

Audio style conversion using deep learning

Aakash Ezhilan, R. Dheeksha, S. Shridevi*

Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, India

ABSTRACT

Style transfer is one of the most popular uses of neural networks. It has been thoroughly researched, such as extracting the style from famous paintings and applying it to other images thus creating synthetic paintings. Generative adversarial networks (GANs) are used to achieve this. This paper explores the many ways in which the same results can be achieved with audio related tasks, for which a plethora of new applications can be found. Analysis of different techniques used to transfer styles of audios, specifically changing the gender of the audio is implemented. The Crowd sourced high-quality UK and Ireland English Dialect speech data set was used. In this paper, the input is the waveforms belonging to a male or female and waveforms belonging to the opposite gender is synthesized by the network, with the content spoken remaining the same. Different architectures are explored, from naive techniques and directly training audio waveforms against convolution neural networks (CNN) to using extensive algorithms researched for image style conversion and generation of spectrograms (using GANs) to be trained on CNNs. This research has a broader scope when used in converting music from one genre to another, identification of synthetic voices, curating voices for AIs based on preference etc.

Keywords: Style transfer, Audio analysis, Neural networks, Dialect transfer.

OPEN ACCESS


Received: January 29, 2021

Accepted: April 7, 2021

Corresponding Author:

S. Shridevi

shridevi.s@vit.ac.in

 **Copyright:** The Author(s). This is an open access article distributed under the terms of the [Creative Commons Attribution License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted distribution provided the original author and source are cited.

Publisher:

[Chaoyang University of Technology](https://www.chaoyang.edu.cn/)

ISSN: 1727-2394 (Print)

ISSN: 1727-7841 (Online)

1. INTRODUCTION

Style conversion of audio has many applications. It can be used in dubbing movies and synthesizing voices for text to speech systems. Voice conversion is a field of speech synthesis which changes elements like accents, emotion, pitch and identity (Mirjam et al, 2016; Lorenzo-Trueba et al, 2018). The main barrier in this field has been making the voice sound real and not machine generated. Griffin-Lim algorithm has been used to create realistic sounding audio in this paper. In this paper we explore the implementation of different architectures on the Crowd-sourced high-quality UK and Ireland English Dialect speech data set. In this we will be using southern English male and female, welsh English male and female datasets. It provides audios and `line_index.csv` containing the text voiced in the audio files. The expected outcome is to develop a voice translation style transfer using convolution neural network architecture, by using spectrograms (Pasini, 2019) and implementing generative adversarial networks. This network is compared against the naive implementation for accuracy. When the network is trained, it can identify if a male or female voice is speaking, converting one style of English speaking to another while keeping the content constant like changing the identity of the audio from male to female or vice versa (Sisman et al., 2020). The same architecture can be used for accent transfer also. The male/female identity change of the audio has been demonstrated in this paper. An analysis of how accurate the translated audio is comparing the spectrograms, and analyzing if the machine translated can be differentiated from the original audios as well is noted (Verma and Smith, 2018). The paper starts with an introduction to audio representation as waveforms and how different the waveform's properties are with respect to the individual's voice and whether it belongs to a male or female. Next, visual representation of these waveforms as spectrograms has been demonstrated. On performing linear transformation to the spectrograms, they were

converted to Mel-spectrograms to convert the values similar to what humans perceive. The differences between Mel-spectrograms of different genders have been highlighted. The paper then discusses three approaches in detail, with their architecture, implementation details and results of each along with the cause of success/failure in audio style transfer (Pasini, 2019). The first naive approach to the problem was by implementing a simple auto encoder (Hsu et al., 2017) which took the male spectrogram as the input and represented it as a latent vector. The architecture comprised of an encoder and a decoder. The encoder reduces dimensionality while the decoder gets the latent function and generates the spectrogram. Failure of pixel wise loss function for this approach has been explained. The second approach was the custom loss based auto-encoder. In this approach, the output is checked if it has the same content as the input spectrogram as well as if it resembles a female spectrogram or not (if the input is male). A Siamese network as well as a classifier was implemented here. While the results were comparatively better, the accuracy was still not great. The clarity of the spectrograms was bad and thus the audio was also not clear. Reasons of failure and methods of improvement have been discussed in detail. The final approach included a GAN (Generative adversarial network) (Kameoka et al., 2018). The GAN used the same generator along with the Siamese network and classifier in parallel appended with a fully connected network as the discriminator. This produced a realistic audio as its result with the content fully preserved and style transferred.

