I.J. Image, Graphics and Signal Processing, **2019, 5, 1-12**
Published Online May 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2019.05.01

An Extensive Review of Feature Extraction Techniques, Challenges and Trends in Automatic Speech Recognition

Vidyashree Kanabur[1] and Sunil S Harakannanavar[2]
[1,2]Department of Electronics and Communication Engineering, S. G. Balekundri Institute of Technology, Belagavi-India
Email: vidyashreerk1992@gmail.com and sunilsh143@gail.com

Dattaprasad Torse[3]
[3]Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology, Belagavi, India
Email: datorse@git.edu

Received: 07 January 2019; Accepted: 28 January 2019; Published: 08 May 2019

***Abstract*—**Speech is the natural mode of communication between humans. Human-to-machine interaction is gaining importance in the past few decades which demands the machine to be able to analyze, respond and perform tasks at the same speed as performed by human. This task is achieved by Automatic Speech Recognition (ASR) system which is typically a speech-to-text converter. In order to recognize the areas of further research in ASR, one must be aware of the current approaches, challenges faced by each and issues that needs to be addressed. Therefore, in this paper human speech production mechanism is discussed. The various speech recognition techniques and models are addressed in detail. The performance parameters that measure the accuracy of the system in recognizing the speech signal are described.

Index Terms—Automatic Speech Recognition, Feature Extraction, Acoustics, Phonemes, Pattern Recognition, Artificial Intelligence.

I. Introduction

Human-machine interaction is gaining high attention in the past few decades due to advances in technology and applications which demand the computers to be able to perceive, interpret and respond to the command given by the user via the voice signal. Hence, understanding the human speech production model and devising an accurate model to recognize speech is necessary.

A. Mechanism of human speech production

Speech is the transformation of thoughts into words. Human ears perceive the sound signal and there is a conversion of pressure signal into electrical signal. The message is conveyed to the brain where the processing is done and appropriate decision is taken. If the response is to be verbal then it is conveyed to the speech model through the motor system of the human body. The articulation and the formation of meaningful messages are carried out by the speech model. The human speech production model is shown in figure 1. Here, sound is produced when the air pressure is applied to the lung via the muscles and then this pressure signal passes through the vocal tract.

The vocal tract has vocal folds that are characterized by having different resonant frequencies. The opening and closing of the vocal folds produce varying words or sound signal. The sound signal may pass through the oral cavity producing oral sounds or the nasal cavity producing nasal sound, depending on the closing or the opening of the velum respectively.

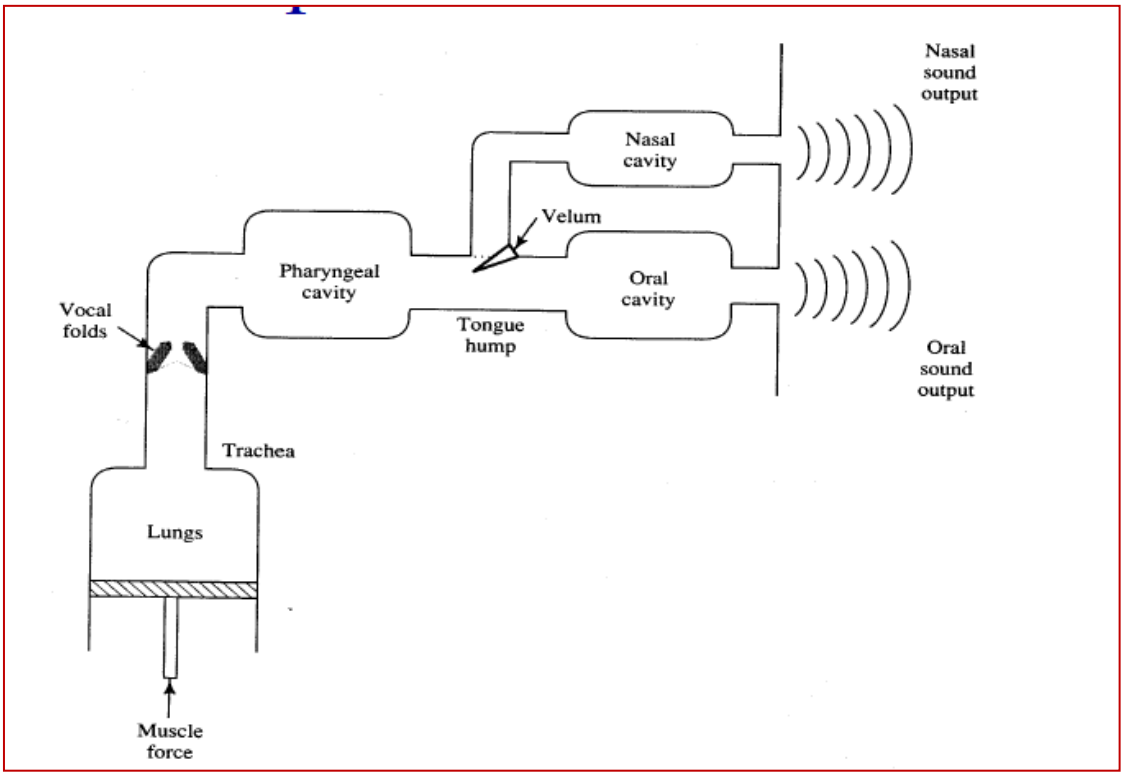


Fig.1. Human Speech Production model

B. Classification of Speech signal

There are various ways for classifying speech signal. One of the methods is based on the periodicity of the signal. Three broader classes are identified and their descriptions are discussed below.

- ***Voiced Speech*:** If the speech signal is quasi-periodic then the signal is said to be voiced which is produced by the periodic vibration of vocal folds in the vocal tract. Voiced signal carries important information about the utterances and hence recognizing the same becomes important.
- ***Unvoiced speech:*** Unvoiced sounds are aperiodic signals which are produced as turbulence due to constriction of the signal at some point in the vocal tract (especially near the mouth end). Unvoiced signal may be the noise that is present in the actual spoken sentences.
- ***Silence:*** Silence is the period of pauses between the words or sentences. It has zero information and has to be neglected while processing.

