Representing the Timbre of Traditional Musical Instruments Based On Contemporary Instrumental Samples Using DDSP

Yousang Kwon
UNIST
Ulsan, Republic of Korea
yk7244@unist.ac.kr

Seonuk Kim
UNIST
Ulsan, Republic of Korea
d02reams@unist.ac.kr

Taeyoung Ko
UNIST
Ulsan, Republic of Korea
tyk0506@unist.ac.kr

Juhyeok Yoon
UNIST
Ulsan, Republic of Korea
heok95@unist.ac.kr

Kyungho Lee
UNIST
Ulsan, Republic of Korea
kyungho@unist.ac.kr

ABSTRACT

This project explores the potential of Differentiable Digital Signal Processing (DDSP) to represent and synthesize the timbre of five different notes of the Korean traditional musical instrument, Geomungo, using digital instrumental samples of the bass guitar, which has a similar mechanism to produce the sound. To evaluate the feasibility and quality of the digital recreation process, we compared hand-played Geomungo audio samples with digitally recreated audio samples using DDSP. The MFCC, spectral contrast, chroma features, and raw signal comparison, were used for assessment. Our findings show the possibility of applying DDSP to represent and synthesize the nuances of pitch and dynamics for expressive aspects of Geomungo's five different notes effectively. We also propose three audio features that can be used to evaluate the results quantitatively under the context of neural sound synthesis.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**.

KEYWORDS

traditional instrument, neural sound synthesis, audio style translation, digital signal processing, timbre, sound quality

ACM Reference Format:
Yousang Kwon, Seonuk Kim, Taeyoung Ko, Juhyeok Yoon, and Kyungho Lee. 2023. Representing the Timbre of Traditional Musical Instruments Based On Contemporary Instrumental Samples Using DDSP. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3586182.3616678

1 INTRODUCTION

A traditional musical instrument is an exceptional manifestation of the culture from which it originates, having been created and refined by musicians and craftsmen to cater to the requirements of those who engage with it and its particular role within a society. However, traditional music is facing the risk of disappearance as contemporary values become more prevalent in our society. This not only threatens the physical instruments themselves but also endangers the distinctive timbres and tones associated with them, which often carry ethnic or national characteristics.

Since many traditional musical compositions, expressions, and timbre have not been represented and transformed digitally well [8], it is difficult to access and utilize such traditional musical sounds and expressions in the contemporary music scene driven by digital tools. One of the challenges in the digital representation of a traditional music sound is to capture and synthesize the sound quality produced by the instrument – the timbre as Smalley [17] noted. Representing and recreating timbre from a signal-processing perspective presents several challenges due to the multidimensional and subjective nature of timbral perception, such as (1) High-dimensional data: Timbre analysis often involves extracting a wide range of features from audio signals. This complexity can make it difficult to process and interpret the data effectively; (2) Non-linear Transformations: Timbral attributes can undergo nonlinear transformations. This can lead to nontrivial relationships between physical sound properties and perceptual timbre qualities; and (3) Lack of Objective Metrics: Unlike pitch and loudness, which have relatively objective metrics, there is no universally accepted set of objective metrics to quantify timbre and evaluate its quality.

To address the gap, we explore the potential of Differentiable Digital Signal Processing (DDSP) [4] to represent and synthesize the timbre of a traditional acoustic music instrument sound based on a modern electric music instrument audio sample and evaluate its similarity between the samples. Specifically, we investigated the problem of timbre style transfer for synthesizing a traditional Korean musical instrument, Geomungo, from digital instrumental samples of the bass guitar, which has a similar mechanism (plucked and finger-picking techniques) to produce the sound. For our experiments, DDSP was considered as it is proven effective in morphing audio into a range of different instruments while it preserves the nuances of pitch and dynamics as a form of neural audio synthesis [6]. Also, it is agnostic to any given network architecture. Therefore, DDSP can be used as one of the neural network components rather than a specific generative model built.

