ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747358

THE 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

ABSTRACT

Experiments to understand the sensorimotor neural interactions in the human cortical speech system support the existence of a bidirectional flow of interactions between the auditory and motor regions. Their key function is to enable the brain to 'learn' how to control the vocal tract for speech production. This idea is the impetus for the recently proposed "MirrorNet", a constrained autoencoder architecture. In this paper, the MirrorNet is applied to learn, in an unsupervised manner, the controls of a specific audio synthesizer (DIVA) to produce melodies only from their auditory spectrograms. The results demonstrate how the MirrorNet discovers the synthesizer parameters to generate the melodies that closely resemble the original and those of unseen melodies, and even determine the best set parameters to approximate renditions of complex piano melodies generated by a different synthesizer. This generalizability of the MirrorNet illustrates its potential to discover from sensory data the controls of arbitrary motor-plants.

Index Terms— Autoencoder, Audio synthesis, Music synthesis, DIVA synthesizer, Unsupervised learning

1. INTRODUCTION

Most organisms function by coordinating and integrating sensory signals with motor actions to survive and accomplish their desired tasks. For instance, visual and auditory signals guide animals to navigate their surroundings [1, 2]. Similarly, auditory and proprioceptive percepts are essential in skilled tasks like playing the piano or speaking. The difficulty of learning to perform these tasks is enormous. It stems from the fact that to control such actions, one needs harmoniously to close the loop between sensing and action. That is, it is necessary to map the desired sensory signals to the correct commands, which in turn produce exactly the desired sensory signals when executed.

But to learn the necessary mappings and interactions between the perception and action domains, standard Artificial Intelligence (AI) methodology typically relies on creating large databases that map the input sensory data to their corresponding actions, and then train intervening Deep Neural Networks (DNN) to associate the two domains [3, 4]. Humans and animals however never learn complex tasks in this way. For instance, human infants learn to speak by first going through a "babbling" stage as they learn the "feel" or the range and limitations of their articulatory commands. They also listen carefully to the speech around them, initially implicitly learning it without necessarily producing any of it. When infants are ready to learn to speak, they utter incomplete malformed replica of the speech they hear. They also sense these errors (unsupervised) or are told about them (supervised) and proceed to adapt the articulatory commands to minimize the errors and slowly converge on the desired auditory signal. In other words, learning these complex sensorimotor mappings proceeds simultaneously and often in an unsupervised manner by listening and speaking all at once [5, 6, 7].

Motivated by such learning of complex sensorimotor tasks, a new autoencoder architecture, referred to as the "Mirror Network" (or MirrorNet) was recently proposed in Shamma et al. [5]. 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. In this paper we extend and demonstrate the efficacy of the MirrorNet architecture in learning audio synthesizer controls/parameters to synthesize a melody of notes using a commercial, widely available synthesizer (DIVA) developed by U-He[1].

MirrorNet is fundamentally different from the Differentiable Digital Signal Processing (DDSP) based models [8, 9] which effectively learn a differentiable music synthesizer, whereas the goal of the MirrorNet is to learn controls to drive a given non-differentiable, off-the-shelf music synthesizer. Previous work with DNNs on determining music and speech synthesizer controls are all based on at least partially supervised techniques which often involve large databases of audio and control parameter pairs (order of 1000s) [10, 11, 12, 13]. Furthermore, previous efforts have mostly demonstrated the ability to compute the controls for single notes or single vowels for speech [11, 14]. In this paper we propose an alternative approach model which is fundamentally unsupervised, in that it does not require matched pairs of input melodies and their corresponding control parameters. The proposed model can predict synthesizer controls for a melody composed of several notes demonstrating the scalability of the model for real world applications. The true potential of the MirrorNet is further validated by showing how well it can predict synthesizer controls not only for DIVA generated melodies, but for other off-the-shelf synthesizer-generated melodies.

2. MIRRORNET MODEL

2.1. Model Architecture

The MirrorNet was initially proposed as a model for learning to control the vocal tract and is based on an autoencoder architecture. The structure of this network is shown in Figure 1a [5], depicting the biological structures and experiments that motivated the network. The goal of the model is 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). For simplicity, we use auditory spectrograms [15] generated from the audio streams as the input and output representations, but other representations may prove more versatile (e.g., cortical representations [16]). The "motor" parameters in this study are the parameters needed to synthesize the closest possible

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

audio signals matching the inputs. The primary difference between this MirrorNet and the previously studied model in [5] is the use of the music synthesizer (DIVA) with its unique set of parameters.

