# CHAPTER 2
# LITERATURE REVIEW

This chapter presents the research works that are related to the speech recognition, gender classification and ML. It starts by describing the overview of ASR and benefits of speech-based applications. Characteristics of human speech are considered in this chapter. And the general review of feature extraction and gender classification techniques and concept of ML in computer science are also discussed.

## 2.1. Overview of Automatic Speech Recognition (ASR)

Speech recognition is an interdisciplinary subfield of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as ASR, computer speech recognition or speech to text (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields. Some speech recognition systems require "training" (also called "enrollment") where an individual speaker reads text or isolated vocabulary into the system. The system analyses the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent".

The term voice recognition or speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process [8].

From the technology perspective, speech recognition has a long history with several waves of major innovations. Computing power and artificial intelligent are largely behind the advances in this area. ASR by machine has been a field of research for more than 60 years. The industry has developed a broad range of commercial products where ASR as user interface has become ever more useful and pervasive. Consumer- centric applications increasingly require ASR to be robust to the full range of real-world noise and other acoustic distorting conditions. However, reliably recognizing spoken words in realistic acoustic environments is still a challenge.

### 2.1.1. Methodology of ASR Systems

Speech recognition aims at deriving the sequence of speech sounds that best match the input speech sound by using pattern recognition technology. On the other hand, speaker recognition is the process of automatically recognizing who is speaking, rather than what is being spoken. Basic ASR system operates in two distinct modes: training and recognition. Robust features of speech utterance are extracted both in training and recognition mode. These features contain information in time and frequency domain. The template generation module uses possible variations in the utterances of the same word to find the template of feature vectors for different words, during the training phase. In the recognition phase the feature vectors are extracted to be compared with stored templates for the words and a best match is found to recognize word [9].

ASR integrated with embedded technology is gaining popularity. Embedded ASR systems have been classified by software, hardware and combined hardware software design. ASR has been widely used in human–machine interaction, such as mobile robots, consumer electronics, manipulators in industrial assembly lines, automobile navigation systems, and security systems. The important highlights of ASR systems which make them preferable in various application areas are:

1. Users do not need a specialized skill, like typing, to use speech recognition systems. For most people, speech is an inherent skill that comes natural by and is cultivated from an early age,
2. Using speech is significantly faster than other forms of communication like typing or writing. A user can communicate with speech up to ten times faster than writing on paper,
3. ASR systems allow the use of multiple modalities, i.e., users can speak while doing other activities with their hands, legs, eyes, or ears,
4. The input methods of automatic speech recognition systems are economical. Specifically, microphones and telephones are very affordable.

### 2.1.2. Applications of ASR Systems

Research on automatic speech recognition began in the early fifties with the attempt to extract the significant features from acoustic data and to classify and recognize them, by using methodologies developed in the area of pattern recognition. Later, in the seventies ,the artificial intelligence technologies were applied for the design of speech understanding systems [10]. An important motivation for the research in this area is the attractive perspective of gathering a deeper understanding about the complex mechanism underlying human perception of spoken language and the characterization of speech sounds in terms of physically detectable features in the brain. The research oriented towards this goal is of great utility in psychology and in the field of development of hearing aids for the handicapped. Moreover research in speech understanding offers a good opportunity for investigating complex parallel processing systems capable of modeling human perception.

Further motivation for research in automatic speech recognition lies in its industrial application, to simplify the communication between humans and machines. Avoiding intermediate keying and handwritten steps, leads towards multiple task capability to communicate with the machine while hands or eyes can carry out other functions. Further speech input is inherently faster than other methods. Hence, automatic speech recognition systems are gaining industrial acceptance in various sectors ranging from defence and medical to consumer products. These systems are more cost effective and often more efficient in real world applications that need to command and control with reduced human efforts. The development of faster computational capabilities of processors has also promoted in cost effective applications of ASR systems useful for industries. Presently, ASR systems find a wide variety of applications in the following domains:

1. Medical Assistance,
2. Industrial Robotics,
3. Forensic and Law enforcement,
4. Defence & Aviation,
5. Telecommunication Industry,
6. Home Automation and security Access Control,
7. I.T. and Consumer Electronics.

