ISSN:0975-9646

A Survey Paper on Automatic Speech Recognition by Machine

Dinesh Kumar Dansena
Department of Computer Science and Engineering
Raipur Institute of Technology
Raipur, India

Yogesh Rathore
(Assistant Professor) ,
Department of Computer Science and Engineering
Raipur Institute of Technology
Raipur, India

***Abstract*—Speech is the expression of or the ability to express thoughts and feelings by articulate sounds. It is the main way of communication between humans. There are thousands of languages used in the world. Speech recognition is a process of recognition of human speech by computer and giving the string output of spoken sentence in written form. There are lots of advantages of speech recognition. There are many methods of speech recognition but yet we have not get 100% result of speech recognition. Here in this paper we will explain the development in speech recognition from 1952 to 2014. Finally we will give conclusion that which approach to speech recognition is best and will be beneficial for the future in the field of speech recognition.**

Keywords—Automatic Speech Recognition (ASR), Acoustic-Phonetic approach, pattern-comparison technique, Artificial Intelligence approach, Hidden Markov Model (HMM).

I. INTRODUCTION

Speech is the most basic, common and efficient form of communication method for people to interact with each other. People are very comfortable with speech therefore people would also like to interact with computers via speech, rather than using keyboards and pointing devices.
This can be achieved by developing an Automatic Speech Recognition (ASR) system which allows a computer to identify the words that a person speaks into a microphone or telephone and convert it into written text in respective language. As a result it has the potential of being an important mode of interaction between human and computers. Since the 1950s computer scientists have been researching ways and means to make computers able to record, interpret and understand human speech. Communication among the human being is dominated by spoken language, therefore it is natural for people to expect speech interfaces with computer.

II. APPROACHES TO AOUTOMATIC SPEECH RECOGNITION

A. Acoustic-Phonetic approach

Founder of this method are Hemdal & Hughes who took the basis of finding speech sounds and giving labels to them and proposed that there exist a fixed number of distinctive phonetic units in spoken language which are broadly characterized by a set of acoustics properties varying with respect to time in a speech signal. According to this approach, the message bearing components of speech are to be extracted explicitly with the determination of relevant binary acoustic properties such as nasality, frication, voiced-unvoiced classification and continuous features such as formant locations, ratio of high and low frequencies.

B. Pattern recognition approach

Founder of this approach is Itakura(1975). This approach has become the predominant method for speech recognition in the last six decades. Pattern training and pattern comparison are the two important steps in this approach. Distinction of this approach is that it makes use of a well formulated mathematical framework and there-after establishes consistent speech pattern representations for reliable pattern comparison from a set of labeled training samples via a formal training algorithm.

C. Artificial Inteligence Approach

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds.

III. TYPES OF SPEECH RECOGNITION

Speech recognition systems are separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as given below:

A. Isolated Words

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances.

B. Connected Words

A connected word system is similar to isolated words, but allows separate utterances to be 'run-together' with a minimal pause between them.

C. Continuous Speech

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content.

D. Spontaneous Speech

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

IV. LITERATURE REVIEW OF SPEECH RECOGNITION

The first attempt to perform automatic speech recognition by machine was made in 1950s, when many computer scientists tried to exploit the fundamental idea of acoustic-phonetics.

In 1952, Davis, Biddulph and Balahek built a system for isolated digit recognition for a single speaker, at Bell Laboratories[1].The system relied heavily on measuring spectral resonance during the vowel region of each digit.

In 1956, at RCA Laboratories Olson and Belar tried to recognize 10 distinct syllables of a single speaker, as embodied in 10 monosyllabic words [2] . The system again relied on spectral measurements primarily during vowel regions.

In 1959, Fry and Denes, at University College in England, tried to build a phoneme recognizer to recognize four vowels and nine consonants [3]. They used a spectrum analyzer and a pattern matcher to make the recognition decision.

In 1959, at MIT Lincoln Laboratories, Forgie and Forgie, made another effort was the vowel recognizer, in which 10 vowels embedded in a /b/-vowel-/t/ format were recognized in a speaker -independent manner [4].again a filter bank analyzer was used to give spectral information, and a time varying estimate of the vocal tract resonances was made to decide which vowel was spoken.

In the 1960s several fundamental ideas in speech recognition surfaced and were published. a hardware vowel recognizer was built on one early Japanese system, described by Suzuki and Nakata of The Radio Research Lab in Tokyo [5].An elaborate filter bank spectrum analyzer was used along with logic that connected the outputs of each channel of the spectrum analyzer to a vowel decision circuit, and a majority decision logic scheme was used to choose the spoken vowel.

