International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Voice Translation and Modification using Convolutional Neural Networks

Shivanand[1], Guruprasad K[2], Merin Meleet[3]

[1,2,3] R.V. College of Engineering, Bengaluru, Karnataka, 560059

ABSTRACT

A person who imitates the voices and speech patterns of others is known as an impressionist. Such a task has been accomplished by humans over the years. Thousands of speakers can be distinguished by the human auditory system, but little is known about the specific qualities that enable this. Although Fourier Transforms can record the pitch and harmonic composition of the speaker, this alone is insufficient for uniquely identifying speakers. We knew very little about the remaining structure, often known as timbre, which is essential for distinguishing speakers. We used the conventional approaches for neural networks after being inspired by recent work on neural network picture production. By altering the speaker's timbre as well as pitch, we were able to change the voice of one speaker into another using recent developments in neural networks. Our initial findings are positive when transferring voices from one speaker to another. After doing the subsequent study, we have come to the conclusion that Convolutional Neural Networks are capable of changing one speaker's voice into another. Our findings demonstrate the effectiveness of our suggested model in creating a superior synthesised speech. When used to identify synthetic voices, translate music from one genre to another, select voices for AIs based on preferences, etc., this research has a wider use.

Keywords: CNN, style transfer, voice conversion,spectrum.

1. Introduction

Voice conversion is a crucial element of artificial intelligence (VC). It is the study of voice conversion, or imitating another person's voice without changing the meaning of the language. Three qualities of a speaker can be distinguished generally: There are three types of characteristics: 1) linguistic (sentence structure, lexical choice, and idiolect); 2) supra-segmental (prosodic elements of a speech signal); and 3) segmental (related with short-term properties and including spectrum and formants). When the linguistic content is fixed, the supra-segment and segmental elements are important aspects relevant to speaker individuality. Both the supra-segment and the segmental factors should be converted via an efficient voice conversion approach.. Voice conversion is still far from ideal despite significant advancements. The segmental and supra-segmental elements should both be transformed using a successful voice conversion strategy. Voice conversion has come a long way, but it's still far from perfect. Voice conversion pipelines, also known as analysis-mapping reconstruction pipelines, are where you'll often find modules for speech analysis, mapping, and reconstruction. The reconstruction module resynthesizes time-domain speech signals, while the mapping module maps the source speaker's speech signals to the target speaker. The speech analyzer separates a source speaker's speech waveforms into features that stand for supra- and segmental information. In many of the studies, the mapping module has assumed the lead role. These methods can be divided into various categories, such as those based on the training data used—parallel vs. non-parallel; the type of statistical modelling method—parametric vs. non-parametric; the level of optimization—frame level vs. utterance level; and the conversion process—direct mapping vs. inter-lingual. A deep neural network trained on discrimination can roughly flip the representation it learns, converting it from a classification model to a generator. Even though precise inversion is impossible, the backpropagation technique can be utilised to find the inputs that lead the network to behave as desired. This approach has been used in the computer vision field to understand how networks function, identify adversarial examples that slightly change picture inputs to change the predictions provided by the network, synthesis textures, and renew an image in line with another's aesthetic (basically matching the low-level texture).

2. Literature Review

The goal of [1] Voice Conversion using Convolutional Neural Networks by Shariq A. Mobin and Joan Bruna was to employ current developments in neural networks to alter the voice of one speaker into another by transforming not only the speaker's pitch but also the timbre. The model does a good job of capturing the speaker's harmonic structure, but the frequency resolution is not great. As a result, training generative adversarial networks has proven to be highly challenging in practice, and further research is required to determine how to best optimize the model created here.

The goal of [2] Convolutional Neural Networks-based Continuous Speech Recognition utilizing Raw Speech Signal was to apply the CNN-based method to a problem requiring recognition of a vast vocabulary. To contrast the standard ANN-based technique with the CNN-based approach more specifically. Studies on the Wall Street Journal corpus demonstrated that the CNN-based system outperforms the ANN-based system, which uses input from traditional

cepstral characteristics, in terms of performance. For example, the word error rate (WER) for the CNN-based system is 6.7%, while the WER for the ANN-based system is 7%. The CNN-based technique has a language dependence.

