AUDIO STYLE TRANSFER

Aysenur Karagoz
ak121@rice.edu

Youssaf Menacer
ym31@rice.edu

Yuan Yang
yy87@rice.edu

November 10, 2021

1 Background

With the emergence of convolutional neural networks (CNN), neural style transfer (NST) for images becomes a very active topic in the literature and achieves popularity due to its applications in social media. The goal is to synthesis a high-quality artistic image by combining the content of an image with the appearance of well-known artwork. With the advent of CNN, revisiting this domain has resulted in great success. Particularly, the pre-trained object recognition CNN presented in [1] is a seminal work in this research area. A recent review on NST for images can be found in [2].

Although NST for images is a widely studied topic in the literature, NST for audios is a more recent research domain that is open to any development. Similar to NST for images, NST for audio aims to combine the content of audio with the style of another audio. However, the terms content and style are rather ambiguous for audios. While content can be defined as the notes of a music piece, style can be a music genre or an instrument type.

A spectrogram is a visual representation of the frequencies of a signal. Therefore, as input to this problem, content and style audio spectrograms are used. To transform an audio into a spectrogram, a common method is to use a short-time Fourier Transform (STFT), which has a drawback of a fixed resolution. As an alternative conversion method, continuous wavelet transform (CWT) can provide better results as it can use a size-adjustable window.

Introductory approaches to solving this problem are inspired by [3] and try to extend the model architectures used for image style transfer to the audio domain [4]. However, complex patterns of audio often lead to low-quality results. To capture the complexity, another approach is to use recurrent neural networks (RNNs), which provide poor results in distinguishing the content and style [5]. More promising results are obtained by using deep generative methods such as variational autoencoders (VAEs) ([6], [7], [8]) and generative adversarial networks (GANs) ([6], [9]) due to their flexibility to handling different styles of audio. Another approach that can provide promising results is using a wide-shallow-random network as discussed in [10] and [11].

2 Broader Scope and Goal

The concept of style transformation is general because it aims to learn mappings between entire fields, such as images or music. For example, style conversion is able to take any input from style A and change it so that it looks like style B. The styles that switch between A and B can be music genres, such as jazz and country music, or audios produced by different natural sources, such as sounds caused by the motion of liquid water or wind. Our project is inspired by the great success that image style transfer has achieved so far ([1], [12], [13]), and our team believe it is reasonable and riveting to extend similar approaches to realize NST in audio-formed data. Certain attention has been recently given to audio style transfer, and some studies have shown promising results ([6]-[11], [14]), but there is still much potential for further enhancement.

This project aims to build a valuable tool to realize NST for audios and convey a comprehensive understanding of neural styles in audio-based NST. We will integrate existing techniques from our explorations on image style transfer that can be successfully applied to audio and combine them with current methods already implemented to audio to achieve the desired results. The final transfer should be obviously identifiable while preserving enough of the original melody and texture so that the source audio remains recognizable. We hope that our project will be able to provide some insights or inspiration for future research.

3 Proposed Methods

The methodology to generate a new audio from an input audio is similar to the Image Style Transfer method, where authors in [11] used CNN for transforming the content image into a generated image rendered with the Style image. In the following, we briefly explain the three stages of our method.

First, in the pre-processing stage, we use a Continuous Wavelet Transform (CWT), which is the inner product of the audio signal with the family of wavelets, to convert the 1D audio signal to a spectrogram. The spectrogram is a visual representation of the frequencies of a signal. The result of this transformation is a 2D spectrum for time-frequency analysis. Moreover, the advantage of using CWT is the size-adjustable window, which gives optimal time and frequency resolutions.

Next, is the Image Style Transfer (IST), where we use a wide shallow network due to the small amount of available data. IST consists of a layer with 4096 distinct filters which has 4096 feature maps, each of size M equal to the width times the height of the feature map. We store the response in a matrix C where the entries of this matrix C_{ij} represent the activation of the i^{th} filter at the j^{th} position. Furthermore, to correct the output, we calculate the content loss, which is the difference between the content audio and the generated audio's feature maps at every time step. We use the gram matrix G of the convolutional feature map to obtain a representation of the style audio. We also define the output audio by a matrix A. Then, the style loss is the distance between these two matrices. The last step in this phase is calculating the total loss, which is a linear combination of the content loss and the style loss. To get a better image synthesis, we use L-BFGS [15] as our numerical optimization strategy. Then we resize the style and content images to extract image information on comparable scales.

