WEARABLE VIBRATION-BASED DEVICE FOR HEARING-IMPAIRED PEOPLE USING ACOUSTIC SCENE CLASSIFICATION

Anju. L[1], Aniruddh Aiyengar[2], Tamil Selvan H[3], Vishnuvaradhan Moganarengam[4]

[1]Assistant Professor, Department of Electronics and Communication Engineering, Sri Venkateswara College of Engineering, Chennai, Tamilnadu, India.
[2]Student, Department of Electronics and Communication Engineering, Sri Venkateswara College of Engineering, Chennai, Tamilnadu, India.
[3]Student, Department of Electronics and Communication Engineering, Sri Venkateswara College of Engineering, Chennai, Tamilnadu, India.
[4]Student, Department of Electronics and Communication Engineering, Sri Venkateswara College of Engineering, Chennai, Tamilnadu, India.

---***---

Abstract - *Numerous people around the world experience various hearing difficulties. A sense of sound is critical for the quality of life of people with hearing difficulties, including deaf people. Wearable devices have substantial potential for sound recognition applications owing to their low cost and lightweight. They can assist deaf and physically impaired people to perform their daily activities more easily without requiring assistance from others. In this present study, A wearable assistive device has been designed and developed for the hearing impaired that informs the users of important sounds through vibrations, thereby understanding what kind of sound it is. The aim is to provide assistance for deaf people who cannot move around easily without any support. The classical Mel-frequency cepstral coefficients (MFCC) are used for feature extraction. Mel Frequency Cepstral Coefficient (MFCC) is for extracting the features from Audio. So overall MFCC technique will generate 13 features from each audio signal sample which are used as input for the Classification model. The performance of three Hidden Markov Model (HMM) models are compared with respect to various parameters and an improved HMM approach is proposed. An ensemble modelling approach is used to combine the three models to improve the accuracy of classification of the environmental sounds. This method is effective, robust, and well-suited for hearing aid applications. The model is deployed on Raspberry Pi which produces the vibration output based on the model prediction.*

Keywords: Hidden Markov Model (HMM), Mel-frequency cepstral coefficients (MFCC), Raspberry Pi, Wearable processing, Hearing aid, environmental sound.

1.INTRODUCTION

1.1. Overview

The global burden of hearing impairment is estimated at 466 million people (6.1% of the world's population) where 432 million (93%) of these are adults (242 million males, 190 million females) and 34 million (7%) children. It is judged that the number of people with deafness will grow to 630 million by 2030 and maybe over 900 million by 2050.

Hearing impairment has an important bearing on many aspects of an individual's life, including their socioeconomic status, mental well-being, education and employment opportunities. Older people with moderate or more severe hearing loss were more likely to feel depressed and suffer from poor mental health. The deaf child cannot listen to her or his mother and focus on an activity at the same time since both inputs must be processed visually. In addition, the deaf child is unaware of sounds of the outside environment, and thus, is centred on self and own activities. This has consequences on the child's development of language, social skills and cognition. [1].

Sound is a biometric feature used to difference people or species from one another. By filtering an existing audio signal, it is possible to detect whether this sound comes from a human or another object. Since voices differ from one person to another, they can be used for voice accolade purposes. It is also possible to power or direct various devices through words obtained from sound signals. Therefore, the processing and use of an audio signal are very important. Hearing is a very important sensory task for people. Developing a device that can perceive and classify various voices at home, and thereby raise the quality of life for hearing impaired people, is regarded as a basic requirement. A fire alarm or a phone alert that warns against danger are some of the sounds that should be perceived to encourage urgent action.

People with hearing loss have strain in hearing and understanding speech. Despite significant advances in hearing aids and inner ear implants, these devices are frequently not enough to enable users to hear and understand what is being communicated in different settings. Even with the latest technology, hearing aids have a limited effective range, basically amplifying almost all sounds, and usually can't separate the background noise from the voices and sounds that the user actually wants to hear, making the users not much satisfied with the use of hearing aids. [2].

1.2. Causes for hearing loss

1.2.1. Aging

Deterioration of inner ear structures happens over time.

1.2.2. Loud noise

Subjection to loud sounds will harm the cells of your inner ear. Damage will occur with long-run subjection to loud noises, or from a brief blast of noise, such as from a gunfire.

1.2.3. Heredity

The genetic makeup may make one more susceptible to ear damage due to sound or deterioration from ageing.

1.2.4. Occupational noises

Jobs where loud noise is a regular part of the working environment, such as farming, construction or factory work, can lead to damage to the ear.

1.2.5. Recreational noises

Subjection to unstable noises, like from firearms and airplane, may cause immediate and permanent hearing impairment. Other distraction activities with hazardously high noise levels include motorcycling, listening to loud music.

1.2.6. Some medications

Drugs like sildenafil (Viagra), the antibiotic gentamicin and certain chemotherapy drugs, can damage the inner ear. Temporary effects on hearing ringing in the ear (tinnitus) or hearing loss may occur if very high doses of aspirin, alternative pain relievers, antimalarial drugs or loop diuretics are taken.

1.2.7. Some illnesses

Diseases or illnesses that end in high fever, such as meningitis, can harm the cochlea.

1.3. Problems faced by hearing impaired people

People who are hearing impaired face considerable challenges. They experience and navigate the world very differently compared to those who possess perfect hearing. To gain an understanding of the difficulties they may face, ten situations that make life more challenging are shown below.

1.3.1. Public announcements

Hearing impaired people can't grasp information transmitted through public address systems.

1.3.2. Slow talkers

When someone realizes they're interacting with a hearing-damaged person, they usually switch to a slower kind of speech. Hearing impaired will have learned to understand words when they are spoken naturally, so slowing it down intentionally may result in miscommunication.

1.3.3. Being in the dark

The absence of light in the surrounding environment makes it difficult for hearing impaired people to interact with others. They generally rely on visual stimuli, such as lip-reading or sign language, so dark environments present a challenge.

1.3.4. Relying on touch

When a person is hearing impaired, they won't hear their name called. That's why in deaf society, firm but civil tapping on the shoulder is normal in order to gain observation. However, those not familiar with the hearing impaired community may be unaware of this, leading to confrontation.

1.3.5. Sign language misunderstandings

Sign language is not universal, and different standards are present in different countries. In addition, regional areas have their own specific variations similar to accents or slang causing additional difficulty. There are many instances of professional interpreters using the wrong words due to the variations across regions and countries; while this may not seem like a big deal, it has led to lasting harm, such as in legal situations or accidents during hospital visits.

