Automatic Speech Recognition System: A Review

Neerja Arora
Assistant Professor
KIIT College of Engineering
Gurgaon,India

ABSTRACT
Speech is the most prominent & primary mode of Communication among human beings. Now-a-days Speech also has potential of being important mode of interaction with computers. This paper gives an overview of Automatic Speech Recognition System, Classification of Speech Recognition System and also includes overview of the steps followed for developing the Speech Recognition System in stages. This paper also helps in choosing the tool and technique along with their relative merits & demerits. A comparative study of different techniques is also included in this paper.

Keywords
Automatic Speech Recognition (ASR), ASR classification, Speech Analysis, Feature Extraction, Modelling Techniques, Language Modelling, Testing, ASR Tools.

1. INTRODUCTION
Voice is a part of biometric and it contains behavioural information. Voice processing can be done by two types:

- Speech Recognition
- Speaker Identification

In speech recognition, it recognizes the speech what user is speaking whereas in speaker identification, it identifies the user, who is speaking.

Speech recognition system is a natural way for the human to machine interaction. Automatic Speech Recognition is advance way to operate computer without much efforts through speech only. Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Rudimentary speech recognition software has a limited vocabulary of words and phrases and may only identify them, if they are spoken very clearly. More sophisticated software has the ability to accept natural speech. Speech recognition applications include call routing, speech-to-text, voice dialing and voice search.

The terms "speech recognition" and "voice recognition" are sometimes used interchangeably. However, the two terms mean different things. Speech recognition is used to identify words in spoken language. Voice recognition is a biometric technology used to identify a particular individual's voice.

1.1 Automatic Speech Recognition System
Automatic Speech Recognition **or** ASR**,** as it's known in short, is the technology that allows human beings to use their voices to speak with a computer interface in a way that, in its most sophisticated variations, resembles normal human conversation. It can be defined as the independent, computer-driven transcription of spoken language into readable text in real time.

2. TYPES OF ASR SYSTEMS
Speech recognition systems can be classified into number of classes as follows:

2.1 On the Basis of speaking mode
It means that how the words are spoken whether in isolated or in connected**.**

2.1.1 Isolated Words Recognizers
They accept one word at a time and good for situations where the user is required to give only one word responses or commands

2.1.2 Connected Words Recognizers
They are similar to isolated word Recognizers, but allow separate utterances to be run together˝ with a minimal pause amongst them.

2.2 On the Basis of speaking Style
It includes that whether the speech is continuous or spontaneous.

2.2.1 Continuous Speech Recognizers
This type of Recognizers allows users to speak almost naturally, while the computer basically determines the content. It includes a great deal of "co articulation", where adjacent words run together without pauses or any other apparent division between words.

2.2.2 Spontaneous Speech Recognizers
They recognize speech that is not rehearsed but natural and allow us to speak spontaneously. They are able to handle a wide variety of natural speech features such as words being run together and even slight stutters. Spontaneous (and unrehearsed) speech may also include mispronunciations, false-starts, and non-words.

2.3 On the Basis of Enrolment
The voice of every speaker is unique due to their specific physical body and personality. Enrolment is by two ways one is Speakerdependent and other is Speaker independent.

2.3.1 Speaker dependent Recognizers
They are designed for a specific speaker and are generally more precise for the particular speaker, but much less precise for other speakers. These systems require a particularuser to train the system according to his or her voice.

2.3.2 Speaker independent Recognizers
They are developed for variety of speakers. It recognizes the speech patterns of a large group of people. These systems do not require a particular user to train the system i.e. they are developed to operate for any speaker.

2.4 On the Basis of Vocabulary

The size of vocabulary influences the complexity, processing requirements and the accuracy of the system. It is simple to discriminate a small set of words, but error rates increase as the vocabulary size increases. In ASR systems the types of vocabularies can be categorized as follows:

1) *Small vocabulary set* - tens of words
2) *Medium vocabulary set* - hundreds of words
3) *Large vocabulary set* - thousands of words
4) *Very-large vocabulary set* - tens of thousands of words.
5) *Out-of-Vocabulary*- Mapping a word from the vocabulary into the unknown word.

