Speech And Speaker Recognition: A Review

Karthika Kuppusamy, Chandra Eswaran

ABSTRACT: Essential thrust research domain of digital signal processing is speech processing. The framework of speech recognition allows the normal people to converse to the computer to fetch information whereas the framework of speaker recognition aims to determine the speaker alone. Automatic speech recognition is considered as the concept of science invention and now it is said to be the significant branch of information and communication technology. This paper gives the overview of speech and speaker recognition, role techniques namely feature extraction and classification which were discussed with its recent study. Finally, the paper concludes with the security issues and applications of speech and speaker recognition.

Index Terms- Speech Recognition, Speaker Recognition, Feature Extraction, Classification, Security Issues, Challenges, Applications and Tools.

1. Introduction

Communication is the most important part of the human behavior where it is made by using natural form of languages like speaking and writing. Human beings find easy and undemanding to converse and express their ideas by means of speech. [1] Speech is the vocalized form of human communication which contains information that is produced in speaker’s intellect. It is bimodal in nature. [2] The speech signal is altered with rapid and dynamic transform both in terms of frequency spectrum and the intensity. [4 The speech signal conveys the linguistic information and the information like age, sex, societal, location, physical condition and state of emotion. Signal Modeling and Pattern Matching are the basic functional operations of speech recognition systems where signal modeling converts the speech signals into a set of parameters and pattern matching discovers the parameter sets that are closely matched with the parameter sets from the input speech signal. [5] Globally, most recognized specific instruments used by individuals for identification is their voice. For these kinds of reasons, Automatic Speech Recognition (ASR) is considered as a prominent research area. Thus, a reasonable amount of research has been committed for the. The better distinct one is recognition of speech by automatic; its objective is to decode a taped speech articulation into its relating progression of words. Distinctive applications consolidate:

- speaker recognition, where the objective is to choose either the attested character of the speaker (checking) or who is talking (ID),
- speaker analysis, where the objective is to segment (or divide) an acoustic sequence regards to the underlying speakers. Even though huge number of researches has been resolved to speech processing, still it has some form of choices with respect to the essential gadgets used to approach the issues. [6]

1.1 Speech Recognition:

The measure of which artificial intelligence enabled device to analyze and classify the spoken words denotes the term speech recognition. Essentially, it recognizes what the user is speaking to a computer. The primary terms help in understanding the speech recognition include: Accuracy, Vocabularies, Utterance, Speaker Dependence and Training. [7]

![Fig. 1. Block Diagram of Speech Recognition System](image)

1.1.1 Types of Speech Recognition:

Isolated Words- Isolated Words (also called as segregated words) more often needs every utterance to be silent. It doesn’t imply that it acknowledges single words, yet it requires one utterance at any given moment. Connected Words - It is much as like Isolated Words, but it permits independent utterances to flow with very low pause time between them. Continuous Speech - Recognizer of the Continuous Speech permits the user to speak freely and can determine the content. Spontaneous Speech - It is the speech without any rehearsal or practice. In speech recognition, the system is ought to have the capacity to deal with natural speech. [8]

1.2 Speaker Recognition:

Speaker Recognition (also called as voice recognition) is the way towards recognizing the speaker from a given expression by analyzing the voice biometrics of the utterance along with expression models that have been collected previously. [10] It can also be said as, Speaker Recognition is the procedure of automated analyzing and detecting of who is talking by utilizing the unique information which is incorporated in sound waves, where it confirms the identification of individuals. The important aspect of speaker recognition is extracting the information to characterize it [9] from the prerecorded or live speech.
1.2.1 Types of Speaker Recognition:
Speaker recognition is divided into two types, which are:

- **Speaker Identification:** It is said to be a process that identify the user (already registered) who speak the statement or phrase or utterance [11].
- **Speaker Verification:** It is said to be a process that accept or reject the uniqueness of the claimed speaker [12].

- **Text Dependent** - It denotes the text that is used in the stages of testing and training [13]. Recognizing the speaker provides increased accuracy in order to identify the speaker, but this will not solve the real time problems in the current world [14].
- **Text Independent** - There is no need for the speaker to use the same phrase (or utterances). The text that is used to train and test is entirely different. It is assumed that it considers the real time problems in the current world, which is expected to be solved [15].
- More times the user gets confused with speech recognition and speaker recognition, where human voice is the common input for both types but different process are done with them [16].

