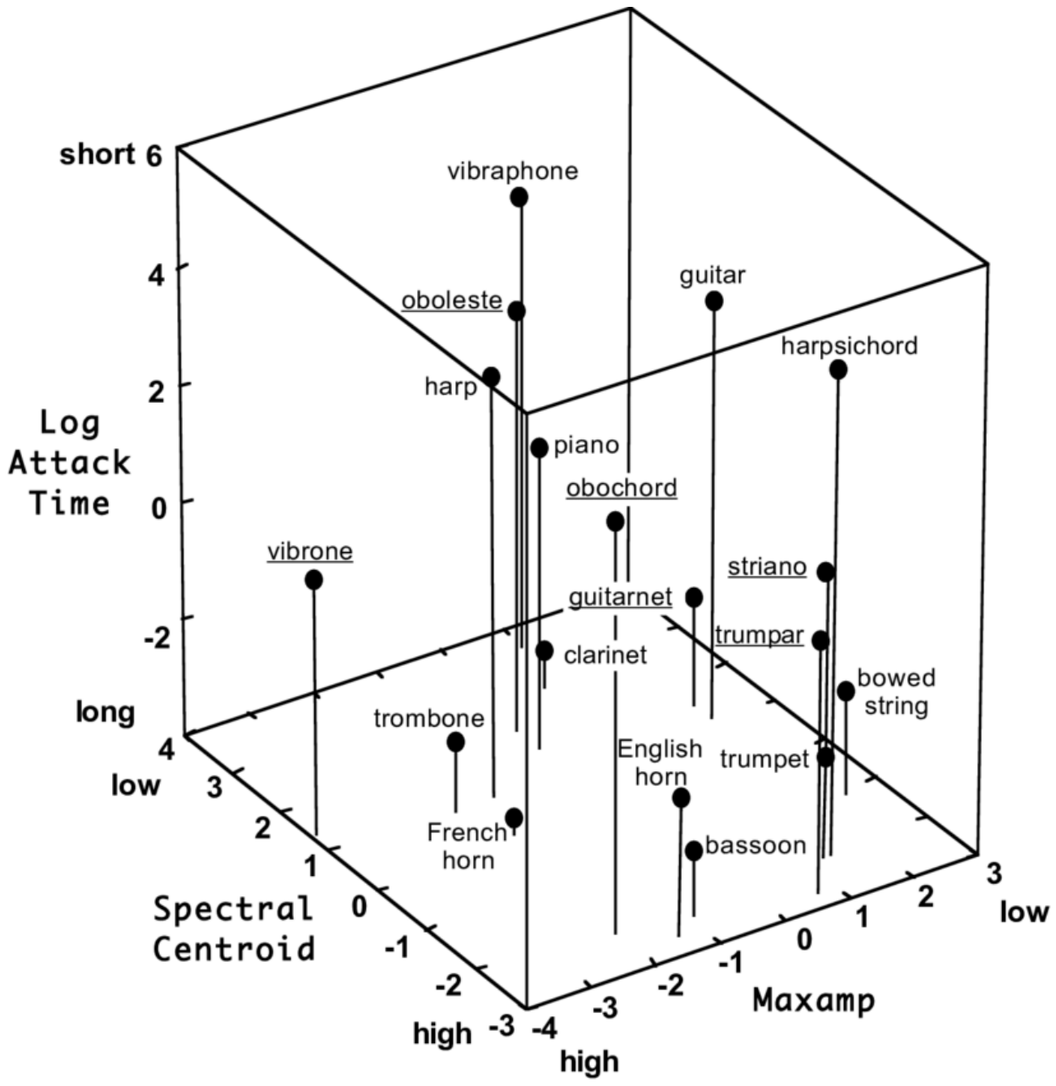
Supervisor(s): Prof Andrew McPherson, Prof Rebecca Fiebrink

AIM Summer School 2023

Expressive Performance Rendering for Music Generation Systems

Tyler Howard McIntosh

2022



Finding

Performance data is heavily concentrated in single musical contexts, such as classical piano, which makes modelling performance in general terms more challenging. A potential approach to this problem lies in domain adaptation, which may be used to learn performance on other instruments from the perspective of piano.