It is observed that a careful analysis is performed by ASR model to determine the voiced, unvoiced or silence portions of the speech signal. This model formulates algorithms which trains the computer to perform the processing of the information and determines what is the uttered word or sentence. It also performs a kind of mapping from continuous time speech signal to a sequence of discrete time samples.

C. Objectives of Automatic Speech Recognition

The main objective of speech recognition is to understand what is being said. ASR develops models and algorithms that can precisely simulate the physical system of human speech production. Text conversation between human and machine needed the user to type in the data to the machine at the speed that would match the speed of the utterance, which is difficult. If the machine is able to take in the instructions directly by speech without text then a lot of time is saved. In order to achieve this objective ASR is deployed as it converts speech signal into text form.

D. Challenges of Automatic Speech Recognition

Speech is a non-stationary continuous time signal whose implementation for speech recognition poses many challenges. Few of them are described as follows.

- Speakers have different language of communication. So the machine has to be fed with the database of each language which will be huge and this increases the cost of implementation and the search time.
- Human interact not just with words but with the help of gestures as well. The need to inculcate gestures and emotion recognition with speech becomes challenging task.
- Recording of voice signal will introduce background noise which will affect the accuracy of recognition. So a de-noising process must be adopted to improve performance of ASR.

- Speech recognition will give accurate results when the processing is carried on single phoneme, word, phrase or sentences. But in real-time, continuous speech has to be processed which is very difficult to implement.
- Vocal tract characteristics vary with age, gender, accent used while speaking, emotional state of the speaker while recording, etc. Also the language used for communication will be different. So, generalizing the speech recognition model will need huge database and processing time.

E. Motivation

The demand for human-machine interaction is increasing day-by-day. Therefore, the challenges of ASR need to be addressed with an efficient algorithm for speech recognition. The knowledge of human speech production, nature of speech signal and synthesizing speech using machine, the current speech recognition techniques available, and their loopholes and to address them becomes necessary so that areas of improvements can be highlighted.

The rest of the work is organized as follows. Section II deals with the related work on existing ASR systems. Section III deals with the general block diagram of ASR. Section IV is dedicated to the study of performance parameters used to measure the recognition rate of ASR. Section V discusses the applications and advantages of ASR. The conclusion and scope of the work on ASR is described in section VI.

II. Existing Models of Automatic Speech Recognition

Researchers and scientists have contributed a lot in surveying and proposing new techniques that will help in increasing the speech recognition accuracy. Vijayalakshmi et al., [5] highlights the architecture of ASR system using different approaches. Li Deng and Xiao Li [9] discussed Machine Learning (ML) paradigms which are motived by ASR applications. The cross-pollination of ML and ASR techniques are introduced to obtain improved results. Pandey et al., [13] discussed various databases generated for recognition of Indian languages. Swati et al., [22] explained various speech feature extraction and classification techniques and compared the various existing techniques for speech recognition. Itoh et al., [28] gives the performance measuring metrics normally used in speech recognition techniques. It is seen that Word Error Rate (WER) is widely used accuracy measurement metric for ASR.

Considering the recent advances in speech recognition, it would be no surprise to predict that by 2050 fifty percentages of the searches will be voice searches. Table 1 shows the growth of the speech recognition technology over the years.

Table 1. Major Advancements in Speech Recognition Technology

Year	Major Advancements
1784	Wolfgang invented Acoustic-Mechanical Speed Machine
1879	Thomas Edition invents the first dictation machine.
1952	Bell Labs releases Audrey to recognize only voice.
1962	IBM Shoebox understands sixteen English words.
1971	Harpy can comprehend 1,011 words and some phrases.
1986	IBM Tangora predicts upcoming phonemes in speech.
2006	NSA is able to isolate key words in recorded speech.
2008	Google brings speech recognition to mobile devices
2011	Apple launches Siri to facilitate voice-enabled assistant.

III. Overview of Automatic Speech Recognition

In this section the brief classification of speech recognition and architecture of automatic speech recognition system are explained in detail.

A. Classification of Speech Recognition Approach

Speech Recognition is the mapping of speech signal to text. Figure 2 shows the broad classification of speech recognition approaches. [5] [6].

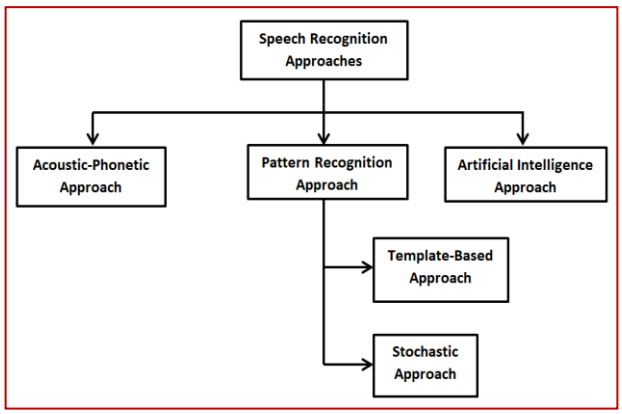


Fig.2. Classification of Speech Recognition Approaches

- ***Acoustic Phonetic Approach:*** Acoustics and Phonetics [7] [8] are the study of different sounds and phonemes of the language respectively. Acoustic phonetic approach of speech recognition relies on the postulates that every spoken language will have finite, definite and distinctive phonetic units whose properties are determined either by the time-domain signal or the spectrum of the signal. This approach will be useful for languages which are under- resourced.