1.4. Comparing loudness of common sounds

Table - 1: Comparing loudness of common Sounds

Decibels	Noise Source
30	Whisper
40	Refrigerator
60	Normal conversation
75	Dishwasher
85	Heavy city traffic, School cafeteria
95	Motorcycle
100	Snowmobile
110	Chain saw
115	Sandblasting
120	Ambulance
140 - 165	Firecracker

2. DESCRIPTION

2.1. Introduction to the proposed methodology

The project is aimed at the development of a wearable assistive device for the hearing impaired that informs the users of important sounds through vibrations, thereby understanding what kind of sound it is. The classical Mel-Frequency Cepstral Coefficients (MFCC) is used for feature Extraction. The Hidden Markov Model (HMM) approach, Gaussian HMM and GMM-HMM are used to classify the environmental sounds. [3].

In recent years, research, especially on speech data, has gradually increased to meet the demands of the developing world. This is because speech data such as audio and voice data are those that are closest to human life, and they can best express daily life. Recently, study on audio-based systems and deep-learning technologies has improved rapidly. Examples of sound sources are traffic noise. The output data from the sound source models are therefore further processed by a hierarchical HMM in order to determine the current listening environment.

An acoustic thumbprint is a condensed digital summary, a thumbprint, deterministically generated from an sound signal, that can be used to identify an sound sample or quickly locate similar items in an sound database. A robust acoustic fingerprint algorithm must take into account the perceptual characteristics of the sound. If two files sound alike to the human ear, their acoustic thumbprints should match, even if their binary representations are quite different. Such sound fingerprinting techniques can be used to identify different voices.

The classified sound signals can be categorised into classes depending on the nature of sound and the importance of the sound. Then, the vibration motor sends a warning to the user, which is perceived through the sense of touch. For each type of sound, a different vibration stimulus is transmitted to the hearing-impaired person so that different sounds can be identified. Thus, this type of wearable device would be beneficial for the aged, disabled, and hearing impaired patients who can understand immediately the type of repetitively occurring sounds in their daily life so easily by means of vibrations which they feel in their skin. [4].

2.2. Dataset description

The dataset consists of 1091 audio files annotated with 5 labels. All audio samples are provided as uncompressed PCM 16-bit, 44.1 kHz mono audio files. The categories are: “Applause”, “Bark”, “Knock”, “Laughter”, “Telephone”. The dataset is created by extracting and restructuring data from the FSDKaggle2018 dataset and the Sound Event Classification(Kaggle) dataset.

2.3. Block diagram

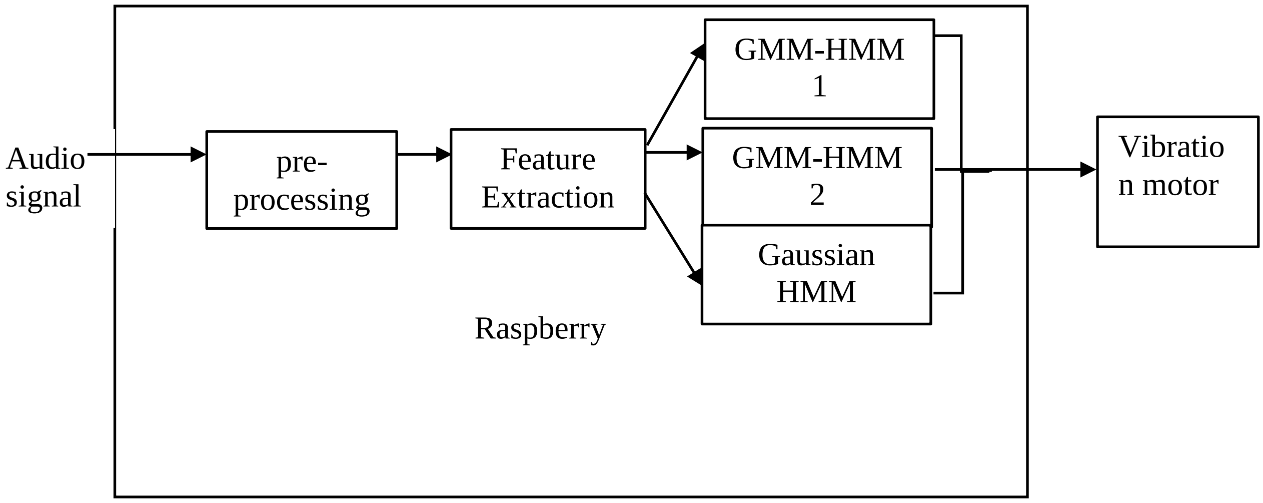

Fig -1: Block diagram

The audio signal is observed by the microphone and sent to the Raspberry Pi where pre-processing, feature extraction and classification processes occur. The output is given to the vibration motor corresponding to the predicted class.

2.4. Preprocessing

The audio signals received are of unequal lengths making classification difficult. So, the audio signals are converted to exactly 7 seconds in length with shorter audio signals repeated until it reaches the 7 second mark and longer audio signals cut short.

2.5. Feature extraction

Feature Extraction is performed using Mel Frequency Cepstral Coefficients (MFCC) technique. It represents the audio signal with as a set of 39 cepstral coefficients which concisely describe the overall shape of the spectral envelope. The frequency bands are equally spaced on the mel scale, which approximates the response of the human auditory system more closely compared to the linearly-spaced frequency bands in the normal spectrum. This frequency warping allows for better description of audio.

2.6. Classification

Classification scores are simultaneously calculated using a Gaussian HMM model and 2 GMM-HMM models. The overall maximum score among the 3 models is found and the class corresponding to the score is the final prediction.

2.6.1. Gaussian HMM

The Gaussian hidden Markov model (Gaussian HMM) is a type of finite-state-space and homogeneous HMM where the measurement probability distribution is the normal distribution,

$$Y_t|S_t \sim N(\mu S_t, \Sigma S_t) \quad ......\ (2.1)$$

where μS_t and ΣS_t are mean and covariance parameters at state S_t , $S_t = 1,\ldots,K$. Hence, the initial state probability vector (ISPV) π , the transition probability matrix (TPM) A, and the observation parameter $B(\equiv \{\mu_I , \Sigma_i\}_{i=1,\ldots,K}$, which consists of mean and covariance parameters) together specify the Gaussian HMM; that is, the parameter θ of the Gaussian HMM is $\{\pi, A, B\}$.