3. WORKING OF ASR SYSTEM

Figure 1, shows the basic block diagram of the Automatic Speech Recognition System which explains the working of ASR system. It is divided into number of phases which are explained below:

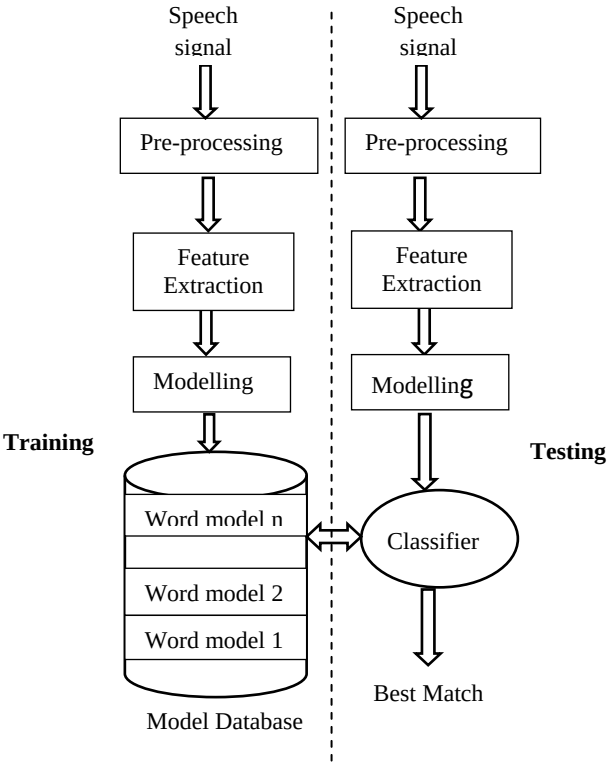

Fig 1: Overview of working of ASR system

3.1 Data Collection

The first step of ASR system is to collect the data which is to be processed. Initially the raw data (i.e. sentences to be spoken by the speakers) are collected from various resources. Each sentence is then uttered by every speaker in the form of soundwaves. These sound waves are then captured by a microphone and converted into electrical signals, which are then converted into digital form by using recording software and stored in the form of wave file (with .wav extension).

3.2 Pre-Processing

At the time of recording the speech, the interference due to noise mainly occurs, due to which the performance of the system can be degraded. Before feeding the speech signal to feature extraction block the noise contained in speech signal must be removed. Preprocessing does this task. It removes the noise based on zero-crossing rate and energy. The output speech contains the desired information and noise is eliminated.

3.3 Pronounciation Dictionary

To start training of the speech recognition system, pronounciation dictionary is first created. Phoneme is the basic unit of sound in any language. The Pronounciation dictionary contains words and mapping to their phonetic contents. The words covered in the dictionary can only be recognized by the recognizer.

3.4 Annotation

In this, we first create transcriptions of the sentences (i.e. segmentation) spoken by the speaker into words or phonemes by carefully listening the sentence, make necessary changes in the transcriptions so that resulting transcriptions match the utterances in the audio files and then annotate all the transcriptions by using software and the pronounciation dictionary created earlier. The resulted transcription file is then stored as (.lab file).

3.5 Feature Extraction

After Annotation, next step is to extract the features from (.lab files) created in above step and representing them using an appropriate data model of the input signal. Extracted features are assumed to contain only the relevant information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of irrelevant information such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone) from the recorded speech.

3.5.1 Feature Extraction Techniques

Various feature extraction techniques that are commonly used in speech recognition are:

i. Mel-Frequency Cepstrum Coefficients (MFCC)
ii. Linear Predictive Coding (LPC)
iii. Linear Prediction Cepstral Coefficients (LPCC)
iv. Perceptual Linear Prediction (PLP)
v. Linear Discriminant Analysis (LDA)
vi. Discrete Wavelet Transform (DWT)
vii. Relative Spectral (RASTA-PLP)
viii. Principal Component analysis (PCA)

3.6 Modelling

The objective of modelling is to generate speaker specific models using speaker specific feature vector. These models can be word-based or phoneme-based.

3.6.1 Acoustic model

Acoustic model typically refers to the file containing statistical representationsfor the feature vector sequences computed from the speech waveform. In case of phoneme based acoustic model, statistical representations of each of the distinct sounds that makes up a word are stored in a file.

It can be implemented by using different approaches such as HMM, ANNs, dynamic Bayesian networks (DBN), support

vector machines (SVM). HMM is the approach which is most commonly used.

3.6.2 Language model

The main goal of generating language model is to improve the performance of various speech recognition systems. Basically, Language model is a file which contains a large list of words and their probability of occurrence, this means finding the best possible estimate for P(Wi), that is the probability for the word Wi to occur. The most commonly used method in speech recognition for estimating P(Wi) is n-gram language modelling.

3.7 Testing

It is the most important and last step in the speech recognition process. Testing is performed for finding the best match for the incoming acoustic model with the trained model (generated as a result of Training). Number of algorithms available for matching and to take actual decision about recognition of a speech utterance by combining and optimizing the information conveyed by the acoustic and language models.

4. VARIOUS TOOLS USED FOR ASR

Number of tools are available for developing ASR system, some of them are:

4.1 For speech recording and editing

Audacity: It is free, open source software available with latest version of 2.0 which can run on wide range of OS platforms and meant for recording and editing sounds.

