Arabic Spoken Language Identification System

Submitted by:

Zeyad Sulalman Alhoti	438850187
Ibrahim Abdulmajeed Alsaif	439050466
Mohammed Ibrahim Alabdulhadi	441850756

Supervised by:

Dr .Mohammed Alatiyyah

Department of Computer Science

Shawwal,1443 H

Contents

Chapter 1

Chapter1 : introduction

In this chapter, the research problem will be clarified, how long it will take to solve this problem, and what is the system that will help us improve and develop a solution to this problem.

1.1 Problem statement(إبراهيم)

Many voice recognition applications are not specialized in recognizing the voice of the Quran reader, In this project improve the problem of these applications by collecting certain data about Quran readers and identifying them. Through voice recognition algorithms.

1.2 Proposed system(زياد)

The application contains many services available to users for free, and special work will be done to provide this application to meet the needs of users who are searching for the voice of a specific reciter and the verse in the Book of God Almighty, so that the process of searching and extracting the reader is a quick process in terms of the algorithms provided, and also It will provide some services such as the Noble Qur'an, hadiths of the Prophet, prayer times, morning and evening remembrances, and the direction of the qiblah.

1.3 Project plan(محمد)

In the section the tasks will be divided into weeks to complete this project according to a specific time plan

Task name	WEEK 1	WEEK 2	WEEK 3	WEEK 4	WEEK 5	WEEK 6	WEEK 7	WEEK 8	WEEK 9	WEEK 10	WEEK 11	WEEK 12	WEEK 13	WEEK 14	WEEK 15	WEEK 16
Chapter1 : introduction	▓	▓														
Chapter 2: Literature Review			▓	▓	▓											
Chapter 3: requirements					▓	▓	▓									
Chapter 4: System design							▓	▓	▓	▓						

1.4 conclusion

In general, the project is to solve the problem of recognizing the voice of the Quran reciter through an audio clip for reading, and it collects data for the readings of the reciters

Chapter 2

Chapter 2: Literature Review

In this chapter, we will talk about similar applications, the advantages and disadvantages of each application, the proposed system that solves this problem, the algorithm and the Features extraction that will help us in the solution as well.

2.1 introduction(محمد ال عبدالهادي)

in section 2.2 will talk about the application are similar to our application, and also in section 2.3 will talk about algorithm are used in Identification who speakers and also in section 2.4 will talk about what tools are used in Features extraction and also in section 2.5 will talk determine the proposal system are solve the problem and also in section 2.6 will be compared between similar to our application, in finally the chapter will write conclusion.

2.2 application(محمد ال عبدالهادي)

In this section, we will talk about the applications that are similar to our applications, and we will explore each application and its features

2.2.1 Soundhound(محمد ال عبدالهادي)

SoundHound: has applied audio processing and machine learning on millions of songs to extract features that are characteristic of each song,this is used to identify who speakers,the application works with natural language understanding technology.

Features of the SoundHound App

- Immediately identifies the song.
- Extract words from audio.
- Getting to know the Quran reciter but not well

Disadvantages of the SoundHound app

- Doesn't always play the correct song

- Has a slight learning curve.
- Does not recognize the sound in Arabic well

2.2.2 Shazam(إبراهيم السيف)

Definition of Shazam application:-

    Shazam is an application that can identify music, movies, TV shows and clips in general by taking a sample of the audio clip, and this is done using the device's microphone.

Who are the users of the application?:-

    Users of the Shazam application are the ones who want to find the name of the music or audio track that was searched, the Shazam application allows them to find the name of the music or audio track that they want.

Shazam application advantages and disadvantages:-

- **Advantages:**
  - ❖ easy to use.
  - ❖ Easy to discover music.
  - ❖ Recognizes the voice in ten second.
- **Disadvantages**:
  - ❖ The user can not sing the song and search for it.
  - ❖ It is not easy to recognize the reader.

Technology used in Shazam application:-

    The Shazam application uses the microphone of the mobile phone or the device in which the application is located to identify the audio clip, and Shazam uses fingerprint technology to puts it in a graph through which the audio clip is fragmented and searched in databases.

2.2.3 MusixMatch(زياد الحوطي)

How does the program work and howdoes it work?

