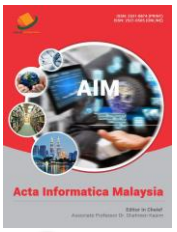
ZIBELINE INTERNATIONAL PUBLISHING

Acta Informatica Malaysia (AIM)

DOI : http://doi.org/10.26480/aim.01.2019.01.03

ISSN: 2521-0874 (Print)
ISSN: 2521-0505 (Online)
CODEN: AIMCCO



CrossMark

REVIEW ARTICLE

SPEECH RECOGNITION SYSTEMS - A COMPREHENSIVE STUDY OF CONCEPTS AND MECHANISM

Neha Jain[1], Somya Rastogi[2]

[1]Vidya College of Engineering, Meerut, India
[2]Vidya College of Engineering, Meerut, India
Corresponding Author Email: njnehajain330@gmail.com, rastogi.somya8@gmail.com

ARTICLE DETAILS

Article History:

Received 15 November 2018
Accepted 17 December 2018
Available online 2 January 2019

ABSTRACT

Speech Recognition Systems now-a-days use many interdisciplinary technologies ranging from Pattern Recognition, Signal Processing, Natural Language Processing implementing to unified statistical framework. Such systems find a wide area of applications in areas like signal processing problems and many more. The objective of this paper is to present the concepts about Speech Recognition Systems starting from the evolution to the advancements that have now been adapted to the Speech Recognition Systems to make them more robust and accurate. This paper has the detailed study of the mechanism, the challenges and the tools to overcome those challenges with a concluding note that would ensure that with the advancements of the technologies, this world is surely going to experience revolutionary changes in the near future.

KEYWORDS

Interactive Electronic Agent, Dynamic Type Warping, Hidden Markov Model, Voice Activity Detector, Feature Vector

1. INTRODUCTION

Human beings come in this world with their innate abilities to learn different things. Suppose the human beings have to interact with each other by writing messages to them, it would be a real pain. And that is how the human beings interact with computers, what if they could just talk with them to make things done, it would be so much easier if the computers can understand what human beings are saying and for that the human beings need a really good Speech Recognition Systems.

Speech Recognition is a technology with the help of which a machine can acknowledge the spoken words and phrases, which can further be used to generate text. Speech Recognition System works using techniques popularly termed as acoustic modeling and language modeling. Acoustic modeling represents statistical relationship between the linguistic segments of audio signals and phonemes, on the other hand language modeling represents probability distribution of word segments in a given word sequence [1].

Performance of the machines capable of implementing Speech Recognition technology is evaluated using two parameters:

- Accuracy (percentage error in converting spoken words and phrases into text)
- Speed (the extent to which the machine can keep up the pace with a human speaker)

Speech Recognition system is something that has been dreamt about and worked for decades. A variety of software products are available that allow users to dictate their systems and get words and phrases converted to text in a word processing document. Human beings are now able to access function commands, such as opening files, accessing menus, etc. with their voice instructions.

Speech Recognition Systems have now helped many disabled people, who at times are not able to write and have now adopted speech recognition systems. So now people can just speak certain words to get what they need and the system that makes this possible is a kind of speech recognition script implementing speech recognition technology.

2. EVOLUTION OF SPEECH RECOGNITION SYSTEMS

In 1784, a scholar created the first Acoustic Mechanical Speech Machine in Vienna. After that in 1879 Thomas Edison invented the first dictation machine. Continuing with the chain, the next speech recognition system was developed at Bell Laboratories in 1952, capable of recognizing spoken digits with 90% accuracy, but it could only recognize numbers spoken by its inventor. In 1970 a scholar came out with a Harpy System, which was able to recognize over 1000 words and could recognize different pronunciations and some phrases [2,3]. Speech Recognition continued in 80s with the introduction of Hidden Markov Model, which used a more mathematical approach of analyzing sound waves and led to many of the breakthroughs.

