Major Challenges of Voice Command Recognition Technique

Bhavneet Kaur

Abstract— Human listeners are capable of identifying a speaker, over the telephone or an entryway out of sight, by listening to the voice of the speaker. Achieving this intrinsic human specific capability is a major challenge for Voice Biometrics. Like human listeners, voice biometrics uses the features of a person's voice to ascertain the speaker's identity. The best-known commercialized form of voice biometrics is Speech Recognition System (SRS). Speech recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voice. This paper gives a brief introduction of SRS describing how the technology works, and then discusses the general architecture of SRS, methodologies, merits of using this system, major technological perspective and appreciation of the fundamental progress of speech recognition. It gives an approach to the recognition of speech signal using frequency spectral information with Mel frequency for the improvement of speech feature representation in a HMM based recognition approach and also gives overview of techniques developed in each stage of speech recognition along with the current and future researches on the same. This paper describes the major challenges for SRS system which have been came across by users feedback and various researches which has to be resolved as soon as possible for better performance outcome.

Index Terms— SRS (Speech recognition system), ASR (automatic speech recognition), MFCC (Mel frequency cepstral coefficient), DTW (dynamic time wrap), HMM (hidden markov models), WER (word error rate), WRR (word recognition rate).

1 Introduction

Speech is the expression of or the ability to express thoughts and feelings by articulate sounds. This is what is known for humans. But today it is not confined to just expressing one's feeling; it has many advance uses too. Today it is one of the security measures in many areas. And the credit completely goes to Speech recognition system.

Speech Recognition is the ability of a machine or program to receive and interpret dictation, or to understand and carry out spoken commands. SRS is a technology that involves generating a sequence of words that best matches the given speech signal (Digitized speech sample). It is also known as "automatic speech recognition", (ASR), "Voice Command Recognition", "computer speech recognition", "speech to text", or just "STT".

Speech interfacing involves two distinct areas, speech synthesis and automatic speech recognition (ASR). Speech synthesis is the process of converting the text input into the corresponding speech output, i.e., it acts as a text to speech converter. Conversely, speech recognition is the way of converting the spoken sounds into the text information being conveyed by these sounds. Among these two tasks, speech recognition is more difficult but it has variety of application such as interactive voice response systems for physically challenged persons.

SRS is quite helpful for illiterates who face many problems in human-computer interaction. For that user simply speaks their query in system and it will display the result of their query moreover because of speech synthesis technique they can listen to the corresponding displayed results too. This technique has been developed for around 7300 existing languages.

There are many soft wares for SRS like CMU Sphinx, Julius, Kaldi, iATROS, Dragon dictation, SILVIA, HTK etc. [1] and its leading vendors are Mac speech, SRC, High bridge communications, SRT distribution, Lumnenvox and many more.

This paper presents different speech features extraction techniques and their decision based recognition through artificial intelligence and statistics techniques, including the detail discussion on the major challenges faced by voice command recognition along with the current progress and future work.

1.1 Mathematical Representation of Automatic Speech Recognition

An utterance in ASR system statistically represented by some sequence of acoustic feature observations O, which is derived from the sequence of words W, and it is needed to find the most likely word sequence as given below [2]:

W= arg max P (W|O). (1)

In equation (1) argument of P (W|O), that is, the word sequence W, is found which shows maximum probability, of given observation vector O.

Using Bayes' rule it can be written as:

W = arg max P (O|W).P (W) / P (O) (2)

In equation (2), P (O) is ignored as it is constant with respect to

- *Bhavneet Kaur, Completed Post Graduation in MCA from Sikkim Manipal University, India, PH-9953125008. E-mail: daisy.singh.14@gmail.com*

W. Hence,

$$W = \arg\max P\ (O|W)\ P\ (W). \qquad (3)$$

In equation (3), P (W) is determined by a language model like grammar based model and P (O|W) is the observation likelihood and is evaluated based on an acoustic model.

2 VARIOUS KINDS OF SPEECH

On the basis of the utterances a system can recognize different classes of SRS. [3]:-

1) Isolated Word

Here recognizer accepts single word or single utterance at a time, and requires that it must be present at both sides of the sample window. It is simple and easiest to implement because word boundaries are obvious and the words tend to be clearly pronounced. The main disadvantage is in choosing different boundaries that affects the results.

2) Connected Word

Planned speech can be a good example of it. Here system allows separate utterance to be run together with minimum pause between them.

3) Continuous speech

Here recognizer allows user to speak almost naturally. It utilizes special methods to determine an utterance boundary that is why it is the most difficult one to be recognized. Here confusion between word sequences grows with the growth in vocabulary.