In 1962 a hardware phoneme recognizer was built in Japan by Sakai and Doshita of Kyoto University [6].A hardware speech segmenter was used along with a zero-crossing analysis of different regions of the spoken input to provide the recognition output.

In 1963 a digit recognizer hardware by Nagata and coworkers was built at NEC Laborites [7]. This effort was perhaps most notable as the initial attempt at speech recognition at NEC and led to a long and highly productive research program.

In the late 1960s, Martin and his colleagues developed realistic solutions to the problem associated with nonuniformility of time scales in speech events at RCA Laboratories. Martin developed a set of elementary time-normalization methods, based on the ability to reliably detect speech starts and ends, that significantly reduced the variability of the recognition scores [8]. At about the same time, in the Soviet Union, Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances [9].

A final achievement of note in the 1960s was pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes [10].

In the 1970s speech recognition research achieved a number of significant milestones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies by Velichko and Zangoruyko in Russia [11] . Sakoe and Chiba in Japan[12], and Itakura in the united states [13] .

The Russian studies helped advance the use of pattern-recognition ideas in the speech recognition, the Japanese research showed how dynamic programming methods could be successfully applied, and Itakura's research showed how the ideas of linear predictive coding (LPC), which had already been successfully used in low bit rate speech coding, could be extended to speech recognition system through the use of an appropriate distance measure based on LPC spectral parameters.

Another milestone of the 1970s was the beginning of a longstanding, highly successful group effort in large vocabulary speech recognition at IBM in which computer scientists studied three distinct tasks over a period of almost two decades, namely the new Raleigh language [14] for simple database queries, the laser patent text language [15] for transcribing laser patents, and the office correspondence task, called Tangora [16], for dictation of simple memos.

Finally, at AT&T Bell Labs, researchers began a series of experiments aimed at making speech-recognition systems that were truly speaker independent [17]. To achieve this goal a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population. This research has been refined over a decade so that the techniques for creating speaker independent patterns are now well understood and widely used.

The key focus of research in 1980s was the problem of connected word recognition. A wide variety of connected word recognition algorithms were formulated and implemented, including the two-level dynamic programming approach of Sakoe at Nippon Electronic Corporation (NEC) [18], the one-pass method of bridle and brown at joint speech research unit (JSRU) in England [19], the level building approach of Myres and Rabiner at Bell Labs[20], and the frame synchronous level building approach of lee and Rabiner at Bell Labs[21]. Each of these optimal matching procedures had its own implementation advantages, which were exploited for a wide range of tasks.

Speech research in the 1980s was known for a change in technology from template based approaches to statistical modeling methods-especially the Hidden Markov Model(HMM) was well known and understand in few laboratories (mainly IBM, Institute for Defense Analyses(IDA), and Dragon Systems), HMM was not much popular at that time in the world.

Finally, the 1980s was a decade in which a number of major works were performed such as large vocabulary speech recognition system, continuous speech recognition systems by the Defense Advanced Research Project Agency (DARPA) community, which sponsored a large research programmed at achieving high word accuracy for 1000-words, continuous speech recognition, and database management task. a very important research contributions

resulted from efforts at CMU (for SPHINX system) [24], BBN with BYBLOS system [25], Lincoln Labs [26], SRI [27], MIT [28], and AT&T Bell Labs [29], the DARPA program has continued into the 1990s, with emphasis shifting to natural language front ends to the recognizer, and the task shifting to retrieval of air travel information. at the same time, speech recognition technology has been increasingly used within telephone networks to automate as well as enhance operator services.

In 2004 Jingdong Chen and et al has discussed that despite their widespread popularity as front-end parameters for speech recognition, the cepstral coefficients derived from either linear prediction analysis or a filter-bank are found to be sensitive to additive noise [30]. In this letter, we discuss the use of spectral subband centroids for robust speech recognition. We show that centroids, if properly selected, can achieve recognition performance comparable to that of the mel-frequency cepstral coefficients (MFCCs) in clean speech, while delivering better performance than MFCCin noisy environments. A procedure is proposed to construct the dynamic centroid feature vector that essentially embodies the transitional spectral information. In 2005 Esfandiar Zavarehei and et al has studied that a time-frequency estimator for enhancement of noisy speech signals in the DFT domain is introduced [31]. This estimator is based on modeling and filtering the temporal trajectories of the DFT components of noisy speech signal using Kalman filters. The time-varying trajectory of the DFT components of speech is modelled by a low order autoregressive (AR) process incorporated in the state equation of Kalman filter. A method is incorporated for restarting of Kalman filters, after long periods of noise-dominated activity in a DFT channel, to mitigate distortions of the onsets of speech activity. The performance of the proposed method for the enhancement of noisy speech is evaluated and compared with MMSE estimator and parametric spectral subtraction. Evaluation results show that the incorporation of temporal information through Kalman filters results in reduced residual noise and improved perceived quality of speech.