### 2.1.3. Challenges of ASR Systems

Speech is the nature’s gift to the human being which contributes towards the intelligence and discrimination from rest of the animal kingdom. Taking into consideration technological aspects, speech recognition is the buzzword today, as communication and hands free computing are evolving day by day. Speech is a very important mode of the communication and interaction with the digital computer. Speech recognition along with the wide range of applicability in domain of computer science, medical science, psychology, sports, and neurology has many challenges while developing. Developing real time speech recognizer may hurdle from adverse environment to anatomy of the human body. It also involves linguistic aspects too. This article explores various challenges in developing a robust ASR system.

#### 2.1.3.1. Human understanding of speech

Human has their own knowledge base which is resulted from the reading, experiments, experiences, examination, situation, interaction and communication. They may hear more than the speaker speaks to them. While speaking speakers have its own language model of native language. Human may understand and interpret the words or sequence of words, they never heard before.

In automatic speech recognition, the annotated corpus and language model which provide the system with the limited grammatical structures is needed to develop. To increase the chance of pattern matching one can use the statistical models like Hidden Markov model. At the level of human understanding the knowledge can be represented by the exhaustive content. In ASR, the limited knowledge needs to build and can be trained accordingly. Vocabulary size plays a crucial role in speed of the system. Search time and resources get increase due large vocabulary. Limited vocabulary has the restricted application domain.

#### 2.1.3.2. Background noise

In a natural environment, human speaks the words and sentences which incorporate the background noise. The noise is unwanted information in the clearest speech signal. It may be an echo effect, another speaker, speaking in the background, playing instruments in the background and so on. The speech signals mixed with noise is difficult to recognize accurately. Noise contaminated speech signal gives rise to speech variability and recognition become difficult in real time events. Reverberation and noise elimination are most challenging tasks. A robust noise reduction algorithm can be deployed dynamically for this type of challenge.

#### 2.1.3.3. Continuous speech

The real time ASR system works with the continuous speech. Phonetic, syllabic and word level recognition is not complex as recognition performed in isolation, but when the sequence of words/sentence is spoken out recognition become more critical. The word boundary ambiguity is a difficult problem for ASR. One way to address this difficulty is to give the pauses while speaking so that adjacent words can be identified clearly, but it tends to loss naturalness in the speaking style and also it needs the training of the speaker while speaking for increased length of the sentences. Current continuous speech recognition (CSR) systems with large vocabulary are strictly based on the principles of statistical pattern recognition [11, 12].

#### 2.1.3.4. Gender of the speaker and anatomy of vocal tract

Men and women have different fundamental frequencies while speaking. It is because women have shorter vocal tract than men have. The fundamental frequency of a woman's voice is roughly two times higher than a man’s voice. The shape and length of the vocal cords, formation of cavities and size of the lungs may change over time and it may cause the speaker variability. Children have shorter vocal tract and vocal folds compared to adults. This results in higher positions of formants and fundamental frequency. The high fundamental frequency is reflected in a large distance between the harmonics, resulting in poor spectral resolution of voiced sounds. The difference in vocal tract size results in a non-linear increase of the formant frequencies. In order to reduce these effects, previous studies have focused on the acoustic analysis of children's speech [13]. This variability issue has been addressed by vocal tract length normalization [14] as well as spectral normalization [15].

#### 2.1.3.5. Speed of the speech

The speed while speaking may vary and depends upon the situations, physical stress. Rate-of-speech (ROS) is considered as an important factor which makes the mapping process between the acoustic signal and the phonetic categories more complex. Timing and acoustic realization of syllables are affected due in part to the limitations of the articulatory machinery, which may affect pronunciation through phoneme reductions, time compression/expansion, changes in the temporal patterns, as well as smaller-scale acoustic phonetic phenomena. Due to rate of speech, accuracy of performance has been degraded. [16].