2. LITERATURE REVIEW

Several existing papers on style transfer on audios exist which use convolution networks, generative adversarial networks and deep neural networks (Huang et al., 2020). Audio style transfer (Grinstein et al., 2018) using shallow convolutional networks and random filters (Chen et al., 2020) proposes the use of convolution neural networks for style transfer in audio domain to generate a new audio. They have used continuous wavelet transfer to convert the audio waveforms into spectrograms. These were then used to create new images and then generate new audios using iterative phase reconstruction with Griffin-Lim algorithm. This method managed to successfully transfer audio like plain music but had yielded poor results in transferring audio that contained lyrics and metrics such as emotion or tone. This paper also proposes several measures to improve the quality of audio.

Neural style transfer for audio spectrograms (Verma and Smith, 2018) presents a method for creating new sounds using a GATYS approach. In this approach they have started from a random-noise input and have used back-propagation

to optimize the sound iteratively to conform to the outputs from a neural network. Audio style transfer for accents (Deshpande et al., 2019; Demirsahin et al., 2020) draws inspiration from work using generative models and auto encoders to transfer styles in images. This project proposes architecture with separate generators and discriminators, which is like a GAN-like structure but without the use of random noise for generation of fake data unlike the previous architecture. In this architecture is made up of a generator, discriminator and a vocoder (Tamamori et al., 2017; Hayashi et al., 2017).

MelGAN-VC: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms (Pasini, 2019; Huang et al., 2020) firstly computes spectrograms from the waveforms then translates domain using GAN. An additional Siamese network was used to preserve speech content in the translation process, without sacrificing the ability to change the style of the speaker's voice. Audio style transfer (Chen et al., 2020) explores two approaches, a neural based approach and an auditory based approach. The paper presented four sound texture models - SoundNet CNN, VGG-19 CNN, Shallow CNN and McDermott auditory model with several types of audio content and styles of sounds. Singing style transfer (Wu et al., 2018) using cycle-consistent boundary equilibrium generative adversarial networks was implemented using cycle GAN and CycleBEGAN.

3. SPECTROGRAM AND MEL SCALE

1. Audio represented as waveforms

Audio files such as human speech are essentially a waveform generated over a period of time. This can be represented as an array of float values which can be manipulated with to generate different sound forms and stylistic effects. We are considering two audios, where two narrator reads out the same sentence but in different dialect or the narrators have different gender (Miyoshi et al., 2017). The difference in the waveform generated for different narrators can be associated with features of an individual's voice such as pitch, frequency etc. In Fig. 1 two waveforms generated by two narrators is shown. We can see the difference in amplitude and shape between these waveforms as they belong to different gender.

2. Converting audio to spectrogram

Spectrograms are visual representation of waveforms, depicting the spectrum of frequencies as the signal varies over time. In this particular scenario the signal will be the audio files of human voice. For a given audio file (waveform) the associated spectrogram will be generated as shown in Fig. 2 alongside the wave waveform.

