Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 1, No. 1, 2022

Advancements in Voice Conversion: Spectrogram-Based Speech Style Transfer Using Convolutional Neural Networks

Emily Johnson

Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA

Abstract: Voice Conversion (VC) transforms the phonetic style of a source speaker to a target speaker while preserving semantic content. This technology has applications in communication, healthcare, entertainment, and security. Traditional methods using neural networks have enhanced speech quality, but current research aims to reduce training data requirements. Inspired by image style transfer, this paper uses convolutional neural networks (CNNs) to extract and stylize spectrogram features from speech signals. The proposed model achieves high-quality speech style transfer, demonstrating CNNs' effectiveness in voice conversion with reduced data dependency.

Keywords: Image style transfer, Speech style transfer, Convolutional neural network, 2D spectrogram.

1. Introduction

Voice Conversion (VC) [1] refers to the conversion of the phonetic style features of the Source Speaker to those of the Target Speaker, while keeping the semantic information of the Source Speaker unchanged. Actually, Speech style transfer can be applied to communication, medical care, entertainment and other fields: In the text-to-speech (TTS) model [2], the synthesized speech might sound more like the voice delivered by a real person if it is processed by speech style transfer at the meantime; For the sake of confidentiality and security, speech style transfer technology can be adopted to change the style characteristics of speaker's voice [3]; In the medical field, the speech signal phonated by patients with damaged larynx can be repaired with the help of voice style transfer technology [4]; In the film dubbing field, especially, when another language is used in film dubbing, speech style transfer technology can make the voice style of the dubbing actor the same as that of the film actor, and the ideal dubbing effect is achieved finally[5].

It can be seen from the aforementioned examples that speech style transfer technology, as a subject with strong interdisciplinarity, has extremely important research role and value, and attracts numerous researchers to explore. In 1988, Abe's group [6] firstly proposed speech style transfer based on Vector Quantization (VQ) and codebook mapping. In 1992, Savic[7] et al. improved the codebooks mapping into a neural network on the basis of Abe's research, which greatly enhanced the quality of converted speech. This is the first time to apply the artificial neural network model in the study of speech style transfer and has made a certain breakthrough. Subsequently, the research of speech style transfer based on neural network becomes the mainstream research direction. In 1995, Narendranath [8] et al. took the formant parameters of the original and stylized speech signals as the inputs and outputs of the artificial neural network (ANN) respectively and trained them by dint of BP method. Ultimately, stylized speech signal was synthesized with the average gene frequency and stylized formant parameters, which were converted by the formant parameters of the tested speech. In 2002, scholars Watanabe [9] et al. come up with an algorithm for LPC spectral envelope transformation using RBF neural network. In 2007, Guldo [10] et al. put forward a speech style conversion algorithm based on wavelet transform and artificial neural network model. In 2014, Nirmal [11] et al. took advantage of the generalized regression neural network (GRNN) to transform the style characteristic parameters for speech signals. In 2015, Ghorbandoost [12] et al. combined the personality characteristics of two kinds of speeches into a new speech personality characteristic, and realized the speech style transformation through the classical gaussian mixture model and artificial neural network model.

The above references show that the performance and stability of generated speech have immensely improved based on neural network in the research of speech style transfer. Therefore, how to use less training data or even no other data to study the speech style transfer model is the research innovation of this article. Inspired by the study of image style transfer, the paper utilizes the convolutional neural network to extract the features of the spectrograms for speech signals, so as to generate the stylized spectrogram and acquire corresponding stylized speech.

2. The Spectrogram of Speech Signal

2.1. The 2D Spectrogram of Speech Signal

Before the experiments of speech style transfer, the feature extraction in the 2D spectrogram of speech signal is usually required. That is to say, the discernable information contained in speech signals can be achieved through the extracted feature information. Moreover, speech signals are generated through the vocal tract, the shape of the vocal tract determines what kind of speech is phonated, to some extent. The shape of the vocal tract can be shown in the envelope of speech short time power spectrum, and the characteristic information of speech signals - 2D spectrogram can describe the envelope exactly.

