Deep learning for Audio Style Transfer

Zhixian Huang
Department of ECE

Shaotian Chen
Department of ECE

Bingjing Zhu
Department of CS

ABSTRACT

Style transfer is the technique of recomposing inputs using the style of other inputs, which has gained increasing popularity recently. The success in image style transfer has inspired people to use similar methods to do style transfer in audio domain.

In this project, we implement the architecture in [2]. We test it on several audio samples, and find the result of speech style transfer is the best, so we do further experiments on VCTK dataset to prove its performance in speech style transfer.

Figure 1 Image style transfer example

INTRODUCTION

Image style transfer can be achieved using deep CNN models. Gatys [1] proposes to generate images of specific artistic style from white noise using CNN.

In [1], content is extracted as feature maps outputted by some pre-trained VGG layers; Style is defined as correlation between the feature maps in different channels and is given by Gram Matrix, which is the inner product of feature maps and their transpose. The white noise is gradually optimized by back propagation of both content loss and style loss.

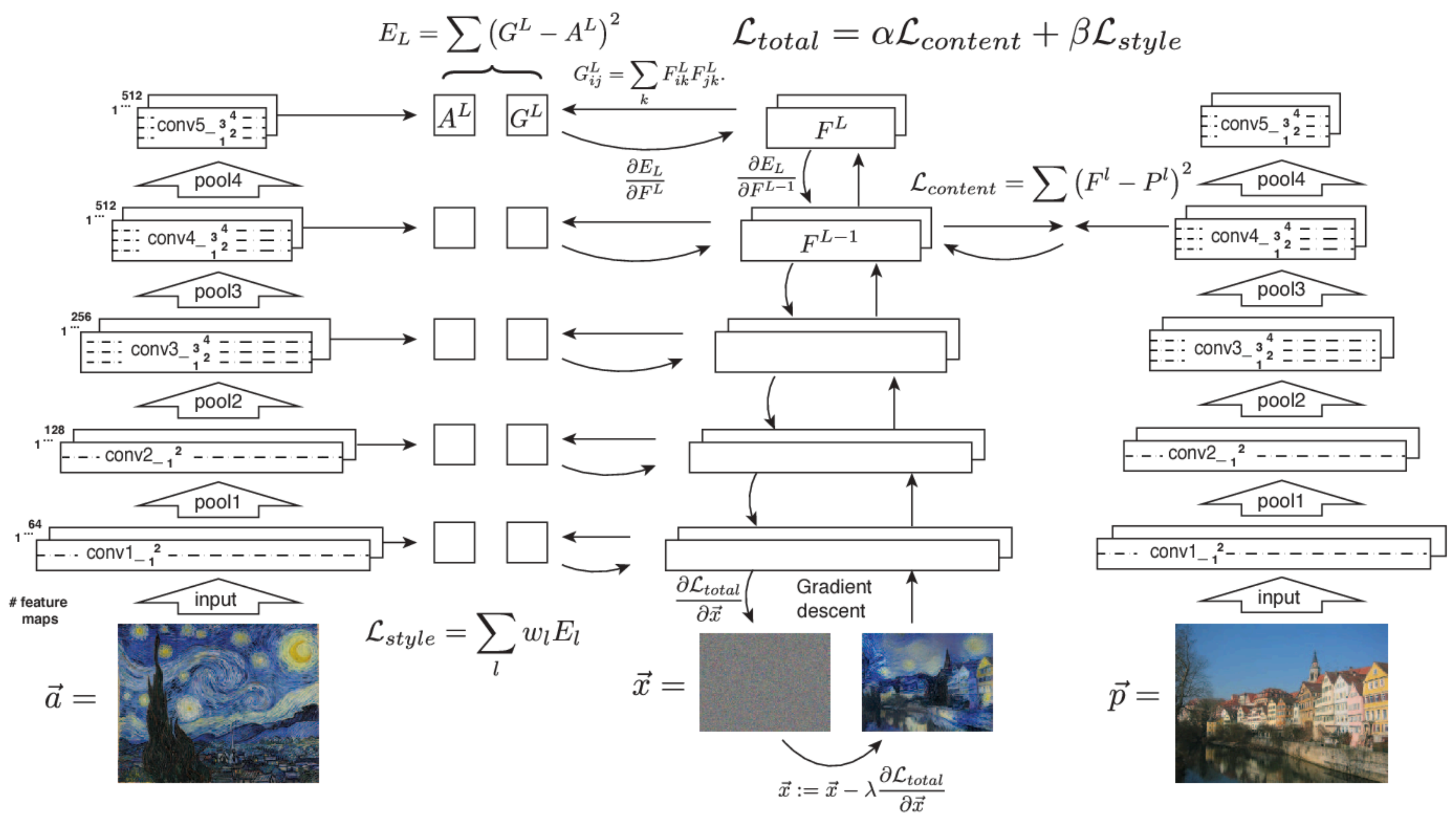
$E_L = \sum (G^L - A^L)^2$
$\mathcal{L}_{total} = \alpha\mathcal{L}_{content} + \beta\mathcal{L}_{style}$
$G^L_{ij} = \sum_k F^L_{ik}F^L_{jk}$
A^L
G^L
F^L
$\frac{\partial E_L}{\partial F^L}$
$\frac{\partial E_L}{\partial F^{L-1}}$
$\mathcal{L}_{content} = \sum (F^l - P^l)^2$
F^{L-1}
512
256
128
64
conv5_1 2 3 4
pool4
conv4_1 2 3 4
pool3
conv3_1 2 3 4
pool2
conv2_1 2
pool1
conv1_1 2
input
# feature maps
$\frac{\partial \mathcal{L}_{total}}{\partial \vec{x}}$
Gradient descent
$\mathcal{L}_{style} = \sum_l w_l E_l$
$\vec{a} =$
$\vec{x} =$
$\vec{p} =$
$\vec{x} := \vec{x} - \lambda \frac{\partial \mathcal{L}_{total}}{\partial \vec{x}}$

