IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Review and Analysis of Speech Recognition Techniques for Mobile Devices

Gulbakshee J. Dharmale[1*], Dipti D. Patil[2], Vilas M. Thakare[3]

[1,3] Computer Science Dept., SGB Amravati University, Amravati, INDIA.
[2] Information Technology Dept., MKSSS's Cummins College of Engineering for Women, Pune, INDIA.

Available online at: www.ijcseonline.org

Accepted: 21/Jan/2019, Published:31/Jan/2019

Abstract— The most widely correspondence modes for individuals is speech. People use speech as another mode to convey information without lifting a finger with the help of speech recognition applications. This paper presents a study about execution of speech recognition debilitates inside considering resounding and even low levels of including clamor. Speech recognition interfaces in local lingo which connects the people to use the advancement to more imperative degree without the data by operating with a PC. An extraordinary measure of research has done in different areas of speech recognition and its applications since three decades. A specific overlook on speech recognition has been done in this paper, introducing the design, databases, speech parameterization, techniques, attributes, apparatuses, applications and issues. Also analyzed performance with respective accuracy for different available techniques from various research work.

Keywords—Automatic Speech Recognition (ASR), Mel Frequency Cepstral Coefficient (MFCC), Hidden Markav model (HMM), Gaussian Mixture Model (GMM)

I. INTRODUCTION

Speech is the general and well-known method for correspondence among general population. It can expectedly trade it's communication via speech. The modern digitized world has made new improvements. The percentage of HCI or human computer interface alliance leads to the progress in investigating on recognition of speech (SR) [1]. Speech recognition has been an efficient method which is tested over a wide range among other algorithms.

Speech evaluation or SR is the farthest point of a computer system or program to obtain the commitment via speech which includes words and enunciations and after that modify or convert it to an understandable by machine setup. This is customarily used to evaluate a device, execute orchestrates on the devices by press any gets or with no utilizing a console and mouse. The devices contain a program called Automatic Speech Recognition (ASR) which perceives and see the voice and then change over the speech to text and to activate some activity [1].

The basic ASR device was used a bit in 1952 and recognized a single digit spoken by a user. Nowadays, automatic speech recognizers or ASR systems have been utilized in various zones to evaluate this speech recognition process. ASR is most routinely used in one or many of following divisions [1]:

a) Medical Center
b) Border Organizations
c) Communications
d) Robot development

Fundamental development in Speech Recognition is a specific extreme target to build up a correspondence between the overall public and the machine. Various analysts clear more vitality on this range from various fields.

The following portrayals are, wherever this speech recognition may be used:

a) Automated Phone Systems
b) Google Speech
c) Apple Siri (Apple Phone)

As shown in below structure figure 1. Speech recognition divided in two stages i.e. Speaker Identification and Speaker Recognition which further classified in various sorts, for instance,

a) Supervised Speech Recognition
b) Unsupervised Speech Recognition
c) Natural Language
d) Continuous Speech Recognition
e) Isolated Speech Recognition

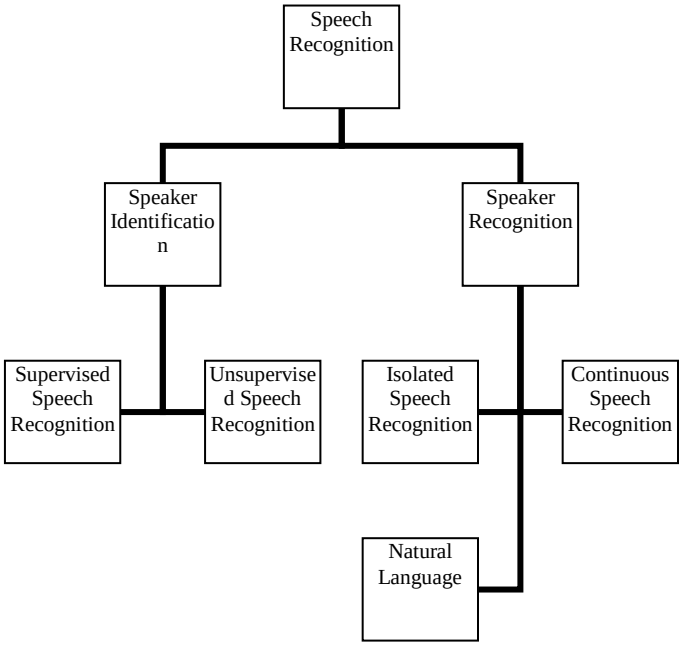

Figure 1. Types of Speech Recognition

In Supervised speech recognition the thing sees only the pre-processed speech of user; however, In Unsupervised speech recognition, the structure perceives and sees all voice samples. In Natural language, it sees the voice and answers the demand. In Continuous sort, the framework acknowledges the words unendingly at regular speed. In Isolated sort, the framework gets a handle on the voice just by observing single separate word [1].