4) Spontaneous speech

A non-rehearsed speech is a spontaneous speech. For this kind of speech, system should be able to handle variety of natural speech features. A care should be taken by system as spontaneous speech may include mispronunciations, non-words or false starts.

3 AUTOMATIC SPEECH RECOGNITION CLASSIFICATIONS

Research into the concepts of speech technology began as early as 1936 at Bell Labs. As the Speech recognition describes it recognizes the speech patterns, i.e. it is a special case of pattern recognition.

There are two phases for classification in supervised pattern recognition, that are Training and Testing. During the training phase, the parameters of the classification model are estimated using a large number of class examples (Training Data). Training phase can also by describe as enrollment. During the testing or recognition phase, the feature of test pattern (test speech data) is matched with the trained model of each and every class. The test pattern is declared to belong to that whose model which matches the test pattern on best. Testing phase can also be described as verification.

4 WORKING AND METHODOLOGIES OF SRS

The underlying premise for speech recognition is that each person's voice differs in pitch, tone, and volume, which are enough to make it uniquely distinguishable.

How the Speech Recognition System works is described below:-

1) Speech recognition fundamentally functions as a pipeline that converts PCM (Pulse Code Modulation) digital audio from a sound card into recognized speech.
2) This is done by asking each person to speak out a word or any kind of utterance in microphone.
3) After this the digitalization of the speech signal is followed by some signal processing.
4) This creates a template for the speech pattern which is enrolled in the memory.
5) In order to recognize the speaker's voice a comparison is done by the system between the utterance and the template stored respectively for that utterance in the memory.

Fig. 1 shows the diagrammatic view of the working of the speech recognition process.

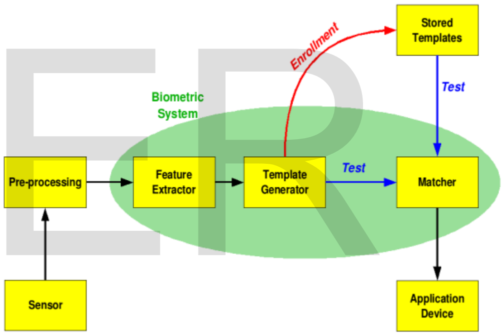

Fig. 1: Outline of Speech Recognition System

5 SPEECH RECOGNITION TECHNIQUES

An ASR technology is becoming the best source for human-computer interaction. Today it is being used in every other source of facility of human. This adoption of ASR technologies has led to more conveniences for citizens that have ASR in their languages.

The goal of automatic speaker reorganization is to analyze, extract characterize and recognize information about the speaker`s identity [4]. The complete ASR works in 4 stages which are described below.

5.1 Analysis Technique

The analysis of speaker`s recognition is done from the information obtained by him that is embedded in signals which are carried out for further procedures. A visual representation of voice is very useful for analysis, which is called spectrogram also known as voiceprint, voice gram, spectral waterfall or sonogram. A spectrogram displays the time, frequency of vibration of the vocal cords (pitch) and amplitude (volume). It has been studied that the pitch is usually higher of females as compare to males [5]. The speech analysis stage deals with suitable frame sizing, for segmenting speech signal, that is further used for analysis and extraction. Analysis can be done from one of the three techniques mentioned in below table 1-

TABLE1
TYPES OF ANALYSIS TECHNIQUES

Type of Analysis Points of differentiation	**Seg-mentation Analysis**	**Sub Segmental Analysis**	**Supra Segmental Analysis**
SPEECH IS ANALYZIES USING	Frame size and sift of 10- 30 ms.	Frame size and sift of 3-5 ms.	Frame size and sift of 100- 300 ms
EXTRACTS	Extract speaker information which is further used to extract vocal tract information of speaker recognition	Extract the characteristic of the excitation state.	Extract speaker information which is used mainly to analyze and characteristic speech on the basis of behavior character of the speaker
IMPORTANT POINT	-	The small frame size and shift are required to best capture the speaker information because source information varies fast as compared to vocal tract information.	-

5.2 Feature Extraction Technique

The main goal of the feature extraction technique is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal, which is done by speech feature extractor who collects the feature vectors from input audio waveforms using a sequence of signal processing steps in a data flow framework.

The number of training and testing vectors needed for the classification problem grows with the dimension of the given input in formation of speaker identification and verification system. Therefore feature extraction technique is used for reducing the dimensionality of the input vector while maintaining the discriminating power of the signal i.e., the transformation of the incoming sound into an internal representation such that it is possible to reconstruct the original signal from it.