Figure 2 Architecture of image style transfer [1]

METHOD

In [2], they use same method as Gatys [1], taking magnitude spectrogram via STFT (Short Time Fourier Transform) as input instead of image. In this project, our method is based on [2] and we use several pairs of music to test the architecture. For quantitative evaluation, we design an experiment and use VCTK[10] dataset to prove the effectiveness.

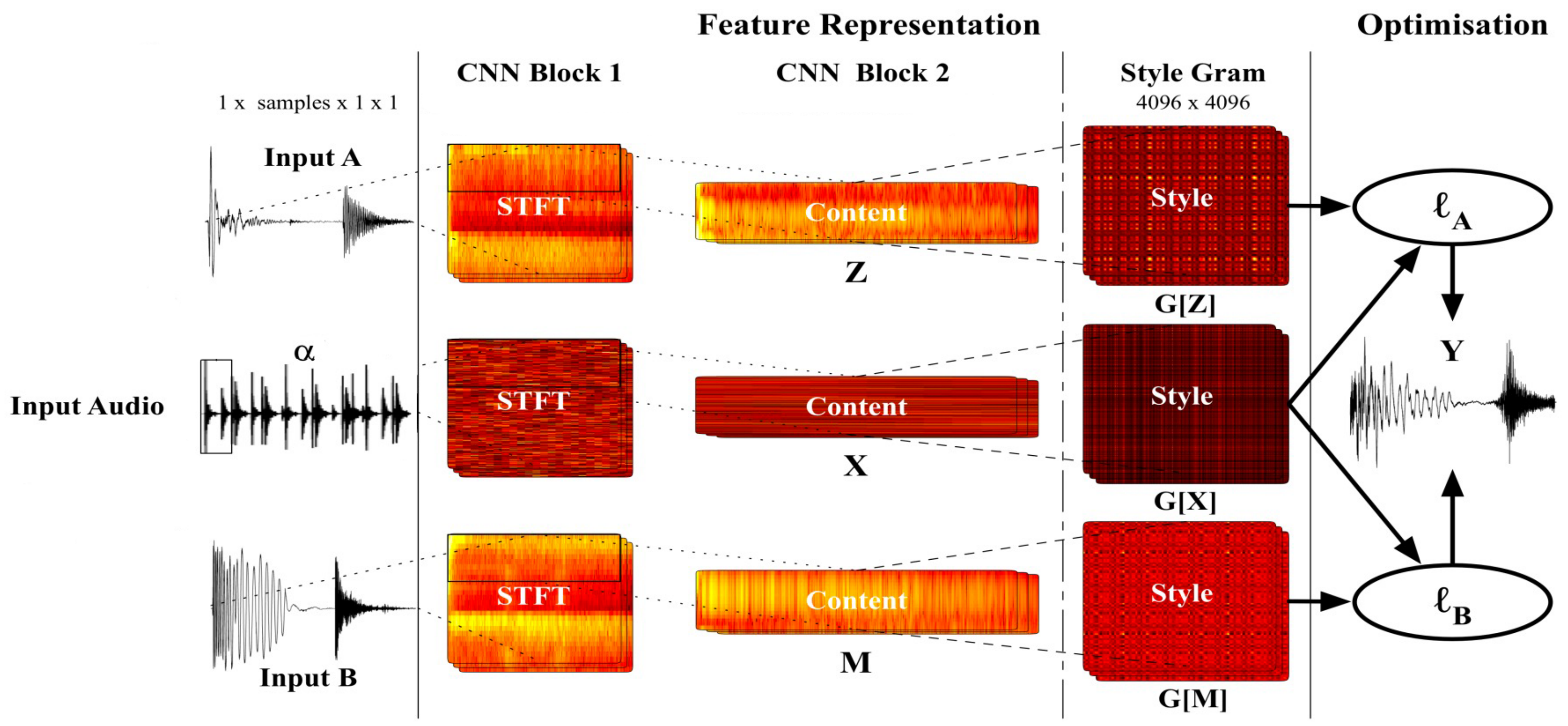

Figure 3 Architecture of this method

- Network Selection: Wide-Shallow-Random Network, 1-layer CNN with 4096 random filters.
- Pre-processing: Hanning window of n samples (2048) with a n/2 hop size to segment it into T frames.
- Feature Representation and Loss Function:
  - Content loss: MSE between output and input audio

$$L_{content}\ (\vec{p}, \vec{x}, l) = \frac{1}{2}\sum_{i,j}(F^l_{ij} - P^l_{ij})$$

