Singing Style Transfer Using Cycle-Consistent Boundary Equilibrium Generative Adversarial Networks

Cheng-Wei Wu 1,2  Jen-Yu Liu 1,2  Yi-Hsuan Yang 2  Jyh-Shing R. Jang 1

Abstract
Can we make a famous rap singer like Eminem sing whatever our favorite song? Singing style transfer attempts to make this possible, by replacing the vocal of a song from the source singer to the target singer. This paper presents a method that learns from unpaired data for singing style transfer using generative adversarial networks.

1. Introduction
Figure 1 shows a two-stage framework combining singing voice separation with singing style transfer. The audio from a source singer is separated into accompaniment and vocal first, and then the singing style of the separated vocal is changed to that of a target singer. Finally, the separated accompaniment is integrated with the style-transferred vocal. Our focus here is on the singing style transfer model. We assume that the input to this model is clean, meaning that it has been perfectly separated from the source audio.

We evaluate the proposed model through a listening test under the inside test scenario, meaning that the training and test sets have overlaps. We show audio result for both inside and outside tests on our project website: http://mirlab.org/users/haley.wu/cybegan/.

2. Methodology
The method proposed by Gatys et al. (2016) leads to the earliest successful examples of image style transfer. The method has been extended to perform audio style transfer, either to transfer sound texture (Ulyanov & Lebedev, 2016) or to stylize an acapella cover to match the original vocal (Bohan, 2017). However, the method requires the availability of paired data for training (i.e., the source and target singers need to have sung the same songs), which is not readily available for arbitrary target singers.

We experiment with a few extensions of this basic model. First, we use a U-Net architecture (Ronneberger et al., 2015) and add symmetric skip connections between the encoding and decoding layers to preserve the details of the given source audio and to increase the sharpness of the vocal. Second, we use BEGAN (Berthelot et al., 2017) instead of GAN in an attempt to stabilize the training process. Moreover, in this CycleBEGAN design, we replace the original PatchGAN-based discriminator (Li & Wand, 2016; Ledig et al., 2017; Isola et al., 2017) with an auto-encoder having identical architecture with the generator. Finally, being inspired by (Liu & Yang, 2018), we add a recurrent layer (using GRU (Cho et al., 2014)) before the output convolution layer to account for the temporal dependency in music.

3. Experiments
We train our model on the iKala dataset (Chan et al., 2015), which comes with 252 30-second excerpts of clean vocal recordings. We cut the excerpts into 5-second clips, with 4-second overlaps. We randomly select 2,800 segments from each gender as the training set, and 100 segments from each gender as the test set. As part of each test clip may have been observed in the training set, this is an inside test.
Singing Style Transfer Using Cycle-Consistent Boundary Equilibrium Generative Adversarial Networks

To simplify the problem, we train our model to perform "gender transfer", i.e., either male-to-female or female-to-male. We implement and compare in total five different models (denoted as m1–m5 in Figure 2), including the basic 'CycleGAN-CNN' and the most sophisticated 'CycleBEGAN-CNN+skip+recurrent.'

We conduct a subjective listening test to evaluate the performance of style transfer. The subjects are asked to listen to the source and transferred audio for six 15-second clips, including 3 clips for male-to-female and 3 clips for female-to-male. A subject has to compare the result of two randomly chosen models for each set, and evaluate the results in terms of the following six metrics, using a five-point Likert scale:

- **The sharpness** of the transferred singing voice.
- **Whether the lyrics** (main content of the song) is intelligible and is consistent with that in the source.
- **The perceived pitch** accuracy (relative pitch) of the transferred singing voice.
- **Naturalness:** whether the transferred singing voice is generated by human not by machine.
- **Whether the gender** of the singer has been changed.
- **Overall:** whether the transferred voice achieve the goal of style transfer and its overall perceptual quality.

Figure 2 shows the mean opinion scores (MOS) in these six metrics from 69 adults (19 females). The following observations are made. First, in terms of sharpness and lyrics, all the models with symmetric skip connections (m2, m4, m5) outperform those without them (m1, m3). This shows that the audio details propagated via the skip connections from the encoder to the decoder help generate high-resolution audio. Second, in terms of pitch, the CycleBEGANs (m3, m4, m5) outperform the simple CycleGAN model (m1), but not the CycleGAN with skip connections (m2). By listening to the result, we found that m2 works like a normal auto-encoder which only reconstructs the input without transferring the style. Therefore, it preserves the pitch well but performs very poorly in gender.