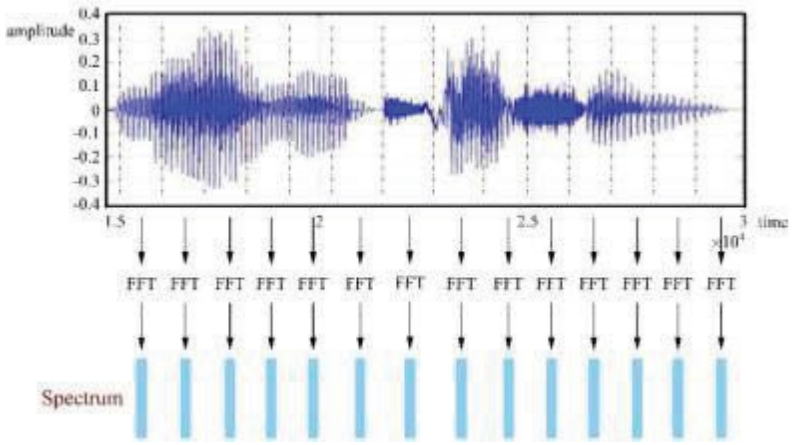

Figure 1. The speech frames

First, the speech signal displayed in figure 1 is divided into 14 frames. Then, the spectrums corresponding to those 14 frames of speech can be obtained by calculating Short Fast Fourier Transformation. The spectrum represents the relationship between frequency and energy amplitude, and the specific information is plotted on the left graph in figure 2.

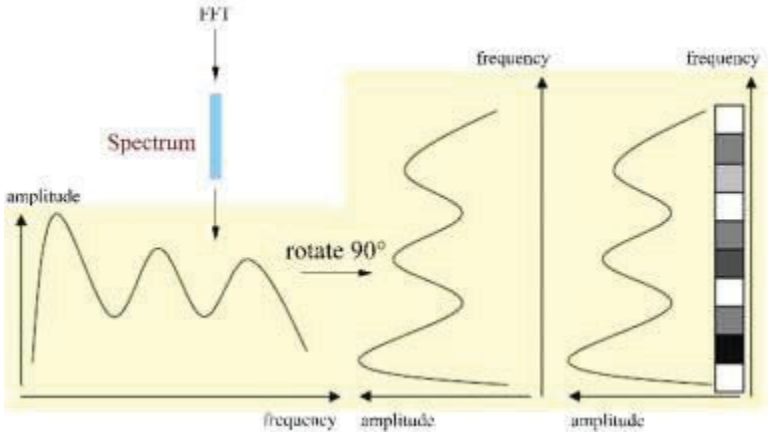

Figure 2. The spectrum

Next, Rotating the spectrum curve by 90 degrees to get the middle graph in figure 2. Furthermore, the amplitudes in the middle graph are mapped to a range of gray level, with gray level 255 and gray level 0 represented by the black area and the white area, respectively. In other words, the larger the amplitude value, the darker the corresponding area is. Thus, the right-most graph is acquired in figure 2.

The aim above is to add the time dimension, so that the spectrum of a speech, rather than a frame, can be manifested integrally. Finally, we obtain a spectrum diagram over time, which is the 2D spectrogram describing the speech signal, as shown in figure 3.

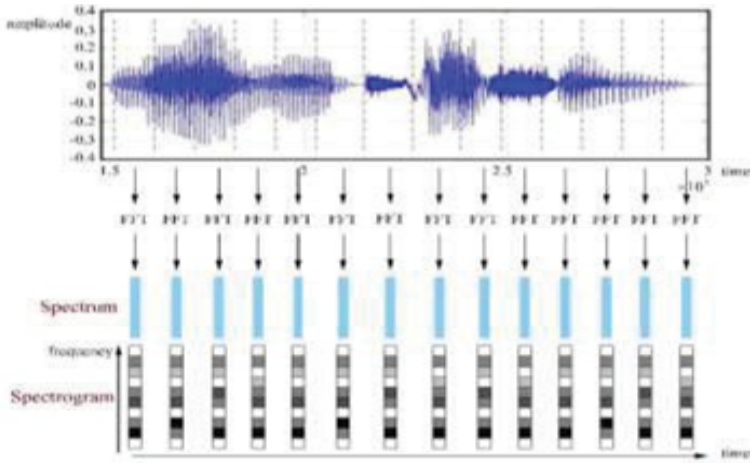

Figure 3. The 2D spectrogram

3. Speech Style Transfer Model Based on Neural Network

This paper proposes an innovative solution - the speech style transfer model based on convolutional neural network.

3.1. The Speech Style Transfer Model Based on convolutional Neural Network

The principle of image style transfer model can be demonstrated through the following flow figure 4, roughly.

