MirrorNet : Learning Audio Synthesizer Controls Inspired by Sensorimotor Interactions

icassp 2022 Singapore

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

University of Maryland · The Institute for Systems Research · Laboratoire des Systèmes Perceptifs

Paper #4418

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

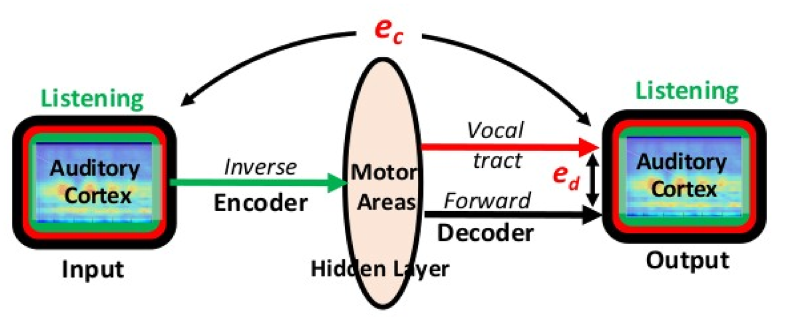


Figure 1: Autoencoder Architecture

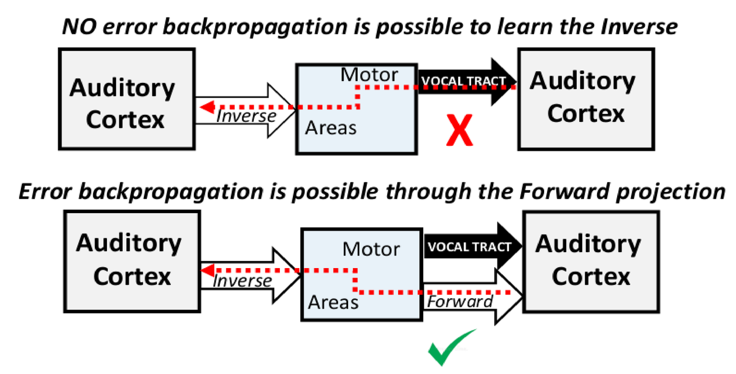
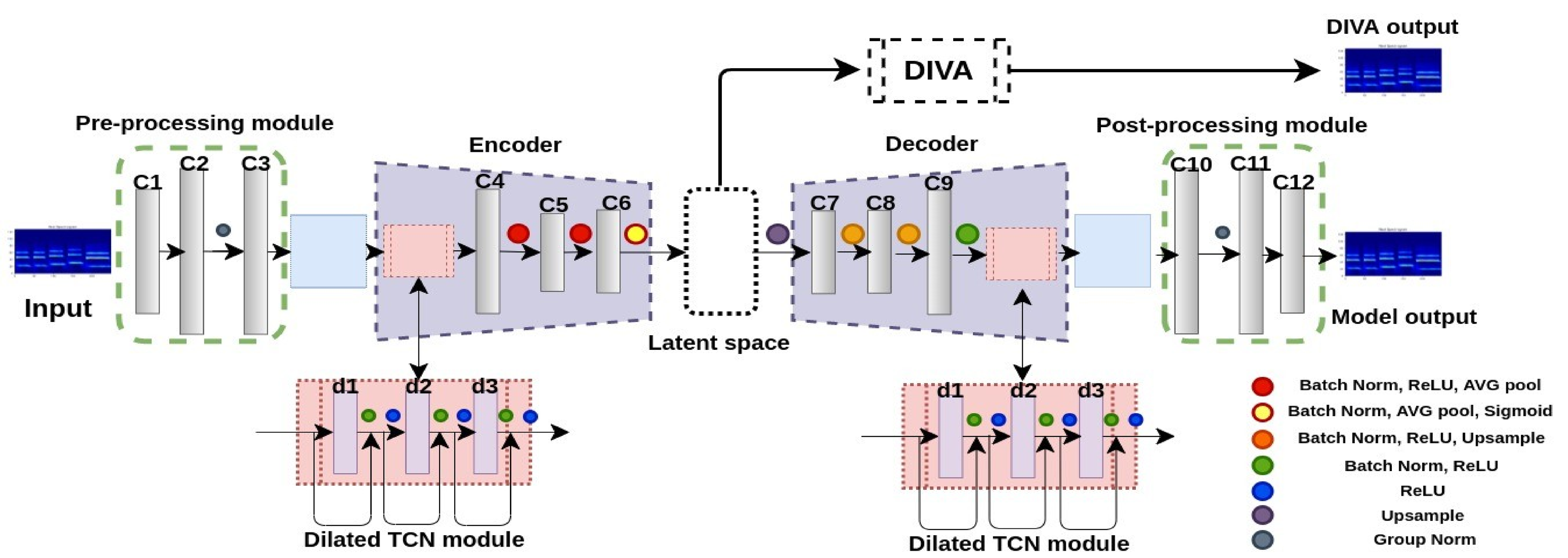