Goldwave: It is a free, highly rated, professional digital audio editor available withlatest version of v6.24. It's fully loaded to do everything from the simplest recording and editing to the most sophisticated audio processing, restoration, enhancements, and conversions. It is easy to learn and use.

4.2 For Annotation

Wavesurfer: It is an open source software for sound visualization and manipulation. It is mainly used for speech/sound analysis and sound annotation/transcription. It is a customizable, extensible, embeddable and Multi-platform tool.

4.3 Other tools for Recognition

PRAAT: It is free software with latest version 6.0 which can run on wide range of OS platforms and is a very flexible tool to do speech analysis. It offers a wide range of standard and non-standard procedures, including spectrographic analysis, articulatory synthesis, and neural networks.

CSL: Computerized Speech Lab is a highly advanced speech and signal processing workstation (software and hardware). It possesses robust hardware for data acquisition and a versatile suite of software for speech analysis.

HTK: The basic application of open source Hidden Markov Toolkit (HTK), written completely in ANSI C, is to build and manipulate hidden Markov models.

SPHINX: Sphinx 4 is a latest version of Sphinx series of speech recognizer tools, written completely in Java programming language. It provides a more flexible framework for research in speech recognition

KALDI: Kaldi is an open-source toolkit for speech recognition written in C++ and licensed under the Apache License v2.0. The goal of Kaldi is to have modern and flexible code that is easy to understand, modify and extend. Kaldi is available on SourceForge. The tools compile on the commonly used Unix-like systems and on Microsoft Windows. It provides a speech recognition system based on finite-state transducers.

5. PERFORMANCE OF ASR SYSTEMS

The recognition performance evaluation of an ASR system must be measured on a database different from the training database of speech. The performance of speech recognition system is usually specified in terms of accuracy and speed. Accuracy is computed by word Recognition rate, whereas speed is measured with the real time factor.

$$\text{(WRR) Accuracy} = \frac{\text{No. of words correctly Recognized} * 100}{\text{Total no. of words}}$$

When reporting the performance of a speech recognition system, sometimes Word Error Rate (WRR) is used instead of Word Recognition Rate (WRR), which can be calculated as,

$$\text{(Word Error Rate)WER} = 1 - \text{WRR}.$$

6. CURRENT CHALLENGES IN SPEECH RECOGNITION

The performance of the ASR Systems degrades due to the occurrence of noise from the outer sources. Accuracy and reliability of the system is affected by the unwanted input and low output result from the system. The fault tolerance capacity lacks in this case. User responsiveness is also one of the challenges, it happens when the resources are not ready and user starts to speak the command and then it leads to problem of synchronizing the data with multiple applications (media, phone, navigation).

Table 1: Comparision Of Various Speech Recognition Techniques

Method	Description	Merits	Demerits	Applications
DTW (Dynamic Time Warping)	• It is a statistical approach used to recognize speech. • Its main principle is to compare two dynamic patterns and measure its similarity by calculating a minimum distance between them.	• It is powerful for measuring similarity between two time series which may vary in time or speed. • The training procedure in DTW is very simple and fast.	The main problem is to prepare the reference template. Single template is not sufficient.	It is used in small scale embedded speech recognition system such as those embedded in cell phones.
VQ (Vector Quantization)	• It is the pattern classification technique applied to speech data to	The density matching property of this Vector quantization technique is	It does not take into account the temporal evolution of the signals	It is useful for speech coders, i.e., efficient data

	forms a representative set of feature vectors. • It is a fixed-to-fixed length algorithm	very powerful, mainly in the density of large and high dimensional data identification	(speech, signature, etc.) because all the vectors are mixed up.	reduction. This method is also appropriate for lossy data compression.
SVM (Support Vector Machine)	• It is a simple and effective method for classification of speech or speaker recognition. • SVM is a binary nonlinear classifier capable of guessing whether an input vector x belongs to a class-1 or class-2 category.	• Minimize the structural risk which results in better generalization ability. • Increase the robustness of the system. • Training is relatively easy. • Local optimality is not needed. • It scales relatively well for high dimensional data	• Good kernel function is needed. • Requires full labeling of input data • The SVM is only directly applicable for two-class tasks.	The application of SVMs can be speaker and language recognition. They are helpful in text and hypertext categorization.Classification of images can also be performed using SVMs
HMM (Hidden Markov Model)	• It is a mathematical framework or statistical model of a sequence of feature vector observations. • In HMM state sequences are hidden and the observations are probabilistic functions of the state.	• It is fast in its initial training, and when a new voice is used in the training process to create a new HMM model. • Performs quite well in noisy environments because every sound entity is treated separately.	• large priori modelling assumptions about the data have to make • Amount of data and no. of parameters that need to be set during training is huge. • It does not minimize the probability of observation of instances from other classes.	It has been used with success in low level NLP processes such as phrase chunking, extracting necessary information from documents and part-of-speech tagging.
Neural networks	• They are also statistical model and similar to HMM. • They use connection strengths and functions for state transitions • Neural networks are fundamentally parallel. • Frequencies in speech, occur in parallel, while syllable series and words are essentially serial.	• Performs very well at learning phoneme probability from highly parallel audio input. • It can handle large amount of data sets. • It can learn according to discriminative criteria • Do not require strong assumptions about the input data. • It produces reasonable outputs for inputs which have not been taught before how to deal with.	• Does not perform well for more complex tasks as continuous speech recognition. • Inability to build speech model even though the recurrent structures are defined.	They are well suited for classification problems, character recognition, image compression problems.