In the last stage, pro-processing, we apply the Griffin-Lim algorithm [16] which is based on Inverse Continuous Wavelet Transform (ICWT), to convert the audio back to the time domain from the spectrogram.

4 Dataset

Our team will evaluate our model on the dataset named AudioSet [17], which comprises an extended ontology of 632 audio event classes and a collection of over two million manual-tagged sound clips from YouTube videos. Each audio clip is a stereo signal with a 10-second duration and 44100 Hz sampling frequency.

We determine to test on two types of data for audio style transfers:

- Music Genre: In terms of music genre, AudioSet has various representative styles or classifications of music, such as Hip hop music, Rock music, Rhythms and Blues. Each genre has thousands of human-labelled audio data. We plan to take several types of music as target styles, utilize our model to learn their most significant and salient features, input the other types of music as content and convert them to the target style of music.
- Music Mood: This kind of music category denotes the excessive emotional effect of music, regardless of genre or instrumentation. Our team will attempt to apply NST between different music moods, such as transferring a piece of music that evokes a happy feeling to a melody that conveys sadness. However, it should be noticed that the judgments of tags or annotation for such datasets are intrinsically subjective and likely to vary substantially across different musical cultures.

5 Project Execution Plan

The main accomplishments of this project are:

1. Exploring the state-of-art methods for audio style transfer to understand the network structures that provide a high quality synthesis.
2. Implementing a wide-shallow-random network for audio style transfer.
3. Utilizing some other networks or key findings of the state-of-art methods mentioned in the Background part.
4. Testing and comparing those networks by using AudioSet dataset.

6 Potential Impact

There are many possible real-world applications of style transfer in audio. For example, professional musicians often re-write or cover different versions of songs, often as a new interpretation of another musician's song. In many cases, the author of the song and the cover singer have radically different styles. Manually transferring musical styles would require recomposing the music to make the new-styled one pleasing enough to the audience. This process would be very arduous and time-consuming. However, our design could potentially speed up this procedure or even automate it completely. Moreover, those who are not experts in musical fields can also quickly obtain a song containing the same content but in the desired style different from the original piece.

From an academic perspective, neural style transfer for audios is still in the developmental stage. We hope to obtain interesting and valuable results that will provide understanding and inspiration for future related research.

References

[1] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[2] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019.

[3] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[4] Eric Grinstein, Ngoc QK Duong, Alexey Ozerov, and Patrick Pérez. Audio style transfer. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590. IEEE, 2018.

[5] Iman Malik and Carl Henrik Ek. Neural translation of musical style. *arXiv preprint arXiv:1708.03535*, 2017.

[6] Gino Brunner, Andres Konrad, Yuyi Wang, and Roger Wattenhofer. Midi-vae: Modeling dynamics and instrumentation of music with applications to style transfer. *arXiv preprint arXiv:1809.07600*, 2018.

[7] Ondřej Cífka, Umut Şimşekli, and Gaël Richard. Groove2groove: one-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2638–2650, 2020.

[8] Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. A universal music translation network. *arXiv preprint arXiv:1805.07848*, 2018.

[9] Marco Pasini. Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms. *arXiv preprint arXiv:1910.03713*, 2019.

[10] Ivan Ustyuzhaninov, Wieland Brendel, Leon A Gatys, and Matthias Bethge. Texture synthesis using shallow convolutional networks with random filters. *arXiv preprint arXiv:1606.00021*, 2016.

[11] Jiyou Chen, Gaobo Yang, Huihuang Zhao, and Manimaran Ramasamy. Audio style transfer using shallow convolutional networks and random filters. *Multimedia Tools and Applications*, 79(21):15043–15057, 2020.

[12] Xiaoyan Zhang, Xiaole Zhang, and Zhijiao Xiao. Deep photographic style transfer guided by semantic correspondence. *Multimedia Tools and Applications*, 78:34649 – 34672, 2019.

[13] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NIPS*, 2017.

[14] Dhruv Ramani, Samarjit Karmakar, Anirban Panda, Asad Ahmed, and Pratham Tangri. Autoencoder based architecture for fast & real time audio style transfer. *ArXiv*, abs/1812.07159, 2018.

[15] Lu P et al Zhu C, Byrd RH. Fortran subroutines for large-scale bound constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550 – 560, 1997.

[16] Lim J Griffin D. Signal estimation from modified short-time fourier transform[j]. *IEEE Trans Acoust Speech Signal Process*, 32(2):236–243, 1984.

[17] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.