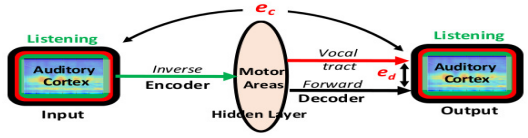
As shown in Figure 1a, the MirrorNet model is optimized simultaneously with two loss functions namely the 'encoder loss'(e_c) and the 'decoder loss'(e_d). The encoder loss is the typical autoencoder loss - the Mean Squared Error (MSE) between the input auditory spectrogram and the reconstructed auditory spectrogram from the decoder (forward path). The decoder loss is the MSE between the auditory spectrograms generated by the DIVA (the motor plant path) and the decoder (forward path). It is the 'decoder loss' that constrains the latent space to converge to the expected control parameters while simultaneously reducing (e_c), and this is the key feature of the MirrorNet architecture.

Figure 1b shows the role of the 'forward' path in the model, namely to back-propagate the errors computed to learn the 'inverse' mapping and hence the control parameters. In general, directly computing a vocal-tract or an audio synthesizer inverse is difficult if not impossible because of its complexity, nonlinearity, and our incomplete knowledge of its workings. The MirrorNet in Figure 1b (bottom panel) solves this problem by adding the forward projection that serves as a parallel, "neural" model of the vocal tract or the audio synthesizer, or any motor-plant to be used. The critical importance of this "neural" projection is that it readily provides a route for the e_c errors to back-propagate to the motor areas (latent space), enabling the training of the inverse mapping (Encoder).

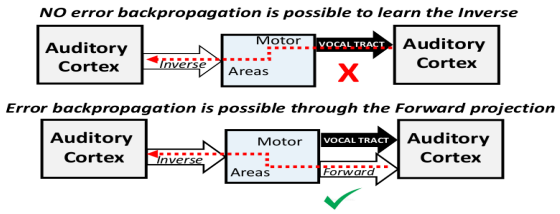
2.2. Model Implementation and Training

The MirrorNet for audio synthesizer control is implemented in PyTorch with 1-D convolutional (CNN) layers modeling both the encoder and decoder. The complete network is inspired by the multi-layered Temporal Convolution Network (TCN) [17]. Figure 2 shows the complete DNN model architecture with its sub-modules used for pre/post processing and dilated TCN. The pre/post processing modules are symmetrically matched (C1≡C12, C2≡C11, C3≡C10) and have 128, 256 and 256 filters respectively with 1×1 kernels. d1, d2 and d3 dilated CNN layers have a kernel size of 3 with 1,4 and 16 dilation rates respectively. The CNN layers in the encoder and decoder are also symmetrically matched and the C4, C5 and C6 layers have 256, 128 and 7 filters respectively with 1×1 kernels. The latent space dimensions are chosen to match with the number of parameters to be learned and the number of notes in each melodic segment. For example to learn 7 controls of the DIVA synthesizer to generate a melodic segment of 5 notes, we use a latent space of (7×5) dimensions. Average pooling is done after C4, C5 and C6 layers (window sizes of 5, 5 and 2 respectively) while upsampling is done before C7, C8 and C9 layers (window size of 2, 5 and 5 respectively). The auditory spectrograms used as inputs (and outputs) of the model are of dimension (128×250). We use auditory spectrograms which have a logarithmic frequency scale, simply because they provide a unified multi-resolution representation of the spectral and temporal features likely critical in the perception of sound [15, 16].