Next, in terms of naturalness, the outputs of the models with skip connections (m2, m4, m5) score higher than the others (m1, m3) as expected. Similar to pitch, the outputs of m2 sound more natural than the outputs from the other models, except for m5. Finally, the scores in gender is consistent with the result of the overall performance. This is expected, because the goal is to transfer the singing style male → female or female → male. CycleBEGAN models (m3, m4, m5) again outperform the CycleGAN ones (m1, m2), which verifies that the integration of CycleGAN with BEGAN effectively improves the result of style transfer. The most sophisticated model m5 outperforms the other models in all the metrics, reaching scores that are close to or over 4 on the five-point Likert scale. Figure 3 shows an example of the transferred spectrograms by m5.

4. Conclusion and Future Work

We have proposed a new approach for singing style transfer without paired data by combining CycleGAN with the training strategy of BEGAN. We have shown that symmetric skip connections increase the sharpness of the result, and that BEGAN plays an important role in transferring the singer properties. Future work will be done to transfer singer identities instead of only gender, and to include a singing voice separation model (Jansson et al., 2017; Liu & Yang, 2018) to deal with real songs instead of clean vocals.
Singing Style Transfer Using Cycle-Consistent Boundary Equilibrium Generative Adversarial Networks

References

Berthelot, David, Schumm, Tom, and Metz, Luke. BEGAN: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717, 2017.

Bohan, O. B. Singing style transfer, 2017. URL http://madebyoll.in/posts/singing_style_transfer/.

Chan, Tak-Shing, Yeh, Tzu-Chun, Fan, Zhe-Cheng, Chen, Hung-Wei, Su, Li, Yang, Yi-Hsuan, and Jang, Roger. Vocal activity informed singing voice separation with the ikala dataset. In Proc. IEEE. Int. Conf. Acoustics, Speech and Signal Processing, pp. 718–722, 2015.

Cho, Kyunghyun, van Merriënboer, Bart, Gülçehre, Çalar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In Proc. Conf. Empirical Methods in Natural Language Processing, pp. 1724–1734, 2014.

Griffin, Daniel and Lim, Jae. Signal estimation from modified short-time fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2):236–243, 1984.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. Image-to-image translation with conditional adversarial networks. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2414–2423, 2016.

Jansson, Andreas, Humphrey, Eric J., Montecchio, Nicola, Bittner, Rachel, Kumar, Aparna, and Weyde, Tillman. Singing voice separation with deep U-net convolutional networks. In Proc. Int. Soc. Music Information Retrieval Conf., pp. 745–751, 2017.

Ledig, Christian, Theis, Lucas, Huszar, Ferenc, Caballero, Jose, Cunningham, Andrew, Acosta, Alejandro, Aitken, Andrew, Tejani, Alykhan, Totz, Johannes, Wang, Zehan, et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690, 2017.

Liu, Chuan and Wand, Michael. Precomputed real-time texture synthesis with markovian generative adversarial networks. In Proc. European Conf. Computer Vision, pp. 702–716. Springer, 2016.

Liu, Jen-Yu and Yang, Yi-Hsuan. Event localization in music auto-tagging. In Proc. ACM Multimedia, pp. 1048–1057, 2016.

Liu, Jen-Yu and Yang, Yi-Hsuan. Denoising auto-encoder with recurrent skip connections and residual regression for music source separation. arXiv preprint arXiv:1807.01898, 2018.

Long, Jonathan, Shelhamer, Evan, and Darrell, Trevor. Fully convolutional networks for semantic segmentation. In Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition, pp. 3431–3440, 2015.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. In Proc. Int. Conf. Medical Image Computing and Computer-assisted Intervention, pp. 234–241. Springer, 2015.

Ulyanov, D. and Lebedev, V. Singing style transfer, 2016. URL https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/.

Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, and Efros, Alexei A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proc. IEEE Int. Conf. Computer Vision, 2017.