Question

A general model of performance may be required to generate performances for any number of instruments, and performances should form a cohesive whole when combined. This requires a system that is unbounded in the number of input sequences, and is able to facilitate some form of communication or high-level planning between sequences in performance generation.

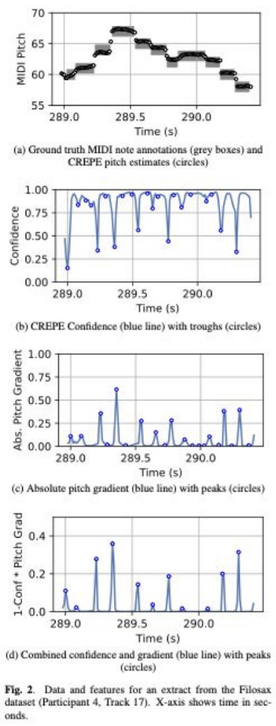
Supervisor(s): Simon Dixon, Doon MacDonald, Nadine Kroher, Aggelos Pikrakis

AIM Summer School 2023

Transcribing the Jazz Ensemble - towards automatic transcription of small jazz groups

Xavier Riley
2020

Transcribing the Jazz Ensemble
Xavier Riley & Simon Dixon



(a) Ground truth MIDI note annotations (grey boxes) and CREPE pitch estimates (circles)

(b) CREPE Confidence (blue line) with troughs (circles)

(c) Absolute pitch gradient (blue line) with peaks (circles)

(d) Combined confidence and gradient (blue line) with peaks (circles)

Fig. 2. Data and features for an extract from the Filosax dataset (Participant 4, Track 17). X-axis shows time in seconds.

Work under review (June 2023)

CREPE Notes:
monophonic note segmentation

	CNt	CN	PYIN	BP	MT3
Recall	88.26	**88.61**	50.32	80.62	40.67
Precision	**77.18**	76.91	69.50	71.18	45.78
F-measure	**82.31**	**82.31**	58.28	75.54	42.97
Overlap	88.54	**89.91**	87.36	83.45	72.96
Parameters	0.5M	22M	N/A	17M	77M

Table 1. Results on the Filosax dataset. Mean scores are shown for each metric. Abbreviations are CNt (Crepe Notes "tiny" model, proposed), CN (Crepe Notes "full" model, proposed), PYIN (PYIN Notes), BP (Basic Pitch). Parameter counts for each model are shown for reference. For the proposed models we quote the size of the CREPE model which was used to provide the f0 and confidence estimates.

	CNt	CN	PYIN[5]	BP	MT3
Recall	**66.66**	65.79	36.58	55.56	23.87
Precision	66.73	**67.18**	64.83	64.92	28.35
F-measure	**66.58**	66.35	46.44	59.58	25.47
Overlap	79.96	80.53	**82.50**	77.33	69.02
Parameters	0.5M	22M	N/A	17M	77M

Table 2. Results on the ITM Flute 99 dataset, showing mean scores for each metric. Abbreviations are given in Table 1.

Beyond Piano - scaling transcription models through accurate polyphonic score alignment

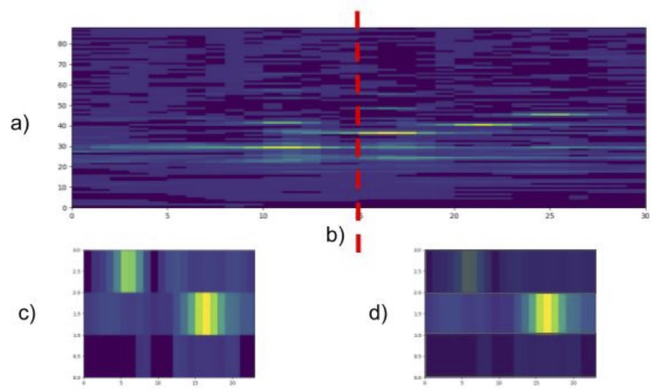


Figure 2: Aligning polyphonic scores to transcription model activations

	P_{50}	R_{50}	F_{50}	P_{25}	R_{25}	F_{25}
Basic Pitch [23]*	67.26	87.52	75.29	63.62	82.94	71.27
Omnizart [28]*	63.11	67.41	63.55	51.44	55.92	52.23
MT3 [6]*	95.97	95.00	95.45	95.22	94.26	94.70
Kong et al. [2]	67.48	49.69	54.79	58.41	42.45	47.02
Kong et al. (augmented)	80.61	44.04	50.32	72.59	38.78	44.57
Our approach	85.51	88.58	86.75	77.36	80.12	78.49

Table 2 - Results of our trained model on guitarset (unseen). 86.75% accurate - within 9% of larger, overfitted models

Finding

We find that fine grained score alignment accurate enough to train music transcription models. Working with guitar, we trained a model (under review) which achieves SOTA zero shot performance on guitarset with as little as 25ms tolerance.