It should be observed that the extracted features must follow some criteria like it must be easy to measure extracted speech features, it should not be susceptible to mimicry, it should show little fluctuation from one speaking environment to another, it should be stable over time and it should occur frequently and naturally in speech [9].

Recognition performance depends heavily on the feature extraction phase therefore feature extraction requires much more attention as compare to others. LPC, MFCC, AMFCC, PCA,ICA, Wavelet, Mel Cepstrum Analysis, RAS, DAS, ΔMFCC, Higher lag autocorrelation coefficients, PLP, MF-PLP, BFCC, RPLP are the different techniques for feature extraction. It was found that higher lag autocorrelation algorithms gave the better results [6]. The researches have been made and have been found that noise robust spectral estimation is possible on the higher lag autocorrelation coefficients. Therefore, eliminating the lower lags of the noisy speech signal autocorrelation leads to removal of the main noise components.

Some of the commonly used speech feature extraction techniques are discussed below-

5.2.1 Spectral Analysis Techniques

Spectral analysis techniques are mainly required to recognize a time domain signal when it is in its frequency domain representation. This is basically done by performing a Fourier transform over it. It is used on the basis of Spectrogram and it has the property of Robust Feature extraction method.

5.2.2 Cepstral Analysis

It has the property of Static feature extraction method, and Power spectrum.

This is an important analysis technique, by which excitation and vocal tract can be set apart. The speech signal is given as [7]-

$$s\ n = g\ n \times v\ n \qquad (1)$$

In equation (1) where, $v\ n$ is the vocal tract impulse response, and $g\ n$ is the excitation signal.

Also, the frequency domain is represented as-

$$S\ f = V\ f \qquad (2)$$

Logarithmically it can be represented as-

$$\log S\ f = \log G\ f + \log V\ f \qquad (3)$$

Therefore, it can be observed that excitation and vocal tract can be set apart and can be superimposed if logarithm is taken in the given frequency domain.

5.2.3 Mel-Frequency Cepstral Coefficient (MFCC)

It has been proposed by Preeti Sahni [11] that Mel-Frequency Cepstral Coefficient (MFCC) are used because it is designed using the knowledge of human auditory system and is used in every state of speech recognition system or art speech. It's a standard method for feature extraction in speech recognition tasks. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed.

A new modification of MFCC feature has been proposed by Om Prakash Prabhakar and Navneet Kumar Sahu [8] for extraction of speech features for Speaker Verification (SV) application. Now, Extracted features are compared with both original MFCC method and one with the recent modification.

MFCC has one limitation that it does not consist an outer ear model due to which it cannot represent perceived loudness precisely. The most commonly used cepstral coefficients are MFCCs and LPCCs, because of less intra-speaker variability and also availability of spectral analysis tools. However, the speaker-specific information represents different aspects of speaker information due to excitation source and behavioral tract. The main limitation for the use of excitation source and behavioral tract is the non – availability of suitable feature extraction tools.

5.3 Modeling Technique

The objective of modeling technique is to generate speaker models using speaker specific feature vector.

The two commercialized applications of speaker recognition technologies are SIS and SVS as discussed in [8]-

Speaker Identification System (SIS) – The goal is to identify the speaker from set of known voices. It is a 1: N matches where the voice is compared against N templates.

Speaker Verification System (SVS) – It is the process of accepting or rejecting the speaker claiming to be the actual one. It is a 1:1 match where one speaker's voice is matched to one template.

Speaker Recognition (verification) divided on the basis of speaker as well as on the basis of text.

On the basis of speaker it is divided in:

1) **Speaker Independent** - In this mode, the system ignores the speaker specific characteristics of the speech signal and extracts the intended message only.
2) **Speaker Dependent**- In this mode, system should extract speaker characteristics in the acoustic signal.
3) **Speaker adaptive**- A third type of speaker models is now emerging, called speaker adaptive which has been discussed in [3]. These systems usually begin with a speaker independent model and adjust these models more closely to each individual during a brief training period.

On the basis of text it is divided in:

1) **Text Dependent**- In this method the speaker speaks the key words or sentences having the same text for both training and recognition trials (i.e. for both verification and identification).
2) **Text Independent**- It does not rely on a specific texts being spoken.