The Supervised and Unsupervised speech recognition comes under Speaker Identification. Also other classifications, these are Natural language, Isolated and Continuous speech recognition are lying under Speech Recognition.

In Figure 2., the discourse can be perceived and seen by the framework through two or three strategies, for instance, Analysis, Feature Extraction, and reorganization. While receiving the speech as a data, it has to see first the responsibility by inquiring about it with appropriate interims. In the stir of isolating the information, it has to expel the words and a brief span later redesign it as showed by the contraption to remember its purposes of intrigue and obstructions in speech recognition technology [2].

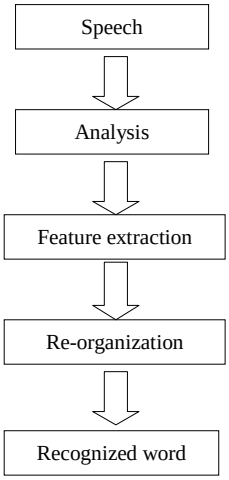

Figure 2. Speech recognition process

The different specialists explain some speech recognition improvement while utilizing the speech recognition technology. The speech recognition improvements and Cons:

a) Increases profitability.
b) Easy to utilize and quickly accessible.
c) More clear.
d) Now and again introduced in PCs and different telephones, considering fundamental access.

In like way, this progression contains several cons, for example,

a) Inability to get words in perspective of varieties of address.
b) Lack of help for most dialects other than English.
c) Inability to oversee foundation racket.
d) Sometimes it may induce messes up.

Considering each and every one of those properties, it has to make and recommend a structure related to speech recognition progression to contract with devices or sorts of contraption from wherever and at whatever point [2].

This paper describes a review on Automatic Speech Recognition and covers the outlining, frameworks, applications within it's introduction. Next section describes the different ASR techniques. In analysis section, a survey on major research works in the development of automatic speech recognition has described. Finally, this paper concludes with the assessment on feature trends for developing technique in ASR.

II. TECHNIQUES FOR ASR

A. Acoustic-Phonetic Approach

Speech recognition relies upon acoustics phonetics (AP) approach which hypothesizes that availability of the constant

and discrete phonetic unit in talk's dialects. Phonetic units are distinguished through a strategy of phonetic parameters which showed up inside the discourse signal or its fluctuation. The steps of this approach are;

a) Spectral analysis
b) Segmentation and Labeling
c) Determining valid word

In the first step spectral analysis of discourse joins with a characteristic identifier that improved spectral dimensions into a collection of characteristics which delineate the wider acoustic properties of the gathered phonetic units. Segmentation and labeling phase present that the discourse signal is segregated into constant acoustic zones, trailed through association too many phonetic units to each appropriated state, accomplishing in phoneme lattice classification of the speech.
In third step advance amid this algorithm finds to decide a suitable string of words from the phonetic label succession created by the segmentation to labeling [3].

B. *Artificial Intelligence Approach*

AI approach is crossbred of pattern recognition and acoustic phonetic method. This algorithm exploits the conceptions and thoughts of pattern recognition and acoustic phonetic (AP) approach. Two basic ways that to subsume superintend direct, arrange orchestration is everything seen as used as a piece of ASR – developed model overseeing in setting of Dynamic Time Wrapping (DTW) and Stochastic approach to assist in using hidden markov models. Every category to be seen is talked by perhaps few or couple setups in DTW [4]. Utilizing as slightly of a way over one reference imagine every category could also be better with a specific real objective to upgrade the verbalization/speaker drive showing within the interior of confirmation, a package between a watched speak progress and category mean is calculated to minimize the impact of the size bewilder among assessment and mention outlines, extended and reshaped understandings of suggested procedure also used out there estimation.

The acknowledged word at appears the course through the model that constrains the accumulated portion. Increase the amount of class sample alternate and loosening enclosed limitations might upgrade dynamic time wrapping based acknowledgment execution to inadequacy of the process inquisitive and storage space. At best, in classification systems, hidden markov model based pattern matching is rested rather than dynamic time wrapping in context of lower memory requirement and higher speculation properties.

