

A PROJECT REPORT ON
Audio Source Separation

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY

IN THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF ENGINEERING
IN
INFORMATION TECHNOLOGY

BY

Ananth Narasimhan (71700040M)
Hiten Agarwal(71700229C)
Shirish Khairnar (71700302H)
Mubassir Patel (71700433D)

Under the guidance of
Prof. Varsha Naik



DEPARTMENT OF INFORMATION TECHNOLOGY
MAHARASHTRA INSTITUTE OF TECHNOLOGY, PUNE

2019-2020



MAHARASHTRA INSTITUTE OF TECHNOLOGY, PUNE

Department of Information Technology

CERTIFICATE

This is to certify that the Project entitled “**Audio Source Separation**”

Submitted by

Name Of Candidate

Ananth Narasimhan

Hiten Agarwal

Shirish Khairnar

Mubassir Patel

Exam Seat No:

71700040M

71700229C

71700302H

71700433D

Is record of bonafied work carried out by them under the supervision of **Prof. Varsha Naik** and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University for the award of the degree of Bachelor of Engineering in Information Technology.

This project report has not been earlier submitted to any other institution or University for the award of any degree or diploma.

Date: 02/06/2020

Place: Maharashtra Institute of Technology, Pune

Prof. Varsha Naik
(Internal Guide)

(External Examiner)

Prof. Sumedha Sirsikar
(Head, Dept. of IT)

Dr. L.K Kshirsagar
(Principal, MIT Pune)

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give to all professors whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this report.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of all the staff, who gave the permission to use all required equipment and the necessary materials to complete the project. A special thanks goes to my team mates, who helped me to overcome the problems faced in the process of completion of project. Last but not least, many thanks go to the guide of the project, Prof. Varsha Naik, who have invested his full effort in guiding the team in achieving the goal. I have to appreciate the guidance given by other supervisor as well as the panels especially in our project presentation that has improved our presentation skills thanks to their comment and advices.

Ananth Narasimhan	406032
Hiten Agarwal	406031
Shirish Khairnar	406055
Mubassir Patel	406067

B.E. (Information Technology)
MIT, Pune 411038

INDEX

Abstract.....	i
List of Figures.....	ii
01. Introduction	1
1.1 Introduction to Project	1
1.2 Relevance	1
1.3 Project Undertaken	1
1.3.1 Aim	2
1.3.2 Objectives	2
1.3.3 Motivation	2
1.3.4 Domain Area of Project	2
1.4 Organization of Report	3
02. Background	4
2.1 Literature Review	4
03. Specification	6
3.1 Problem Statement	6
3.2 Introduction	7
3.2.1 Purpose	7
3.2.2 Product Scope	7
3.3 Overall Description	7
3.3.1 Product Functions	7
3.3.2 User classes and Characteristics	8
3.3.3 Operating Environment	8
3.3.4 Design and Implementation Constraints	8
3.3.5 Assumptions and Dependencies	10
3.4 System Features	11
3.5 External Interface Requirements	11

3.5.1	User Interfaces	11
3.5.2	Hardware Interfaces	11
3.5.3	Software Interfaces	12
3.5.4	Communication Interfaces	12
3.6	Non- Functional Requirements	12
3.6.1	Performance Requirements	12
3.6.2	Safety and Security Requirements	12
3.6.3	Software Quality Attributes	13
3.7	Other Requirements	13
3.7.1	Legal Requirements	13
04.	Design	14
4.1	Architecture Design	14
4.2	High Level Project Design	15
4.2.1	Data Flow Diagram	15
4.2.2	Activity Diagram	18
4.2.3	Sequence Diagram	29
05.	System Implementation	20
5.1	System Architecture	20
5.2	Implementation and Deployment	24
06.	Evaluation and Results	25
6.1	Feasibility study	25
6.2	Risk Management	25
6.3	Experimental Setup	26
6.4	Testing Strategy	26
6.5	GUI and Results	27
07	Conclusion and Future Work	30
08	References	31

Abstract

Audio source separation is the method of separation of a set of source signals from a set of mixed signals, without the aid of information about the source signals or the mixing process. In this project we carry out blind audio source separation from a stereo sound signal. This audio source separation is carried using a pre-learned universal NMF dictionary, GCC-NMF operates in a frame-by-frame fashion by associating individual dictionary atoms to target speech or background interference based on their estimated time-delay of arrivals (TDOA). We evaluate GCC-NMF on two-channel mixtures of speech and real-world noise from the Signal Separation and Evaluation Campaign (SiSEC). We demonstrate that this approach generalizes to new speakers, acoustic environments, and recording setups from very little training data. We provide the input of a mixed audio signal of various speakers talking together and obtain the output of individual speakers' audio files. We have used the Generalized Cross Correlation over Basic Cross Correlation as it provides better output due to weighting function. A flexible, soft masking function in the space of NMF activation coefficients offers real-time control of the trade-off between interference suppression and target speaker fidelity. This helps us obtain a robust system to carry out audio source separation.

Index Terms—*unsupervised machine learning, speech enhancement, source separation, phase based, multi-channel, GCC, NMF.*

List of Figures

1	System Block Diagram	15
2	DFD Level 0	16
3	DFD Level 1	17
4	DFD Level 2	18
5	Activity Diagram	19
6	Sequence Diagram	20
7	Waterfall Plot	22
8	NMF Illustration	23

Chapter 1

INTRODUCTION

1.1 Introduction to Audio Source Separation

In this project we are tackling the famous “Cocktail Party Problem”. This problem states that in an audio or audio-video file there are a number of speakers talking together. Our goal is to separate the audio of each speaker from the mixture of audio of various speakers. There are many examples where we need to separate audio of individuals from mixed signals like meetings, news debates, social events recordings, etc. Humans are unable to do this task efficiently and correctly without interference from machines. So there was a need to do this task with the help of machines which can give us the result as per the expectations. Our proposed system successfully separates the audio of each person involved in conversation and displays it to the end in front end in minimum response time.

1.2 Relevance

The primary goal of audio source separation is to separate the audio of each person from a mixed audio which is not possible to perform by humans efficiently. The system will help to separate out the audio of each person involved in the conversation like meetings, news debates or social events, etc. correctly and efficiently and display them to the end user in the front end in minimum response time and maintenance.

1.3 Project Undertaken

Audio source separation using GCC-NMF methodology and Time difference of arrival (TDOA).

1.3.1 Aim

- The aim of the project is to develop a system that will help us to listen to each person individually from a audio file which contains multiple speakers speaking concurrently.

- To develop a low computation model which can also be used offline and work efficiently in terms of computation and time required to run the project.
- To develop a system in which the background noise is as less as possible of the individual speaker's audio file.

1.3.2 Objectives

- To develop an unsupervised model
- To develop a low computational model
- To achieve optimum audios of individual speakers

1.3.3 Motivation

- In current times the surveillance needs to be upgraded for the better security of citizens.
- Help catch criminals faster by understanding what people are saying.
- To help foil the plans of potential criminal activities.
- Develop a system which uses low computational requirements and hence can be used everywhere.

Hence arises a need to develop an audio source separation.

1.3.4 Domain Area of project

The main domain of the project is Machine learning.

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data.

1.4 Organization of Report

The project report is organized as follows

Chapter 1 gives an Introduction to the system 'Audio Source Separation' and the need for developing it to overcome the drawbacks of the existing systems in place. It also gives an insight into how the system shall overcome these drawbacks.

Chapter 2 gives a background about existing systems and previously done research on audio separation techniques...

Chapter 3 includes project specifications, a detailed description of objectives for development of the project. It lists goals, functionality, interfaces and other information that is required for successful completion of the project.

Chapter 4 includes system design where architecture, modules and data are defined. It includes detailed system design through the various diagrams.

Chapter 5 focuses on implementation of the system and how it can be deployed.

Chapter 6 focuses on evaluation of the project including assessing the relevance, effectiveness, efficiency, impacts and sustainability of the project and its activities. The various risks associated with the system are described and the ways to manage those risks.

The last part of the report gives an account as to how the use of the developed system will have an impact on the society and help overcome the drawbacks of the existing systems.

Chapter 2

BACKGROUND

➤ “STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement “

In this paper, they present a method to reconstruct the spectral phase of voiced speech from only the fundamental frequency and the noisy observation. They show that, when the noisy phase is enhanced using the proposed phase reconstruction, instrumental measures predict an increase of speech quality over a range of signal to noise ratios, even without explicit amplitude enhancement. Their research has mainly focused on the estimation of the clean speech spectral amplitudes from the noisy observation.

Advantage-The paper gives the analysis of sound using STFT and how we can reduce the noise considerably using STFT.

Disadvantage-The paper uses a single mic and has only one person speaking whereas in the cocktail party problem we have multiple people speaking together.