Question

Is it possible to combine source separation, transcription models and sheet music layout models to transcribe an entire jazz ensemble accurately enough for real consumers?

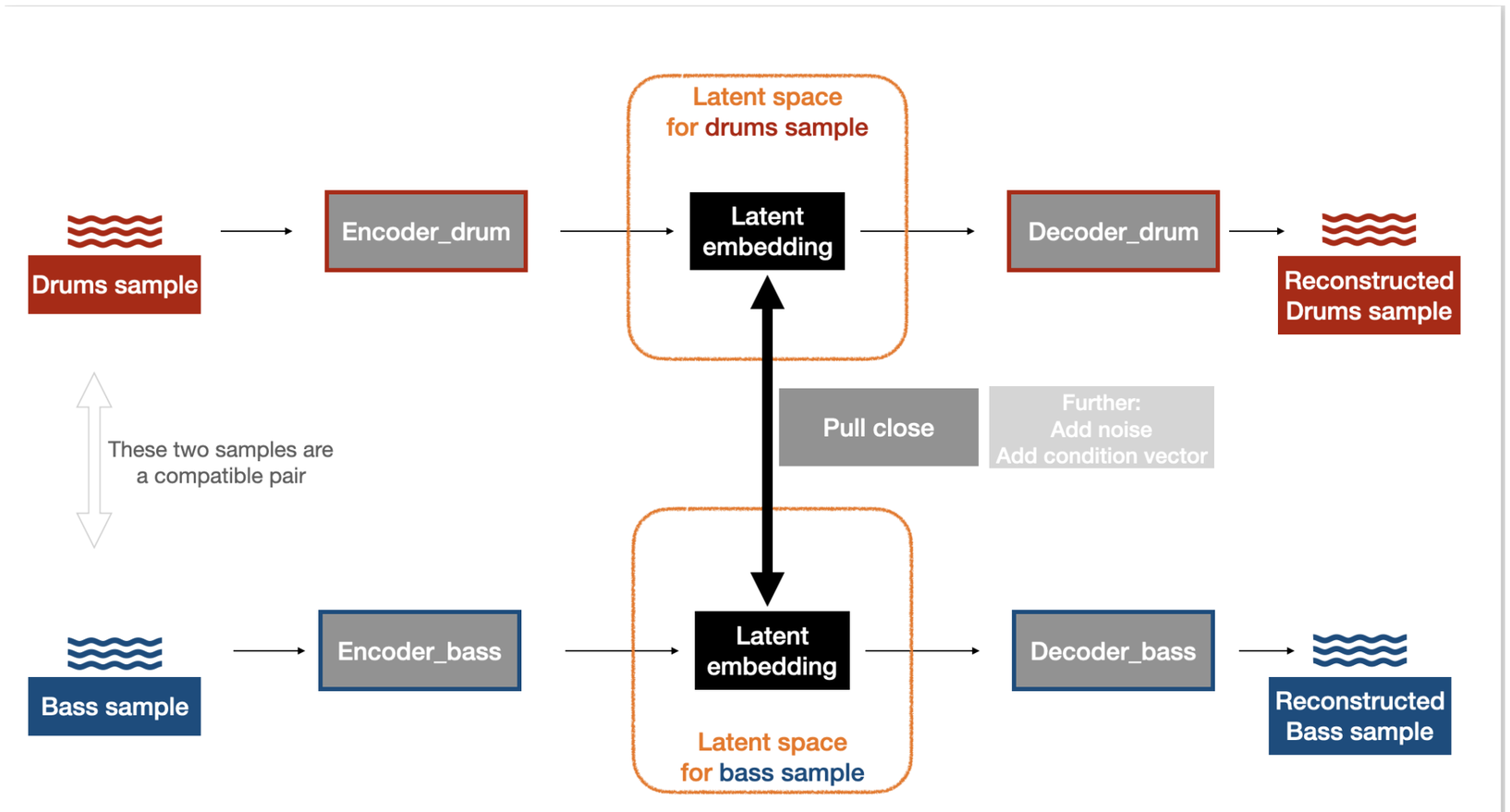
Supervisor(s): Simon Dixon

AIM Summer School 2023

A Generation-based Pairwise Sample Compatibility Modeling Framework

Xiaowan Yi

2021



Finding

We can possibly use one framework to do similarity-based and compatibility-based searching and generating for drums samples and bass samples. In our experiment, some generated bass samples adhered to the input drums sample’s rhythmic structure without being explicit conditioned on information like tempo or genre.