  - Feature Correlation (Gram Matrix):

$$G_{ij} = \sum_{k} F_{ik}F_{jk}$$

  - Style Loss: MSE between the gram matrices of style audio and the generated audio.

$$L_{(\vec{a},\vec{x})style} = \frac{\sum_{i,j}(X_{ij} - A_{ij})^2}{4N^2M^2}$$

  - Total Loss = content loss + style loss
- Phase Reconstruction: Griffin-Lim algorithm

EXPERIMENT

- Dataset:
  - Dataset 1: 6 pairs of audio samples: single instrument music, symphonic music and speech
  - Dataset 2: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit: 109 native speakers of English
- Result
  - Experiment 1: The result of rap music (content) and speech (style) sounds like transfer voice in rap music onto voice in speech music.

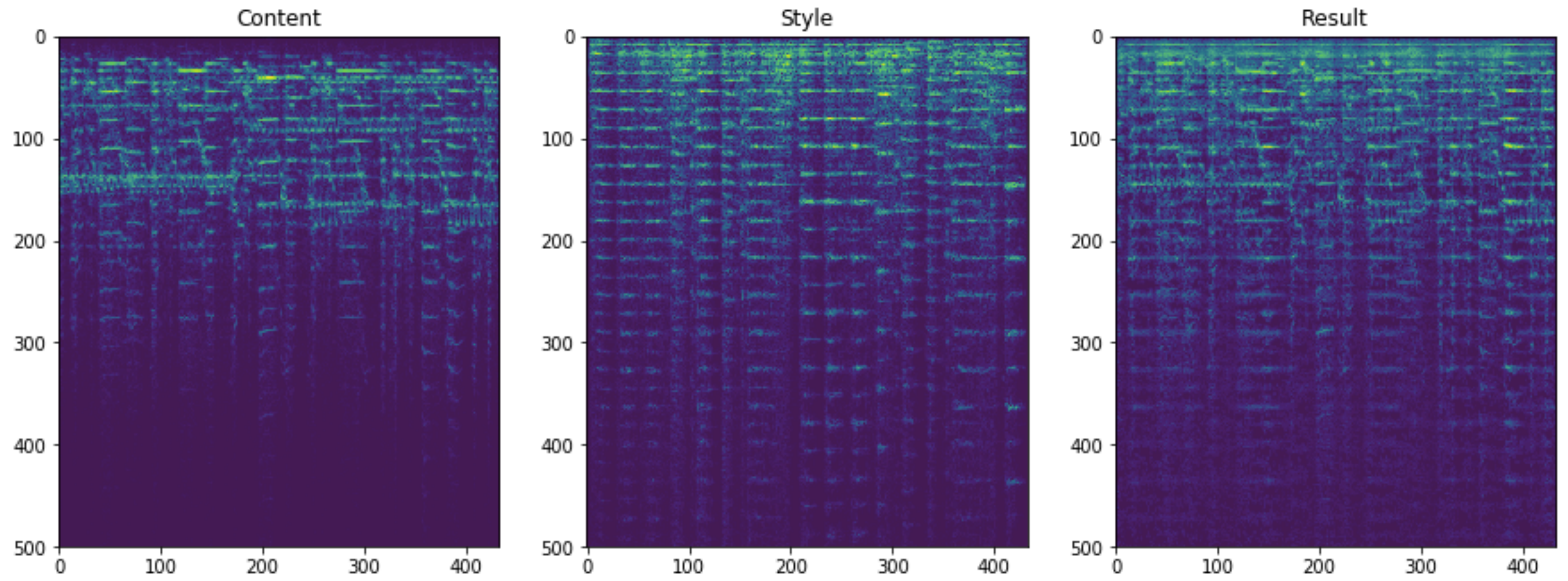

Figure 4 Content audio: rap; Style audio: speech

  - Experiment 2:

Speaker Number	Train/Test	Accuracy
4	80/80	0.925
10	80/80	0.485

DISCUSSION

The model is inspired by texture network, which is designed for image style transfer. We regard spectrogram as image in order to use the network in image style transfer. Instead of using pre-trained CNN (like VGG, SoundNet), we use Wide-Shallow-Random Network to extract "content" and "style". Surprisingly, the random CNN shows a good performance in speech recognition, which means it can capture features of speaker identity.

REFERENCE

[1] Gatys et al. "Image style transfer using convolutional neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
[2] Tomczak et al. "Audio style transfer with rhythmic constraints."