➤ The generalized cross-correlation Method for time delay estimation of infrasound signal

The work described herein discusses the application of generalized cross-correlation method to time delay estimation of infrasound signal. Before this paper, the basic cross-correlation technique was mainly used to estimate the time delay estimation of infrasound signals. However, the estimation accuracy based on this method is not high. In order to improve the accuracy, the generalized cross-correlation method is applied in this paper. The generalized cross-correlation method can be viewed as applying a weighting or a window function to the cross-power spectrum. The result shows that the estimation accuracy by the improved window function is high and the performance is stable under different signal-to-noise ratio.

Advantage- The result shows that the performance is stable under different signal-to-noise ratio. This paper helps us to understand Generalized Cross Correlation and how to assign

weights to cross-power platforms. We can also see how the Generalized Cross Correlation is better than Basic Cross Correlation.

Disadvantage-In this paper the work is carried out on infrasound signals. In the cocktail party problem, we have humans speaking and the frequency is different than that of infrasound. This paper helps us to know the localization of the source whereas in our problem we need to also separate the signal from the sources.

➤ **Analysis of the GCC-PHAT technique for multiple sources**

In this paper, they have derived the cross-correlation function by GCC method with the PHAT weighting function for multiple sources and obtained the relationship between the correlation value and source characteristics. They have also compared the GCC function they have obtained and the real GCC function calculated by actual signals.

Advantage-The paper gives an idea of how to calculate GCC values using PHAT. The method used to calculate GCC values is also accurate. It helps understand how to calculate GCC values.

Disadvantage-This paper does not have the separation of those audio signals.

➤ **Speech Enhancement using Non negative Matrix Factorization and Enhanced NMF**

This paper gives us a brief idea of Non-Negative Matrix Factorization. This paper also talks about the importance of Non negative matrix factorization and purpose of NMF in speech separation and enhancement. Various NMF techniques such as Enhanced NMF, BNMF-HMM and the NMF in speech enhancement systems used for musical source separation in a single channel speech enhancement.

Advantage-This paper describes various NMF techniques and also compares the performance of each NMF technique.

Disadvantage-The methodology explained in this paper can be used for speech enhancement and not blind speech separation.

Chapter 3

SPECIFICATIONS

3.1 Problem Statement

The essence of the cocktail party problem can be formulated as a deceptively simple question: “How do we recognize what one person is saying when others are speaking at the same time?” Finding answers to this question has been an important goal of human hearing research for several decades. At the root of the cocktail party problem is the fact that the human voices present in a noisy social setting often overlap in frequency and in time, and thus represent sources of direct acoustic interference and “energetic masking” that can impair the perception of speech. In addition, recent research has revealed that even those components of concurrent speech that do not overlap in frequency or time with those of the target signal can dramatically affect speech intelligibility via so-called “informational masking”. The ability of concurrent speech and speech-like noise to impair speech perception is well-documented in the literature on human hearing. Understanding the sensory solutions to the cocktail party problem has been a goal of research on human hearing and speech communication for several decades. In this project we are trying to separate the audio of each speaker from a mixture containing many people speaking together. In most of the cases we may not have the voice of each person talking so we need to find a way in which blind audio separation can be done. Also, the model needs to be computationally efficient so that this system can be run in most of the systems. This problem has been present since the 20th century to study human as well as animal behavior. There have been huge strides in audio analysis techniques. We made use of this audio analysis technology to carry out audio source separation.

3.2 Introduction

3.2.1 Purpose

Nowadays, we come across many situations like surveillance, meetings, debates or social event recordings where we need Audio Source Separation.

These meetings, debates or social event recordings contain multiple speeches overlapped on one another along with background noise due to which we are unable to understand a specific person's point at a time.

The purpose of 'Audio Source separation' is to develop a system that tries to precisely detect the people involved in conversation, separate their respective audio from mixed audio and reduce background noise.

3.2.1 Scope

The goal of the project is to separate the audio of the persons involved in the conversation having overlapped speeches. As the task of source separation cannot be performed by humans that much efficiently and accurately, there was a need to introduce the interference of machines to do this task.

As a result, we have introduced the system which can do this task to the required expectations. The system requires frameworks and libraries updated in a timely manner. The system needs an audio file which is in acceptable format. The system hopes to be practically implemented by any user with minimum response time and maintenance at server side. The system intends to help in

fulfilling the goal of any user concerned with the solution provided by our system

3.3 Description

3.3.1 Product Functions

The system 'Audio Source Separation' acquires an audio file, detects the number of speakers involved in the conversation, separates the audio of each individual and displays the respective files on the GUI accordingly.

3.3.2 User Classes and Characteristics

The system comprises two user classes as follows: -

1. Normal user.
2. System maintenance person.

The system is directly used by the people (Normal user) who want the solution proposed by our system. The normal user can interact with the proposed system with the help of simplified

GUI and instruction given in it with ease. The second user class include System maintenance person which is extremely important because the whole functioning of the software is dependent on the work of this user class, who is responsible for overall maintenance part of the system like updating of libraries and framework's versions if newer versions are available. If this user class fails in its work the whole system may crash. So proper working of this class is crucial.

3.3.3 Operating Environment

Operating System: Windows or Ubuntu

Platform: PyCharm

Language: Python, HTML, JavaScript, CSS, SQL

Framework: Flask, Bootstrap

3.3.4 Design & Implementation Constraints

A. Hardware Design Constraints

The proposed system is expected to function for a long-time span with minimal maintenance efforts. As the proposed system is entirely software based which needs minimal hardware for storing data generated as output. But if a vast number of user's use this application the data generated will be in large amounts which will be a challenge to store in the application. So there will be a need for additional storage devices to store the data and for efficient working of the system.

The storage devices can be locally available or can be virtually available such as cloud storage.

- **Cloud Storage**

Cloud storage is a storage option that utilizes remote servers and is accessible from any computer within internet access. It is kept up, worked and overseen by a cloud storage service provider on storage servers that are based on virtualization strategies. In case of shortage of storage devices, the application can be deployed on the cloud where a number of instances of

databases with required memory can be made available for applications which are charged dynamically based on the usage time efficiently.

- **Hard Disks**

Hard disks can be made locally available for minimal use for storage of data. Hard disk is a secondary storage device, which holds the data in bulk, and it holds the data on the magnetic medium of the disk. Hard disks have a hard pattern that holds the magnetic medium, the magnetic medium can be easily erased and rewritten. Data stored onto the disk is in the form of files.

B. Software Design Constraints

- **Operating system**

An operating system is required for communication between database & software programs acting as an interface and to run the software program. Windows or Ubuntu operating systems can be used for running the proposed system with proper installation of all required libraries.

- **Language Compiler/Interpreter**

A programming language can provide all required features to develop software programs for the proposed system. Interpreter helps in execution of software program of proposed system.

C. Implementation Constraints

- **Choice of Platform**

The platform to be used for development and it will decide the performance of the system as some platforms are faster in terms of efficiency which is a desirable property for the proposed system for faster response. Speed of calculation and other operations makes a significant impact on working of the system. For the proposed system, we are using PyCharm platform for development which easing the load of the installation of required libraries without actual installation commands. It also provides access to the terminal inside it which makes execution of software programs easier.

- **Choice of development framework/environment**

A framework will help in quick development of the system by providing repeated patterns in development on a single command. As some frameworks specialize in specific types of systems it is important to choose a framework that suits best to the nature of the system to be developed. An environment (IDE) can help in coding and also be used to promote good standards of coding and bring consistency in different programs developed for the software part of the system.

3.3.5 Assumptions & Dependencies

It is assumed that the system is open source which performs the following tasks: -

- Accepting the input data (i.e. Audio file) from the user and passing it to the main audio source separation module for execution.
- Generate separate audio files of all involved individuals.
- Generate pop up error message box in case of incorrect input.

Assumption that the input files provided by the user will be in either .mp3 or .wav format to the proposed system which is the only acceptable format for the proper and error free execution of the software program.

System heavily depends on the platform and software specific libraries for the operation. For example, the Flask provides the libraries which provide the communication and data flow between working main source code and front end.

3.4 System Features

- **Dynamic behavior**

The system generates a different number of output audio files in case depending on the number of individuals involved in the input audio file. For displaying these varying outputs, frontend provides dynamic behavior in each case.

- **Error reporting**

This is responsible for reporting the errors occurred in execution of a software program. The system will generate the pop-up error message box in case the user provides incorrect input to the software program.

- **Adaptiveness**

This system is open to multiple input formats which can be provided by the system user. The user can provide .mp3 as well as .wav files as input to the system program for correct execution. The proposed system can adapt to any of the above-mentioned input formats.

3.5 External Interface Requirements

3.5.1 User Interfaces

- The system will be provided with the GUI for communication between users at a client side and software program acting as an interface.
- The simplified GUI will be easy to use and understandable to the user as the instructions for the user will be included in the GUI itself.

3.5.2 Hardware Interface

- As the proposed system is using the local hardware for storing the user data, no additional hardware interface will be needed.
- If the number of users increased, then additional hardware resources may be required to handle the load. Resources can be made available either locally or remotely for use.

