NEURAL STYLE TRANSFER OF ENVIRONMENTAL AUDIO SPECTROGRAMS

by

Dejan Milacic
BCogSc, Carleton University, 2014
MA, McGill University, 2018

A Major Research Paper
presented to Ryerson University
in partial fulfillment of the requirements for the degree of

Master of Science
in the program of
Data Science and Analytics

Toronto, Ontario, Canada, 2020

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PAPER (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

NEURAL STYLE TRANSFER OF ENVIRONMENTAL AUDIO SPECTROGRAMS

Dejan Milacic

Master of Science 2020

Data Science and Analytics

Ryerson University

ABSTRACT

Neural Style Transfer is a technique which uses a Convolutional Neural Network to extract features from two input images and generates an output image which has the semantic content of one of the inputs and the “style” of the other. This project applies Neural Style Transfer to visual representations of audio called spectrograms to generate new audio signals. Audio inputs to the style transfer algorithm are sampled from the Dataset for Environmental Sound Classification (ESC-50). Generated audio is compared on the basis of input spectrogram type (STFT vs. CQT) and pooling type (max vs. average). Comparison is done using Mean Opinion Scores (MOS) calculated from ratings of perceptual quality given by human subjects. The study finds that STFT spectrogram inputs achieve high MOS when subjects are given a description of the style audio. The audio generated using CQT spectrogram inputs raises concerns about using visual domain techniques to generate audio.

Key words:

neural style transfer, convolutional neural network, audio, spectrogram, environmental sound

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Sri Krishnan for his guidance and helpful comments on this project. Thanks also to Garima Sharma for her valuable input in the early stages, and the other members of the Signal Analysis Research (SAR) group at Ryerson for their participation in the surveys. An honourable mention goes to all of my friends and family members who also contributed by completing the surveys. Lastly, I thank my parents for their unwavering support.

Contents

1 Introduction 1

2 Background 1

3 Literature Review 3

4 Methodology 4

4.1 The Neural Style Transfer Algorithm 4

4.2 Network and Hyperparameters 5

5 Results and Discussion 6

5.1 The Experiment 6

5.2 Survey 1: Style Reference 7

5.3 Survey 2: Content Reference 8

6 Conclusions and Future Work 9

List of Tables

1 Descriptions of style and content audio used in the experiment 7

2 Mean Opinion Scores calculated from Survey 1 8

3 Mean Opinion Scores calculated from Survey 2 8

List of Figures

1 An example of Neural Style Transfer applied to images 1
2 Two spectrogram representations of the same audio clip 2
3 A comparison of max pooling and average pooling 2
4 Overview of the network layers used in this study 6
5 An example of Neural Style Transfer applied to audio spectrograms 7

1 Introduction

Neural Style Transfer is a technique which uses a Convolutional Neural Network to extract features from two input images and generates an output image which has the semantic content of one of the inputs and the "style" (colour, texture) of the other (Fig. 1). This project extends the Neural Style Transfer technique to audio data using spectrogram representations of audio in place of images. The rest of the paper is structured as follows. Section 2 provides background knowledge necessary to understand the study. Section 3 gives a summary of the relevant literature. Section 4 explains the Neural Style Transfer algorithm and the neural network used in the study. Section 5 documents the experiment and discusses the results. Section 6 concludes.

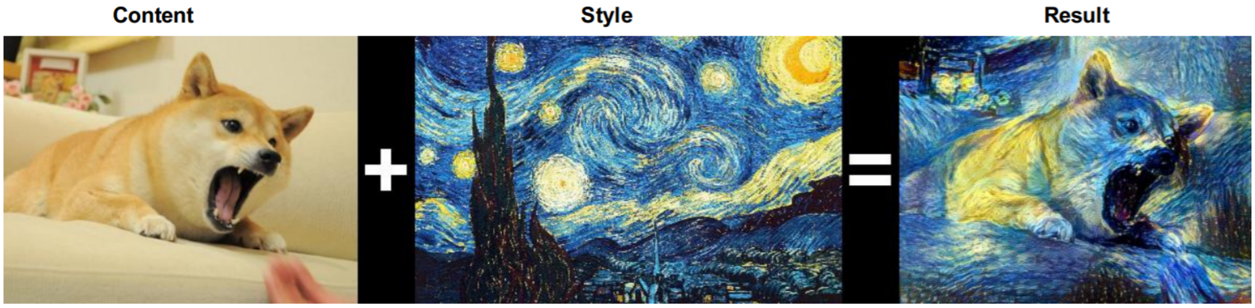

Figure 1: An example of Neural Style Transfer applied to images [8].

