KMITL ENGINEERING PROJECT DAY 2020

IMSE_Y4_07

Department of Computer Engineering
(Music Engineering and Multimedia)

The analysis and synthesis of Thai musical instruments using deep neural networks

Tantep Sinjanakhom[1], Dr. Nachanant Chitanont[2] and Asst. Prof. Dr. Sorawat Chivapreecha[3]

Abstract

This project aims to develop an intelligence signal processing technique that can synthesize the sound of Thai musical instrument including Thai xylophone and Thai flute. The convolutional neural network is used for pitch estimation and the test result is 5.34 in RMSE. The MLP and RNN are used for synthesis parameters generation. The system takes pitch and loudness of any monophonic signal as inputs and generate the sound of the instrument that it was trained for. The trained synthesizer was tested by resynthesizing the sound of Thai xylophone and Thai flute inputs. Another experiment is timbre transfer or synthesis of Thai instrument sound with other instruments used as input. The result can be further optimized in the future.

Introduction

Synthesizing musical instrument sound is one of the challenges in the field of both digital audio signal processing and artificial intelligence. Google's DDSP [1] is a model that combines deep neural networks and spectral modeling synthesis technique to give a natural sounding output. This project implements a similar architecture for Thai xylophone and Thai flute neural synthesizer. Also, the loudness tracker and pitch tracker model similar to CREPE [2] were implemented.

Methodology

The whole synthesizer system has the encoder, decoder, harmonic synthesizer, and filtered noise synthesizer.