A program that can use some algorithms to obtain music through the microphone, after which it is read and extracted the results from the sounds of the Qur'an, poem or music in general.

Who is using the app or who needs it?

Those who use the app or who need it are people who have some music and want to know more about it.

Features and Disadvantages:

-Some of the features

- Some words can be captured in offline mode.
- Tracking the fast download feature of the subscribed user.
- Can read static characters from within a page.

-Some of the Disadvantages

- Sometimes the lyrics are not accurate.
- Classical music instruments cannot be read.
- It may not be possible to get all the lyrics correctly.

2.2.4 Rateel (محمد ال عبدالهادي)

Rateel : Application to identify the reader of the Qur'an,use the Application who want to search for the voice of the reciter of the Qur'an..

Features of the Rateel App

- Get to know the reciter of the Qur'an

Disadvantages of the Rateel app

- Does not recognize the reader in the short voice

- It takes a long time to get to know the reader

- Has a slight learning curve.

2.3 Proposal System(محمد وابراهيم)

This system aims to identify and know the voice of the reciters in the Qur'an and make this application recognize quickly and correctly and make the application learn to help identify the new reader

2.4 Algorithms(محمد وابراهيم)

Hidden Markov model is one of the successful techniques of voice modeling in speech recognition systems. The reasons for the success of the model are the analytical ability to recognize speech and its accuracy in the systems.
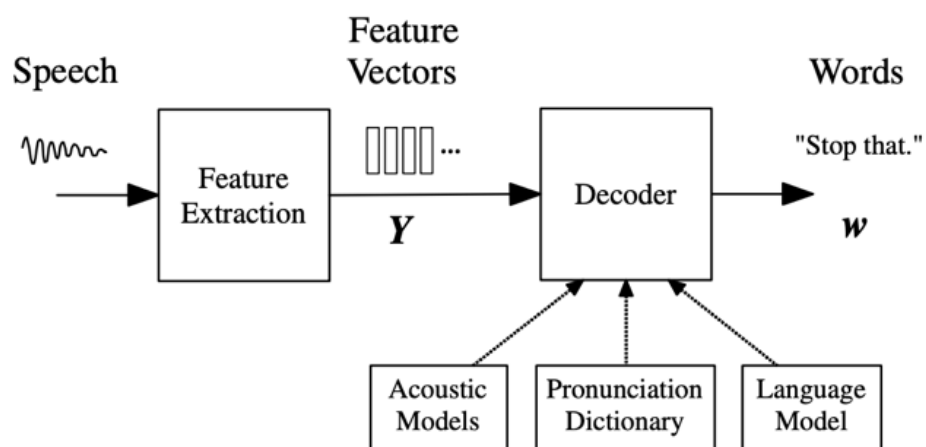

Figure 1 Architecture of a HMM-based Recogniser

2.5 Features extraction(محمد وابراهيم)

The first step in speech recognition system is to extract features here we will use Mel Frequency Cepstral Coefficients (MFCCs) ,it is a feature widely used in speech and speaker recognition.
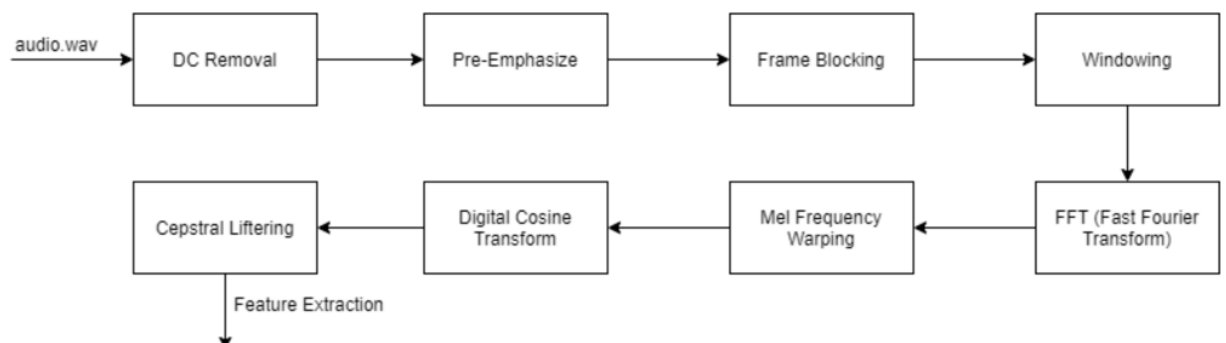

