

**Deep Learning**

***Song Conversion Using Speech-to-Speech Neuro-Style Transfer***

***Team 7***

*Arnav Gupta*

*Parv Bhatt*

**School of Graduate Professional Studies**

Data Analytics

DAAN 570 – Deep Learning

*Spring, 2022*

**Document Control**

**Work carried out by:**

| **Name** | **Email Address** | **Exhaustive list of Tasks** |
| --- | --- | --- |
| Arnav Gupta | arg5956@psu.edu | Audio to spectrogram and style transfer of audio. |
| Parv Bhatt | pnb5078@psu.edu | Voice Isolation, Spectrogram to original audio and overlaying of audio. |
|  |  |  |

**TABLE OF CONTENTS**

[**1**](#_heading=h.3znysh7) **Introduction 3**

[**2**](#_heading=h.2et92p0) **Problem Statement 3**

[2.1](#_heading=h.tyjcwt) Challenges 3

[2.2](#_heading=h.3dy6vkm) Related Works 3

[2.3](#_heading=h.1t3h5sf) Importance 3

[**3**](#_heading=h.4d34og8) **Data Collection 3**

[**4**](#_heading=h.2s8eyo1) **Data Preprocessing 3**

[**5**](#_heading=h.17dp8vu) **Methodology 4**

[**6**](#_heading=h.3rdcrjn) **Results 4**

[**7**](#_heading=h.26in1rg) **Discussion of Results 4**

[**8**](#_heading=h.35nkun2) **References 5**

# Introduction

* What if we could go back to 1967, knock on The Beatles Studios' front door holding an 'All Access' badge, and have the privilege of listening to those signature harmonies A-Capella? Well, thanks to Convolution Neural Networks (CNNs), audio source separation is possible. What if we take this even further. What if we tell you that it's possible to clone your voice on a Beatles song or any song of your voice with just a 15-second voice sample. Neural Style transfer of images has been effectively applied by many models. It has successfully captured the style of an image which is defined by its colors and localized structures and applied it to a content image. Taking inference from this in this project we are going to take an input song, a target voice, and go to style the song using the target voice.

# Problem Statement

* As mentioned in the introduction we are going to style target audio to a content song using style transfer. For this, we are first going to split the song into two parts. Vocals and Music. After this, we are going to convert the vocals and target audio into spectrogram representation using the librosa library. These spectrograms are image representations of audio where the y axis is frequency scale and the x axis is time. The frequencies are also not plotted linearly but are done in a logarithmic way to capture the sounds effectively. After creating the spectrograms we style the vocals spectrograms using the target audio spectrogram using the Deep learning model. After this mixture of spectrograms is created we will convert it back into a wav audio file and then combine it with the music to produce our final output.

# Challenges

* Music can be represented in multiple ways - it can be read, listened to, or performed, and it all depends on whether we are relying on the score, sound, or control and no end-to-end system can deal with all levels of music representation together in an elegant manner.
* It's difficult to visualize sound. You can convert them into a spectrogram but on a spectrogram, discrete sound occurrences do not divide into layers; instead, they all add together to form a distinct whole. A specific measured frequency in a spectrogram cannot be believed to belong to a single sound because its amplitude could have been formed by any number of accumulated sounds or even by complex interactions between sound waves such as phase cancellation. In spectrogram representations, this makes it difficult to distinguish between simultaneous sounds.
* The two dimensions in spectrograms indicate fundamentally different units, one being frequency intensity and the other being time. As a result, the spatial invariance provided by 2D CNNs may not be as effective for this type of data.
* Additionally, when creating the model, style transfer is done by reducing the loss rather than comparing it with a test dataset as no datasets are available on songs to optimize model
* Another challenge is the conversion of a spectrogram back to the audio file as while converting spectrogram back to audio sound is distorted to quite an extent producing undesirable output.
* As two models are being used in the architecture, if there is an issue in even one of the models, it could create a hindrance and inept output.

## RELATED WORKS

* Introduce related projects and how your project and your solution differ from existing projects?
* There are various projects such as (Simpson et al., 2015)'s, (Jansson et al., 2017) are some of the researches that focus on voice Isolation, whereas (S. H. Mohammadi et al., 2015), (Srinivas Desai et al., 2015), (Lifa Sun et al., 2015) focus on voice cloning. The solution we have presented for this research is using both voice isolation separating vocals and music from songs and then cloning vocals using the voice transfer procedure and then finally joining both the audios.

