# Timbre Transformation

**Yu-Chen Huang**
Carnegie Mellon University
California, CA 94035
yuchenhu@andrew.cmu.edu

**Hung-Kuang Han**
Carnegie Mellon University
California, CA 94035
hungkuah@andrew.cmu.edu

**Yi Wang**
Carnegie Mellon University
Pittsburgh, PA 15213
wangyi@andrew.cmu.edu

## 1 Introduction

Timbre is the characteristic of instruments. People use timbre to identify different musical instruments. Sometimes, people may want to hear a piece with different instrumentation. Timbre transformation would expand the scope of sounds of any existing music, for instance, we could transform a piece of classical guitar performance into one performed with electric guitar without having the musician playing it all over. Together with the technique of instrumentation separation, we could apply timbre transformation on each individual instrument and yield a high-quality remix of any existing piece of music.

## 2 Related Work

In the previous work of [1], Settel, et al.(1994) use FFT/IFFT in real time to conduct digital signal processing in Max programming environment, which requires no compilation for digital signal processing(DSP). They use what's called overlap-add technique, including the following steps: (1) windowing input signal (2) transformation of the input signals into the spectral domain using FFT (3) operate on signal's spectra (4) resynthesis of modified spectra using IFFT (5) windowing the output signal. Their operation in the spectral domain includes convolution, addition, square root. We want to apply similar procedures for our timbre transformation project on data from Megenta's NSynth.

## 3 Dataset

After loading NSynth Dataset of guitar family[4], we obtain train, test and validation sub directories. Within each sub directory, audio files is formatted as: guitar_source_identifier_pitch_velocity.wav, (e.g.,guitar_acoustic_000-021-025.wav)

In each sub directory, we have 3 sources: acoustic, electronic and synthetic; number of identifiers varies between sources, pitch ranges from 21 to 108 and 5 velocity options: 25, 50, 75, 100, 127.

## 4 Working Experiment

Our goal is to transform audio of original source $S$ to target source $T$. To simplify the problem for now, we take $S$ and $T$ to be of the same family, which is the guitar family in our experiment.

We model our goal as we would like to obtain a transformation matrix $W$ that would transform the original source $S$ to an approximated spectrum of target source as $T'$ which would be close to $T$.

To obtain such transformation matrix, our idea is to find spectrum centroid of $S$ and $T$, $s$ and $t$. Then use $s$ and $t$ to compute the transform matrix $W$ between the two spectrum centroids. Our method to compute such transformation $W$ is taking the pseudo inverse of one:

$$W = Pinv(s) \times t. \tag{1}$$

Preprint. Under review.

Since we want our centroids to focus on timbre information and not be diverged by varing pitches in our training data set. We pitch-shifted our training set to pitch 69, which is A4 (A440) and used the pitch-shifted data for computing centroids.

Our method for computing centroids is quite naive and this is definitely the part we want to improve on. For now, the centroid of each source is computed by adding all spectrum in the pitch-shifted training data set then take its mean. The resulting centroid audio suffered recognizable loss phase cancellation.

In our experiment, we take the original source $S$ as acoustic guitar samples and target source $T$ as electronic guitar samples.

After getting the transformation matrix $W$ between spectrum centroids of $s$ and $t$, we would apply $W$ to transform acoustic source to electronic source in the testing data set.

We tested our transform on the testing set by randomly selecting acoustic audio file $S$ and electronic audio file $T$ of the same pitch. Then conduct the transform on the $S$ and collect the transformed audio $T'$. We then calculated the average norm distance between MFCC feature of $T'$ and the electronic source audio of same pitch, $T$. The average norm distance between $T$ and $T'$ is 849.

Figure 1 is the spectrum of one of the electronic source audio from testing data set, figure 2 is the spectrum of the transformed electronic from one of the acoustic source audio of the same pitch from testing data set.