Because the Gaussian HMM is a type of finite-state-space and homogeneous HMM, the six common problems like filtering, smoothing, forecasting, evaluating, decoding, and learning problems can be solved using the three algorithms found in the section Hidden Markov Model.

2.6.2. GMM-HMM

The GMM can be observed as a mixture between parametric and non- parametric density models. Like a parametric model, it has structure and parameters that power the actions of density in known ways. Like non-parametric model it has many degrees of release to allow arbitrary density modelling. The GMM density is given as weighted sum of Gaussian densities:

$$P_{GM}(x) = \sum_{m=1}^{M} W_m g(x , \mu_m , C_m) \quad ......\ (2.2)$$

W_m are the component probabilities or weights ($\sum W_m = 1$). The K-dimensional densities so the argument is a vector $x = (x_1, \ldots , x_K)T$. The $g(x, \mu_m, C_m)$ is a K-dimensional Gaussian Probability Density Function (PDF).

$$g(x , \mu_m , C_m) = (1/(2\pi)^{K/2} |C_m|^{1/2}) e^{-1/2 (x-\mu_m)^T C_m^{-1} (x-\mu_m)} \quad ......\ (2.3)$$

where C_m denotes the covariance matrix and μ_m denotes the mean vector. Now, a Gaussian mixture model probability density function is completely given by a parameter list by,

$$\theta = \{w_1, \mu_1, C_1 \ldots w_1, \mu_1, C_1\}m ; \quad m=1\ldots M \quad ......\ (2.4)$$

Sorting the data for input to the GMM is weight since the components of GMM play a major role in the creation of word models. For this purpose, The K- Means clustering technique is used to break the data into 256 cluster centroids. These centroids are then collected into sets of 32 and then passed into each component of GMM. As a result a set of 8 components are obtained for GMM. Once the component inputs are decided, the GMM modelling can be implemented. [5].

The GMM/HMM hybrid model has the power to find the joint maximum probability among all viable reference words W given the sequence O. In real case, the union of the GMMs and the HMMs with a weighted coefficient may be a good scheme because of the variance in training methods.

3. REQUIREMENTS

3.1. Software requirements

Table - 2: Software Requirement for the system.

Operating system	Windows/Linux/MacOS
IDE	Jupyter Notebook for python IDE

3.2. Hardware requirements

- Laptop with minimum 8GB RAM and installed with above mentioned software.
- Raspberry Pi 4
- Vibration Motor
- Microphone
- Connecting wires

4. COMPONENT DESCRIPTION

4.1. Raspberry PI 4

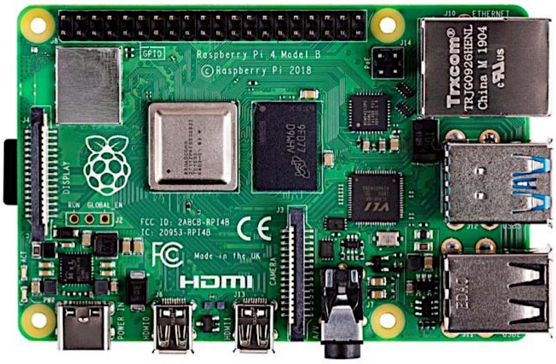

Fig -2: Raspberry Pi 4

Raspberry Pi 4 Model B is the new product in the popular Raspberry Pi series of computers. It offers incredible increases in processor speed, multimedia performance, memory, and connectivity compared to the previous generation Raspberry Pi 3 Model B+, while still maintaining backwards compatibility and having similar power consumption. For the end user, Pi 4 Model B gives desktop performance equal to the entry-level x86 PC systems.

The Raspberry Pi 4 Model B (Pi4B) is the first of the next generation of Raspberry Pi computers supporting additional RAM and having greatly improved GPU, CPU and I/O performance all within a similar power envelope, form factor and cost as the previous generation Raspberry Pi 3B+. The Pi4B is obtainable with 1, 2 and 4 GB of LPDDR4 SDRAM.

This product's key details include an improved performance 64-bit quad-core processor, dual-display with resolutions up to 4K via a pair of micro-HDMI ports, hardware video decode at up to 4Kp60, up to 4GB of RAM, dual-band of 2.4 and 5.0 GHz wireless LAN, Bluetooth 5.0, Gigabit Ethernet, USB 3.0.

The dual-band wireless LAN and Bluetooth have modular compliance certification, allowing the board to be designed into end products with significantly reduced compliance testing, thus improving both cost and time to market.

4.1.1. Raspberry PI hardware

- Possesses Quad core 64-bit ARM-Cortex A72 running at frequency of 1.5GHz
- Different option possessing 1, 2 and 4 Gigabyte LPDDR4 RAM
- H.265(high efficiency video coding) hardware decoding (up to 4Kp60)
- H.264 hardware decode (up to 1080p60)
- Video Core VI 3D Graphics
- Dual HDMI display output up to 4Kp60 is supported

4.1.2. Raspberry PI software

- ARMv8 Instruction Set
- Mature Linux software stack
- Actively developed and maintained

  ⇒ Recent Linux kernel support
  ⇒ Many drivers up streamed
  ⇒ Stable and well supported userland
  ⇒ GPU functions using standard APIs are available

4.1.3. Interfaces

- 802.11 b/g/n/ac Wireless LAN
- Bluetooth 5.0 with BLE
- 1x SD Card
- 2x micro-HDMI ports which support dual displays up to 4Kp60 resolution
- 2x USB2 ports
- 2x USB3 ports
- 1x Gigabit Ethernet port (supporting PoE with add-on PoE HAT)
- 1x Raspberry Pi port for camera (2-lane MIPI CSI)
- 1x Raspberry Pi port for display (2-lane MIPI DSI)
- 28x user GPIO pins where various interface options are supported:

  ⇒ Up to 6x UART
  ⇒ Up to 6x I2C
  ⇒ Up to 5x SPI
  ⇒ 1x SDIO interface
  ⇒ 1x DPI (Parallel RGB Display)
  ⇒ 1x PCM
  ⇒ Up to 2x PWM channels
  ⇒ Up to 3x GPCLK outputs

4.1.4. Peripherals

4.1.4.1. GPIO interface

The Pi4B has 28 BCM2711 GPIOs available through a standard Raspberry Pi 40-pin header. It has backwards compatibility with all previous Raspberry Pi boards with a 40-way header.