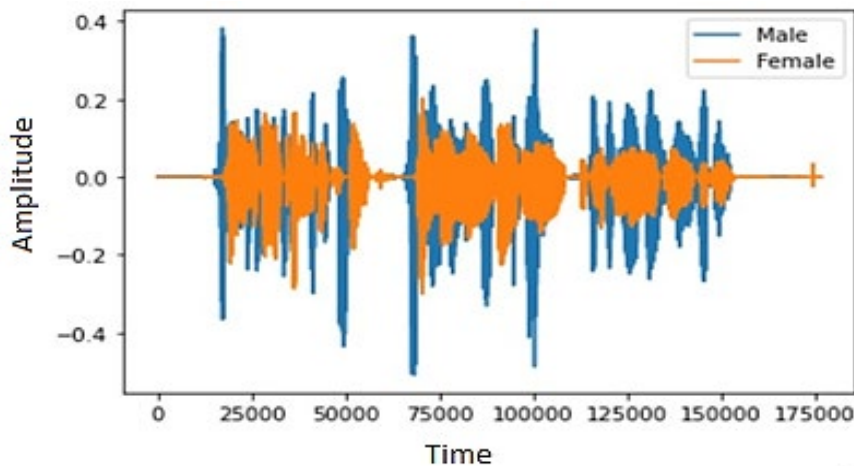


Fig. 1. Comparison of male and female voice

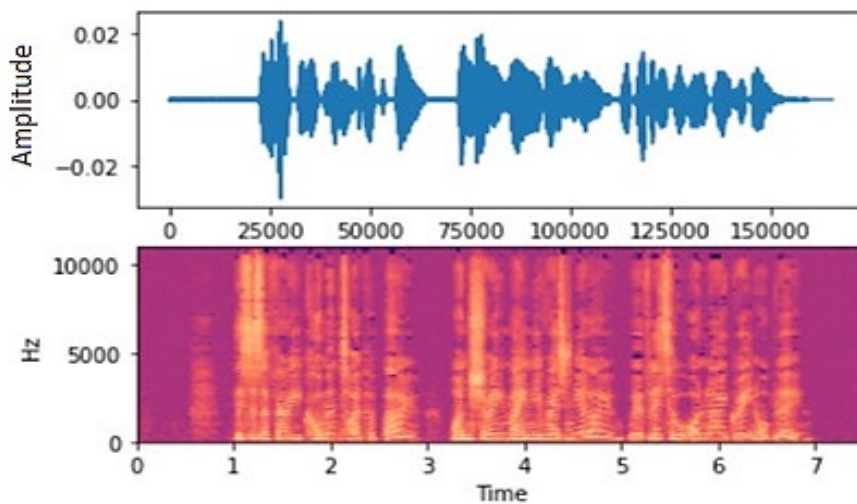


Fig. 2. Waveform (top) and the generated spectrogram of it (below)

3. Mel-scale and spectrogram

In the form of spectrogram it becomes easier to process the information in the audio file. The spectrogram generated however can be made more focused on human audience by converting them into mel-spectrogram. Mel-spectrogram is the output on performing a non-linear transformation thereby converting the given values similar to what humans perceive. Equation (1) converts a given frequency f to the corresponding mel-frequency. (Wester et al., 2016).

$$mel(f) = 2595 * \log_{10}(1 + f/700) \quad (1)$$

Mel-spectrogram is being considered here to represent the audio, as the data are going to be based on how humans hear the sound and distinguish the frequencies. On generating the mel-spectrogram for various audio signals of male and female voice, there is a noticeable pattern that can be seen.

We can observe the differences between the mel-spectrograms of male and female voices in Fig. 3. Two sentences have been represented, each voiced by a male and a female. In both the sentences a similar pattern is noticeable; the shading of the higher frequencies is darker in the female

spectrograms, while the shading of the lower frequencies is lighter in the female spectrograms. This corresponds to the fact that a female's voice is generally an octave higher than the male's voice. Differences like these will be learned by the networks while training and be projected leading to transformation of the audio's personality like its accent or gender.

From the crowd sourced high-quality UK and Ireland English Dialect speech data set, southern English male and female audios were chosen to use for gender based audio style transfer. In an effort to simplify the problem male to female and female to male was tried. This can also be used for accent style transfer. Dataset provided audios and a csv containing the text voiced in the audio files. Pickle files were generated to store the wave files of male and female in separately with each file named with the id provided in the dataset for each sentence. Then two pickle files of mel-spectrograms were created. The spectrograms are then reshaped, converted to mel-spectrograms and added to a numpy array to perform encoding. This was used in both the implementations described in this paper.