Soon with the invention of Hidden Markov Model in 80s IBM Tangora, in 1986 used the Hidden Markov Model which was able to predict the upcoming phonemes in speech. In 2006, The NSA (National Security Agency) started using speech recognition systems to segment keywords in recorded speech.

Then there arrived a boom of speech recognition technology when the world's top IT companies comprising of Facebook, Google, Amazon, Microsoft and Apple started offering this functionality in various devices through services like Google Home, Amazon Echo, Apple Siri and many more [4]. The goal of these top tech companies is to make voice assistants response and reply with more accuracy.

3. MECHANISM OF SPEECH RECOGNITION SYSTEMS IN DISTRIBUTED REALTIME

When human beings speak, it leads to creation of vibrations in air. To convert speech to an on-screen text, a computer has to pass through several complex steps. In the process of speech conversion the query from the client extracts a sufficient number of vectors in the form of acoustic speech vectors representing utterances via communication channel [5]. The system then filters the digitized sound to remove unwanted signals and noise and sometimes to separate it into different bands of frequency. The filtered signals are then normalized and adjusted to a constant volume level so that they can be easily divided into small segments as short as hundredths of a second, or even thousandths in case of occlusive consonant sounds. The sounds are then sent to the server, and using Hidden Markov Model, grammars, dictionaries the signals are decoded into text using Natural Language Processor and Database Processor, in this the optimized SQL (Structured Query Language) statements perform full text search in the database. Further processing is performed on the searched statements to get a single stored question to which answers are taken via file path and sent to the client in the compressed form. Reaching the client side, the answer to the user's query is now conveyed to the user using a text to speech engine in his or her native language. The whole architecture of the above describes process through HMM model is given in figure 1.

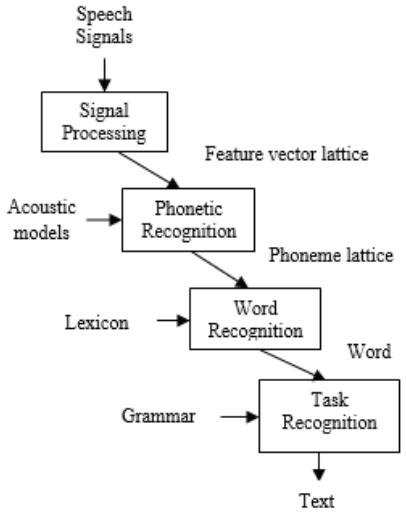


Figure 1: Speech Recognition in mechanism using HMM model

4. ADVANCEMENTS ADAPTED IN SPEECH RECOGNITION SYSTEMS

Speech Recognition Systems today are good at transcribing what human beings say; they can talk about travels, they can talk about their contacts, and the list is endless. Technology has made tremendous improvements very quickly. Years, ago one needs to be an engineer to interact with computers but today everybody can interact. One thing, that is still at infancy is the understanding, for that the current generation needs a far more sophisticated language understanding models that understands what the sentence means. Advancements have been adapted to catch up the fast-moving pace by using IEA (Interactive Electronic Agent) that provides a prompt to the user during the said interactive speech based sessions accompanied with suggestions and queries [6].

Moreover, Interactive Electronic Agent provides a confirmation confirming the substance of the said NLP (Natural Language Processing). Continuing with the confirmation, these agents provide the said response to the user from the Natural Language Engine and Database Processor.

At present, the human brain and the learning algorithms at human brain are far better at things like language understanding. To support this, developers and researchers have now moved to neural networks that work better than this existing technology basically just a table look up. The brain has several billion of neurons working in parallel and all of the knowledge in the brain is the result of the connection among neurons in the brain. In the same way, a neural network is developed to work the same way as our brain is working [7]. With the advent of neural network technology, they recognize the feature on their own, they can learn the features, then they can learn features of features and then they can learn features of features of features. And that has led to huge improvements in speech recognition systems.

