Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 1, 2023

Research on speech style transfer algorithm combined with image processing perspective

Yuanqi Chen

School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, Anhui 233030, China

Abstract: Speech, as the acoustic expression of language, is one of the most natural and effective means of human information communication. With the rapid development of the Internet and communication technology, the function of robot voice interaction is more and more popular among people. However, the robotic pronunciation cannot meet people's demand for personalized voice interaction. At the same time, style transfer technology, which is widely used in image and video processing, has been relatively mature. By studying the theoretical methods of generalized style transfer technology (including the style transfer of images and video signals), and comparing and analyzing various machine learning algorithms used by the current voice style transfer technology, this paper draws the following conclusions: First, various models generally have the problem of large demand for training data and difficulty in training. Second, an algorithm model shows the alienation effect in different usage scenarios. Finally, based on the above problems, suggestions for the development of voice style transfer are put forward.

Keywords: Style transfer; Machine learning; Speech; Image.

1. Introduction

1.1. Research background and significance

As a bridge of human communication, voice information contains richer emotional elements than text information, and is an indispensable part of communication in today's network information era. In the era of voice calls and voice interaction devices everywhere, voice signals are so common but so important. With the rapid development of electronic technology products and people's dependence on audio-visual entertainment is only increasing, the degree of intelligence of voice interaction can no longer meet people's needs.

With the rapid development of computer performance and the rise of machine learning, style transfer technology is widely used in image, video and audio signal processing, defined as the extraction of the style features of subject A, applied to subject B, to obtain subject B with features of subject A. Speech signal processing is embodied in the extraction of the target speaker's speech features (timbre, tone or emotion) information, applied to another speech signal, producing speech with the target speaker's voice characteristics. The feature, which we often see in voice assistants on smartphones, records a piece of audio on request to generate a unique voice style. With the development of voice style transfer technology, it is widely used in the medical field. Speech style transfer technology is expected to play a positive role in the adjuvant therapy of autistic patients[1], and relevant studies have shown that synthetic speech with specific speech styles can better achieve the purpose of communication in the caregivers of the elderly and children.

The technique of style transfer was originally used in image processing. After a large number of scholars' long-term research and the continuous improvement of various algorithms, the technique of image style transfer has been relatively mature. In order to provide development opportunities and innovation space for speech style transfer technology which is still in long-term development, it is of great significance to learn and summarize the general rules of style transfer in image processing. At the same time, there are many kinds of machine learning algorithms about style transfer technology at home and abroad, which correspond to different use environments, and the effect of each algorithm is obviously different. This paper tries to study the realization principle of image and video style transfer technology, and summarizes its key. Secondly, in the face of complex and changeable environmental influences, the effect of different conversion algorithms to achieve voice style transfer is compared and analyzed, and the advantages and disadvantages of each algorithm in different scenes are summarized. Finally, in view of the existing problems of speech style transfer, such as low naturalness of speech, high cost of modeling resources or easy to be disturbed by environment, some suggestions are provided for the development of speech style transfer technology.

1.2. Research status at home and abroad

The concept of style transfer originated from the field of image processing in the early stage, and the research on voice style transfer began in the 1980s. In the early stage, Abe et al. proposed using vector quantization codebook mapping to realize speech conversion[2]. In China, Chu Min et al.[3] proposed a method based on Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) to realize the conversion of male voice and female voice. Johnson et al. [4] proposed a feedforward neural network with perception loss function, which consists of two parts: One is the image generation network, which is a deep residual network, which can be trained to get a specific style. The other is the loss network (also called feature extraction network), which is used to calculate the size of the gap between the input content image and the style image, and provides the gradient for the parameter update of the image conversion network, although the speed is much improved. But the network needs to be retrained for different styles. Shortly after this research, Goodfellow et al. [5]proposed a generative adversarial network (GAN)model. The network consists of two parts, generator and discriminator, and learns the generative model by means of adversarial game. Zhu et al.[6] proposed the

CycleGAN framework, which used generators and discriminators to map and reflect images from the source domain to the target domain, and obtained source images similar to the target image through 1L norm constraints.

Although style transfer is presented in more diversified ways under the framework of deep neural network, there is still a lot of room for improvement. In the aspect of application, we can consider applying style transfer to speech signal to realize the personalized function of speech so as to make its application more extensive.

2. The application of style transfer in image processing

Image style conversion is an increasingly popular topic in the field of artificial intelligence and computer vision, using deep learning models to transform images into different visual styles. The technology has the potential to expand creative horizons, allowing artists to create a variety of new visual styles of art quickly and easily.