## 2.2. Gender Classification

Gender contains a wide range of information regarding to the characteristics difference between male and female. Gender classification is to determine a person's gender, e.g., male or female, based on this person's biometric cues. There are a number of biometrics which may be used to classify gender such as the face, eyes, fingerprint and hand shape, speech etc. Automatic gender classification is receiving increasing attention, since gender carries rich and distinguished information concerning male and female social activities. The aim of gender classification is to recognize the gender of a person based on the characteristics that differentiate between masculinity and femininity. In the area of artificial intelligence, gender classification is considered to be one of the most important applications of pattern recognition method. The progress of gender classification research has driven many potential applications. For instance, a computer system with gender recognition functions has a wide range of applications in fundamental and applied research areas including: human-computer interaction (HCI), the security and surveillance industry demographic research, commercial development, and mobile application and video games [17].

### 2.2.1. Opportunities and Challenges of Gender Classification

In practice, gender classification is a two-class problem in which the given information is assigned as male or female. Gender classification is a relatively easy task for humans but a challenging task for machines. Human beings are often able to make accurate and fast decisions on gender through visual inspection. For example, on the most basic characteristic, gender is a relatively invariant aspect of faces. Humans can readily determine gender for most faces, and additional information from hairstyle, body shape, clothing, eyebrows, and posture will support the evidence gained from the visual image [18]. Acoustical differences between male and female voices have been discussed by various investigators. Besides, the differences between males and females in physical features, psychological and neural signals provide a dynamic cue for gender analysis. Previous psychological and neural studies [19], [20] indicated that different gender exhibits diversity in changing facial expressions and head movement when they are stimulated. In addition, according to physiological measurements including electroencephalograph (EEG) and deoxyribose nucleic acid (DNA), as well as the daily social information such as handwriting, blog, email, etc , males and females also reveal different features or properties for the machine to classify.

Some researchers have explored these aforementioned cues and made significant progress for gender classification. Some of these features are in the form of images and can be analyzed by image processing. Research on gender classification using facial images started at the beginning of the 1990s. In 1990, Golomb et al. [21] used a multi–layer neural network to classify gender based on human faces. They reported a gender classification error rate of 8.1% by using a Cottrell-style back-propagation image compression network. Since then, applications of gender classification were developed extensively in various domains, bringing the emergence of gender classification approaches, from which iris [22], hand shape [23] and eyebrows [24] were considered as features in the literature. In recent years, clinical signal processing techniques have been proposed to serve in the field of gender recognition using EEG [25] and electrocardiograph (ECG) [26] signals. Although some progress is reported, gender classification is still challenging work for the machine because of the variation in gender features, particularly; gender characteristics of the face may result in problems with automatic gender detection. Variation in gender information extraction occurs due to changes in illumination, pose, expression, age, and ethnicity. Similarly, during the image capturing process the factors of image quality like dithering, noise, and low resolution also make image analysis a challenging task. On the other hand, the choice of features is one of the most critical factors. The classification techniques are also affected by feature extraction and classification algorithms. In order to distinguish gender from a voice signal, a set of techniques have been employed to determine relevant features to be utilized for building a model from a training set. This model is useful for determining the gender (i.e., male or female) from a voice signal. Machine learning algorithms give promising results for classification problem in all the research domains.

### 2.2.2. Gender Classification Techniques

Gender recognition is very difficult job started from the last decade. Gender classification is a two class problem (Male or Female) in which the given face image or voice is assign to any one of the class. It is an easy task for humans to classify gender but challenging task for machines. Over the period of time, automated classification of gender has gained enormous significance and has become an active area of research.  Many researchers have put a lot of effort and have produced quality research in this area. Still, there is an immense potential in this field because of its utility in many areas like monitoring, surveillance, commercial profiling and human-computer interaction. Security applications have utmost importance in this area. Gender classification can be used as a part of face recognition and speech recognition. Generally gender classification consists of three main steps: pre-processing, feature extraction and classification.

#### 2.2.2.1. Gender classification by face

Most the researcher focuses on classify gender using face images. Every face database needs some pre-processing, like normalization of illumination and face detection etc. After performing pre-processing, facial features are extracted. Extraction of different facial features is an important sub-task of gender classification. Gender classification approaches are categorized into two classes based on feature extraction. In this approach, global facial features and local facial features are used. Some facial points like face, nose, mouth and eyes are called local features and extracting local features are called geometric-based feature extraction. Some useful information may loss using geometric-based techniques. Next method is appearance-based feature extraction and it is also called global feature extraction. In appearance-based method, features are extracted from the whole face part instead of extracting features from facial points. Classification is the last step of gender classification in which the face is successfully classified as that of a male or female. For this purpose, different types of classifiers are used. E.g. K-nearest neighbor (KNN), neural network (NN) and support vector machine (SVM).