Unlike a regular autoencoder, the MirrorNet is trained in two alternating stages in each iteration. The decoder is trained first (to minimize e_d) for a chosen number of epochs. Then, the encoder is trained (to minimize e_c) for a given number of epochs and this alternation of training is continued until both losses converge to a minimum. Learning rates of 1e-2 and 1e-3 were used for the encoder and decoder networks, respectively. The best learning rates were determined based on a grid search testing all the combinations from [1e-2, 1e-3, 1e-4, 3e-4] for both the encoder and decoder which result in the lowest training errors at convergence. The two objective



(a) MirrorNet: Autoencoder Architecture



(b) Role of the forward pass

Fig. 1: MirrorNet Model Architecture for speech and the critical role of the forward projection (taken from *Learning Speech Production and Perception through Sensorimotor Interaction* by Shamma et al. in *Cerebral Cortex Communications.*)

Table 1: Set of Audio controls/parameters used. Here MIDI note and MIDI duration are parameters set in RenderMan library to drive the synthesizer patch.

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

functions were optimized using the ADAM optimizer with an 'ExponentialLR' learning rate scheduler and a decay (gamma) of 0.5. All the models were trained using NVIDIA Quadro P6000 GPUs and on average the models converged after around 32 hours of training. For further implementation information of the network, the PyTorch project is publicly available in GitHub [2]. Sample audio reconstructions can also be found in the supporting web page hosted in the GitHub repository.

2.3. DIVA audio synthesizer

We use DIVA, an off-the-shelf commercial synthesizer as our audio synthesizer for the MirrorNet model. DIVA has almost all its parameters MIDI-controlled. A python library called RenderMan [3] is used to batch-generate audio files using a fixed set of parameters. We built a software layer with RenderMan to drive DIVA to synthesize a melody of notes by concatenating individual notes synthesized by DIVA. All the melodies used in this paper are 2 seconds long and sampled at 44.1 kHz. The parameters are all continuous and normalized between [0,1]. Table 1 lists the set of parameters selected for the learning experiments with the MirrorNet, and the corresponding parameter labels from DIVA where applicable.

[2]https://github.com/Yashish92/MirrorNet-for-Audio-synthesizer-controls
[3]https://github.com/fedden/RenderMan

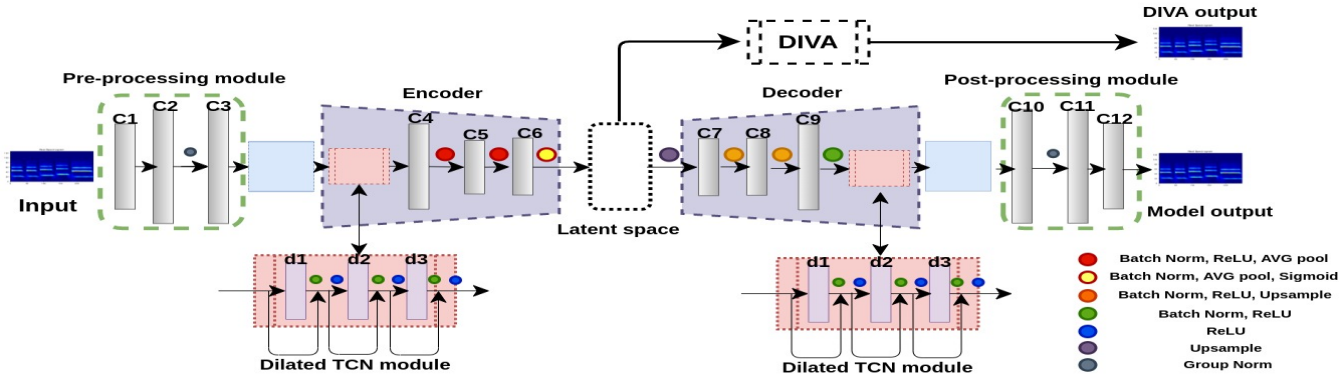


Fig. 2: DNN architecture of the MirrorNet model. Here C1-C12 represent 1D-CNN layers where d1-d3 represent 1D dilated CNN layers.

3. EXPERIMENTS AND RESULTS

3.1. Learning DIVA parameters from melodies synthesized with the same set of parameters (set 1)

In this first experiment, we use 400 melodies (set 1) to train the MirrorNet and test with 80 melodies, all originally synthesized by DIVA. The advantage of this set of melodies is that we have its ground-truth parameter values, and hence we can assess how accurately the MirrorNet rediscovers them and reconstructs the melodies. Each melody contains 5 notes and is 2 seconds long. The train and test set of melodies were synthesized by randomly sampling a total of 7 parameters (the first 7 parameters in Table 1) using a defined range and keeping a pre-defined set of other parameters constant across all notes and melodies. The pre-defined set of parameters used for the experiments can be found in the GitHub repository of the project.

