UNIVERSITY OF MARYLAND

The Institute for Systems Research

LSP Laboratoire des Systèmes Perceptifs

icassp 2022 Singapore

MirrorNet : Learning Audio Synthesizer Controls Inspired by Sensorimotor Interaction

Yashish M. Siriwardena[1], Guilhem Marion[2], Shihab Shamma[1,2]

[1]Institute for Systems Research, University of Maryland, College park, USA

[2]Laboratoire des Systèmes Perceptifs, École Normale Supérieure, PSL University, France

Background

- Existence of bidirectional flow of interactions between the auditory and motor regions

- Learning complex sensorimotor mappings proceeds simultaneously and often in an unsupervised manner by listening and speaking all at once [1,2,3]

- Inspired by such learning of complex sensorimotor tasks, a new autoencoder architecture has been proposed to model this mechanism, and is referred to as the “Mirror Network” (or MirrorNet) by Shamma et al. [1]

- The essence of this biologically motivated algorithm is the bidirectional flow of interactions (‘forward’ and ‘inverse’ mappings) between the auditory and motor responsive regions, coupled to the constraints imposed simultaneously by the actual motor plant to be controlled.

- We used the the MirrorNet architecture to learn controls/parameters of a commercial and a widely available synthesizer (DIVA) in a completely unsupervised fashion

MirrorNet Model Architecture

- ***Goal of the model***: To learn two neural projections, an inverse mapping from auditory representation to motor parameters (Encoder) and a forward mapping from the motor parameters to the auditory representation (Decoder)
- Encoder and Decoder optimized simultaneously with two loss functions namely the 'encoder loss'(e_c) and the 'decoder loss'(e_d)
- The role of the 'forward' path is to back-propagate the error to learn the inverse mapping that is used to estimate the control parameters