The three main approaches for modeling analysis are discussed below-

5.3.1 Acoustic-Phonetic Approach

Acoustic-phonetics is the earliest approaches to speech recognition which were based on finding speech sounds and providing appropriate labels to these sounds. This method is indeed viable and has been studied in great depth for more than 40 years. This approach is based upon theory of acoustic phonetics and postulates. Acoustic-phonetic approach exploits the theory of acoustic-phonetics of existence of finite phonetic units which are characterized by set of properties in the speech signal. It decodes the speech signal based on the known relationship between acoustic features of the signal and phonetic unit. This approach consists of three processes, which are: spectral analysis and features extraction, segmentation and labeling, and valid word/ string identification. Acoustic phonetic approach is also known as rule-based approach. It has been stated that despite being one of the earliest approaches acoustic-phonetic has not recorded wider use in commercial applications [10].

5.3.2 Pattern Recognition Method

Speech recognition is one in which the speech pattern are required directly without explicit feature determination and segmentation. Pattern training and pattern comparison are the two essential steps in this approach stated in [6].

The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern so that direct comparison between unknown test patterns and trained patterns can be done, so that identity of similarity between them can be made accordingly and the goodness of matched pattern can be determined [11]. The only problem associated with the pattern recognition approach is that the system`s performance is directly dependent over the training data provided. The block diagram of shown below in Fig. 2.

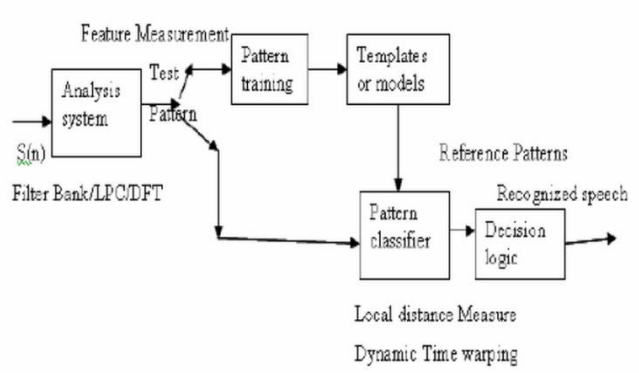

Fig. 2: Block Diagram of Pattern Recognition.

5.3.2.1 Template Matching Approach

In template matching approach a collection of prototypical speech patterns are stored as reference patterns representing the dictionary of candidate`s words. Recognition is then carried out by matching an unknown spoken utterance with each of its reference templates and selecting the category of the best matching pattern. Therefore, it has the advantage of using perfectly accurate word models. Usually, templates for entire words are constructed. This has the advantage that, errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided. As a consequence, every word must have its own full reference template. Test pattern and reference pattern are represented by sequences of feature measurements. Pattern similarity is determined by aligning test pattern with reference pattern with distortion. Decision rule chooses reference pattern with smallest alignment distortion.

Template preparation and matching become prohibitively expensive or impractical as vocabulary size increases. Wiqas [6] have made an attempt to overcome the key problems of HMM framework, i.e. discarding the information about time dependencies and over-generalization, by applying template based continuous speech recognition with Dynamic Time Warping.

Dynamic Time Warping (DTW) [12] is used to compute the best possible alignment between test pattern and reference pattern and the associated distortion. Any data which can be turned into a linear representation can be analyzed with DTW. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other [13].

In DTW let the two time series are Q and C.

$Q= q_1, q_2, \ldots, q_n$ and $C=c_1, c_2, \ldots., c_n$

Now construct n X m matrix D with distances $D\, j=d\ (q_i, c_i)$.

Here warping path W is a contiguous set of matrix elements $w_k = (i, j)_k$.

Defining warping between Q and C. $W=w_1, w_2, \ldots., w_k$, where $\max (n, m) \leq K \leq m+n-1$.

Therefore, DTW (Q, C) = min $\sqrt{\sum wk}$

A general view of DTW is shown in Fig. 3.

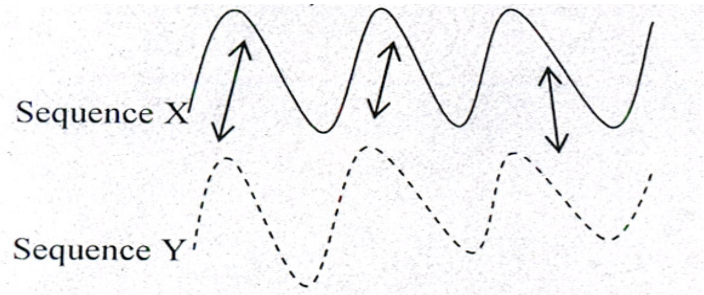

Fig.3: View of Dynamic Time Warping Technique.

This sequence alignment method is often used in the context of HMM. It`s quite efficient for isolated word recognition and can be adapted to connected word recognition.