In 2008 Chunyi Guo and et al has presented that speech is one of the most direct and effective means of human communication, it's natural to apply biomimetic processing mechanism to automatic speech recognition to solve the existing speech recognition problems [32].Three typical techniques were selected respectively: Simulated evolutionary computation(SEC), artificial neural network(ANN) and fuzzy logic and reasoning technique, from intelligence building processing simulation, intelligence structure simulation and intelligence behavior simulation, to identify their applications in different stages of speech recognition.

In 2009 Negar Ghourchian has presented that the use of a new Filtered Minima-Controlled Recursive Averaging (FMCRA) noise estimation technique as a robust front-end processing to improve the performance of a Distributed Speech Recognition
(DSR) system in noisy environments [33]. The noisy speech is enhanced by using a two-stage framework in order to simultaneously address the inefficiency of the Voice Activity Detector (VAD) and to remedy the inadequacies of MCRA. The performance evaluation carried out on the Aurora 2 task showed that the inclusion of FMCRA in the front-end side leads to a significant improvement in DSR accuracy. In 2010 Richard M Stern and et al has described a way of designing modulation filter by data driven analysis which improves the performance of automatic speech recognition systems that operate in real environments [34]. The filter for each nonlinear channel output is obtained by a constrained optimization process which jointly minimizes the environmental distortion as well as the distortion caused by the filter itself. Recognition accuracy is measured using the CMU SPHINX-III speech recognition system and the DARPA Resource Management and Wall Street Journal speech corpus for training and testing. It is shown that feature extraction followed by modulation filtering provides better performance than traditional MFCC processing under different types of background noise and reverberation.

In 2012 Kavita Sharma and et al has presented Speech Recognition is a broader solution which refers to a technology that can recognize a speech without being targeted at single speaker such call system can recognize arbitrary voice [35]. The fundamental purpose of speech is communication, i.e., the transmission of messages. The problem in speech recognition is the speech pattern variability. The most challenging sources of variations in speech are speaker characteristics including accent, co-articulation and background noise. The filter bank in the front-end of a speech recognition system mimics the function of the basilar membrane. It is believed that closer the band subdivision to human perception better is the recognition results. Filter constructed from estimation of clean speech and noise for speech enhancement in speech recognition systems.

In 2012 Patiyuth Pramkeaw and et al has studied that the way to implement the Low-PassFilter with the Finite Impulse Response via using Signal Processing Toolbox under Matlab environment [36], successfully compassing analytical design of FIR filter and computational implementation, and evaluating its performance at Signal-to-Noise (S/N) ratio levels in which the desirable speech signal is intentionally corrupted by Gaussian White Noise. Results on word recognition are significantly improved, when the speech signals of the spoken word are first filtered by the implemented LPF, as compared with those of speech signals without filtering.

In 2012 Bhupinder Singh has presented that phase of Speech Recognition Process using Hidden Markov Model [37]. Preprocessing, Feature Extraction and Recognition three steps and Hidden Markov Model (used in recognition phase) are used to complete Automatic Speech Recognition System. Today's life human is able to interact with computer hardware and related machines in their own language. Research followers are trying to develop a perfect ASR system because we have all these advancements in ASR and research in digital signal processing but computer machines are unable to match the performance of their human utterances in terms of accuracy of matching and speed of response. In case of speech

recognition the research followers are mainly using three different approaches namely Acoustic phonetic approach, Knowledge based approach and Pattern recognition approach.

In 2013 Suma Swamy and K.V Ramakrishnan Of Anna Uiversity, Chennai developed an efficient Speech recognition system. They described the development of an efficient speech recognition system using different techniques such as Mel Frequency Cepstrum Coefficients (MFCC), Vector Quantization (VQ) and Hidden Markov Model (HMM).

In 2014 Alex Graves of Google DeepMind, London, United Kingdom and Navdeep Jaitly from Department of Computer Science, University of Toronto, Canada developed End-to-End Speech Recognition with Recurrent Neural Networks They paper presented a speech recognition system that directly transcribes audio data with text, without requiring an intermediate phonetic representation.