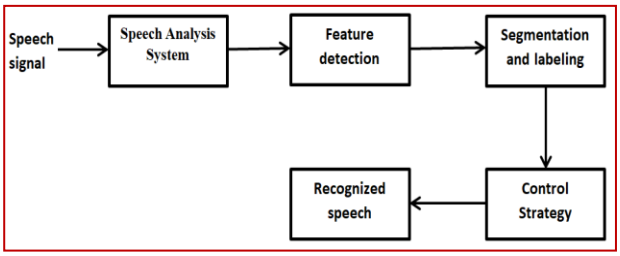


Fig.3. Acoustic Phonetic Approach

Figure 3 shows the general block diagram of the acoustic phonetic approach used for speech recognition. The speech signal when given as input to the speech analysis system will give the spectral representation of the time-varying signal. The features that provide speech acoustics are obtained from feature detection stage. Speech signal is quasi-stationary and its analysis is performed by segmentation of signal into region of stable behavior. The segmented regions are grouped into classes having similar behavior. Each class is labeled by a phoneme of the language whose behavior match with each other. Some of the areas of applications of this approach are multilingual speech recognition, accent classification, speech activity detection system, speech recognition for Asian languages, speech recognition for Indian languages, vocal melody extraction and so on.

- ***Pattern Recognition Approach:*** In this scheme there are two steps namely pattern training and pattern comparison. A training algorithm is selected to generate training samples which are correctly labeled using certain measurement parameters such as zero-crossing rate. A mathematical framework is developed to establish speech pattern representation in the form of a template or a stochastic model. The spoken words are compared with the possible patterns obtained in training phase for determining the identity of the unknown. Figure 4 shows the block diagram of pattern recognition approach and the classification of the Pattern recognition approach is tabulated in table 2.

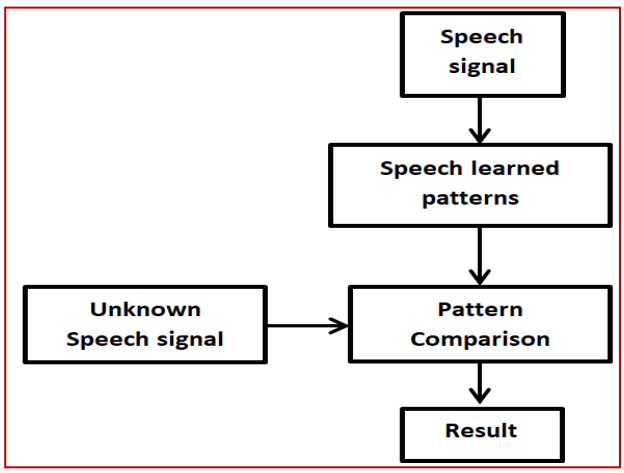


Fig.4. Pattern Recognition Approach

The spoken words are compared with the possible patterns obtained in training phase for determining the identity of the unknown. Figure 4 shows the block diagram of pattern recognition approach and the classification of the Pattern recognition approach is tabulated in table 2.

Table 2. Classification of Pattern Recognition Approach

Type	Description
Template Based Approach	A collection of prototypical speech patterns are stored as reference patterns. An unknown spoken utterance is matched with each of these reference templates and a category of the best matching pattern is selected.
Stochastic Approach	This approach is based on the use of probabilistic models so that uncertain or incomplete information, such as confusable sounds, speaker variability, contextual effects and homophone words can be addressed.

- ***Artificial Intelligence Approach:*** Artificial Intelligence [12] is the emerging technology having applications in almost every field of day-to-day life. Speech recognition remains one of the challenging areas in artificial intelligence. Artificial intelligence provides flexible recognition with a self-learning program to understand the voice modulation by listening and updating different pronunciations. The recognition of speech happens by using frequency depths to differentiate background noise from actual speech signal.

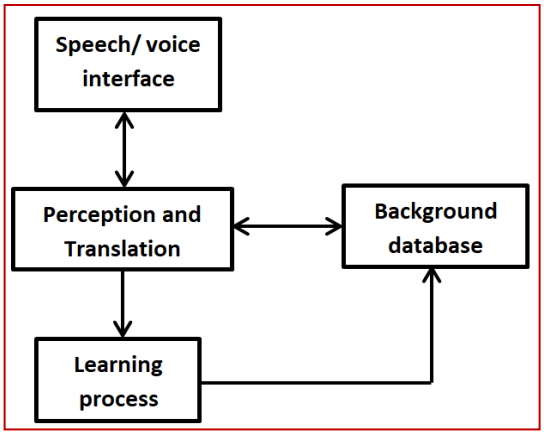


Fig.5. Artificial Intelligence Approach

Figure 5 shows the block diagram of speech recognition using artificial intelligence approach. Speech/ voice interface will analyze, learn and reply based on the voice input given to it. The perception and translation block will convert the signal to machine understandable form. The translated data helps in gathering properties of the voice signal and thereafter appropriate action are taken. Finally the comparison of database with the translated and learned information is performed and database will be updated based on the obtained results. In

this way, artificial intelligence approach is very effective in large vocabulary processing systems.

B. Architecture of Automatic Speech Recognition System.

Speech recognition is carried out in two stages namely training stage and testing stage. In training stage, the input to the system is known speech obtained from a database. The database sample is preprocessed and the significant features are extracted by applying various feature extraction techniques. In testing stage, the test sample is unknown and the acoustic analysis is performed. In order to discuss the stages involved in ASR architecture [6] [10], it is very important to be aware of the databases that serve as input to the ASR system. Figure 6 shows the architecture of ASR system.

