International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-6, November 2021
https://doi.org/10.55524/ijircst.2021.9.6.27
Article ID IRPS12038, Pages 118-122
www.ijircst.org

The Comprehensive Analysis Speech Recognition System

Dr. Anubhav Soni[1], and Dr. Jitendra[2]

[1, 2] SOMC, Sanskriti University, Mathura, Uttar Pradesh, India

Correspondence should be addressed to Dr. Anubhav Soni; anubhavs.somc@sanskriti.edu.in

Copyright © 2021 Dr. Anubhav Soni et al. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT- Automated computer speech detection has been a goal of research for almost four decades. The idea underlying the development of a voice recognition system is that it is beneficial for people to communicate with a computer, robot or another device via speech. Researchers want to utilize people's voices in an integrated environment to provide customers comfort and flexibility in today's era. Speech recognition is dependent on the input voice signal. It enables the computer to transform the voice signal to text or instructions using the technique of recognition and understanding. This article describes the fundamental concept of speaking recognition, types of voice recognition, difficulties with speech recognition, several ways to gather characteristics, and different pattern-matching algorithms to identify the different speakers. The technology of voice recognition is inspired by a wide vocabulary automated speech recognition system which allows language-independent operation and continuous speaking in a foreign language.

KEYWORDS- ANN, Feature Extraction, Markov Model, Speech Recognition, Voice Recognition.

I. INTRODUCTION

Human beings have long been driven to construct a computer that can understand and talk like human people. Since the 1960s, data scientists studied various techniques and means to create computer databases, read and decode human speech. The key component of voice recognition is the converting of sound into text and commands. Spoken recognition is the process by which computers map a kind of speech to an auditory signal. The sound has to be compared with preserved sound bits, which need more study and frequently do not match the pre-existing sound bits. This technique is very tough. This method is very difficult. Different techniques for the extraction of features and patterns are utilized to make voice recognition systems of better quality. Feature extraction methods and pattern matching technology play an essential part in the voice recognition system to optimize performance[1].

Speech is the most frequent method of human communication. There are many languages spoken by people throughout the globe for communication. Researchers attempt to build the system to analyze the voice signal and to categorize it. In many fields, including agriculture, health care and government, the computer system, which can comprehend the spoken language may be extremely helpful. Speech recognition refers to the capacity to hear spoken words and recognizes the different sounds contained in them. Speech signals are almost stationary. When speaking signals are examined over a short period of time (5-100 m/sec), they are stationary, but change the characteristics of the signal for a longer period; they reflect the difference in speaking sounds. Features of the voice signals are retrieved on a short-term amplitude basis (phonemes). The most crucial step of voice recognition is feature extraction. During the extraction procedure, there are various difficulties due to the diversity of the speakers[2].

Speech recognition is the machine or program's capacity to recognize words and sentences from spoken language and transform them into machine-readable formats. It is sometimes called Automatic Speech Recognition or Computer Speech Recognition. The primary purpose of language recognition is to develop techniques and systems for machine speech input. Speech is the principal method of communication among people and the dominance of this medium drives research attempts to enable speech to become a feasible computer interface between human beings. Automatic speech recognition (ASR) is thus seen as an intrinsic component of human-computer interfaces, envisioned for using speech to get natural, prevalent and omnipresent computing, among other things. Two types for isolated and ongoing language recognition typically exist: speech-dependent and speaker-independent. The Speaker Dependent Method involves training a system that recognizes each lexical word spoken by a specific number of speakers individually or several times, while the self-employed methods of training are usually impracticable and the words are recognized by their acoustic properties. With computer hardware and software rapidly evolving, voice recognition technology is relatively important to computer information technology. It is extensively utilized in voiced phone exchange, medical services, financial services, industrial management of all aspects of society and the lives of individuals[3]. Figure 1 illustrates basic voice

recognition system in a single equation including function extraction, database, network training, testing or decoding.