**Papers on voice Isolation.**

* A convolutional deep neural network can differentiate vocal sounds from regular musical blends. (Simpson et al., 2015)'s convolutional DNN has roughly a billion parameters and was trained using a tiny amount of data (and relatively few iterations of SGD). In the context of a trade-off between artifact and separation quality, they compared this performance to a like-for-like (suitably scaled) NMF technique, indexed by confidence in the statistical predictions made. (Jansson et al., 2017) investigated the U-Net architecture in the context of singing voice separation and discovered that it outperforms the current state-of-the-art. Compared to ordinary convolutional encoder-decoders, the advantages of low-level skip connections were demonstrated. (Petermann et al., 2020) presented an adaption of the U-Net design in which some network parameters are conditioned on the base frequency contour of each of the SATB mixed sources. Preliminary findings indicated that using domain knowledge throughout the training process increased performance on both of their proposed use-cases. The oracle F0 was employed as external control input data to the network for the evaluation described in the paper.

**Papers on Voice cloning.**

* The research by (S. H. Mohammadi et al., 2015) tries to extend the initial DNN model’s capabilities by adding a Deep Autoencoder model so that it will result in pre-trained weights of the final DNN and also help the model effectively learn the speech spectral features. Also after using this architecture another important contribution by this research was forming of effective target speech with little deviations from the expected output with an input of only two sentences of the target audio being given as input.

Another research by (Srinivas Desai et al., 2015) extended RNN models by using a Bidirectional recurrent connection that can make full use of the context information in both forward and backward directions elegantly. The LSTM network architecture including memory blocks and peephole connections makes it possible to store information in linear memory cells over a longer period and to learn the optimal amount of contextual information for the task.

The research by (Lifa Sun et al., 2015) wanted to use parallel data but as it is not always present would generate two methods not involving parallel data. In the first method, a set of nonparallel utterances from the source and the target speakers are collected and a unit selection approach is employed to find corresponding parallel frames. In the second method, a voice conversion model is trained on preexisting parallel datasets and speaker adaptation techniques are used to adapt this voice conversion model to a particular pair of source and target speakers.

## IMPORTANCE AND IMPACTS

* The goal of this project is to reproduce a specific target singer’s voice, given input such as a musical score with lyrics. In this context, the ability to create a new voice from a small number of recordings, e.g., 2 min, is desirable for many use cases. There may, for instance, be cases where no new material can be obtained because the target singer is already deceased. Other cases might require modeling a large number of vocal timbres, where recording a full dataset for each voice would be impractical, such as for large choir synthesis.

# Data Collection

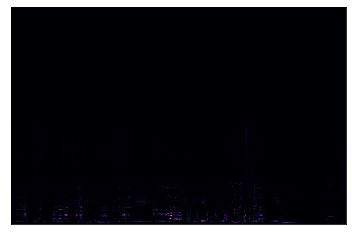
* For voice isolation, Spleeter library is used. There is no need for data.
* Style transfer is a type of abstract use case. You can train a model to copy a style. Instead, we tried to minimize the loss. Again no data is needed for style transfer.

# Data Preprocessing

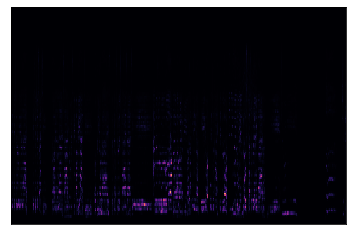
* After isolation of vocals, the next step was to convert them into spectrograms.
* We tried using different scales and methods like - Linear scale, Magnitude Phase Linear Scale, Amplitude to Decibel. However, the Logarithmic Scale to create the mel-spectrogram gave the best results.
* Logarithmic Scale generates less noise and is optimized for capturing the essence of style transfer.
* The morphed spectrogram generated at the end was effectively converted back to audio from the logarithmic scale as compared to other scales.

Displaying all the spectrograms.

