DIFFUSION-BASED MUSIC AUDIO EDITING SYSTEM USING DIFFERENTIABLE DIGITAL SIGNAL PROCESSING MIXTURE MODEL

Kengo Takemoto[1,2], *Tomohiko Nakamura*[2], *and Hiroshi Saruwatari*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
takemoto-kengo2262@g.ecc.u-tokyo.ac.jp

ABSTRACT

This paper proposes a music audio editing system that enables source-wise editing of harmonic instrument mixtures without explicit source separation. It builds on our previously proposed score-informed method for estimating source-wise synthesis parameters, i.e., time-varying controls used to synthesize each source, such as fundamental frequency and loudness. The method directly estimates these parameters from a mixture signal and the corresponding musical score in an analysis-by-synthesis framework. Using the estimated parameters, the proposed system allows users to edit individual sources by modifying note sequences and instrument types, and then re-synthesizes the edited mixture. Through demonstrations on two-instrument mixtures, we show that the system supports note-level phrasing modification and instrument conversion of selected sources.

1. INTRODUCTION

Music audio editing aims to modify a given musical audio signal according to user preferences and has advanced with the introduction of deep neural networks (DNNs) [1]. This task requires not only high audio quality but also controllability, i.e., faithfulness to the specified edit while preserving the musical content that should remain unchanged. To improve controllability through intuitive user instructions, text-conditioned and multimodal music generation methods have been actively explored, but achieving precise control over user-specified attributes remains challenging [2]. This motivates models that expose control variables directly corresponding to musical attributes to be edited.

One promising approach to such controllable music audio editing is differentiable digital signal processing (DDSP) [3]. It implements signal-processing modules, such as additive and subtractive synthesizers, as differentiable modules and trains them jointly with DNNs in an end-to-end manner. As intermediate representations, it provides physically interpretable parameters (*synthesis parameters*), such as fundamental frequency (F_o) and loudness. These parameters allow DDSP-based models to relate user-specified edits to synthesis controls more explicitly than purely black-box generation models [4]. Despite their success with isolated musical instrument sounds, most DDSP-based models are not designed to handle mixture signals directly [5], making source-wise editing of musical mixtures difficult without explicit source separation.

In this paper, we present a system that estimates synthesis parameters of individual sources directly from a mixture signal and enables users to edit the sources through the estimated parameters. It builds on our previously proposed score-informed DDSP-based synthesis parameter estimation method [6]. Given an observed mixture and its musical score, the system estimates the synthesis parameters through an analysis-by-synthesis approach, where the parameters are inferred so that the re-synthesized mixture matches the observed mixture. The system then regenerates the parameters according to user-modified score information. We show two illustrative examples of the proposed system: note-level phrasing modification and instrument conversion of an individual source.

2. PROPOSED MUSIC AUDIO EDITING SYSTEM

Figure 1(a) shows a system workflow consisting of synthesis parameter estimation and music audio editing stages. The former estimates source-wise synthesis parameters from an observed mixture and its musical score, while the latter regenerates the parameters according to user-modified score information. We first describe the system architecture and then explain the two stages.

2.1. System Architecture

The proposed system comprises a dynamic time warping (DTW) module, a DDSP mixture model (DDSPMM) module, and and a diffusion-based estimator. The DTW module temporally aligns a musical score with the performance in the input mixture signal. For this alignment, we use memory-restricted multiscale DTW [7].

The DDSPMM module is based on the DDSP mixture model, which models a harmonic sound mixture using source-wise DDSP synthesizers [8]. Each DDSP synthesizer is the decoder of a pretrained DDSP autoencoder, which reconstructs isolated harmonic instrument sounds from frame-wise synthesis parameters such as F_o, loudness, and timbre features [3]. Given source-wise synthesis parameters, the module generates the corresponding source signals and sums them to obtain a synthesized mixture. This representation provides a synthesized mixture for analysis-by-synthesis while retaining synthesis parameters as meaningful controls.

The diffusion-based estimator is based on a diffusion model defined in the synthesis parameter domain. The model is pretrained to generate the synthesis parameter trajectory of a single source conditioned on its note sequence and instrument type [6]. By generating the parameters as a trajectory, the model captures temporal dependencies among synthesis parameters and produces musically plausible parameter sequences. The note condition specifies the temporal structure of the source, while the instrument condition

This work was supported by JSPS KAKENHI under Grant JP23K28108.

(a) Workflow

(b) Graphical user interface

Figure 1: Workflow and graphical user interface of the proposed music audio editing system. The workflow consists of a synthesis parameter estimation stage and an interactive music audio editing stage.