Figure 3 depicts auditory spectrograms of a given melody at various stages in the fully-trained MirrorNet. The spectrogram (b) suggests how well the decoder has learned to generate an identical spectrogram to the one generated with DIVA for the exact same controls. The spectrogram (d) suggests how well predicted DIVA controls are from the encoder to synthesize an identical melody to the input.

We performed preliminary statistical tests to evaluate the robustness of the MirrorNet in predicting the 7 parameters. Plot in Figure 5a validates that the predicted and ground truth parameters are significantly closer together than would result from a random set of values. A second test was performed to check how well the predictions of each parameter are compared to a random prediction. For that we performed a Levene's test that confirmed that all parameter predictions were significantly better than chance. Plot in Figure 5b shows the parameter difference distributions for the test set. The distributions also suggest that critical parameters like pitch, bandpass filter, filter resonance and duration are predicted with significant accuracy where as volume and envelope attack parameters are predicted with comparatively lower accuracy.

3.2. Learning DIVA parameters from melodies synthesized with extra unknown DIVA parameters (set 2)

In this experiment, we use a train set of 400 and a test set of 80, both DIVA generated melodies (set 2) which are synthesized in similar fashion to set 1 except for the fact that they now use all the 10 parameters in Table 1. The MirrorNet is still trained to predict 7 parameters as in previous experiment. The goal here is to demonstrate that the MirrorNet can approximate the input melodies even if they have additional sound/musical qualities that are impossible for the restricted set of 7 DIVA parameters to reproduce, e.g., vibrato in this case. The top panel in Figure 4 illustrates the original (vibrato) notes and the successfully regenerated melody captured with only 7 parameters (vibrato not included).

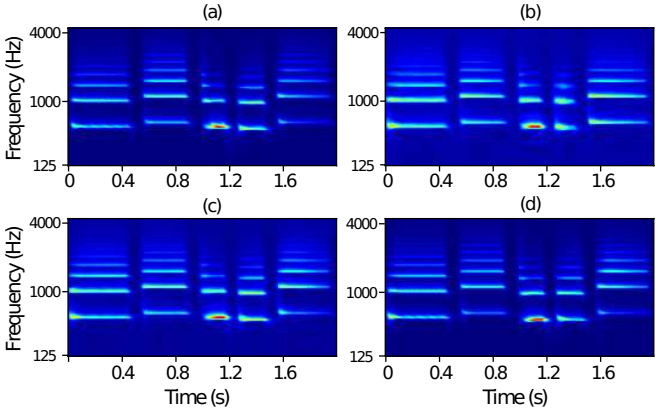


Fig. 3: Auditory spectrograms from the model learned with DIVA synthesized melodies (set 1). (a) Input melody (b) Decoder output from true DIVA parameters (c) Final output from the decoder (d) DIVA output from the learned control parameters

3.3. Learning DIVA parameters to synthesize melodies generated from other synthesizers

A fundamental advantage of the MirrorNet is its ability to discover the DIVA parameters corresponding to music generated by other sources and synthesizers by finding parameters that allow the DIVA output to be as close as possible, given the constraints of the number of parameters (here 7 are used), to the original input. The experiment utilized 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). The network successfully reproduces accurate renditions of the piano music from unseen samples (test set of 80 samples) using the decoder/encoder mappings learned during the training. The bottom panel in Figure 4 shows such an example where the DIVA produces a melody which closely resembles the input piano melody.

4. DISCUSSION

We described a MirrorNet model inspired by cortical sensorimotor interactions measured when humans speak or play a musical instrument [5]. The first two experiments utilized DIVA generated melodies for training, and this allowed us evaluate the effective-

Table 2: Mean and variance of Mean Square Errors (MSE's) across multiple model training runs

Input melody type	Train/Test for Input vs DIVA(learned)	Parameter-Train	Parameter-Test
DIVA melodies (set 1)	2.995±.21/3.596±.15	0.0666±.003	0.0671±.002
DIVA melodies (set 2)	6.380±.34/8.101±.20	0.0832±.007	0.0857±.004
Piano melodies	4.585±.25/4.751±.22	-	-