5.3.2.2 Stochastic Approach

This approach is based on the use of probabilistic models so that uncertain or incomplete information which may arise from many sources, such as confusable sounds, speaker variability`s, contextual effects, and homophones words, can be dealt with. It collects a large corpus of transcribed speech recording and train the computer to learn the correspondences. At run time, statistical processes are applied to search for all the possible solutions & pick the best one from all [11]. The most popular stochastic approach today is hidden Markov modeling.

Hidden Markov model (HMM) modeling is more general and possesses firmer mathematical foundation in comparison to template based approach [13]. HMM is characterized by a finite state markov model and a set of output distributions [14].

So, the transition parameters in the Markov chain models are temporal variability's, while the parameters in the output distribution model, spectral variability's. A template based model is simply continuous.

To overcome the disadvantage of the HMMs machine learning methods could be introduced such as neural networks and genetic algorithm programming. In these machine learning models there is no need for explicit rules or other domain expert knowledge as they can be learned automatically through emulations or evolutionary process.

5.3.3 Artificial Intelligence Approach

The Artificial Intelligence approach is a hybrid of the acoustic phonetic proposed by M.A.Anusuya [13]. In the AI, an expert system implemented by neural networks is used to classify sounds. The basic idea is to compile and incorporate knowledge from a variety of knowledge sources with the problem in hand.

Thus AI is divided in two processes that are automatic knowledge acquisitions learning and adaptation. Neural networks have many similarities with Markov models, both are statistical models which are represented as graphs. The only key difference between neural networks and markov chain is that the earlier one is fundamentally parallel while latter is serial. Expert systems are used widely in this approach.

5.4 Matching Technique

In this technique the obtained unknown word is matched with known word using one of the following techniques-

5.4.1 Whole-Word Matching

In this digital-audio signal is compared against prerecorded template of the word by the search engine. It is quite faster than sub-word matching. It requires huge amount of storage (between 50 and 512 bytes per word) because there are many words which are formed from the combination of two morphemes and here they are stored as a whole. For example a speaker utters a word "misunderstand" then in whole word matching the system will compare the whole word with prerecorded template "misunderstand".

5.4.2 Sub-Word Matching

In this matching, search for sub-words usually phonemes is done and then performs further pattern recognition is carried out. This technique takes more processing than whole-word matching, but it requires much less storage (between 5 and 20 bytes per word) [15]. Let`s take the same example as earlier of a word "misunderstand" here system will divide it in two "miss" and "understand" and then perform the further matching with prerecorded templates.

6. APPLIACTIONS OF SRS AND FUTURE RESEARCHES

Automatic speech recognition enables a wide range of current and emerging applications such as automatic transcription, multimedia content analysis, and natural human-computer interfaces. Speech recognition systems are making rapid advancements described in [6] and are being used in nearly every major industry like in transportation, communication, education, military, mining, manufacturing, police-terrorist interactions, prisons, courts processing, construction, space, vehicle navigation, self-driving car interface and many more.

Future researches in this field are also on its track and some on track researches as follows-

1) **Body Language + Facial Expression + Voice Recognition:**

Facial expression or mirroring is very popular. The goal here is to create an emotional bond with the machine through human-machine interaction. Voice Recognition systems that can also read body languages and facial expressions can also be used for threat detecting assessments, which can be used at airports, border crossings etc. because of which human workers at those locations or choke points can be replaced by SRS. Various research projects on android robotics are being carried in Japan and US.

2) **Vocal Cord Vibration Recognition + Current Voice Recognition:**

Advance research in the US Military is going on that allows vocal cords to be read, without actual speech spoken or voice occurred; these systems will be in working soon. This is done with a device placed near the larynx that picks up sensitive vibrations, which is coupled to a transmitter for sending the signals and the human receiver on the other side has a tiny ear piece for hearing hear the speech.

3) **Understanding a Joke and Responding with Another One**:

Artificial Intelligence is getting better all the time, soon, AI software engineers will create joke recognition systems, where the computer will understand irony and know when the human is telling a joke, then reciprocate with a joke of their own, perhaps creating a joke from scratch. The system would be pre-loaded with all the jokes common to human interaction in all cultures. It will be able to pick one that has most likely not been heard by the human they are working with at the time; also put in memory that it has been told to that individual so it does not repeat it.

4) **Speech Recognition in Radiology Information System**:

The Radiology report is the fundamental means by which radiologists communicate with their clinicians and patients. The traditional method of generating reports were time consuming and expensive but with recent advances in computer hardware and software technology with improved Speech Recognition systems are used for radiology reporting. Likewise more advancement in field of medical along with speech recognition is on the way [3].