C. Pattern Recognition Approach

Two main stages of pattern recognition approach are training and evaluation of a pattern. Pattern training phase involves pattern learning. The comparison of learning pattern and unidentified speeches is prepared in the evaluation phase to find out uniqueness of unknown to the probable pattern match. For the last six decades, the pattern recognition approach has changed into the incredible system for speech acknowledgement. This approach uses an engineered numerical framework and sets up reliable speech sample depictions, for comparison of the consistent pattern, from a strategy of prepared samples through proper preparation algorithm. A discourse plan depiction can be a discourse design or mathematical form (e.g., HMM) then it can connected to a noise more diminutive than utterance or a phrase [4].

D. Discriminative Learning – HMM-ANN

The perspective of different machine analysis unites utilizing a discriminative learning approach or evaluation thus making a generative model. In 1990, usage of neural frameworks as Multilayer Perceptron (MLP) and the different nonlinear kernel purposes of last layer were admired. The neural framework organized by back-improvement fault imitative appeared as a captivating acoustic demonstrating approach for discourse recognition. From the result of the MLP, it is taken as restrictive possibilities [5], and after yield is empowered into a hidden markov model, an incomprehensible discriminative development model, or crossbreed MLP-HMM can be made. Mostly to the difficulty in MLPs learning, this area of research is changed to another course where the multilayer perceptron basically makes a compartment of characteristic arrays and combined with the standard sets characteristics in hidden markov models [6]. As opposed to HMM, Neural frameworks develop no inquiries regarding characteristic statistical assets.

Neural system permitted discriminative learning in an ordinary and capable way when used to measure the percentages of a speech characteristic part, However, paying little identity to their feasibility in dealing with brief time units, for instance, solitary phones and secluded words, due to the lack of ability to model temporal dependencies of neural systems, it unsuccessful in sequential recognition task [7]. This type of superficial structure has been demonstrated all around sensible or uncomplicated issues, yet their constraint, showing up and genuine power can cause challenges while coordinating more tangled conformable applications including human talk. Hence, a neural network is used as an alternative algorithm for pre-processing such as measurement deductions for hidden markov model based recognition and characteristic alteration.

E. Generative Learning Approach - (HMM-GMM)

The generative learning approach relies on GMM dependant HMM which is used to address the successive arrangement of dialect signals. Speech is usually a static strategy in the concise time period. In speech recognition hidden markov model are used because discourse signal usually a piecewise inactive signal or a concise time inactive signal [8].

Discourse is usually taken off as an HMM for some random purposes. Generally, every HMM state uses a mix of GMM to show a spectral depiction of reverberation wave. Mixture of these are factored through ().is an array of random state percentages; P = (p_{ij}) is a state improvement probability network; Q = {(q_1,…, q_n} where q_j addresses the GMM of state j. The state is mostly considered associated with a part of a phone with a talk [9].

HMM is used to accomplish enormous levels of execution. Hidden Markov model is most commonly used because it hastily manages the changeable size data groupings which result from assortments in speaker rate, pronunciation and word arrangement. Speech recognition structures can be planned simple, consequently and computationally conceivable to utilize. Despite the way that the GMM-HMM approach had changed into the usual mechanical assembling in ASR, it has its own particular positive condition furthermore stacks. Though, the disadvantage of GMM is that it is to a great degree inconvenient for showing data that's based on an undirected complex in the information space.

F. Deep Learning -HMM-DNN

Deep learning is so often suggested as invalid feature learning or depiction learning, this is the latest zone in machine learning. This approach is changing into the standard improvement in speech recognition. It has adequately swapped feature coding and GMM for speech recognition at a inevitably higher range [10]. The standard sort consolidates generative deep structure, which proposed to depict the request relationship properties in information and its quantifiable dispersals in obvious information and related instances. This structure converts to discrete one by utilizing Bayes rule. Instances of this sort are deep Boltzmann machines, sum-product networks, different deep auto-encoders, and the total thing organizes it to the deep learning or deep belief networks [11] thus it is an improvement to the considered higher order Boltzmann machine at its base level.

Another sort of Deep learning is proposed to give discrete energy to pattern sorting and to do at that point of describing the back scatterings of instance names framed on the noticeable data. Cases include unit -MLP framework, identity-based ASR architecture, deep-framed CRF, deep curve or arranging framework and its structure assortment.

The main aim or third sort of crossbreed deep learning model is separation, but it is assisting through the results of generative structures. The Ultimate goal of the crossbreed structure is the generative module which typically battered assist with isolation [13].