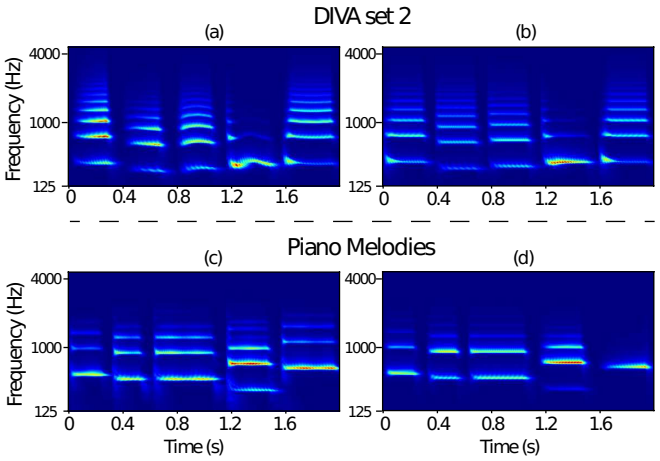


Fig. 4: (Top panel) Auditory spectrograms from the model learned with DIVA synthesized melodies (set 2) (a) Input melody (b) DIVA output from the learned control parameters. (Bottom panel) Auditory spectrograms from the model learned with piano melodies. (c) Input melody (d) DIVA output from the learned control parameters.

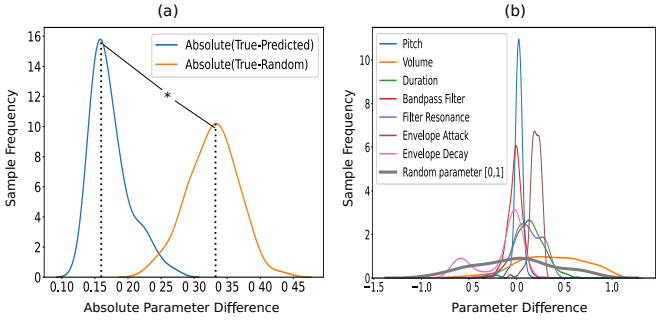


Fig. 5: Evaluating statistical significance of the predicted DIVA parameters with respect to a set of random parameters on the test set (a) Distributions for absolute parameter differences across all parameters (b) Distributions of parameter differences (ground truth - predicted) for 7 parameters and the distribution for a random parameter difference (ground truth - random)

ness of the MirrorNet given the ground truth parameters to compare against, e.g., to perform preliminary tests to validate the MirrorNet predictions of the synthesizer controls across all the training and test sets, as shown in Table 2. The MSE values for the test set compared to the train set in Table 2 also give an idea on how well the model generalizes for the unseen input melodies.

Taking the MirrorNet to the next level in the last experiment, we demonstrated how the MirrorNet could closely approximate a set of controls for DIVA to synthesize a set of piano melodies generated by a completely different synthesizer. This idea opens up a whole new area of applications in music synthesis as it describes a tool to find parameters for an arbitrary synthesizer that maximally approximate an arbitrary sound without being necessarily capable to exactly reproduce it (reproduce a violin using a guitar for instance). It should also be noted that this paper only discusses results in synthesizing fixed duration melodies with a fixed number of notes, but it is a step in the right direction to synthesizing a piece of music which can have a variable number of notes in a fixed frame of audio.

The inspiration of the MirrorNet also comes from the area of computational neuroscience and especially to learning and predictive processing. Our brain is able to extract strong relations between sensory stimuli and their corresponding motor parameters that enable children to learn to speak by mere passive exposure to speech without any proper external teaching. In addition, after learning to control their own vocal tract, adults can, without any additional training, produce sounds they hear even if the acoustic target is not reachable by their specific vocal tract (case of the experiments 2 and 3). However, the brain is able to find a set of motor parameters that approximate well the target sound while being produced by the specific vocal tract. Such predictive mechanism can also be seen in music production when humans learn how to play an instrument by mapping the auditory stimulation to the motor commands to a specific instrument. Even music perception rely on similar predictive pathways where high-order cortical areas constantly predict activation in the auditory cortices in order to modulate attention and emotions, for instance[18, 19].