7. CHALLENGES OF ASR

Automatic Speech Recognition is an application that consistently exploits advances in computation capabilities. With the availability of a new generation of highly parallel single-chip computation platforms, ASR researchers are faced with the question of unlimited computing to make speech recognition better. The goal of the work reported here is to explore plausible approaches to improve ASR in three ways:

1) Improve Accuracy: Account for noisy and reverberant environments in which current systems perform poorly, thereby increasing the range of scenarios where speech technology can be an effective solution.

2) Improve Throughput: Allow batch processing of the speech recognition task to execute as efficiently as possible, thereby increasing the utility for multimedia search and retrieval.

3) Improve Latency: Allow speech-based applications, such as speech-to-speech translation, to achieve real-time performance, where speech recognition is just one component of the application.

Some goal challenges in ASR which will enable progress on multiple promising research areas at a variety of levels are noted below:-

1) "Every day audio" is a term that represents a large range of speech, speaker, channel, and environmental conditions which people typically encounter and routinely adapt to in responding and recognizing speech signals. Currently ASR systems are challenged, and deliver significantly degraded performances, when they encounter audio signals that differ sometimes even slightly from the limited conditions under which they were originally developed. New techniques and architectures are been proposed whose focus will be exploring alternatives for automatically adapting to changing conditions in multiple dimensions even simultaneously.
2) Today's systems deliver best performances by building complex acoustic and language models using a large collection of domain-specific speech and text examples. This set of language resources is often not available for rarely-seen languages. So, the challenge is to create rapid portability spoken language technologies. Research is also needed to study the minimum amount of supervised label information required to bring up a reasonable system that will serve for bootstrapping purposes.
3) There is a need for learning at all levels of speech and language processing to cope with changing environments, non-speech sounds, speakers, pronunciations, dialects, accents, words, meanings, and topics, to name but a few sources of variation over the lifetime of a deployed system. Like human, system will engage in automatic pattern discovery, active learning and adaptation. Research in this area must address both the learning of new models, as well as the integration of such models into pre-existing knowledge sources. Thus, an important aspect of learning is being able to discern when something has been learned, and how to apply the result. Therefore, it's a challenge to create self-adaptive (or self-learning) speech technology.
4) Current speech recognition systems have difficulty in handling unexpected out-of-vocabulary words, and in languages for which there is relatively little data with which to build the system's vocabulary and pronunciation lexicon. A common outcome in this situation is that high-value terms are overconfidently misrecognized as some other common and similar-sounding word. Yet, such spoken events are key to tasks such as spoken term detection and information extraction from speech. Their accurate detection is therefore of vital importance. So, the challenge is to create a system that detect the word when they actually don`t know the correct one.
5) A key cognitive characteristic of humans is their ability to learn and adapt to new patterns, and stimuli. Although this behavior is very important and relatively well-understood in humans, but very little of this knowledge has been found its way into automatic speech and language systems. The challenge is to understand and emulate relevant human capabilities and to incorporate these strategies into ASR. Since it is not possible to predict and collect separate data for any and all types of speech, domains, etc. So, it is important to enable automatic systems to learn and generalize even from single instances or limited samples of data, so that new or changed signals can be correctly understood. It has been well demonstrated that adaptation in ASR is very beneficial.
6) To achieve a broad level of speech understanding capabilities, it is critical that the community explores building language comprehension systems that can be improved by gradual accumulation of knowledge and language skills, as todays systems are designed to transcribe spoken utterances only. Therefore, the goal is to facilitate language comprehension enabling technologies in systems. It is clear that such evaluations will emphasize accurate detection of information-bearing elements in speech rather than raw word error rate. Collaboration between speech and language processing communities is a key driver to the success of this program. The outcomes of this research will provide a paradigm shift for building domain-specific language understanding systems, and will impact the education and learning communities too.

Relevant issues of ASR design*:* Main issues on which recognition accuracy depends have been presented in the Table 2-

TABLE 2
ISSUES OF ASR DESIGN

Environment	Type of noise; Signal/noise ratio; working conditions
Transducer	Microphone; telephone
Channel	Band amplitude; distortion; echo
Speakers	Speaker dependence/independence Sex, Age; physical and psychical state
Speech styles	Voice tone(quiet, normal, shouted); Production(isolated words or continuous speech read or spontaneous speech) speed
Vocabulary	Characteristics of available training data; specific or generic vocabulary.

8 SUMMARY OF THE TECHNOLOGY PROGRESS

The complete progress of ASR technology from past to current present has been described in the table-3.

