Voice Speech and Recognition—An Overview

Ayush Anand and Raju Shanmugam

Abstract In the present scenario, when the automation has gain so much strength, all the support and machinery to human conversation are getting linked with voice controls. In all the field be it Internet of things, artificial intelligence or communication. In order to reduce the efforts on giving instructions and implementing them, the world has changed its behavior toward voice recognition. The voice speech recognition allows the user to understand the spoken voice commands and generate the result on the basis of that. This deals with the voice received and generating outputs based on the hearing capability of the machine. The result is generated and is made available for the further. Voice-based devices or applications are growing a lot. It uses state art process in speech-to-text, natural language understanding, deep learning and text-to-speech. The first step to build a voice application is to listen for user voice constantly and then understand the voice to understand and implement things. This deals with understanding in various languages under the section and reflects the work assigned to it. Speech recognition which is also known as automatic speech recognition (ASR) and voice recognition recognizes the spoken words and phrases and converts them to a machine-readable text, and speech recognition technology let users control digital devices by speaking instead of using conventional tools such as keystrokes, buttons or keyboards. From automated phone systems to Google voice to digital assistant, i.e., voice recognition, ASR helps in daily life activities. Even the bluetooth used in cars uses ASR. Voice recognition has taken a complete scan all over the arena. The tool of automation in voice further deals with the communication among the machines without the code access and suggestively talking with the help of compatible recognizable words. This would strengthen the compatibility and speed up the automation process by decreasing the amount of efforts made. Such work would overcome the error and would possibly give a better way of transformation.

A. Anand (✉) · R. Shanmugam (✉)
Unitedworld School of Computational Intelligence, Karnavati University, Gandhinagar, Gujarat, India
e-mail: ayushananad636@gmail.com

R. Shanmugam
e-mail: srajuhere@gmail.com

A. Abraham et al. (eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks*, Lecture Notes in Networks and Systems 167,
https://doi.org/10.1007/978-981-15-9712-1_29

As Edmund L. Andrews researcher from Stanford recently on March 23, 2020 said that the possibility of error could be minimized to a great extent with the evolution under this process.

Keywords Speech to text · Text to speech · ASR · HMM · Neural network

Abbreviations

ASR	Automatic Speech Recognition
HMM	Hidden Markov Model
NN	Neural Network
WER	Word Error Rate

1 Introduction

In voice speech recognition, the users are made to make the voices in the customized language which has to be understood by the system, and interpretation is made what was said. This enables the microphone and stresses on the recognition factor of the different voices generated by different people by picking up the word from the speech made. The search can be performed on any device with input as voice. Language is the most important factor for a system to understand and give the most accurate results of what the user search as output. This covers across languages, dialects and accents, as users want a voice assistant that understands both of them and speaks to them with a better understanding. The search method is similar to performing normal search on the Web site; the only difference is the search that is conducted using speech, rather than text.

This covers the ability of a machine to decode the voices made by human as an input signal to the system. The main purpose includes operating a device, performing the commands assigned by the most common is text-to-speech format wherein the system understands whatever the human is saying, and after understanding by differencing it on the language basis and various pitch and difference in voice, it recognizes and writes the suitable text which is further ready for other various activities. The result can be searched each time by calling out the program and is limited to ten words. Further, chunk size and sampling rate are adjusted accordingly in order to get the complete out of the system.

It is expected to meet the demands by hearing out the same sound with different frequency and pitch with different voice, and it has been made understandable where the system understands the voice easily and shows quick implementation and the result of it from the voice made for the system (Fig. 1).