Outlining how ML techniques are employed is the main objective of [3] On Using BACKPROPAGATION for Speech Texture Generation and Voice Conversion by Jan Chorowski, Ron J. Weiss, Rif A. Saurous, and Samy Bengio. a CTC model that has been trained on a 13-layer CONV architecture. model that generates a statistical description of the target voice using the activations of a deep convolutional neural network trained for speech recognition. The main benefit of the suggested technique is its ability to utilise only a small amount of information from the target speaker. Recognizable components of the target voice can be combined in just a few lines of speech. The slowness of the proposed strategy was a drawback. The synthesised utterances' quality is also very subpar.

In [4] A Survey on Voice Conversion Using Deep Learning by Benjamin Meier, This study describes an efficient system for the indicated purpose and provides a summary of the most recent deep learning-based voice conversion techniques. Impressive results for style-transfer images demonstrate that neural networks can even handle challenging transfer/conversion jobs. This survey provides an overview of the approaches that are now available and may be used to solve the voice conversion challenge using an end-to-end learning network architecture, or at the very least, to introduce the subject. Being a survey, there were no gaps.

In [5], Aakash Ezhilan, R. Dheekksha, and S. Shridevi discuss the use of deep learning for audio style conversion. The goal was to use spectrograms to construct a vocal style transfer utilising CNN architecture. various architectures are investigated 1. Simple auto encoder was implemented using a naive approach, using the male spectrogram as input and representing it as a latent vector. 2.a unique loss-based auto-encoder that checks the output to determine if the content matches the input spectrogram and if it resembles a female spectrogram (if the input is male). Simple mentality We developed and evaluated a straightforward auto encoder architecture that produces the output for a given male-like spectrogram.

The goal of [6] Voice Conversion Using Voice-to-Speech Neuro-Style Transfer by Ehab A. AlBadawy, and Siwei Lyuthe was to employ the CNN-based approach to a speech recognition problem with a wide vocabulary. To contrast the standard ANN-based technique with the CNN-based approach more specifically. It was done using a novel voice conversion technique built on a neural style transfer model of the mel-spectrograms. The recent advancements in neural network models for visual style transfer are utilized by this technique.

3. Hypothesis

Style transfer, which was initially presented in the visual domain, is the process of altering a picture or video so that it takes on the visual style of an additional image or video. Parallel initiatives in the audio realm have evolved as attempts to conduct visual style transfer have grown more successful with the use of deep learning models. Style transfer for the audio modality refers to altering the style of an audio sequence to make it appear as though it were created by another audio source. Often, tasks like making music played on one instrument sound like it was played by another or making a spoken utterance sound like it was said by someone else are explored (voice conversion).

The intriguing issue of style transfer has drawn a lot of attention recently.

Recent years have seen a huge increase in the use of convolutional neural networks in many deep learning applications. The transfer of image styles is one of these applications. In line with this development, we may investigate how this method might be used with audio data. The technique mentioned includes fusing the stylistic and content characteristics of two different audio samples. An audio signal's features can be extracted using a convolutional neural network.

Multiple speech processing techniques, including speech analysis, spectral conversion, prosody conversion, speaker characterization, and vocoding, are used in voice conversion. We can now produce human-like voice quality with great speaker similarity thanks to recent breakthroughs in theory and practise.

4. Methodology

Audio Pre-processing

The audio files used were in the 44100 Hz sampled WAV format. Utilizing formats like WAV is essential because they store data in an uncompressed manner that makes data extraction simple. This method doesn't work with stereophonic audio; it only works with single channel audio files. Two audio files—the style audio and the content audio—make up the inputs. First, the time domain to frequency domain conversion of the audio signal is performed.

Modelling the Convolutional Network

In order to extract high-level characteristics from the audio inputs, a convolutional layer is necessary. In this implementation, a single convolution layer with 4096 filters is used. The matrices acquired after preprocessing the content and style audio signals are appropriately reshaped to comply with the convolutional network's input shape requirements. To produce the content feature map and the style feature map, the content and style matrices are each individually fed into the convolutional neural network. We employ a Gram matrix to extract the high-level style traits. The inner-product of the style feature map with itself is a Gram matrix.

