Morphing of Drum Loop Sound Sources Using CNN-VAE

Mizuki Kawahara[1], Tomoo kouzai[1], and Tetsuro Kitahara[1⋆]

College of Humanities and Sciences, Nihon University, Japan
{kawahara,kouzai,kitahara}@kthrlab.jp

Abstract. In this paper, we attempt a morphing technique that combines a convolutional neural network (CNN) and a variational autoencoder (VAE) in order to produce a variety of sound sources of drum loops. Although there have already been studies related to sound or music morphing, and some of them have focused on drum sound synthesis, morphing of sound sources of drum loops has not been attempted. Our system trains the spectrograms of the drum loop sound sources using CNN-VAE and generates a new source by interpolating two sources in the latent space. Preliminary experiments using commercially available sound sources show promising results.

Keywords: morphing, spectrogram, convolutional neural networks (CNN), variational autoencoder (VAE), drum sound source

1 Introduction

A loop sequencer is commonly used in music production, with which the creator concatenates and mixes various loop sound sources. However, it is often difficult to find the sound sources that they desire from a limited set of sound sources. For this reason, research has been conducted to generate a variety of sound sources by morphing. For example, Primavera et al. [1] proposed a method to achieve smooth transitions between different sound sources in sound morphing. Nistal et al.[2] and Aouameur et al.[3] proposed a method for the synthesis of drum sounds, in which they have explored methods for extracting features of drum sounds and generating new sound sources based on the extracted features.

In this paper, we propose a sound sources morphing method that combines a convolutional neural network (CNN) and a variational autoencoder (VAE). First, the features of given drum loop sound sources are extracted using a CNN and are mapped to the latent space using a VAE. Then, a new sound source is generated by morphing two given sound sources in the latent space.

⋆ This work was supported by JSPS Kakenhi Nos. JP22H03711 and JP21H03572.

2 Proposed System

Our proposed method uses a model based on CNN and VAE (hereinafter referred to as CNN-VAE) to achieve morphing in a low-dimensional latent space representing the features of sound sources.

In the training phase, the sound sources are first transformed into spectrograms by the Fourier transform. Then, a convolutional layer is used to map the sound sources into a low-dimensional latent space. Then, the inverse convolution layer is used to reconstruct the spectrogram of the original sound source. The CNN-VAE model is trained so that the reconstructed spectrogram is equivalent enough to the original spectrogram.

At the generation phase, two sound sources are selected from the trained one. A new vector in the latent space is generated by interpolating the two vectors corresponding to the selected sources. Then a spectrogram is generated by using the decoder.

2.1 Generation of spectrograms

An input audio signal is transformed into a spectrogram with the short-time Fourier transform (STFT). A Hamming window is used with a window width of 2048 and a hop size of 1/4 of the window length. Since the sampling frequency is assumed to be 22050 Hz and the length of the audio signal is about 3.43 seconds (two measures at 140 BPM), the spectrogram is a 1025-by-148 matrix.

2.2 Building a CNN-VAE model

The CNN-VAE model is built and trained with given spectrograms of drum loop sound sources. This model has an encoder consisting of three convolution layers to compress spectrograms and them to the 16-dimensional latent space. The decoder consists of three deconvolution layers that reconstruct spectrograms from the 16-dimensional latent vector. The ReLU function is used as the activation function, the batch size is 64, and the number of epochs is 3000. The mean square error is used as the loss function.

2.3 Morphing

The user selects two sound sources (denoted as s_i and s_j) from the trained ones. Let $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ be the latent vectors obtained from s_i and s_j, respectively. Then, a new latent vector $\boldsymbol{z}_{\mathrm{new}} = (1-\alpha)\boldsymbol{z}_i + \alpha\boldsymbol{z}_j$ is calculated, in which $\boldsymbol{z}_{\mathrm{new}}$ represents an internally dividing point of $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ in the ratio $\alpha : 1-\alpha$. Finally, the spectrogram is transformed into an audio signal by inverse Fourier transform and phase restoration.

3 Preliminary Experiment

3.1 Method

We trained our model with 74 drum loop sound sources. The morphing parameters α were set to 0.00, 0.25, 0.50, 0.75, and 1.00. The 74 drum sounds were taken from a commercial loop sound dataset "Sound PooL vol.2" [1](genre: Techno & Trans).

