Humming-to-Instrument Conversion based on CycleGAN

1st Wen-Hsing Lai
dept. of Computer and Communication Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan
lwh@nkust.edu.tw

2nd Siou-Lin Wang
Ph.D. Program in Engineering Science and Technology, College of Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan
0015901@nkust.edu.tw

3rd Zhi-Yao Xu
dept. of Computer and Communication Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan
f108110130@nkust.edu.tw

2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) | 978-1-6654-1951-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISPACS51563.2021.9651076

***Abstract*—In this research, we propose a humming-to-instrument conversion system based on cycle-consistent adversarial networks (CycleGAN) to convert human humming to the sound of viola. This research adjusts the weight of cycle-consistency loss and identity loss to successfully convert the sound. From the experimental results of objective RMSE, the converted audio is more similar to viola compared to the similarity to humming. From the results of subjective MOS, the quality of the converted sound is fair to listeners.**

Keywords—CycleGAN, humming-to-instrument, audio conversion, viola

I. Introduction

Voice conversion (VC) is a technique for converting one voice into another voice with a different timbre. It consists parallel and non-parallel methods. For example, parallel voice conversion systems based on Gaussian mixture model [1] and neural network [2] have been developed. On the other hand, non-parallel voice conversion can be divided into categories of feature disentangle [3] and direct transformation. One example of direct transformation is cycle-consistent adversarial networks (CycleGAN) based non-parallel voice conversion, named CycleGAN-VC [4], which is a voice conversion system based on CycleGAN.

In this research, a non-parallel humming-to-instrument conversion system based on CycleGAN-VC is built to perform humming-to-viola.

II. Proposed Method

CycleGAN-VC is able to transform audio without feature extraction and learn the input-to-output mapping without relying on parallel data.

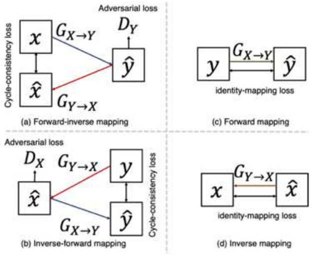

Fig. 1. The training process of CycleGAN-VC.

The training process of CycleGAN-VC is shown in Fig. 1. Three types of losses, namely adversarial loss, cycle-consistency loss and identity-mapping loss, are used to control the process of conversion. $G_{X\to Y}$ is the forward generator, which converts X to Y; $G_{Y\to X}$ is the inverse generator, which converts Y to X. D_X and D_Y are the discriminators of X and Y respectively. Fig. 1 (a) and (b) are forward-inverse and inverse-forward mapping respectively. Both mapping contain adversarial loss and cycle-consistency loss, and their main work is to find the optimal pseudo pair from the unpaired audio data. Fig. 1 (c) and (d) are forward mapping and inverse mapping. Both mapping use identity-mapping loss to retain linguistic information.

978-1-6654-1951-2/21/$31.00 ©2021 IEEE

The architectures of the generator and discriminator of CycleGAN-VC are shown in Fig. 2 and Fig. 3.

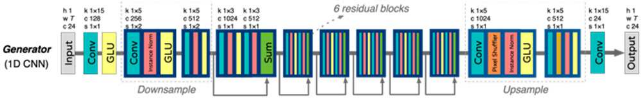

Fig. 2. The network architecture of generator.

In Fig. 2, CycleGAN-VC uses one-dimensional convolutional neural network (CNN) to design the generator. In the input and output layers, *h* represents for the height, *w* represents for the width, and *c* represents for the number of channels. In each convolutional layer, *k* represents for the kernel size, and *s* represents for the stride size. Since the generator is fully convolutional, it can accept inputs of arbitrary length *T*.

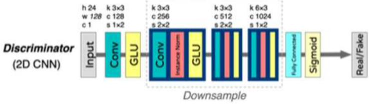

Fig. 3. The network architecture of discriminator.

In Fig. 3, the discriminator is designed as a two-dimensional CNN to focus on spectral texture. CycleGAN-VC refers to the work of Johnson et al. [5] and adds down-sampling, residual, and up-sampling layers to the generator and discriminator. Pixel shuffler for up-sampling is used. In addition, like CycleGAN, CycleGAN-VC also uses instance normalization.

Suppose we have an audio source x that we want to convert into a target audio y, where $x \in X$ and $y \in Y$. The adversarial loss, cycle-consistency loss, and identity-mapping loss are defined as in Eqs. (1)-(4).

$$\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = \mathbb{E}_{y\sim P_{Data(y)}}[logD_Y(y)] + \mathbb{E}_{x\sim P_{Data(x)}}\left[\log\left(1 - D_Y\left(G_{X\to Y}(x)\right)\right)\right] \quad (1)$$

$$\mathcal{L}_{adv}(G_{Y\to X}, D_X) = \mathbb{E}_{x\sim P_{Data(x)}}[logD_X(x)] + \mathbb{E}_{y\sim P_{Data(y)}}\left[\log\left(1 - D_X\left(G_{Y\to X}(y)\right)\right)\right] \quad (2)$$