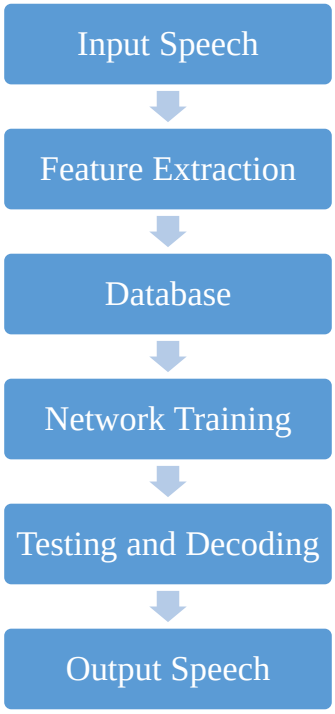

Figure 1: Elementary Prototype of Speech Recognition System

A. System of Speech Recognition Based on Events

1) Isolated Words

A solitary word recognition device understands a single word. For situations where only one-word responses or instructions are required from the user, isolated word recognition is adequate, but not suitable for numerous inputs. It is smoother and clearer for implementation, since word boundaries are obvious and the words look articulated, which is the main advantage of this kind. The disadvantage of this approach is the choosing of separate limits that affect the results.

2) Words Connected

A linked world structure is similar to independent sentences, but allows various expressions to occur together with a minimal latency between them. The word is the vocalization of the device of a single word or phrases.

3) Continuous Address

A continuous voice recognition system encourages people to virtually automatically speak, but the content is determined on by the computer. This is a computer dictation. In this manner, words flow together without pause or any other separation between words. The ongoing technique of voice recognition is difficult to construct.

4) Spontaneous Discourse

The spontaneous speech recognition system recognizes natural expression. Spontaneous speech suddenly coming from the mouth is natural. A random ASR speaker may handle many natural speaking features, such as words running together. Spontaneous expression may include mispronunciation, incorrect beginnings, and non-words.

B. Speech Recognition Based On the Model of the Speaker

Every speaker has a unique voice because of his particular physical physique and personality. The voice recognition technique is divided into two main categories:

1) Dependent Speaker Models

Speaker-based systems are designed for a particular speaker type. They are usually more detailed for the same speaker, although they may be less accurate for certain speaker. These systems are usually cheaper, easier to create and more accurate, but they are not as flexible as speaker-independent systems.

2) Independent Models Speaker

The Speaker Independent gadget can detect many speakers without prior expertise. A speaker-independent device is developed to work with any particular type of amplifier. It is utilized in the IVRS, which needs a large number of distinct consumers to provide input. But the disadvantage is the reduction in the number of terms in the vocabulary. The most difficult is the introduction of the Autonomous Speaker Method. It is also expensive and its performance is worse than speaker-dependent devices. Different methods of extraction used in speech recognition:

- Analysis of the main component
- Linear Analysis of Discrimination
- Analysis of independent components
- Predictive linear coding
- Wavelet

C. Model-Based Approach

In the template-based method there are a number of preset speech patterns. These patterns are treated as comparison patterns that describe the word dictionary. By comparing an unknown spoken word with one of these reference models and selecting the best fitting kind of pattern, speech is recognized. For all words, templates are typically created. Errors related to tighter acoustic segmentation or grouping may be prevented by more variable units like as phonemes. The template-based approach to voice recognition has produced a family of methods that have significantly advanced the field over the last two decades. This is a simple process. It is the way to match new language to a set of pre-recorded words or models. The usage of exact word models offers the benefits of this approach, but that has the disadvantage of fixing the already registered templates. Therefore, speech disparities, which certainly become inefficient, can only be changed by employing more models each word. Modeling and matching with vocabulary size above a few hundred words are prohibiting expensive or inefficient. This method is typically wasteful in terms of the required storage and processing resources needed to

conduct the matching. In addition, continuous speech detection is not possible with this technique[4].

D. Approach Based On Knowledge

Several researchers proposed the use for voice recognition using a knowledge-based approach and used it for speech recognition. The method to knowing uses linguistic, phonetic and spectrometric data. Specific experts' expertise of speech variation is coded into a system. It uses a number of features from the voice and then teaches the computer to generate a set of sample output rules automatically. These rules are the result of factors that give important information about categorization. This technique offers the advantages of modelling speech variation precisely, but unfortunately, such expertise is difficult to gather and utilize effectively, so that this approach is considered to be selected rather than nonpractice and automated learning methods.

E. Neural Network Approach

The usage of neural networks is another method in the voice recognition system for matching patterns. Neural networks can handle more complicated recognition problems; however, they cannot accomplish as well when it comes to wide vocabulary as the Markov Hidden Model (HMM). They can handle low output, noisy data and speaker flexibility. This kind of system may be more precise in case of training outcomes than HMM-based systems and the vocabulary is decreased. Phoneme detection is a method more known with neural networks. This is a varied area of study, but normally its results are better than HMM. There is also a hybrid NN-HMM approach that combines the neural network and the HMM as part of language modelling for phoneme identification. Artificial neural network technology is utilized in voice recognition since it reduces the modelling device and energy[5].