2 Background

An audio spectrogram is a visual representation of the spectrum of frequencies in an audio signal as it varies over time. A spectrogram can be calculated from a raw time-domain signal using a mathematical transform which decomposes the signal into its constituent frequencies. Two such transforms are the Short-Time Fourier Transform (STFT) and the Constant-Q Transform (CQT). These transforms differ in their frequency scaling: STFT uses linear frequency scaling, whereas CQT uses logarithmic frequency scaling (Fig. 2). This study compares the performance of the resulting spectrograms as inputs to an audio style transfer algorithm.

A Convolutional Neural Network (CNN) is a class of neural networks which was designed to analyze visual imagery. The CNN model introduces two types of network layers: convolutional layers and pooling layers. In a convolutional layer, several filters pass over the input image and each filter extracts a feature map from the input. In a pooling layer, a filter passes over each feature map and reduces its size (Fig. 3).[1] The most used type of pooling is max pooling, which outputs the

[1] The pooling visualizations are from Deep Learning lecture notes by Kanchana Padmanabhan. The sample feature map is from "An Intuitive Explanation of Convolutional Neural Networks" by Ujjwal Karn at `https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/`.

maximum value of each region the filter passes over. Another type of pooling is average pooling, which outputs the average value of each region the filter passes over. This study compares the effects of max pooling and average pooling on the perceptual quality of audio generated by the Neural Style Transfer algorithm.

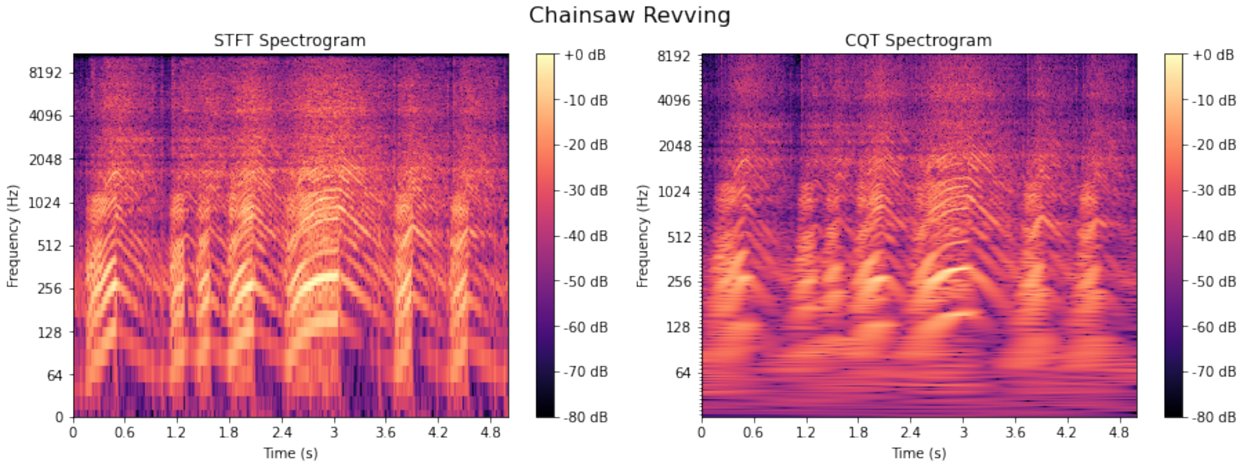

Figure 2: Two spectrogram representations of the same audio clip. The CQT spectrogram has better frequency resolution at lower frequencies.