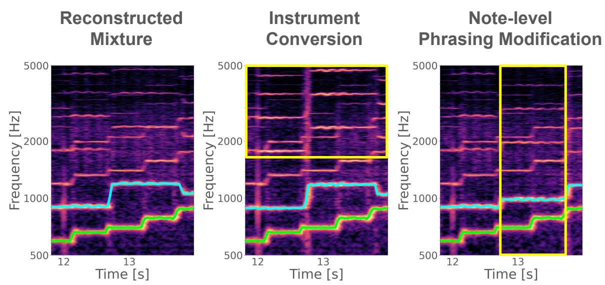

Figure 2: Examples of source-wise music audio editing using the proposed system. Blue and green curves denote the F_0 trajectories of the respective sources. The blue part was edited while the green part is left unchanged.

specifies the instrument identity. Following the denoising diffusion probabilistic model (DDPM) framework [9], the model generates samples by iteratively denoising Gaussian noise through the reverse diffusion process.

The proposed system also includes a graphical user interface (GUI) for parameter visualization and score-level editing. Figure 1(b) shows the GUI of the proposed system. It visualizes the estimated F_0 and loudness trajectories together with time-aligned note blocks, which users can interact with for score-level editing. For each source, users can also select a target instrument from a dropdown menu. The interface supports separate playback of the estimated source signals and the observed mixture.

2.2. System Workflow

In the synthesis parameter estimation stage, the user first provides an mixture audio and a musical score in Musical Instrument Digital Interface (MIDI) format. The DTW module then aligns the score with the mixture. Using the time-aligned score, source-wise synthesis parameters are estimated from the observed mixture following the method in [6]. During estimation, the diffusion-based estimator and the DDSPMM module work in tandem. At each reverse diffusion step, the current source-wise synthesis parameter estimates are passed to the DDSPMM module, and the synthesized mixture is compared with the observed mixture using a multi-scale spectral loss [3]. This loss is incorporated into the reverse diffusion process so that the estimator produces parameters whose synthesized mixture is closer to the observed mixture.

Once the synthesis parameters are estimated, the system presents them to the user together with the corresponding synthesized audio, note sequence, and instrument type through the GUI. In the music audio editing stage, the user edits the note sequence or instrument type of a selected source. Given the edited conditions, the diffusion-based estimator regenerates the synthesis parameters of the selected source, while the parameters of the other sources are left unchanged. The DDSPMM module then re-synthesizes the edited mixture from the updated synthesis parameters.

2.3. Music Audio Editing Examples

Figure 2 illustrates examples of music audio editing using the proposed system. The DDSP autoencoder and diffusion model were pretrained on the instrument performances in the University of Rochester multimodal music performance dataset [10], following the training setup in [6]. For a given mixture of two flute sounds, the system was used to edit only one part while leaving the other part unchanged. In the instrument conversion example, the blue part was converted from flute to violin, resulting in stronger high frequency harmonic components. In the note-level phrasing editing example, one of the F_0 trajectories was changed according to the edited note sequence. These examples illustrate that the proposed system can apply source-level and instrument-level edits to an individual source in a mixture.

3. CONCLUSION

We presented a music audio editing system that enables source-wise editing of a harmonic sound mixture based on the DDSPMM and the synthesis-parameter-domain diffusion model. The system first estimates source-wise synthesis parameters from an observed mixture and then allows users to edit individual sources by modifying score-level conditions such as note sequences and instrument types. By combining the interpretability of DDSP with the generative capability of the diffusion model, the system supports controllable music audio editing through source-wise parameter estimation and regeneration.

4. REFERENCES

[1] M. Huzaifah and L. Wyse, "Deep generative models for musical audio synthesis," in *Handbook of Artificial Intelligence for Music*. Springer, 2021.

[2] S. Li, S. Ji, Z. Wang, S. Wu, J. Yu, and K. Zhang, "A survey on music generation from single-modal, cross-modal, and multi-modal perspectives: Data, methods, and challenges," *ACM Computing Surveys*, vol. 58, no. 11, 2025.

[3] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proc. Int. Conf. Learn. Representations*, 2020.

[4] Y. Wu, E. Manilow, Y. Deng, R. Swavely, K. Kastner, T. Cooijmans, A. Courville, C.-Z. A. Huang, and J. Engel, "MIDI-DDSP: Detailed control of musical performance via hierarchical modeling," in *Proc. Int. Conf. Learn. Representations*, 2022.

[5] B. Hayes, J. Shier, G. Fazekas, A. McPherson, and C. Saitis, "A review of differentiable digital signal processing for music and speech synthesis," *Front. Signal. Process.*, vol. 3, 2023.

[6] K. Takemoto, T. Nakamura, and H. Saruwatari, "Differentiable digital signal processing mixture model-guided diffusion for synthesis parameter estimation from harmonic sound mixtures," 2026, submitted to Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.

[7] T. Prätzlich, J. Driedger, and M. Müller, "Memory-restricted multiscale dynamic time warping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 569–573.

[8] M. Kawamura, T. Nakamura, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Differentiable digital signal processing mixture model for synthesis parameter extraction from mixture of harmonic sounds," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 941–945.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[10] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multi-track classical musical performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 522–535, 2019.