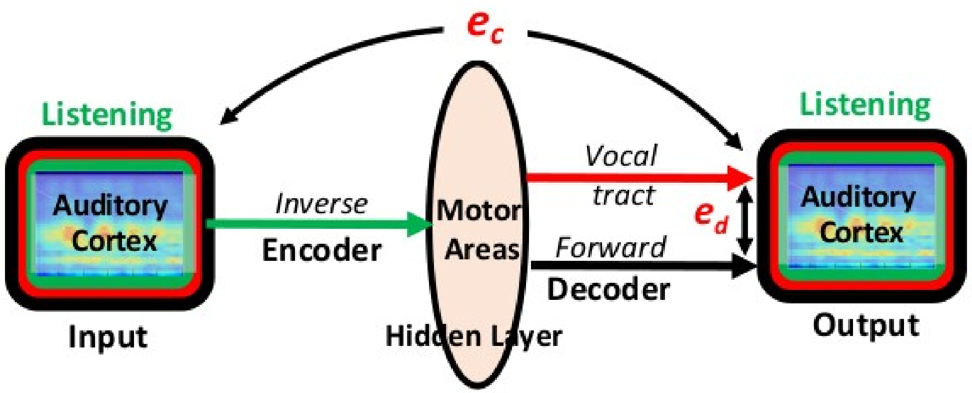


Figure 1: Autoencoder Architecture

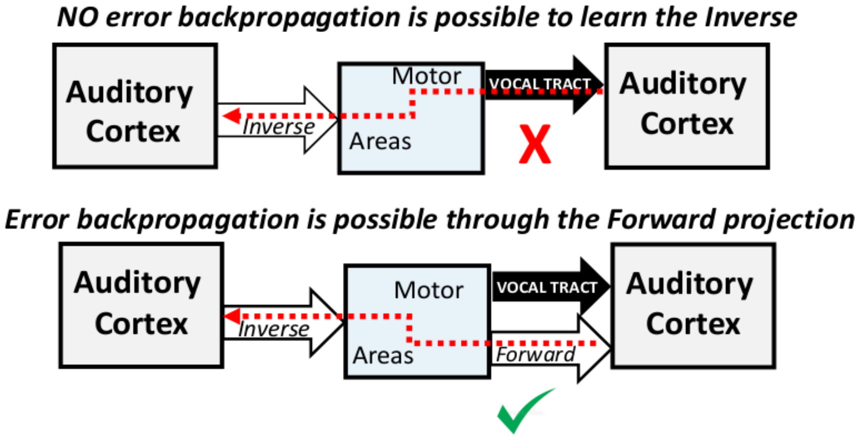
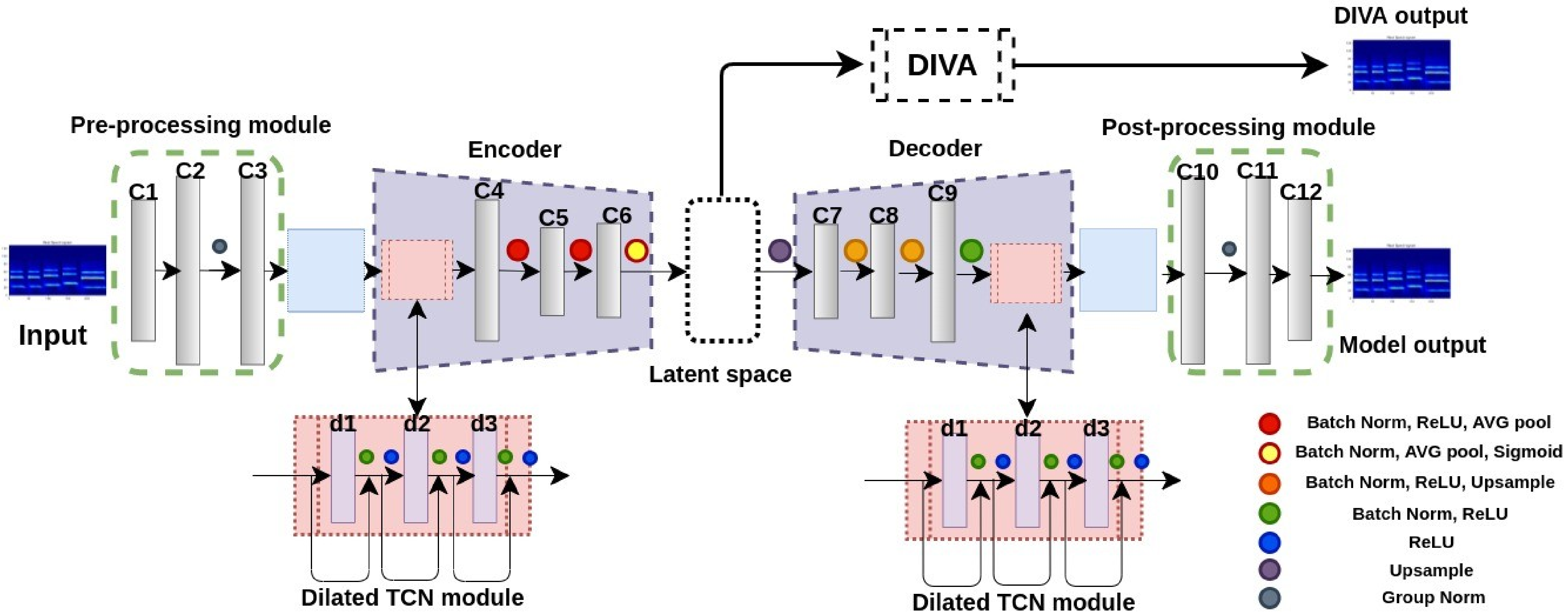


Figure 2: Role of the Forward Pass

Deep Neural Network (DNN) Architecture

- 1-D convolutional (CNN) layers modeling both the encoder and decoder. The complete network is inspired by the multi-layered Temporal Convolution Network (TCN) [4]



[4] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager, “Temporal convolutional networks for action segmentation and detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017

Learning Audio Synthesizer Controls

➢DIVA[1] synthesizer and List of Parameters

- DIVA, an off-the-shelf commercial synthesizer as the audio synthesizer for the MirrorNet model.
- 2 seconds long melodies with 5 notes and sampled at 44.1 kHz
- Continuous parameters normalized between [0,1]
- Table lists the set of parameters selected for the experiments and the corresponding parameter labels from DIVA where applicable
- The Encoder of the MirrorNet predicts the first 7 parameters in Table (shaded in yellow)