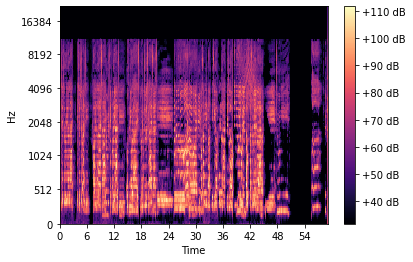
* Linear Scale



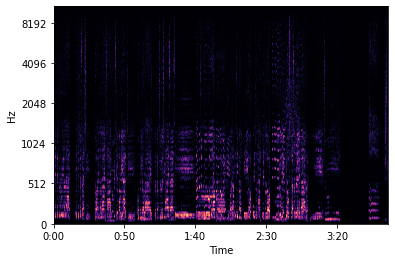
* Magnitude Phase Linear



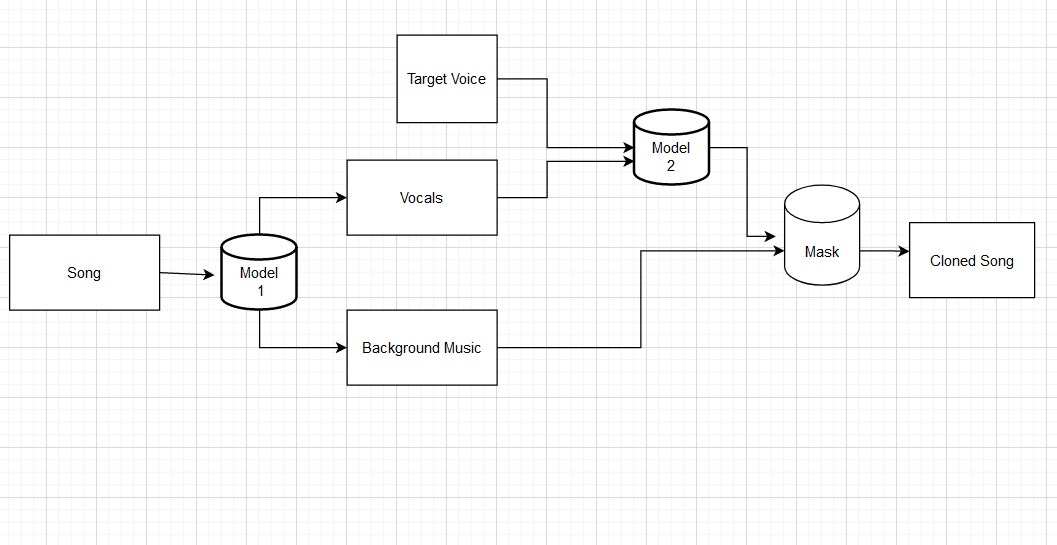
* Amplitude to Decibel



* Logarithmic Scale



# Methodology

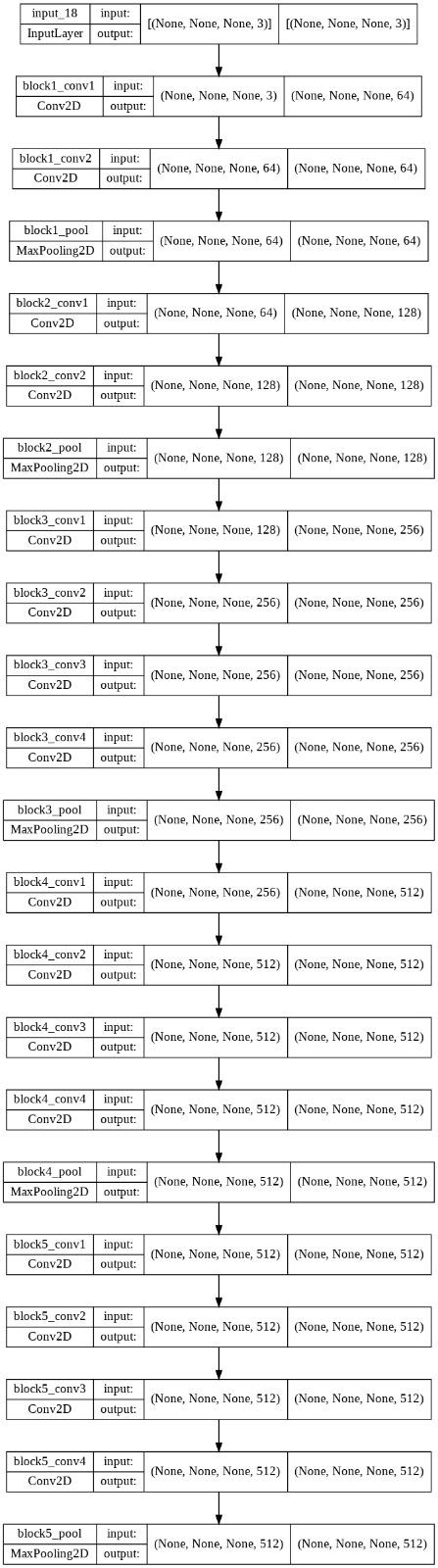


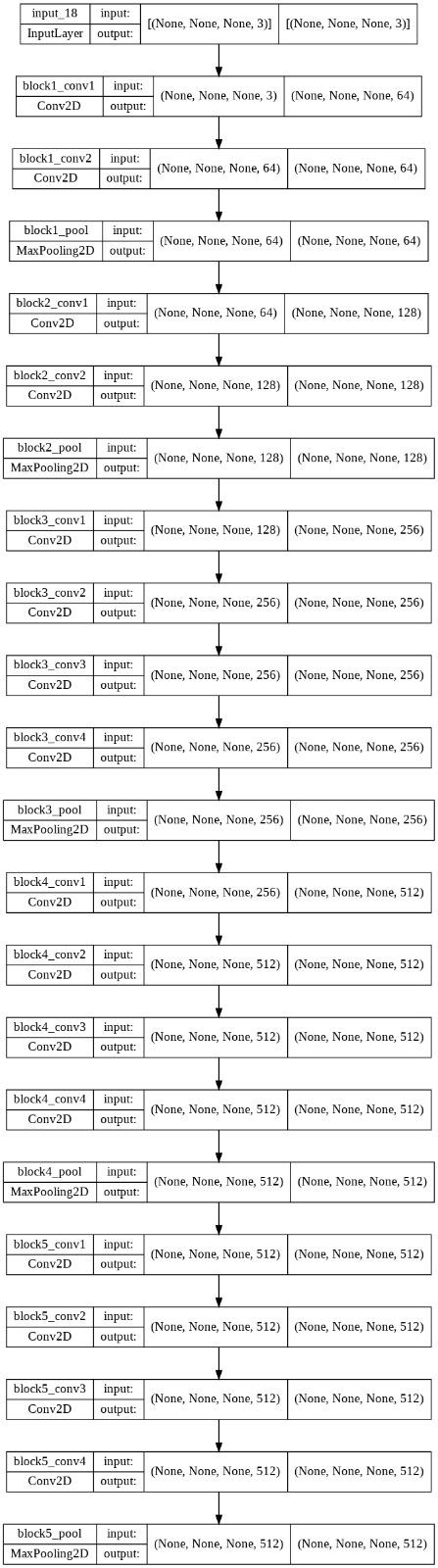
* We have used the spleeter model to separate the vocals and music from a song. To do voice cloning we have used a Convolution Neural network model
* We have chosen the spleeter model because it effectively works well with Audio separation and it was already being trained/tested on the MUSDB18 dataset. As we were going to use style transfer of target audio on content audio, we used a convolution neural network as it has already been used for different types of style transfer in images and we knew that it could thus be done effectively on Melspectograms.
* For the CNN model, we have created a single-layer model with a rectified linear unit activation function. Weights and biases are being randomly initialized in this model and as it is a single layer model the computation time isn’t going to be too much.

As we have used a spleeter model to isolate the song vocals and music we have used a pre-trained model and the validation of the model is being performed by listening to the audio samples being created.

* We have created another model for style transfer using transfer learning from VGG-19.

The model summary is shown below.