4.1.4.2. GPIO pin assignments

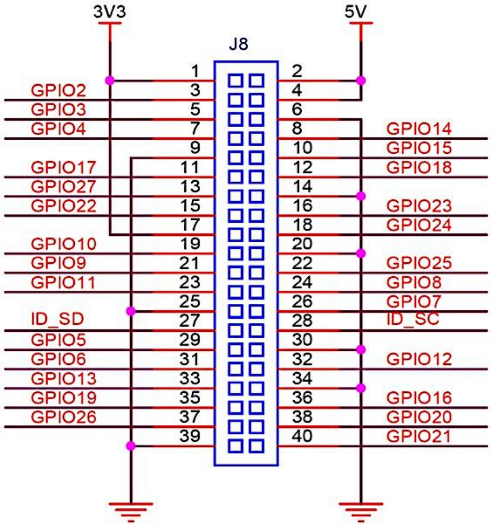

Fig - 3: GPIO Connector Pinout

Other than being capable of being used as straightforward software controlled input and output (with programmable pulls), GPIO pins can be multiplexed into various other modes backed by dedicated peripheral blocks such as I2C, UART and SPI.

Extra I2C, UART and SPI peripherals have been added to the BCM2711 chip in addition to the standard peripheral options found on legacy Pis and are available as further mux options on the Pi4. Much more flexibility is given to users when attaching add-on hardware when compared to older models.

4.1.4.3. Display Parallel Interface (DPI)

A standard parallel RGB (DPI) interface is available to the GPIOs. This can support a secondary display up-to a 24-bit parallel interface.

4.1.4.4. SD/SDIO Interafce

The Pi4B has a dedicated SD card socket which supports 1.8V, DDR50 mode (at a peak bandwidth of 50 Megabytes / sec). A legacy SDIO interface is also available on the GPIO pins.

4.1.5. Camera and display interfaces

The Pi4B has one Raspberry Pi 2-lane MIPI CSI Camera and one Raspberry Pi 2-lane MIPI DSI Display connector. These connectors have backwards compatibility with legacy Raspberry Pi boards, and provide support to all of the available Raspberry Pi camera and display peripherals.

4.1.6. USB

The Pi4B has two USB2 and two USB3 type-A sockets. Downstream USB current is limited to approximately 1.1A in total over the four sockets.

4.1.7. Temperature range and Thermals

The recommended ambient operating temperature range is 0 to 50 degrees Celsius. The Pi4B reduces the CPU clock speed and voltage to reduce thermal output when idling or under light load. The speed and voltage (and hence thermal output) are increased during heavier load. The internal governor will throttle back both the CPU speed and voltage to ensure that the CPU temperature never exceeds 85 degrees Celsius.

The Pi4B can operate perfectly well without any extra cooling and is designed for sprint performance, that is, expecting a light use case on average and ramping up the CPU speed as needed. Further cooling may be needed if a user wishes to load the system continually or operate it at a high temperature at full performance.

4.1.8. HDMI

The Pi4B has two micro-HDMI ports, both of which support CEC and HDMI 2.0, with resolutions up to 4Kp60.

4.1.9. Audio and Composite (TV OUT)

The Pi4B supports near-CD-quality analogue audio output and composite TV-output through use of a 4-ring TRS 'A/V' jack. The analog audio output can be used to drive 32 Ohm headphones directly.

4.1.10. Power requirements

A good quality USB-C power supply capable of delivering 5V at 3A is required. If the attached downstream USB devices consume less than 500mA, a 5V, 2.5A supply can be used.

Table - 3: Absolute Maximum Ratings

Symbol	Parameter	Minimum	Maximum	Unit
VIN	5V Input Voltage	-0.5	6.0	V

4.2. Mobile phone vibration motor

The Mobile Phone Vibration Motor is a shaftless vibration motor that is fully enclosed with no exposed moving parts. It has a small size of10 mm diameter and3.4 mm height and can be mounted onto a PCB or placed into a pocket to add quiet, haptic feedback to any project. [6].

The motor contains a 3M adhesive backing on it for easy mounting and 1.5 inch leads for making quick connections.

This small, button-type vibrating motor vibrates with a vibration amplitude of 0.75g and draws approximately 60 mA when 3V is applied to its leads.

Fig - 4: Vibration Motor

4.2.1. Mobile phone vibration motor feature

- Rated voltage: DC 3.0V
- Working Voltage: DC 2.5V – 4.0V
- Working Temperature: -20^{0}C – 60^{0}C
- Minimum rated rotate speed: 9000RPM
- Maximum Rated current: 90mA
- Maximum starting current: 120mA
- Starting Voltage: DC 2.3V
- Insulation Resistance: 10MΩ
- Terminal impedance: 31Ω ± 15% (single posture), 59Ω ± 15% (double posture)
- Cable length: 20mm

4.3. Microphone

A mike is a device that used to translates sound vibrations within the air into electronic signals and scribes them to a recording medium or over a speaker unit. Microphones are used in many types of audio recording devices for various purposes including communications of many kinds, as well as music vocals, speech and sound recording.

Fig - 5: Microphone

Some features of this microphone are:

- Omnidirectional mic with LED indicator
- Noise filter to obtain crisp and clear audio
- Frequency Response 100Hz to 10KHz
- Sample Rate: 44.1KHz
- Compatible OS:- Win XP/Vista/7/8/10
- Impedance: 2.2k

5. FEATURE EXTRACTION

5.1. Introduction

Mel-Frequency Cepstral Coefficients (MFCC) and their derivatives are used for feature extraction techniques. This cepstral coefficient is derived in terms of mean and correlation coefficient using MFCC and its derivatives. The pre-processing, feature extraction technique, selection of a feature, feature reduction as well as classification of features are used for calculating the efficiency. MFCC is the feature extraction technique used in this project.

The first step in any automatic speech recognition system is to extract options i.e. identify the parts of the audio signal that are good for recognizing the linguistic content and removing all the unimportant stuff which carries information like emotion, background noise etc. The most purpose to know concerning speech is that the sounds generated by a personality's square measures filtered by the form of the vocal tract together with the tongue, teeth etc. This shape determines what sound comes out. If one can determine the shape accurately, this should give an accurate representation of the phoneme being produced. The shape of the vocal tract is shown in the envelope of the short-time power spectrum, MFCCs are required to correctly represent this envelope.

Mel Frequency Cepstral Coefficients (MFCCs) are widely used automatic speech recognition (ASR). Before the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) were the predominant feature type for ASR, especially with HMM classifiers.