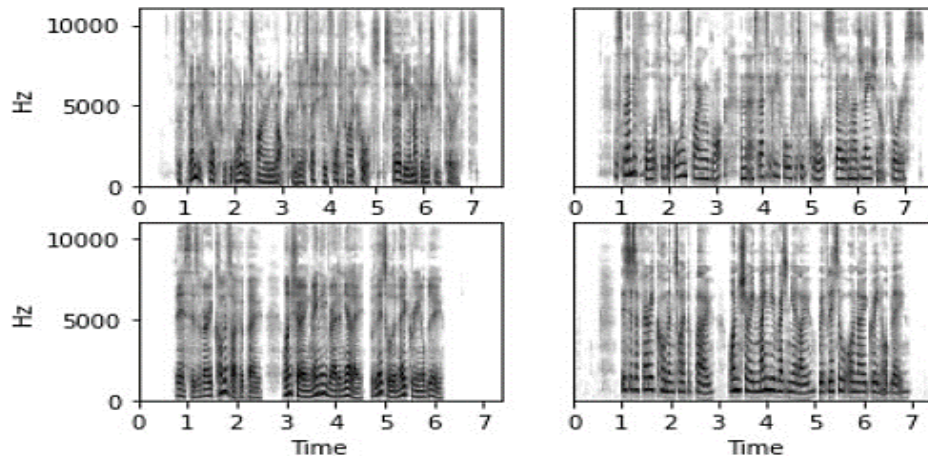


Fig. 3. Visualizing male (images on the left) and female (images on the right) Mel-Spectrograms

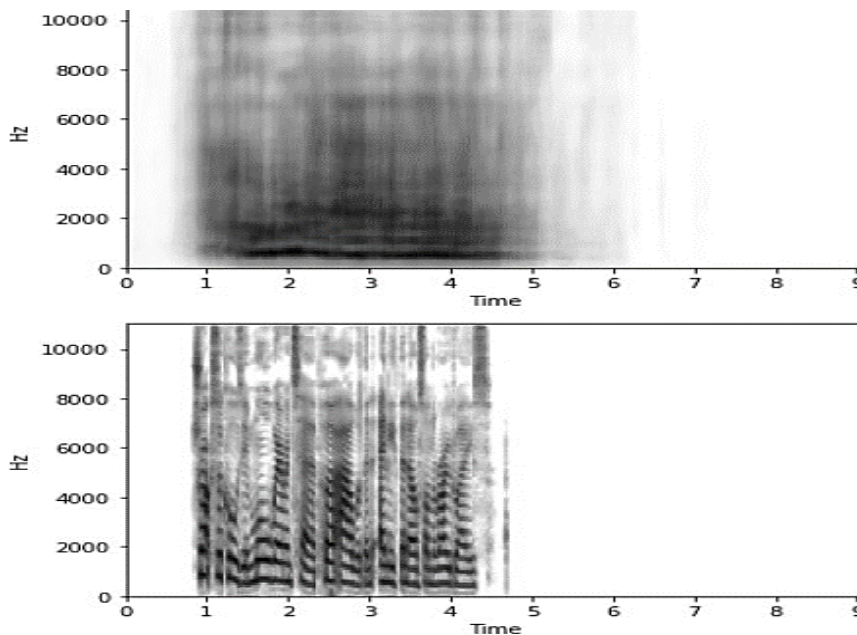


Fig. 4. Output using pixel-wise loss (top) compared to input female spectrogram (bottom)

4. IMPLEMENTATION

Initially a simple auto encoder was implemented which takes the male spectrogram as input and represents the same as latent vector of size 1024.

1. Encoder

A simple convolutional encoder was implemented to reduce the dimensionality of the input spectrogram. This has been done by have repeating blocks of convolution layer, batch normalization layer and maxpooling layer. The dimensionality of the input spectrogram (128x391x1) is reduced to (13x46x128). This is then flattened and connected to a dense layer of size 1024.

2. Decoder

The decoder does the converse operation of the encoder, by getting the latent function and generating the spectrogram matrix. This is done by repeated blocks of

transpose convolution layer and up scaling layer. The final output of this network is of the size (128x391x1) which is the same as the input spectrograms. The above discussed architecture was trained over a dataset of twenty thousand spectrograms with binary cross-entropy as the loss function, as the predicted and input values range between 0 and 1. The notable issue with such a model however is that the loss function takes into account only the difference in corresponding pixels and not the nature of spectrogram as a whole.

3. Difference in spectrograms based on speaker

Mel-spectrogram is a graphical representation of the particular audio file and therefore will vary based on the nature of an individual voice, making it nearly unique for each person, even when they are reading out the same sentence. Therefore when one audio file in male voice corresponds to multiple audio files in female voice, the pixel-wise cross entropy function will try to update the

weights such that the average loss for all the entries is minimal. However, the objective of the architecture was to transfer the male audio style to female and therefore minimizing the loss for each individual entry, such that the output is the mean of all female audio entries of the given sentence is not valid. Fig. 4 gives the output using pixel-wise loss compared to female spectrogram. The loss function should be able to determine if the generated spectrogram seems female or not and whether the content of the generated spectrogram matches with input spectrogram.