Figure 2: Role of the Forward Pass

Deep Neural Network (DNN) Architecture



DIVA control Parameters

Parameter Name	DIVA preset
MIDI note (Pitch)	-
MIDI duration	-
Volume	OSC : Volume2
Filter(center frequency BPF)	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

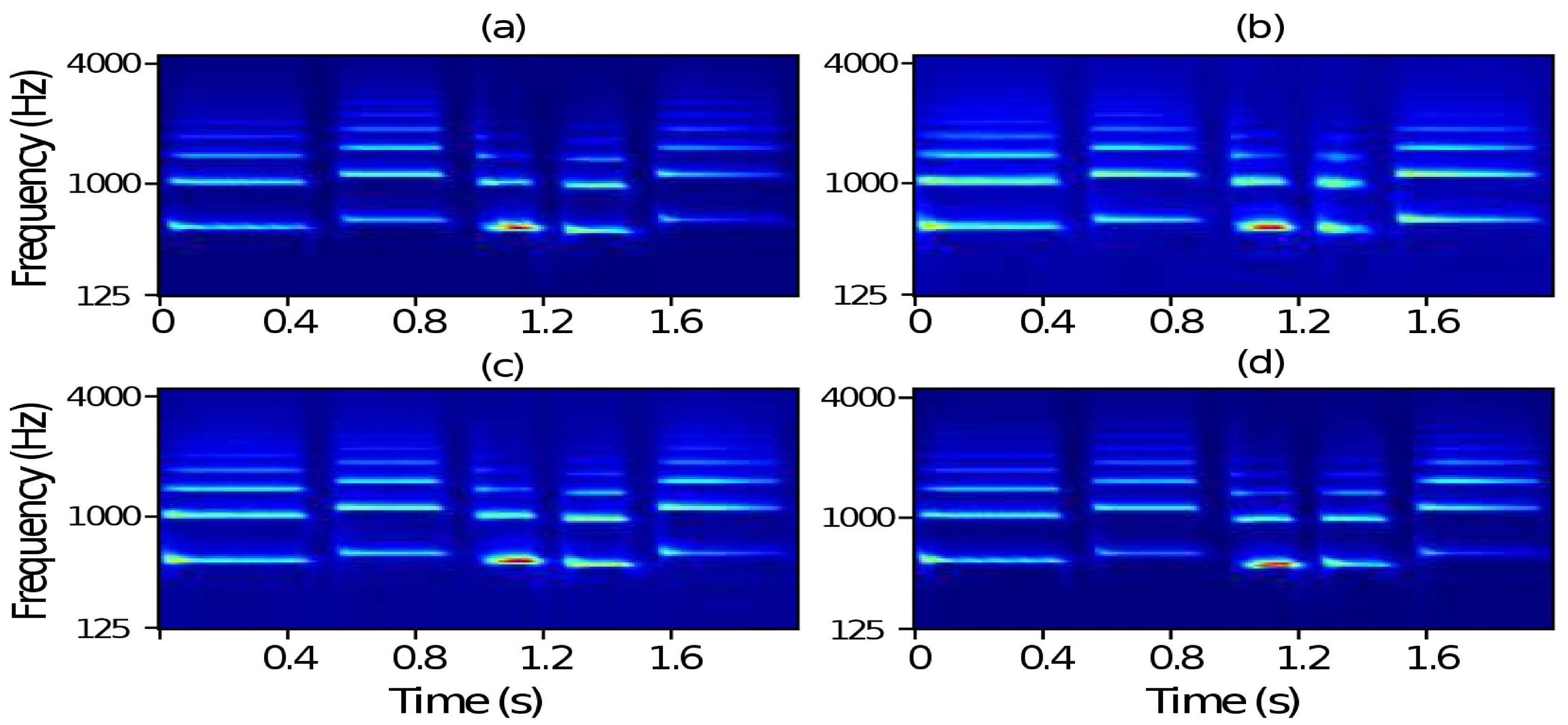
- ➢ MirrorNet predicts the first 7 parameters in Table (shaded in yellow)