5.2. Steps involved in MFCC

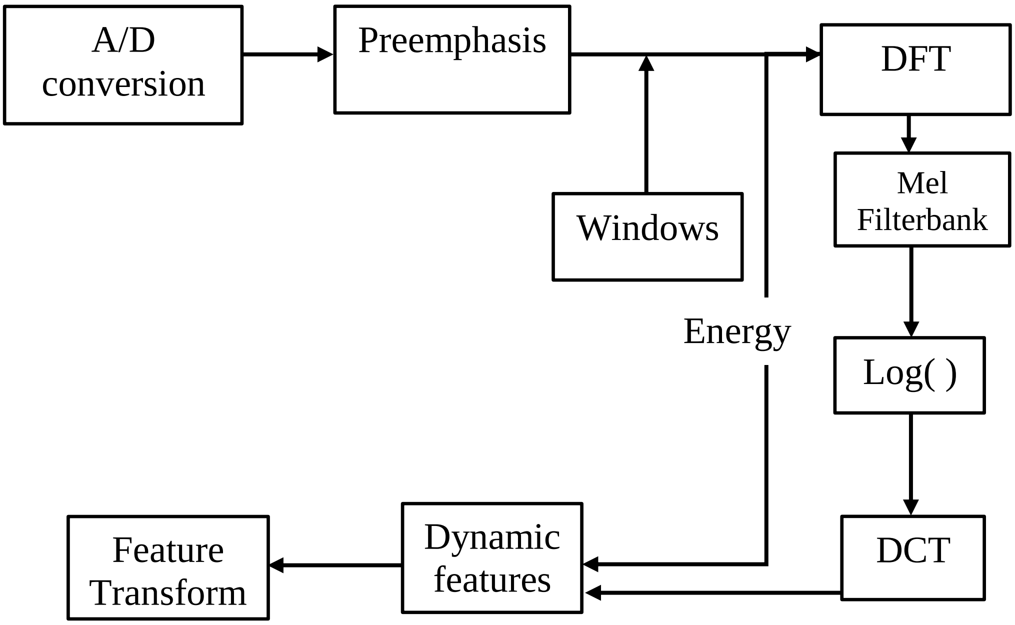

Fig - 6: Steps involved in MFCC

5.2.1. A/D conversion

In this step, the conversion of audio signal from analog to digital format with a sampling frequency of 8kHz or 16kHz is performed.

5.2.2. Pre-emphasis

Pre-emphasis significantly increases the magnitude of energy in the higher frequencies. On observing 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 phoneme detection accuracy thereby improving the performance of the model. The first order high-pass filter is used for pre-emphasis.

5.2.3. Windowing

The MFCC technique aims to develop the features from the audio signal which can be used for detecting the phonemes in the speech. But in the given audio signal there will be many phonemes, so the audio signal is broken into different segments with each segment having 25ms width and with the signal at 10ms apart. On an average, a person speaks three words per second with 4 phonemes and each phoneme will have three states resulting in 36 states per second or 28ms per state which is close to our 25ms window. From each segment, 39 features are extracted. Moreover, while breaking the signal, if the signal is chopped directly at the edges of the signal, the sudden fall in amplitude at the edges will produce noise in the high-frequency domain. So instead of a rectangular window, Hanning windows are used to chop the signal which won't produce the noise in the high-frequency region.

Framing is the method of blocking the audio signal into frames of "n" samples in the time domain. After the framing step, each individual frame is windowed using the window function. The window function is a mathematical function that is used to minimize signal discontinuities at the beginning and at the end of each frame by taking the block of the next frame in consideration and integrating all closet frequency lines. This step makes the end of each frame connects smoothly with the beginning of the next frame after the window function is applied. [7].

5.2.4. Discrete Fourier Transform (DFT)

The signal is converted from the time domain to the frequency domain by applying the DFT transform. Analysing audio signals is easier in the frequency domain than in the time domain.

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi nk/N} \ ; 0 \le k \le N-1 \quad ...... (6.1)$$

5.2.5. Mel filter bank

The perception of sound by human ears is different from that of machines. Human ears have higher resolution at lower frequencies compared to higher frequencies. So if a sound is heard at 200 Hz and 300 Hz, it is easily differentiated by humans when compared to the sounds at 1500 Hz and 1600 Hz even though the difference is the same. However, machines have the same resolution at all frequencies. It is noticed that modelling the human hearing property at the feature extraction stage will improve the performance of the model. So, the Mel scale is used to map the actual frequency to the frequency that human beings will perceive.

$$Mel(f) = 1127ln(1+f/700) \quad ...... (6.2)$$

Filter banks can be implemented in both the time and frequency domains. To compute the MFCC coefficients, the filter banks are usually implemented in the frequency domain. The centre frequencies of the filters are normally evenly spaced on the frequency axis.

5.2.6. Applying log

Humans are less sensitive to changes 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 a high value of input gradient value is less. So the log is applied to the output of Mel-filter to mimic the human hearing system.

5.2.7. Discrete Cosine Transform (DCT)

The vocal tract is smooth, therefore, the energy levels in adjacent bands tend to be correlated. The DCT is applied to the transformed Mel frequency coefficients and produces a set of cepstral coefficients. The Mel spectrum is usually represented on a log scale before computing DCT. The output is a signal in the cepstral domain with a que-frequency peak corresponding to the pitch of the audio signal and a number of formants indicating low que-frequency peaks. Since most of the audio signal information is represented by the first few MFCC coefficients, the system can be made robust by ignoring or truncating higher order DCT components and extracting only those first few coefficients.

$$C(n) = \sum_{m=0}^{M-1} log_{10}\,(s(m))\cos((\pi n(m-0.5))/M) \ ; \ n = 0, 1, 2,....., C-1 \qquad ...... (6.3)$$

Where, s(m) is log-energy at the output of each filter
M is the number of filter outputs.

5.2.8. Dynamic MFCC features

Cepstral coefficients are called as static features since they only contain information from a given frame. Further information regarding the temporal dynamics of the audio signal is obtained by computing the first and second order derivatives of the cepstral coefficients. The first-order derivatives are known as delta coefficients, and the second-order derivatives are known as delta-delta coefficients. Delta coefficients tell about the speech rate, and delta-delta coefficients and the second-order derivative is called delta-delta coefficients. Delta coefficients indicate the audio rate, and delta-delta coefficients indicate the acceleration of the audio signal.

$$\Delta C_m(n) = (\sum_{i=-T}^{T} k_i C_m(n+i)) / (\sum_{i=-T}^{T} |i|) \quad ...... (6.4)$$

where $C_m(n)$ indicates the mth feature for the nth time frame, T is the number of successive frames used for computation, and k_i is the ith weight. Generally, T is taken as 2. Taking the first-order derivative of the delta coefficients gives the delta-delta coefficients. Therefore, the MFCC technique will generate 39 features from each audio signal sample. These features are used as input for the audio recognition model.