5. AUDIO STYLE TRANSFER USING NON-PIXEL-WISE LOSS

As pixel-wise loss based architecture does not satisfy the problem in hand, a custom loss function was designed to satisfy the important requirements for the given problem. The conditions for output to be realistic are:

1. Female-like spectrogram

The generated spectrogram should be similar to the spectrograms generated for a female English audio file. To satisfy this particular condition, a classifier network was implemented and trained such that for a given audio spectrogram the classifier determines if it is a female-like spectrogram or not. The output of this network is a value between 0 and 1, where zero corresponds to female-like spectrogram and one corresponds to non-female spectrogram.

2. Content preservation

It must be noted that the content of the input and output spectrograms should be the same, this means the sentence spoken by the user should be comprehensible in the output. For this the network must be able to understand how similar the input and output are in terms of the content. The content in spectrogram of English sentences can be described as the patterns in the same. Therefore a Siamese network was implemented which takes in both the input and output of the generator and determines whether they belong to the same sentence or not. The output for this network also is value between 0 and 1, where zero means that the spectrograms belong to the same sentence, and therefore have the same content. These two networks were then used to design a custom loss function to train the previously explained generator architecture. The loss is determined by the Equation (2), where α is a constant ranging from 0 to 1, L is the calculated loss, X and Y are input and output respectively. Thus the loss function will also be ranging from 0 to 1, with zero meaning that the output is valid and one meaning that the output is not valid.

$$L = \alpha * (\text{classifier}(Y)) + (1 - \alpha) * (\text{Siamese}(X, Y)) \quad (2)$$

The classifier and Siamese networks were trained to an accuracy of 99% and 94% respectively using the dataset in hand. Fig. 5 represents the learning curve of the Siamese network over the first fifty epoch with a learning rate of 0.00001 and binary cross entropy considered as the loss function. Followed by which various values of α was tested to identify the best value, such that output of generator seems genuine.