3.5.3 Software Interface

- The PyCharm is used which provides a proper directory structure which is easier for accessing and it also provides easy installation of required libraries without any actual commands.
- A language interpreter is used to develop and run software programs.
- Libraries specific to communication between main module and frontend are used.

3.5.4 Communication interface

- Flask libraries will be used as a communication interface between the main module code and the frontend.
- These libraries also provide the proper data flow between main module code and frontend code.

3.6 Non-Functional Requirements

3.6.1 Performance Requirements

- **Adaptability**-The system developed should be adaptive to and should be able to handle multiple file input formats and not fail if an incorrect input file is provided by displaying the pop-up error message box at frontend.
- **Accuracy**-The system should accurately perform the audio separation task and display them to the user.

3.6.2 Safety and Security Requirements

A physical housing should be provided for the hardware components (if used) which is durable and provides longer life for the components and the working system should be secure and software programs should also be provided with security such that no external entity can alter the functioning of the system.

3.6.3 Software Quality Attributes

- **Availability:** The system should be working if big sized files are uploaded by the user, where it is required for the module code to be executed correctly and in estimated time without any delay providing the expected output to the user on the front end.
- **Correctness:** The system should perform audio separation task correctly by providing the equal number of detected number of individuals and their respected audio files without any noise.
- **Maintainability:** The proposed system is composed of and is using multiple frameworks and libraries. These frameworks and libraries are available in multiple versions, so proper updation should be done in time if newer versions are available which contain required functionality for correct functioning of the system.
- **Usability:** The user will be provided with a simplified GUI with usage instruction which will be easy to understand and adapt.

3.7 Other Requirements

3.7.1 Legal requirements

- **Technology related**

The software used is PyCharm and technologies used for development are Python, Flask, HTML, etc. which is free and open source so there are no legal requirements relating to development of software.

- **Real life implementation related**

As the proposed software will be open source, anyone can use it. Any audio recording from the sources like meetings, news debates or social events can be used as input to the proposed system for performing the separation task.

Chapter 4 DESIGN

4.1 Architecture design

The following architecture diagram focus mainly on flow of dictionary matrix W and coefficient matrix H . The GCC-NMF has an encoder-decoder block and an atom masking block. The encoder-decoder block is a two-layer structure consisting of an STFT transformation and NMF decomposition. This encoding decoding block is interfered by a speech separation block that makes use of spatial information of separate NMF atoms, at each point of time too mask attributes to interfering sources.

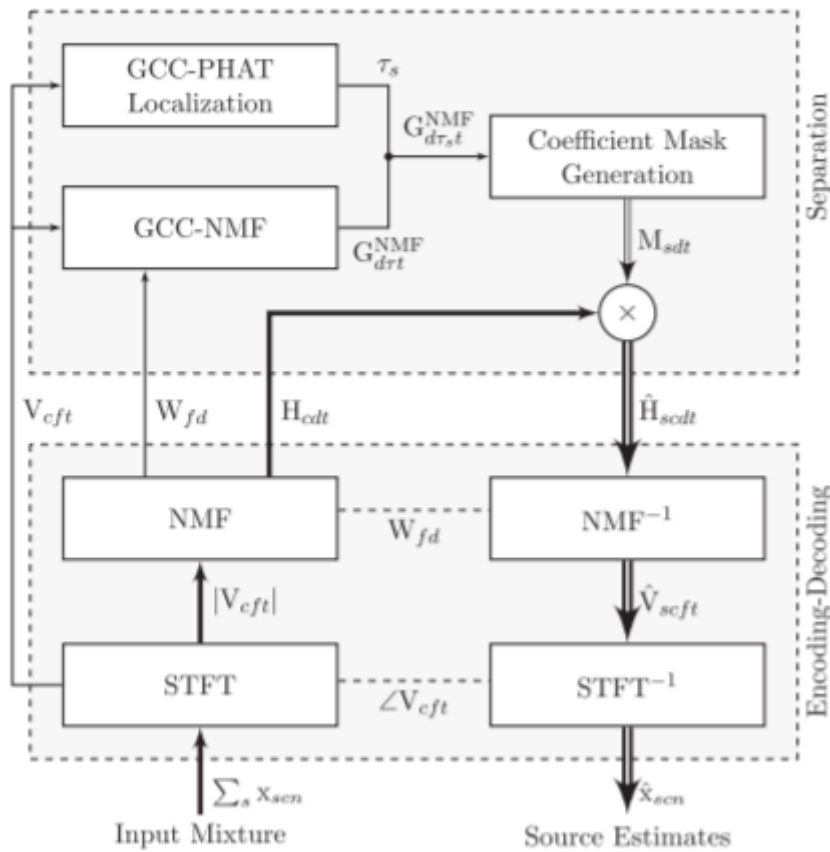


Figure 1- System Block Diagram

The GCC function introduces a generalized frequency weighting function to the TDOA estimation process. We may use the fact that NMF dictionary atoms are themselves non-negative functions of frequency to define a set of GCC frequency-weighting functions.

4.2 High level Project Design

4.2.1 Data Flow Diagram

The Data flow diagram represents the flow of data through the system and provides information about the outputs and inputs of each entity and the process itself.

- **DFD level 0**

DFD Level 0 is also called a Context Diagram. It's a basic overview of the whole system or process being analyzed or modeled. It's designed to be an at-a-glance view, showing the system as a single high-level process, with its relationship to external entities. It should be easily understood by a wide audience, including stakeholders, business analysts, data analysts and developers.

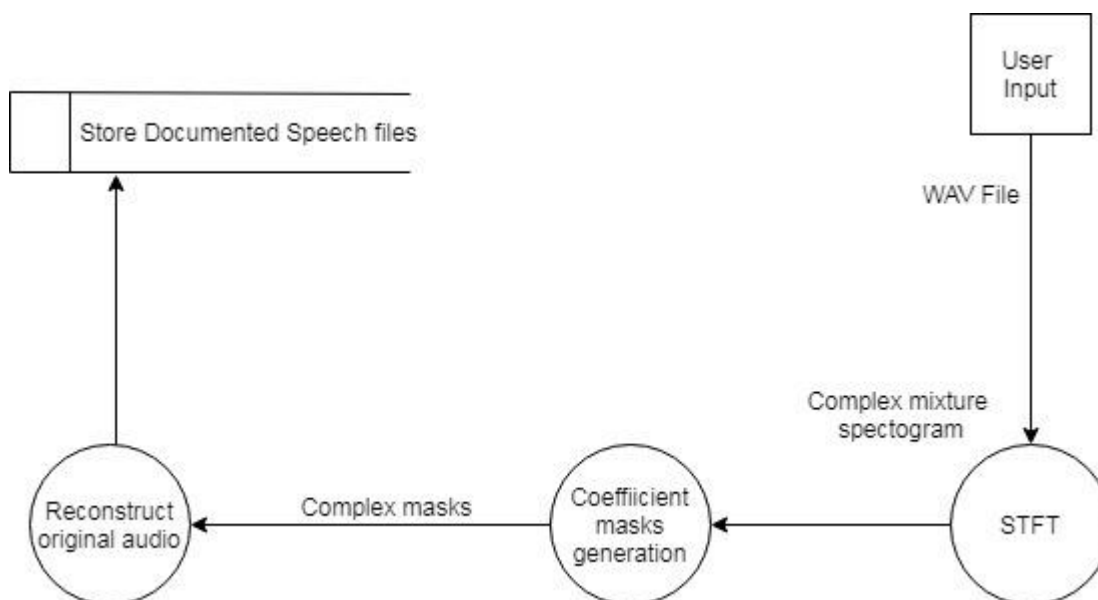


Figure 2 – Data Flow Diagram Level-0

- **DFD level 1**

DFD Level 1 provides a more detailed breakout of pieces of the Context Level Diagram. You will highlight the main functions carried out by the system, as you break down the high-level process of the Context Diagram into its sub processes.