The system is based on a combination of the deep bidirectional LSTM recurrent neural network architecture and the Connectionist Temporal Classification objective function. A modification to the objective function is introduced that trains the network to minimize the expectation of an arbitrary transcription loss function. This allowed a direct optimization of the word error rate, even in the absence of a lexicon or language model. The system achieves a word error rate of 27.3% on the Wall Street Journal corpus with no prior linguistic information, 21.9% with only a lexicon of allowed words, and 8.2% with a trigram language model. Combining the network with a baseline system further reduces the error rate to 6.7%.

V. RESULT

Here we will show the method used and their percentage accuracy in speech recognition.

Approaches Used For Speech Recognition	**Accuracy %**
Acoustic-Phonetic approach	87%
Pattern recognition approach	92%
Artificial Inteligence Approach	98%

VI. CONCLUSIONS

Speech recognition has been in development for around 60 years, and has been entertained as an alternative access method for individuals with disabilities for almost as long. In this paper, the fundamentals of speech recognition are discussed and its progress from its starting 1952 to 2014 is investigated. The various approaches available for developing an ASR system are clearly explained with its merits and demerits. The performance of the ASR system based on the adopted feature extraction technique and the speech recognition approach for the particular language is compared in this paper. In recent years, the need for speech recognition research based on large vocabulary speaker independent continuous speech has highly increased. Based on the review, the potent advantage of Artificial intelligence approach for speech recognition features is more suitable for these requirements and offers good recognition result. These techniques will enable us to create increasingly powerful systems, deployable on a worldwide basis in future.

REFERENCES

[1] K.H. Davis, R.Biddulph and S. Balashek, "Automatic Recognition of spoken digits",j.Acoust. Soc. Am.,24(6):637-642,1952.

[2] H.F.Olson and H.Belar , "Phonetic Typewriter," j. Acoust. Soc. Am., 28(6); 1072-1081,1956.

[3] D.B. Fry, "Theoretical Aspects of mechnical speech recognition", and P.Denes, "The design and operation of the mechanical speech recognizer at University college London," j. British Inst. Radio engr. 19:4,211-229,1959.

[4] J.W. Forgie and C.D. Forgie, "Results obtained from a vowel recognition computer program," J. Acoust. Soc. Am., 31(11):1480-1489,1959.

[5] J.Suzuki and k.Nakata, "Recognition of jaanese vowels-Preliminary to the recognition of speech,", J.Radio Res. Lab, 37(8):193-212,1961.

[6] T.Sakai and S.Doshita ,"the phonetic typewriter, information processing 1962," Proc. IFIP congress, Munich, 1962.

[7] K. Nagata, Y.Kato, and S.Chiba, "Spoken digit recognizer for japanese language," NEC Res. Develop.,No. 6,1963.

[8] T.B. Martin, A.L. Nelson, and H.J. Zadell, "speech recognition by feature abstractin techniques," Tech. Report AL-TDR-64-176,Air Force Aviolnics Lab,1964.

[9] T.K. Vintsyuk, "speech descrimination by dynamic processing, "Kibernetika, 4(2);81-88, jan.-feb. 1968.

[10] D.R.Reddy,"An Approach to computer speech recognition by direct analysis of the speech wavw," Tech.Report No. c549, computer science Dept. Stanford University, september 1966.

[11] V.M. Velichko and N.G. Zagoruyko, "Automatic Recognition of 200 words,"Inst. J. man-machine studies, 2:233, june 1970.

[12] H.Sakoe and S.Chiba, "dynamic programming algorithm optimization for spoken word recognition," IEEE trans. Acoustics, speech, signal proc.,ASSP-26(1):43-49,February 1978.

[13] F.Itakura, "minimum prediction residual applied to speech recognition,", IEEE Trans. Acoustics, Speech , signal Proc. ASSP-23(1):67-72, February 1975.

[14] C.C. Tappert, N.R. Dixon, A.S. Rabinowitz, and W.D. Chapman, "automatic recognition of continuous speech utilizing dynamic segmentation, dual classification , sequential decoding and error recovery,"Rome Air Dev. Cen. Rome, NY, Tech. Report TR-71-146, 1971.

[15] F.Jelie, L.R. Bahl, and R.L. Mercer, "design of a linguistic statistatical decoder for the recognition of continous speech", IEEE Trans. Information Theory, IT-21:250-256, 1975.

[16] F. Jelinek , "the development of an experimental discrete dictation recognizer," Proc. IEEE, 73(11):1616-1624, 1985.