Question

How to incorporate condition information into this framework? For instance, how to condition the latent mapping module with genre classes, tempo, and semantic information?

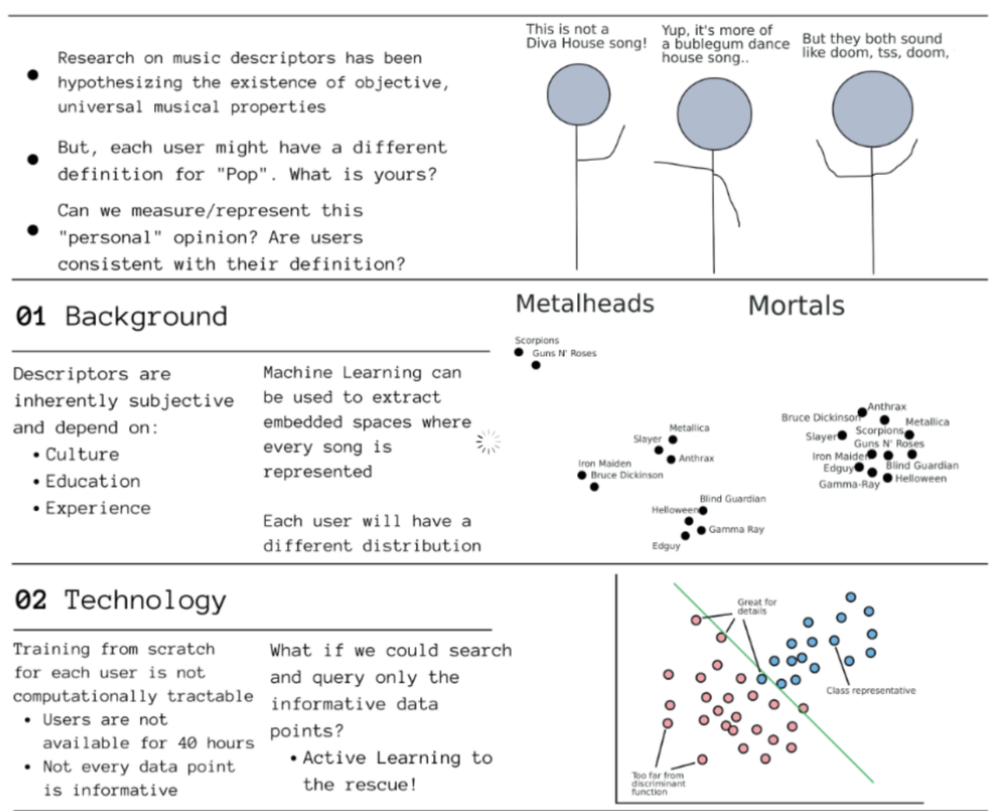
Supervisor(s): Mathieu Barthet

AIM Summer School 2023

Multi-modal User Adaptation for Automatic Music Tagging

Yannis Vasilakis

2022



Finding

Nothing

Question

What is the least amount of song annotations needed to define a tag for a specific user? Do tags (or music descriptors in general) mean the same thing for different users?