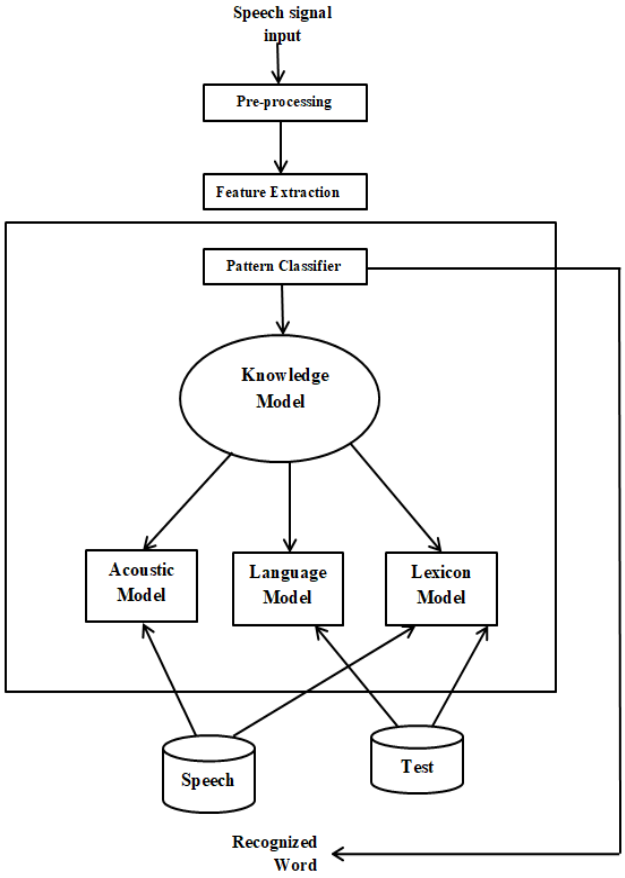


Fig.6. Architecture of Automatic Speech Recognition System

The various steps involved in ASR are database, preprocessing, feature-extraction and classification.

- ***Database:*** To achieve accurate speech recognition, databases used must be precisely collected and their scope has to be wide enough to provide coverage to all the acoustic-phonetic units of the speech. There are two main reasons to develop speech database. One is the use of database in research areas covering phonemes, acoustics, lexical and expressions of the language; and the other is in differentiating the speakers on the basis of age, gender, environment etc., while designing a speech database, the following factors must be taken into account.

- *Dictionary:* The vocabulary used in training must be of the same class for recognition. It should have coverage of all the phonemes of the speech class under consideration to obtain satisfactory results.
- *Number of sessions:* This gives the details of the entries made by a particular speaker in certain period of time. Also the time for recording of each speaker can be determined simultaneously. The smaller and efficient recording time is desirable.
- *Technical aspects:* The recordings are done in real-time and there is every possibility that it will have noise added to it either by the recording environment or by the technical equipment used for recording the speech.
- *Population of participating individuals:* The larger the population of participating individuals better will be the coverage of all the speech units and greater will be the recognition rate. However, huge databases will require more storage. Therefore, there has to be a tradeoff between the database size and the number of speech recordings.
- *Intended use:* Based on the application, speech corpus has to be obtained.
- *Intra-speaker and inter-speaker variability:* The database should have the recordings of the speaker in different emotions, expressions, accent etc., which is intra-speaker variability.

Collecting database for Indian languages is challenging as there are many languages and variations among the same language depending on the culture and geographical region of the speaker. Most of the time, there is no standard database available and hence the databases are designed as and when the need arises. Table 3 gives the overview of some of the Indian databases generated for various applications [13].

Table 3. Overview of Indian speech database

Database	Recording environment	Recording device used	Application of database	Languages
TIFR Mumbai and IIT Bombay	Noisy Environment	Cell Phone and Voice Recorders	Speech Recognition System for Agriculture Purpose	Marathi
Government P.G. College, Rishikesh	Noisy Environment	Standard Microphones	Speech Recognition System	Garhwali
TIFR Mumbai and C-DAC Noida	Noise-free environment	Standard Microphone	General Purpose	Hindi
IIT Kharagpur	Studio Environment	Standard Microphone	General Purpose	Hindi, Telugu, Tamil and Kannada

There are also some of the standard databases available for ready usage in speech recognition. Most of them are for English and foreign languages. Table 4 gives an overview of some of the standard and frequently used speech databases [14].

Table 4. Overview of Standard Databases

Database Name	No. of speakers	Language	Description	Advantages	Disadvantages
TIMIT and its derivatives	630 (438 Male/ 192 female)	American English	Most widely used database used for the development and analysis of the ASR systems. It also provides speech data useful in acquiring acoustic-phonetic knowledge.	Since the database is large it can be used for both speaker and speech recognition.	The database is not suitable for inter-variability analysis since it is of single-session.
ELSDSR	22 speakers (12Male/ 10Female)	English	It is recorded in a noise-free environment using a fixed microphone in a single-session and is suitable for speaker as well as speech recognition.	Useful for recognition of system where the intended speaker is not of American origin.	It is a single session database. Since recording is done in noise-free environment, recognition rate in noisy-environment will be low.
Polycost	134 speakers (74 Male /60 Female) from 13 European countries	Both native and non-native English	In the POLYCOST case, the English prompts are fully annotated in terms of word boundaries. The mother tongue prompts are just labeled at the word level with no segmentation	• Useful for intra as well as inter-speaker recognition. • Useful for language and accent recognition over Telephony line.	The database is not available free of cost and its cost is high.
Yoho	138 speakers (106 Male/32 Female).	English	The *YOHO* corpus was designed for evaluating speaker verification in text-dependent situation for secure access applications.	• Useful for digit recognition. • Used in secure access applications	• Accurate results are not obtained for word recognition. • Not useful for text-independent situations of speech recognition.
PBCM	100 speakers (50 Male/50 Female)	Non-native speakers (Hindi) of English	A phonetically balanced read speech corpus for code-mixed Hindi-English automatic speech recognition.	It is one of the first phonetically balanced corpus of code-mixed speech in an Indian language pair.	----

- ***Pre-processing:*** Speech signal, being continuous signal, has to be digitized using an analog-to-digital converter and given as input to the processing system [15] [16]. The first step is pre-processing of speech signal and it involves various stages as shown in figure 7.