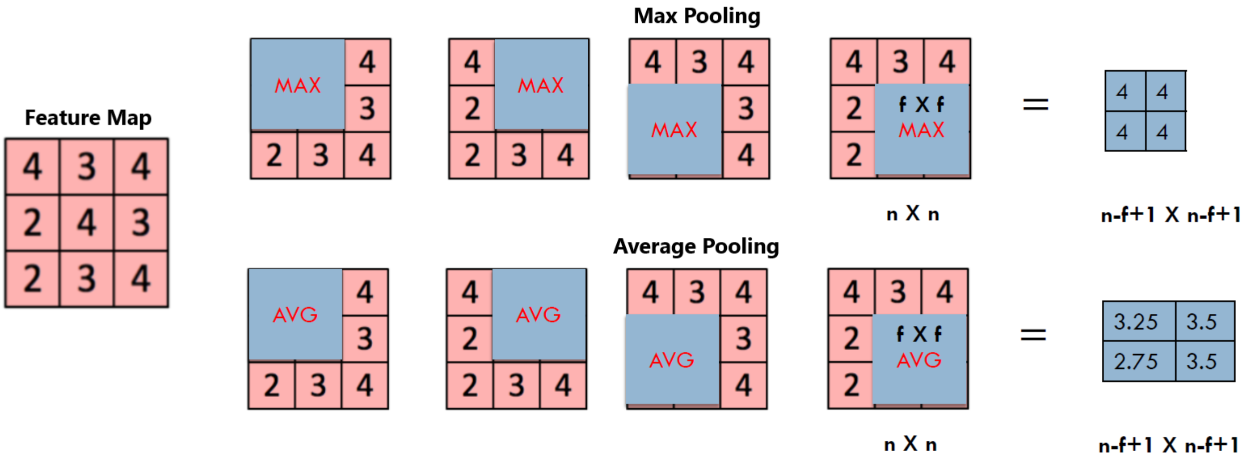

Figure 3: A comparison of max pooling and average pooling performed on the same 2D feature map with a 2D filter.

The performance of the Neural Style Transfer algorithm is measured by the perceptual quality of the generated output. Since perception is a subjective experience, different individuals may assign different scores to the same stimulus when given a predefined perceptual quality scale. Mean Opinion Score (MOS) is a measure which can be used to represent the overall perceptual quality of a stimulus. It is the arithmetic mean of all individual ratings of perceptual quality given by human subjects on a predefined scale. Given ratings R by N subjects:

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \tag{1}$$

This study will survey several human subjects and use MOS to compare the effects of different hyperparameter settings on the perceptual quality of audio generated by Neural Style Transfer.

Audio data can be categorized as music, speech, or environmental sound. Environmental sounds are the audio stimuli other than music and speech which form our daily surroundings, such as sounds of nature or technology. The Dataset for Environmental Sound Classification (ESC-50) [6] is a set of 2,000 environmental sound recordings. Each recording is 5 seconds long and belongs to one of 50 classes (40 per class). Examples of classes include toilets flushing, babies crying and birds chirping. This study uses environmental sound recordings from ESC-50 as inputs to the Neural Style Transfer algorithm.

3 Literature Review

This project extends the Neural Style Transfer technique developed for images to audio data using audio spectrograms. Many recent studies have focused on applying Neural Style Transfer to audio spectrograms to generate new audio signals. Some studies have focused on music, while others have looked at audio data more generally. In past studies, the relative perceptual quality of the audio generated by different models has generally been determined by the authors. Few researchers have employed measures such as MOS to compare the performance of different models. In this section, I give an overview of some previous studies, making note of the models and types of audio data used.

The Neural Style Transfer technique for images was first developed by Gatys et al. [2][3]. They use a deep CNN which is pre-trained for image classification. They note that using average pooling "yields slightly more appealing results" than max pooling in the visual domain.

Ulyanov and Lebedev [7] made a preliminary attempt at audio style transfer for music. They use a shallow 1-layer network with random weights and 1D convolutions only along the time dimension of the spectrograms. Grinstein et al. [4] compare this shallow random network to deep trained networks for audio data more generally. They conclude that the shallow random network shows more promising results than the deeper models. Chen et al. [1] look at music, speech, and environmental audio data. They use the random shallow network and compare the performance of STFT spectrogram inputs to spectrograms generated by Continuous Wavelet Transform (CWT), which can achieve better time and frequency resolution. They survey 30 human subjects and find that audio generated using CWT spectrograms achieves a higher MOS.

Huzaifah and Wyse [5] use a 3-layer architecture and compare a network with random weights to a network which was trained on ESC-50. They conclude that using randomly initialized weights results in the most stable style transfer for all types of audio, even audio taken from the trained network's training set. They explore the use of CQT spectrogram representations in place of STFT spectrograms, but they do not report a direct comparison of their use in style transfer.