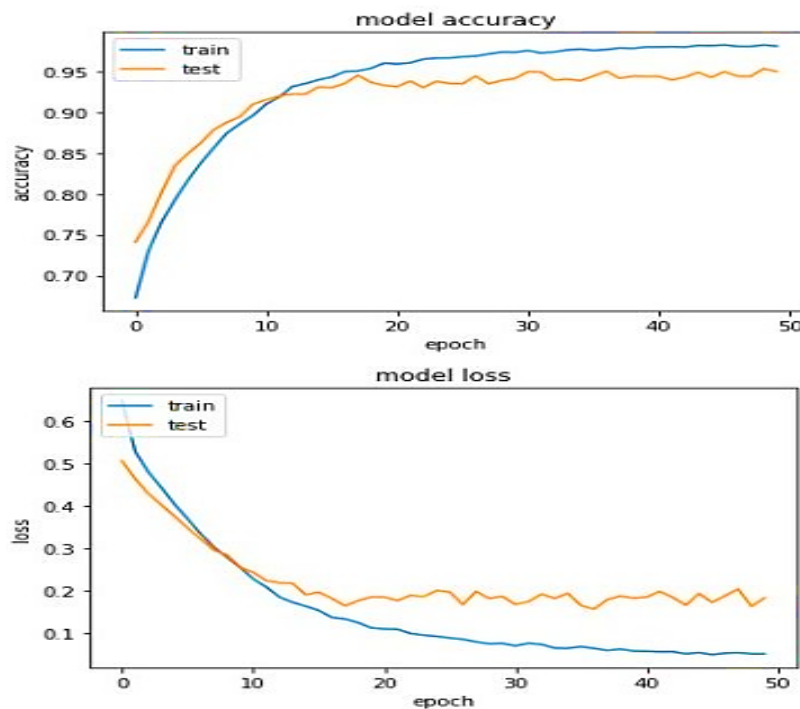


Fig. 5. Training of Siamese network over 50 epochs

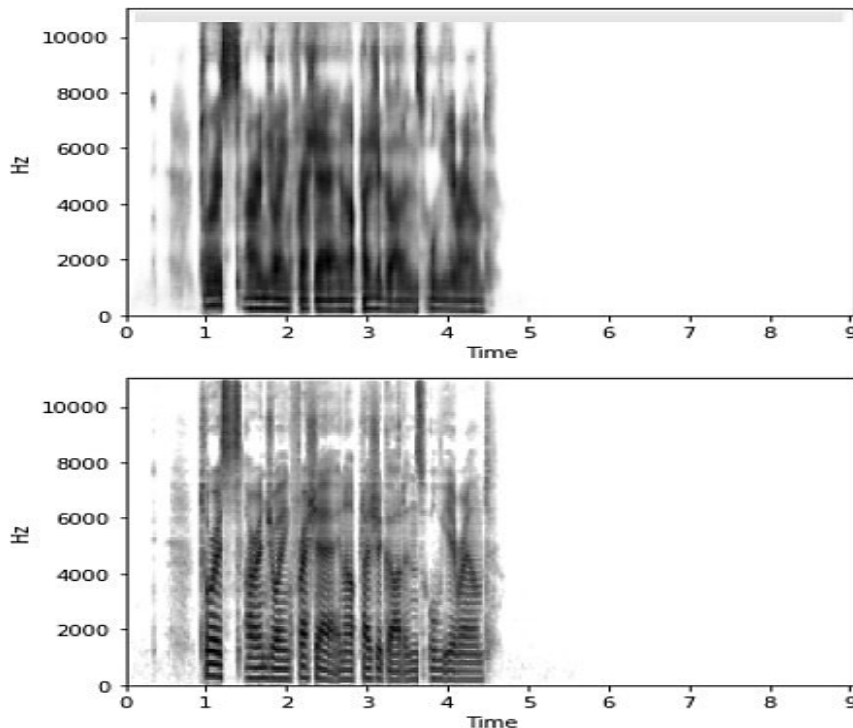


Fig. 6. Output using custom loss function (top) with $\alpha = 0.75$ on given input (bottom)

The output spectrogram when training with α as 0.5, was able to identify certain aspects of female-voice spectrogram. However there was still a significant amount of noise and the frequency bands were not well formed. As the Siamese network deals with content preservation in this particular architecture, the α value is varied to 0.75 and the generator was trained again. Fig. 6 shows the output of the generator on training with $\alpha = 0.75$. As we can see the frequency bands are more prominent and the content is similar to the original spectrogram provided. However the generated image still lacks clarity and therefore makes the corresponding audio file illegible, this is because the Siamese network and the classifier are not being updated to account for the fake entries generated. To tackle this problem a generative adversarial network was implemented where the discriminator takes the classifier and Siamese network architecture and updates their weights to account for the fake spectrograms.

6. GAN IMPLEMENTATION FOR AUDIO TRANSFER

The generative adversarial network implemented uses the same generator architecture as discussed in the previous sections and has the Siamese and classifier network running in parallel followed by a fully connected network as the discriminator. The implemented architecture is a combination of below explained networks:

1. Generator

A convolution based generator takes the input of a male audio spectrogram and generates the corresponding female

spectrogram. This is being done by reducing dimensionality of the spectrogram using convolution layers and representing the input as a latent vector. This latent vector is scaled up using transpose convolution layers. The output dimension of the generator is the same as the input. This generated input is then passed into a discriminator architecture that has been discussed below.

2. Classifier

The classifier architecture, like in the previous experiment, is a convolution layer which is used to determine whether the image that is being generated by the generator is similar to that of the spectrograms of a female English speaker. The classifier network uses convolution, batch-normalization and Max pooling layers, to extract the features from the image and the output is represented as a vector of dimension 256×1 .

3. Siamese network

The Siamese network is similar to the one present in the previous experiment and is intended to determine whether the input male audio spectrogram and the generated female audio spectrogram are similar to each other or not. The similarity check essentially is to determine whether the audio content, such as words spoken and the order of them is preserved while performing the style transfer. The output of the Siamese network is vector of dimension 256×1 denoting the similarity between the input and output.

