* + - 1. ***Project Overview:***
  1. ***Context:***

The field of Music Information Retrieval (MIR)[1] concerns itself, among other things, with the analysis of music in its many facets, such as melody, timbre or rhythm. Singing is used to produces musically relevant sounds by the human voice, and it is employed in most cultures for entertainment or self-expression. The singing voice becomes immediately the main focus of attention when we listen to musical pieces with a voice part. Now a days, in multimedia technology various audio editor software’s are available. Mixture of singing voice and music accompaniment known as a song. Music recording are either monaural (single channel) or stereo (two channel) basis. Speech is an acoustic signal produced from a speech production system. Sound is a representation of an audio signal. 20 Hz to 20 kHz are the audio frequency range. The human auditory system has a better capability in separating sounds from different sources.

Speech separation is a very challenging task in signal processing. An Audio signal classification system detecting the audio type of a signal (speech, background noise and musical genres). A singing voice separation system has its applications in areas such as automatic lyrics recognition and alignment, singer identification, musical information retrieval, karaoke, musical genre classification, melody extraction, audio signal classification, ect. As for commercial applications, it is evident that the karaoke industry, estimated to be worth billions of dollars globally, would directly benefit from such technology.

* 1. ***Purpose:***

In this project, I focus on the purpose of isolating the singing voice from the source contains the mixture of vocals and musical instruments sound, showed by the diagram below.

A screenshot of a cell phone

Description automatically generated

Figure 1 Diagram of Separation system.

As I mentioned on the Overview part, the system will isolate the singer’s vocal sound from provided music mixture, and return back the result of vocal-only file to users.

* 1. ***Project scope:***

1. ***Structure of the thesis:***

***INTRODUCTION -*** This chapter gives information about the context and purpose of the project as well as giving the scope of the problems which will be focused on the thesis.

***Chapter 1: THEORIES AND TECHNOLOGIES –*** This chapter introduces about all knowledge theories and technologies used in this project.

***Chapter 2: ANALYSIS AND DESIGN –*** This chapter covers the main features, software requirement specifications and database design of the project.

# THEORIES AND TECHNOLOGY

First of all, we have to understand about sound and signal representation, and some methods that I am using in this work.

* 1. ***Signal:***
     1. ***Signal representation in modern computer:***

The sound or signal data are saved in our devices as a waveform - a graph that shows a wave's change in displacement over time. A waveform's amplitude controls the wave's maximum displacement. Below is a sample of waveform visualization.