This study makes use of the 3-layer architecture of [5] to compare the effects of STFT vs CQT spectrograms and max pooling vs average pooling on the perceptual quality of audio generated by Neural Style Transfer. Previous studies using 1-layer networks were not able to explore the effect of pooling in the audio domain. This study joins [1] in using MOS to compare the perceptual quality achieved by different models and is the first to focus on style transfer for environmental audio in particular.

4 Methodology

The aim of this study is to apply the Neural Style Transfer algorithm to audio spectrograms and determine how different hyperparameter settings affect the perceptual quality of the generated audio. In this section, I give an outline of the Neural Style Transfer algorithm and describe how it is modified and evaluated in the study.

4.1 The Neural Style Transfer Algorithm

Given a content image c and a style image s, the Neural Style Transfer algorithm aims to generate an image x which matches the "content" features of c and the "style" features of s. These features are extracted from the reference images using a CNN and are matched by those of the generated image using a loss function.

Each layer of the network defines a set of filters which extract feature maps from the input. An input is encoded in each layer of the CNN by the feature maps extracted. A layer l with C_l distinct filters has C_l feature maps of size U_l, where U_l is the product of the height, width and batch size of the feature map. The filter responses in a layer l can be stored in a matrix $F^l \in \mathcal{R}^{C_l \times U_l}$ where F^l_{ij} is the activation of the ith filter at position j in layer l.

Style is represented in terms of feature correlations given by a Gram matrix $G^l \in \mathcal{R}^{C_l \times C_l}$ where G^l_{ij} is the inner product of the vectorized feature maps i and j in layer l, normalized by U_l:

$$G^l_{ij} = \frac{1}{U_l} \sum_k F^l_{ik} F^l_{jk} \tag{2}$$

Given a set of style layers $\mathscr{S}$, the contribution of a layer l to the style loss function is defined as the squared Frobenius norm of the difference between the style representations of the reference style image and the generated image.

$$L_{style}(s,x) = \sum_{l \in \mathscr{S}} \frac{1}{C_l} \| G^l(s) - G^l(x) \|_F^2 \tag{3}$$

Content is represented directly by the feature maps. The corresponding content loss function is calculated from the content representations of the reference content image and the generated image in content layers $\mathscr{C}$.

$$L_{content}(c,x) = \sum_{l \in \mathscr{C}} \frac{1}{C_l} \| F^l(c) - F^l(x) \|_F^2 \tag{4}$$

The total loss function used to update the generated image is a weighted sum of the style and content loss functions at the corresponding layers $\mathscr{S}$ and $\mathscr{C}$ of the CNN. The generated image x is initialized using either random noise or a clone of the content image c. The loss function is then minimized by updating x using gradient descent with backpropagation through a CNN with fixed weights, resulting in an image with the style features of s and the content features of c. The relative influence of the content and style representations can be adjusted using the weighting hyperparameters α and β.

$$\begin{aligned} x &= \underset{x}{\operatorname{argmin}}\, L_{total}(c,s,x) \\ &= \underset{x}{\operatorname{argmin}}\, \alpha L_{content} + \beta L_{style} \end{aligned} \tag{5}$$

4.2 Network and Hyperparameters

The base network architecture used in this study is composed of three stacks, each a sequence of a convolutional layer, a batch normalization layer, ReLU and max pooling (Fig. 4). The network is run with random weights and the generated image initialized as a clone of the content image, since this combination has already been shown to give stable results [4][5]. Since the weights are not being trained, the fully-connected layer is not used. Layers used for the content and style representations are $\mathscr{C} = relu3$ and $\mathscr{S} = relu1, relu2$.[2] Loss function weights are set to α = 1, β = 1e8.

The factors that are varied in this study are the input spectrogram and the pooling layers. The algorithm is first run with the base network using STFT and CQT input spectrograms. The max pooling layers are then replaced by average pooling layers and the experiment is repeated. The experimental design is factorial, with all four possible combinations of factor levels being tested.

The performance of the Neural Style Transfer algorithm is measured by the perceptual quality of the generated audio. Since perceptual quality is subjective and may vary between individuals,

[2] Following [5], $relu1$ refers to the ReLU layer in the 1st stack, $conv3$ refers to the convolutional layer in the 3rd stack, etc.

the response variable is a Mean Opinion Score (MOS) calculated from ratings of perceptual quality given by several human subjects.