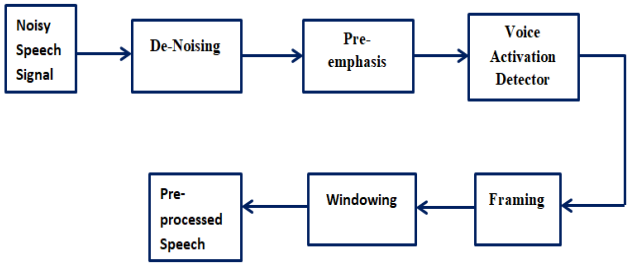
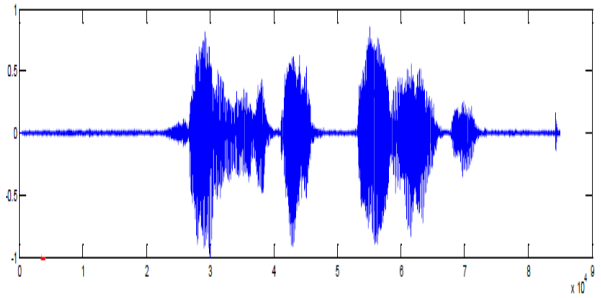
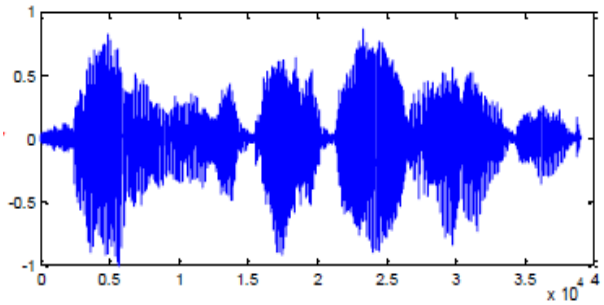


Fig.7. Stages of Preprocessing

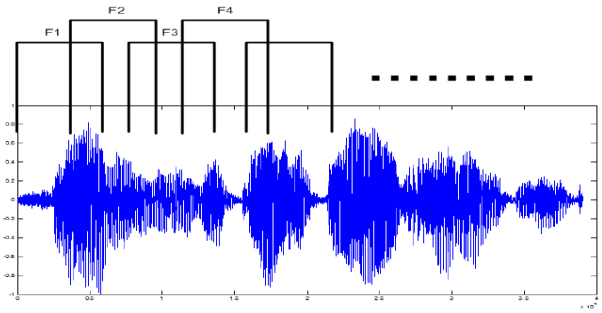
- Background or Ambient noise removal: Background noise is any unwanted signal and it has to be suppressed for recognizing the actual speech. Therefore, suitable filters are designed to remove unwanted interferences or noises present in the signal.
- Pre-emphasis: It removes high-frequency components, which are usually of little or no significance in processing.
- Voice Activity Detection: The Voice Activity Detectors (VADs) are designed to select voiced/ speech, the unvoiced/ non-speech sections and the silence regions in the signal using some of the classic parameters like zero-crossing rate, energy of the signal and autocorrelation function.
- Framing: Speech is a quasi-stationary signal and hence processing of speech has to be done for short period of time and also these frames are overlapped before analyzing so that vital information is not lost. To achieve this, speech signal is normally divided into frames of 20-30ms and overlapping of 30-50% of each frame with adjacent frames is performed.
- Windowing: The speech signal frames are multiplied with the windows of varying shapes so as emphasize the portions of greater importance and suppress the portions that are of little importance. The window co-efficient should be such that its weight is less at the region of discontinuity and more in the region of continuous speech.



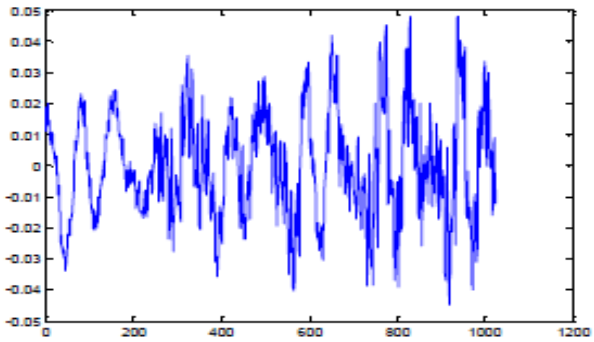
(a) : Original speech signal



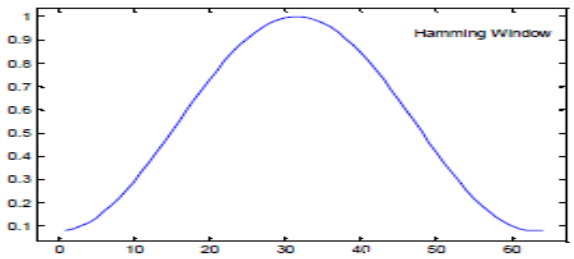
(b): After removal of silence region



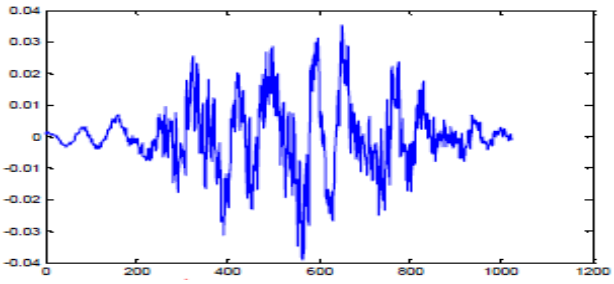
(c) : Pre-emphasis



(d): Framing 50% overlapping frames duration 23ms



(e): Hamming window



(f): Output of the Windowing block.

Fig.8. Effects of pre-processing of speech signal.

Figure 8 shows the effect of applying various stages of preprocessing on speech signal. Figure 2.(a) is the original waveform, figure (b) is the waveform obtained after removing the silence regions, figure (c) is the output of the pre-emphasis stage after suppressing high frequency components, figure (d) shows the waveform obtained after framing is done. Here frames of 23ms are chosen and the overlapping between frames is 50%. Figure (e) shows the graphical representation of Hamming window. Figure (f) shows the effect of applying hamming window on one of the frames [15] [16]. At the end of preprocessing stage, speech signal is in the form suitable for extracting useful features and hence aid in accurate recognition.