The different variation in gender face may cause to affect the Image processing techniques. Variation in face images occurs due to illumination change, poses, occlusions, age and ethnicity. The classification techniques are also affected by different masks on face (i.e. glasses, jewelry and hats). Similarly during the image capturing process the factors like blurring, noise and low resolution also make the face image analysis a challenging task. Similarly during the image capturing process the factors like blurring, noise and low resolution also make the face image analysis a challenging task [27].

#### 2.2.2.2. Gender classification by iris

The human iris is an annular part between the pupil and the white sclera. The iris has an extraordinary structure and includes many interlacing minute features such as freckles, coronas, stripes, furrows, crypts and so on. These visible features, generally called the texture of the iris, are unique to each individual. Previous work on gender classification from iris images has focused on handcrafted feature extraction methods using normalized NIR iris images.

The latest research proposed the use of the same pipeline that is used for iris recognition systems. The input image is segmented in a pre-process step. The iris region is then transformed to a polar space and codified using modified binary statistical image feature (MBSIF).

Several classification algorithms were used to test gender information from iris texture images. Those algorithms are:Adaboost M1, LogitBoost, GentleBoost, RobustBoost, LP-Boost, TotalBoost and RusBoost. Additionally, a Random Forest classifier with 500 trees, a Gini Index, and a LIB-SVM classifier with Gaussian Kernel (RBF) were also used. This is particularly important when large amounts of data need to be processed such as gender classification in highly populated countries (i.e. India, china) [28].

#### 2.2.2.3. Gender classification by fingerprint

A fingerprint is epidermis of finger consist of the pattern of interleaved ridges and valleys. The endpoints and bifurcation points of ridges are called minutiae. Fingerprint minutiae patterns of ridges are determined as unique through the combination of genetic and environment factors.

In the era of technology and science, gender classifications have immense value. Fingerprint based gender classification helps to analyze the data in easy way and help to sort out the data. Now a days finger print based gender classification is seen in civilian, industrial, commercial, and unique Id of nation as AADHAR ,a Hindi word meaning 'foundation' or 'base', card. As the fingerprints are unique, gender classification helps to minimize the large data. It is utilized by using simple scanner which is available in affordable prices.

The performance of minutiae extraction algorithms and fingerprint recognition technique relies heavily on the quality of the input fingerprint images. Since the fingerprint images acquired from sensors or other media are not assured with perfect quality. The importance of fingerprint enhancement algorithm is to improve the clarity of the ridge structures in the recoverable region and mark the unrecoverable region for further processing.

Only a Region of Interest (ROI) is useful to be recognized for each fingerprint image. The image area without effective ridges and valleys is first discarded since it only holds background information. And then minutia points are identified from finger print images. Fuzzy C- Means (FCM), Linear Discriminant Analysis (LDA), and Neural Network (NN) are generally used for the classification using the most dominant features [29].

#### 2.2.2.4. Gender classification by palm print

A palm print refers to an image acquired of the palm region of the hand. There are many differences between the male and female palm, such as palm color and texture. According to the study of Zhong et al. [30], there is a definite color difference between healthy men and women. For example, a male's palm usually shows a reddish color and dark partial state. Furthermore, the grip strength of a man is generally greater than that of women’s, so his palm print line will be deeper and will contain different texture features. Overall, it is feasible to classify gender by palm print imaging. Palm print gender classification can revolutionize the performance of authentication systems, reduce searching space and speed up matching rate.

A common method of gender classification is to first extract features, and then use feature vectors to train a classifier, which requires a combination of several features vectors to train a classifier, which requires a combination of several methods, such as, Fisher linear discriminant analysis [31] and principal component analysis (PCA) [32].