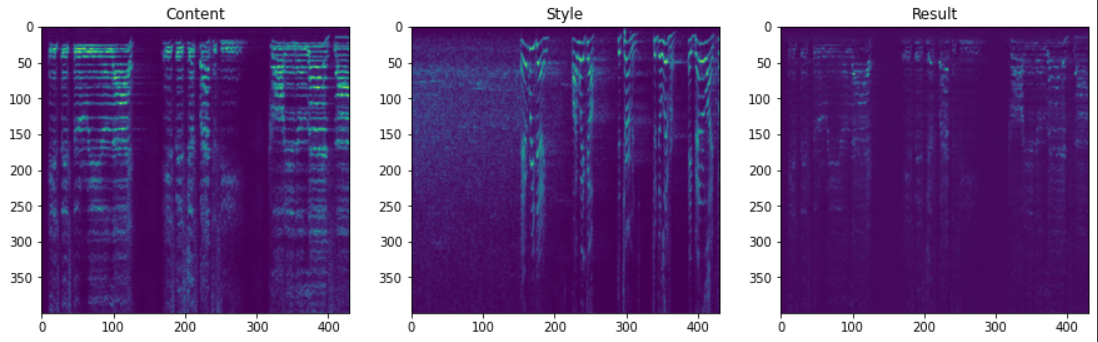
* The model has five blocks each consisting of Conv2D and Max Pooling layers.
* The model has a total of 20,024,384 trainable parameters.
* The Spleeter model is very effective in separating the vocals from the songs, whereas there is some sort of style of target audio being converged in the content audio but the results are not quite effective. Finally, after mixing both the results it does feel like there is some style being transferred.

# Results and INterpretation

* The outcome of the model for this problem will be determined by the loss function result of the skip-gram matrix. The total loss of the model as low as possible would mean the model is a very good model and has high performance while determining it for the style transfer model. If the loss function is high it will lead to a result of the model performing badly or not as effective as it could have been.
* The loss function for the model after 300 iterations was 709 which for style transfer is comparatively a good result but does also leave room for improvement that the loss of the overall style conversion model could do better to have a loss of around 100 which is observed in style transfer for images.
* The content of an image is represented by the values of the intermediate feature maps. The style of an image can be described by the means and correlations across the different feature maps. For this, we calculated the gram matrix.
* We also used cosine similarity to compare the output with the input spectrograms.

# Discussion of Results

* The result is not the most optimal result, but it is actually a good result which can be made more effective with more time being put in the model and using more layers and hyperparameter tuning.
* The spectrograms for content and style audio and similarity between the two is shown below.



* The output will be effective if for some reason a singer could not sing the song and hence using this model a song in his/her voice could be converted. Or if a singer would like to test which song will be best suited on him to perform for a concert then it the model could be used. Also if someone wants to try and sing a song but is not sure which one to use he can effectively use this model and then find which one of the songs will be his choice.
* The limitation of this project is the effectiveness of the output as the loss is around 700 and the model is not able to perfectly give the results of the target audio sample.
* Future work could be improving the voice cloning part using a better neural network model so that the target audio is effectively loaded and any human person could not identify whether this audio object is tampered or original.

# Your Feedback

* Using the Spleeter model and a Convolution Neural network model using a song and a target audio as an input we have successfully isolated vocals and music from a song, cloned the target audio on, the vocals and then after this successfully merged the music with the cloned target vocal output and the result is a song converted in the voice of a target audio.
* As mentioned if someone wants to try using this audio sample to test whether the song would look particularly good on his/her voice then it could be done or if a singer is unable to sing and wants to have a song in his voice then this model will be effective in such scenarios. A deep CNN architecture with more neural networks and loss function which could help build a good effective model would have some future scope and improvements which can be done on this project.

# References

* Simpson, A.J., Roma, G., & Plumbley, M.D. (2015). Deep Karaoke: Extracting Vocals from Musical Mixtures Using a Convolutional Deep Neural Network. LVA/ICA.
* Jansson, A., Humphrey, E.J., Montecchio, N., Bittner, R.M., Kumar, A., & Weyde, T. (2017). Singing Voice Separation with Deep U-Net Convolutional Networks. ISMIR.
* Voice conversion using deep neural networks with speaker-independent pre-training S. H. Mohammadi, A. Kain. [1]
* Spectral Mapping Using Artificial Neural Networks for Voice Conversion Srinivas Desai, Alan W. Black, B. Yegnanarayana, and Kishore Prahallad. [2]
* VOICE CONVERSION USING DEEP BIDIRECTIONAL LONG SHORT-TERM MEMORY BASED RECURRENT NEURAL NETWORKS by Lifa Sun, Shiyin Kang, Kun Li and Helen Meng. [3]
* <https://github.com/CorentinJ/Real-Time-Voice-Cloning>
* <https://github.com/coqui-ai/tts>
* <https://github.com/mazzzystar/randomCNN-voice-transfer>
* <https://github.com/ebadawy/voice_conversion>
* <https://www.youtube.com/watch?v=zbVQwqx-kYk>