JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Comparative Study of AI Techniques for Speech-to-Text Conversion

Gurvinder Singh
University Institute of Engineering
Chandigarh University
Mohali, India

Ayush Bhardwaj
University Institute of Engineering
Chandigarh University
Mohali, India

Bhupen Garg
University Institute of Engineering
Chandigarh University
Mohali, India

Manish Kumar Singh
University Institute of Engineering
Chandigarh University

Nivedita Vats
University Institute of Engineering
Chandigarh University Mohali, India

***Abstract*- Speech-to-text conversion, also known as automatic speech recognition (ASR), has become increasingly important in many applications, including virtual assistants, captioning, and transcription. However, building accurate and robust ASR systems is challenging due to the variability of speech signals, including background noise, speaker accents, and speech styles. In recent years, transfer learning has emerged as a promising technique for improving the performance of deep learning models in various tasks, including natural language processing. In this paper, we investigate the effectiveness of transfer learning in the context of speech-to-text systems.**

the potential of transfer learning in improving the accuracy and robustness of speech-to-text systems. This approach has practical implications for building ASR systems in various domains and scenarios.

Keywords: Speech-to-text, automatic speech recognition, transfer learning, fine-tuning, feature extraction, deep learning.

I. INTRODUCTION

Speech-to-text, also known as automatic speech recognition (ASR), is a technology that enables machines to convert spoken language into written text. The use of ASR has grown rapidly in recent years, driven by advancements in deep learning and the increasing need for efficient and reliable transcription in various fields. ASR has significant implications for accessibility, communication, and productivity, making it a valuable tool for individuals and organizations.

Despite its potential benefits, building accurate and robust ASR systems is a complex and challenging task. Speech signals are inherently variable and affected by numerous factors, such as speaker accents, dialects, background noise, and speaking rate. These factors can significantly impact the quality of the transcribed text, making it difficult to achieve high accuracy in ASR systems.

To address these challenges, researchers have proposed various techniques to enhance the performance of ASR systems. Utilizing, deep learning models is one strategy that has drawn a lot of interest because it produces cutting-edge outcomes for many NLP jobs. Deep learning models are trained on large amounts of data, allowing them to learn complex patterns in the input and make accurate predictions.

Another approach that has shown promise in ASR is transfer learning. Transfer learning is a technique that leverages pre-trained models and adapts them to new tasks or domains. This approach has proven effective in various NLP tasks, such as sentiment analysis, machine translation, and text classification. Transfer learning has the potential to improve the performance of ASR systems by leveraging knowledge from pre-trained models and adapting them to new domains with limited labeled data. The goal of the proposed system is to create and deploy a system for converting speech to text for college and universities.

II. THE SPEECH RECOGNITION TECHNOLOGY'S DEVELOPMENT PROCESS AND CURRENT SITUATION

The first ten English digits are recognised by the Bell Labs Audrey voice recognition system, which dates back to the 1950s when speech recognition research initially got underway. However, it made significant strides and became a crucial topic for research in the late 1960s and early 1970s.[1] The HMM model and artificial neural network (ANN) were successfully employed for voice recognition in the 1980s.[2] 1988ˈFULEE The speaker-independent continuous voice recognition system-SPHINX, which has 997 vocabulary, is created by Kai and colleagues using the VQ/Iü IMM approach. It is a high-performance, continuous speech recognition system with a broad vocabulary that is the first of its kind in the world. Finally, people are able to overcome the three main challenges of a big vocabulary, continuous speaking, and non-specific. Additionally, it recognised the widely used statistical techniques. and theories in language processing and speech recognition. Speech recognition technology has already advanced from the lab to the real world; there are more established market products. Many wealthy nations, including the United States, Japan, and South Korea, as well as well-known corporations like IBM, Apple, Microsoft, AT&T, and others, have made significant investments in the study and development of useful voice recognition systems..

