DATA AUGMENTATION FOR SINGING VOICE SEPARATION USING MUSICAL INSTRUMENT TRANSFER AND RESYNTHESIS

Jaekwon Im
GSCT, KAIST
jakeoneijk@kaist.ac.kr

Eunjin Choi
GSCT, KAIST
jech@kaist.ac.kr

Wootaek Lim
GSCT, KAIST
wtlim@kaist.ac.kr

ABSTRACT

The singing voice separation is to separate the vocal and accompaniment from the music mixture. In recent years, the singing voice separation system has been improved significantly with advances in deep learning. However, despite these advances, the performance of source separation is still limited. Therefore, in this paper, we propose a new data augmentation method that maintains the musical structure to improve the performance of sound source separation. According to the experimental results, it is confirmed that the proposed method outperforms the baseline separation model.

1. INTRODUCTION

Music Information Retrieval (MIR) is the area of analyzing music information. This field is growing rapidly because we consume a lot of audio and music in our daily lives. In the field of MIR, lots of background knowledge is used together, such as musicology, psychoacoustics, psychology, education, signal processing, informatics, machine learning, computational intelligence, and music analysis [1]. Among them, music analysis is one of the largest research fields in MIR. Generally, popular commercial music is characterized by vocal and accompaniment parts. This is an oversimplification of music in a variety of genres, but we focus on the analysis of vocal melodic lines and accompaniment [2]. However, these two components have completely different characteristics and need to be analyzed separately. In this aspect, audio source separation is the process of separating a mixture into isolated sounds from individual sources. The sound source extracted through the source separation model can be used for various audio analysis tasks such as speech recognition, MIR, audio remixing, etc. The extracted audio sources from mixed audio are often used in MIR tasks because the separated source enables more accurate analysis and it has various applications. Therefore, people want to separate the mixed audio into a stem file, using some techniques such as source separation. However, since most source separation models extract the sound source with distortion, artifacts, or reverb, there are limitations. Therefore, several approaches have been proposed to address these problems, but research is still ongoing due to performance limitations.

In this paper, we propose a novel data augmentation method using audio synthesis which utilizes the differentiable digital signal processing (DDSP) network to improve the separation performance. The U-Net based singing voice separation network was used as a baseline. The performance of the proposed model was evaluated on the pop music dataset and verified the generalization performance using cross-dataset evaluation.

2. RELATED WORKS

2.1 Audio Source Separation

Music recordings are usually a mix of several individual instrument tracks. On the contrary, the goal of music source separation is to separate the original single source from the audio mixture. Audio source separation has been studied for decades, and there are many different approaches. Recently, neural network-based models that outperform conventional methods have been proposed [3-9]. These methods perform source separation in the spectrogram domain [3] or end-to-end separation in the waveform domain [4]. In addition, various approaches have been studied to improve the performance of source separation [5-9]. However, it is still limited. Therefore, in this paper, we proposed a method to improve source separation quality by using a new augmentation method that can be used without any model structure or separation domain constraints.

2.2 Data Augmentation in Source Separation

Data augmentation is a common strategy for improving the performance of deep neural networks. In MIR, augmentation involves audio specific modification such as pitch shift, time stretch, loudness, frequency filter, and mixing [10]. Specifically, for source separation tasks, several studies investigated mixing related augmentation. For example, the random-mixing augmentation method is proposed in [11]. It randomly swapped left/right channels for each instrument, scaled the amplitude, chunked into sequences for each instrument, and mixed instruments from different songs. However, the result of random mixing augmentation is not realistic, since it does not consider the musical relations between the tracks. In [12],

musically coherent mixing specific data augmentation method for violin/piano source separation was proposed. They used chroma distance and correlation-based pairing to consider the tonic, harmonic and rhythmic relations between the real paired stems. It also utilized modern music production routine to data augmentation, such as equalization, contrast, reverb, and addition of pink noise. However, these augmentation methods are limited in that they are mainly performed by selecting correlated pairs of stem files or giving some variation to the existing audio stem files. Our approach differs in that we augment our dataset with new audio, which is a rearranged version of the original song. By taking advantage of the powerful timbre transfer performance and expressiveness of DDSP, various timbre transformations are possible for each instrument which enables the model to be trained with more diverse data. Moreover, this method can augment the data without harming any tonic, harmonic and rhythmic coherence between the tracks.