6. HIDDEN MARKOV MODEL

6.1. Introduction

Hidden Markov Model (HMM) is a statistical model where the modelled system is assumed to be a Markov process with hidden states. It has a set of states each of which has limited number of transitions and emissions.

6.2. Markov chains

The HMM is dependent on augmenting the Markov chain. A Markov chain is a model provides information about the probabilities of sequences of random variables and states, each of which can take on values from a particular set. The set may be words, or tags, or symbols representing something such as the weather. A Markov chain makes a very strong assumption that to predict the future in the sequence, all that matters is the current state. The states before the current state have no impact on the future except via the current state.[1]

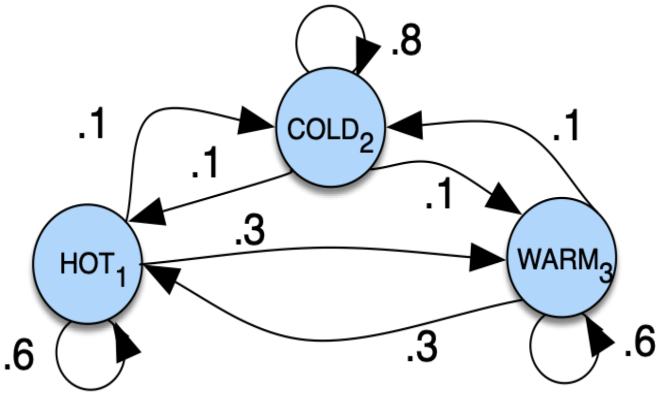

Fig - 7: A Markov chain for weather showing states and transitions.

Consider a sequence of state variables $q_1,q_2,...,q_i$. A Markov model incorporates the Markov assumption on the probabilities of this sequence: when predicting the future, the past doesn't matter, only the present.

$$\text{Markov Assumption: } P(q_i = a|q_1...q_{i-1}) = P(q_i = a|q_{i-1}) \quad ...... (6.1)$$

The nodes in the graph represent various states and the edges indicate transitions along with their probabilities. Since the transitions are probabilities, the sum of values of arcs leaving a given state must be equal to 1 .

6.3. The Hidden Markov Model

A Markov chain is generally useful when there is a need to calculate the probability for a sequence of observable events. However, in many cases, the events that one is interested in are hidden: they cannot be observed directly. For example, part-of-speech tags in a text can't be observed directly. Rather, the words are seen the tags need to be inferred from the word sequence. Thus the tags can be called as hidden because they are not observed.

A hidden Markov model (HMM) incorporates both observed events (like words that are seen in the input) and hidden events (like part-of-speech tags) that can be thought as causal factors in a probabilistic model.

Two simplifying assumptions are used by a first order Hidden Markov Model.

First, same as a first-order Markov chain, the probability of a particular state is dependent only on the previous state:

$$\text{Markov Assumption: } P(q_i|q_1...q_{i-1}) = P(q_i|q_{i-1}) \text{ ...... } (6.2)$$

Second, the probability of the output observation o_i is dependent only on the state that produced that observation, q_i and not on any other states or observations.

$$\text{Output Independence: } P(o_i|q_1 ...q_i,...,q_T ,o_1,...,o_i,...,o_T) = P(o_i|q_i) \text{ ...... } (6.3)$$

Given a sequence of observations O (where each integer represents the number of ice creams eaten on a given day) , to find the 'hidden' sequence Q of weather states (H or C) which caused Tom to eat the ice cream.

The two hidden states (H and C) indicate hot and cold weather, and the observations (drawn from the alphabet O = {1, 2, 3}) indicate the number of ice creams eaten by Tom on a given day.

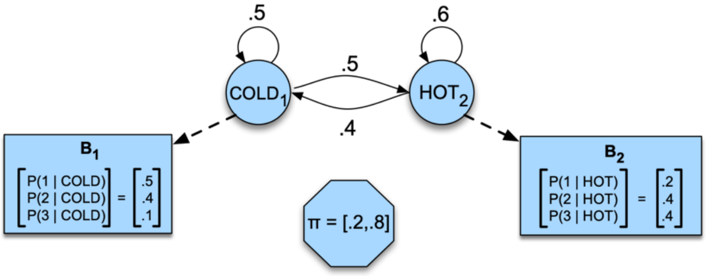

Fig - 8: Sample HMM for the ice cream task.

7. RESULTS

7.1. Performance metrics

The performance of the 3 models are analysed using 2 tools:

- Classification report
- Confusion matrix

7.1.1. Classification report

It is sone of the performance evaluation metrics used for a classification-based machine learning model. It shows a model's precision, F1 score, recall and support. It gives a better understanding of the overall performance of a trained model.

7.1.2. Confusion matrix

A confusion matrix or an error matrix is a specific table layout that enables visualization of the performance of the. model. Each row of the matrix represents the instances of an actual class while each column represents the instances of a predicted class, or vice versa. It visualises and summarises the performance of the model.

7.1.3. True Positive (TP)

These are cases where the predicted class was correct.

7.1.4. True Negative (TN)

These are cases where the class was correctly not predicted.

7.1.5. False Positive (FP)

These are cases where the class was incorrectly predicted.

7.1.6. False Negative (FN)

These are cases where the actual class was not predicted.

7.1.7. Accuracy

It is defined as the ratio of sum of true positives and true negatives to the total number of predictions. It shows how often the classifier prediction is correct.

$$\text{Accuracy} = (TP+TN)/\text{total} \quad \ldots\ldots (7.1)$$

7.1.8. Precision

Precision can be defined as the ratio of true positives to the sum of true and false positives. Out of all the positive predicted, it shows what percentage is truly positive.

$$\text{Precision} = TP/(TP+FP) \quad \ldots\ldots (7.2)$$

7.1.9. Recall

Recall can be defined as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = TP/(TP+FN) \quad \ldots\ldots (7.3)$$

7.1.10. F1 score

It is calculated as the harmonic mean of recall and precision. Both false negatives and false positives are taken into account. Thus, it performs well on an imbalanced dataset.

$$\text{F1 Score} = 2*(\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad \ldots\ldots (7.4)$$

7.1.11. Support

Support shows the number of actual occurrences of a class in the dataset. It doesn’t differ between models.
The performance of 3 different HMM models on MFCC features produced from audio data are studied and **the classification reports and the confusion matrices for the Gaussian HMM Model and the GMM-HMM models are given below:**

7.2. Gaussian HMM model

In the Gaussian HMM model, the number of components is 7.