III. ANALYSIS

The execution of speech recognition is generally shown with respect to speed and accuracy. Speed is assessed with the consistent factor, whereas accuracy is assessed as execution precision that is generally calculated as Word Error Rate (WER) [14]. Command Success Rate and Single Word Error Rate are other measures of accuracy. Word Error Rate and Word Recognition Rate (WRR) are used to measure exactness of speech recognizer. Word insertion (X), substitution (Y), deletion (Z) and number of words present in reference (R) are used to calculate word error rate and word recognition rate which is calculated by given equations.

$$\text{Word Error Rate (WER \%)} = \frac{X+Y+Z}{R} * 100 \quad (1)$$

$$\text{Word Recognition Rate} = 1 - \text{WER} = \frac{R-Y-Z-X}{R} \quad (2)$$

Comparison of various parameters in speech recognition used by different automatic speech recognition techniques is shown in the given table.

Table 1: Comparative Study of Various Methods of Speech Recognition

Sr. No.	Author and year	Name of Activity Referred	Method used for recognition	Techniques used	Parameters for recognition	Parameters for comparison	Parameters for performance	Overall output status
1	Karpagavalli S.and Chandra E. 2016 [1]	Correspondence amongst human and machines	Resonation, and attributes of the transducer	ANN	Noise and resonance	Resonance and attributes like frequency	Accuracy of recognition	61
2	Shweta Doda and Mehta R. 2014 [2]	Point by point survey of Speech Recognition	Perceive secluded words	Dynamic Time Warp and Hidden Markov Model	PCA and LDA	Point by Point	Accuracy of recognition	88
3	Ramakrishnan K.V, Swamy S. 2016 [3]	Advancements in Speech Recognition	Speed Recognition	VQ and HMM	MFCC	Precisely	Accuracy of recognition	98
4	Vimal Krishnan, V. R., Babu Anto 2009 [4]	Issue in Speech Recognition	Position Based Speech Recognition	MLP	Position	Time, position, speech	Accuracy of recognition	83
5	Ben Messaoud, Ben Hamida. 2010 [6]	Enhancing Traditional ASR	Classes of speech and its portrayal	SOM	Speech classes, speech portrayal, highlight extraction strategies, database, and execution assessment	Precision	Accuracy of recognition	63
6	Li, Z., Chen, R., Liu, L. and Min, G 2015 [7]	Pattern Similarity for Large-Scale Social Internet of Things	Dynamic Resources	MFCC, WPT	HMM	Classifiers	Accuracy of recognition	48
7	Toth, L. 2011 [8]	Low dormancy and precise ASR	Low memory	LSTM	CTC	SVD	Accuracy of recognition	78
8	Anusuya M. A and S. Katti 2011[9]	Short View cost of Speech Recognition	Speech Recognition based on Phases	RNN	Energy Features	Energy features based on testing fields	Accuracy of recognition	56
9	Ahad, A., Fayyaz, Mehmood. 2002 [10]	Overcoming their innate constraints and deficiencies	World View	MLP	Persistent speech features	Unsupervised clusters	Accuracy of recognition	89
10	Venkateswarlu, R.L.K, Kumari. 2011 [11]	Social Internet of Things	Inclination Closeness	3D Cartesian Coordinates	Cartesian Coordinate	Precision and recall	Accuracy of recognition	89
11	Pour, M. M, Fronkhi 2009 [12]	Reading the Marathi language	Interfacing Human and PC interface	RBF	Acoustic Features	Precision and Recall	Accuracy of recognition	77
12	Zhou, P, Tang. 2009 [13]	Robotization of Framework	Voice recognition module	SVM, DAGSVM	Binary Classifiers	Classifiers	Accuracy of recognition	94

IV. CONCLUSION

The learning layer of HMM-DNN Model is adaptive and maps input speech into text by scanning the context of speech, which provides better understanding of the user input. Hence, HMM-DNN based techniques outperform other techniques in terms of accuracy and speed of conversion. HMM and SVM are considered to be most dominant speech recognition techniques used in the field of speech recognition. HMM-DNN method can be extended to other languages as well as test the accuracy of the deep learning model for local languages to check its performance.

REFERENCES

[1] S. Karpagavalli., E. Chandra.," *A Review on Automatic Speech Recognition Architecture and Approaches*", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. **9**, No. **4**, pp. **394-398, 2016**.

[2] S. Doda., R. Mehta.,"*Speech Recognition Techniques: A Review*", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. **4**, No. **8**, pp. **944-947, 2014**.

[3] S. Swamy., K. V. Ramakrishnan.,"*An Efficient Speech Recognition System*", International Journal of Computer Science and Engineering, Vol. **3**, No. **4**, pp. **23-24, 2013**.