Figure 2 MFCC Process

2.6 Comparison of the Research

	SoundHound	Shazam	MusixMatch	Rateel
recognition speed	✔	✔	✔	
Arabic speaker recognition	✔	✔	✔	✔
Often recognized correctly		✔	✔	
recognition the reader of the Qur'an	✔	✔	✔	✔
high learning curve		✔		
easy to use.	✔	✔	✔	

2.7 Conclusion

In this chapter we talked about similar applications and compared them and searched for some algorithms to identify the voice of the reciter in Arabic.

Chapter 3

Chapter 3: requirements:(ابراهيم)

In this chapter, we will talk about the functional and non-functional requirements of the proposed system and user, what quality is required of the system, and what is the response time.

3.1 Functional requirements:(ابراهيم)

They are the features and functions of the system that the system developer must implement to enable the user to use the system as required. The functional requirements are divided into two parts:

1- User requirements.

2- System requirements.

In this section of the chapter we will talk about what are the user requirements and what are the system requirements.

3.1.1 User requirements:(ابراهيم)

User requirements are what the user shall or should have to use the system as required, in this section we will review the points of the user requirements:

1- The user shall have the microphone.

2- The user shall have small clip from the voice of his Quran reader.

3- The user shall test specific Quran readers.

4- The user should test the system in a quiet place to capture the voice of the reader.

3.1.2 System requirements(زياد)

- If the system takes the voice of a reader, it sends it to the database, and then the voice of the reader is recognized and sent to the system.

- If the system does not recognize the voice of the reader, it will be saved in the database and will be trained in the upcoming passages.

- Mostly there are votes for the reader, but if it is new, a fingerprint is created for it based on the sample and it is compared in the central database for later matching

3.2 non - Functional requirements(محمد)

In the non-functional requirements we talk about additional features of

the program, such as Quality and response time, which are requirements for the overall

quality of the program

3.2.1 Quality(محمد)
It is the quality in extracting the correct and accurate reciter of the Qur'an.

3.2.2 response time(محمد)
It is the response time of the program through sound and recognition of the reader.

Chapter 4

Chapter4 : System design

4.1 Introduction

Speech is one of the ancient ways to express ourselves, and today's speech signals are also used in biometric recognition and machine communication techniques.One of the theoretical aspects is that we can recognize speech directly from the digital wave, and with the great variation in the speech signal, it is better to extract features that reduce this contrast, and we can get rid of sources that produce some different information.In this chapter, we will talk about Feature extraction, as it includes a lot of topics, and we will use LPC and MFCC , The Hidden Markov Model and Speech Recognition.

4.2 Feature Extraction(زیاد)

Feature extraction is the process of obtaining various features such as energy composition, tone and vocal tracts from a speech signal. Parameter conversion is the process of converting these features into signal parameters,In speaker independent speech recognition, a premium is placed on extracting features that are somewhat invariant to changes in the speaker. So future extraction involves analysis of speech signal. Broadly the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis.

LINEAR PREDICTIVE CODING (LPC)

LPC is one of the most powerful speech analysis techniques and is a useful method for encoding quality speech at a low bit rate. The basic idea behind linear predictive analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples.

Mel Frequency Cepstral Coefficients MFCC
The use of Mel Frequency Cepstral Coefficients can be considered as one of the standard method for feature extraction.