TABLE 3
PROGRESS SUMMARY

Sr. no	Past	Present
1	Template matching	Corpus -based modeling e.g. HMM and n-grams
2	Filter bank/spectral resonance	Cepstral features, Kernel based function, group delay functions
3	Heuristic time normalization	DTW/DP matching
4	Distance –based methods	Likelihood based methods
5	Maximum likelihood approach	Discriminative approach e.g. .MCE/GPD and MMI
6	Isolated word recognition	Continuous speech recognition
7	Small vocabulary	Large vocabulary
8	Context Independent units	Context dependent units
9	Clean speech recognition	Noisy/telephone speech recognition
10	Single speaker recognition	Speaker-independent/adaptive recognition
11	Monologue recognition	Dialogue/Conversation recognition
12	Read speech recognition	Spontaneous speech recognition
13	Single modality(audio signal only)	Multimodal (audio/visual) speech recognition
14	Hardware recognizer	Software recognizer
15	Speech signal is assumed as quasi-stationary in the traditional approaches. The feature vectors are extracted using FFT and wavelet methods etc.	Data driven approach does not possess this assumption i.e. signal is treated as nonlinear and non-stationary. In this features are extracted using Hilbert Haung Transform using IMFs.[24]

9 PERFORMANCE OF ASR

Accuracy and Speed are the two criterions for measuring the performance of an automatic speech recognition system. Latest research says that the use of artificial neural networks (ANNs), mathematical models of the low-level circuits in the human brain, improves the speech-recognition performance, through a model known as the ANN-Hidden Markov Model (ANN-HMM) which has shown promise for large-vocabulary speech recognition systems.

9.1 Accuracy

Accuracy is measured in terms of performance accuracy which is rated with word error rate (WER). Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR).

The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence.

In order to analyze the system performance, Hidden Markov Model Tool Kit (HTK) provides a tool known as HResult. From which accuracy of the system can easily be computed. It compares the machine transcription of the test utterances with the corresponding reference transcription files [2].

The performance of speech system is evaluated as shown in (1):

$$\%\text{correct} = \frac{N-D-S}{N}\ X\ 100 = \frac{H}{N}\ X\ 100 \qquad (1)$$

In equation (1) where N is the number of words in test set, D is the number of deletions, S is number of substitutions and H is the number of correct labels.

%correct gives the percentage of word correctly recognized.

The word error rate is working at the word level instead of the phoneme level. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment [11].

It can be computed as given in (2):

$$WER = \frac{S-I-D}{N} = 100 - \%\ accuracy \qquad (2)$$

The accuracy or Word Recognition Rate (WRR) is computed as given in (3): [4]

$$\%accuracy = \frac{N-D-S-I}{N}\ X\ 100 = \frac{H-I}{N-I}\ X\ 100 \qquad (3)$$

In equation (2) and (3) I is the number of insertions.

WRR= 1-WER (4)

WRR can also be computed as done in (4).

9.2 Speed

Speed is measured with the real time factor [9]. If it takes time P to process an input of duration I, the real time factor is defined in (5)-

RTF = P / I (5)

For example, Let real time factor is 2, if it takes 6 hours of computation time to process then a recording of duration is 3 hours. RTF ≤ 1 implies real time processing.

10 CONCLUSIONS

An attempt has been made in this paper to presented an extensive survey of Speech Recognition Systems i.e. the identification of the person who is speaking (i.e. speaker recognition) by characteristics of their voices (voice biometrics) with what is being said along with the detail discussion of the major challenges of the system . SRS has been categorized into different modules and discussed different approaches for each module i.e. various techniques for speech recognition which includes processes for the feature extraction and pattern matching with their properties. In addition to this, a study of the various typical applications of SRS, current research being carried out in this field. There is a wide use of SRS in the field of writing, translation, designing, word processing and recording and discussed that how the performance of system can be measures so that it can be improved in future.

REFERENCES

[1] *"List of speech recognition software"*, en.wikipedia.org/wiki/List_of_speech_recognition_software. 2014.

[2] Kuldeep Kumar, Ankita Jain and R.K. Aggarwal,*" A Hindi Speech Recognition System for Connected Words using HTK*", *IJCS*, vol. 1, no. 1, pp. 25-32, 2012.

[3] Parwinder pal Singh, Er. Bhupinder Singh, *"Speech Recognition as Emerging Revolutionary Technology"*, *IJARCSSE*, vol. 2, issue. 10, pp. 410-413, Oct. 2012.

[4] Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar, *"A Review on Speech Recognition Technique"*, *IJCA*, vol. 10, no.3, pp. 16-24, Nov 2010.