Table 1 : List of Parameters

Parameter Name	DIVA preset
MIDI note (Pitch)	-
MIDI duration	-
Volume	OSC : Volume2
Band pass filter(center frequency)	VCF1: Frequency
Filter Resonance	VCF1: Resonance
Envelope Attack	ENV1: Attack
Envelope Decay	ENV1: Decay
Vibrato Rate	LFO1: Rate
Vibrato Intensity	OSC : Vibrato
Vibrato Phase	LFO1: Phase

[1] https://u-he.com/products/diva/

Learning Audio Synthesizer Controls

- **Experiment 1:** Learning DIVA parameters for melodies synthesized with DIVA (set1)
  - 400 melodies to train the MirrorNet originally synthesized by DIVA using the first 7 parameters in Table 1
  - Availability of ground-truth parameters to assess the MirroNet predictions
  - Pitch, bandpass filter (center frequency), filter resonance and duration are predicted with significant accuracy where as volume and envelope attack parameters are predicted with comparatively lower accuracy

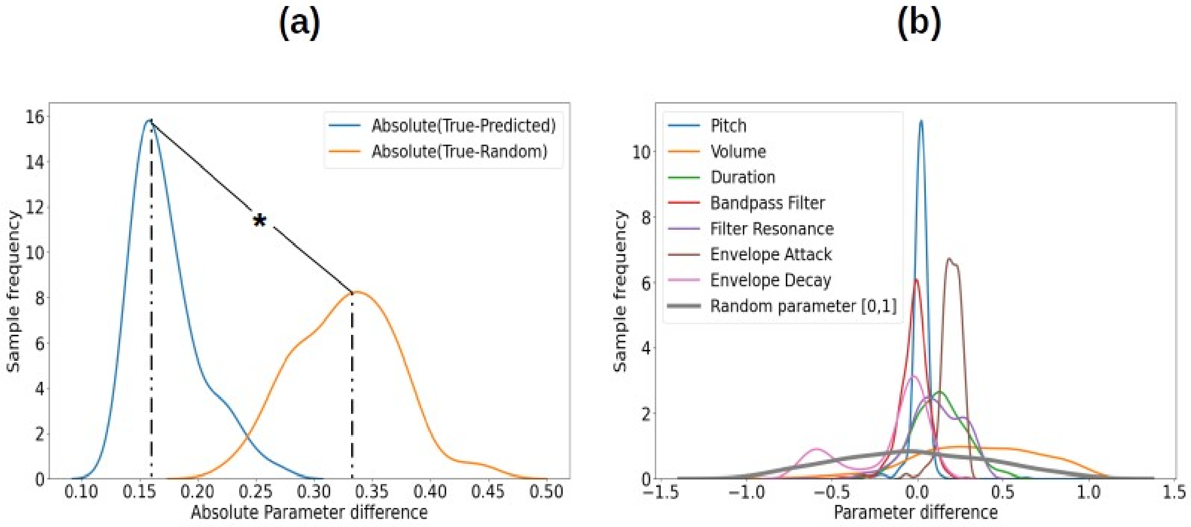
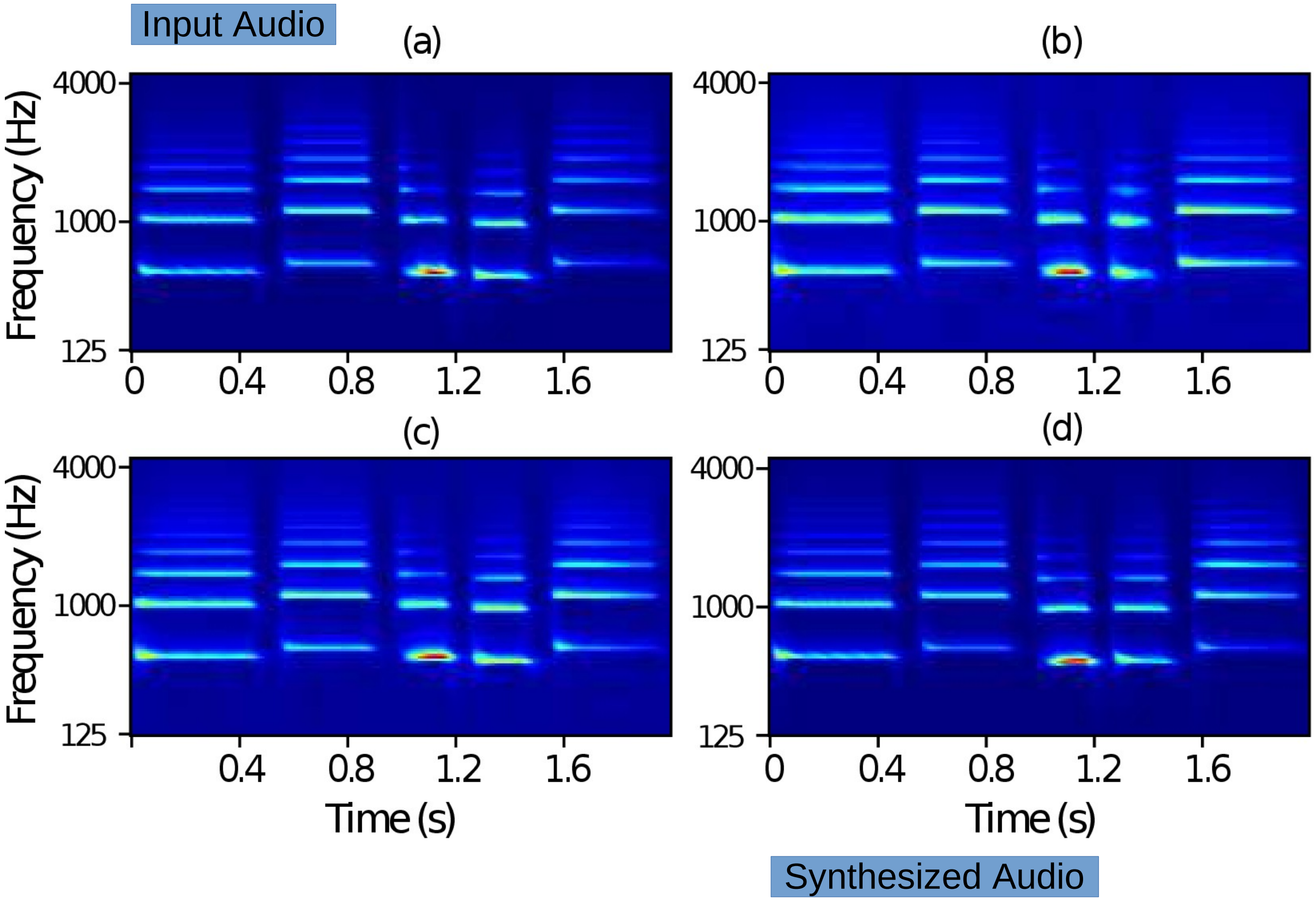