5. COMPARISON BETWEEN DYNAMIC TYPE WARPING ALGORITHM AND HIDDEN MARKOV MODEL

Since the era of evolution of speech recognition systems many different approaches have been adapted to implement speech recognition technology in machines and systems, out of which DTW (Dynamic Type Warping) Algorithm and HMM (Hidden Markov Model) have been the center of attraction for all time. The comparison between DTW and HMM given in table 1 is based on [8].

Table 1: Comparison of DTW and HMM

DTW	HMM
Dynamic Type Warping is an algorithm used for pattern matching between the two signals that may vary in speed and time.	Hidden Markov Model is a model which is used to study hidden or unobserved states.
DTW has been applied to temporal sequence of audio, video and graphic data.	HMM provide a framework for modeling using mathematical computations.
DTW is widely used in areas of speaker recognition and online signature recognition and partially used in shape matching application.	HMM is widely used in areas of speech conversion, gestures and tagging of POS (Part of Speech).
DTW algorithm has proved helpful in coping up with different speeds in automatic speech recognition systems.	HMM provides an estimate (via probability score) of the given sequence matching with the string of phonemes.

6. CHALLENGES WITH SPEECH RECOGNITION SYSTEMS

Speech is an essential mode of communication with computers as well as human beings. Speech Recognition has a wide range of applicability in the domain of computer science, medical science, etc. Developing a real time speech recognizer may get effected from adverse environment to anatomy of human body, involving human aspects too. Some of the key challenges faced with Speech Recognition Systems are discussed below [9].

6.1 Noisy Environment

Studies have proven that the drawback affecting most of the Speech Recognition Systems is the environmental noise and its adverse effect on systems performance. It is a challenge for these systems to extract feature during conversion of speech to the on-screen text.

6.2 Intensive Use of Computer Power

Running the statistical models needed for speech recognition requires the computer's processor to perform a lot of heavy work. One of the reasons for this is the necessity to remember each stage of the word recognition search, in case the system needs to backtrack to come up with the right word.

6.3 Accent

The speaking accent differs according to the social and personal situations (e.g., physiological and cultural aspects). Indeed, compared to native speech recognition, performance degrades when recognizing accented speech and non-native speech. Studies have shown that the human vary in

accent while speaking to parents and when speaking to friends.

6.4 Speed of Speech

The speech recognition systems find difficulty separating segments of continuous speedy speech signals. The speed of speech while speaking may vary and depend upon situations and physical stress. The pace of speaking may affect in pronunciation through phoneme reduction, time expansions and compressions.

6.5 Recognition of Punctuation Marks

It has been observed that while conversion of speech to on screen text the punctuation marks are not recognized as they are, instead proper words are recognized. Many different strategies are made to overcome this challenge involving the speed dictating these punctuation marks, and so on. But the better solution is yet to come.

6.6 Homophones

Homophones are the words that have different meanings but sounds same when pronounced (e.g., "There" and "Their", "Be" and "Bee"). In Speech Recognition Systems, it is very difficult at the word level to recognize which one is the correct intended word. The current observations show that, the Speech Recognition Systems on an average achieve 94 to 99 percent accuracy due to different accents.

7. TOOLS TO OVERCOME CHALLENGES FACED WITH SPEECH RECOGNITION SYSTEMS

A number of noise reduction techniques have been engineered to extenuate the effect of noise on systems performance and often require the estimate of noise statistics. One technique that has been engineered is to design a database that may be used for the evaluation of feature extraction at the front end using a defined a Hidden Markov Model employed at backend.

7.1 Voice Activity Detector

Voice Activity Detector is a useful technique for enhancing the performance of Speech Recognition Systems employed in noisy environmental conditions [10]. Voice Activity Detector is used in Speed Recognition Systems for feature extraction process thereby resulting in enhancement of speech by the systems. The dependency factor of Voice Activity Detector lies on pitch detection, energy threshold, periodicity measure, and spectrum analysis. One major challenge that the detector face is while making a decision about extraction of Feature Vector (FV); selection of feature vector for signal detection and strong decision rule is a challenging problem, affecting the performance rate of Speech Recognition Systems.