[4] Vimal Krishnan, V. R. Babu., P. Anto., "*Features of wavelet packet decomposition and discrete wavelet transform for Malayalam speech recognition*", Recent Trends Eng., Vol. **1**, No. **2**, pp. **93–96, 2009**.

[5] S.K. Gaikwad., B.W. Gawali., P. Yannawar.,"*A Review on Speech Recognition Technique*", International Journal of Computer Applications, Vol. **10**, No. **3**, pp. **17-21, 2010**.

[6] Ben Messaoud, Z. Ben., H. amida, "*CDHMM parameters selection for speaker-independent phone recognition in continuous speech system*", MELECON 2010 – IEEE Mediterranean Electrotechnical Conf., Valletta, Vol. **15**, pp. **253–258, 2010**.

[7] Z. Li., R. Chen., L. Liu., G. Min.,"*Dynamic resource Discovery Based on Preference and Movement Pattern Similarity for Large-Scale Social Internet of Things*", IEEE Internet of Things Journal, Vol. **3**, No. **4**, pp. **583-587, 2015**.

[8] L. Toth., "*A hierarchical, context-dependent learning network architecture for improved phone recognition*", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP, pp. **5040–5043, 2011**.

[9] M. A. Anusuya., S. K. Katti.,"*Speech Recognition by Machine: A Review*", International Journal of Computer Science and Information Security, Vol. **6**, No. **3**, pp. **184-187, 2009**.

[10] A. Ahad., A. Fayyaz., T. Mehmood., "*Speech recognition using multiside perceptron*", IEEE Proc. Students Conf., ISCON'02, pp. **103–109, 2002**.

[11] R. L. K. Venkateswarlu., R.V. Kumari., G.V. Jayasri., "*Speech recognition using radial basis function learning network*", Third Int. Conf. on Electronics Computer Technology ICECT, Kanyakumari, pp. **441–445, 2011**.

[12] M.M. Pour., F. Farokhi.,"*A new approach for Persian speech recognition*", IEEE Int. Advance Computing Conf., IACC, Patiala, pp. **153–158, 2009**.

[13] P. Zhou., L.Z. Tang., D.F. Xu., "*Speech recognition algorithm of parallel sub band HMM based on wavelet analysis and learning network*", Inf. Technol. J., pp. **796–800, 2009**.

[14] P. Saini., P. Kaur.,"*Automatic Speech Recognition: A Review*", International Journal of Engineering Trends and Technology, Vol. **4**, No. **2**, pp. **133-135, 2013**.

Authors Profile

Mrs. Gulbakshee J. Dharmale pursued B.E. Computer Science & Engineering from SGB Amravati University, Amravati in 2006 and M.tech. Computer Engineering from Dr. Babasaheb Ambedkar Technical University, Lonere in year 2011 and pursuing Ph.D. Computer Engg. From SGB Amravati University, Amravati. She is life member ISTE since 2011. She has published 3 research papers including in IEEE and its also available online. Her main research work focuses on Artificial Intellegence and Machine learing.

Dr. Dipti D. Patil pursued M.E. Computer Engineering from TSEC Mumbai University, Mumbai in 2007 and Ph.D. Computer Engineering from SGB Amravati University, Amravati in 2014. She is currently working as Associate Professor in MKSSS's Cummins College of Engineering for Women, Pune since 2014. She is member of BoS-Information Technology, SPPU, LMISTE, LMCSI. She has published more than 37 research papers in reputed journals including in Scopus and conferences including in IEEE and it's also available online. Her research work focuse on Machine Learning , Pattern Recognition, Classification, Neural Networks and Artificial Intelligence.

Dr. Vilas M. Thakare is a senior most professor of Computer Science among all the 11 Government Universities in Maharashtra. Continue as a Head, Post Graduate Department of Computer Science, Faculty of Engineering & Technology, SGB Amravati University, Amravati for more than 10 years. Guided more than 5 PhD scholars and more than 300 M.E./ M.S. /M.Phil./ M.C.A. Thesis. Guiding 8 PhD scholars. Published more than 350 research papers. Member, Expert Committees like AICTE (WR), CEDTI, YCMOU. Member Advisory Committee, IICC, Nagpur University, Chairman of BOS of Science as well as Engineering Faculty, BOS Member of SRTM University Nanded, Nagpur University, Nagpur and Dr.BAMU, Aurangabad. Member of Academic Council, NAAC, BUTR, ASU, DRC, RRC, SEC, CAS, NSD Recognized supervisor for Computer Science, Computer Engineering, Electronics Engineering. Delivered hundreds of keynote addresses and invited talks throughout India on variety of subjects related to computer science and engineering. Organized number of International and National conferences, workshops and seminars.