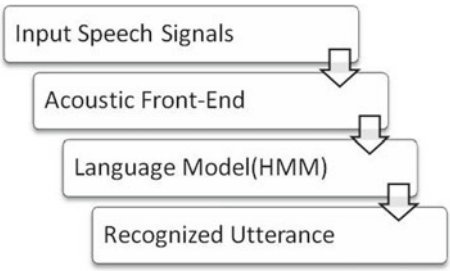

Fig. 1 Hierarchy automation model

Among the technology gaining the market, in 2018, Amazon alone sold around ten million of echo devices marking as one of the biggest selling devices. Further, the paper suggests the development in the automation field where the system could communicate among each other and deal with the solution to the problem in real quick time. This became very problematic and a difficult scenario to deal with, yet with the inducement in technology, the program was called off. Wherein it was seen that the system earlier started communicating in their own encrypted format and the project had to shut down, this being the challenge where the encryption should be held with the developer so that they can auto tune the faults. Various fields such as call centers and IVR systems use speech recognition tool for many organizations and for self-service as well. The load of the customer and client is cleared off with the help of this. Further, it has a lot of practical deployment in the field of dictation solutions. The speech-to-text technology has given the opportunity to express the ideal without typing anything. Even the field of education is entirely covered up with the recognition technology, where the students are creating and documenting the word files without actually typing. This has made the work and load for the students much easy and has reduced a great amount of load from them. For the disabled, the applications serve as a new hand where they are able to learn things more quickly and adapt the learning habitat as fast as any other people. This paper deals with the growth of automatic voice recognition in the field of automation where the system communicates within them creating a sentimental analysis for them and leading to the easier work.

2 Literature Survey

Voice speech recognition roots can be traced on from an early time and could be majorly understood by understanding how we as human are able to recognize the statements made. These happen under a set of parameters of speech classification which includes the isolated words which are the single words that are interpreted one at a time, and then, according to that, it is interpreted. Following are the connected and continuous words. Under connected words, it allows the users to speak naturally, and the computers are made to examine the words, but it allows the separate words

to run together with creating a minimum pause in between. Another classification of speech is continuous, similar to connect words, it allows under to speak naturally by creating boundaries and various difficulties aroused, while speech is implemented. Lastly, spontaneous speech has the ability to take control of the various voice tools and connected words and making them work altogether.

As the first speech recognition was mainly focused on numerology rather than building words. In 1952, the system was developed which recognized a dingle voice speaking digits that too in loud frequency. From that, it took 10 more years for IBM (Shoebox) to understand English language and responded to 16 words. Much later came the concept of hidden Markov model (HMM), and it took as a storm by the 80s. Further with due course of time, it was made to consumer availability in the later 90s where ASR was introduced known as automatic speech recognition.

The best extract for speech recognition is ASR. However, other there were other recognition systems which were speaker-dependent system, under which it requires training before any word or sentence is made to system. Speaker-independent system is the software recognizes voice with minimal training. In discrete speech recognition, the user has to take a pause before speaking the other word in order to let the system understand. In continuous speech recognition, the system understands in the normal speaking flow, and lastly, in natural language, not only understanding is there but an automated response is also generated based on that.

With the involvement of recognition among the automation for making the workload easier, it was to be stated the condition of work and the related work in this field. In the year 2009, Anusuya [1] stated the field of acoustic–phonetic approach to the recognition along with artificial intelligence and discriminative learning—HMM-ANN. In the year 2012, Singh [2] stated the explanation of hidden Markov model involvement in the recognition field to reduce the load. It showed the development of a speech model which was created on the platform of machine interface. Sahu [3] in May 2013 further added to the field the concept of connected words where the different lengths of the word took different time for the system to apt the voice. Lleida [4] stated that the utterance of the similar would be taken as single word to reduce the load on the machine and the paper dealt with the making the machine more optimized to get only the required data and ignore the rest of the body. In October 2017, Cambria [5] under his article on the review of research paper developed a theory which stated that about reduction of 30% in the work would be done if the sentiment analysis of the systems is done. Burt [6] in October 2019 stated the quality configuration in speech where the receiver was given the priority as the interference from the source might cause problem. Shobha Dey [7] in April 2014 did the work in the intelligent recognition of the words where the predictive tests were converted into the closest similar word so that better and accurate results could be achieved. Sharma and Hakar [8] in 2012 kept the main focus on speech denoising using different types of filters which could help in the clarity of the tone and volume of the content. Acharjee et al. [9] presented a paper on voice recognition system: Speech-to-text, where the continuous format of speaking style was directly converted into speech with intellectually calculating the pause between the words and at the same time prediction analysis was performed on the words to get the accurate sentence. Paliwal [10] in the year

2003 developed the use if the spectral co-bands of centroids where it displayed the dynamic approach to SSC coefficients revealing that they are stronger enough to deal with the noise than the MFCC specifications.