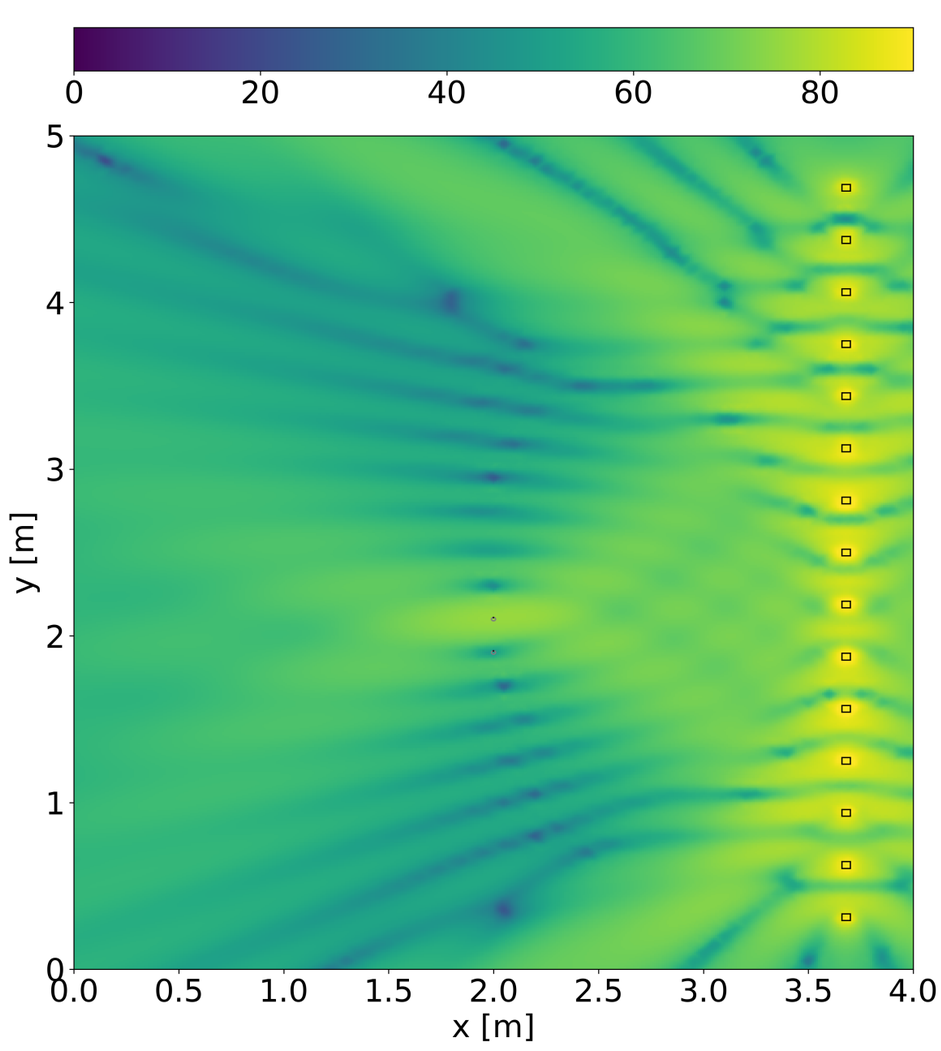
Supervisor(s): Johan Pauwels, Rachel Bittner

AIM Summer School 2023

Personal and Spatial Audio Reproduction using Loudspeaker Arrays for Home Applications

Yazhou Li

2021



Finding

Accurate room acoustic modeling plays an important role for designing a transaural reproduction system. I investigate the impact of mismatched room acoustic modeling on the performance of the system in reverberant environments and find that modeling the early reflections using the image source method can provide satisfactory binaural listening experience.

Question

To improve the performance in different reverberant rooms, room equalization may be utilized. Whether we can estimate room impulse responses at different positions accurately enough (for example, based on sound field interpolation and other data-driven methods) for room equalization and personal audio remains unclear.

