Separating Speech(Vocals) From Sound(Songs)

Group 15
Aman Agarwal and Jayesh Lahori
IIIT-Hyderabad

Abstract - Separating singing voice from music accompaniment is an important task in many applications such as music information retrieval, lyric recognition, singer identification and alignment, etc. This bears close resemblance to speech recognition but is quite a difficult and challenging task as compared to normal speech detection due to the presence of music accompaniment which are often non-stationary and harmonic. In this project we have proposed different techniques for vocal removal. It includes filtering frequency range of human voice(i.e. Bandpass filtering), cancellation of common frequencies between stereo channels(i.e. Stereo cancellation) and finally masking time frequency spectrogram (i.e audio blind source separation). In our case, we focused on extracting the vocals track from the mix consisting of the rest of the instruments. In this process to obtain the spectrogram of signals, we are going to take Short Time Fourier Transform (STFT) of signal.

I. INTRODUCTION

The project's aim was to design a method so as to separate speech from speech+music signal. In this project,we have analysed various techniques for separation of speech from speech+music..

II. ISSUES INVOLVED

While separating target voice from monaural mixture of different sound types seems effortless for humans, it is a very difficult task for the machines. It has been a major challenge since decades. Singing voice separation is somewhat a special case of speech recognition and also has similar applications. They perform substantially bad in presence of background noise or music accompaniment.Singing voice bears many similarities to speech. For example, they both consist of voiced and unvoiced sounds. But the differences between singing and speech are also significant. A well known difference is the presence of an additional formant, called the singing formant, in the frequency range of 2000-3000 Hz in operatic singing. This singing formant helps the voice of a singer to stand out from the accompaniment.From the sound separation point of view, the most important difference between singing and speech is the nature of other concurrent sounds. In a real acoustic environment, speech is usually contaminated by interference that can be harmonic or non-harmonic. Interference in most cases is independent of speech in the sense that the spectral contents of target speech and interference are uncorrelated. For recorded singing voice, however, it is almost always accompanied by musical instruments that in most cases are harmonic, broadband, and are correlated with singing since they are composed to be a coherent whole with the singing voice. This means separation of singing voice from music accompaniment is more difficult than speech separation.

III. PROPOSED METHODS

A. BANDPASS FILTERING & PREDOMINANT PITCH DETECTION

We have analysed a simple model. So, we have tried to separate speech from speech + music using a band-pass filter which will pass the range of frequency under which human voice falls and rejecting others. The human voice has a distinct frequency range between 300Hz and 3Khz.As there are instruments also , whose frequency lies between this range, they get passed by the filter and thus we do not get a pure vocals. But the vocals component of song is much higher in the result.

Since , This does not gives an clear output more advanced methods like predominant pitch detection to detect the pitch contour of singing voice for vocal portions. The algorithm first decomposes a vocal portion into its frequency components with a 128- channel gammatone filterbank. A normalized correlogram is then computed for each channel and each frame to obtain periodicity The peaks in the normalized correlogram contain the periodicity information of the input. However, due to the presence of music accompaniment, some peaks may give misleading information. To alleviate the problem, channel and peak selection are applied to all channels to extract reliable periodicity information. The algorithm uses **Hidden Markov Model (HMM)** to describe the pitch generation process. In each frame the observation probability of a pitch hypothesis is calculated by integrating the periodicity information across all frequency channels. The transition probability between two consecutive frames is determined by training. In order to reduce the interference of other harmonic sounds from accompaniment, the HMM tracks up to 2 predominant pitch contours simultaneously. Finally the Viterbi algorithm is used to find the most likely sequence of pitch hypotheses and the first pitch contour of this optimal sequence is considered as the pitch contour of the singing voice.