Over the year's speech recognition has gained a physical importance and in the technological field, the advancements are seen much more in information recovery where the information is recovered using the concept of speech recognition tool. Artificial intelligence comes into the picture which is a very important field in relation to the work done in speech-to-formatting or text-to-speech, the ultimate goal is to translate and judge the automation just like a human would do, and the automation here just increases the capabilities with it. Further, with the artificial intelligence as base, it is done on the basis of bottom up, top down and the blackboard ways. In the case of bottom up, the cases with the less priority are induced first, and then, the cases with the higher priority are instantiated. The reverse of this takes place in the top down approach. This sums up the involvement of artificial intelligence in speech recognition tool. However, the blackboard approach takes the inclusion of various approaches among the acoustic, semantic, surrounding and the easy methodology process to act.

3 Problem Statement

In the early voice recognition tool, it had capabilities of hearing out only numerical values and further extends to hearing a sentence that too with a limited training to the system and mainly focused in one language. Then came the advancement where multiple language support was possible and automated response could be generated.

Whenever we say a word or call out a sentence that is understood by the system, it interprets in its own way, by generating the various results, as different users have different voices. Filtering out all the disturbances and noise cancelation of the voice, the sentence is interpreted as a natural language, and based on that, result is generated, with the system having a functioning microphone and for an automated response a speaker. The search is made on the various search engines by making the speech.

Further, the utterances are the area to be dealt of, and it got grouped into short and long utterances where more than 10 words say up to 100 words can be taken and least specific which means ignoring the basic pronunciation errors in favor of fast customer experience with taking less interruptions in the code. While under short, it is focused on very specific variation of speech. Various training data of different voices are assigned to overcome this and in order to provide an accurate result.

4 Proposed System

The basic working of the voice speech recognition is the intake of speech signals which goes under preprocessing, and language modeling is performed on it with the

help of various models such as uniform model and stochastic Model. Speech signals from sources are captured with the help of microphone, where generally sampling of frequency ranges between 8 and 20 kHz. In the need of calling signals of voice, it is advised for the signals to have a sampling rate of 8 kHz, and the general microphones are rated to be at the frequency of 16 kHz. Modeling of language is done in order to spot the right word sequence by predicting the nth words using preceding word technique.

Our proposed system is automatic sound recognition where it uses the concept of hidden Markov model, in which an alternative approach speech recognition is to construct a model of each word to recognize the vocabulary and make the system understand working of it. This is a modeling approach which consists of three major parameters namely the model, the method of computing the probability of the model and the methods of computing the parameters of the model. Our system here deals with listening to the user and understanding it to the different levels of connected, isolated and continuous word. On taking these values as input, the result is generated which shows the result in the form search made. The model thus suggests the capabilities of using the automation search recognition techniques in order to communicate between the systems to remove the chances of error and help to grow it faster and makes it more reliable and data is accurate. The workload is also reduced with the inducement of such act. The recognition saves time, and it helps in understanding the speech in different voice and sentiments to judge them as a question or an imperative remark. These studies are done by making the system learn by providing them a different set if training data. The length of words further creates an another set of test cases for the system to understand. All these tasks are performed on a unit testing basis to increase the readability and decrement in chances of error by the system.