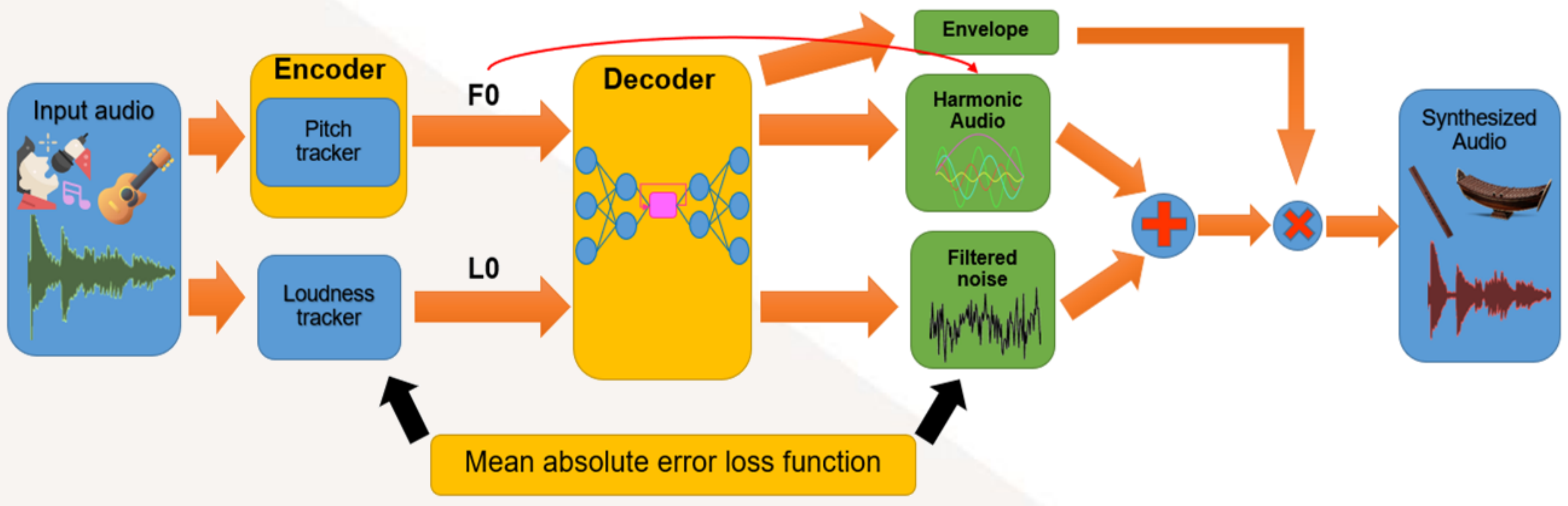

Figure 1 Synthesizer system architecture

Pitch tracker

- Based on residual convolutional neural network (CNN) [3].
- Takes 1024 audio samples at a time and process them through 8 hidden layers.
- It can predict the pitch between 65.41 Hz - 1975.53 Hz (C2-B6) with 50-cent intervals.
- Dataset includes synthetic pure sinusoid wave, complex sinusoid wave, flute, piano, guitar, and exponentially decaying complex sinusoids wave sound.
- The audio has length of 1 second and 16 kHz of sampling rate.

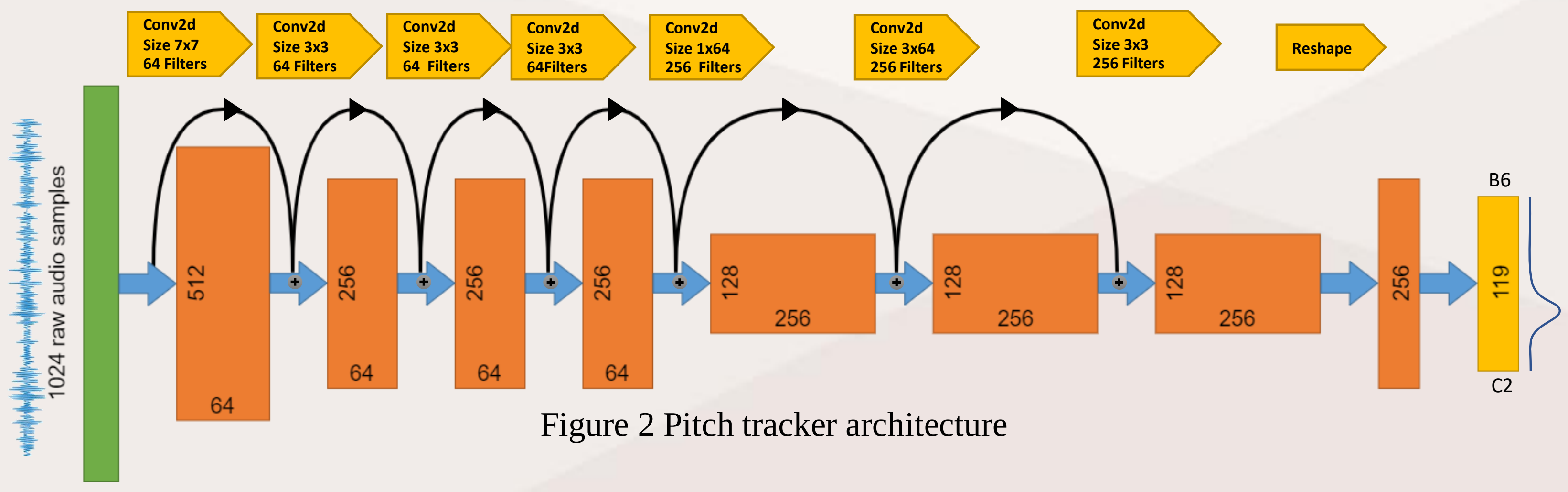
Conv2d Size 7x7 64 Filters
Conv2d Size 3x3 64 Filters
Conv2d Size 3x3 64 Filters
Conv2d Size 3x3 64Filters
Conv2d Size 1x64 256 Filters
Conv2d Size 3x64 256 Filters
Conv2d Size 3x3 256 Filters
Reshape
1024 raw audio samples
512
64
256
64
256
64
256
64
128
256
128
256
128
256
256
119
B6
C2

Figure 2 Pitch tracker architecture

Loudness tracker

- Based on short-time Fourier transform (STFT)
- The loudness is obtained by averaging the magnitude each frequency bin of the STFT of the signal that filtered with A-weighting filter.