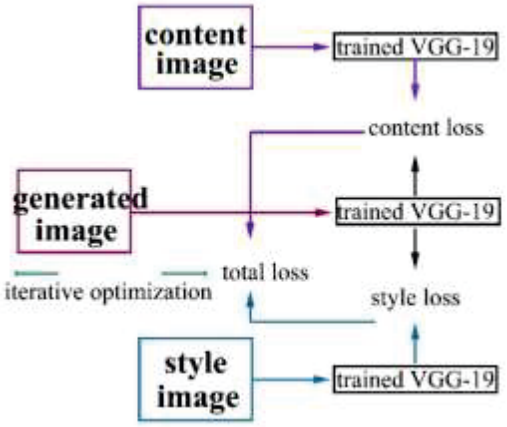

Figure 4. The image style transfer model

Since the convolutional neural network is good at dealing with picture type data, processing speech signal into the picture type data is key point and innovation point of speech style transfer. According to the introduction of 2D spectrogram in section 2.1, it can be known that spectrogram can be regarded as 2-dimensional picture type data, to some extent. Therefore, when the research object is speech signal, content image, style image and generated image in figure 4 are replaced by the 2D spectrograms of content speech, style speech and generated speech, respectively. Theoretically, we can achieve the speech style transfer model. The speech style transfer model is shown in figure 5.

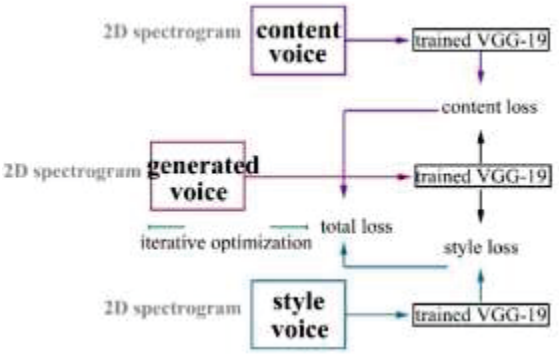

Figure 5. The speech style transfer model

In figure 5, the essential function of convolutional neural network is to extract the feature information of the input (2D spectrogram) layer by layer. After the layered extraction of convolutional layer, pooling layer, full connection layer and other network layers, the feature information of 2D spectrogram will become more and more advanced and abstract. That is to say, the low-layer convolution filters in the convolutional neural network tend to extract content feature information (edge, texture and color etc.) for 2D spectrogram,the high-level convolution filters tend to extract style feature information (a rough skeleton or layout etc.) for 2D spectrogram.

The 2D spectrograms of content speech, style speech and generated speech are denoted by $\overrightarrow{C_v}$ 、$\overrightarrow{S_v}$ and $\overrightarrow{G_v}$ respectively in this paper. Next, this paper will introduce the relevant principles of the creative research scheme - The Speech Style Transfer Model Based on Convolutional Neural Network. Figure 6 is a schematic diagram of feature map of the l th layer extracted from the spectrogram by convolution filters. Detailedly, n_W 、n_H and n_C represent width,height and number of channels respectively in the feature map of the l th layer. And $a^{[l]}_{i,j,k}$ Represents the activation value at the coordinate point (i, j,k) in the l th layer feature map, $i=1,2,\ldots,n_H, j=1,2,\ldots,n_W, k=1,2,\ldots,n_C$.

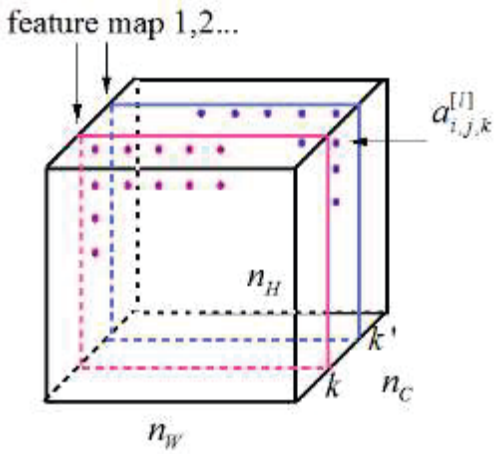

Figure 6. The l th layer feature map of 2D spectrogram

3.1.1. The Extraction of Content Features from The 2D Spectrogram of Content Speech

The features of the l th high layer's feature map extracted by the convolution filters are selected as content features in the spectrogram of generated speech.