Eqs. (1) and (2) are the adversarial losses. In Eq. (1), discriminator D_Y aims not to be deceived by maximizing the loss, and $G_{X\to Y}$ aims to generate indistinguishable audio by minimizing the loss.

$$\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{x\sim P_{Data(x)}}[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1] + \mathbb{E}_{y\sim P_{Data(y)}}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1] \quad (3)$$

Eq. (3) is the cycle-consistency loss, where $G_{X\to Y}$ is the forward generator and $G_{Y\to X}$ is the inverse generator. The purpose of cycle-consistency loss is to measure the difference of the input audio and the audio after forward-inverse or inverse-forward.

$$\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{y\sim P_{Data(y)}}[\|G_{X\to Y}(y) - y\|_1] + \mathbb{E}_{x\sim P_{Data(x)}}[\|G_{Y\to X}(x) - x\|_1] \quad (4)$$

Eq. (4) is the identity-mapping loss of $G_{X\to Y}$ and $G_{Y\to X}$. It encourages the generator to preserve composition between the input and output. Finally, the full objective loss is shown in Eq. (5) as follows.

$$\mathcal{L}_{full} = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) + \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) + \lambda_{id}\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) \quad (5)$$

The weighting of cycle-consistency loss and identity-mapping loss are controlled by λ_{cyc} and λ_{id} respectively. Since humming and viola are very different types of sounds, and the cycle-consistency loss aim to encourage $G_{X\to Y}$ and $G_{Y\to X}$ to find (x, y) pairs with the same contextual information, if the weighting of cycle-consistency loss is increased, the generator may fail to preserve composition between the input and output. Therefore, unlike the original design of CycleGAN-VC, which increases the weight of cycle-consistency loss, the proportion of cycle-consistency loss in this research is lower than the identity loss, so that the identity loss is more important in the process of loss convergence, and the composition and timbre of the viola can be preserved in the humming to viola. In this research, we use λ_{cyc}=1 and λ_{id}=5.

III. Experiments

A. Corpus

MIR-QBSH dataset is used, which contains 48 midi files and 4431 humming clips. We selected 22 MIDIs and used the SONAR synthesizer developed by Cakewalk to convert them into viola in 16kHz. In addition, 22 humming clips (including 17 humming and 5 singing clips), which are recorded by the same singer, were included in the experiment. The songs of the 22 humming and viola were the same, but the length of the humming was shorter. Each humming clips is the first 8s of the corresponding song, and the total length of the 22 songs is 2 minutes and 56 seconds. The viola audio is the entire song with lengths ranging from 11 seconds to 201 seconds, and the total length of the 22 songs is 13 minutes and 47 seconds.

B. Evaluation Measures

In this research, both objective and subjective measures are applied. Root-Mean-Square Error (RMSE) of the Mel-frequency Cepstrum Coeffients (MFCCs) is used as the objective measure. The dimension of MFCC is 14. The analysis frame length is 1024 samples, and the frame shift is 256 samples. Since the RMSE must be calculated in the same length, the original viola is cut to fit the length.

In addition to RMSE, we use the Mean Opinion Score (MOS) as the subjective measure. The MOS scores are 1 to 5, meaning 'Bad', 'Poor', 'Fair', 'Good', and 'Excellent'. Ten converted audio were used and five listeners attended. Listeners were asked to rate the audio based on its clarity, timbre, melody, and similarity to viola.

C. Experimental Results

TABLE I. The average RMSE of the converted audio versus the original viola / humming

	vs. Original Viola	vs. Original Humming
Converted	1.2373	1.6274

From Table I, the average RMSE is 1.2373 when the converted is compared with the original viola, and the average RMSE is 1.6274 when the converted is compared with the original humming. Hence, the converted audio is more similar to the original viola.

TABLE II. The MOS scores of the converted audio

	Bad	Poor	Fair	Good	Excellent
Listener 1	0	2	7	1	0
Listener 2	0	1	8	1	0
Listener 3	0	0	10	0	0
Listener 4	0	2	8	0	0
Listener 5	0	1	9	0	0

Table II shows the MOS of 5 listeners and 10 converted audio. From Table II, the converted viola has a higher percentage of Fair, accounting for 84%. Converted audio received an average MOS of 2.92, which is close to Fair. From the MOS results, the conversion from humming to viola is successful, however, there is still room for improvement in the converted quality.

IV. Conclusion

In this research, we built a CycleGAN based humming-to-instrument conversion system to convert human humming to viola. RMSE and MOS are used as objective and subjective measures. From the results of RMSE and MOS, the converted audio is more close to viola than to humming, and the quality of the converted sound is fair to listeners.

References

[1] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, November 2007.

[2] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice Conversion Using RNN Pre-Trained by Recurrent Temporal Restricted Boltzmann Machines," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 580–587, March 2015.

[3] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet Vocoder with Limited Training Data for Voice Conversion," Interspeech, September 2018, pp. 1983–1987.

[4] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," European Signal Processing Conference, September 2018, pp. 2100–2104.

[5] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," arXiv:1603.08155 [cs], March 2016, Accessed: Jun. 05, 2021. [Online]. Available: http://arxiv.org/abs/1603.08155