Extraction Mel-frequency cepstral coefficients (MFCC) from the audio recording signals.
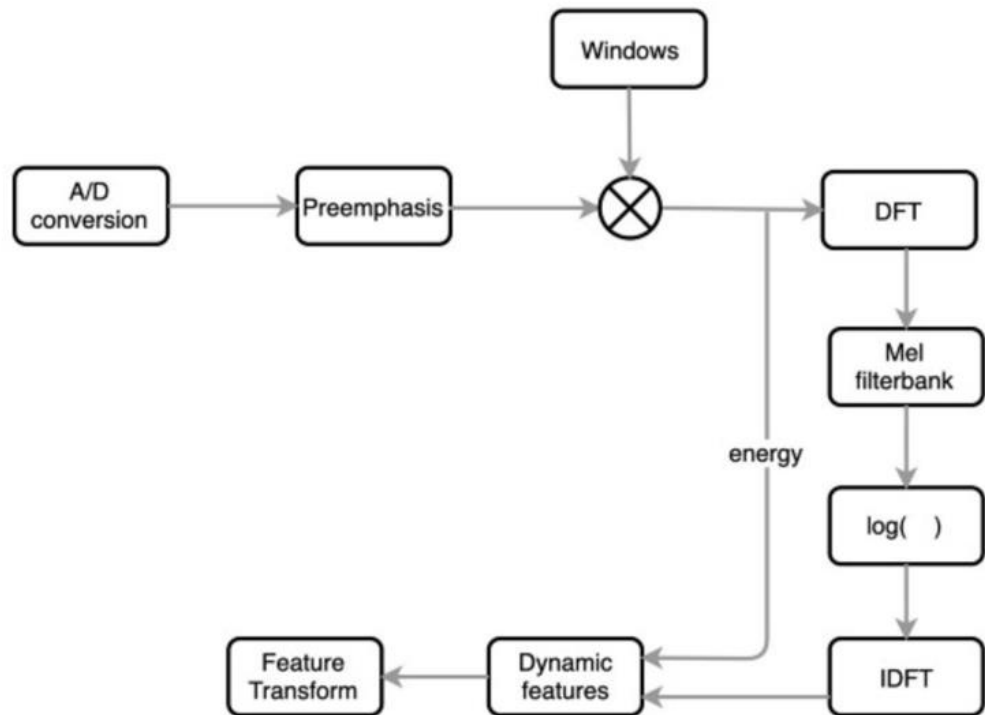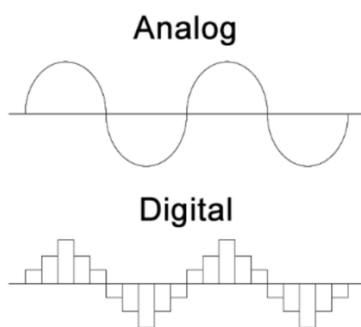

A/D Conversion:

In this step, we will convert our audio signal from analog to digital format with a sampling frequency of 8kHz or 16kHz.

Preemphasis:

Preemphasis increases the magnitude of energy in the higher frequency. When we look at the frequency domain of the audio signal for the voiced segments like vowels, it is observed that the energy at a higher frequency is much lesser than the energy in lower frequencies. Boosting the energy in higher frequencies will improve the phone detection accuracy thereby improving the performance of the model.

Windowing:

The MFCC technique aims to develop the features from the audio signal which can be used for detecting the phones in the speech. But in the given audio signal there will be many phones, so we will break the audio signal into different segments with each segment having 25ms width and with the signal at 10ms apart

DFT (Discrete Fourier Transform):

Convert the signal from a field to a field (dft) for engineering signals, and the analysis is easier.

Mel-Filter Bank:

The way our ears will perceive the sound is different from how the machines will perceive the sound. Our ears have higher resolution at a lower frequency than at a higher frequency.

Applying Log:

Humans are less sensitive to change in audio signal energy at higher energy compared to lower energy. Log function also has a similar property, at a low value of input x gradient of log function will be higher but at high value of input gradient value is less. So we apply log to the output of Mel-filter to mimic the human hearing system.