Content loss function $J_{Content}(\overrightarrow{C_v},\overrightarrow{G_v})$ is defined as the sum of squared errors of activation values from the content speech spectrogram $\overrightarrow{C_v}$ and the generated speech spectrogram $\overrightarrow{C_v}$ in the l th layer feature map:

$$J_{Content}(\overrightarrow{C_v},\overrightarrow{G_v}) = \frac{1}{2}\left\| a^{[l](\overrightarrow{G_v})} - a^{[l](\overrightarrow{C_v})} \right\|^2 = \frac{1}{2}\sum_{i=1}^{n_H}\sum_{j=1}^{n_W}\sum_{k=1}^{n_C}\left(a^{[l](\overrightarrow{G_v})}_{i,j,k} - a^{[l](\overrightarrow{C_v})}_{i,j,k}\right)^2 \quad (1)$$

The content loss function $J_{Content}(\overrightarrow{C_v},\overrightarrow{G_v})$ measures the similarity between content speech spectrogram $\overrightarrow{C_v}$

generated speech spectrogram $\overrightarrow{C_v}$ in content features such as skeleton and layout.

3.1.2. The Extraction of Style Features from The 2D Spectrogram of Style Speech

The features of the low layers' feature maps extracted by the convolution filters are selected as style features in the spectrogram of generated speech.

Step 1: The style matrix of spectrogram, also known as Gram matrix, is used to measure the associativity between different sections (1, 2...) in a certain feature map. Define the style matrix $G^{[l](\overrightarrow{S_v})}$ or the style speech spectrogram $\overrightarrow{S_v}$:

$$G_{kk'}^{[l](\overrightarrow{S_v})} = \sum_{i=1}^{n_H^{[l]}} \sum_{j=1}^{n_W^{[l]}} a_{ijk}^{[l](\overrightarrow{S_v})} a_{ijk'}^{[l](\overrightarrow{S_v})}; \quad (2)$$

Define the style matrix $G^{[l](\overrightarrow{G_v})}$ for the generated speech spectrogram $\overrightarrow{C_v}$:

$$G_{kk'}^{[l](\overrightarrow{G_v})} = \sum_{i=1}^{n_H^{[l]}} \sum_{j=1}^{n_W^{[l]}} a_{ijk}^{[l](\overrightarrow{G_v})} a_{ijk'}^{[l](\overrightarrow{G_v})}. \quad (3)$$

Figure 7 is a schematic diagram of Gram matrix calculation process including multiplication and sum operations for activation values $a_{i,j,k}^{[l]}$ and $a_{i,j,k'}^{[l]}$ at corresponding positions.

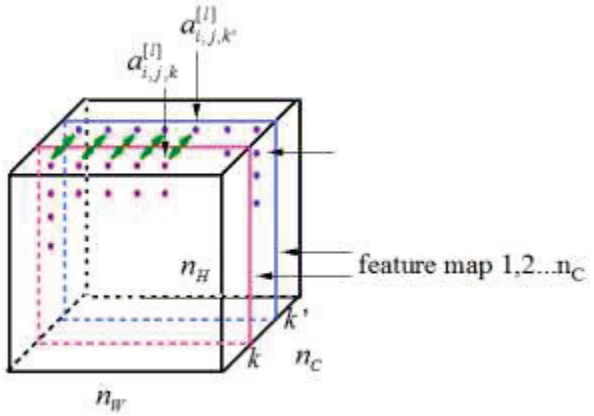

Figure 7. The Gram matrix calculation process

The Gram matrix of spectrogram measures whether or not two features appear simultaneously in the spectrogram and show the response between two features when they appear together.

Step 2: Define style loss function $E_l = J_{style}^{[l]}(\overrightarrow{S_v}, \overrightarrow{G_v})$ for the l th layer feature map:

$$E_l = J_{style}^{[l]}(\overrightarrow{S_v}, \overrightarrow{G_v}) = \frac{1}{(2n_H^{[l]} n_W^{[l]} n_C^{[l]})^2} \left\| G^{[l](\overrightarrow{S_v})} - G^{[l](\overrightarrow{G_v})} \right\|_F^2 = \frac{1}{(2n_H^{[l]} n_W^{[l]} n_C^{[l]})^2} \sum_{k=1}^{n_C} \sum_{k'=1}^{n_C} (G_{kk'}^{[l](\overrightarrow{S_v})} - G_{kk'}^{[l](\overrightarrow{G_v})})^2 \quad (4)$$

Step 3: Finally, style loss function $J_{style}(\overrightarrow{S_v}, \overrightarrow{G_v})$ of speech spectrogram is defined as the weighted sum of multilayer style loss functions , 1, 2,...

$$J_{style}(\overrightarrow{S_v}, \overrightarrow{G_v}) = \sum_l w_l E_l. \quad (5)$$

The style loss function $J_{style}(\overrightarrow{S_v}, \overrightarrow{G_v})$ measures the similarity between style speech spectrogram $\overrightarrow{S_v}$ and generated speech spectrogram $\overrightarrow{C_v}$ in style features such as edge, texture and color.