The concept behind image style conversion is based on the idea that an image's content can be separated from its style. Machine learning algorithms are used to identify features in an image and then transmit those features to another image. This allows the content of one image to be combined with the style of another image to create an entirely new image.

To achieve a successful transformation of image styles, several factors must be considered. First, in order to achieve a high-quality migration effect, you need to accurately represent the source image and the target image. This includes making sure you use the same resolution, aspect ratio, and color. Second, the machine learning model must be trained on the original image to learn how to extract the visual features of the image. Finally, the target image needs to be provided so that the model can add the style of the source image to the target image.

The most commonly used image Style Transfer algorithms include Neural Style Transfer, Feature Transfer Learning, CycleGAN and Instance normalization And Normalization, along with Texture Networks.

2.1. Neural Style Transfer

By speeding up the process of deep learning, neural style transfer algorithm can stylize images quickly and effectively. It can be compared with the more advanced algorithms in style transfer effect, while keeping the complexity of algorithm implementation low. However, this method requires high computer performance and computing power, so it may consume a lot of time and resources in the computing process if it is not properly operated. At the same time, this approach is relatively difficult to implement when faced with the task of repainting, stylizing photos with complex patterns or height variations.

2.2. Feature Transfer Learning

Feature transfer learning is an effective method to train deep neural networks, and research shows that this method is effective in the application of style transfer. Feature transfer learning can transform source image styles into target images quickly and accurately. In addition, it tends to learn the most salient features of the source image and apply them to the target image. The disadvantage of feature transfer learning is that it is difficult to find the correct configuration or hyperparameter to control feature transfer. In addition, since the target image may not accurately transmit all the features of the source image, the effect of style transfer is unstable.

2.3. Loop Generation Network (CycleGAN)

Cyclic generation network algorithm is one of the most commonly used algorithms in image style transfer. The ability to realize style transfer between different style images makes this algorithm very useful for practical applications. However, in order to get a better effect, the cyclic generation network needs to carry out several iterations, which makes the algorithm take a long time to execute. In addition, it may not accurately capture more subtle features in the image, such as texture or gloss.

2.4. Instance Normalization

The case normalization method enables the neural network to better learn the feature differences between each image. It also makes the network training process more stable and relatively easy to implement, but at the same time, this method will also reduce the effectiveness of style transfer, because the instance normalization method allows the network to be more affected by the source image content than the target image.

2.5. Texture Networks

Texture networks accurately capture finer details of an image, such as color and texture. And the algorithm can learn the style of the target image in a way similar to that of a human artist. However, training a mature texture network model requires a large amount of training data, which leads to the high cost of this method.

3. Analysis of speech style transfer algorithm

3.1. Introduction to Common Algorithms

Widely used audio style migration algorithms include convolutional neural networks (CNN), generative adversarial networks (GAN) and recursive neural networks (RNN). In addition, several researchers have implemented audio style transfer algorithms for transfer learning, such as CycleGAN and Domain Adversarial Neural Networks (DANN).

Convolutional neural networks (CNN) can efficiently extract features from raw audio data and generate high-quality migrated audio. However, this model requires a large amount of training data and is more difficult to train than other models. In contrast, generative Adversarial networks (GAN) can generate high-quality migrating audio while requiring less training data than some other algorithms. However, GAN model may be difficult to train and prone to overfitting. The second is recursive neural network (RNN), which can capture the long-term time dependence of audio signals that are difficult to be captured by other models. Its disadvantage is that the algorithm is relatively complex, which makes the model training difficult. CycleGAN is relatively easy to train and can be used to bridge the gap between two different audio domains, but if the domains are too different, CycleGAN will produce distorted audio, and the migrated audio may not adequately reflect the content of the original signal. Domain adversarial neural networks (DANN) can be very effective in adjusting the content of an audio signal while ensuring its stylistic transfer. However, DANN also requires a large amount of training data and requires a large amount of calculation.

3.2. GAN-based voice style migration

Compared with the various speech style transfer models introduced above, the speech style transfer technology based on Generative Adversarial Networks (GAN) is characterized by a small amount of training data and low difficulty. At the same time, the quality of the generated speech is high. This model was proposed by Goodfellow et al in 2014. The network consists of two parts, generator and discriminator, and learns to generate the model through the way of antagonistic game.