4. Discriminator

The discriminator network uses the above mentioned Siamese and classifier networks in parallel. The output of these two networks is concatenated to get vector of size 512. This is then passed through a fully connected network to

determine whether the provided image to the discriminator is real or fake. The final layer consists of only one node with sigmoid activation. If the output of discriminator network is 1 then the spectrogram provided to it has been classified as fake whereas if the output is 0 then the spectrogram has been classified as real. The generator and discriminator networks were trained in parallel, where for a particular batch; the generator is used to develop the female spectrograms. These are then passed on to the discriminator along with the real spectrograms of the female audio files. The discriminator is now trained to predict zeros for the real files, whereas ones for the generated files. The generator is then trained based on the output of the discriminator of the generated images. Let us consider a male spectrogram X and corresponding female spectrogram Y. The generated image for X is then Y'. The discriminator now passes X and Y' through the Siamese network while only Y' goes through the classifier network. The output of these two networks combined, go through a fully connected network which generates a final output between 0 and 1. The output of the discriminator and the ideal value 1 is used to calculate the loss, for which binary cross entropy was used. This is now back propagated through the discriminator and the weights are updated. The same process is done with real female spectrogram Y and X, with the expected output being zero as the spectrogram is real.

7. RESULTS

1. Pixel-wise loss based auto encoder

In this approach, simple auto encoder architecture was implemented and tested, which generates the output for a given male-like spectrograms. The loss function used here was binary cross entropy on each pixel. The auto encoder however was not able to generate legible spectrograms as it tried to minimize the loss function such that the generated spectrogram is the relative average of all the female audio spectrogram associated with the given input spectrogram. This generated a blurred spectrogram without any legible structure to it.

2. Custom loss based auto encoder

In this approach, the auto encoder was trained using a custom loss which determined whether the generated spectrogram has the same content as the input and if it is female-like spectrogram or not. This custom loss was implemented using a Siamese network and a classifier network. While the results of this spectrogram were better than that of the pixel-wise loss, it still failed to generate accurate spectrograms, as the custom loss functions was not getting updated to determine whether the output is real or fake.

3. GAN based approach

In the GAN based approach, the Siamese network and the classifier network were used to develop a discriminator architecture which determines whether the generated image is real or fake. The output of the discriminator was trained using binary cross entropy loss. On training this architecture, we can see the generated spectrogram is more realistic and on converting the same to audio file, we can see that the contents are preserved and are legible while the speaker sounds higher pitched than in the original. We can see in the Fig. 7, the output of GAN architecture proposed in the paper is able to generate the corresponding female spectrogram with greater efficiency than the previous method such as pixel-wise loss based auto encoders and also the custom loss method that was discussed. In the output represented in Fig. 7, the frequency bands are more prominent while we can see that the contents of the male and female spectrogram are very similar. Furthermore, the higher frequencies are prominent giving the generated audio file higher pitch. On converting the generated spectrogram to audio and applying griffin lim algorithm, we can see the audio file is legible and sounds female with respect to the audio file given as input to the generator. This shows the successful working of the GAN architecture proposed in this paper.

8. CONCLUSION

In this paper, we have proposed a model to generate female-like spectrograms given a male-like spectrogram. Three different approaches have been discussed, starting with simple pixel-wise loss, followed by custom loss function and finally a GAN based approach. While the experiment and testing was done over a dataset of female and male British audio files, the proposed architecture can be used to convert any two set of audio files from one to another. This is because the architecture proposed is a generic one. The proposed architecture can be trained over a dataset of audio files spoken by two people, to transfer the style of speaking of one person to another. However, some of the notable flaws with the proposed model include that the generated spectrograms do not vary in terms of pace at which the words are spoken. The second at which a word is spoken in the input and the generated audio file are generally the same. There the proposed model will not be successful when the style of talking has to be varied by speed. The proposed architecture can be extended to transfer the style of talking among two individuals; however it must be noted that such application can be used for unethical activities. Furthermore the present model can be improved further by identifying ideal hyper parameter settings and training it over a larger dataset, containing audio of various accents.