[1] https://www.ah-soft.com/soundpool/

Table 1. Pairs of sound sources used in the experiment

degree of similarity	0.0013	0.2505	0.5101	0.7501
sound source pair	s17, s24	s15, s28	s34, s73	s08, s51

Table 2. Morphing result (similarity to original source)

(a) s17 — s24

α	0.00	0.25	0.50	0.75	1.00
s17	1.0000	0.8510	0.1120	0.0041	0.0013
s24	0.0013	0.0211	0.3877	0.9331	1.0000

(b) s15 — s28

α	0.00	0.25	0.50	0.75	1.00
s15	1.0000	0.9221	0.6264	0.3368	0.2505
s28	0.2505	0.4165	0.6708	0.8931	1.0000

(c) s34 — s73

α	0.00	0.25	0.50	0.75	1.00
s34	1.0000	0.5736	0.5838	0.5830	0.5101
s73	0.5101	0.6714	0.6954	0.9250	1.0000

(d) s08 — s51

α	0.00	0.25	0.50	0.75	1.00
s08	1.0000	0.9492	0.8218	0.7596	0.7501
s51	0.7501	0.7818	0.7873	0.9773	1.0000

After these sound sources were trained, pairs were selected for morphing. The selected pairs are shown in Table 1. The "s + number" (e.g., s01, s02, $\cdots$) represents sound sources. The pairs of sound sources used for morphing were selected so that the cosine similarity of the spectrograms is 0.00, 0.25, 0.50, and 0.75 in order to try a wide range of pairs, including those similar and dissimilar to each other.

When α is close enough to 0 (or 1), the generated sources should become similar to s_i (or s_j) as α. We confirm this by calculating the cosine similarity between the spectrograms of the generated sources and the original sources.

3.2 Results

The results of the morphing are shown in Table 2. Some of the generated results are posted at the following URL: https://sites.google.com/kthrlab.jp/en-drum-morphing

It is confirmed that the original sound sources are reconstructed with sufficient accuracy for α=0.00 and α=1.00. In the cases of α=0.25 and α=0.75, the similarity is intermediate between the values when $\alpha = 0.00$ and when $\alpha = 1.00$. This indicates that generated sound sources that contain both features of the two sources.

For the pairs in Table 2 (b) and Table 2 (c), when α=0.50, sound sources with low similarity to both original sources were generated. It implies the possibility that our model can generate novel sources. In fact, Fig 1 shows that the spectrogram with $\alpha = 0.50$ is different from those with $\alpha = 0.00$ and $\alpha = 1.00$.

4 Conclusion

In this paper, we proposed a CNN-VAE model to achieve sound source morphing in the latent space. When we changed α (a morphing ratio), the model generated different sound sources accordingly. Future work includes larger-scaled experiments with various sound source pairs and subjective evaluation through listening experiments.

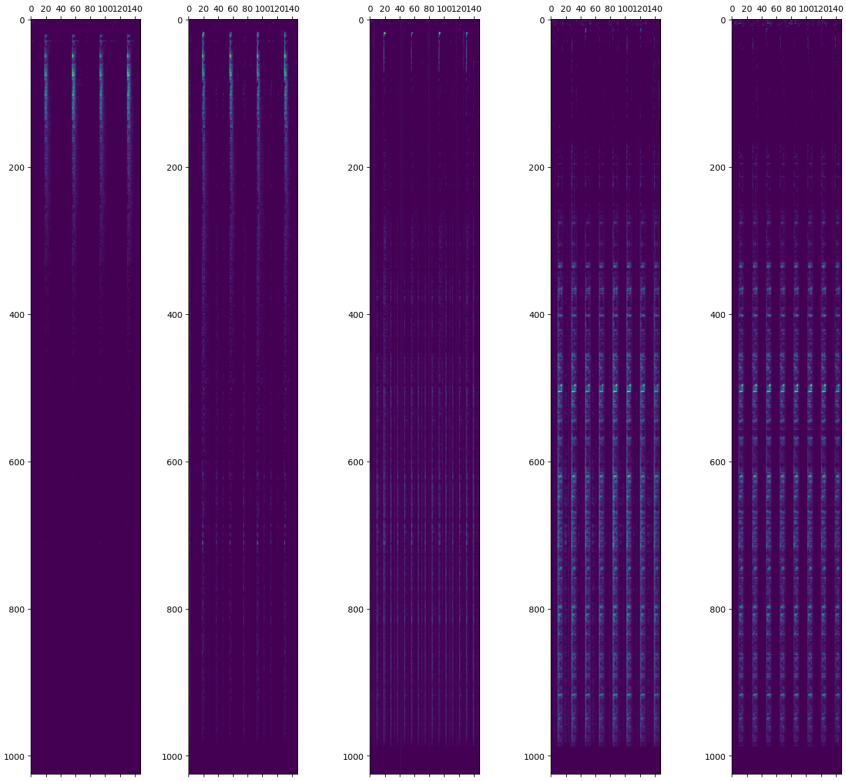

Fig. 1. Spectrogram of morphed source from s17 and s24 (from left to right $\alpha = 0.00, 0.25, 0.50, 0.75, 1.00$)

References

1. Andrea Primavera, Francesco Piazza, and Joshua D. Reiss: Audio Morphing for Percussive Hybrid Sound Generation, *Proceedings of the 45th AES Conference* (2012).
2. Javier Nistal, Stefan Lattner, and Gaël Richard: DrumGAN: Synthesis of Drum Sounds With Timbral Feature Conditioning Using Generative Adversarial Networks, arXiv:2008.12073 (2020).
3. Cyran Aouameur, Philippe Esling, and Gaëtan Hadjeres: Neural Drum Machine: An Interactive System for Real-time Synthesis of Drum Sounds, arXiv:1907.02637, (2019).