Separation of Vocals from Audio Sample

Soham Rahul Inamdar
210100149
Electrical Engineering
IIT Bombay

Vighnesh Hareesh Nayak
210100169
Mechanical Engineering
IIT Bombay

Tanmay Ganguli
210100156
Mechanical Engineering
IIT Bombay

***Abstract*—Any piece of music produced generally has two major components vocals (the voice of the singer/s) and instrumentals (the sound produced by various instruments like drums, bass etc.). We have implemented the paper which uses U-Net architecture of neural networks to separate the vocal and instrumental components of a given audio file.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Music Information Retrieval (MIR) is a broad field , involving many aspects like melody and timber. One of the numerous topics is to separate the singing voice from a given melody. We have implemented the methodology followed by the cited paper to separate the vocal component from a given audio file. The audio clippings on which the voice separation is performed are taken from the MUSDB18 dataset. Next, a Short Time Fourier Transform (STFT) is applied on them to obtain a spectrogram. A neural network involving U-Net architecture is trained on it and finally the performance on training and test datasets is evaluated.

II. Approach

A. Dataset Used

The training and testing data used for the project is taken from the MUSDB18 dataset. It has over 150 full length music tracks of different genres along with their isolated 'drums', 'bass', 'vocals' and 'others' stems. It has two folders: a folder 'train' with over 100 songs, to be used for training ML models, and a folder 'test', with over 50 songs, to be used for testing the models. The 'drums','bass','vocals' and 'rest of the accompaniment' are stereo streams, and the 'mixture' is the sum of these four signals.

B. Preprocessing

The 'mixture' corresponding to each song in the MNDUB18 dataset is the melody for which we have to perform voice separation, and the 'vocals' , which contains only the singing voice, can be considered as the desired output of the entire process. The first step is to perform the Short Time Fourier Transform (STFT) of the given audio signal, that is 'mixture'. The STFT is defined as:

$$X(\tau,\omega)=\int_{-\infty}^{\infty}x(t)w(t-\tau)e^{-i\omega t}dt \tag{1}$$

for continuous-time systems and

$$X(m,\omega)=\sum_{m=-\infty}^{m=\infty}x[n]w[n-m]e^{-i\omega n} \tag{2}$$

for discrete-time systems. In the above equations, one must note the difference between ω, the angular frequency and $w(t-\tau)$, the window function. The window function is a function used to select values of $x(t)$ in an interval symmetric about the origin which means it will be non-zero over $(-a,a)$ and 0 at all other points on the real line. The default window function used by librosa is the **Hann Window**.

The STFT is a useful tool for analysing a time varying signal with respect to both frequency and time, unlike the Fourier Transform, using which we can only analyse a signal in frequency domain.

There are two audio channels corresponding to each stream in the MNDUB18 dataset. The output of the STFT on the 'mixture' signal is a spectrogram, which is a 2-D function in both time and frequency. The spectrogram is essentially a visualisation of the magnitude of the Short Time Fourier Transform(STFT) with respect to frequency and time. We get two spectrograms corresponding to the two audio channels. The processing of a spectrogram can be considered to be somewhat similar to processing a image.

C. U-Net Architecture

The U-Net architecture is followed for the purpose of vocal separation. It has components from a convolutional and a deconvolutional network. In the first half, which is the encoding stage, successive convolutional hidden layers are applied on the spectrogram, which extract the various features from the spectrogram. In the second half, the decoding stage, in successive stages, deconvolutions are performed and the decoded matrices are concatenated behind the encoded matrices at the same level and passed on for further decoding. An illustration of the U-Net architecture used by us is the following:

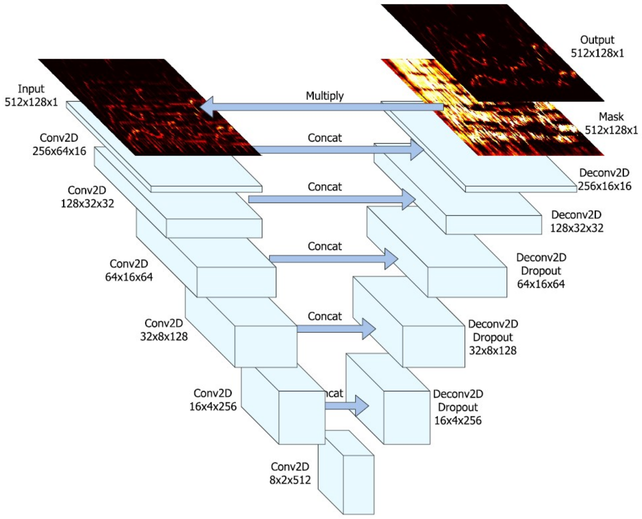

Fig. 1. U-Net Architecture

We have minimised Mean Squared Error loss in our model.

The final output after passing the spectrogram through the neural network is also a spectrogram corresponding to only the vocal component. Performing the Inverse Short Time Fourier Transform on this should yield the vocal part separated from the song, depending on the accuracy of the model.

D. *Observations and Results*

The results of the experiments conducted by us were quite unsatisfactory as we got a rather underwhelming accuracy of 23.30% on the training dataset as result of which we did not proceed further with the validation dataset. A few images of the spectrograms obtained by us are as follows:

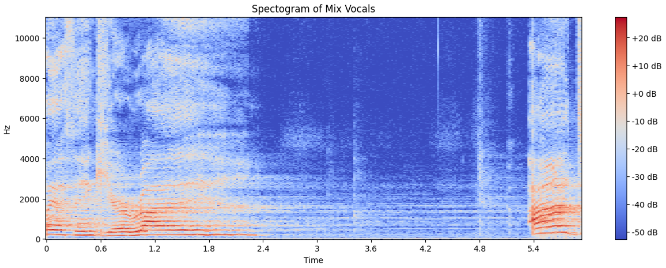

Fig. 2. U-Net Architecture

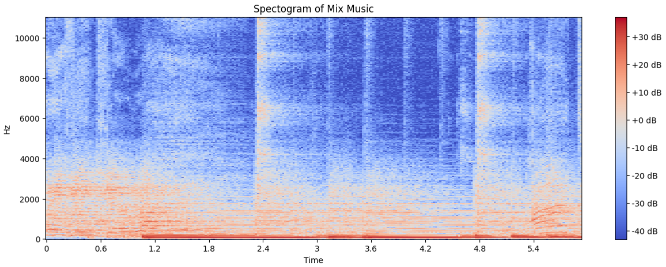

Fig. 3. U-Net Architecture

E. *References*

- Singing Voice Separation With Deep U-Net Convolutional Networks, by Andreas Jansson and co-authors https://drive.google.com/file/d/1ACfBSXTizIy6MaYWO1E-Ty3Tb6uryiQu/view?usp=share_link
- MUSDB18 Dataset https://zenodo.org/record/1117372/files/musdb18.zip?download=1