III SYSTEM OVERVIEW

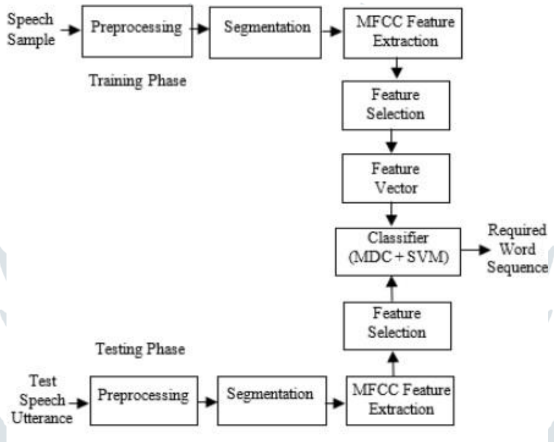

Fig. 1. Block diagram of system

A block diagram of a speech-to-text conversion system is shown in Fig. 1. Training and testing is the two processes that make up the system's operation. Each sentence's initial speech utterances are recorded during the instruction period. Word segmentation and preprocessing are done on the speech signal. The MFCC method is used to extract acoustic information for each syllable. These features for every word in the feature vector are saved for future use. The test spoken utterance is preprocessed, divided into words, and features are extracted for each word during the testing phase. These features are contrasted with the training phase's reference feature vector. SVM and the Minimum Distance Classifier are used in tandem to accomplish this. The term with the fewest differences is designated as a recognised word.

In order to match the voice template stored on the computer and the characteristics of the input voice signal, a computer is used in the speech recognition process.[3] Search and matching techniques to determine the best vocal range for input that matches the pattern. The computer recognition results can be provided in accordance with the definition of this template through the lookup table. The following approaches for speech recognition are typical: dynamic temporal warping (DTW), hidden Markov model (HMM), vector quantization (VQ), artificial neural network (ANN), support vector machine (SVM), and others. The two techniques of the hidden Markov model (HMM) and artificial neural network (ANN) are the main topics of the article.

A. Hidden Markov Model (HMM)

A statistical model used to describe a series of observations is called a hidden Markov model (HMM).[4] It is based on the idea of a Markov process, a stochastic process in which the probability of the following state depends solely on the present state and not on the past states. The hidden states and the visible states are the two main parts of an HMM.

The hidden states are the states that the system goes through, but which are not directly observable. For example, in speech recognition, the hidden states could represent the different phonemes (units of sound) that are being spoken, while the observable states are the acoustic features of the speech signal, such as the energy or frequency of the signal at different time intervals.

The HMM assumes that the sequence of hidden states is a Markov process, and that the observable states are generated from the hidden states according to a probability distribution. This probability distribution is often modeled using a Gaussian mixture model, which is a weighted sum of multiple Gaussian distributions. The parameters of the model, such as the means and variances of the Gaussian distributions, are learned from training data using the Expectation-Maximization algorithm.

Many different applications, including bioinformatics, handwriting recognition, and speech recognition, make extensive use of HMMs. In order to discover the most likely sequence of hidden states given the known acoustic features, the Viterbi algorithm is employed in speech recognition.[5] HMMs are used to describe the probability distribution of various phonemes.

B. *Artificial Neural Network (ANN)*

Artificial neural networks (ANNs) can be used in speech to text applications to convert spoken language into written text. This process is known as automatic speech recognition (ASR).

ASR systems typically consist of multiple ANNs that work together to perform various tasks such as feature extraction, acoustic modeling, and language modeling. [6]The input to the system is the raw audio signal, which is first transformed into a sequence of feature vectors representing the speech signal. These feature vectors are then used as input to the acoustic model, which predicts a sequence of phonemes (the basic units of sound in a language) that correspond to the spoken words.

The output of the acoustic model is then fed into the language model, which uses probability-based techniques to predict the most likely sequence of words that correspond to the phoneme sequence. Finally, the output of the language model is transformed into written text.