![Block diagram of Speaker Recognition](image)

**Fig.2:** Block diagram of Speaker Recognition

The three major differences are:
- Speech recognition detects the words, where Speaker recognition detects the individuals by neglecting the language and its meaning.
- Speech recognition is dependent on language, where Speaker recognition is independent from language.
- Biometric devices are available for speaker recognition, but not for speech recognition. [17]

### 2. FEATURE EXTRACTION

Feature extraction is the procedure of remodeling the input data into a set of features which can very well highlight the input data. It identifies the speech features, which is spoken by various speaker. It also captures speaker specific properties.[18]

<table>
<thead>
<tr>
<th>Table: 1 Comparison of Different Feature Extraction Methods[7],[19]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technique</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Mel Frequency Cepstral Coefficient (MFCC)</td>
</tr>
<tr>
<td>Linear Predictive Coding (LPC)</td>
</tr>
<tr>
<td>Linear Predictive Cepstral Coefficient (LPCC)</td>
</tr>
<tr>
<td>Perceptual Linear Prediction (PLP)</td>
</tr>
<tr>
<td>Relative Spectra (FASTA)</td>
</tr>
</tbody>
</table>

#### 2.1 RELATED WORK

**2.1.1 Speech Recognition**

- F.de-la-Calle-Silos and R. M. Stern.,2017, proposed the temporal prototype and application pattern of auditory-nerve firings to enhance the strength of the automatic speech recognition systems. A new feature extraction technique for noise removal which is based on noise suppression is developed for increasing the accuracy of speech recognition in the presence of additive noise.[20].
- Y. Huang et al.,2015, have proposed a feature extracting approach for wavelet packet to recognize the emotion in speech by automate manner. Further, optimization concept is utilized to increase the performance of emotion recognition by utilizing the feature of Mel Frequency Cepstrum Coefficient [21].
- N. Moritz et al., 2015, presented the concept of Amplitude Modulation Filter Bank to extract the feature to describe data from human psychoacoustics. This proposal was to exhibit the error rate of significant word; also it has exposed the increased robustness in the presence of (i) noise, (ii) various characteristics of the channel used for data transmission, and (iii) reverberation of room [22].
- Zhen-Tao Liu et al., 2018, proposed a method for selecting the feature by analyzing the correlation of data. The main aim was to eliminate the features that are not necessary by checking the correlation between all the features. Further, a method for emotion recognition was proposed based on extreme learning machine in order to increase the performance of recognition [23].
2.1.2 Speaker Recognition

- Sharada V. Chougule et al., 2015, proposed a feature selection method in recognizing the speaker namely normalized dynamic spectral feature. It aimed to recognize the speaker even there exist a nose in an additive manner. The process of extracting the feature was performed with the proposed feature to recognize the speaker in a automated manner [24].
- Sourjya Sarkar et al., 2014, proposed a method for compensating the stochastic feature to verify the speaker. It utilizes the mixture model of Gaussian method to restrict the tasks in recognizing the speech. An application towards the proposed algorithm in verifying the speaker, where it concentrates to compensate the background noise [25].
- Latha., 2016, To avoid loss of high frequency region characteristics in speaker identification, a method to divide the samples into voice segments (i.e., unvoiced and voiced) was proposed. The segment of voiced speech is penetrated by utilizing the mel filter bank concept to produce speech signal with low level frequency, and the unvoiced speech is done with vice versa by using inverted mel filter bank [26].
- Suma Paulose et al., 2017, focused on features of voice source and spectro temporal features. The research work was proposed with the main of making classification by using two classifiers to increase the accuracy of recognition. It relay on the feature extraction methodologies and classifiers to recognize the speaker [27].

3. CLASSIFICATION

Classification is a process of predicting a specific result based on input given. It can also be said as, process of formulating the data into a categorical label with the condition given for the effective and efficient use [18].

<table>
<thead>
<tr>
<th>Technique</th>
<th>Characteristics</th>
<th>Vantages</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machine</td>
<td>Supervised</td>
<td>Simple operations, Higher accuracy</td>
<td>Not applicable for large data</td>
</tr>
<tr>
<td>Gaussian Mixture Model</td>
<td>Unsupervised</td>
<td>Requires less training data component</td>
<td>Issue in estimating themixture</td>
</tr>
<tr>
<td>Hidden Markov Model</td>
<td>Unsupervised</td>
<td>Simple, feasible to use</td>
<td>Computational more complex</td>
</tr>
<tr>
<td>Dynamic Time Wrapping</td>
<td>Unsupervised</td>
<td>Less storage spaces</td>
<td>Limited number of templates</td>
</tr>
<tr>
<td>Vector Quantization</td>
<td>Unsupervised</td>
<td>Used for data compression</td>
<td>Complex in encoding</td>
</tr>
</tbody>
</table>