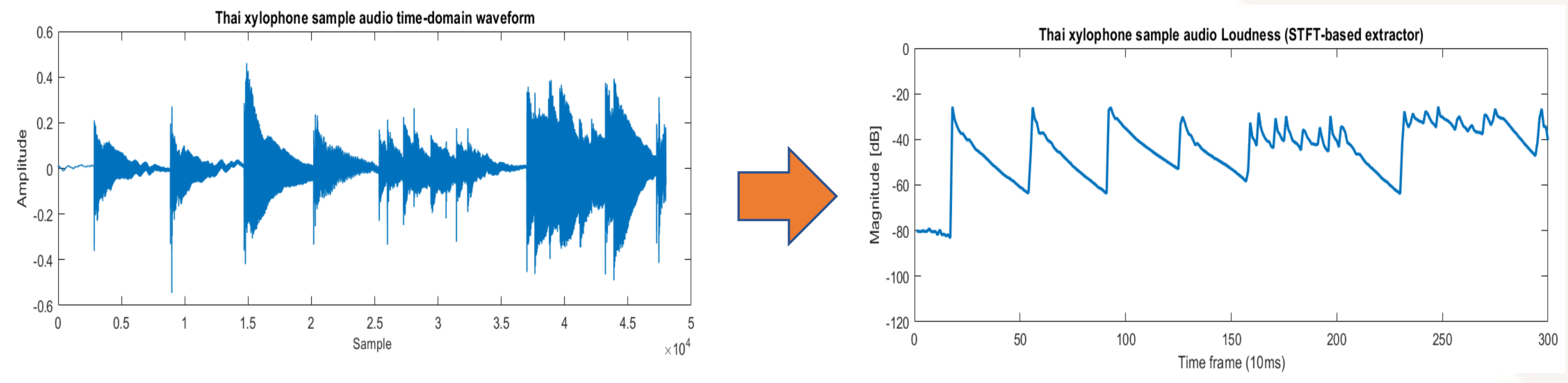

Figure 3 Example of loudness extraction

Thai xylophone recording session

- Recorded with Earthworks M30 microphone.
- The recorded contents include
  - Single note
  - Octaves
  - Octaves with tremolo/rolls
  - Songs and melodic phrases.

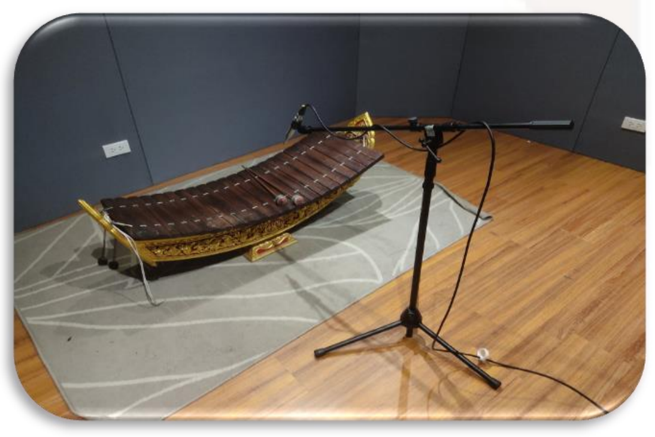

Figure 4 Recording session setting

Decoder

- Consisted of 3 identical neural networks including 5 hidden layers of multilayer perceptron and 2 GRU (recurrent) layers in the middle.
- It takes the inputs from the pitch tracker and loudness tracker.
- Generates 3 parameters: envelope, harmonic distribution, and filter magnitude response.
- Synthesizes a combination of harmonic and stochastic signal from the parameters.

Results

For the pitch tracker models, the least testing root mean squared error is 5.34. The decoder was tested on synthesis where the inputs are their own sound and timbre transfer in which the inputs are other types of instrument.