7. CONCLUSION

Speech recognition based applications are getting enormous popularity as they prove to be essential for both the civilized and weakly educated people. These days, lot of research is being carried out and further lot of work needs to be done in the context of ASR for different languages.In this paper, brief description of Automatic Speech Recognition System, its working and classification of ASR systems on the basis of various factors have been discussed and also put forth performance measures, current challenges, some of the tools available and the approaches used for developing ASR system by different researchers with their merits and demerits.

8. REFERENCES

[1] Dr. Kavitha, Nachammai, Ranjani, Shifali., "Speech Based Voice Recognition System for Natural Language Processing", "In International Journal of ComputerScience and Information Technologies", Vol. 5, 2014.

[2] Anjali Bala, Abhijeet Kumar, Nidhika Birla, "Voice Command Recognition System Based on MFCC and DTW", "International Journal of Engineering Science and Technology", Vol. 2 (12), 7335-7342, 2010.

[3] Jayashree Padmanabhan ,Melvin Jose Johnson Prem kumar, "Machine Learning in Automatic Speech Recognition: A Survey", "IETE Technical Review, Taylor & Francis", pp-1-13, 2015.

[4] Rashmi C R,"Review of Algorithms and Applications in Speech Recognition System", "International Journal of Computer Science and Information Technologies", Vol. 5 (4) , 5258-5262, 2014.

[5] Manav Bhaykar, Jainath Yadav, and K. Sreenivasa Rao, "Speaker Dependent, Speaker Independent and Cross Language Emotion Recognition from Speech Using GMM and HMM", IEEE, 2013.

[6] Preeti Saini, Parneet Kaur, "Automatic Speech Recognition: A Review", "International Journal of

Engineering Trends and Technology- Vol 4, Issue 2-2013.

[7] Mayur R Gamit, Prof. Kinnal Dhameliya, Dr. Ninad S. Bhatt, "Classification Techniques for Speech Recognition: A Review", "International Journal of Emerging Technology and Advanced Engineering",Vol 5, 2250-2459,2015.

[8] C. Sunitha Ram, Dr. R. Ponnusamy, "An Effective Automatic Speech Emotion Recognition for Tamil Language using Support Vector Machine", "International Conference on Issues and Challenges in Intelligent Computing Techniques", 2014.

[9] Ahmad A. M. Abushariah, Teddy S. Gunawan, Mohammad A. M. Abushariah, "English Digit Speech Recognition System Based on Hidden Markov Model", "International Conference on Computer and Communication Engineering", IEEE, May 2010.

[10] UtpalBhattacharjee, "A Comparative Study of LPCC and MFCC Features for the Recognition of Assamese Phonemes", "International Journal of Engineering Research and Technology (IJERT)", Vol.2, Issue 1, January 2013.

[11] Jorge MARTINEZ, Hector PEREZ, Enrique ESCAMILLA, Masahisa Mabo SUZUKI, "Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) Techniques", IEEE, pp 248-251, 2012.

[12] R K Aggarwal and M. Dave, "Markov Modeling in Hindi Speech Recognition System: A Review", "CSI Journal of Computing", vol. 1, no.1,pp. 38-47, 2012.

[13] Kuldeep Kumar, Ankita Jain and R.K. Aggarwal, "A Hindi speech recognition system for connected words using HTK", International Journal of Computational Systems Engineering, vol. 1, no. 1, pp. 25-32, 2012.

[14] Jay Patadia, Alpa Reshamwala, "Feature Extraction Approach in Emotional Speech Recognition System", "International Journal of Advanced Research in Computer Science and Software Engineering", 2277 128X, Vol 6, Issue 5, 2016.

[15] Prof. Pisal Ranjeet, Thite Prakash, Satpute Amruta & Shingade Monali, "Automatic Speech Recognition System", "Imperial jounal of Interdisciplinary Reasearch (IJIR)", 2454-1362, Vol-2, Issue-3 , 2016.

[16]Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique", "International Journal of Computer Applications",0975-8887,Vol 10, 2010.