7.2 AURORA Experimental Framework

The AURORA framework was designed as a contribution to the ETSI STQ-AURORA DSR Working Group. AURORA is developing standards for Distributed Speech Recognition (DSR) where speech analysis is done in telecommunication terminal and the recognition is performed at the central location in the telecom network [11]. In AURORA framework the idea was to design a database that may either be used for feature extraction along with backend defined Hidden Markov Model in Speech Recognition Systems. The TIDigits Database is used as the basis to develop the original database in AURORA framework.

8. CONCLUSION

The objective of this paper is to review various aspects related to Speech Recognition Technology and the systems implementing this technology to engineer Speech Recognition Systems. The paper first describes what actually lead to the emergence of developing Speech Recognition Systems, continuing with how the mechanism of conversion of speech takes place in distributed real-time systems. This paper reviews the advances that have taken place after the development of traditional Speech Recognition Systems, also this paper briefly makes a quick comparison between the algorithms and the models that were and now are being used to implement these Speech Recognition Systems. Many tools and framework have been developed to overcome the challenges with Speech Recognition Systems like Voice Activity Detector, AURORA framework, etc. Technological advances in computation has led the technology of Speech Recognition reach the state where situations are far better as they were used to be years back and definitely in the coming few years the world is about to experience much better language understanding by the machines.

ACKNOWLEDGEMENT

This research paper has been equipped as a result of discussion among the faculty members of my college. The authors like to thank Dr. Rajendra Kumar (Head of Department CSE/IT) of Vidya College of Engineering, Meerut, India for their constant guidance and support during the research work time period.

REFERENCES

[1] Rouse, M. Speech Recognition, Available: https://searchcrm.techtarget.com/definition/speech-recognition.

[2] Boyd, C. The Past, Present and future of Speech Recognition technology, Available: https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf.

[3] A short history of Speech Recognition, Available: https://sonix.ai/history-of-speech-recognition.

[4] van der Valde, N. Speech Recognition technology overview, Available: https://www.globalme.net/blog/the-present-future-of-speech-recognition.

[5] Bennett, I.M., Babu, B.R., Morkhandikar, K., Gururaj, P. 2015. Distributed Real-time Speech Recognition, Naunce Communication Inc., Patent No. US 9,076,448.

[6] Bennett, I.M., Babu, B.R., Morkhandikar, K., Gururaj, P. 2014. Speech Recognition System Interactive Agent, Naunce Communication Inc., Patent No. US 8,762,152 B2, 24 June 2014

[7] Deng, L., Hinton, G., Kingsbury, B. 2013. New type of Deep Neural Network learning for speech recognition and related applications: an overview, IEEE International Conference on Acoustics, Speech and Signal Processing (13886524), ISBN-(978-1-4799-0356-6), 26-31 May 2013.

[8] Chadha, N., Gangwar, R.C., Bedi, R. 2015. Current Challenges and Applications of Speech Recognition Process using Natural Language Processing: A Survey, International Journal of Computer Applications (0975-8887), 131(11), 28-31.

[9] Petkar, H. 2016. A review of Challenges in Automatic Speech Recognition", International Journal of Computer Applications (0975-8887), 151(3), 23-29.

[10] Ramírez, J., Górriz, J.M., Segura, J.C. 2007. Voice Activity Detection, Fundamentals and Speech Recognition Systems Robustness, University of Granada, Spain.

[11] Hirsch, H.G., Pearce, D. 2000. The Aurora Experimental framework for the performance evaluation of Speech Recognition System under noisy conditions, ASR-2000 Automatic Speech Recognition: Challenges for the new Millennium Paris, France, 181-188.