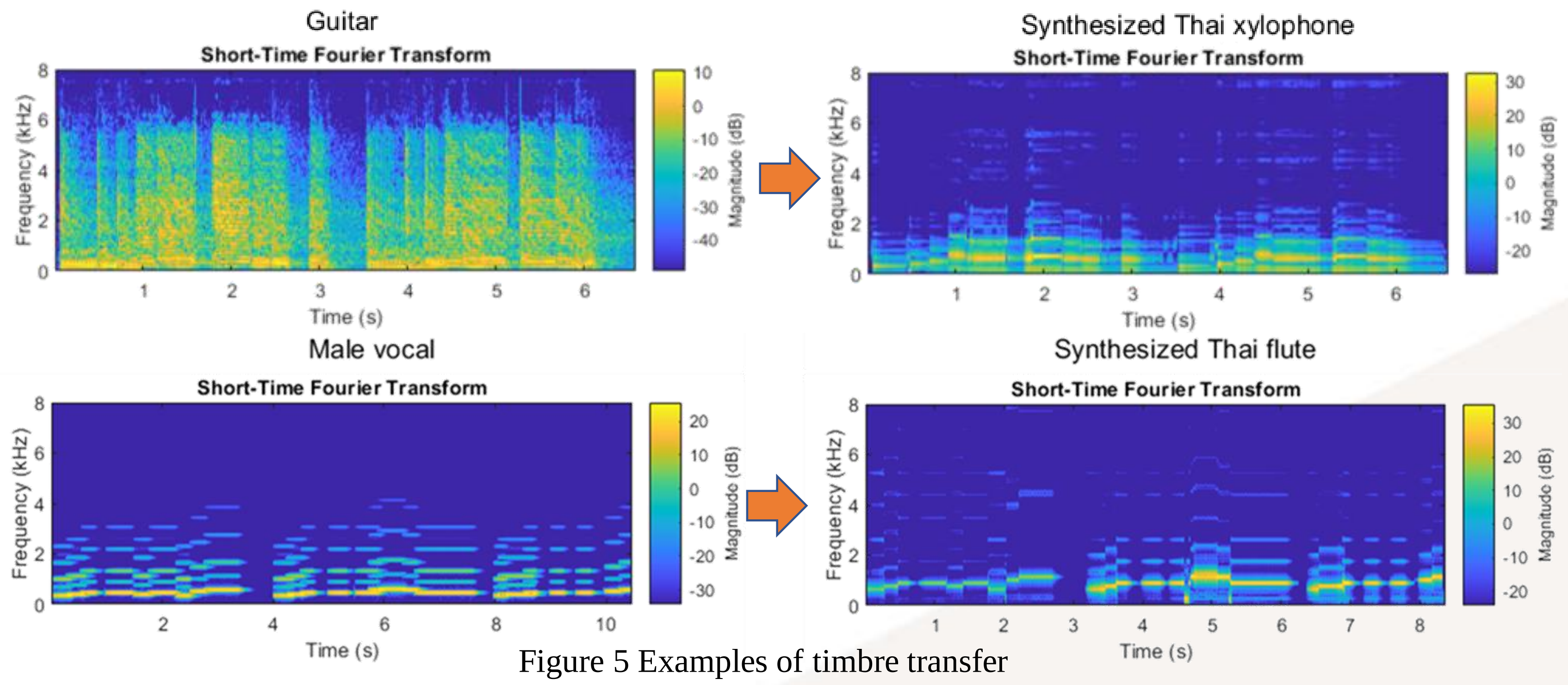

Figure 5 Examples of timbre transfer

Conclusion

- A neural Thai instrument synthesizer was implemented with sub-models inside the construction including residual CNN–based pitcher tracker, STFT-based loudness tracker, and the MLP, RNN-based decoder.
- The decoder was tested under two circumstance including resynthesis of the original signal and timbre transfer to transform other instruments into Thai instrument sound.
- The result can be further optimized to improve the naturalness.
- These models can be developed in other music applications.

References

[1] J. Engel et al., "Differentiable Digital Signal Processing," ICLR, 2020.
[2] J. W. Kimet al., "A Convolutional Representation for Pitch Estimation," 2018.
[3] K. He et al., "Deep residual learning for image recognition," CVPR, 2016.

E-mail: nachanant.ch@kmitl.ac.th[1], sorawat.ch@kmitl.ac.th[2]

KMITL 60 BEYOND THE LIMIT

SMART FACULTY