IDFT:

Here he performs the inverse conversion of the output from the step before it, and we have to understand how the sound is produced by humans

[iy] (tea)   [ae] (cat)   [uw] (moo)
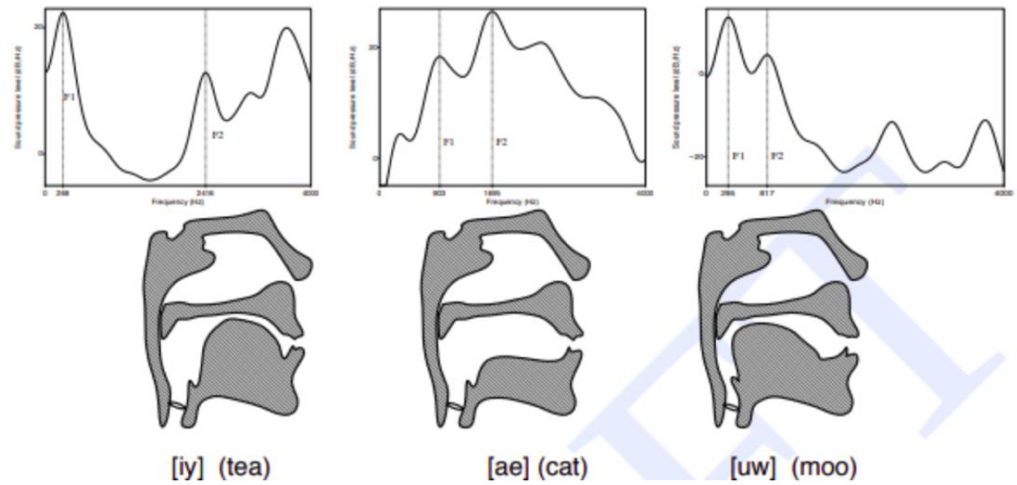
Dynamic Features:

It computes derivatives by coefficients among audio signal samples and helps understand the occurrence of the transition.

4.3 The Hidden Markov Model(ابراهيم)

Hidden Markov models depend on state sequences and their probability, and the probability sequence problem can be solved and simplified by Markov Assumption and depends on the transition probability on the current state ,P=Transition probabilities ,S=States:

$$P(S_4,S_3,S_2,S_1)=P(S_4|S_3,S_2,S_1)\times P(S_3|S_2,S_1)\times P(S_2|S_1)\times P(S_1)$$

In this equation, a representation of the Markov Assumption, so that the transition possibilities are divided in order with the state, which expresses this:

$$P(S_n, ..., S_1) = \prod_{i=1}^{n} P(S_i|S_i - 1)$$

There is such a thing as one fair coin that explains the states in two states 1 and 2

$$P(H) = 1.0 \qquad P(H) = 0.0$$
$$P(T) = 0.0 \qquad P(T) = 1.0$$
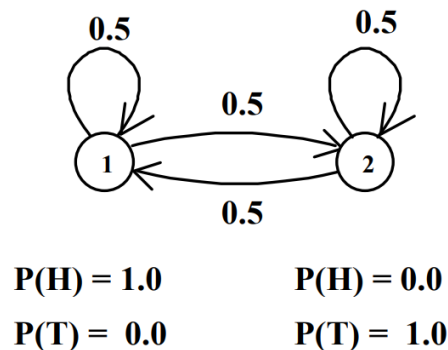
In this diagram, there are two nodes, and for each node, there are two possibilities, and these two possibilities are divided by 100, that is, for every 50 that node is 1 or 50 percent that it is node 2.