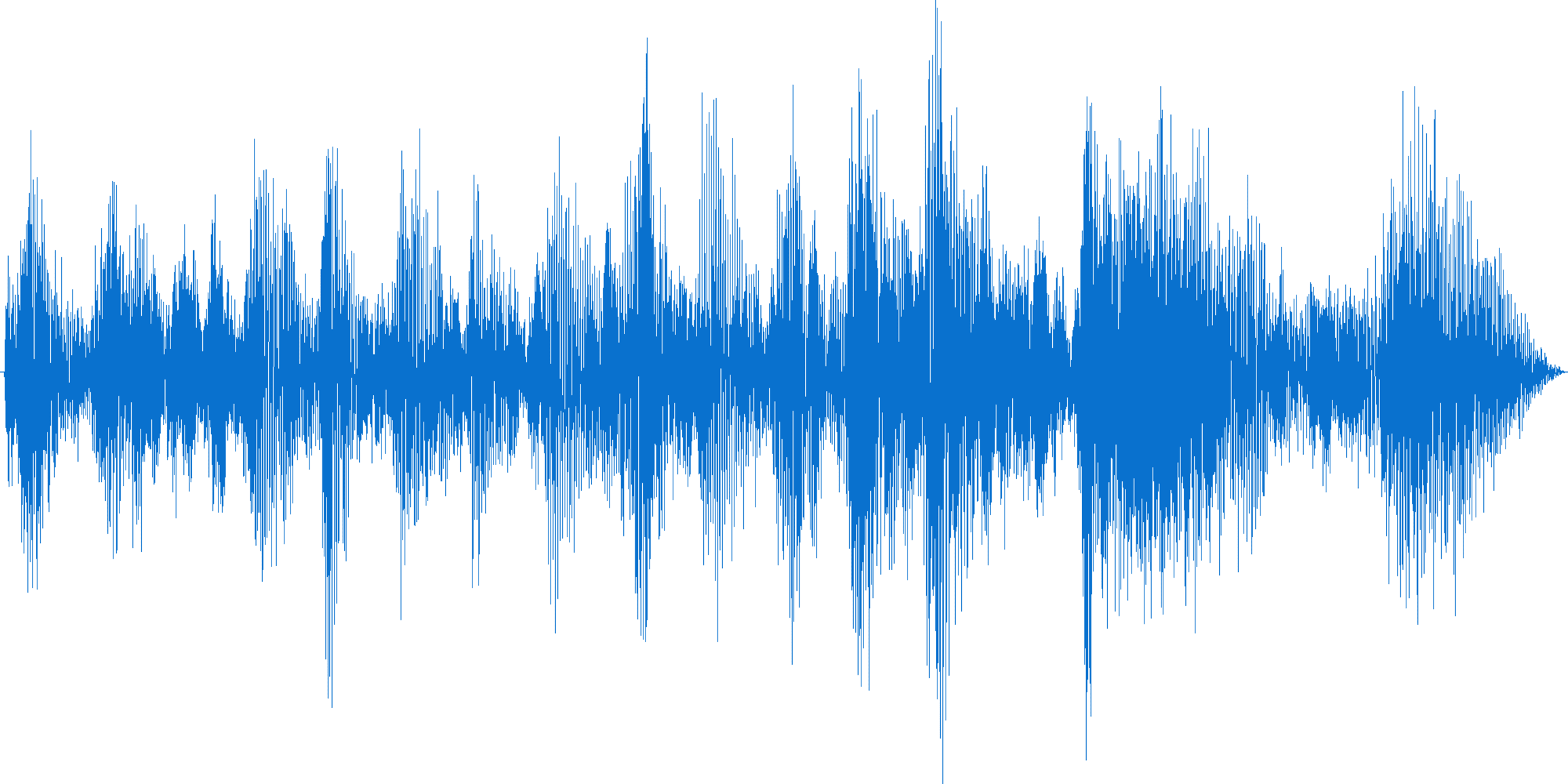


Figure 2: Waveform sample

Because the waveform or *time-domain signal* can just only provide us the access to the amplitude values of the signal over time. Which is not enough for analysing work!

We have to transform these representation to another one to get access to more value and features from the source file.

* + 1. ***Short-time Fourier transform (STFT):***

The short-time Fourier transform (STFT), is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. It defines a particularly useful class of *time-frequency distributions*, which specify complex amplitude versus time and frequency for any signal.

A close up of text on a white background

Description automatically generated

Figure 3: Fourier Transform (Picture from towardsdatascience.com)

The upper figure describe how our waveforms can be converted to Frequency Domain, therefore we can get Amplitude vs. Frequency of Audio clips. If we have the window size for time domain small enough, we can get more information about Time. That is the idea of STFT. The below figure will show more about the representation of the sound after do the STFT.

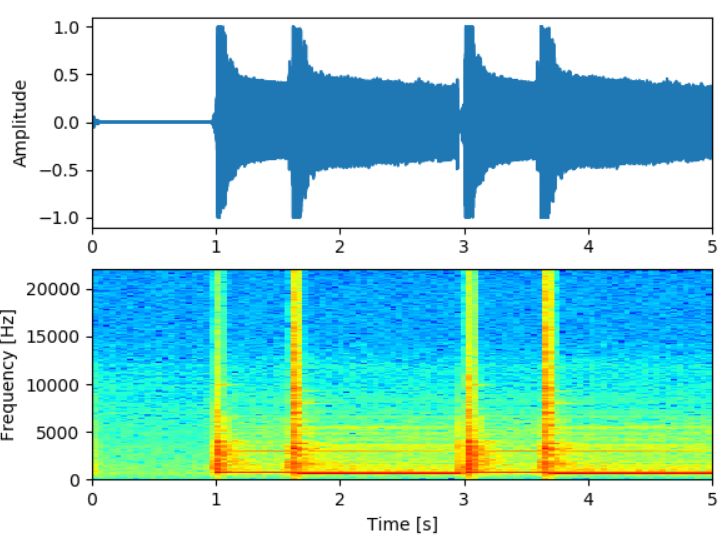


Figure 4: Waveform and STFT (picture from stackexchange.com)

The x-axis represents time, y-axis for Frequency and inside the picture, the warmer color, the higher amplitude of corresponding frequency is in the certain time. After doing STFT, we can have all information of Time, Frequency and Amplitude.

As we saw the output of

# References

[1] Haas, W. B., & Wiering, F. (2010). Hooked on Music Information Retrieval. http://cs.uu.nl/groups/mg/multimedia/publications/art/emr2010.pdf