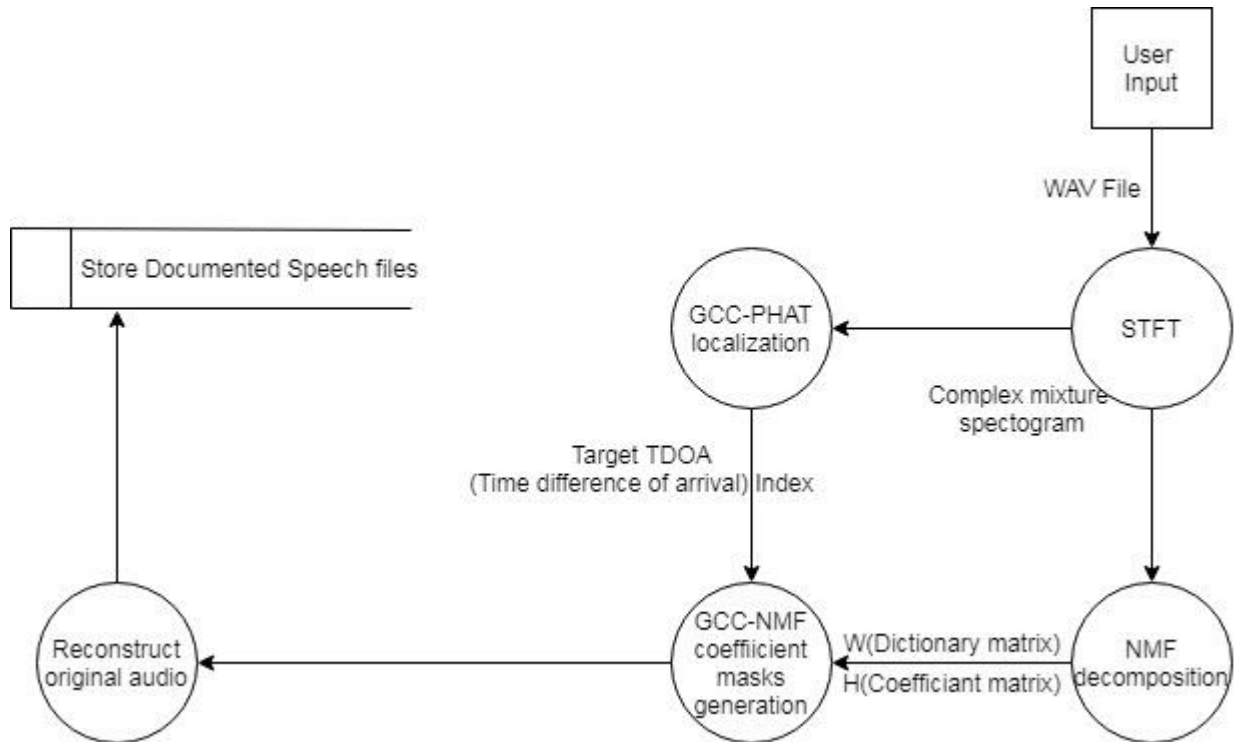


Figure 3 – Data Flow Diagram Level-1

- **DFD level 2**

A level 2 data flow diagram (DFD) offers a more detailed look at the processes that make up an information system than a level 1 DFD does. It can be used to plan or record the specific makeup of a system.

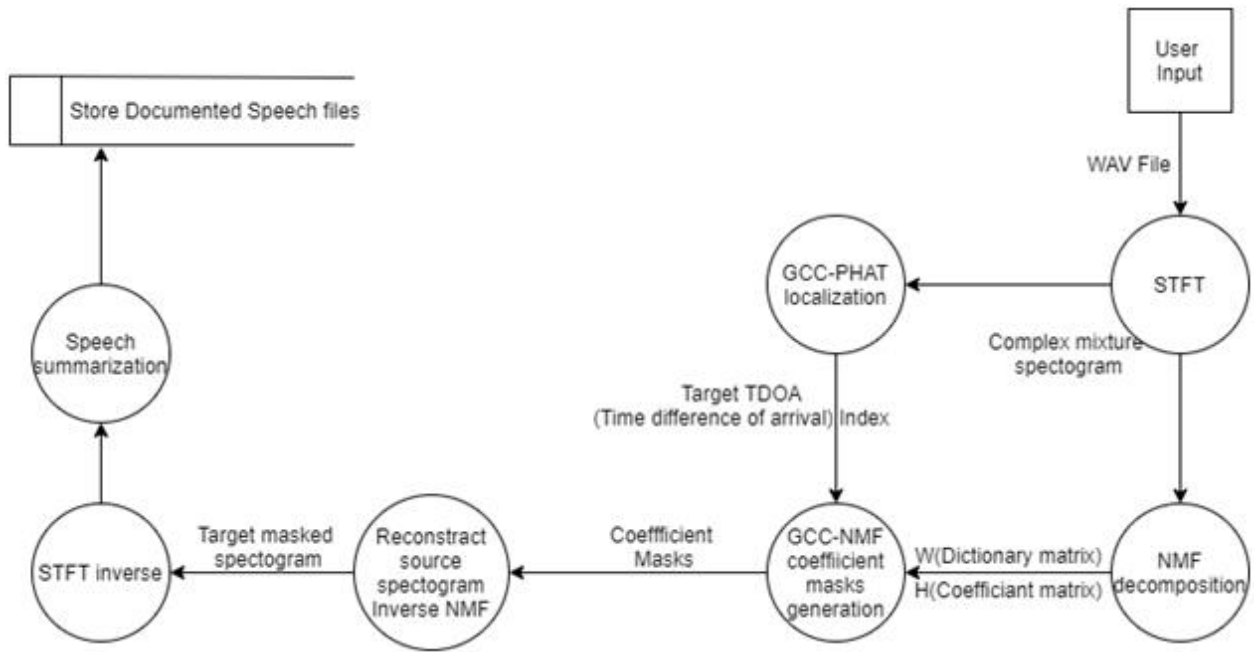


Figure 4 – Data Flow Diagram Level-2

The user input data is a stereo recorded sound of multiple speakers speaking concurrently. STFT converts the audio file into a frequency temporal file format. GCC-PHAT computes the time difference of arrival of sound of different speakers to the microphone. NMF decomposition converts the audio spectrogram into dictionary and coefficient matrices. The GCC-NMF module generates masks which can be used to separate the speech signals of different sources which is carried out in the Inverse NMF module. STFT Inverse plays the role of converting the separated spectrograms back in audio signals. Speech summarization focuses on creating textual interpretation of each speaker.