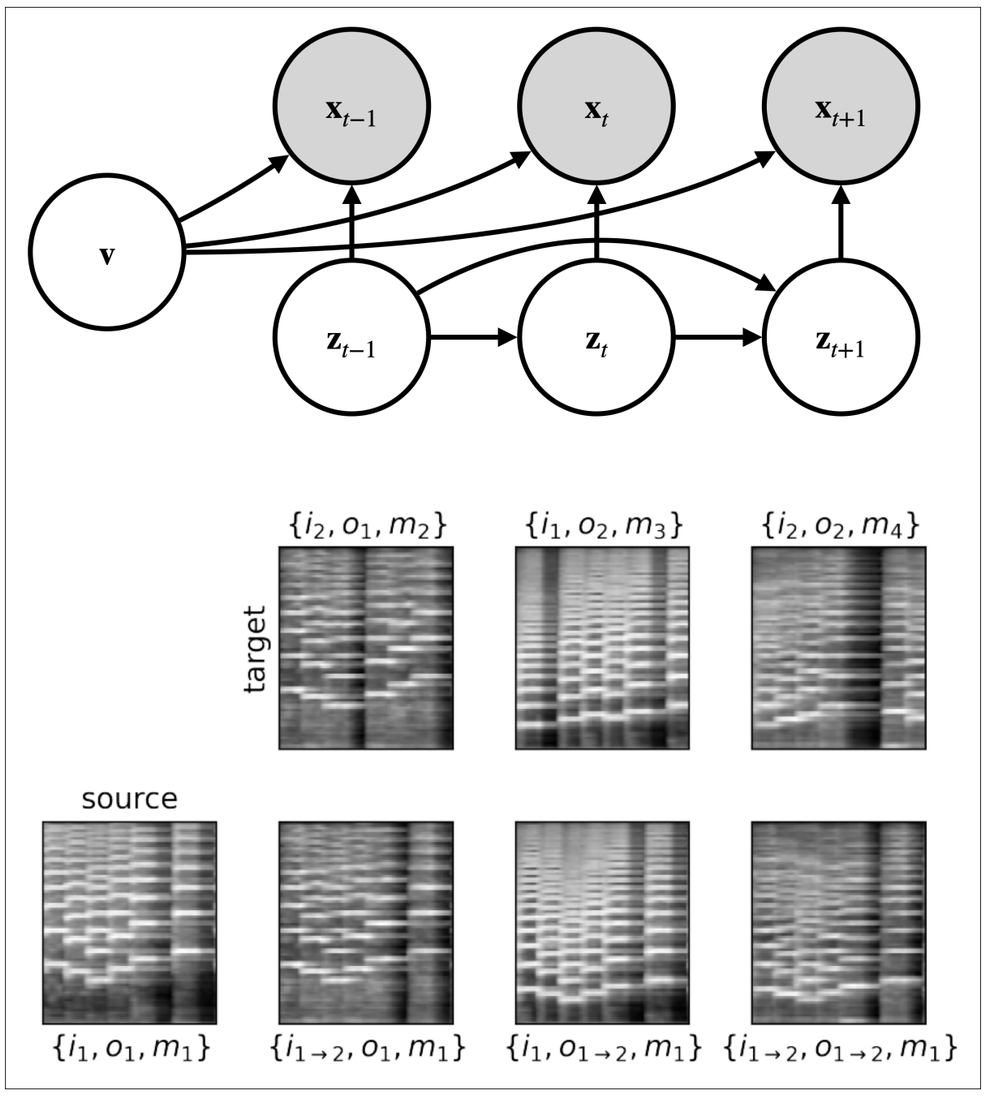
Supervisor(s): Joshua Reiss, Lin Wang

AIM Summer School 2023

Unsupervised Disentangled Representation Learning for Sequential Data

Yin-Jyun Luo

2020



Finding

How to learn semantically meaningful feature representations from sequential data without any supervision and domain-specific knowledge? With a balanced information bottleneck and an optimisation strategy that priorities features of lower capacity, a deep generative model can learn disentangled representations without annotations.

Question

It is common to sample a sequence given a sequence-level and a frame/segment-level latent variables. It can probably be seen both as a necessary inductive bias for the unsupervised learning, or a limited assumption that hinders the optimisation. How to strike a balance in this potential trade-off?