One of the challenges of ASR systems is dealing with variations in speech, such as accents, background noise, and speaking style.[7] ANNs can be trained on large datasets of speech recordings to learn to recognize and adapt to these variations. Additionally, recent advancements in deep learning techniques, such as recurrent neural networks and convolutional neural networks, have led to significant improvements in the accuracy of ASR systems.

ASR systems have numerous applications, including transcription of audio and video content, voice-activated assistants, and closed captioning for broadcast and streaming media.

Figure 2 illustrates the speech recognition process using artificial neural networks, together with the e-learning process. Speech signal is known as a learning sample in the network 200 learning process, a self-learning neural network, and eventually a collection of connection weights and bias. The voice signal is tested as network input throughout the speech recognition process, and the results of the recognition are acquired using an association network. The key to these two procedures is to balance neural network learning with voice characteristic parameters.

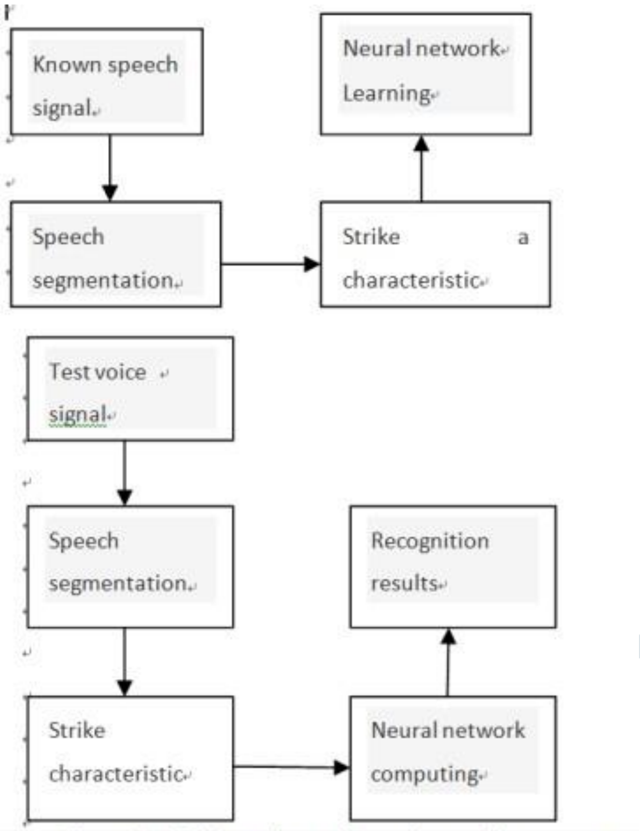

Figure 2 Artificial neural network speech recognition process

A recent hot topic is the use of artificial neural networks for voice detection. We can anticipate that in the coming decade, artificial neural network- based speech recognition system products will hit the market. As a result of artificial neural network technology's success in solving pattern classification problems and demonstration of its tremendous power, people will modify their speech patterns to accommodate a variety of detection systems.

IV. THE PROBLEMS AND THE USE OF SPEECH RECOGNITION TECHNOLOGY

A. Use of speech recognition technology

A future where research and development are expedited The voice recognition system's commercial operation has begun thanks to the speech recognition program. Traditional speech recognition technology - VRCP technology It was created by AT&T in 1992.[8] When used with AT&T Communications Online, the system is 5 words (collect, person, third number, operator, phone card), a speech recognition system for general little words. Instead, it offers an automated operator- assisted call. Five different call types have been finished by the operator. Charles Schwab developed the first extensive commercial voice recognition software system in September 1996[9]. System of stock prices. system was moreover the first in the banking industry to use speech recognition. From The quality of services can be improved with this system. enhancing client pleasure while cutting call center expenses. Schwab began his speech by discussing the stock trading mechanism.

It is a division of Sprint PCS, a significant US carrier that also boasts the biggest digital wireless network.

Recognised for its outstanding and creative customer service. We have been providing our clients with access to audio systems since 2000.[10] This system provides customer support, voice dialing, number verification, address change, and other services. Additionally, China Telecom introduced the voice value-added system (VOICEVALUE-ADDED SYSTEM) CELL-

VVAS value added service system. The system's effective use is supported by a superior distributed, reliable detection engine and a fully integrated, multi-service communications exchange network. a program that offers consumers a variety of functionalities.

