ALTERATION OF LATENT REPRESENTATIONS IN AUDIO/SPEECH STYLE TRANSFER

by

Remizova Anna

A Thesis

Submitted in Partial Fulfillment of the

Academic Advisor:

Ph. D. Andrey Grabovoy

Bachelor of Science

Major: Applied mathematics and Computer Science

Moscow Institute of Physics and Technology

May 2024

Abstract

Motivation: Style transfer in audio is the field of computer science which is not yet thoroughly studied and, therefore, is of great interest. Moreover, it finds its applications in many practical tasks like dubbing movies, investigating voice abilities and music writing without any effort. In this paper we've been working on the task of voice/music style transfer. While there are many approaches to the task: Variational Auto Encoder(VAE), Gaussian Mixture Models(GMM), Linear Homodomain Transformations, it can be noticed that Image Style transfer is one of the most developed branches of Machine learning. That's why the image-like approach to the audio style transfer is altered in this paper.

Keywords: Machine Learning, Audio Style transfer, Speech Style transfer, Image Style Transfer, Deep Convolutional NN, Gram Matrix, Linear Transformation

Table of Contents

Chapter Page

1 Introduction 1

Chapter 1

Introduction

Audio style transfer is a technique to captivate style of one source of sound and content of another. This can be applied to music, for instance, a piano composition may sound in a different pitch or even different instrument while keeping the structure of the composition.

Other approaches:

1. Voice conversion based on DNN time-valiant linear transformation for

· Linear transform is enough because we work with homo-domain mapping

1) Cepstrum - result of inverse Fourier transform of the log of the estimated Signal spectrum

2) Signal spectrum of a time series x(t) - distribution of power into freq. components f comp. that signal

The main idea of the approach is the wave transformation via warp function(which is a linear transformation in cepstral space).

2. Learning Latent Representations for Style control and transfer in end toend Speech synthesis

. Main idea - learn latent representations of the source audio to control style transfer

· Variational Autoencoder - reveals relationship between latent z and observed x. True posterior $p_\theta(z|x)$ is impossible to find, therefore, Likelihood $p_\theta(x)$ is indifferent. So $p_\theta(z|x)$ can be approximated by $q_\phi(z|x)$ and variational lower bound for $\mathcal{L}(\theta, \phi; \mathbf{x})$:

$$\begin{aligned} \log p_\theta(\mathbf{x}) &= KL\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z} \mid \mathbf{x})\right] + \mathcal{L}(\theta, \phi; \mathbf{x}) \\ &\geq \mathcal{L}(\theta, \phi; \mathbf{x}) \\ &= E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - KL\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right] \end{aligned}$$

Loss $= KL\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right] - E_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{t})\right] + l_{\text{stop}}$

Both 1. and 2. methods work with feature vectors x_t, y_t, which are retrieved from data via specially pretrained for this task NN. But there is another way to work with music data - STFT(Short Time Fourier Transform), which converts discrete audio signal into frequency domain. Next method along with many others uses this representation.

3. Deep learning for Audio style transfer

· main idea - learning Content and Style separately using Gram Matrix for style retrieval and convolutional NN's for content

· preprocessing - STFT (Short time Fourier transform)

· Using different losses for style and content

So let $\vec{p}$ and $\vec{x}$ be the original image and the image that is generated and P^l and F^l their respective feature representation in layer l. We then define the squared-error loss between the two feature representations

$\mathcal{L}_{\text{content}}\left(\vec{p}, \vec{x}, l\right) = \frac{1}{2} \sum_{i,j} \left(F_{ij}^l - P_{ij}^l\right)^2$

Gram matrix $G^l \in \mathcal{R}^{N_l \times N_l}$, where G_{ij}^l is the inner product between the vectorised feature map i and j in layer l: G_{ij}^l = $\sum_k F_{ik}^l \, F_{jk}^l$

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l - A_{ij}^l\right)^2$$

$$\mathcal{L}_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l$$

All these approaches have their strong and weak sides, so this topic has a lot to offer for the further research. Therefore, I decided to start investigating Image-like approach, as one of the most developed field in the topic.

Datasets used in this research are collections of .wav files - recorded english utterances of different speakers or one speaker audio(for one-shot voice conversion) cut into short intervals.

- a public domain speech dataset consisting of 13,100 short audio clips of a single speaker reading passages
- a corpus of approximately 1000 hours of 16kHz read English speech
- english multi-speaker corpus