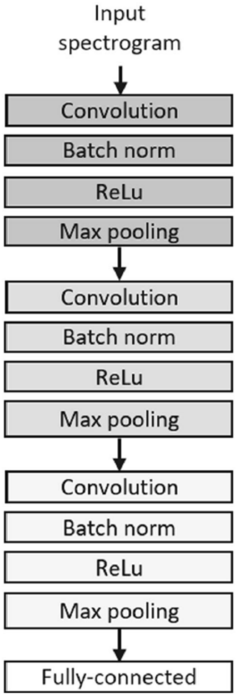

Figure 4: Overview of the network layers used in this study [5].

5 Results and Discussion

This section documents the experiment that was performed and discusses the results of two surveys of perceptual quality.

5.1 The Experiment

Audio clips were selected from ESC-50 in pairs of style and content audio (Table 1). All audio clips were sampled at 22,050 Hz. STFT spectrograms were generated using a window length of 1024 samples and hop length of 128, resulting in a size of 513 frequency bins $\times$ 862 time bins. CQT spectrograms were generated using a hop length of 128, with 512 frequency bins and 64 bins per octave. These parameter settings yield spectrograms of similar size and frequency coverage for a fair comparison. Magnitudes were normalized with a natural log function $\log(1 + |magnitude|)$ before running the algorithm.

Examples of audio style transfer were generated for each of the 5 style-content pairs in Table

1 using each of the 4 hyperparameter combinations for a total of 20 generated audio clips.[3] The audio generated by the Neural Style Transfer algorithm matches the timbre of the style audio and adopts the rhythmic structure of the content audio (Fig.5). This behaviour is consistent with the findings of Huzaifah and Wyse [5].

Style	**Content**
Snoring	Hand saw
Chicken	Car horn
Toilet flush	Chainsaw
Frog	Keyboard typing
Water dripping	Applause

Table 1: Descriptions of style and content audio used in the experiment.

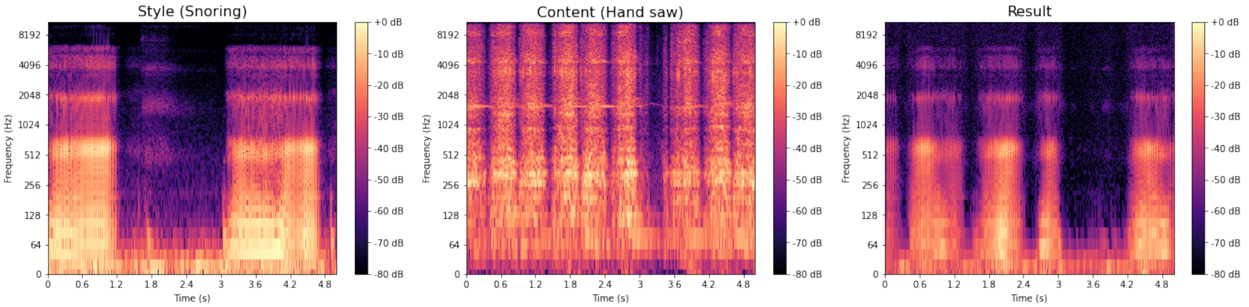

Figure 5: An example of Neural Style Transfer applied to STFT spectrograms using max pooling. The contributions of the style and content audio can be seen in the result.

Two surveys were conducted to determine how the different hyperparameter settings affect the perceptual quality of the generated audio.

5.2 Survey 1: Style Reference

In the first survey, each generated audio clip was paired with a text description of the style audio used to generate it as in Table 1. Subjects were asked to rate the perceptual quality of the generated audio on a scale of 1 to 5 (1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent). The text description was included in order to give subjects some frame of reference to base their ratings on. They were told to base ratings on the presence of noise or distortions, and the similarity of the audio to the text description. The ratings were used to calculate MOS for each combination of hyperparameters (Table 2). Given 18 participants and 5 ratings per model, each score is calculated from 90 ratings.

[3] All style, content and generated audio clips discussed here can be heard on the companion website: `https://sites.google.com/view/mrp-audio-neural-style/home`.

Input	Pooling	MOS
STFT	max	4.26
	average	4.19
CQT	max	2.79
	average	2.84

Table 2: Mean Opinion Scores calculated from Survey 1.