7.2.1. Classification report

Table - 4: Gaussian HMM Model Classification Report.

	Precision	Recall	F1-score	Support
Applause	0.83	0.96	0.89	25
Bark	0.70	0.70	0.70	27
Knock	0.83	0.87	0.85	39
Laughter	0.81	0.65	0.72	26
Telephone	0.73	0.70	0.71	23
Accuracy			0.79	140
Macro Avg	0.78	0.78	0.78	140
Weighted Avg	0.78	0.79	0.78	140

It can be seen that Precision is highest for Applause and Knock and lowest for Bark.
For the F1- score, Applause being the highest and Bark the lowest with 0.89
0.70 respectively.

7.2.2. Confusion matrix

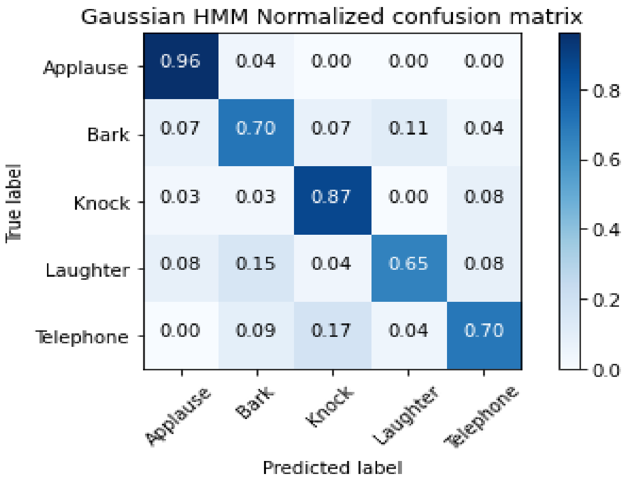

Fig - 9: Gaussian HMM Model Confusion Matrix.

While comparing Recall for various classes, Applause is having the highest value of 0.96 followed by Knock with 0.87, Laughter being the lowest with 0.65

7.3. GMM-HMM Model-1

In this GMM-HMM Model, the number of components) is 7 and number of mixtures is 5.

7.3.1. Classification report

Table - 5: GMM-HMM Model-1 Classification Report.

	Precision	Recall	F1-score	Support
Applause	**0.88**	**0.92**	**0.90**	**25**
Bark	**0.70**	**0.70**	**0.70**	**27**
Knock	**0.82**	**0.85**	**0.84**	**39**
Laughter	**0.82**	**0.69**	**0.75**	**26**
Telephone	**0.72**	**0.78**	**0.75**	**23**
Accuracy			**0.79**	**140**
Macro Avg	**0.79**	**0.79**	**0.79**	**140**
Weighted Avg	**0.79**	**0.79**	**0.79**	**140**

It can be seen that Precision is highest for Applause and Knock and lowest for Bark.
For the F1- score, Applause is the highest and Bark the lowest with 0.90 and 0.70 respectively.

7.3.2. Confusion matrix

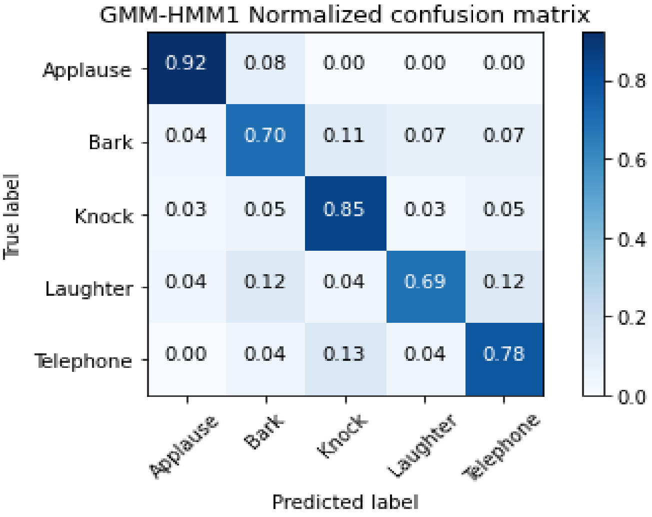

Fig - 10: GMM-HMM Model-1 Confusion Matrix.

While comparing Recall for various classes, Applause is having the highest value of 0.92 followed by Knock with 0.85, Laughter is the lowest with 0.69.

7.4. GMM-HMM model-2

In this GMM-HMM Model, the number of components is 4 and number of mixtures is 3.

7.4.1. Classification report

Table - 6: GMM-HMM Model-2 Classification Report.

	Precision	Recall	F1-score	Support
Applause	0.88	0.92	0.90	25
Bark	0.68	0.70	0.69	27
Knock	0.80	0.82	0.81	39
Laughter	0.86	0.69	0.77	26
Telephone	0.72	0.78	0.75	23
Accuracy			0.79	140
Macro Avg	0.79	0.78	0.78	140
Weighted Avg	0.79	0.79	0.79	140

It can be seen that Precision is highest for Applause and Laughter and lowest for Bark.
For the F1- score, Applause is the highest and Bark the lowest with 0.90 and 0.69 respectively.

7.4.2. Confusion matrix

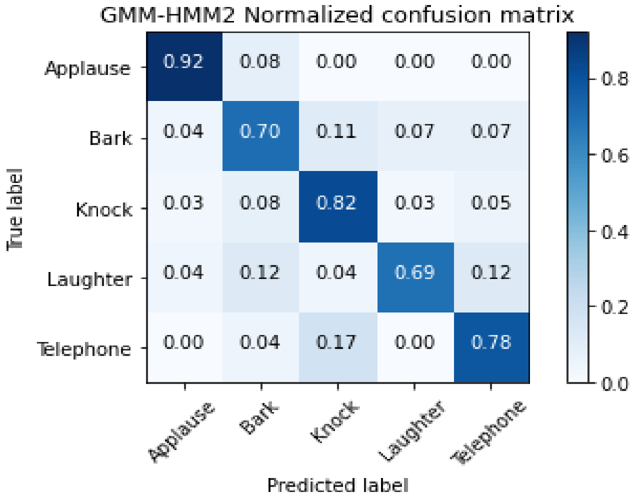

Fig - 11: GMM-HMM Model-2 Confusion Matrix.