3.1 RELATED WORKS

3.1.1 Speech Recognition

- Pribil J et al., 2014 Dealt with evaluating the quality enabled with synthetic speech by reversing the speech recognizing process of core speakers in which their voices were used by various text to speech conversion systems. It also aims to evaluate the control the transformation of voice in the process of recognizing the inventive speaker [28].
- Jiri Pribil and Anna Pribilova., 2013 find the correctness of emotion classification by using the various kinds of features line spectral and prosodic, where it is used to classify the emotional speech by depending on the parameters (i) count, (ii) its order vector of the input feature, and (iv) complexity of computation [29].
- Verkholyak O and Karpov A., 2018 A theoretical model was proposed to acquire the feature representation in a low level to feed the descriptor sequence of names to the network of long lasting short term memory. This combines the concepts of (i) Principal Component Analysis, (ii) representing the feature at the level of utterance, (iii) prediction of logical regression classifier [30].
- Mansour Alsulaiman et al., 2014 made a study to develop systems for diagnosing the patients by using speech and explored the utilization of the feature called relative spectral transform perceptual linear predictive, which is utilized in the pathology of speech. The proposed work aimed to detect and classify the disorder in voice [31].
- Dennis Norris et al., 2016 proposed a cognitive based prediction method for classification which implies the activation of processes between lexical and pre-lexical in the model of interactive-activation [32].

3.1.2 Speaker Recognition

- Enrique M. Albornoz et al., 2017 presented an methodology to use state-of-art features proposed for recognizing the state of speech and speaker. It ensembles the techniques of different classifier and to show the proposed classifier was best [33].
- Zakariya Qawaqneh et al., 2017 proposed a classification method to detect speaker's age and gender by using Bottle-Neck Feature (BNF) extractor together with Deep Neural Network (DNN), whereby regularizing the classes in DNN is made by using the shared class labels among misclassified classes and transformed Mel Frequency Cepstral Coefficients (MFCC) feature set is generated [34].
- Yanxiong Li et. al., 2017 proposed an unsupervised technique to analyze the role of speaker roles conversation speech with the presence of multiple participant. The features that were used to characterize the dissimilarity of various roles were extracted. The outcome of speaker recognition depends in the extraction of feature. The clustering method utilized to increase the inter-cluster distance was proposed to achieve the roles count...
and to concatenate the utterances related to the similar role into single cluster [35].

Ankita Jain et al., 2018 presented an approach to classify the gender using user's information of gait which was captured by utilizing the sensors that were inbuilt inside the smart phone. Histogram of Gradient method was proposed for gait feature extraction, which comprise group of signals that were gathered from accelerometer and gyroscope sensors that were inbuilt in smart phone [36].

Dong-Yan Huang et al., 2014 focused on investigating the consequences of Simple Partial Least Squares in binary classification (i.e., unbalanced). A classifier was proposed to increase the accuracy of prediction with minimum data count, where the dimension reduction and low computational complexity exist. Also, another classifier was proposed to increase the performance maximum data count [37].

4. Security Issues
Security is the condition of being liberated from threat and hazard which facilitate to ensure that the verification gets to succeed. The demand for getting the identity and authority for the users is increased in the recent times which leads and compel the user to memorize the passwords, pin codes etc., A better solution to overcome this kind of scenario is to employ biometric based verification scheme which is based upon the physical characteristics like iris, face, finger print, palm print, nose shape, voice which are distinct. According to the current situation, both the voice characteristic and the other physical characteristics of the speaker can be taken into account without the user knowledge. So the researchers started focusing on this research area to answer the questions:

(i) “who is the speaker?”
(ii) “Is the speaker who they claim to be?”

4.1. Challenges Faced in Speaker Recognition
Imitation or Mimicry: It is the process of endeavor of an impostor to mirror a subject that is enlisted in the framework, to access the framework by means of the outside record. Hence there exist contrasts amongst skilled and unskilled imitators. Speech Synthesis: The attacker makes a fabricated voice of the target person. Replay Attacks: It is one of the primary types of attack in speaker recognition. Here, the voice of the target speaker is recorded without their awareness and is used for the recognition process Unit Selection: It is the advancement of replay attacks, where the prerecorded audio of target voice is partitioned into number of sectors (i.e., units). These units will be made to play in selected orders to reach the target to cheat the recognition of speaker. Low-Quality voice samples: There exists a maximum change in the voice of the speaker because of background noise, health condition, mood, long period of time, digital & analogue, using different microphones. Accuracy: Accuracy is always a big question mark, where users are not able to trust whether the speaker recognition have detected the correct person or not. Vocal Stress: User may need to speak louder than regular. User voice will get strain and hoarseness due to speaking loudly for longer periods. Transformation of Voice: In the transformation of voice attacks, the speech signal of the impostor is altered for the similarity of a target person. Responsiveness of the User: User may start speaking or giving the command to recognize the voice before the system gets ready. Fault Tolerance: There arises a situation of algorithm getting work perfectly in speaker recognition but the hardware and software may not support due to getting operated over a long time, and vice-versa also may get happen [38]. Some of the recent proposals for speaker verifications are:

(i) To recognize the speaker, a fuzzy hidden markov model was proposed, where it uses the concept of kernel fuzzy c-means to extend the calculation of memberships of fuzzy while training the samples. [39]
(ii) To detect the activity of speech, a methodology was proposed by recording the specific Gaussian mixture modeling of speech and non-speech, where the frames were tend to extend the existing expectation maximization algorithm to train the mixture model by utilizing the semi-supervised learning [40].

Analysis of the unsupervised binaural scene to perform parallel operations like localizing, detecting and recognizing the specific speaker reverberant noise with inferential environment. It consists of three steps, which were: (i) localizing the source of sound, (ii) recognizing the speaker, and (iii) performing the indexing system [41].

Importance of speaker identification is investigated and proposed a method to utilize the feature of speaker recognition based on:

(i) Formants, (ii) Wavelet Entropy, and (iii) Neural Networks [42]

5. Applications
Even though there are various tasks that communicates with a computer which are capable to make use of ASR system, the following are the most frequently used applications. Dictation: Dictation is one of the most common processes in current ASR systems. It is incorporated medical transcription, business and word processing applications. In special cases, specific vocabularies are used to enhance the accuracy. Command and Control: Today smart devices are designed to function based on the user command. For example, in smartphones, the user can just command the names and make a call, instead of typing the contact numbers or names in dialer applications. Medical Disabilities: Due to physical limitations some people feel difficult to operate the devices. For example, user who has the difficulty to hear can use a system that is connected to their telephone to convert the caller's speech to text. Embedded Applications: This enables the user to communicate with systems only with some predefined words. Personalized User Interface: It denotes the interaction between user and computer by enabling the concept of personalization. For example, in voice-mail, the system could accommodate his/her needs and preferences. Multi Speaker Tracking: In this, more than one speaker is included in the conversation and allowing the system to detect which speaker is speaking. Example: Conference calls. Forensic Speaker Recognition: It is the act of proving the identity of a prerecorded voice which can help to identify a criminal in court. Biometric
Applications: One of the authentication techniques to authenticate the authorized user for accessing data [43].

6. MEASURES OF PERFORMANCE
Accuracy and speed are used to measure the performance of any speech recognition system. Accuracy can be calculated by using word error rate (WER) and speed with real time factor. Other measures include Single word error rate and Command success rate. The word error rate and recognition rate can be computed using.

\[
\text{Word Error Rate} = \left( \frac{I + S + D}{N} \right) \times 100
\]

\[
\text{Word Recognition Rate} = 1 - \text{WER} = \frac{N - S - D - 1}{N}
\]

7. TOOLS FOR AUTOMATIC SPEECH RECOGNITION
Hidden Markov Toolkit (HTK): It is written in ANSI C and is mainly used for building and manipulating hidden markov models. Initially it is build for English language and therefore it uses 8-bit ASCII standard code. SPHINX: The latest version of sphinx series is Sphinx 4. It is written in Java programming language and it provides flexible framework for speech recognition. JULIUS: It is an open source decoder software for continuous speech recognition and is developed for linux environment. SCARF: This toolkit is designed and developed for speech recognition with segmental restricted random fields. PRAAT: This software is popular, as it runs on broad range of operating system platform. It is mainly used for recording and analyzing the human speech in mono recording and stereo records. AUDACIETY: Mainly, it is used for recording and editing sounds and is free open source software. [45]

8. CONCLUSION
Integration of computers and telecommunication system has brought the issue of convenient computer interfaces for remote access to the fore. A computer with speech interfaces enables ordinary people to reap the benefit of information revolution. The ability to interact with computer faces multiple challenges. This paper reviewed the speech and speaker recognition, with feature extraction and classification used. Also the paper discussed the security issues that this research area faces and finally the applications available for this research area. Acknowledgment I am grateful to all kinds of support provided by Prof. Dr. E. Chandra Eswaran for guiding me for my research work. Thanks are also extended to all the higher authorities of Bharathiar University for giving me opportunity for doing my research work.

REFERENCES:
[7]. Michelle Cutajar et.al.(2013),“ Comparative study of automatic speech recognition techniques”, IET Signal Processing.


[42]. Khaled Daqrouq & Tarek A. Tutunji(2015), “Speaker identification using vowels features through a combined method of formants,