F. Approach Based on Dynamic Time Warping (DTW)

Dynamic Duration Warping is a method which may vary in speed or time in order to determine correlations between two sequences. It is used in ASR to handle changing vocalization rates. Generally, it is a method that enables a programme to optimally meet certain restrictions for two sequences, which means that the sequences are "warped" so as to match each other non-linearly. In general, DTW is a technique for finding an optimum match between two sequences with certain restrictions. In principle, this method is helpful for the identification of independent words and may also be updated to detect linked terms.

G. Approach Based on Statistics

Variations in expression are statistically modelled using training methods in this methodology. Present techniques for the universal recognition of speech are based on acoustic and linguistic mathematical models. A large amount of acoustic and linguistic data are needed for parameter estimation, acoustic and language models for ASR in an unlimited domain. The processing of enormous amounts of training data is a key element in the development of effective ASR systems. The main drawback of predictive models is the assumption that the prior modelling may not be accurate and may restrict the effectiveness of the approach[6].

H. Hidden Markov Model (HMM)

The Secret Markov Model technique is useful for the recognition of speech. Since HMM may be learned automatically and utilized with computer viability. HMMs are simple networks that may utilize many statements for each model to create a speech and model for each state's short-term continuum. The parameters of the model are the probabilities of change in state, the methods, variances and mixture weights that define the distributions of state production. An HMM for a number of words or phonemes is created by connecting an HMM trained individual to the distinct words and phonemes. Each word or phoneme would have a different distribution of output. Modern HMM-based, broad-based speech recognition systems are typically practiced over hundreds of hours using acoustic input. The word set, pronunciation dictionary and HMM training technique will automatically determine the word. This indicates that it is very simple to use large training data. It is the core value of HMM to substantially decrease the time and complexity of the recognition process for wide vocabulary learning[7].

II. DISCUSSION

Speech is one of the old methods of expressing oneself. These voice signals are now widely utilized in biometric recognition and machine communication technologies. These signals are carefully timed with different signals (quasi-stationary). The properties are relatively stationary when evaluated over a suitable short length of time (5-100 m/sec). But if the signal qualities vary over a period of time, it represents the various speech sounds. The information in the voice signal is conveyed by the short-term amplitude of the spoken wave shape. This enables us to extract characteristics from the voice based on the short term amplitude (phonemes). The main problem to recognize the speech is that, owing to various speakers, nt speech speeds, content and sound circumstances, the speech signal is extremely changeable.

Speech has evolved as one of human communication's most amazing methods. Speech is considered to be the most natural and basic type of communication in parallel communications, such as writing, body language and gesture. Since we are satisfied with the speech, we naturally want the interface to the computers via voice medium, without the need of ancestor interfaces such as keyboards or pointing devices. This is made feasible simply by the use of an Automatic Speech Recognition (ASR) method that is used to modulate a voice signal with the help of an algorithm conducted by a computer programme. It is beautifully equipped with the abilities needed to emerge as an important interface between people and machines. In recent decades the automated voice recognition (ASR) has

greatly increased the pinnacle of winning in a wide range of genuine applications, including simple digit identification, gigantic vocabulary transcript news, reading a style voice dictation, impulsive conversation, etc. This significant breakthrough has been made using a variety of prominent ASR-recognized statistical modelling methods, such as speech signals and spoken language documents collected from practical applications. The basic objective of the speech detection domain is to provide new language input techniques and processes to the computer. The ASR research by machines has invited a soaring excitement for a gigantic period of 60 years. The ASR is currently widely used for tasks requiring the human machine interface such as automated call processing and also the computer that can talk and recognize speech in a local language. In line with the exciting progress in statistical voice modelling, the ASR technology offers many applications for the activities that need a human-machine interaction today. These applications include automatic telephone call processing and data systems that provide the latest travel data, stock price quotes, weather reports, data entry, speech dictation, data access, travel, banking, commands, the automotive portal, speech transcript, supermarkets for persons with disabilities and reservations for railways etc. Spoken word recognition technology was used widely to mechanize and modernize operator services in the telephone networks. In addition, speech understanding methods are today able to understand the voice input in the functional situations with vocabulary of thousands of words[8].