Figure : Evaluating statistical significance of the predicted DIVA parameters with respect to a set of random parameters on the test set (a) Distributions for absolute parameter differences across all parameters (b) Distributions of parameter differences (ground truth - predicted) for 7 parameters and the distribution for a random parameter difference (ground truth - random)

Learning Audio Synthesizer Controls

➢ **Experiment 1:** Auditory spectrograms

a) Input Melody
b) Decoder Output from ground-truth parameters
c) Final output from Decoder
d) DIVA output from learned control parameters

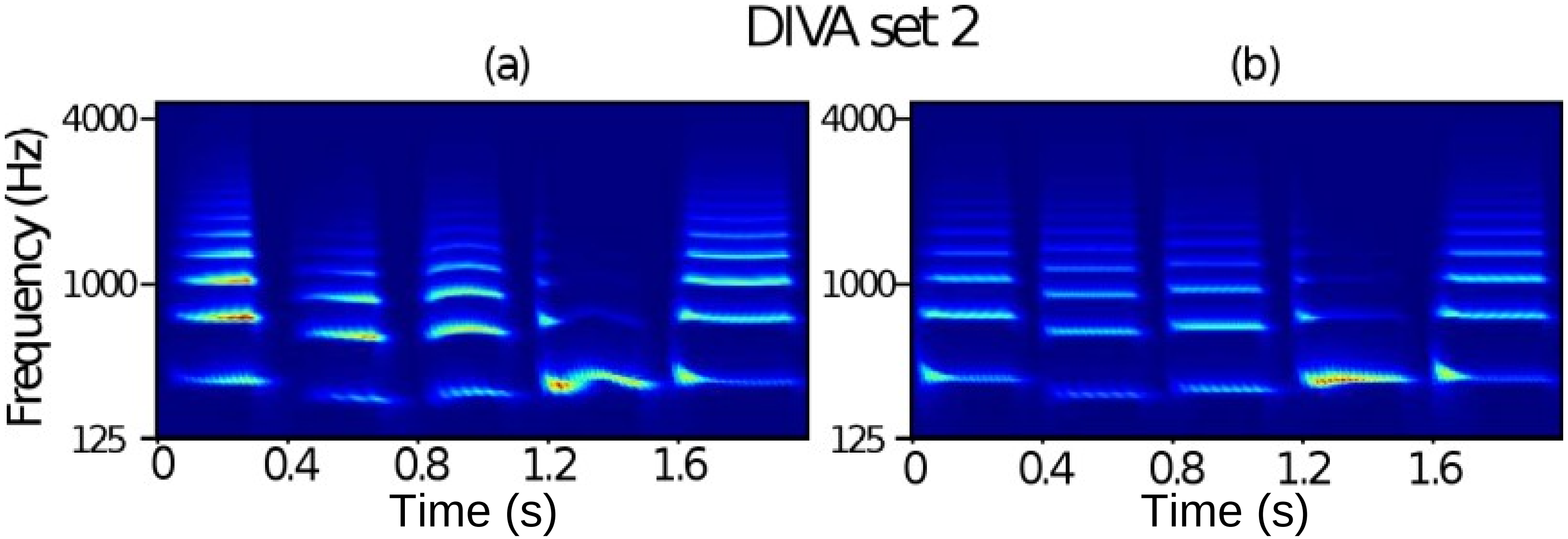


Audio samples for all experiments available at :
https://yashish92.github.io/MirrorNet-for-Audio-synthesizer-controls/

Learning Audio Synthesizer Controls

➢**Experiment 2:** Learning DIVA parameters for melodies synthesized with extra unknown DIVA parameters (set 2)

- 400 melodies to train the MirrorNet originally synthesized by DIVA using all the 10 parameters in Table 1
- MirrorNet is still trained to predict 7 parameters as in previous experiment
- Evaluates how the well MirrorNet can approximate the input melodies even if they have additional sound/musical qualities, eg. Vibrato