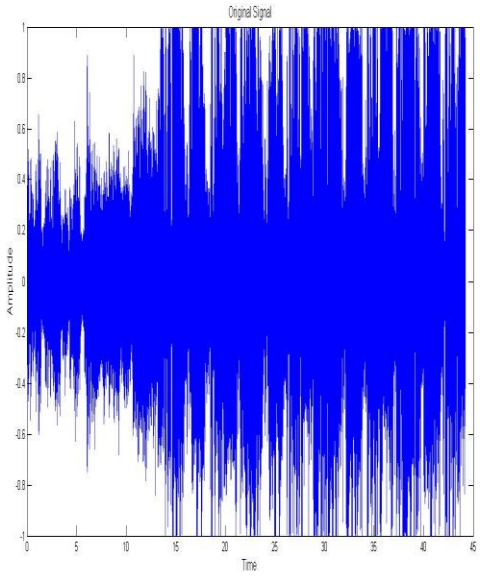

Fig:1-Y-T graph of original signal

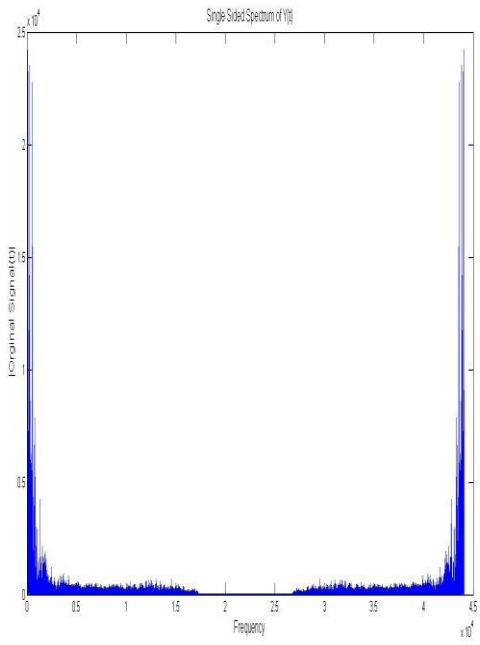

Fig:2-FFT of original Signal

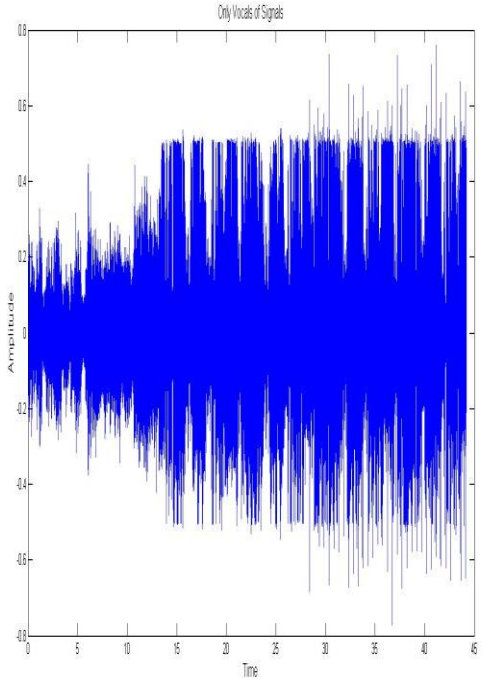

Fig:3-Signal after Band-Pass Filering

B. STEREO CANCELLATION

Stereo cancellation requires stereo tracks as name suggests. This technique involves cancellation of common frequencies from left and right channels. As voice is generally distributed symmetrically in left and right channel (center-panned) this technique works most of the times. As we are subtracting both channels result of this technique is mono track. The result is that since vocals are spread across both channels symmetrically, so vocals part gets subtracted and only music is left, and after subtracting this from the original signals we are left only with the vocals. This technique gives better results than band-pass filtering.