The development of voice-activated telephones is a distinct area of speech recognition technology research. In terms of cognitive technology, Bell Labs is a pioneer.

B. ***The problems of Speech recognition technology*** Speech recognition research is currently making slow, primarily theoretical, progress. Despite the fact that fresh modifications of all kinds are appearing, there is also a lack of universal applicability. primarily in The voice recognition system's poor adaptability is mostly manifested in its reliance on the environment, If you utilise a speech training system to gather data under specific conditions, you can only use it in that environment; otherwise, system performance will rapidly deteriorate. Another issue is that this system does not react appropriately to user input errors. Additionally, it is particularly challenging to advance speech recognition in loud contexts because at this because of the Lombard effect, when people's pronunciation fluctuates significantly over time, it is necessary to develop a new signal analysis and processing strategy. Second, the utilization of the existing achievements of this aspect in speech recognition is still a challenging process. Understanding of the human auditory comprehension, the accumulation of information and learning mechanism, and the system of the brain control mechanism are still unclear.

V. CONCLUSION

The issues with speech recognition include The usage of speech recognition technologies is still widespread. Numerous things may be done better. However, as technology develops, speech recognition will soon be a reality thanks to voice recognition. The application of voice will make the system more complex. The identifying process will be more thorough. People will modify their speech patterns to fit into various human recognition systems when they emerge on the market, which is also a temporary solution. People that are akin to speech cannot be created. It still takes a lot of work to develop a belief system. Humanity has a hurdle; we can only advance incrementally and in the right ways in terms of speech recognition technology.

VI. REFERENCES

[1] Yogita H. Ghadage, Sushama D. Shelke , "Speech to Text Conversion for Multilingual Languages",International Conference on Communication and Signal Processing, April 6-8, 2016,India

[2] Jianliang Meng, Junwei Zhang, " Overview of the Speech Recognition Technology", 2012 Fourth International Conference on Computational and Information Science

[3] Huang Shan. Voice recognition systems in the telecom prepaid business applications [J]. Information Science, 2010.

[4] Umar Nasib Abdullah, Kabir Humayun, Ahmed Ruhan, Uddin Jia., "A Real Time Speech to Text Conversion Technique for Bengali Language," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), pp. 1-4, 2018

[5] L. Wan, "Extraction Algorithm of English Text Summarization for English Teaching," 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Xiamen, China, 2018, pp. 307-310.

[6] Saiyed S., Sajja P. S., "Review on text summarization evaluation methods," Indian Journal of Computer Science and Engineering, vol. 8, no. 4, pp. 497, 2017.

[7] R. Alguliyev, R. Aliguliyev and N. Isazade, "A sentence selection model and HLO algorithm for extractive text summarization," 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), Baku, pp. 1-4, 2016.

[8] Jain D. Bhatia and M. K. Thakur, "Extractive Text Summarization Using Word Vector Embedding," 2017 International Conference on Machine Learning and Data Science (MLDS), Noida, pp. 51-55, 2017.

[9] K. Vythelingum, Y. Estève and O. Rosee, "Error detection of grapheme-to-phoneme conversion in text-to-speech synthesis using speech signal and lexical context," 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, pp. 692-697, 2017

[10] J. Zenkert, A. Klahold and M. Fathi, "Towards Extractive Text Summarization Using Multidimensional Knowledge Representation," 2018 IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, pp. 0826-0831, 2018.

[11] S. R. Rahimi, A. T. Mozhdehi and M. Abdolahi, "An overview on extractive text summarization," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, pp. 0054-0062, 2017

[12] A. Vimalaksha, S. Vinay, A. Prekash and N. S. Kumar, "Automated Summarization of Lecture Videos," 2018 IEEE Tenth International Conference on Technology for Education (T4E), Chennai, India, pp. 126-129, 2018.