Recently, convolutional neural network (CNN) has achieved great performance in image classification [33, 34] after the pioneering work by Krizhevsky et al. [35]. Some studies [36, 37] used CNN for face gender classification and achieved promising results. Thus, it motivates researchers to apply CNN for palm print gender classification.

Compared to iris or fingerprint, palm print images have a large area containing rich information and can be easily collected. Furthermore, texture features are much more stable. The acquisition equipment is simple and the cost is much lower than iris recognition acquisition devices. Moreover, scientific studies have shown that biological characteristics of the hands are the least infringed when they are collected. These demonstrate that palm prints have an advantage over other biometrics [38].

#### 2.2.2.5. Gender classification by speech

Gender classification by speech analysis basically aims to predict the gender of the speaker by analyzing different parameters of the voice sample. This investigation mainly concentrates on short time analysis of the speech signals. The techniques used to process speech signals that can be broadly classified as either time domain or frequency-domain analysis. In time domain analysis, the measurements are performed directly on the speech signal to extract information. In frequency-domain analysis, the information is extracted after the frequency content of the speech signal computed to form the spectrum.

Gender Classifier from speech is a part of automatic speech recognition system to enhance speaker adaptability and a part of automatic speaker recognition system. The need for gender classification from speech also arises in several situations such as sorting telephone calls by gender for gender sensitive surveys. It is also a part of modern voice password technology [39].

There are numerous machine learning and deep learning models to classify whether the person is male or female, based on speech. Male and female voice acoustic features are commonly extracted by using Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficient (MFCC), Zero Crossings with Peak Amplitudes (ZCPA), Dynamic Time Warping (DTW) and Relative Spectral Processing (RASTA) for speech recognition system. Linear Discriminant Analysis (LDA), K-Nearest Neighbour (KNN), Classification and Regression Trees (CART), Random Forest (RF), and Support Vector Machine (SVM), Neural Network and so on is applied for classification from speech [40].

## 2.3. Human Speech Analysis

The human voice consists of sound made by a human being using the vocal tract, such as talking, singing, laughing, crying, screaming, shouting, yelling etc. The human voice frequency is specifically a part of human sound production in which the vocal folds (vocal cords) are the primary sound source. Generally speaking, the mechanism for generating the human voice can be subdivided into three parts; the lungs, the vocal folds within the larynx (voice box), and the articulators [41].

Each person has a unique voice because so many factors work together to produce that voice. A human’s voice starts down in the lungs, where air is exhaled to create an airstream in the trachea and across the larynx, which is often called the voice box. Stretched horizontally across the larynx are vocal folds, which are also known as vocal cords. As air passes over them, the vocal cords vibrate very quickly to produce sounds. The higher the rate of vibration, called frequency, the higher the pitch will be. The pitch of voice is largely determined by the length and tension of vocal cords. By themselves, the vocal cords produce just a buzzing sound. The parts of human body between the vocal cords and the outside world, such as the throat, nose, and mouth, act as a resonating chamber to turn those buzzing sounds into the unique human voice.

There are many different parts of the body involved in producing voice. Each of those parts is unique in each person, which is what gives each person a unique voice. Moreover, those parts can change over the years and even from day to day, so our voice itself can change over time and even day to day. For example, as boys grow older, their voices tend to deepen. This is because testosterone in boys makes their vocal cords longer and thicker. Girls' voices will change a bit as they grow older, too, but not to as great an extent as with boys. In general, men will have deeper voices than women, because their vocal cords tend to be larger and vibrate at lower frequencies. Moreover voice may change when people have a cold. The raspy voice common to people with colds is caused by swelling in the vocal cords caused by the cold virus. Coughing can also cause further irritation and swelling in the vocal cords.

Emotional communication is an important part of social interaction because it gives individuals valuable information about the state of others, allowing them to adjust their behaviors and responses appropriately. Emotions can also play a role in voice changing. When people get excited, nervous, or scared, the muscles around their larynx often tighten up, causing increased tension in vocal cords. That increased tension translates to the higher pitch when they're excited or stressed. Likewise, human voice can change from time to time as a result of anything that affects vocal cords, the larynx, or any of the other parts of the body that help to produce voice. Some of these factors include pollution, climate, smoking, and shouting or screaming too much [42].