The initial stage in the process of language recognition is the parameterization of an analogue voice stream. Several prominent methods for signal analysis have become de facto norms in literature. These algorithms are designed to create a parametric "perceptually meaningful" representation of the voice signal: parameters that imitate certain behaviors of the human auditory and perceptual systems. These algorithms are also, of course, and perhaps more crucially, intended to optimize recognition performance. The origins of several of these methods may be traced to research on speaker-dependent technologies early on. Although major parts of research on voice recognition now concentrate on the issue of independent speaker identification, many of these parameters remain relevant. A premium is given to create descriptions which are relatively invariant in terms of changes to the speaker in an independent speech recognition. Parameters that reflect the sound's outstanding spectrum energy are sought instead than specifics of the voice of the individual speaker. In this article we are of the opinion that two basic operations comprise of a syntactic pattern recognition approach to language recognition: signal modelling and network research. Signal modelling is the act of turning voice sequences into observer vectors that reflect occurrences within a probability space. The goal of network search is to identify the most likely sequence of such occurrences in view of certain syntax restrictions[9].

Speech has been an essential means of communication among people from antiquity. Speech Recognition is the process by which an auditory discourse is converted and/or a speaker is identified. Due to growing connection between people and computers or automated systems, it has become an important element of our lifestyle throughout the years with current technological advent. A system developed in 1952 at Bell Laboratory, the first word recognition system to be taught to identify numbers. Some of the most commonly used voice recognition systems are speech recognition systems. Speakers, independent speakers, Isolated Word Recognizers, Connected Word Recognizers and Spontaneous Recognition systems are all part of this system[10].

Over the years, speech recognition systems have been very important owing to the established requirement for voice operating systems. The procedure has guaranteed its existence. However, a lot has to be done. Most study to date is based on the notion that speech is a highly subjective phenomenon. Speaker variation, background noise and continuous speech character are the common issues. Perhaps the most obvious cause of deterioration in voice recognition is noise. Noise. Noise may be categorized as environmentally friendly, including traffic, rain, conversation or speakers, such as coughing, sneezing, swallowing, breathing, knocking, etc.

III. CONCLUSION

In this article on analysis the foundations of the speech recognition technique and various approaches used for the extraction and matching of features were given. These many methods may improve the level of language recognition and better quality voice recognition can be produced. In the future, focus will be placed on the development of a wide vocabulary voice recognition system and an autonomous speech recognition system. The Artificial Neural Network (ANN) and the Hidden Markov Model (HMM) will be utilized to enhance these systems in the future as these methods have been used in recent years to recognize the voices.

Speech recognition is a difficult issue to address. In this article, we have tried to examine how far this technology has advanced in past years. The performance of the speech acknowledgement system depends largely on the quality of the signal processing stage. The quality of preprocessing has the greatest effect on the performance of speech classification. EPD, Fileting, Framing, Windowing, Echo Cancellation, etc. are the pre-processing signals. Improving the overall system performance in any particular component. More efforts should be made in front-end processing to operate effectively on the back-end. MFCC is favored by Feature Extraction Technology because it produces training vectors by converting voice data into frequency domain.

REFERENCES

[1]Naziya S. S, Deshmukh RR. Speech Recognition System – A Review. IOSR J Comput Eng. 2016;

[2]Këpuska V. Comparing Speech Recognition Systems

(Microsoft API, Google API And CMU Sphinx). Int J Eng Res Appl. 2017;
[3]Continuous Speech Recognition System A Review. Asian J Comput Sci Inf Technol. 2014;
[4]Washani N, Sharma S. Speech Recognition System: A Review. Int J Comput Appl. 2015;
[5]K.Saksamudre S, Shrishrimal PP, Deshmukh RR. A Review on Different Approaches for Speech Recognition System. Int J Comput Appl. 2015;
[6]Swamy S, K.V R. An Efficient Speech Recognition System. Comput Sci Eng An Int J. 2013;
[7]Rami M, Svitlana M, Lyashenko V, Belova N. Speech Recognition Systems : A Comparative Review. IOSR J Comput Eng. 2017;
[8]Xiong W, Wu L, Alleva F, Droppo J, Huang X, Stolcke A. The Microsoft 2017 Conversational Speech Recognition System. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2018.
[9]Markovnikov NM, Kipyatkova IS. An analytic survey of end-to-end speech recognition systems. SPIIRAS Proc. 2018;
[10]Kurzekar PK, Deshmukh RR, Waghmare VB, Shrishrimal PP. A Comparative Study of Feature Extraction Techniques for Speech Recognition System. Int J Innov Res Sci Eng Technol. 2014;