The tool of speech recognition in machinery helps in the automation among the machines wherein the system will help in communicating with each other and reduce the human effort of hard coding. The system understands the concept of sentiment analysis and further contributes to each other with the medium of communication, thus helping resolving the queries system is developed in such a way that it understands the sentiment of other devices working along and takes the input signals from them and regulates on the basis of that making it much easier, cost effective and reliable to the system. The idea has been taken from the tech system developed by Facebook, where the system designed to work for the designated work started to talk among themselves and further the project had to call off as the securities issues lacked because in contrast to the work done the outcome showed that they were talking in an encrypted format making it harder for the system to understand the working procedure of the system and thus it turned out to be serious threat as the security concern was not taken. The cause of such an event could be drawn to the sentiment analysis where the understanding is developed and it further takes care of the methodology of the system. Speech recognition is divided into four subcategories on the basis of its properties, and they are speech mode, speaker mode, speaking style, word size and spoken styles. All these deals with the way the voice is enabled in order the system to understand the given data. Under speech mode isolated where the words are separated by a good pause and other mode is continuous where the speaking is

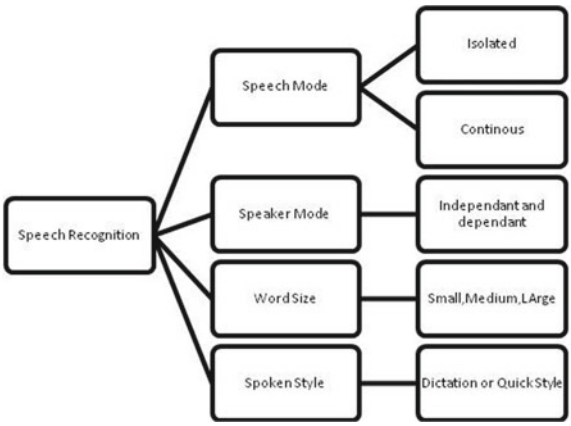

Fig. 2 Speech recognition classification

done in the flow for the system to understand the data. Speaker modes are further categorized as independent and dependent parts. The type of words that are given to the system to analyze may deal with problem with the long, short and medium words. Here, the complexity increases with the increment in the length of the words. The spoken style is however dependent of the type of style which is dictation, the continuous format or the quick style (Fig. 2).

5 Methodology

The proposed method includes the involvement of few classifiers and those are

1. Hidden Markov Model (HMM)
2. Neural Network (NN)

1. **Hidden Markov Model**: The hidden Markov model is a model basically a statistical model where the system is modeled and the modeled system is assumed to be a process called the Markov process where the parameters seem to be unknown as the wholesome tasks lie in generating the hidden parameters from the given data which is also known as observable data. It is the heart of the voice recognition facility and provides a natural framework for connecting such models to it. Further, it helps in finding the unknown variables from the list of known variables.
2. **Neural Networks**: The neural network is a concept of analyzing the data based on certain flow and list of algorithms that explains the flow very similar to as that of a human brain has the capabilities of interacting to the system. The network has an adaptability behavior where it plays as the smartest role in which it can change its behavior according to the condition and mark itself as that to help the situation and paradigm. A neural network is trained by adjusting the inputs taken and based on the network performance. The network classifies the data correctly, which helps increases the probability of correct answers while the other similar

data probability is decreased. This helps in canceling out the other close words that are very similar to the spoken ones and allows the system to make the best of the choices, and it chooses accordingly. The signals sent make a connection sets which compile in accordance with the units and with help of the training set which results in the output (Fig. 3).

Under the proposed path, it requires a general set of skill and proficiency for the user to make the right choice of pronunciation and clear voice as the voice received by the system is filtered under various categories of modeling concepts in the underlying design of the model. The system receives the voice interpretations and on subcategory levels classifies it to the statement received and then is interpreted which is then answered by an automated response on the basic understanding of the system. However, it currently deals with the text-to-speech interpretation, and the link is forwarded to the search engine where the search iOS made and the result is displayed. Pattern recognition deals with acquiring the details from the machine and dealing with the understanding and producing a result dataset. The input speech signal is sent where the speech signals are learned and understood and is followed upon to pattern comparison which takes help from the pattern of unknown signals for the accuracy, and then, the output is generated. The performance of the voice recognition system is solely based on the swiftness and how précised the data is going to be. The accuracy measured in terms of the performance accuracy is usually rated with word error rate (WER), wherein the swiftness measured in the single time factor. Other two factors are important to the contribution and that includes the word error and command success. However, the performance basis of judgment was mainly dependent on the error occurrence of the words connected to it, and the system translation deals with the accomplishment of the recognition in voice. The difficulty was in understanding the difficult words coming through the way as different words have different length. The problem was resolved by first aligning recognized word sequence with the speaker and taking help of the dynamic alignment of the string that is supposed to be the word.