- ***Feature Extraction:*** The goal of feature extraction is to extract the significant information from the enhanced speech signal. The feature extraction techniques convert the preprocessed signal to a set of feature vectors and these vectors characterize the nature of the speech. Table 5 shows the comparison of some of the feature extraction techniques [17-20].

Table 5. Comparison of Feature Extraction Techniques

Method	Filter Bank	Technique Used	Merits	De-merits
MFCC (Mel-Frequency Cepstral Coefficients)	Mel filter bank is used for power spectrum	For the production of cepstral coefficients, MFCC uses logarithmic amplitude compression and IDFT operation.	• Useful for multi-speaker and multi-languages • Reliable for moderate to high sized vocabulary • It is easy to implement.	Noise is not suppressed
PLP (Perception Linear Prediction)	Bark scale is used which is of trapezoid-like filter shape	For estimating the frequency dependent volume sensitivity of hearing, PLP uses equal loudness pre-emphasis and cepstrum is computed by the linear prediction (LP) Coefficients.	Discards irrelevant information of the speech and thus improves speech recognition rate.	It gives less recognition rate than MFCC and RASTA techniques.
PNCC (Power-Normalized Cepstral Coefficients)	Gamma tone filters are used to simulate the behavior of the cochlea	PNCC includes medium time power bias removal which is used to increase the robustness. It is calculated using arithmetic to geometric mean ratio to estimate the reduction in quality of speech caused by noise.	Accuracy rate higher compared to MFCC and RASTA	The implementation is complex as well as complex.
LPC (Linear Prediction Coding)	Linear prediction coding uses autocorrelation method of autoregressive (AR) modeling to find the filter coefficients	It obtains the linear prediction coefficients predicts by minimizing the prediction error in the least square error.	• Low resource required. • Easy implementation	• Unable to distinguish words with same vowel Sounds. • Useful for only single speaker and single language • It is reliable for small vocabulary size.
RASTA (Relative Spectral Filtering)	Band-pass filters are deployed in logarithmic spectrum domain.	RASTA filter will band-pass all feature coefficient due to which noise becomes the additive portion. On low pass filtering the resulting spectra, the noise is suppressed and the spectrum is smoothened.	• Useful for multi- speakers and multi-languages. • Reliable for moderate sized vocabulary.	• It requires moderate to hard implementation.

- ***Pattern classifier:*** The feature vectors obtained from the feature extraction technique are compared with the test feature vector to find out the similarity index. Based on the similarities of feature vectors obtained, the pattern

Classifier makes decision to accept or reject the samples. Table 6 gives the comparison of various classifier techniques [22] [23].

- ***Knowledge model***: It helps in converting the features obtained from feature extraction stage into a sequence of words. This stage has three main sub-models namely acoustic model where individual phonemes or words are determined based on the feature vectors get a sequence of words and lexicon model converts the sequence of words into meaningful sentences using vocabulary of the language available in training database.

Table 6. Comparison of various classifier techniques

Classifier Type	Description	Learning Method used	Merits	De-merits
DTW (Dynamic Time Warping)	Mainly used to find out the similarities between two times based sequences using time normalized distance measurements. The one have minimum distance is classified as the recognized speech.	Unsupervised Learning Method	• Requires less storage space. • Beneficial for variable length.	There is cross channel issues which affect its performance.
HMM (Hidden Markov Model)	HMM is a stochastic process. In Hidden Markov model, the previous state is not directly visible to the observer. But the output depending on that state is visible.	Unsupervised Learning Method	Efficient performance compared to DTW.	• Requires more storage space. • Computational more complex than DTW.
GMM (Gaussian Mixture Model)	GMM is a statistical method for estimating the spectral parameters using Expectation Maximization (EM) Algorithm.	Unsupervised Learning Method	Training and test data requirement is less.	• There is a tradeoff in performance between that of DTW and HMM.
VQ (Vector Quantization)	VQ comes under lossy compression technique. VQ provides multidimensional representation of data. Decision boundaries and reconstruction levels are two important terms used in VQ.	Unsupervised Learning Method	• Computationally less complex. • Effective time utilization	• Encoding in real-time is complex.
SVM (Support Vector Machine)	Basic of SVM is to create a hyper plane. This hyper plane differentiates the features. In binary SVM, features are classified into two classes, each class for recognized and unrecognized speaker.	Supervised Learning Method	Simple operation.	• Binary SVM has limitation in speaker recognition.
Neural Network (NN) Modeling Technique	They are being used in solving complex identification tasks.	May use Supervised or unsupervised learning method.	• Can control low quality speech signal. • Can be used in noisy data environment. • Provides better accuracy than HMM.	• Optimal configuration selection is not easy.

IV. Performance Parameters

A. *Performance Measures*

Two most widely used speech performance measuring parameters are Word Error Rate (WER) and Command Success Rate (CSR). These are discussed in brief here.

- ***Word Error Rate (WER):*** It is the most widely used speech recognition metric. WER [24] is used to measure recognition accuracy at word level. If 'N' is the number of words in reference speech, 'S' is the number of substitutions, 'I' the number of insertions and 'D' the number of deletions, and then the formula for calculating WER is as given by equation 1.

$$WER = \frac{S+I+D}{N} \quad (1)$$

The lower the value of WER, the better is the recognition accuracy.

- ***Command Success Rate (CSR):*** It is defined as the ratio of number of correctly recognized sentences to the total number of sentences used as input. If 'TS' is the total number of sentences and 'RS' is the number of correctly recognized sentences, then CSR [24] is given by equation [2].

$$CSR = \frac{RS}{TS} \quad (2)$$

The higher the value of CSR, the better is the recognition accuracy.

B. Software tools for Automatic Speech Recognition.

Many software tools are available to perform speech recognition which could be either open source or licensed. Some of them are described as follows.

- ***Simon:*** It is one of the popular speech recognition toolkit used for Linux as well as Windows operating system. The biggest strength of Simon toolkit is its friendly user-interface and simple structure. It is an open source speech recognition toolkit and the programming is done in C++.
- ***Julius:*** Julius provides very high performance of speech recognition. It performs two-way large vocabulary continuous speech recognition that helps the speech-related work for scientists and researchers. It is open source software developed for recognizing English and Japanese languages and is available offline.