### 2.3.1. Differences between Children and Adult Speech

It is well known that acoustic and linguistic characteristics of children’s speech are widely different from those of adult speech. For example, children’s speech is characterized by higher pitch and formants frequencies with respect to adults’ speech. Furthermore, characteristics of children’s speech vary rapidly as a function of age due to the anatomical and physiological changes occurring during a child’s growth and because children become more skilled in co-articulation with age. Much has been done in the past in analyzing the acoustic differences between children’s and adult speech, with a particular focus on vocal tract length and its influence on pitch and formant frequency values [43]. Understanding the developmental changes in children’s speech can help devise strategies for dealing with the acoustic mismatch between different age groups, for example in applications such as ASR and in early literacy and reading assessment [44]. Overall, the majority of the previous efforts in children’s speech analysis have dealt with vowel duration, pitch, and formants with little or no work on consonant analysis. The main difference between children and adult speech is the fundamental frequency response. For children, they have shorter vocal track and smaller vocal fold. So children utterance has higher fundamental frequency than adults.

In speech-development research, it is important to know how acoustic parameters of speech such as fundamental frequency, format frequencies, and segmental durations vary as a function of age and gender, and at what age the magnitude and variability of acoustic parameters begin to exhibit adult-like patterns. When properly interpreted, such chronological knowledge of speech acoustics could provide insights into the underlying development of speech organs and speech-motor control in children, and help in creating an accurate developmental model of the vocal tract.

In previous research works, changes in magnitude and variability of duration, fundamental frequency, formant frequencies, and spectral envelope of children's speech were investigated as a function of age and gender using data obtained from children of ages 5 to 17 years, and adults. The results confirmed that the reduction in magnitude and within-subject variability of both temporal and spectral acoustic parameters with age is a major trend associated with speech development in normal children. Between ages 9 and 12, both magnitude and variability of segmental durations decrease significantly and rapidly, converging to adult levels around the age of 12. Within-subject fundamental frequency and formant-frequency variability, however, may reach adult range about 2 or 3 years later. Differentiation of male and female fundamental frequency and formant frequency patterns begins at around the age of 11, becoming fully established around age 15. During that time period, changes in vowel formant frequencies of male speakers are approximately linear with age, while such a linear trend is less obvious for female speakers. These results support the hypothesis of uniform axial growth of the vocal tract for male speakers. The study also shows evidence for an apparent overshoot in acoustic parameter values, somewhere between ages 13 and 15, before converging to the canonical levels for adults. For instance, teenagers around age 14 differ from adults in that, on average, they show shorter segmental durations and exhibit less within-subject variability in durations, fundamental frequency, and spectral envelope measures [45].

### 2.3.2. Study of Speech Features Extraction

The extraction of the relevant and important information from the speech signals of the human voice is an important task to produce a latter recognition performance. The result efficiency of feature extraction step is crucial for the next step like modeling, classification and feature matching since it affects its behavior. Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCC), Zero

Crossings with Peak Amplitudes (ZCPA) and Discrete Wavelet Transform (DWT) are commonly used as feature extraction techniques for speech recognition system.

### 2.3.2.1. Linear prediction coefficients (LPC)

Linear prediction coefficients (LPC) imitate the human vocal tract and gives robust speech feature. It evaluates the speech signal by approximating the formants, getting rid of its effects from the speech signal and estimate the concentration and frequency of the left behind residue. The result states each sample of the signal as a direct incorporation of previous samples. The coefficients of the difference equation characterize the formants, thus, LPC needs to approximate these coefficients. LPC is a powerful speech analysis method and it has gained fame as a formant estimation method [46]. The frequencies where the resonant crests happen are called the formant frequencies. Thus, with this technique, the positions of the formants in a speech signal are predictable by calculating the linear predictive coefficients above a sliding window and finding the crests in the spectrum of the subsequent linear prediction filter [47]. LPC is helpful in the encoding of high quality speech at low bit rate.

Other features that can be deduced from LPC are linear predication cepstral coefficients (LPCC), log area ratio (LAR), reflection coefficients (RC), line spectral frequencies (LSF) and Arcus Sine Coefficients (ARCSIN). LPC is generally used for speech reconstruction. LPC method is generally applied in musical and electrical firms for creating mobile robots, in telephone firms, tonal analysis of violins and other string musical gadgets [48].