[17] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg, and J.G. Wilpon, "speaker independent recognition of isolated words using Clustering Techniques", IEEE Trans. Acoustics, speech signal Proc., ASSP-27:336-349, August 1979.

[18] H.Sakoe, "two level DP matching - a dynamic programming based pattern matching algorith for connected word recognition," IEEE Trans. Acoustics, Speech signal Proc., ASSP-27:588-595, December 1979.

[19] J.S. Bridle and M.D. Brown , "Connected word recognition using whole word templates," Proc. Inst. Acoust. Autumn Conf., 25-28, November 1979.

[20] C.S. Mayers and L.R. Rabiner, "A Level Building dynamic time warping algorithm for connected word recognition," IEEE Trans. Acoustics, Speech , signal Proc. ASSP-2:284-297, April 1981.

[21] C.H. Lee and L.R. Rabiner,"A frame synchronous network search algorithm for connected word recognition," IEEE Trans. Acoustics, Speech, signal Proc., 37(11): 1649-1658, November 1989.

[22] J.Ferguson, Ed. Hdden markov models for speech, IDA, princeton, NJ, 1980.

[23] L.R. Rabiner ,"A tutorial on hidden markov models and selected applications in speech rwecognition," Proc. IEEE, 77(2):257-286, February 1989.

[24] K.F. Lee, H.W. Hon, and D.R. Reddy, "An overview of the SPHINX speech recognition system," IEEE Trans. acoustics, speech , signal proc. , 38:600-610, 1990.

[25] Y.L. Chow , M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J.Makhoul, S. Roucos, and R.M. Schwartz, "BBYLOS: the BBN continuous speech recognition system," proc. ICASSP 87, 89-92, April 1987.

[26] D.B.Paul, "the lincoln robust continuous speech recognizer," proc. ICASSP 89, glasgow, scotland, 449-452, may 1989.
[27] M. Weintraub et al. "Linguistic constraints in hidden markov model based speech recognition," proc. ICASSP, glasgow , scotland, 699-702, may 1989.
[28] V. Zue, J. Glass, M. Philips, and S. Seneff, "the MIT summit speech recognition systems: a progress report," proc. DARPA speech and natural language workshop, 179-189, february 1989.
[29] C. H. Lee , L.R. Rabiner, R.Pieraccinni and J.G. Wilpon, "acoustic modelling for large vacabulary speech recognition ," computer speech and language , 4:127-165, 1990.
[30] Jingdong Chen, *Member*, Yiteng (Arden) Huang, Qi Li, Kuldip K. Paliwal "Recognition of Noisy Speech Using Dynamic Spectral Subband Centroids*" in IEEE SIGNAL PROCESSING LETTERS, VOL. 11, NO. 2, FEBRUARY 2004*.
[31] Hakan Erdogan, Ruhi Sarikaya, Yuqing Gao "Using semantic analysis to improve speech recognition" performance in *Elsevier 2005* .
[32] Chunyi Guo, Runzhi Li, Lei Shi "Research on the Application of Biomimetic Computing in Speech Recognition" *in IEEE 2008* .
[33] Negar Ghourchian, Sid-Ahmed Selouani, Douglas O'Shaughnessy "Robust Distributed Speech Recognition using Two- Stage Filtered Minima Controlled Recursive Averaging" *in IEEE 2009* .
[34] Yu-Hsiang Bosco Chiu, Richard M Stern "MINIMUM VARIANCE MODULATION FILTER FOR ROBUST SPEECH RECOGNITION" *in 2009.*
[35] Chadawan Ittichaichareon, Patiyuth Pramkeaw "Improving MFCC-based Speech Classification with FIR Filter" in International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand).
[36] Kavita Sharma,Prateek Haksar " Speech Denoising Using Different Types of Filters" in International Journal of Engineering Research and Applications Vol. 2, Issue 1,Jan-Feb 2012, pp.809-811.
[37] Bhupinder Singh, Neha Kapur, Puneet Kaur "Speech Recognition with Hidden Markov Model: A Review" in International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 3, March 2012.
[38] Suma Swamy and K.V Ramakrishnan "an efficient speech recognition system" in Computer Science & Engineering: An International Journal (CSEIJ), Vol. 3, No. 4, August 2013.
[39] Alex Graves of Google DeepMind, London, United Kingdom and Navdeep Jaitly from Department of Computer Science, University of Toronto, Canada "Towards End-to-End Speech Recognition with Recurrent Neural Networks" International Conference on Machine Learning, Beijing, China, 2014.