RECOGNIZING SPEECH SYSTEM: AN IN-DEPTH EXAMINATION OF IDEAS ALONG WITH WORKINGS

Ritu Tailor, Surbhi Sharma
E-Mail Id: ritu.tailor@gits.ac.in, surbhi.sharma@gits.ac.in
Department of Computer Applications, Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

Abstract- Nowadays, speech recognition systems make use of a wide range of multidisciplinary technologies, including unifying statistical frameworks, signal processing, natural language processing, and pattern recognition. Applications for these systems are numerous and include signal processing issues among many other things. The aim of this study is to introduce the fundamentals of speech recognition systems, including their development and recent improvements that have been applied to improve their accuracy and robustness. This paper provides a thorough analysis of the mechanism, obstacles, and strategies for overcoming them. It ends with a prediction that, as technology advances, the world will undoubtedly see dramatic changes in the not-too-distant future.
Keywords: Vocal Action Detector, Hidden Markov Model, Interactive Digital The agent, and VariableFont Cracking.

1. INTRODUCTION

Humans are naturally inclined to learn a variety of topics when they enter our planet. It would be rather unpleasant if people had to communicate with one another by writing messages to one another. And that is howpeople now communicate with computers. Imagine if people could just talk to computers to get things done. That would be a lot simpler if computers could comprehend what people are saying, which would need highly developed speech recognition systems on the part of people.

A machine can recognize spoken words and phrases thanks to a technique called speech recognition, which may also be used to create text. The methods used by systems that recognize speech to operate are known as language modeling and acoustic modeling. Language modeling, on the other hand, shows the probability distribution of word segments in a given word sequence. Acoustic modeling represents the statistical link between the linguistic parts of sound waves and sounds [1].

Two metrics are used to assess the speech recognition technology-capable computers' performance:

- Precision (the fraction of errors in textual representation of spoken words and phrases)
- Speed (the degree to which the robot can match a human speaker's cadence)

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 totext within a word processing file. Function tasks, such accessing menus and opening files, may now be accessedby voice commands for humans. Speaking certain words to obtain what they need is now possible thanks to Speech Recognition Systems, which have helped a great deal of disabled people who, at times, are unable to write. A type of speech recognition script that uses speech recognition technology powers this system.

2. EVOLUTION OF SPEECH RECOGNITION SYSTEMS

A professor in Vienna created the first Acoustic Mechanical Speech Machine in 1784. After that, in 1879, Thomas Edison invented the first dictation machine. The next speech recognition system was created by Bell Laboratories in 1952, and it was limited to recognizing numbers that the system's designer said. The recognitionaccuracy of spoken digits by this method was 90%. In 1970, a researcher created the Harpy System, which couldrecognize over a thousand words, different pronunciations, and specific phrases [2, 3]. The creation of the Hidden Markov Model in the 1980s, which used a more mathematical approach to analyze sound waves, was responsible for numerous advancements in the recognition of speech.

In 1986, IBM Tangora used the Hidden Markov Model, which had been developed in the 1980s, to predict the phonemes that will be pronounced next. First employed in 2006 to find keywords in recorded conversation, voice recognition software was developed by the National Security Agency (NSA). Subsequently, voice recognition technology had a surge in popularity as the leading IT firms globally, including Facebook, Google,Amazon, Microsoft, and Apple, began incorporating this feature into a range of devices via services such as Google Home, Amazon Echo, Apple Siri, and many more [4]. These leading tech businesses want to improve the accuracy and responsiveness of voice assistants.