4.2.2 Activity Diagram

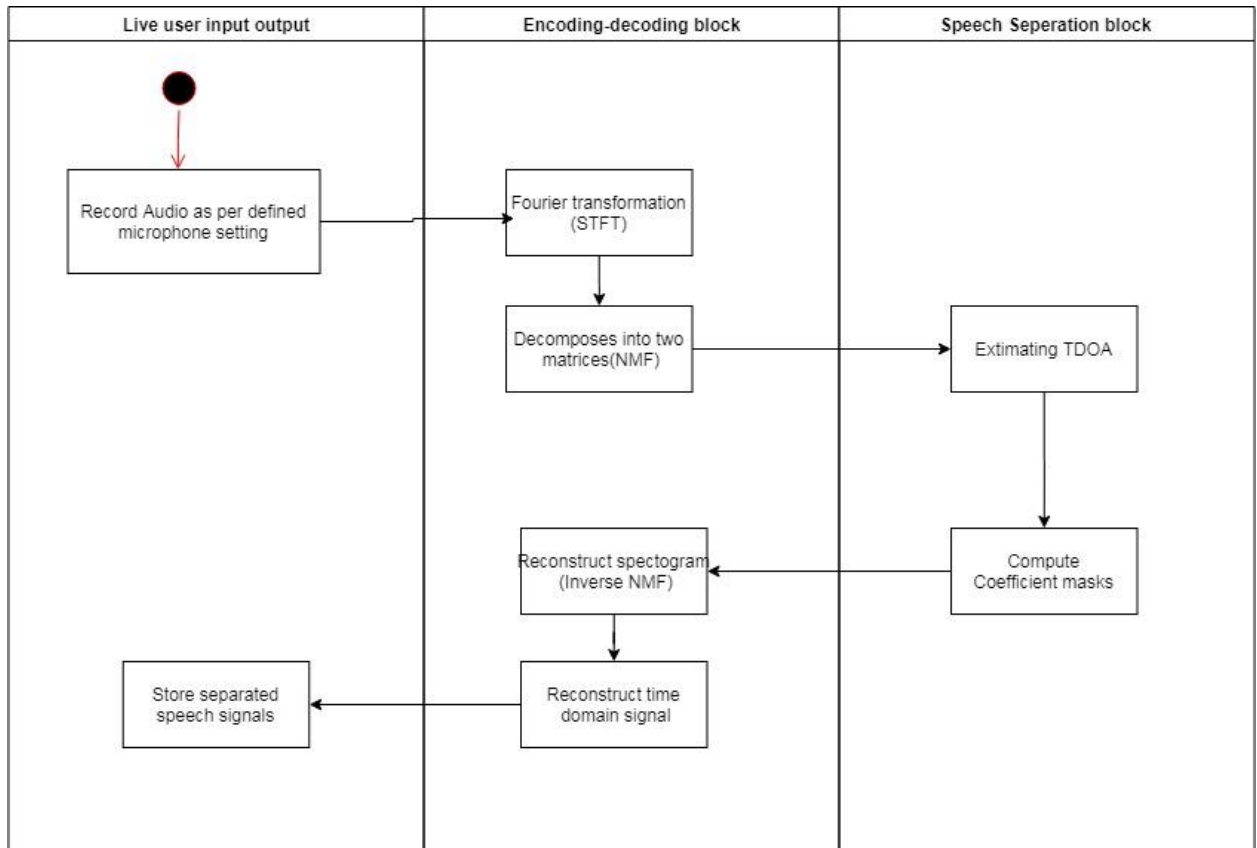


Fig 5- Activity Diagram

4.2.3 Sequence Diagram

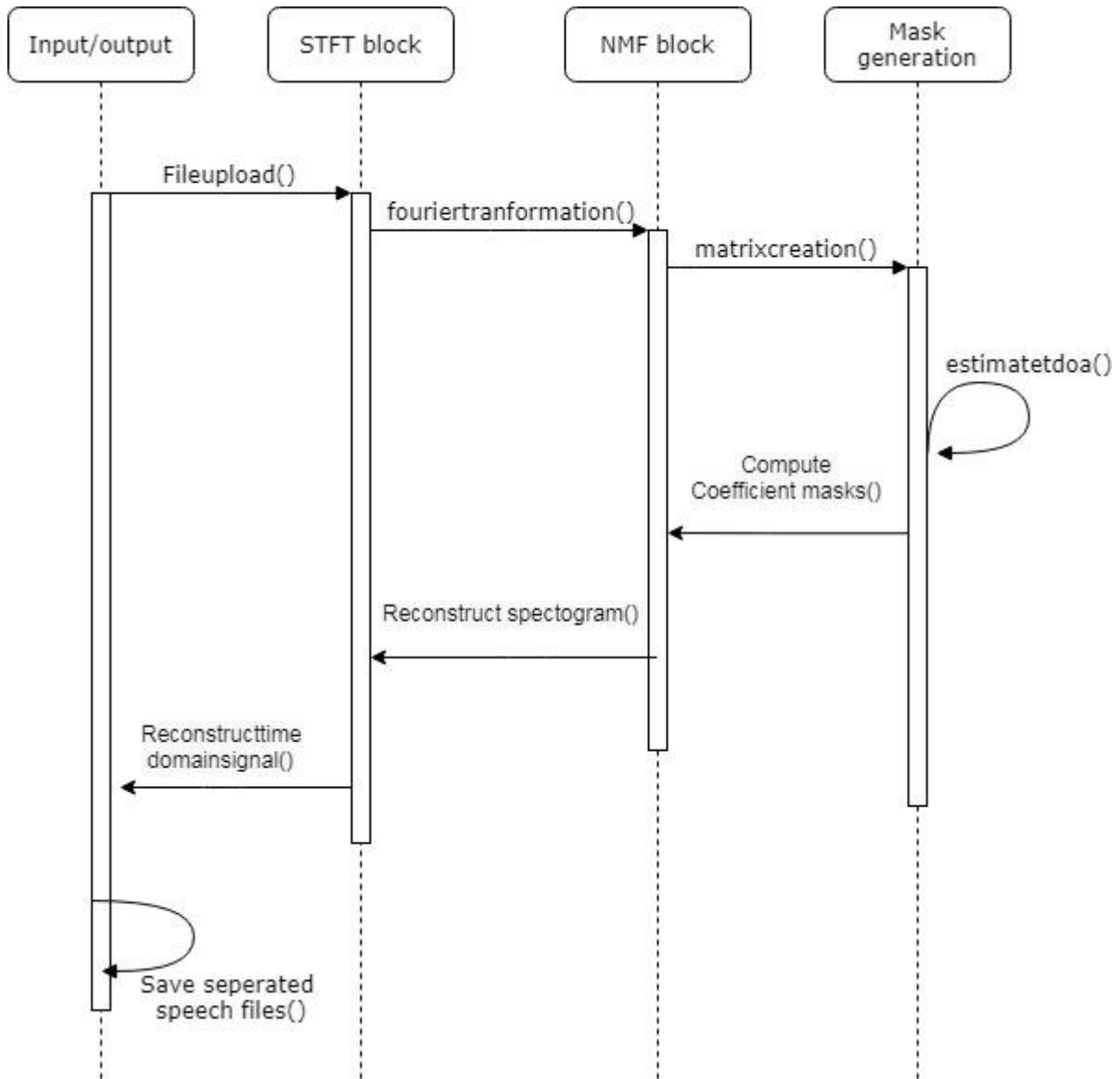


Fig 6- Sequence Diagram

Chapter 5

SYSTEM IMPLEMENTATION

5.1 System Architecture

A conceptual model that defines the structure, behavior, and views of the system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system

- **Description**

The basic building block of the proposed system is defined as a node. As per architecture the node is a structure which consist of the following architectural parts: -

1. STFT Fourier transform
2. NMF decomposition
3. GCC-PHAT Source Localization
4. GCC-NMF coefficient mask generation
5. Reconstruct source spectrogram - Inverse NMF
6. STFT Inverse

The functioning of the nodes mentioned above along with their inputs and outputs are described as follows

STFT Fourier transform

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. In practice, the procedure for computing STFTs is to divide a longer time signal into shorter segments of equal length and then compute the Fourier transform separately on each shorter segment. This reveals the Fourier spectrum on each shorter segment. One then usually plots the changing spectra as a function of time, known as a spectrogram or waterfall plot.

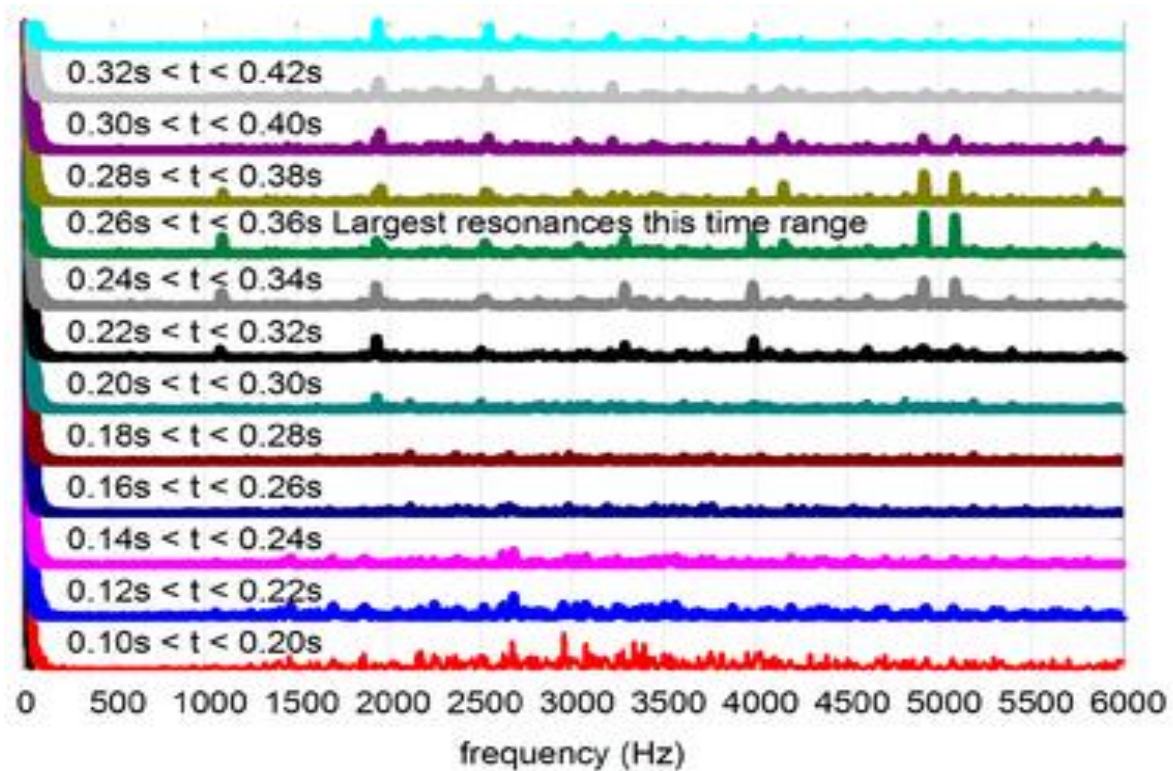


Fig 7- Waterfall plot

Input: Normalized Audio Files

Process: Fourier transforming file into Frequency-Time Domain

Output: STFT NumPy

NMF

It is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H , with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. It is also used in applications such as processing of audio spectrograms or muscular activity, non-negativity is inherent to the data being considered.