3. PROPOSED METHOD

3.1 Data Augmentation using DDSP

Audio synthesis has many practical applications for creative sound design in multimedia production such as music and movies [13]. Conventionally, the acoustic modeling-based spectral modeling synthesis (SMS) approach has been widely accepted. SMS decomposes a target source into sinusoids and a residual using Short-time Fourier transform (STFT) and synthesizes it using sinusoidal oscillators and filtered noise [14]. DDSP is a neural audio synthesis model based on spectral modeling synthesis. It enables an interpretable modular approach to generative modeling without sacrificing the benefits of deep neural networks. This method can directly integrate existing signal processing elements with deep learning and the model is trainable but has a partially deterministic block structure [15].

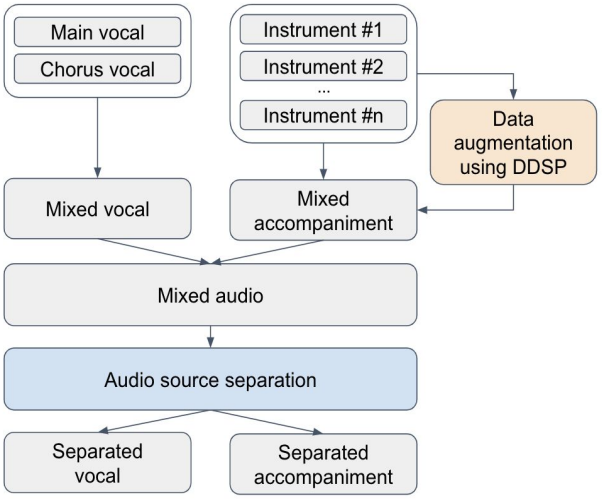

Figure 1. Block diagram of proposed data augmentation and singing voice separation method.

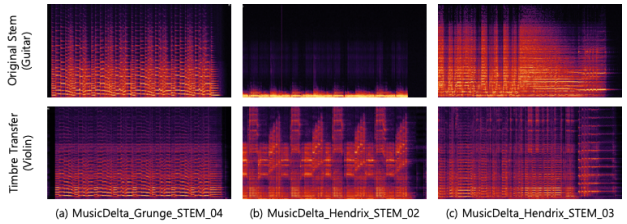

Figure 2. Spectrogram comparison between original stem and timbre transferred audio.

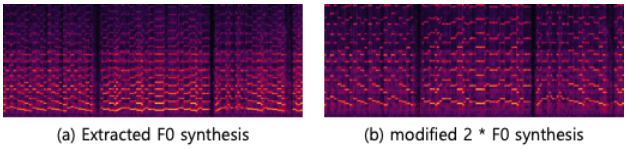

Figure 3. Example of extracted fundamental frequency (F0) resynthesis.

In this work, we utilized the expressive and realistic sound synthesis ability of DDSP as a timbre transfer method of stem files, and remixed them to generate a new audio mixture to be separated. The block diagram of the proposed method is shown in Figure 1. The data generation process is as follows:

(1) Select monophonic stem files from dataset
(2) F0 detection using CREPE [16]
(3) Synthesis of timbre transferred stem file by using pre-trained DDSP with 13-min violin performance [15]
(4) Remixing the timbre transferred stem files with other tracks

A stem file refers to each sound source file that composes a song such as vocal, drum, guitar, and piano. Among them, we considered only the monophonic (e.g. bass, solo guitar, trumpet) instrument for timbre transfer. Also, the chorus track was not selected for DDSP augmentation since we considered chorus as a vocal. For more variability, we also controlled the fundamental frequency (F0) input of the DDSP; F0 augmented tracks have pitches that are one octave higher (2*F0) than the original stem file. For all songs in the training dataset, remixed accompaniments were generated by mixing all augmented tracks with other tracks. An example of data augmentation is shown in Figures 2 and 3.

3.2 U-Net based Singing Voice Separation

U-Net is an architecture composed of encoder-decoder structure with symmetric skip connections. These skip connections help the model to get local information at the decoding stage. Because of this advantage, the U-Net based model showed promising performance in singing voice separation task [3]. Therefore, we used U-Net structure as a backbone architecture to test our data augmentation method. It takes fixed 128 time frames segments of the spectrogram of the mixture music and generates masks the same size as the input. Figure 4 shows the structure of a U-Net based separation network.