3. MECHANISM OF SPEECH RECOGNITION SYSTEMS IN DISTRIBUTED REALTIME

The sound of human speech causes vibrations in the atmosphere. A computer must go through a number of difficult procedures in order to translate voice to text that appears on screen. A sufficient number of vectors in the

form of acoustic speech vectors representing utterances via communication channel are extracted from the client's inquiry during the speech conversion procedure [5]. Subsequently, the digital audio is filtered by the system to eliminate undesirable signals and noise, and occasionally, it is divided into several frequency bands. In order to make the filtered signals easily segment able into brief bursts of as little as hundredths of a second, or even thousandths in the case of occlusive consonant sounds, they are next normalized and set to a constant volume level. The noises are then transmitted to the server, where the signals are translated into text using the Hidden Markov Model, grammatical structures, dictionary entries, and Natural Language Processor. The optimized SQL (Structured Query Language) statements enable complete text search in the database. The searched statements are subjected to further processing, which results in a single stored question. Responses tothis question are obtained via a filename and supplied in compressed form to the client. When the inquiry reaches the client side, the user is now provided with the answer in their language using a text-to-speech engine. Figure1 shows the whole architecture of the process described above using the HMM model.

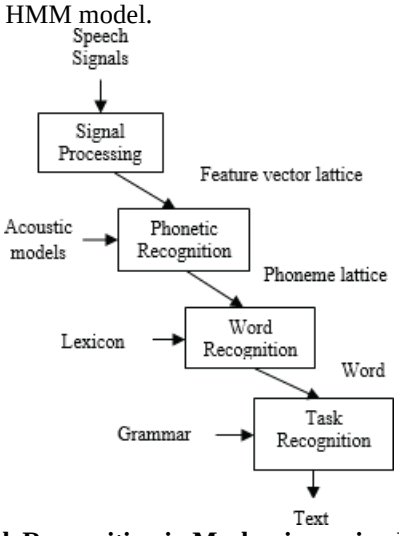

Fig. 3.1 Speech Recognition in Mechanism using HMM Model

4. ADVANCEMENTS ADAPTED IN SPEECH RECOGNITION SYSTEMS

These days, speech recognition systems are rather adept at transcription; they can record conversations concerning contacts, travels, and a plethora of other topics. Technology has advanced rapidly and to great effect. It used to be necessary to be an engineer to communicate with computers, but now days everyone can do it. One area that needs improvement is comprehension; the present generation need significantly more advanced models of language understanding to comprehend the meaning of a phrase. In order to keep up withthe rapid rate of advancement, an Interactive Electronic Agent, or IEA, has been employed. During interactive speech-based sessions, the IEA prompts the user with questions and ideas [6].

Additionally, Interactive Electronic Agent offers a validation verifying the accuracy of the aforementioned NLP. Continuing with the confirmation, these agents give the user the aforementioned answer via the Database Processor and Natural Language Engine.

The way the human mind's learning processes currently work with the human. are significantly more proficient in areas like language comprehension. In order to facilitate this, researchers and engineers have now switchedto neural networks, which perform better than the current technology, which is essentially limited to table lookups. Theconnections between the billions of neurons that make up the brain's network are what allow the brain to store and process information. Neural networks are designed to function in a manner similar to that of the human brain [7]. Since the development of neural network technology, they have been able to identify features on their own and learn features, features of features, and features of features of features. And as a result, voice recognition systems have greatly improved.

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

Since the beginning of the development of voice recognition systems, several methods have been developed to incorporate speech recognition technology into devices and systems. Among them, the Dynamic Type Warping (DTW) Algorithm and the Hidden Markov Model (HMM) have long been the main attractions. Based on [8], the comparison between DTW and HMM is presented in Table 5.1.

Table-5.1 Comparison of DTW and HMM

DTW	HMM
An approach called dynamic type warping is used to match patterns between two signals that could have different timings and speeds.	Covert Markov A modelis a tool used to investigate states that arehidden or unseen.
The temporal sequence of audio, video, and graphic	A structure for predicting with mathematical

data has been subjected to DTW.	computations is providedby HMM.
DTW is somewhat utilized in shape matching applicationsand is extensively employed inthe fields of online signatureand speech recognition.	HMM is extensively utilized in the fields of gestures, speech transformation, and POS (Chapter of Sound) labeling.
The DTW method has been shown to be useful in accommodating varying voice machine learning rates.	Using a likelihood score,HMM offers an estimateof how well the provided sequence matches the sound in the strings.