Word error rate can formulate into

$$\text{WER} = (X + Y + Z)/C$$

Where

X Count of substitutions
Y Count of deletions

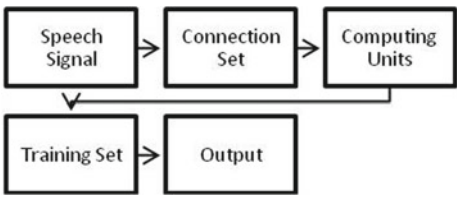

Fig. 3 Neural network approach

Z Count of insertions
C Count of words in the reference.

6 Conclusion and Future Research Directions

Automatic speech recognition has become a strong tool for the system communications to be made automated. The limited availability of the resources could not stop the growth anyway. But with the right techniques and extensive use, it can be made more productive for the customer experience as well as the automation levels in the production machineries as well. The future of the speech recognition for various uses in the various fields will get wider and wider and is going to be one of most vast arenas ahead.

The training data and recognition can be speed up by making it more prone to the environment conditions. Seeking voice of different personalities is tough. However, a major work is being done on that by teaching the system through many dataset available. The platform is going to support the demands of customer as well as automation needs too. The current usage of speech in consumer fields has increased for fold. According to the survey analysis of get elastic in 2019, it was found that around 20% of queries are submitted through voice on Google. 72% of people use voice recognition to manipulate and use devices. The global market has boomed to 187%. Considering the same if this automation is reached in the field of automation design, then the market would grow faster, and the workload can be reduced at a great level.

However, the work has been done on this level as Ford and Toyota have already been introduced with Alexa over few years and Hyundai has teamed up with Google to make it as a virtual assistant. Around 64% of the customers have shown their interest and investment to the technology. Predictions suggest that by the end of the decade around half of the search will be done through voice search. In USA, the rise would be from 13 to 55% in upcoming two years. The field of automation among the machines to communicate themselves has started, and with the proper encrypts voice enabled partnership, they could work as fine as ever.

Not only in the field of automation but the speech recognition has been a boon to impaired patients with difficulty in making out the sentence or words to the people. With the help of tools, these are made possible some of which includes HTK, Julius and Sphinx which are open source toolkits.

References

1. Anusuya MA (2009) Speech recognition by machine. IJCSFS 6(3)
2. Singh B, Kapur N, Kaur P (2012, March) Speech recognition with Hidden Markov Model
3. Prabhakar O, Sahu NK (2013) A survey on voice command recognition technique. IJAR 3(5)

4. Lleida E et al (2000) Utterance verification in decoding and training procedures. IEEE Trans. Speech Audio Process 8(2)
5. Cambria E (2017) A practical guide to sentimental analysis
6. Burt C (2019, October) Ensuring quality in voice biometrics data collection
7. Dey S (2014, April) Intelligent system speech recognition
8. Sharma K, Hakar P (2012) Speech denoising using different types of filters. Int J Eng Res Appl 2(1)
9. Acharjee K, Das P, Prasad V (2015, November) Voice recognition system: speech-to-text
10. Paliwal KK, Gajic B (2003) Conference: acoustics, speech, and signal processing. In: Proceedings. (ICASSP '03). 2003 IEEE international conference
11. Spratt EL (2002) Computers and art in the age of machine learning
12. Yu D, Deng L (2015, March) Automatic speech recognition—a deep learning approach
13. Wankar P (2014) Research paper on basic of neural network
14. Damodaran S (2015) Guidance for hearing impaired people 4(2)
15. Sharma R (2004) Design and implement multi-language converter using machine learning techniques. IJCSMC J
16. Bajorek JP (2000) Guiding principles of voice user research 1(1)
17. Doyle S (2006) Automatically improving a voice recognition system
18. Science in Me (2020, April) Voice and speech recognition market analysis by 2020
19. Magrath D (2020, March) Listening and speaking: teaching hints
20. Singh K (2012) A review of speech literature