As a preliminary exploration, we demonstrate the possibility of effectively using DDSP to represent and synthesize the nuances

UIST '23 Adjunct, October 29–November 01, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0096-5/23/10.
https://doi.org/10.1145/3586182.3616678

of pitch and dynamics for expressive aspects of Geomungo's five different notes. First, we identified the minimum requirements of training data and its specific format to be fed into the network, which helped us understand how much audio data from a traditional instrument is needed as a baseline. Second, we integrated the trained model into a commercial digital composition tool to demonstrate its feasibility of working in real-time. Lastly, we conducted a quantitative audio quality assessment covering all five strings in Geomungo to better understand the sound quality differences between each pitched string. Also, we proposed three audio features that can be used to evaluate the results quantitatively under the context of neural audio synthesis. We compared DDSP-generated sound samples and original 'Geomungo' samples by measuring the Euclidean distance of sound features such as raw signal, MFCC, spectral contrast, and chroma features[18] so that future researchers working on this topic can use the same baseline. To our knowledge, no previous studies have explored the application of Differentiable Digital Signal Processing (DDSP) in synthesizing the timbre of traditional instruments. It is our hope our research can be a starting point to capture, represent, and recreate traditional music digitally so we can use various traditional and ethnic music and its sound to enrich our contemporary music scene.

2 EXPERIMENT

For DDSP timbre training, a minimum of 10 minutes of monophonic audio data from a single recording session is required [4]. In our study, we collected monophonic recordings of 'Geomungo' to ensure consistency in the recording environment. The audio files were recorded at a sampling rate of 44,100Hz, 16-bit depth, and saved as 320kbps mp3 files. We concatenated these files into a single 50-minute file. The training process involved adjusting parameters such as the training step, with an average of 40,000 steps chosen from the range of 30,000 to 50,000 steps.

As the DDSP library offers output files that can seamlessly integrate with widely used software such as Garageband[9] and Logic[10], the trained results can be utilized as DAWs (Digital Audio Workstations)' plugins to transform existing sound samples into distinct Geomungo timbre. We used this approach to generate samples. To compare the timbres, we focused on analyzing signals for basic monophonic notes[14]. The reference sound samples was obtained by recording and sampling the original 'Geo-mun-go' instrument, specifically the virtual instrument developed by Seoul National University [5]. Using DDSP, we transferred the timbre of the Muted Bass instrument, commonly found in Western music, to the Geomungo's timbre. The DDSP-transferred and sampled instruments were played at the same note, specifically G, A, C, D, and E, with consistent length and velocity. These sounds were processed using Garageband's sound tools to demonstrate their potential in the real-time operation context.

One challenging thing in our evaluation is understanding the subtle, qualitative differences in timbre or sound texture in the samples. In our case, a typical triangulation approach could be considered a subjective interpretation as it is often influenced by an individual's cultural background and environment, leading to inconsistent evaluations. Addressing the gap, researchers such as Cao et al.[1], Deng et al.[3], and Karajalainen et al.[13] have proposed various scientific methods using signal processing to analyze timbre more objectively. Therefore, we compared the sound samples in general sound analysis methods by extracting sound features and calculating the Euclidian distance to know the similarity quantitatively, as shown in Table 1. The feature we used: (1) Raw signal: the level of sound in the time domain; (2) Chroma Features[15]: signal's spectrum divided into twelve pitch classes or chroma of the equal-tempered scale; (3) MFCC[16]: cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale; and (4) Spectral Contrast[12]: represents the spectral peak, spectral valley and their difference in each subband. For each feature, the Euclidean distance between original and synthesized audio samples was calculated.

Table 1: Euclidean Distance of Sound Features

Features \ Notes	G	A	C	D	E
Raw Signal	5.32	5.18	7.08	6.58	7.83
Chroma Feature	0.903	0.974	0.967	0.819	0.719
MFCC	30.359	65.575	57.113	39.955	33.500
Spectral Contrast	27.100	24.828	22.971	22.255	29.344

3 FINDINGS & DISCUSSION

As Table 1 suggests, G and A notes are well represented in terms of raw signal comparison, whereas C and E are not well synthesized. In terms of Chroma Feature, D and E notes are well represented where G, A, and C are not well synthesized. G and E can be represented well by using MFCC while A and C are not well reproduced. From a Spectral Contrast perspective, C and D notes are represented well, while there are small differences between G, A, and E. The result implies that although the audio quality seems to be similar, there can be subtle differences. To interpret the results, it is necessary to conduct comparative studies that involve instruments traditionally perceived as similar or belonging to different categories. For instance, a study by Gonzalez et al. [7] measured the Euclidean distance of raw signals between different dynamics of the flute and clarinet, resulting in distances ranging from 1.0 to 7.7 for the same note. Considering the raw signal Euclidean distances obtained in our experiment, ranging from 5.18 to 7.83, it can be inferred that the timbre falls within a similar or same category.