Windows operating system has in-built speech recognition software supporting English, French, Spanish, German, Japanese, Simplified Chinese, and Traditional Chinese languages.

- Smartphones and mobile devices have inbuilt speech recognizers that enable dial-by-voice features. They make use of natural-language processing techniques.
- ***CMU Sphinx:*** Sphinx and its higher versions are open source speech recognition software for English language used for recognizing continuous, speaker-independent speech signal. The coding is done in C/Java/Python and is available in multi-platform operating systems (Sphinx 4 version).
- ***HTK (HMM Tool kit):*** It is a proprietary software toolkit developed to handle HMM. This toolkit is applicable only for recognizing English language and the software is coded using C language. It can also be used for speech synthesis, DNA sequencing and character recognition.

V. Advantages and Applications of Automatic Speech Recognition

Speech Recognition is adopted in almost all the areas of technology. The main aim of speech recognition is to have automation of devices or applications so as to have robust and accurate typing. The major application areas of speech recognition are listed in Table 7 along with their advantages [25] [26].

Table 7. Applications and Advantages of Speech recognition

Application Area	Advantages/ Description
Car Environment	Automatic vehicle operation command systems Speech-controlled navigation systems Speech-controlled cabin systems
Air Traffic Control (ATC)	ASR of Controller utterances for simulator training and pseudo pilots. Offline transcription of simulated and field data for research and forensic purposes. Onboard ASR for Avionic Instruments, Controller workload estimations.
Medical Assistance	To validate the performance of patient in communication disorder To make speech disorder assessment by speech pathologist fast and inexpensive. Physicians dictate to the computers directly for development of reports in areas like radiology, pathology, and endoscopy. Evaluation of voice and speech disorders in head and neck cancer.
Industrial Applications	By endowing the wheelchair with new Human–Machine Interfaces (HMI) and increasing the wheelchair navigation autonomy, wheelchair users can be socially independent. Voice input computers are used in many industrial inspection applications allowing data to be input directly into a computer without keying or transcription.
Forensic and Law Enforcement	FASR (Forensic Automatic Speech Recognition) is used for judicial and law enforcement purposes for identification of speech samples of suspected speaker.
Telecommunications Industry	Generate new revenue by including voice banking services, voice prompter, directory assistance, call completion, information services, customer care, computer-telephony integration, voice dictation. Achieve cost reduction by automation of operator services, automation of directory assistance, voice dialing etc.
Home Automation and Security Access Control	Powered wheelchairs provide unique mobility for the disabled and elderly with motor impairments. ASR can provide an unconventional and more secure means of permitting entry without any need of remembering password, identity numbers, lock combination, etc. or the use of keys, magnetic card or any other device which can be easily stolen.
Education and Learning through ASR	Mobile learning: Speech recognition supported games on mobile devices could help in improving the literacy in the developing countries.
Information Technology and Consumer Electronics	Speech recognition support enables internet access for the people with mobility impairments, people with visual impairments and senior citizens. Google, Apple & Microsoft have successfully demonstrated the use of speech in as search engine and accessing various applications on mobile devices.

VI. Conclusion

In this paper a brief discussion on Automatic Speech Recognition (ASR) system which is typically a speech-to-text converter is presented. In order to recognize the areas of further research in ASR, one must be aware of the current approaches, challenges faced by each and issues that needs to be addressed. In addition to this, human speech production mechanism and the speech recognition techniques are addressed. The performance parameters that measure the accuracy of the system in recognizing the speech signal and software tools employed in speech recognition described in detail.

References

[1] Dani Byrd and Elliot Saltzman, "Speech Production", *The Handbook of Brain Theory and Neural Networks*, pp. 1072-1076, 2002.

[2] Harshalata Petkar, "A Review of Challenges in Automatic Speech Recognition", *International Journal of Computer Applications*, vol. 151, no. 3, pp. 23-26, 2016.

[3] Rashmi M, Urmila S and V M Thakare, "Opportunities and Challenges in Automatic Speech Recognition", *International Conference on Biomedical Engineering and Assistive Technologies*, pp. 1-5, 2010.

[4] Davinder P Sharma and Jamin Atkins, "Automatic speech recognition systems challenges and recent implementation trends", *International Journal on Signal and Imaging Systems Engineering*, vol. 7, no. 4, pp. 219-234, 2014.

[5] Vijayalakshmi A, Midhun Jimmy and Moksha Nair, "A study on Automatic Speech Recognition Techniques", *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 4, no. 3, pp. 614-617, 2015.

[6] Preeti Saini and Parneet Kaur, "Automatic Speech Recognition: A Review", *International Journal of Engineering Trends and Technology*, vol. 4, no. 2, pp. 132-136, 2013.

[7] Bhagat Parabattina and Pradip Das, "Acoustic Phonetic Approach for Speech Recognition A Review", *International Conference of the Neurosurgical Society*, pp. 1-6, 2016.

[8] Rohini Shinde and V P Pawar, "A Review on Acoustic Phonetic Approach for Marathi Speech Recognition", *International Journal of Computer Applications*, vol. 59, no. 2, pp. 40-44, 2012.

[9] Deng, Li, and Xiao Li, "Machine learning paradigms for speech recognition", vol. 21, no. 5, pp. 1060-1089, 2013.

[10] W Ghai and Navdeep Singh, "Literature Review on Automatic Speech Recognition", *International Journal of Computer Applications*, vol. 44, no. 8, pp. 42-20, 2012.

[11] B S Atal, "A Pattern Recognition Approach to Voiced Unvoiced Silence Classification with Applications to Speech Recognition", *IEEE transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 3, pp. 201-212, 1976.

[12] Mohammad A. Bah, Abdusahmad A and M. A. Eyad, "Artificial Intelligence Technigue for Speech Recognition based on Neural Networks", *Oriental Journal of Computer Science and Technology*, vol. 7, no. 3, pp. 331-336, 2014.