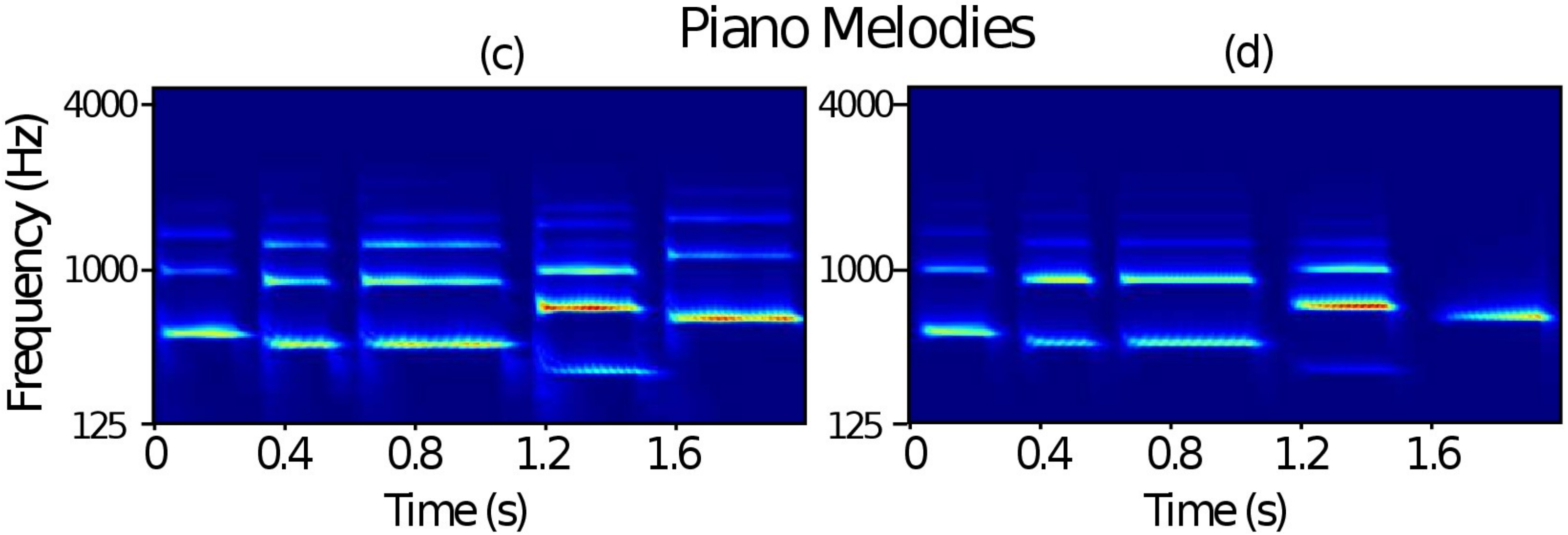


Auditory Spectrograms

a) Input Melody
b) DIVA output from learned control parameters

Learning Audio Synthesizer Controls

- **Experiment 3:** Learning DIVA parameters to synthesize melodies generated from a different syntheszier
  - Fundamental advantage of the MirrorNet is its ability to discover the DIVA parameters corresponding to music generated by other sources and synthesizers
  - 400 5-notes long piano melodies of 2 seconds that are synthesized by a Fender Rhodes digital imitation (Neo-Soul Keys generated trough Kontakt 5)
  - Trained Network successfully reproduces accurate renditions of the piano music from unseen samples



Auditory Spectrograms
c) Input Melody
d) DIVA output from learned control parameters

Summary

- Bidirectional sensorimotor projections enable **unsupervised learning** of vocal tract controls

- An autoencoder architecture with a constrained latent space can be used to simulate the sensorimotor learning algorithm to learn the required ‘inverse’ and ‘forward’ mappings

- The same sensorimotor learning algorithm (MirrorNet) can be generalized beyond speech synthesis

- MirrorNet can accurately estimate control parameters for an off-the-shelf audio/music synthesizer to synthesize a given input melody of notes
    - Learning audio synthesizer controls to synthesize an input melody of notes originally synthesized by the same set of parameters
    - Approximating an input melody of notes with different sound qualities (or synthesized by a different synthesizer) using a limited set of parameters of a given synthesizer

Ongoing and Future Work

- Extending the latent space to predict more parameters to capture richer aspects of sound
- Deploy more advanced and richer representations of sound beyond spectrograms
- Devise more efficient and faster training paradigms
- Synthesis of continuous musical melodies which can have a variable number of notes
- Incorporating a vocal tract model to synthesize a given speech utterance by estimating its parameters

References

1] Shihab Shamma, Prachi Patel, Shoutik Mukherjee, Guilhem Marion, Bahar Khalighinejad, Cong Han, Jose Herrero, Stephan Bickel, Ashesh Mehta, and Nima Mesgarani, “Learning Speech Production and Perception through Sensorimotor Interactions,” Cerebral Cortex Communications, vol. 2, no. 1,2020.

[2] Silvia Pagliarini, Arthur Leblois, and Xavier Hinaut, “Canary Vocal Sensorimotor Model with RNN Decoder and Low-dimensional GAN Generator,” in 2021 IEEE International Conference on Development and Learning (ICDL), 2021, pp.1–8

[3] Patricia K. Kuhl, “Early language acquisition: cracking the speech code,” Nature Reviews Neuroscience, vol. 5, pp. 831–843, 2004.

[4] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager, “Temporal convolutional networks for action segmentation and detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017