### 2.3.2.2. Mel frequency cepstral coefficients (MFCC)

A Mel frequency cepstral coefficients (MFCC) was originally suggested for identifying monosyllabic words in continuously spoken sentences but not for speaker identification. MFCC computation is a replication of the human hearing system intending to artificially implement the ear’s working principle with the assumption that the human ear is a reliable speaker recognizer. MFCC features are rooted in the recognized discrepancy of the human ear’s critical bandwidths with frequency filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to retain the phonetically vital properties of the speech signal. Speech signals commonly contain tones of varying frequencies, each tone with an actual frequency, f

(Hz) and the subjective pitch is computed on the Mel scale. The mel-frequency scale has linear frequency spacing below 1000 Hz and logarithmic spacing above 1000 Hz. Pitch of 1 kHz tone and 40 dB above the perceptual audible threshold is defined as 1000 mels, and used as reference point [49].

MFCC is based on signal disintegration with the help of a filter bank. The MFCC gives a discrete cosine transform (DCT) of a real logarithm of the short-term energy displayed on the Mel frequency scale. MFCC is used to identify airline reservation, numbers spoken into a telephone and voice recognition system for security purpose. Some modifications have been proposed to the basic MFCC algorithm for better robustness, such as by lifting the log-mel amplitudes to an appropriate power (around 2 or 3) before applying the DCT and reducing the impact of the low-energy parts [49].

#### 2.3.2.3. Discrete wavelet transform (DWT)

In numerical analysis and functional analysis, a discrete wavelet transform (DWT) is any wavelet transform for which the wavelets are discretely sampled. The discrete wavelet transform has a huge number of applications in science, engineering, and mathematics and computer science. Wavelet Transform (WT) theory is centered on signal analysis using varying scales in the time and frequency domains [50]. With the support of theoretical physicist Alex Grossmann, Jean Morlet introduced wavelet transform which permits high-frequency events identification with an enhanced temporal resolution [51]. A wavelet is a waveform of effectively limited duration that has an average value of zero. Many wavelets also display orthogonality, an ideal feature of compact signal representation . WT is a signal processing technique that can be used to represent real-life non-stationary signals with high efficiency [52]. It has the ability to mine information from the transient signals concurrently in both time and frequency domains [53].

Continuous wavelet transform (CWT) is used to split a continuous time function into wavelets. However, there is redundancy of information and a huge computational effort is required to calculate all likely scales and translations of CWT, thereby restricting its use. Discrete wavelet transform (DWT) is an extension of the WT that enhances the flexibility to the decomposition process [54]. DWT has been applied to temporal sequences of video, audio, and graphic. DTW is commonly used for measuring similarity between two temporal sequences which may vary in time or speed.

### 2.3.2.4. Zero crossings with peak amplitudes (ZCPA)

This feature extraction technique is based on Human Auditory System. It uses zero crossing interval to represent signal frequency information and amplitude value to represent intensity information, finally frequency information and amplitude information is combined to form the complete feature output. ZCPA is mostly used for development of automatic speech recognition in noisy environments, speaker identification, throat signal analysis, development of noise robust speech recognition system etc.

### 2.3.2.5. Perceptual linear prediction (PLP)

Perceptual linear prediction (PLP) technique combines the critical bands, intensity-to-loudness compression and equal loudness pre-emphasis in the extraction of relevant information from speech. It is rooted in the nonlinear bark scale and was initially intended for use in speech recognition tasks by eliminating the speaker dependent features. PLP gives a representation conforming to a smoothed short-term spectrum that has been equalized and compressed similar to the human hearing making it similar to the MFCC. In the PLP approach, several prominent features of hearing are replicated and the consequent auditory like spectrum of speech is approximated by an autoregressive all–pole model [55]. PLP gives minimized resolution at high frequencies that signifies auditory filter bank based approach, yet gives the orthogonal outputs that are similar to the cepstral analysis. It uses linear predictions for spectral smoothing; hence, the name is perceptual linear prediction [56]. PLP is a combination of both spectral analysis and linear prediction analysis.