While comparing Recall for various classes, Applause is having the highest value of 0.92, Laughter is the lowest with 0.69.

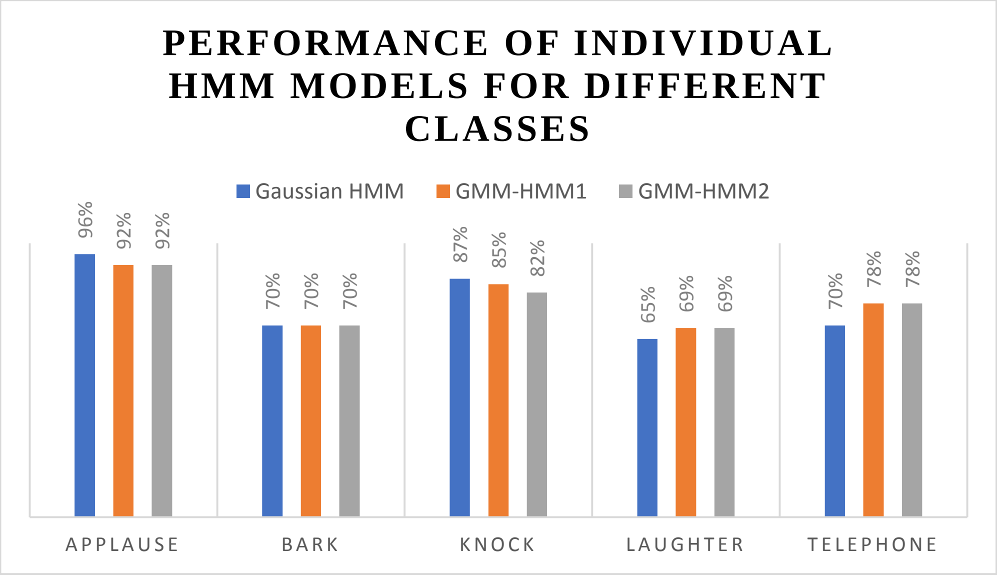

Fig - 12: Performance of Individual HMM Models for Different Classes.

Across the three models it is observed that the performance of the model is best for the class Applause with good performance for Knock as well. The less complex Gaussian HMM shows similar performance to GMM-HMM for the Applause, Bark and Knock classes with the biggest improvement of the GMM-HMM shown for the Telephone class. The two GMM-HMM models with different parameters show minimal difference in performance. All three models show the tendency to misclassify Telephone sounds as Knock sounds. The ensemble model combining these three models ensures that even if one of the models predicts the wrong class, the other two models' predictions are taken into account and so the right class is predicted more often.

8. CONCLUSION AND FUTURE WORK

In this study, the performance of 3 different HMM models on MFCC features produced from audio data are studied and compared and an improved HMM model is proposed by ensemble modelling the 3 models. The classification accuracy of the model drops significantly when deployed in a real-time scenario as compared to checking its performance on the Test dataset. This may be due to the external noise. Use of better noise suppression techniques as well as use of noisy audio similar to what would be found in the deployment environment would improve the performance. One of the main issues when dealing with audio datasets compared to image datasets is the lack of large quantity of data. Further data collection including collecting relevant audio data directly from the deployment environment to create custom datasets would improve the HMM model's performance as well as improve the performance of the deep learning approach. In future works, more classes covering all the major indoor sounds can be added for classification to provide comprehensive coverage of the indoor environment and to reduce misclassification of sounds not covered by this model. Furthermore, better feature extraction techniques including use of deep learning models such as Convolutional Neural Networks(CNN) on mel spectograms may improve the performance of the HMM.

REFERENCES

1. Anders Krogh, " Two methods for improving performance of an HMM and their application for gene finding", Center for biological sequence Analysis Technical University of Denmark, ISMB-97, 2018.

2. Ashish Castellino and Mohan Kameswaran, "Audio-Vestibular neurosensory Prosthetics: Origins, Expanding Indications and future directions, Prosthetics and Orthotics", Electronics, DOI: 10.5772/intechopen.95592 October 2021.

3. D. Clason, "New device helps hearing-impaired feel sounds in their environment", healthyhearing.com, November 23, 2020.

4. M. Yoganoglu, "Real time wearable speech recognition system for deaf persons," Computers & Electrical Engineering, vol. 91, pp. 107026, 2021. Doi:10.1016/j.compeleceng.2021.107026.

5. P. Nordqvist, "Sound Classification in Hearing Instruments," Doctoral thesis, Royal Institute of Technology, Stockholm, 2004.

6. Rashidul Hasan, Mustafa Jamil, Golam Rabbani, Saifur Rahman, Speaker identification using mel frequency cepstral coefficients.

7. Virender Kadyan, Archana Mantri, "Acoustic Features Optimization For Punjabi Automatic Speech Recognition System", Chitkara University, 2018.

8. Hirak Dipak Ghael and L. Solanki, "A Review Paper on Raspberry Pi and its Applications", ijaem.com, 2020.

9. Branko Balon, "Using Raspberry Pi Computers in Education", IEEE, 2019.

10. Latifur Rahman and S.A. Fattah, "Smart Glass for Awareness of Important Sound to People with Hearing Disability", IEEE, 2020.

11. Maraim Alnefaie, "Social and Communication Apps for the Deaf and Hearing Impaired", Talf university, 2017.

12. Mete Yaganoglu, "Wearable Vibration Based Computer Interaction and Communication System for Deaf", Applied sciences, 2017.

13. Cesar Lozano Diaz, "Augmented Reality System to Promote the Inclusion of Deaf people in Smart Cities", ISSN, 2019.

14. Linda Kozma spytex, "Factors Affecting the Accessibility of voice Telephony for people with Hearing Loss: Audio Encoding, Network Impairments, Video and Environmental Noise", Galloudet University, 2021.

15. Hemantha Kumar, "Continuous Speech segmentation and Recognition", electronics, 2018.

16. Jayabalan Kennedy, "Studies in hidden Markov models and related topics", Computer and electrical engineering, 2017.

17. M. Yoganoglu and C. Lose, "Real-time Detection of Important Sounds with a Wearable Vibration Based Device for Hearing-Impaired People," Electronics, vol. 7, pp. 50, 2018. Doi:10.3390/electronics7040050.

18. Luiz W. Biscainho, " Mobile Sound Recognition For the Deaf and Hard of Hearing",electronics,2017.