The use of max vs average pooling in the network did not appear to affect the perceptual quality of the generated audio, with MOS remaining very similar for both settings. An audible difference would probably require a deeper network with more pooling layers. Regarding the input spectrograms, audio generated with STFT inputs was given a higher MOS than audio generated with CQT. This result could be partially due to the conversion of the spectrograms back to time-domain audio signals. In some cases, the algorithm used to convert CQT spectrograms back to audio signals introduces noise not present in the original audio clips. Most notably, the toilet flush audio gains a high-pitched noise when converted back from a CQT spectrogram. However, the conversion alone does not account for the white noise present in most of the audio clips generated using CQT inputs. This white noise is present in the generated audio, but absent from the style and content audio when converted back from CQT spectrograms. This behaviour leads to the conclusion that the white noise arises from some combination of the style transfer itself and the CQT inversion.

5.3 Survey 2: Content Reference

In the second survey, each generated audio clip was paired with a text description of the content audio used to generate it as in Table 1. Subjects were asked to rate the perceptual quality of the generated audio on a scale of 1 to 5 (1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent). The text description was included in order to give subjects some frame of reference to base their ratings on. They were told to base ratings on the presence of noise or distortions, and the similarity of the audio to the text description. The ratings were used to calculate MOS for each combination of hyperparameters (Table 3). Given 18 participants and 5 ratings per model, each score is calculated from 90 ratings.

Input	Pooling	MOS
STFT	max	1.62
	average	1.63
CQT	max	2.10
	average	2.13

Table 3: Mean Opinion Scores calculated from Survey 2.

The scores here are much lower than those from the first survey. This result is expected, given that the generated audio tends to inherit only the rhythmic structure of the content audio, and in general does not sound like the content description. As in the first survey, the use of max vs average pooling in the network did not appear to affect the perceptual quality of the generated audio, with MOS remaining very similar for both settings. Unlike the first survey, audio generated using CQT spectrogram inputs achieved a slightly higher MOS than audio generated using STFT. The audio clips generated using CQT spectrogram inputs tend to have a greater contribution from the content audio clip than audio generated using STFT. This behaviour is notable with audio generated using the keyboard typing content audio, where the content sound can be heard alongside the style. The result is more like a blend of the content and style audio than a "style transfer." If this result is desired, it can be achieved using STFT spectrogram inputs by decreasing the style loss weight β.

6 Conclusions and Future Work

The goal of this project was to apply the Neural Style Transfer algorithm to environmental audio spectrograms and determine how different hyperparameter settings affect the perceptual quality of the generated audio. This was done by generating audio clips using different combinations of hyperparameters and conducting two surveys of perceptual quality of the generated audio. The results of each survey were used to calculate MOS for each combination of hyperparameters.

In survey 1, subjects rated the perceptual quality of the generated audio against a description of the style audio used to generate it. Audio generated using STFT spectrogram inputs achieved MOS of about 4.2, indicating good perceptual quality. Audio generated using CQT spectrogram inputs achieved MOS of about 2.8, indicating poor perceptual quality.

In survey 2, subjects rated the perceptual quality of the generated audio against a description of the content audio used to generate it. STFT and CQT spectrogram inputs achieved MOS of about 1.6 and 2.1 respectively. This result mirrors the relative contribution of the content audio to the overall sound of the generated audio.

Neither survey found that pooling type had any effect on perceptual quality, possibly requiring a deeper network to achieve an audible difference. The poor performance of CQT spectrogram inputs raises questions about the effects of spectrogram inversion algorithms on audio generated using this technique.

Future work may choose to experiment with deeper networks in order to establish the effect of pooling type on perceptual quality, though using very deep networks may render the process too time-consuming to justify the small output. Another possibility is to explore techniques designed specifically for audio data rather than adapting techniques from the visual domain. This option has the benefit of operating directly on raw time-domain audio signals rather than the back and forth conversion of spectrograms.

References

[1] J. Chen, G. Yang, H. Zhao, and M. Ramasamy. Audio style transfer using shallow convolutional networks and random filters. *Multimedia Tools and Applications*, pages 1–15, 2020.

[2] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[3] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[4] E. Grinstein, N. Q. Duong, A. Ozerov, and P. Pérez. Audio style transfer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590. IEEE, 2018.

[5] M. Huzaifah and L. Wyse. Applying visual domain style transfer and texture synthesis techniques to audio: insights and challenges. *Neural Computing and Applications*, 32(4):1051–1065, 2020.

[6] K. J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018, 2015.

[7] D. Ulyanov and V. Lebedev. Audio texture synthesis and style transfer. *https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/*, 2016.

[8] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.