However, it is important to recognize that the Euclidean distance can vary depending on factors such as the number of samples, signal length, and processing methods. Therefore, establishing standardized measures of similarity by conducting a more rigorous comprehensive evaluation incorporating a multi-layered approach such as Chu et al.'s [2]. Additionally, exploring alternative similarity calculation methods, such as Dynamic Time Warping or Cosine similarity, can provide further insights into timbre comparison. It is important to note that timbre perception is subjective and can vary depending on the listener as [11] noted. In future work, a user study will be conducted to evaluate the perception of the DDSP-generated sounds using both qualitative and quantitative methods. This will provide valuable feedback on how listeners perceive the timbre and help assess the effectiveness of the DDSP approach in creating authentic and pleasing instrument sounds.

REFERENCES

[1] Xi-zheng Cao, Hui-li Meng, and Jiu-cheng Xu. 2009. Timbre model of software musical instrument based on sine interpolation. In *2009 International Conference on Image Analysis and Signal Processing*. 358–361. https://doi.org/10.1109/IASP.2009.5054598 ISSN: 2156-0129.

[2] Hyeshin Chu, Joohee Kim, Seongouk Kim, Hongkyu Lim, Hyunwook Lee, Seungmin Jin, Jongeun Lee, Taehwan Kim, and Sungahn Ko. 2022. An Empirical Study on How People Perceive AI-generated Music. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 304–314.

[3] Jeremiah D. Deng, Christian Simmermacher, and Stephen Cranefield. 2008. A Study on Feature Analysis for Musical Instrument Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38, 2 (April 2008), 429–438. https://doi.org/10.1109/TSMCB.2007.913394 Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).

[4] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. 2020. DDSP: Differentiable Digital Signal Processing. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1x1ma4tDr

[5] Center for Arts and Technologies Seoul National University. 2014. *Gugak VSTi*. http://en.catsnu.com/Main/Main.aspx

[6] Francesco Ganis, Erik Frej Knudesn, Søren VK Lyster, Robin Otterbein, David Südholt, and Cumhur Erkut. 2021. Real-time timbre transfer and sound synthesis using ddsp. *arXiv preprint arXiv:2103.07220* (2021).

[7] Yubiry Gonzalez and Ronado C. Prati. 2023. Similarity of Musical Timbres Using FFT-Acoustic Descriptor Analysis and Machine Learning. *MDPI* (Apr 2023). https://doi.org/10.1007/978-3-319-63450-0

[8] Simmon Holland (Ed.). 2013. *Music and Human-Computer Interaction*. Vol. 295. Springer-Verlag, New York, NY, 7–8. https://doi.org/10.1007/978-1-4471-2990-5

[9] Apple Inc. 2023. *Garageband for MacOS*. https://www.apple.com/mac/garageband/

[10] Apple Inc. 2023. *Logic*. https://www.apple.com/logic-pro/

[11] Mads Græsbøll Christensen Jesper Højvang Jensen and Søren Holdt Jensen. 2007. A framework for analysis of music similarity measures. In *2007 15th European Signal Processing Conference*.

[12] Dan Nin Jian, Lie Lu, Hong Jian Zhang, Jian-Hua Tao, and Lian Hong Cai. 2002. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*. https://doi.org/10.1109/ICME.2002.1035731

[13] Matti Karjalainen and Vesa VŠlimŠki. [n. d.]. Towards High-Quality Sound Synthesis of the Guitar and String Instruments. ([n. d.]).

[14] Megan L. Lavengood. 2017. *A New Approach to the Analysis of Timbre*. Ph. D. Dissertation. The City University of New York.

[15] Sebastian Ewert Meinard Muller and Sebastian Kreuzer. 2009. Making chroma features more robust to timbre changes. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1009/978-1-4244-2354-5

[16] Monica S. Nagawade and Varsha R. Ratnaparkhe. 2017. Musical Instrument Identification using MFCC. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*.

[17] Denis Smalley. 1994. Defining timbre — Refining timbre. *Contemporary Music Review* 10, 2 (Jan. 1994), 35–48. https://doi.org/10.1080/07494469400640281 Publisher: Routledge _eprint: https://doi.org/10.1080/07494469400640281.

[18] Mark D. Plumbley Tuomas Virtanen and Dan Ellis. 2017. *Computational Analysis of Sound Scenes and Events*. Vol. 417. Springer-Verlag, New York, NY. https://doi.org/10.1007/978-3-319-63450-0