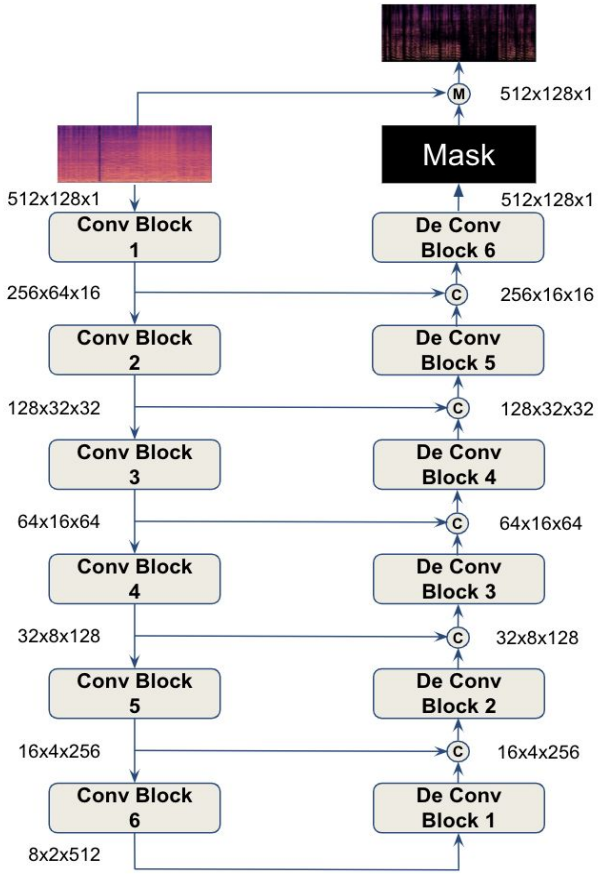

Figure 4. Structure of U-Net based Singing Voice Separation Network.

The loss function for training is the $L_{1,1}norm$ of the difference of the ground truth vocal spectrogram and the element-wise multiplication of inputs and output masks. We used a soft mask (in range [0, 1]) when training the model, and a hard mask (binary) when testing.

4. EXPERIMENTS

4.1 Dataset

MedleyDB is a music database mainly used in the field of MIR. It contains various metadata and annotations along with multi-channel recordings [17]. Therefore, we used the medley dataset in the experiment. For data preparation, we selected 56 audio clips from MedleyDB containing vocal stems and instrument stems for our task. Among them, 48 clips were used for training, and the remaining 8 were used for testing. Moreover, we also selected 105 stem files that have a monophonic sound for data augmentation as mentioned in section 3.1. This is because we used a pre-trained violin DDSP model. In addition, we used a k-pop cross-DB to verify the reliability of the proposed model. This dataset was collected by selecting 8 songs with separated vocal and accompaniment tracks in k-pop music. In the experiments, the audio is resampled to 16kHz sampling rate and down-mixed to mono channel.

4.2 Hyperparameters and settings

The hyperparameters and settings are listed in Table 1. The convolution filter, stride, zero-padding size were set to 5, 2 and 3, respectively. The output of the encoder is normalized by batch normalization and passed through leaky rectified linear units (ReLU) with leakiness 0.2. The output of the decoder is also batch normalized and passed through ReLU activation. 50% dropout was applied to each of the first three decoder blocks. The model was trained with the Adam optimizer. We used validation loss for early stopping and the maximum epoch was set to 700. The batch size and learning rate were set to 16 and 0.001, respectively.

Parameter	Value
Convolution filter size	5 × 5
Activation functions	Leaky ReLU / ReLU
Optimizer	Adam (betas=(0.9, 0.99))
Epochs	700
Batch size	16

Table 1. Hyperparameters and settings for the proposed U-Net based singing voice separation model.

4.3 Experimental Results

To demonstrate the performance of the proposed approach, the U-Net-based singing voice separation model was used as the baseline. The input signal of the network is a mixture of all stem files, and the target signal is the vocal track. The audio signal was converted into a spectrogram using the STFT with a frame size of 1024 and an overlap length of 75%. Then the spectrogram was cropped to 512 x 128 images for network training. The dataset and hyperparameters for training and evaluation are the same as described in 4.1 and 4.2. The performance of the proposed method was verified using three metrics: source-to-distortion ratio (SDR), source-to-artifacts ratio (SAR), and source-to-interference ratio (SIR) [18].

First, we evaluated the performance of the U-Net based baseline and the proposed augmentation method using MedleyDB. We tested two augmentation methods. One is the augmentation using instrument transfer, and another is the method of adding the F0 change. As shown in Table 2, the proposed model achieved 0.95 dB, 0.53 dB, and 2.27 dB better in SDR, SAR, and SIR, respectively, when compared to the baseline system. This confirms that the proposed augmentation methods improve the singing voice separation performance. Moreover, one of the advantages that can be obtained through augmentation is the improvement of model robustness. Therefore, we evaluated the generalization performance of the proposed method using a completely different k-pop cross-DB. As shown in Table 3, the two augmentation methods also outperformed the baseline.