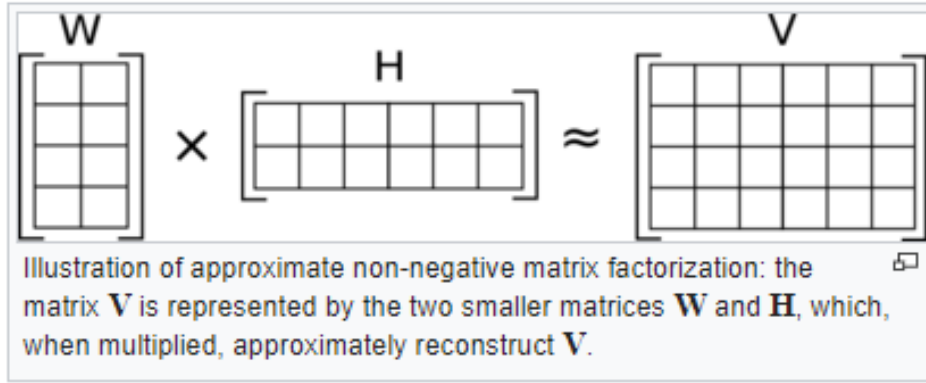


Fig 8- NMF Illustration

Application of NMF in reference to our problem:

Input to the NMF algorithm consists of a magnitude time frequency representation of the mixture signal, represented mathematically as a non-negative matrix V_{ft} with f and t indexing frequency and time respectively. NMF decomposes this spectrogram into two non-negative matrices: a dictionary matrix W_{fd} and a coefficient matrix H_{dt} , such that their product $\Lambda = WH$ approximates V . The d columns of W are referred to as dictionary atoms: non-negative functions of frequency that are combined linearly with the corresponding coefficients at each point in time to reconstruct the corresponding column of the input spectrogram

Input: STFT NumPy

Process: Decomposes STFT NumPy into two non-negative matrices

Output: Dictionary matrix W and a coefficient matrix H

GCC-PHAT Source Localization

The GCC function introduces a generalized frequency weighting function to the TDOA estimation process. The Generalized Cross-Correlation (GCC) is a classic method for estimating TDOAs for an arbitrary set of frequencies. The GCC represents an angular spectrogram: a function of time-delay τ and time t , defined mathematically as:

$$G_{\tau t} = \sum_f \psi_{ft} V_{lft} V_{rft}^* e^{j2\pi f\tau}$$

In order to compute the TDOA (Time Difference of Arrival) between the reference channel and any other channel for any given segment it is usual to estimate it as the delay that causes the cross-correlation between the two signals segments to be maximum. In order to improve robustness against reverberation it is normal practice to use the Generalized Cross Correlation with Phase Transform (GCC-PHAT). Given two signals $x_i(n)$ and $x_j(n)$ the GCC-PHAT is given as: Where $x_i(f)$ and $x_j(f)$ are the Fourier transforms of the two signals and $[\cdot]^*$ denotes the complex conjugate.

Input: STFT NumPy

Process: Estimating TDOA for the spatially different frequencies (Angular Spectrum)

Output: Target TDOA (Time difference of arrival) index.

GCC-NMF coefficient masking

We use the atom-specific angular spectrograms from GCC and NMF to associate each atom, at each point in time, to a single source based on its spatial origin. This is equivalent to defining a binary mask M_{dt} for each source, whose value is 1 if atom d is attributed to the source at time t , and 0 otherwise. A stereo source estimate spectrogram $\hat{V}_{cf}^s(t)$ can then be computed by multiplying the mask M_{dt} with the NMF coefficient matrices H_{cdt} elementwise, prior to reconstruction.

Input: Target TDOA, Dictionary matrix W and coefficient matrix H

Process: Computing the NMF coefficient masks for each target

Output: Coefficient masks

NMF inverse

Input: Coefficient masks, original complex spectrogram

Process: Reconstruction of source spectrogram estimates

Output: Target masked spectrogram

STFT inverse

Input: Target masked spectrogram

Process: Reconstruct time domain target signal estimates (Signal generation)

Output: Separated Speech Signals

5.2 Implementation and Deployment

With the GCC-NMF algorithm we proceed to evaluate our separation performance on real world speech separation tasks. Complex spectrograms are created using 16kHz mixture signals using short-time fourier transform using 1024 sample Hann window and 16-sample hop size(1 ms).

Default NMF parameters are set to 1024 dictionary atoms, 100 iterations, sparsity 0, cost function beta is 1. The GCC nonlinearity is used for 5 cm microphone separation, which results in more accurate localization for concurrent speaker task.

Dataset: Recordings for the concurrent speaker task are taken from SiSEC live speech recording dataset, constructed as static sources Default NMF parameters are set to 1024 dictionary atoms, 100 iterations, sparsity $\alpha = 0$, cost function $\beta = 1$.

The first task is a classic source separation task consisting of mixtures of concurrent speakers in reverberant environments. In this case, all speech signals are to be isolated, with separation quality subsequently averaged over the resulting source estimates.

The second task involves isolating a single target speaker mixed with real-world background noise, a speech enhancement problem that is well-suited for application in assistive listening devices, as well as a preprocessing stage in automatic speech recognition systems.

Finally, we consider a similar speech in noise task in which the speaker is allowed to move over time, reflecting more realistic real-world use cases, where the speaker or listener may not have fixed spatial locations.

Chapter 6

EVALUATION AND RESULTS

6.1 Feasibility Study

In current times the surveillance needs to be upgraded for the better security of citizens. Advanced technical systems should be used for surveillance of suspicious activities. These systems should capture the audios of people for surveillance purposes. Surveillance teams would use this system to separate suspicious audios and help catch criminals faster by understanding what they are saying. These systems can improve the existing systems to a great extent. These systems will help foil the plans of potential criminal activities. These systems should have low computational requirements and hence can be used everywhere.

These systems can be used by every local police department for surveillance of suspicious criminal activities.

It will be feasible to do this project because the existing systems are not good enough for surveillance purposes. Use of these systems will result in reduction in criminal activities and thus provide better security to citizens.

6.2 Risk Management

There are various factors which can pose a risk to the proposed system, These factors are as follows:

1. Input mix audio file:

Sometimes the mix audio file can be corrupt. It may contain excessive background noise which can affect the functionality of the system. The input audio file should be clean and should not contain background noise.

2. Quality of microphone:

The quality of microphone used in recording the input mix audio can affect the input mix audio file. In turn it can give bad results in the quality of separated audio files.

Good Quality microphones should be used while recording the mix audio file.

3. Distance between speaker and microphone:

Depending upon the distance between speaker and microphone, the quality of input audio file may vary. As the distance increases, the quality of input audio may decrease.

6.3 Experimental Setup

We evaluate the RT-GCC-NMF algorithm on the SiSEC 2016 speech in noise dev dataset, consisting of two-channel mixtures of speech and real-world background noise, with microphones separated by 8.6 cm. Unsupervised dictionary pre-learning is performed on a small subset of the CHiME 2016 development set, with randomly selected frames equally divided between isolated speech and background noise signals of a single microphone. The sample rate for both SiSEC and CHiME is 16 kHz, and we use an STFT with 1024-sample windows (64 ms), a 256-sample hop size (16 ms), and a square root Hann analysis and synthesis window functions for the symmetric windowing case. Default RT-GCC-NMF parameters are set to dictionary size = 1024, number of NMF dictionary pre-learning updates = 100, number of NMF activation coefficient inference updates at runtime = 100, number of TDOA samples = 128, and target TDOA window size $3/64$ of the total range, i.e. 6 TDOA samples.

6.4 Testing Strategy

- Unit Testing:

First, we tested the system with a mixed audio of 3 females. The microphone separation distance was 1.0m in this strategy. After execution of the program with proposed experimental setup, separate audio files were generated with good quality of sound. We tested the system with different sets of microphone separation distances and got good results. The Audio Separation unit works fine. Front End is developed using JavaScript. It is able to accept the mix audio file from the system and play the separated audio files in JavaScript.

- Functional Testing:

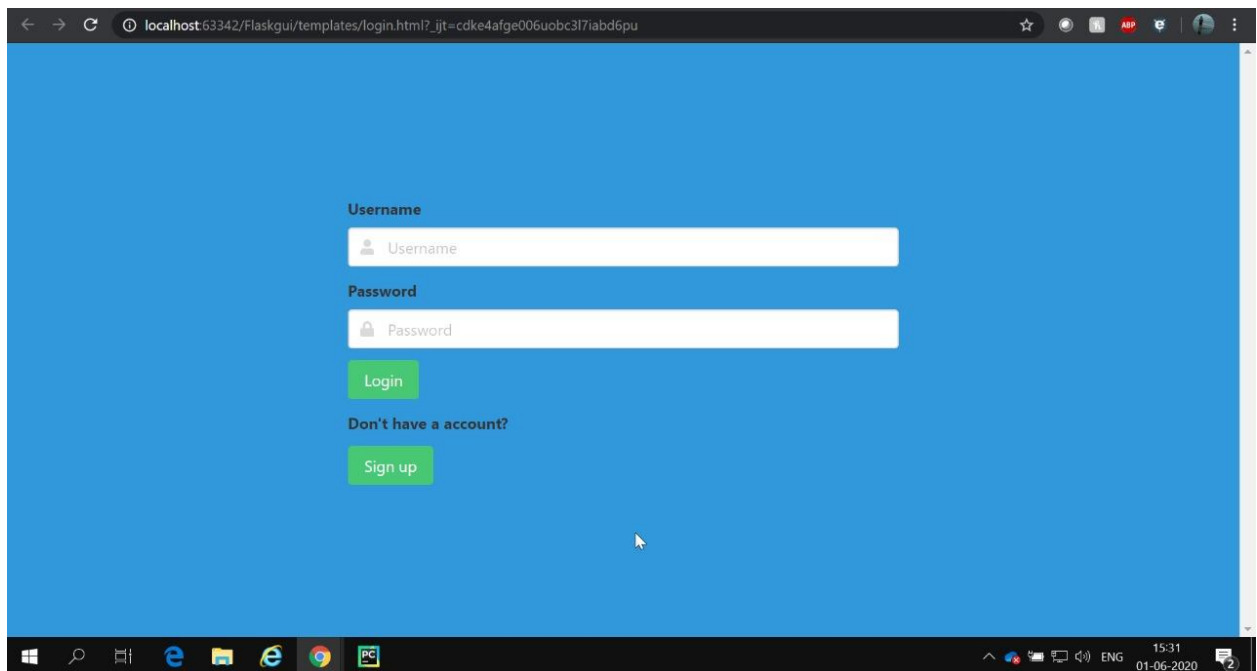
Main aim/requirement of the project is to separate the mix audio of speakers. The proposed system does separate the audio files and thereby passes the functionality test.