Defining the Loss

Loss of both content and style is defined separately. An input sample containing random values is fed into the network to assess the loss, and the network's output is then processed to extract features. These output properties are also represented as a Gram matrix. The network's capacity to produce the same response as the content and style features from the random input sample is determined by the loss.

Training the Network

The network has been taught to minimize overall loss. To get better outcomes, the learning rate can be changed appropriately. The learning rate used in this implementation is 0.001. The maximum number of training epochs can be restricted to save training time, but too few epochs will lead to lower output quality.

Reconstructing the Audio

The network's ultimate output is obtained after the training is finished. It is necessary to process this output in order to extract the audio from it. This is accomplished by first exponentiating each element in the output. from the frequency domain, return the signal to the time domain. The output is written to a WAV file in integer format.

5. Result and Analysis

The generated spectrogram compared with content and style.

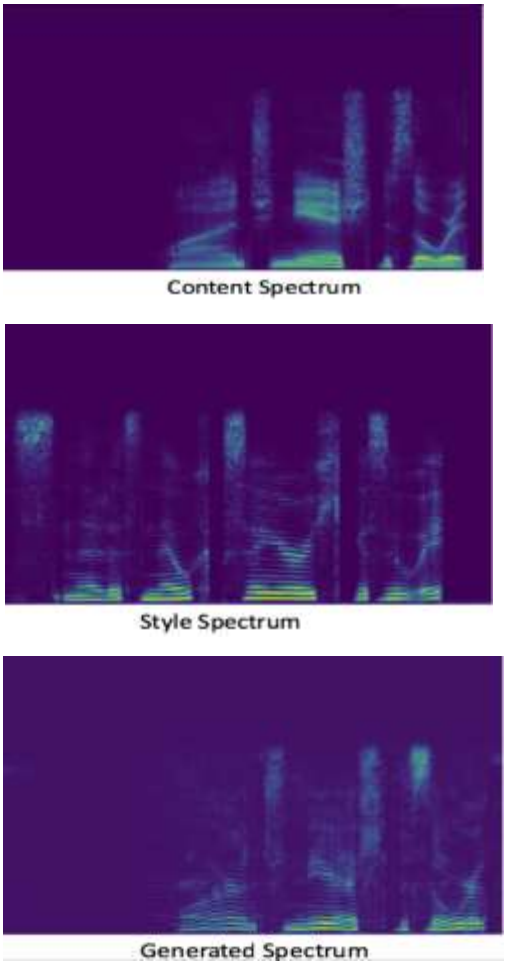

Fig 1: Spectrum images

by contrasting the spectrogram produced with the spectrograms of content and style. (where the X axis represents the Time Domain and the Y axis the Frequency Domain), we can discover that:

- The structure is almost the same as content, and the gap along the frequency axis, which determines the voice texture to a great extent, is more alike to the style.

- The base skeleton is shifted upward a little bit for being similar to the style.

Highlight of work was:

- Use 2-D CONV rather than 1-D for audio spectrograms.
- Compute grams over time-axis.
- quick training. On a single GPU, training and transfer take 5–10 minutes (Tesla P40).
- No need for a dataset! Any two audio files are eligible for transfer. (However, some audio formats could have problems; in that case, do sudo apt-get install libav-tools.)

Loss for content, style and total loss calculated.