3.1.3. Iterating to Achieve Generated Speech Spectrogram, and Converting It to Audio File

In the end, we define total loss function $J(\overrightarrow{G_v}, \overrightarrow{C_v}, \overrightarrow{S_v})$ of the generated speech spectrogram comparing with the content speech spectrogram and style speech spectrogram as:

$$J(\overrightarrow{G_v}, \overrightarrow{C_v}, \overrightarrow{S_v}) = \alpha J_{Content}(\overrightarrow{C_v}, \overrightarrow{G_v}) + \beta J_{Style}(\overrightarrow{S_v}, \overrightarrow{G_v}). \quad (6)$$

Minimizing the total loss function $J(\overrightarrow{G_v}, \overrightarrow{C_v}, \overrightarrow{S_v})$. The stylized spectrogram $\overrightarrow{G_v} := \overrightarrow{G_v} - \lambda \frac{\partial J(\overrightarrow{G_v})}{\partial \overrightarrow{G_v}}$ of generated speech can be obtained iteratively with the aid of gradient descent method. Finally, the audio file of generated speech signal can be acquired according to the stylized spectrogram. Figure 8 is an iteration acquisition diagram for the audio file of generated speech.

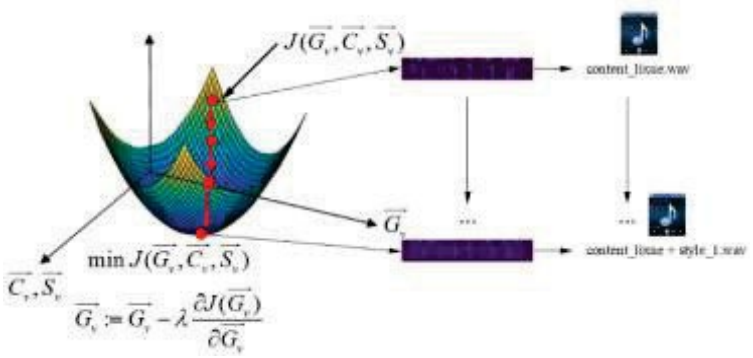

Figure 8. The iteration acquisition diagram for the audio file of generated speech

References

[1] Childers D G, Wu K, Hicks D M, et al. Voice conversion[J]. Speech Communication, 1989, 8(2):147-158.

[2] Byron D K, Pikovsky A, Woods E. Text-to-speech for digital literature, US9183831[P]. 2015.

[3] Schwardt L. C., Du Preez J. A., Voice conversion based on static speaker Characteristics IEEE COMSIG-98, Cape Town, September 1998, 57~62.

[4] Qi Yingyong, Weinbery B. Bi Ning. Enhancement of female esophageal and tracheoesophageal speech, J. Acoust. Soc. Am., Nov. 1998, (5): 2461~2465.

[5] Sundermann D., Ney H., Hoge H., VTLN-based cross-language voice conversion. In IEEE Automatic Speech Recognition and Understanding Workshop, 2003, 676~681.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," ICASSP, 1988:655-658.

[7] M. Savic, and I. Nam, "Voice personality transformation," Digital Signal Process, no.1, pp.107-110, 1991.

[8] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," Speech communication, vol.16, no.2, pp.207-216, 1995.

[9] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on radial basis function networks," International Conference on Spoken Language processing, pp.789-793, 2002.

[10] R. C. Guido, L. Sasso. Vieira, S. Barbon Junior, F. L. Sanchez, C. Dias Maciel, E. Silva Fonseca, and J. Carlos Pereira, " A neural-wavelet architecture for voice conversion," Neurocomputing, vol.71, no.1-3, pp.174-180, Dec.2007.

[11] Nirmal J, Zaveri M, Patnaik S, et al. Voice conversion using General Regression Neural Network[J]. Applied Soft Computing, 2014, 24(24):1–12.

[12] Ghorbandoost M, Sayadiyan A, Ahangar M, et al. Voice conversion based on feature combination with limited training data[J]. Speech Communication, 2015, 67(67):113-128.