Medley DB		**Baseline**	**Data aug.** (transfer)	**Data aug.** (transfer + 2F0)
SDR (dB)	vocal	7.15	7.99	8.10
	accom.	4.15	5.01	5.21
SAR (dB)	vocal	8.26	8.78	8.79
	accom.	5.17	6.12	6.33
SIR (dB)	vocal	15.32	16.76	17.59
	accom.	12.62	12.78	12.92

Table 2. The evaluation of separation performance using MedleyDB.

K-pop DB (Cross-DB)		**Baseline**	**Data aug.** (transfer)	**Data aug.** (transfer + 2F0)
SDR (dB)	vocal	10.35	10.97	10.56
	accom.	2.59	3.17	2.84
SAR (dB)	vocal	10.60	11.27	10.79
	accom.	4.02	4.44	4.30
SIR (dB)	vocal	24.26	24.01	24.93
	accom.	10.24	11.15	10.41

Table 3. The evaluation of separation performance using K-pop DB.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed the data augmentation method for singing voice separation using musical instrument transfer and resynthesis. The DDSP was used for augmentation, and a new audio stem file was generated by transferring timbre and fundamental frequency. The U-Net based singing voice separation network was used as a baseline and the effect of the proposed augmentation method was evaluated on the MedleyDB and K-pop cross-DB. As a result, our proposed augmentation method outperforms the baseline in the aspect of objective evaluation metrics. In the future, because the dataset we used is still not large, so the proposed augmentation method should be verified using a larger dataset. Moreover, if the DDSP model is trained with a variety of instruments (e.g. flute and trumpet), the performance of augmentation can be increased. Also, to expand the available instruments for timbre transfer, using vocoder for vocals, utilizing transcription algorithm and synthesizer for polyphonic instruments such as piano, could be helpful.

6. AUTHOR CONTRIBUTIONS

Jaekwon Im, Eunjin Choi and Wootaek Lim contributed equally to the study. We contributed together in brainstorming, dataset preparation, research progress, experimentation, paper writing, and presentation.

7. REFERENCES

[1] "Music information retrieval," https://en.wikipedia.org/wiki/Music_information_retrieval

[2] P. Tagg: "Analysing popular music: theory, method and practice," *Popular Music*, Vol. 2, pp. 37–67, 1982.

[3] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde: "Singing voice separation with deep u-net convolutional networks," *in Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[4] Y. Luo and N. Mesgarani: "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Trans. Audio, Speech, Language Process, Vol. 27, No. 8, pp. 1256–1266, 2019.

[5] D. Stoller, S. Ewert, and S. Dixon: "Wave-U Net: A multi-scale neural network for end-to-end audio source separation," *in Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[6] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis: "Two-step sound source separation: Training on learned latent targets," *in Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[7] Y. Luo, Z. Chen, and T. Yoshioka: "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," *in Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[8] A. Defossez, N. Usunier, L. Bottou, and F. Bach: "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint*, 2019.

[9] N. Zeghidour and D. Grangier: "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint*, 2020.

[10] J. Schlüter and T. Grill: "Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks," *in Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[11] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji: "Improving Music Source Separation based on Deep Neural Networks through Data Augmentation and Network Blending," *in Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[12] C. Chiu, W. Hsiao, Y. Yeh, Y. Yang and A.W. Su: "Mixing-Specific Data Augmentation Techniques for Improved Blind Violin/Piano Source Separation," *in Proc. of the IEEE International Workshop on Multimedia Signal Processing (MMSP),* 2020.

[13] C. Donahue, J. McAuley, and M. Puckette: "Adversarial audio synthesis," *in Proc. of the International Conference on Learning Representations (ICLR)*, 2019.

[14] X. Serra and J. O. Smith: "Spectral modeling and synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, Vol. 14, No. 4, pp. 12–24, 1990.

[15] J. Engel, L. Hantrakul, C. Gu and A. Roberts: "DDSP: Differentiable digital signal processing," *in proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[16] J. Kim, J. Salamon, P. Li, and J. P. Bello: "Crepe: A convolutional representation for pitch estimation". *in Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[17] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello: "MedleyDB: A Multitrack dataset for annotation intensive MIR research," *in Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[18] E. Vincent, R. Gribonval, and C. Fvotte: "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp. 1462-1469, Jul. 2006.