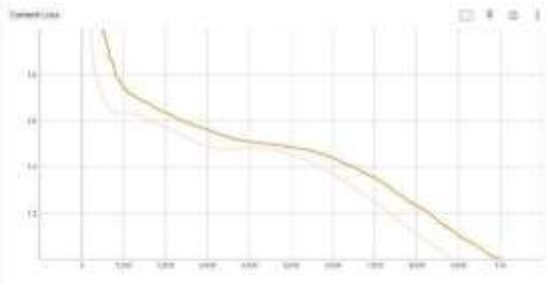

Fig 2: Content loss with number of epochs

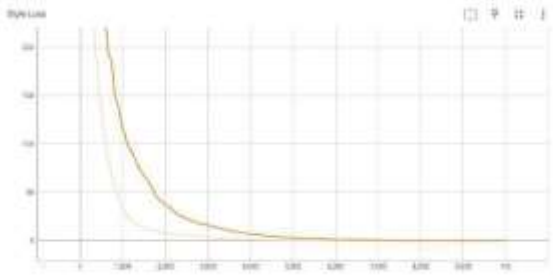

Fig 3: Style loss with number of epochs

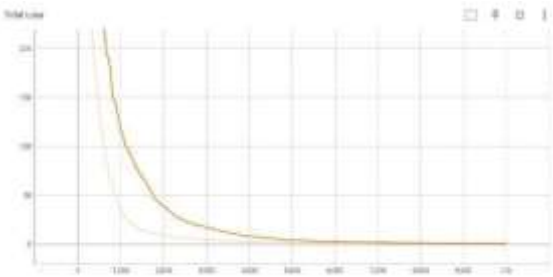

Fig 4: Total loss with number of epochs

6. Conclusion and Future Work

In order to transfer the timbre of one speaker to another, we first created algorithms. Although there is still work to be done, our algorithms were occasionally able to produce speech that was perceptually similar to the target speaker. Due to the audio's mathematical roots rather than being wholly constructed by people utilising various instruments, the outcomes of this work may be easily distinguished from audio that has been traditionally produced. The result demonstrates that it is feasible to transfer the stylistic elements of one audio to another in order to produce a completely new audio that is a distinctive fusion of the two. Depending on the desired effect, the style may be translated differently. This method could be used to change the accent of human speech that has been recorded, remix songs, change the beat, and more.

In our upcoming work, we intend to expand our dataset in order to give it even more variability so that we may test it using various CNN architectures and cross-sensor validation techniques. Our proposed model's use on audio files of many languages is another area of investigation. Increasing the number of convolutional layers to include more high-level characteristics and improve style transfer would be an intriguing future development. Two adjustments

can also be implemented to raise the output's quality. The first is to as much as possible lower the output noise level, and the second is to process multi-channel audio files. To further improve the output audio quality, the number of epochs could be raised, although this is heavily hardware dependent.

References

[1]. Nicholas J. Bryan, Jorge Herrera, and Ge Wang (2012) published "Userguided variable-rate time-stretching via stiffness control" in the proceedings of the 15th international conference on digital audio effects (DAFx).

[2]. Sebastian Bock and Gerhard Widmer, "Maximum filter vibrato suppression for onset identification," Proc. of the 16th International Conference on Digital Audio Effects (DAFx), 2013.

[3]. "Voice morphing system for impersonating in karaoke applications," in Proceedings of the International Computer Music Conference (ICMC), 2000, by Pedro Cano, Alex Loscos, Jordi Bonada, Maarten de Boer, and Xavier Serra.

[4]. Vocalistener is a singing-to-singing synthesis system based on iterative parameter estimation, as described by Tomoyasu Nakano and Masataka Goto in Proceedings of the Sound and Music Computing Conference, 2009, pp. 343–348.

[5]. Maximum filter vibrato suppression for onset detection, S. Bock and G. Widmer, Conference on Digital Audio Effects, 2013.

[6]. "Singing Expression Transfer from One Voice to Another for a Given Song," S. Yong and J. Nam, IEEE ICASSP, pp. 151–155, 2018.

[7]. Highway networks, S. Rupesh Kumar, G. Klaus, and S. Jurgen, arXiv preprint arXiv:1505.00387, 2015.

[8]. P. Senin, “Dynamic time warping algorithm review,” Technical reports, Information and Computer Science Department, University of Hawaii, USA, pp. 1-23, vol. 855, 2008.

[9]. W. Xin, S. Takaki, and J. Yamagishi, “Investigating very deep highway networks for parametric speech synthesis,” SSW-9, 2016.

[10]. Input-to-Output Highway Networks for Voice Conversion, IEICE Transactions on Information and Systems, 100, No. 8, 2017, pp. 1925–1928. S. Takamichi, H. Saruwatari, and Y. Saito