6. CHALLENGES WITH SPEECH RECOGNITION SYSTEMS

Speech is a vital form of communication for both humans and computers. There are several applications for speech recognition in the fields of computer science, medicine, etc. The development of a real-time speech recognizer may be impacted by unfavorable environmental conditions as well as the human anatomy. The nextsection discusses a few of the major difficulties with speech recognition systems [9].

6.1 Noisy Environment

Research has demonstrated that ambient noise has a negative impact on the performance of most speech recognition systems, making it a downside. These systems have difficulty extracting features while voice is being converted to text on the screen.

6.2 Intensive Use of Computer Power

The computer's CPU must work extremely hard to run the statistical models required for voice recognition. Thisis due, in part, to the fact that each step of the word recognition search must be remembered in case the system has to go back and find the correct term.

6.3 Accent

Speaking accents vary depending on social context and individual circumstances (e.g., physiological and cultural elements). In fact, accented and non-native speech recognition ability deteriorates when compared to native speech recognition. Research has demonstrated that people's accents change depending on whether they are chatting to friends or their parents.

6.4 Speed of Speech

Voice recognition systems have trouble distinguishing apart parts of continuous fast-talking voice signals. Speaking at a different pace depends on the context, as well as on physical strain. Pronunciation can be impacted by speech rate through phoneme reduction, temporal compression, and expansion.

6.5 Recognition of Punctuation Marks

It has been noted that when voice is converted to on-screen text, appropriate words are recognized instead of punctuation marks as they are. The difficulty of dictating these punctuation marks quickly is met by a variety of approaches, and so forth. However, the best answer is still to come

7. HOMOPHONES

Homophobic phrases, such as "There" "There," "Be" and "Bee," sound the same when spoken yet possess distinct meanings. It is exceedingly hard to determine at the word level in systems that recognize speech which word is the proper intended word. Due to varying accents, the present observations indicate that speech recognition programs often reach 94 to 99 percent accuracy.

8. TOOLS TO OVERCOME CHALLENGES FACED WITH SPEECH RECOGNITION SYSTEMS

Many methods of reducing noise have been developed to lessen the impact of noise on the operation of systems, and they frequently necessitate the estimation of noise statistics. Using a specified Hidden Markov Model utilized at the backend, a database architecture has been developed as one method for evaluating feature extraction at the front end.

8.1 Voice Activity Detector

When used in loud environments, Voice Activity Detector is a helpful method for improving Speech Recognition System performance [10]. Speech is improved by the Speed Recognition Systems with the usage of Voice Activity Detector in the feature extraction process. Pitch detection, energy threshold, periodicity measure, and spectrum analysis are what the Voice Activity Detector depends on. Selecting the right feature vector for signal detection and a robust decision rule is a difficult problem that impacts the performance rate of speech recognitionsystems. This is one of the main challenges faced by the detector when deciding how to extract the feature vector(FV).

8.2 AURORA Experimental Framework

The AURORA framework was designed as a contribution to the ETSI STQ- Group for Aurora DSR Working. AURORA is creating standards for Distributed voice Recognition (DSR), in which voice recognition is carried

out at a telecom network center while speech analysis is completed at a telecommunication terminal [11]. The goal of the AURORA framework was to provide a database that could be utilized in speech recognition systems for feature extraction in conjunction with a backend-defined hidden Markov model. The original database of the AURORA architecture was developed using the TIDigits Database as a foundation.

CONCLUSION

This paper's goal is to explore a variety of topics pertaining to speech recognition technology and the systems that use it to build speech recognition systems. The paper initially explains the genuine factors that contributed to the development of speech recognition systems before going on to discuss how voice conversion works in distributed real-time systems. In addition to reviewing the developments that have occurred since the creation ofconventional speech recognition systems, this study provides a brief analysis of the differences between the models and algorithms that were and are currently utilized in the implementation of speech recognition systems to address these issues, many frameworks and tools, such as the AURORA framework and Voice Activity Detector, have been created. The field of speech recognition technology has advanced to a point where things arenow considerably better than they were a few years ago. In the next few years, it is certain that robots will be ableto grasp language significantly better than they could a few years ago.

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 forthe new Millennium Paris, France, 181-188.