C. BLIND AUDIO SOURCE SEPARATION

This technique consists of "extracting from an input audio signal a set of audio signals whose mix is perceived similarly to the original audio signal ". In our case, we focused on extracting the vocals track from the mix consisting of the rest of the instruments.
Blind Audio Source Separation (BASS) algorithms attempt to recover one or more sources from a given observation mixture (music track) without prior knowledge or learning of the constituent individual sources. There are many ways to classify the mixture based on the sources that comprise it: broadly as music and speech

The BASS problem can be formulated in its simplest form as follows:

$$\begin{pmatrix} out_l[k] \\ out_r[k] \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ a_{m1} & \dots & a_{mn} \end{pmatrix} * \begin{pmatrix} input_1[k] \\ . \\ . \\ input_n[k] \end{pmatrix}$$

A set of unknown source signals that are mutually independent of each other can be considered and denoted by input1[k], input[2],… which form the source vector 'input'. With respect to audio, these signals would be the various auditory streams from the musical instruments in a music piece. These signals are recorded using sensors and are then linearly mixed using an unknown matrix of mixing filters A to form a music track. Following figure illustrates a basic form of the Blind Audio source Separation problem.

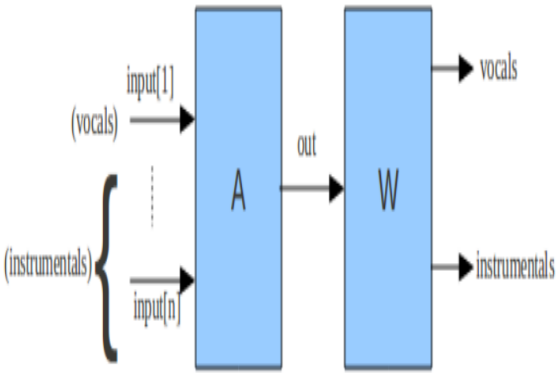

Fig:4- BASS Systems

Our job is to find A but as it's not known a-priory we going to estimate it as W, in our case we just want vocals we so do not want to know whole W.

Methods Applied for it:

Short Time Fourier Transform(STFT)

We, Basically divide the whole song into separate overlapping frames and take a STFT of it, as in the previous case we took the FFT of the whole signal without dividing into frames and it didn't gave a proper overview of the frequency characteristics of the whole signal , but here we have plotted a spectrogram which gives an idea of amplitude of components of different frequency components at each time, which gives a complete overview of the amplitude of various frequency components at each unit time.
Advanced Methods can be applied on the result of this spectrogram to further extract only the frequency components of the song which corresponds to only the human voice frequency range.
The steps taken to get the Vocals back from the result of above procedures are as follows:

- ***Vocal Track Identification***
- ***Time-Frequency Binary Masking***
- ***Getting Track Back Without Vocals(ISTFT)***

After Binary-masking we take Inverse Short-Time Fourier Transform to get back the vocals.

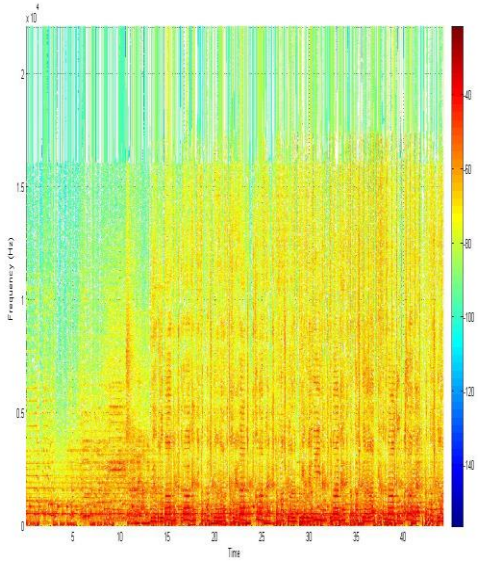

Fig:4-Spectrogram of the original Signal

IV. REFRENCES

- *Singing Voice Separation from Monaural Recordings - Yipeng Li & DeLiang Wang*
- *A Real Time Singing Voice Removal System Using DSP and Multichannel Audio Interface - Hyuntae Kim & Taehoon Kim*
- *Separating a foreground singer from Background Music- Rita Singh*