Finally, from an engineering perspective, the MirrorNet can solve problems where it is hard to find a reasonable number of examples to train a regular feed-forward DNN network, or to learn from examples that may not be exactly similar to the motor-plant outputs, e.g., learning to synthesize a melody from naturally played music. We moreover believe that the MirrorNet can be generalized to design algorithms that can control motor-plants such as self-driving vehicles given various sensory data.

5. CONCLUSION AND FUTURE WORK

This paper presents an autoencoder architecture inspired by sensorimotor interactions to discover and learn audio synthesizer controls. The work is novel in that the proposed MirrorNet can learn the necessary controls to produce a melody in a completely unsupervised way. It can also be potentially generalized to learn the controls for any motor-plant action from the sensory data associated with them. However, to realize all these potentials, many more advances are needed. For example, for the audio synthesizer controls explored here, it is necessary to scale up the current implementations to far more parameters that capture richer aspects of the sound (e.g., vibrato), to deploy more advanced and richer representations of the sound beyond the spectrograms, to devise more efficient and faster training paradigms, and finally to target the synthesis of continuous musical melodies which can have a variable number of notes.

6. ACKNOWLEDGEMENTS

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.

7. REFERENCES

[1] D. Wolpert and Zoubin Ghahramani, "Computational principles of movement neuroscience," *Nature Neuroscience*, vol. 3 suppl. 1, pp. 1212–1217, 2000.

[2] Georg B. Keller, Tobias Bonhoeffer, and Mark Hübener, "Sensorimotor mismatch signals in primary visual cortex of the behaving mouse," *Neuron*, vol. 74, no. 5, pp. 809–815, 2012.

[3] Yiwei Fu, Devesh K. Jha, Zeyu Zhang, Zhenyuan Yuan, and Asok Ray, "Neural network-based learning from demonstration of an autonomous ground robot," *Machines*, vol. 7, no. 2, 2019.

[4] Lei Tai and Ming Liu, "Deep-learning in mobile robotics - from perception to control systems: A survey on why and why not," *CoRR*, vol. abs/1612.07139, 2016.

[5] 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.

[6] 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.

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

[8] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "Ddsp: Differentiable digital signal processing," 2020.

[9] Jesse Engel, Rigel Swavely, Lamtharn Hantrakul, Adam Roberts, and Curtis Hawthorne, "Self-supervised pitch detection by inverse audio synthesis," 2020.

[10] Matthew John Yee-King, Leon Fedden, and Mark d'Inverno, "Automatic programming of vst sound synthesizers using deep networks and other techniques," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 150–159, 2018.

[11] Philippe Esling, Naotake Masuda, Adrien Bardet, Romeo Despres, and Axel Chemla-Romeu-Santos, "Flow synthesizer: Universal audio synthesizer control with normalizing flows," *Applied Sciences*, vol. 10, no. 1, 2020.

[12] Gwendal Le Vaillant, Thierry Dutoit, and Sébastien Dekeyser, "Improving synthesizer programming from variational autoencoders latent space," in *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)*, Sept. 2021.

[13] Marc-Antoine Georges, Laurent Girin, Jean-Luc Schwartz, and Thomas Hueber, "Learning Robust Speech Representation with an Articulatory-Regularized Variational Autoencoder," in *Proceedings Interspeech 2021*, 2021, pp. 3345–3349.

[14] Pramit Saha and Sidney Fels, "Learning Joint Articulatory-Acoustic Representations with Normalizing Flows," in *Proceedings Interspeech 2020*, 2020, pp. 3196–3200.

[15] Kuansan Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 421–435, 1994.

[16] Taishih Chi, Powen Ru, and Shihab A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, 2005.

[17] 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.

[18] Guilhem Marion, Giovanni M. Di Liberto, and Shihab A. Shamma, "The music of silence: Part i: Responses to musical imagery encode melodic expectations and acoustics," *Journal of Neuroscience*, vol. 41, no. 35, pp. 7435–7448, 2021.

[19] Giovanni M. Di Liberto, Guilhem Marion, and Shihab A. Shamma, "The music of silence: Part ii: Music listening induces imagery responses," *Journal of Neuroscience*, vol. 41, no. 35, pp. 7449–7460, 2021.