4.4 Speech Recognition and HMM(محمد)

Speech recognition is a powerful tool of the information exchange using the acoustic signal. Therefore, not surprisingly, the speech signal is for several centuries the subject of research. Speech recognition is a technology that able a computer to capture the words spoken by a human with a help of microphone. These words are later on recognized by speech recognizer, and in the end, system outputs the recognized words, a number of techniques, such as linear-time-scaled word-template matching, dynamic-time-warped word-template matching, linguistically motivated approaches (find the phonemes, assemble into words, assemble into sentences), and hidden Markov models (HMM), were used. Of all of the available techniques, HMMs are currently yielding the best performance
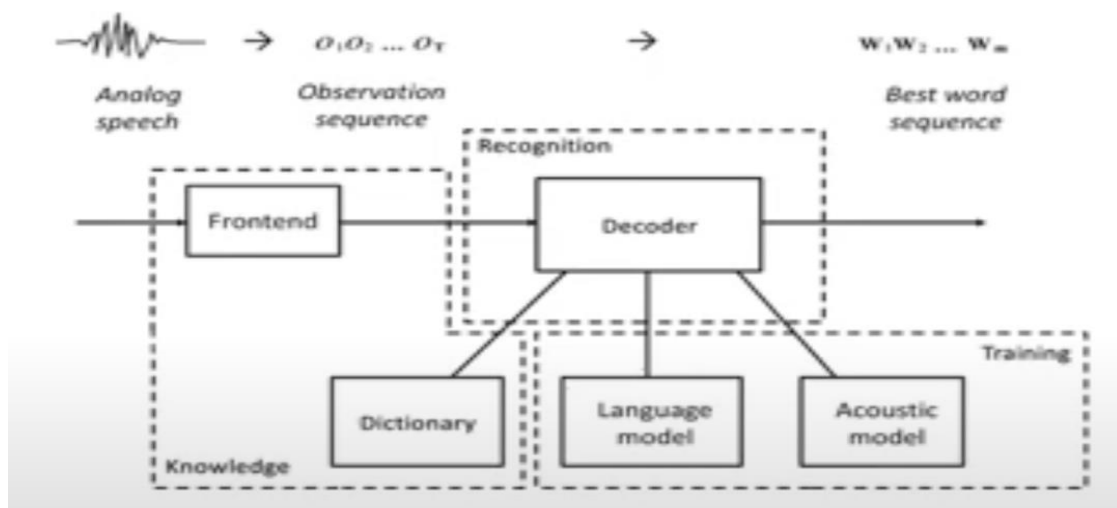

Figure 3 model base recognizer

The frontend converts the analog speech stream into Observation sequence.

The Acoustic model on used to express the BRU(basic recognizer unit) into mathematical mode.

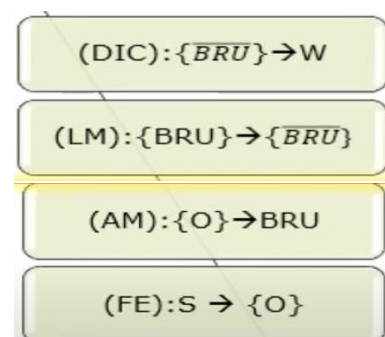Language model converts semantic of the language for the recognition engine.

The dictionary converts the basic recognized units into words (Expected symbols to user) .

BRU(basic recognizer unit) → phone

Frontend → feature extraction

Acoustic model → HMM

$(DIC): \{\overline{BRU}\} \rightarrow W$

$(LM): \{BRU\} \rightarrow \{\overline{BRU}\}$

$(AM): \{O\} \rightarrow BRU$

$(FE): S \rightarrow \{O\}$

Language model →natural Language processing

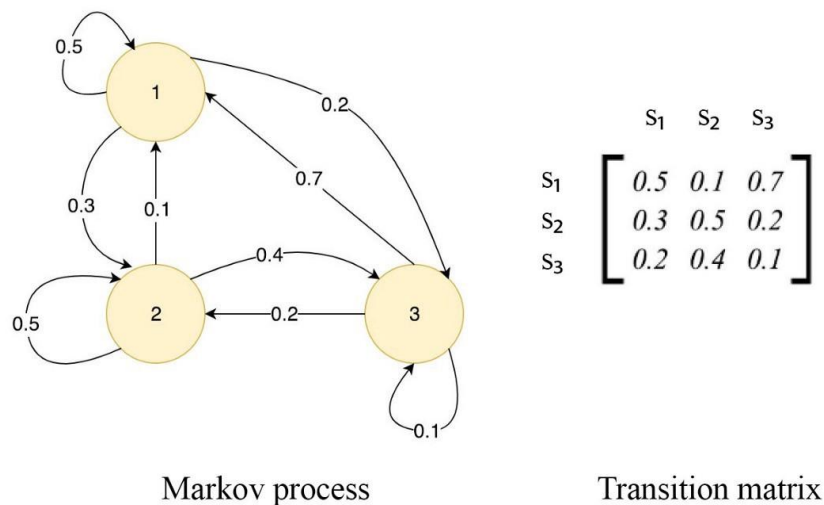Dictionary→ Sequence of phone to word

Acoustic Model

- HMM can be utilized as acoustical model.
- It converts the temporal acoustical features into statistical model.
- BRU(basic recognizer unit) → phoneme. Speech signal is a stream of Phonetics
- Each phoneme is homogeneous features in certain time duration.