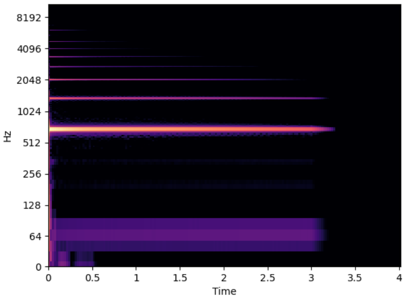

Figure 1: standard sound

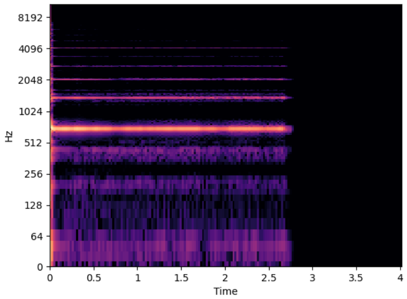

Figure 2: transformed sound

## 5 Not Working Experiment

### 5.1 Linear Regression on MFCC features

We first extracted a set of MFCC features from audio snippets of a specific acoustic guitar identifier as $X$ and an electronic one as $Y$. Each pair of MFCC features $< x, y >$ in $< X, Y >$ is a 1-1 mapping relationship. Then, we used linear regression to learn the MFCC feature transformation matrix $W$ of a specific acoustic guitar identifier to an electronic one.

We got the prediction MFCC features by applying $W$ to the randomly selecting testing data, i.e., another acoustic guitar identifier's MFCC features. The norm distance of MFCC features between the prediction and the ground truth (corresponding electronic identifier's MFCC features) is 49401.864.

### 5.2 Non-negative Matrix Factorization (NMF) on spectrograms

We also experimented with NMF to learn how spectrograms transform from an acoustic guitar identifier to an electronic one. We initialized $V$ with a spectrogram from a random audio snippet of acoustic guitar identifier, $W$ with the corresponding spectrogram from electronic guitar identifier, and $H$ with random values.

Then, we iteratively updated the $H$ by replacing $< V, W >$ with the next pair $< v, w >$ (either $v$ or $w$ represent a spectrogram of an audio snippet of a specific instrument, pitch, and velocity from acoustic and electronic audio snippets).

Table 1: Timeline and division of work

| Task | ETA | Members |
|---|---|---|
| Improve feature extraction of our model | 11/21 | Wang |
| Improve model design of our model | 11/28 | Huang, Han |
| Experiment on our improved model | 12/5 | Wang |
| Finish final report | 12/13 | Huang, Han, Wang |

We get the prediction spectrogram by multiplying the testing data with transformation $H$. Next; we utilized inverse-SFTT to extract the MFCC features from the prediction spectrogram. The norm distance of MFCC features between the prediction and the ground truth is 5855.524.

### 5.3 Possible improvements

Both models fail to learn a generalized transformed matrix because a pair of instrument identifiers cannot represent the domain from acoustic to electronic domain.

To learn a more generalized transformed matrix, we will then train the model with different acoustic and electronic instruments combinations at the same pitch.

## 6 Evaluation metrics

We use Mel cepstral distortion (MCD) [5] as evaluation metric because it's a popular objective measure for evaluating the timbre similarity [6]. MCD represents the distance between the MFCC feature of spectrum of the transformed electric guitar sound set and the standard electric guitar sound set, and the formula is as follows:

$$MCD(y - \hat{y}) = \frac{10\sqrt{2}}{ln10}\|y - \hat{y}\|_2 \tag{2}$$

Where $y$ is the standard sound, $\hat{y}$ is the transformed sound and the coefficient in front of the norm is to convert the unit to decibels.

## 7 Goals, Timeline and Division of work

See Table 1.

## References

[1] Settel, Z., & Lippe, C. (1994). Real-time timbral transformation: FFT-based resynthesis. Contemporary Music Review, 10 (2 ), 171-179.

[2] Wakabayashi, Y., Fukumori, T., Nakayama, M., Nishiura, T., & Yamashita, Y. (2017, March). Phase reconstruction method based on time-frequency domain harmonic structure for speech enhancement. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ) (pp. 5560-5564 ). IEEE.

[3] Yoshii, K., Tomioka, R., Mochihashi, D., & Goto, M. (2013, November ). Beyond NMF: Time-Domain Audio Source Separation without Phase Reconstruction. In ISMIR (pp. 369-374).

[4] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D. &amp; Simonyan, K.. (2017). Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. <i>Proceedings of the 34th International Conference on Machine Learning</i>, in <i>Proceedings of Machine Learning Research</i> 70:1068-1077 Available from https://proceedings.mlr.press/v70/engel17a.html.

[5] Kubichek, R. (1993, May). Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing (Vol. 1, pp. 125-128). IEEE.

[6] Kim, J. W., Bittner, R., Kumar, A., & Bello, J. P. (2019, May). Neural music synthesis for flexible timbre control. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 176-180). IEEE.