- Acceptance Testing:

The separated audio files may have a low volume for some cases, but you can still hear it using headphones. The proposed system meets the requirement of the project.

6.5 GUI and Results

- **Login Page:**



A screenshot of a web browser displaying a login page. The page has a solid blue background. In the center, there is a white login form. The form contains two input fields: 'Username' with a user icon and 'Password' with a lock icon. Below these fields are two green buttons: 'Login' and 'Sign up'. A link 'Don't have an account?' is positioned above the 'Sign up' button. The browser's address bar shows 'localhost:63342/Flaskgui/templates/login.html?_ijt=cdke4afge006uobc317iabd6pu'. The Windows taskbar is visible at the bottom with the time 15:31 and date 01-06-2020.

Username

Username

Password

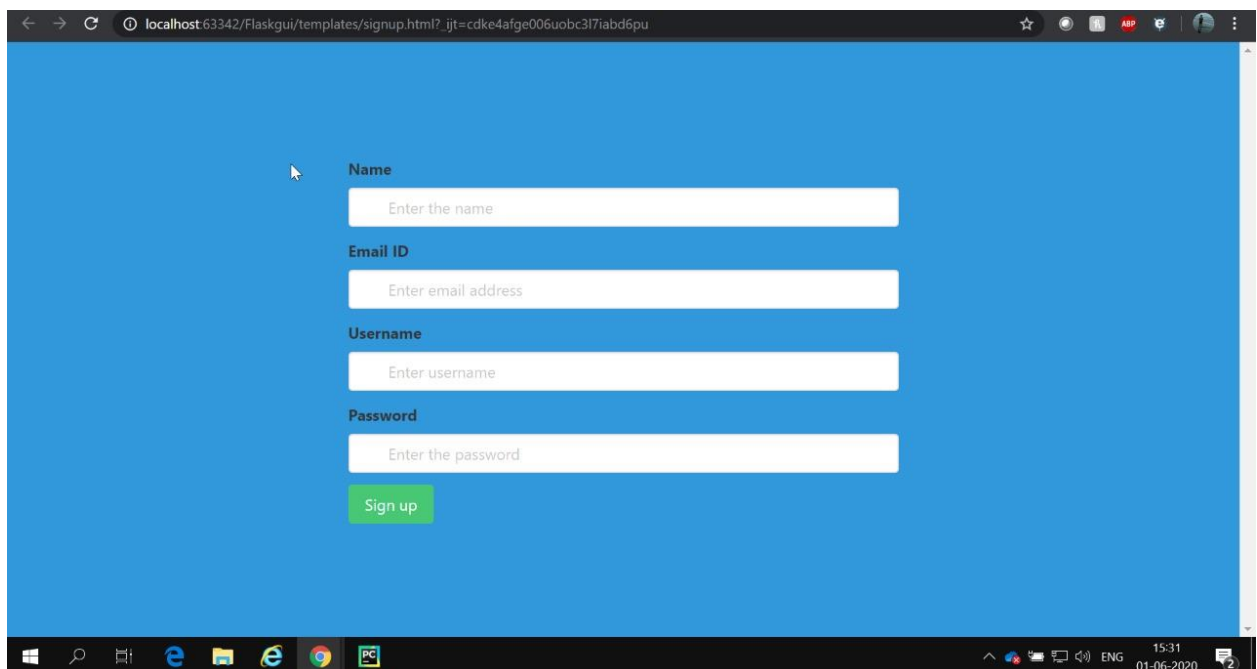
Password

Login

Don't have an account?

Sign up

- **Signup Page:**



A screenshot of a web browser displaying a signup page. The page has a solid blue background. In the center, there is a white signup form. The form contains four input fields: 'Name' (placeholder: 'Enter the name'), 'Email ID' (placeholder: 'Enter email address'), 'Username' (placeholder: 'Enter username'), and 'Password' (placeholder: 'Enter the password'). Below these fields is a green 'Sign up' button. The browser's address bar shows 'localhost:63342/Flaskgui/templates/signup.html?_ijt=cdke4afge006uobc317iabd6pu'. The Windows taskbar is visible at the bottom with the time 15:31 and date 01-06-2020.