[5] Zia Saquib, Nirmala Salam, Rekha P. Nair, Nipun Pandey and Akanksha Joshi, *"A Survey on Automatic Speaker Recognition Systems"*, T.-h. Kim et al. eds., SIP/MulGraB 2010, CCIS 123, Springer-Verlag Berlin Heidelberg, pp.134–145, 2010.

[6] Wiqas Ghai, Navdeep Singh, *"Literature Review on Automatic Speech Recognition"*, *IJCA*, vol. 41, no. 8, pp. 42-50, Mar. 2012.

[7] Shachi Sharma, Krishan Kant, Krishna Kumar Sharma, *"Reviewing Human-Machine Interaction through Speech Recognition Approaches and Analyzing an Approach for Designing an Efficient System"*, *IJCA*, vol. 38, no.3, pp. 26-32, Jan 2012.

[8] Om Prakash Prabhakar, Navneet Kumar Sahu, *"A Survey On: Voice Command Recognition Technique"*, *IJARCSSE*, vol. 3, issue. 5, pp. 576-585, May 2013.

[9] Sanjib Das, *"Speech Recognition Technique: A Review"*, *IJERA*, vol. 2, issue. 3, pp. 2071-2087, May-Jun 2012.

[10] Shahrul Azmi Mohd Yusof, Abdulwahab Funsho Atanda, M. Hariharan *"A Review of Yorùbá Automatic Speech Recognition"*, *IEEE 3rd International Conference on System Engineering and Technology*, pp. 242-247, Aug. 2013.

[11] Preeti Saini, Parneet Kaur, *"Automatic Speech Recognition: A Review"*, *IJETT*, vol. 4, issue. 2, pp. 132-136, 2013.

[12] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov,*"Neural Networks used for Speech Recognition" Journal of Automatic Control, University of Belgrade*, vol 20:1-7. 2010.

[13] M.A.Anusuya, S.K.Katti *"Speech Recognition by Machine: A Review"*, *IJCSIS*, vol. 6, no. 3, pp. 181-205, 2009.

[14] Kuldeep Kumar R. K. Aggarwal *"Hindi speech recognition system using HTK"*, *International Journal of Computing and Business Research*, vol. 2, issue. 2, pp. 1457-1466, Sept 2011, doi 10.1007/s11235-011-9623-0.

[15] Aakash Nayak, Santosh Khule, Anand More, Avinash, Yalgonde, Dr. Rajesh S. Prasad, *"Study of various issues in voice translation"*, *IJARCET*, vol. 2, issue. 1 pp. 188-191, Jan 2013.

[16] Michael Cowling, Renate Sitte, *"Analysis of Speech Recognition Techniques For use in a Non-Speech Sound Recognition System"*, members of IEEE, pp. 15-20.

[17] Takialddin Al Smadi, *"An Improved Real-Time Speech Signal In Case Of Isolated Word Recognition"*, *IJERA*, vol. 3, Issue. 5, pp. 1748-1754, Sep-Oct 2013.

[18] Preeti Saini, Parneet Kaur, Mohit Dua, *"Hindi Automatic Speech Recognition Using HTK"*, *IJETT*, vol. 4, issue. 6, pp. 2223-2229, Jun 2013.

[19] Noelia Alcaraz Meseguer, "Speech Analysis for Automatic Speech Recognition", Master of Science in Electronics Thesis, Department of Electronics and Telecommunications, Norwegian University of Science and Technology, July 2009.

[20] D. Raj Reddy, "Speech Recognition by Machine: A Review", *Proceedings of the IEEE*, vol. 64, no. 4, pp. 501-531, Apr 1976.

[21] Rajesh Kumar Aggarwal , M. Dave "Acoustic modeling problem for automatic speech recognition system: advances and refinements (Part II)", *Int J Speech Technol*, Springer Science Business Media, LLC 2011, pp. 309–320, Aug 2011, doi 10.1007/s10772-011-9106-4.

[22] Laurent Besacier a, Etienne Barnard b, Alexey Karpov c, Tanja Schultz, *"Automatic Speech Recognition for Under-Resourced Languages: A survey"*, Speech Communication 56 (2014) 85–100, Aug 2013.

[23] Ángel de la Torre, Antonio M. Peinado, José C. Segura, José L. Pérez-Córdoba, Ma Carmen Benítez, Antonio J. Rubio, *"Histogram Equalization of Speech Representation for Robust Speech Recognition"*, *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355-366, May 2005.

[24] Sanjivani S. Bhabad, Gajanan K. Kharate, *"An Overview of Technical Progress in Speech Recognition"*, *IJARCSSE*, vol. 3, issue. 3, pp. 488-497, Mar 2013.