Experiments

➢ **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
- Availability of ground-truth parameters to assess the MirroNet predictions

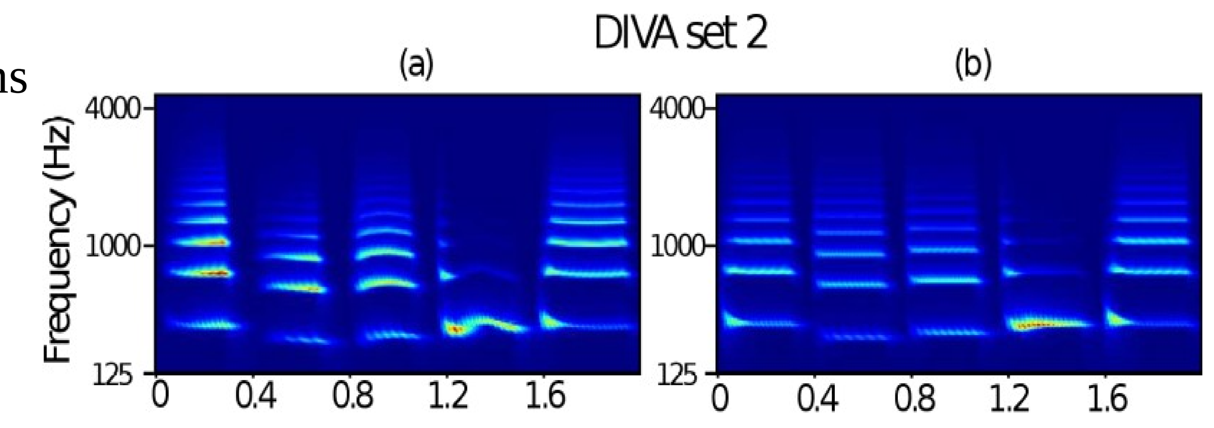
Auditory spectrograms
a) Input Melody b) Decoder Output from ground-truth parameters
c) Final output from Decoder d) DIVA output from learned control parameters



➢**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
- 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

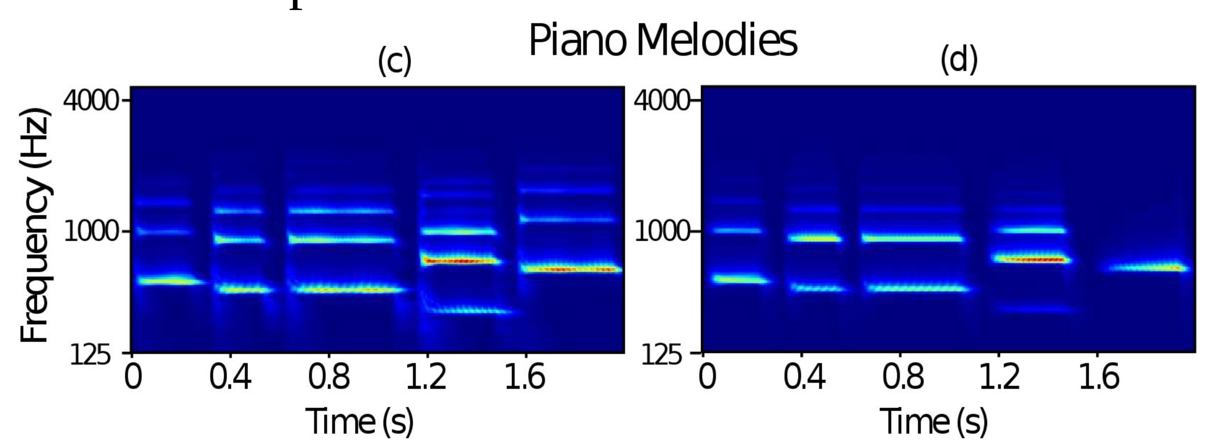
Auditory spectrograms
a)Input Melody
b)DIVA output from learned control parameters



➢**Experiment 3:** Learning DIVA parameters to synthesize melodies generated from a different synthescizer

- 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
- ➢ MirrorNet can accurately estimate control parameters for an off-the-shelf audio/music synthesizer to synthesize a given input melody
  - –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

Acknowledgments

This work was supported by Advanced ERC Grant NEUME 787836 and Air Force Office of Scientific Research and National Science Foundation grants to S.A.S.; and FrontCog Grant ANR-17-EURE-0017, PSL Idex ANR-10-IDEX-0001-02, and a PhD scholarship from the Research Chair on Beauty Studies PSL L'Oréal to G.M.

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.