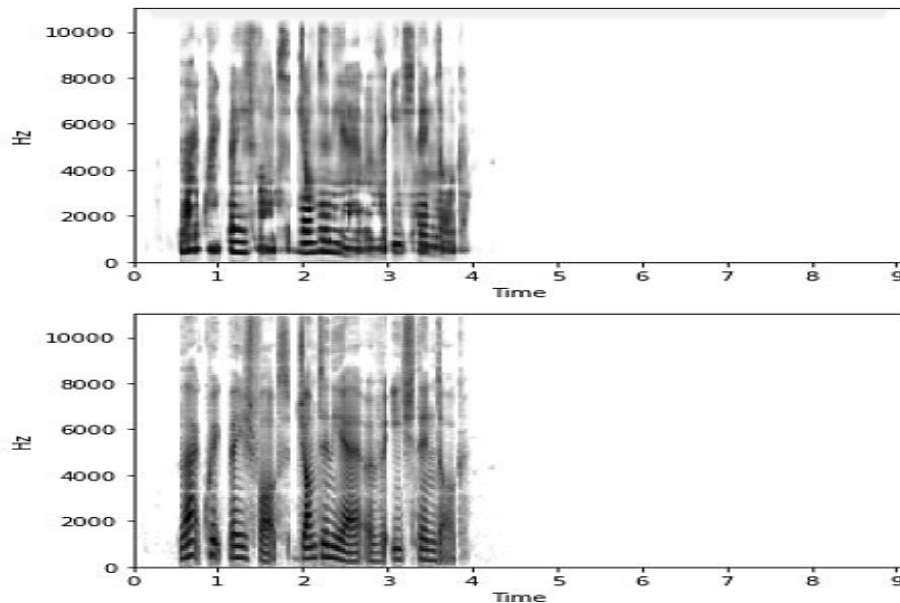


Fig. 7. Output spectrogram of GAN (top) for given input (bottom)

REFERENCES

- Chen, J., Yang, G., Zhao, H., Ramasamy, M. 2020. Audio style transfer using shallow convolutional networks and random filters, *Multimedia Tools and Applications*. Doi :10.1007/s11042-020-08798-6.
- Demirsahin, I., Kjartansson, O., Gutkin, A., Rivera, C. 2020. Open-source Multi-speaker corpora of the English accents in the British Isles, Vol. Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 6532–6541.
- Deshpande, M.S., Chadha, V.S., Lin, V. 2019. Audio style transfer for accents. URL https://shuby.de/files/11-785_project.pdf
- Grinstein, E., Duong, N.Q.K., Ozerov, A. 2018. P. Perez, Audio style transfer, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/icassp.2018.8461711.
- Hayashi, T., Tamamori, A., Kobayashi, K., Takeda, K., Toda, T. 2017. An investigation of multi-speaker training for wavenet vocoder, in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 712–718. <http://dx.doi.org/10.1109/ICASSP.2018.8461711>
- Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.M. 2017. Voice conversion from unaligned corpora using variational auto encoding Wasserstein generative adversarial networks. 3364–3368. 10.21437/Interspeech.2017-63.
- Huang, C., Lin, Y.Y., Lee, H., Lee, L. 2020. Defending your voice: Adversarial attack on voice conversion, 2021 IEEE Spoken Language Technology Workshop (SLT), 552–559, doi: 10.1109/SLT48900.2021.9383529.
- Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N. 2018. StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks, 2018 IEEE Spoken Language Technology Workshop (SLT), 266–273, doi: 10.1109/SLT.2018.8639535.
- Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., Ling, Z. 2018. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods, 195–202. 10.21437/Odyssey.2018-28.
- Miyoshi, H., Saito, Y., Takamichi, S., Saruwatari, H. 2017. Voice conversion using sequence-to-sequence learning of context posterior probabilities, preprint arXiv: 1704.02360.
- Pasini, M. 2019. MelGAN-VC: Voice conversion and audio style transfer on arbitrarily long samples using Spectrograms. <https://arxiv.org/abs/1910.03713>
- Sisman, B., Yamagishi, J., King, S., Li, H. 2020. An overview of voice conversion and its challenges: from statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 29. 10.1109/TASLP.2020.3038524.
- Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K., Toda, T. 2017. Speaker-dependent wave net vocoder. 1118–1122. 10.21437/Interspeech.2017-314.
- Verma, P, Smith, J.O. 2018. Neural style Transfer for audio spectrograms, CoRR abs/1801.01589.
- Wester, M., Wu, Z., Yamagishi, J. 2016. Analysis of the voice conversion challenge 2016 evaluation results. 1637–1641. 10.21437/Interspeech.2016-1331.
- Wu, C.-W., Liu, J.-Y., Yang, Y.-H., Jang, J.-S.R. 2018. Singing style transfer using Cycle-consistent boundary equilibrium generative adversarial network. arXiv:1807.02254