Speech recognition consists of two main modules, feature extraction and feature matching. The purpose of feature extraction module is to convert speech waveform to some type of rep- resentation for further analysis and processing, this extracted information is known as feature vector. The process of converting voice signal to feature vector is done by signal-processing front end module. As shown in above block diagram input to front-end is noise free voice sample and output of it is feature vector.In feature matching, the extracted feature vector from unknown voice sample is scored against acoustic model, the model with max score wins ,and it's output is considered as recognized word. Following are the few method for implementing front-end(for extracting feature factor)

• MFCC (Mel-Frequency Cepstrum Coefficient)

• LPC (Linear Predictive Coding)
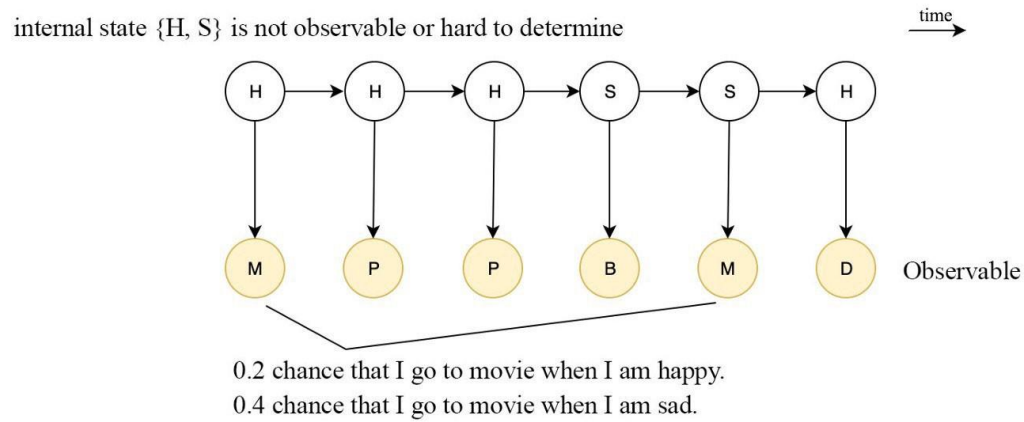
Hidden Markov Model (HMM)

The Hidden Markov Model(HMM) is a powerful statistical tool for modeling generative sequences that can be characterised by an underlying process generating an observable sequence.HMMs have found application in many areas interested in signal processing, and in particular speech processing, but have also been applied with success to low level NLP tasks such as part-of-speech tagging, phrase chunking, and extracting target information from documents,A Markov chain contains all the possible states of a system and the probability of transiting from one state to another.

$$
\begin{array}{c c c c}
 & S_1 & S_2 & S_3 \\
\begin{array}{c} S_1 \\ S_2 \\ S_3 \end{array} &
\left[\begin{array}{c c c}
0.5 & 0.1 & 0.7 \\
0.3 & 0.5 & 0.2 \\
0.2 & 0.4 & 0.1
\end{array} \right]
\end{array}
$$

Markov process   Transition matrix

A first-order Markov chain assumes that the next state depends on the current state only. For simplicity, we often call it a Markov chain.

$$\mathrm{P}\left(X_{n+1} = x \mid \underbrace{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n}_{\text{can be ignored}}\right) = \mathrm{P}\left(X_{n+1} = x \mid X_n = x_n\right)$$

This model will be much easier to handle. However, in many ML systems, not all states are observable and we call these states hidden states or internal states. Some may treat them as latent factors for the inputs. For example, it may not be easy to know whether I am happy or sad. My internal state will be {H or S}. But we can get some hints from what we observe. For example, when I am happy I have a 0.2 chance that I watch a movie, but when I am sad, that chance goes up to 0.4. The probability of observing an observable given an internal state is called the emission probability. The probability of transiting from one internal state to another is called the transition probability.

internal state {H, S} is not observable or hard to determine

time →

Observable

0.2 chance that I go to movie when I am happy.
0.4 chance that I go to movie when I am sad.

For speech recognition, the observable is the content in each audio frame. We can use the MFCC parameters to represent it. Let's see what we can do with an HMM.