[13] Pukhraj P Shrishrimal, Ratnadeep R Deshmukh and Vishal M Waghmare, "Indian Language Speech Database A Review", *Internal Journal of Computer Applications*, vol. 47, no. 5, pp. 17-21, 2012.

[14] Ayushi Pandey, B Srivastava, Rohit Kumar, B Nellore, K Teja and S Gangashetty, "Phonetically Balanced Code-Mixed Speech Corpus for Hindi-English Automatic Speech Recognition", *International Conference* on *Language Resources and Evaluation*, pp. 1-6, 2018.

[15] P L Chithra and R Aparna, "Performance Analysis of Windowing Techniques in Automatic Speech Signal Segmentation", *Indian Journal of Science and Technology*, vol. 8, no. 29, pp. 1-7, 2015.

[16] Nisha, "Voice Recognition Technique Review", *International Journal for Research in Applied Science & Engineering Technology*, vol. 5, no. 5, 28-35, 2017.

[17] Smita Magre, Ratnadeep Deshmukh and Pukhraj Shrishrimal, "A Comparative Study on Feature Extraction Techniques in Speech Recognition", *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 3, no. 12, pp. 18006-18016, 2014.

[18] Shreya Narang and Divya Gupta, "Speech Feature Extraction Techniques Review", *International Journal of Computer Science and Mobile Computing*, vol. 4 no. 3, pp. 107-114, 2015.

[19] Arshpreet Kaur, Amitoj Singh and Virender Kadyan, "Correlative consideration concerning feature extraction techniques for Speech Recognition Review", *International Conference on Circuit, Power and Computing Technologies*, pp. 1-4, 2016.

[20] Anjali Garg and Poonam Sharma, "Survey on Acoustic Modeling and Feature Extraction for Speech Recognition", *International Conference on Computing for Sustainable Global Development*, pp. 2291-2295, 2016.

[21] Saikat Basu, Jaybrata Chakraborty, Arnab Bag and Md. Aftabuddin," A Review on Emotion Recognition using Speech", *International Conference on Inventive Communication and Computational Technologies*, pp. 109-114, 2017.

[22] Swathy M S and Mahesh K R, "Review on Feature Extraction and Classification Techniques in Speaker Recognition", *International Journal of Engineering Research and General Science*, vol. 5, no. 2, pp. 78-83, 2017.

[23] Nidhi Desai, Kinnal Dhameliya and Vijayendra Desai, "Feature Extraction and Classification Techniques for Speech Recognition Review", *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 12, pp. 367-371, 2013.

[24] Ahmed Ali and Steve Renal, "Word Error Rate Estimation for Speech Recognition e-WER", *International conference on Computational Linguistics*, pp. 20-24, 2018.

[25] Pratiksha C Raut and Seema U Deoghare, "Automatic Speech Recognition and its Applications", *International Research Journal of Engineering and Technology*, vol. 3, no. 5, pp. 2368-2371, 2016.

[26] Jayashri Vajpai and Avnish Bora, "Industrial Applications of Automatic Speech Recognition Systems", *International Journal of Engineering Research and Applications*, vol. 6, no. 3, pp. 88-95, 2016.

[27] Sunita Dixit and M Yusuf Mulge, "Speech Processing: A Review", *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 3, no. 8, pp. 2775-2778, 2014.

[28] Nobuyasu Itoh, Gakuto Kurata, Ryuki Tachibana and Masafumi Nishimura, "A Metric for Evaluating Speech Recognition Accuracy based on Human Perception", *International Journal of Information Processing Society of Japan*, vol. 104, no. 11, pp 1-7, 2014.

[29] Trishna Barman and Nabamita Deb, "State of the Art Review of Speech Recognition using Genetic Algorithm", *International Conference on Power, Control, Signals and Instrumentation Engineering*, pp. 2944 – 2946, 2017.
[30] H Gupta and D S Wadhwa, "Speech feature extraction and recognition using genetic algorithm", *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, no. 1, pp. 363–369, 2014.

Authors' Profiles

Ms. Vidyashree Kanabur received the Bachelor degree in Electronics and Communication Engineering at KLE's College of Engineering and Technology, Belagavi and her masters in the stream of Signal Processing at Siddaganga Institute of Technology, Tumkur. She is currently working as Assistant Professor in the department of Electronics & Communication Engineering at S. G. Balekundri Institute of Technology, Belagavi. Her area of research interests includes Signal Processing and Biometrics. She is a life member of International Association of Engineers.

Mr. Sunil S Harakannanavar completed his Bachelor of Engineering in the stream of Electronics & Communication from Sri Taralabalu Jagadguru Institute of Technology, Ranebennur and his masters in the field of Microelectronics and Control Systems from Nitte Mahalinga Adyanthaya Memorial Institute of Technology, Nitte. Presently he is working as Assistant Professor with S. G. Balekundri Institute of Technology Belagavi. He is pursuing his Ph.D at Visvesvaraya Technological University, Belagavi and his area of interests includes Computer Vision, Pattern Recognition and Biometrics. He is a life member of Indian Society for Technical Education, New Delhi and Institute for Exploring Advances in Engineering. He is a life member of International Association of Engineers.

Dr. Dattaprasad Torse has received his Ph.D from Visvesvaraya Technological University, Belagavi. He received his master of engineering from Amravati University in Digital Electronics. He has published over 16 research papers on EEG signal analysis in journal and conferences. He is currently Associate Professor in the Department of Electronics and Communication Engineering, KLS Gogte Institute of Technology, Belagavi, Karnataka, India. He is a member of IEEE.

How to cite this paper: Vidyashree Kanabur, Sunil S Harakannanavar, Dattaprasad Torse, "An Extensive Review of Feature Extraction Techniques, Challenges and Trends in Automatic Speech Recognition", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.11, No.5, pp. 1-12, 2019.DOI: 10.5815/ijigsp.2019.05.01