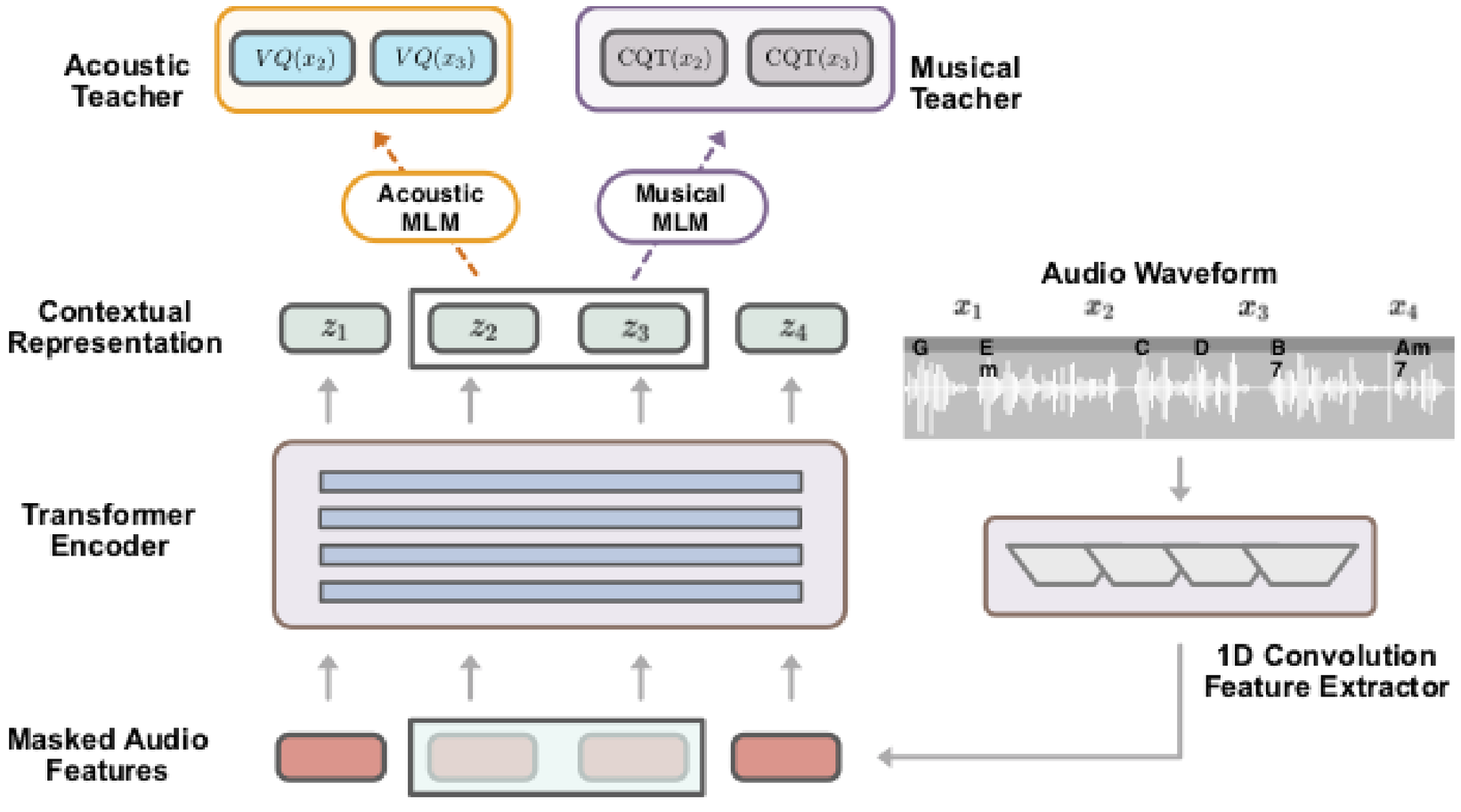
Supervisor(s): Sebastian Ewert and Simon Dixon

AIM Summer School 2023

Self-supervised Learning for Music Information Retrieval

Yinghao Ma

2022



Finding

We propose a Music undERstanding model with large-scale self-supervised Training (MERT), which uses a masked language modelling style for pre-training. We identified a superior combination of EnCodec feature prediction and CQT reconstruction. Results indicate that our BERT-style transformer encoder performs well on 14 MIR tasks, attaining SOTA overall score.

Question

Training a SSL model requires many music recording. One alternative is to use important high-quality data (e.g. English Wikipedia for NLP). Which kind of datasets is important for music? What is the MIR datasets for world music (Chinese music, Indian music etc.), which have some distribution shift from the training dataset (typically Western pop music)?

Supervisor(s): Emmanouil Benetos, Chris Donahue