## 2.4. Machine Learning

Machine learning is the scientific discipline that explores how to predict future events based on documented past events. In order to develop learning machines and software programs, the meaning of learning and what determines success or failure must first be defined. Machine learning uses computational, theoretical, and statistical principles to develop algorithms that model data from real-world phenomena and make accurate predictions about the phenomena. Machine learning operates in supervised, unsupervised and semi-supervised settings to perform classification, regression, visualization, clustering, dimensionality reduction, network modeling, graphical modeling, inference and structured prediction.

The name machine learning was coined in 1959 by Arthur Samuel [57]. Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."[58]. This definition of the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms. This follows Alan Turing's proposal in the paper "Computing Machinery and Intelligence", in which the question "Can machines think?" is replaced with the question "Can machines do what people (as thinking entities) can do?" [59]. In Turing's proposal the various characteristics that could be possessed by a thinking machine and the various implications in constructing one are exposed.

### 2.4.1. Relationships to Other Fields

Machine learning is related to many other fields in its defining principles as well as delivery techniques. Its most important relationships with other fields of study are discussed.

#### 2.4.1.1. Relation to data mining

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process. Machine learning and data mining often employ the same methods and overlap significantly, but while machine learning focuses on prediction, based on known properties learned from the training data, data mining focuses on the discovery of unknown properties in the data (this is the analysis step of knowledge discovery in databases). Data mining uses many machine learning methods, but with different goals; on the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy.

#### 2.4.1.2. Relation to optimization

Machine learning also has intimate ties to optimization: many learning problems are formulated as minimization of some loss function on a training set of examples. Loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances (for example, in classification, one wants to assign a label to instances, and models are trained to correctly predict the pre-assigned labels of a set of examples). The difference between the two fields arises from the goal of generalization: while optimization algorithms can minimize the loss on a training set, machine learning is concerned with minimizing the loss on unseen samples [60].

#### 2.4.1.3. Relation to statistics

Machine learning and statistics are closely related fields in terms of methods, but distinct in their principal goal: statistics draws population inferences from a sample, while machine learning finds generalizable predictive patterns [61]. According to Michael I. Jordan, the ideas of machine learning, from methodological principles to theoretical tools, have had a long pre-history in statistics [62]. The term data science was suggested as a placeholder to call the overall field [63].

Leo Breiman distinguished two statistical modeling paradigms: data model and algorithmic model [64], wherein "algorithmic model" means more or less the machine learning algorithms like Random forest. Some statisticians have adopted methods from machine learning, leading to a combined field that they call statistical learning [65].

### 2.4.2. Types of Learning Algorithms

Machine learning is a large field of study that overlaps with and inherits ideas from many related fields such as artificial intelligence. The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

#### 2.4.2.1. Supervised learning

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs [66]. The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and a desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs [67]. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task [68].

Supervised learning algorithms include classification and regression [69]. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.

In the case of semi-supervised learning algorithms, some of the training examples are missing training labels, but they can nevertheless be used to improve the quality of a model. In weakly supervised learning, the training labels are noisy, limited, or imprecise; however, these labels are often cheaper to obtain, resulting in larger effective training sets.

#### 2.4.2.2. Unsupervised learning

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms therefore learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, [70] though unsupervised learning encompasses other domains involving summarizing and explaining data features.

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to one or more predesigned criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated, for example, by internal compactness, or the similarity between members of the same cluster, and separation, the difference between clusters. Other methods are based on estimated density and graph connectivity.

#### 2.4.2.3. Reinforcement learning

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. Due to its generality, the field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms. In machine learning, the environment is typically represented as a Markov Decision Process (MDP). Many reinforcement learning algorithms use dynamic programming techniques [71]. Reinforcement learning algorithms do not assume knowledge of an exact mathematical model of the MDP, and are used when exact models are infeasible. Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent.

## 2.5. Summary

Firstly, this major introduced the automatic speech recognition system. And then gender classification challenges and techniques are presented. There are many gender classification techniques based on face, palm print, iris, fingerprint and speech. And also nature of adult and children speech is also discussed. Finally, the role of ML in computer science is explained.