Name

Enter the name

Email ID

Enter email address

Username

Enter username

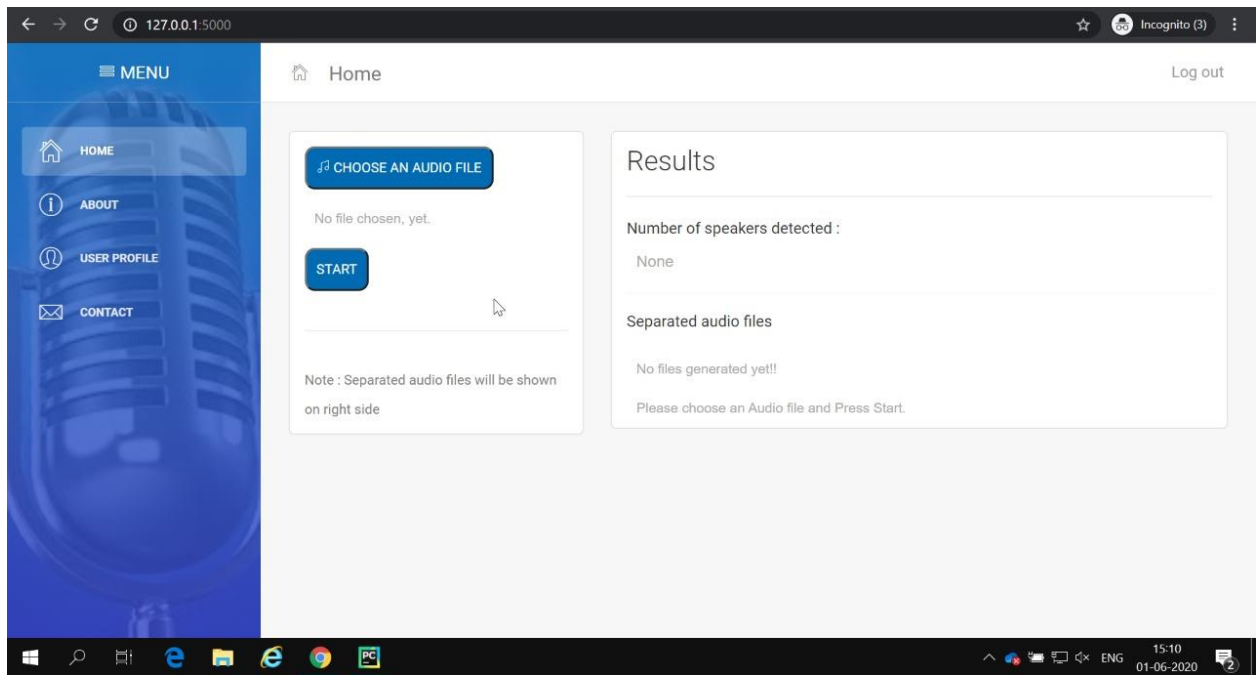
Password

Enter the password

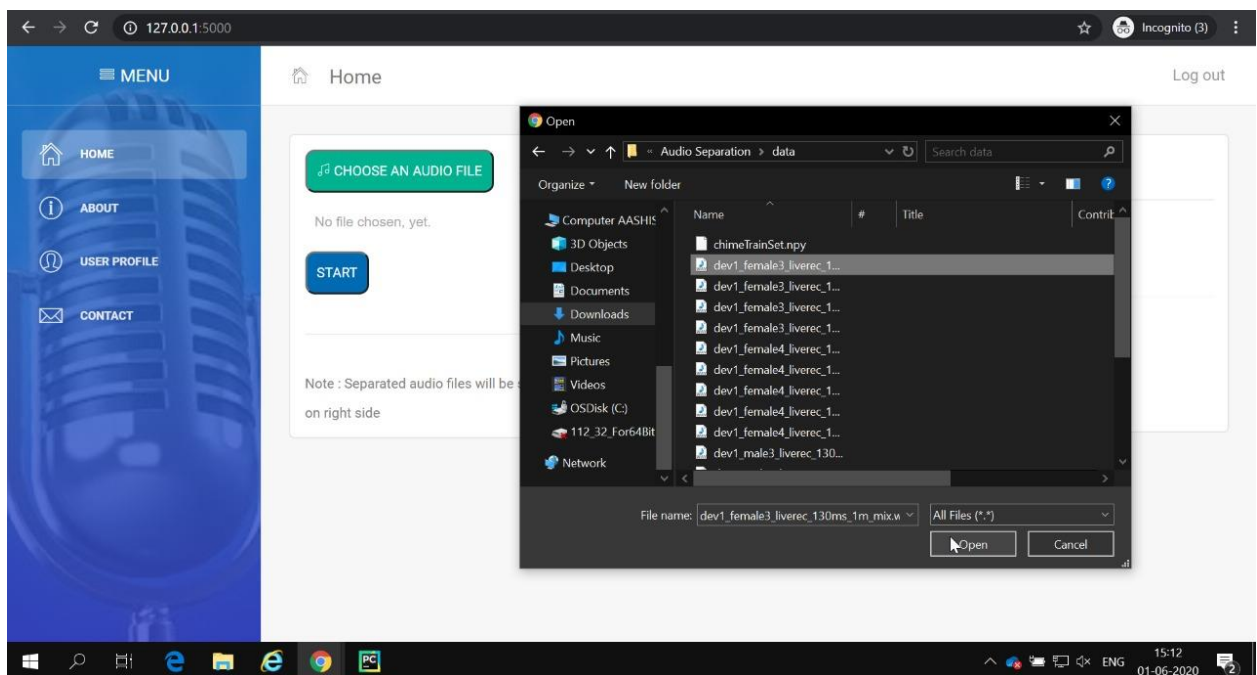
Sign up

- **Home Page:**

Audio Source Separation

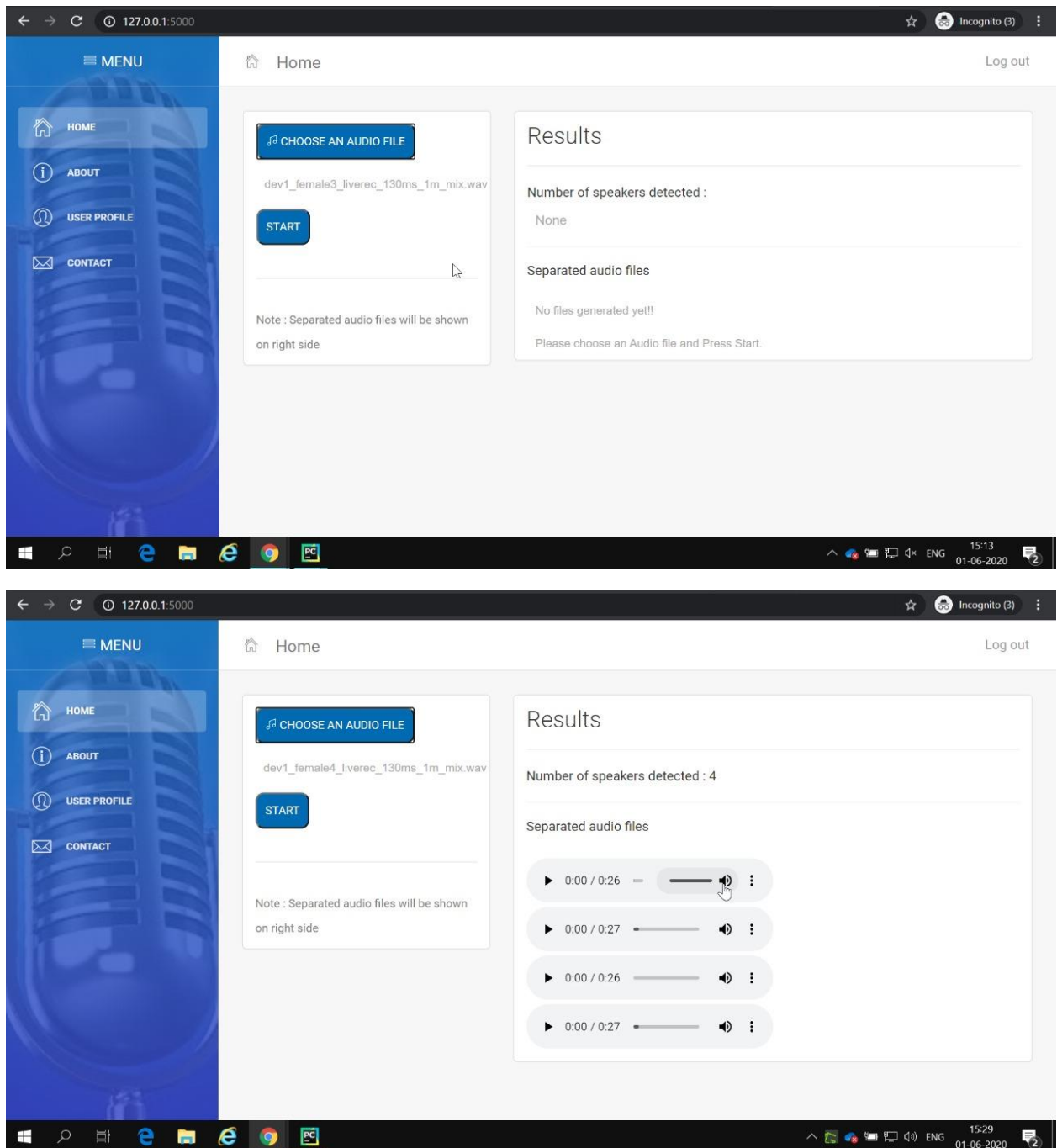


Step 1: First click on the 'CHOOSE AN AUDIO FILE' button.



Step 2: Select a mix audio file of two or more independent speakers from the computer.

Audio Source Separation



Step 3: After clicking on the 'START' button, audio separation system detects the no. of speakers from the mix audio file and then displays the separated audio files on GUI.

Chapter 7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

In this project we have successfully implemented blind audio source separation using STFT and GCC-NMF. The resulting combination of GCC and NMF is completely general, with no prior assumptions about the sources or mixing process, making GCC-NMF simple yet flexible. The flexibility of GCC-NMF was highlighted in applying it to three real-world blind speech separation tasks: mixtures of 3 and 4 concurrent speakers in reverberant environments, speech in real-world background noise, and noisy mixtures of moving speakers. We have introduced a new approach to combining spatial information with NMF for unsupervised source separation. Our system can successfully take in input of an audio source in which many people are talking together and can separate the voice of each speaker and provide his/her audio as the output. The audio of each person can be distinctly heard.

7.2 Future Scope

This system can be extended to include further algorithmic and memory optimizations to run RT-GCC-NMF on lower-power devices suitable for real-world hearing assistive applications. We can also study the use of GCC-NMF as a speech enhancement front-end in automatic speech recognition (ASR) systems, where its effect of word error rate (WER) will be evaluated. Existing audio to text system does not work well in current audio files where many speakers speak together. This reduces the accuracy of the system. Our system can be used to obtain audio of each and every speaker and the accuracy of the audio to text system will improve dramatically.

Our system in combination with audio to text system can be used to obtain automatic minutes of the meeting and by using sentimental text analysis on audio where people are talking which will improve the surveillance. Our application with audio to text software can be used to improve the closed caption subtitles which are currently not very accurate.

Chapter 8

REFERENCES

- [1] STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement: Martin Krawczyk and Timo Gerkmann, Member, IEEE:IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 12,
- [2]The generalized cross-correlation Method for time delay estimation of infrasound signal :Meng Liang Li Xi-Hai Zhang Wan-Gang Liu Dai-Zhi Xi'an Research Institute of High Technology Xi'an 710025, People's Republic of China: 978-1-4673-7723-2/15 \$31.00 © 2015 IEEE DOI 10.1109/IMCCC.2015.283
- [3]Analysis of the GCC-PHAT technique for multiple sources:Byoungcho Kwon , Youngjin Park and Youn-sik Park:International Conference on Control, Automation and Systems
- [4]Speech Enhancement using Non negative Matrix Factorization and Enhanced NMF : Akarsh K.A,Senthamizh Selvi R :2015 International Conference on Circuit, Power and Computing Technologies
- [5] Short-time Fourier Transform Analysis of EEG Signal From Writing C.W.N.F. Che Wan Fadzal, W. Mansor, L. Y. Khuan, A. Zabidi
- [6] Semantic Video Segmentation: A Review on Recent Approaches Mohammad Hajizadeh Saffar¹ . Mohsen Fayyaz² . Mohammad Sabokrou³ . Mahmood Fathy
- [7] Evaluation of bidirectional LSTM for short and long term stock market prediction
- [8] Khaled A. Althelaya, El-Sayed M. El-Alfy, Salahadin Mohammed
- [9] Semantic Video Segmentation: A Review on Recent Approaches Mohammad Hajizadeh Saffar¹ . Mohsen Fayyaz² . Mohammad Sabokrou³ . Mahmood Fathy¹
- [10] Complex Ratio Masking for Monaural Speech Separation Donald S. Williamson, Student Member, IEEE, Yuxuan Wang, and DeLiang Wang, Fellow, IEEE