The implementation process of GAN model in voice style transfer is as follows:

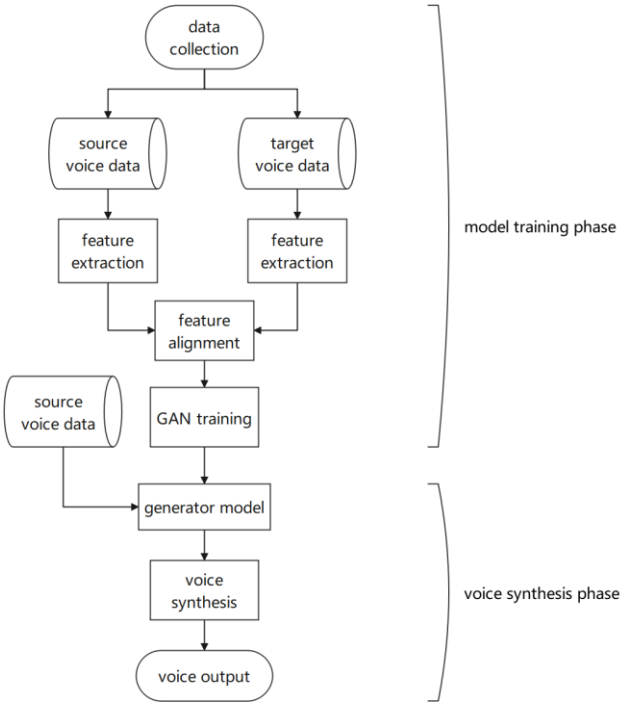

Figure 1. Overall block diagram of voice style transfer based on GAN

Gan-based voice style transfer is a deep learning technique that trains a generative adversarial network (GAN) to learn each speaker's vocal characteristics, including intonation, pitch, pitch, and other vocal characteristics, including the following:

First, the GAN-based voice style transfer model is trained by learning spoken audio in two different domains, such as the mapping between male to female or English to Chinese.

Second, the model utilizes two neural networks to generate the converted speech: a network of generators to generate the converted audio sample, and a network of discriminators to classify whether the audio sample is real or generated.

Third, the model is trained in an adversarial fashion, where generators are trained to test discriminators and discriminators are trained to recognize real samples.

Fourth, the model is unsupervised, that is, it does not require any training on parallel data.

Fifth, models can learn to transform audio to a target style without any explicit comments about the style, such as gender or language.

4. Conclusion

At present, there is a lack of reliable data set to train the algorithm for voice style transfer, and various algorithms have their own advantages and disadvantages, so the implementation effect is very different. Some algorithms are difficult to model, the computer resources are expensive, and the traditional transfer methods have sub-optimal speech perception, or environmental interference factors lead to audio distortion.

In order to improve the quality of synthesized audio, voice style transfer technology needs to continue to develop in four aspects: First, prosody modeling, which is essential for successful speech style conversion. This involves automatically learning style-specific representations of prosody from training data so that the same synthetic model can be applied to different styles. Second, data quality, in order to produce high-quality speech synthesis models, the training data must be of high quality and large enough to capture all the nuances of the style. This can include the use of high-quality recording equipment and the use of large data sets in a variety of styles. Third, expressive speech synthesis, adding expressiveness to synthesized speech is also important because it helps to make synthesized speech more natural. This can include incorporating features such as intonation, pauses, and rhythm into the composite model. Fourth, unsupervised style transformation. Unsupervised style transformation is a promising technique for producing high-quality speech synthesis. This technique uses unsupervised learning algorithms to learn style-independent speech representations, allowing the same model to be used for different styles.

Acknowledgments

This research was supported by the Undergraduate Research Innovation Fund of Anhui University of Finance and Economics (Project number: XSKY22152). The project name was "Research on Speech Style Transfer Technology based on Machine Learning".

References

[1] Ren Qiang. Research and application of speech style transfer technology based on Generative Adversarial Network [D]. Chongqing University of Technology,2019.

[2] Gray R M. Vector Quantization[J]. Readings in Speech Recognition, 1990, 1(2):75–100.

[3] Chu min, Lv Shinan. The invention relates to a synthesis method combining PSOLA algorithm and speech sine model [C]// Proceedings of the Fifth National Conference on Human-Computer Voice Communication. 1998.

[4] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution[C]. European conference on computer vision, 2016: 694-711.

[5] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. Advances in neural information processing systems. 2014: 2672-2680.

[6] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. Proceedings of the IEEE international conference on computer vision, 2017:2223-2232.