International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 10 | Oct 2022 www.irjet.net p-ISSN: 2395-0072

Review On Speech Recognition using Deep Learning

Anushree Raj[1], Sahir Abdulla[2], Vishwas N[3]

[1] *Assistant Professor- IT Department, AIMIT, Mangaluru, anushreeraj@staloysius.ac.in*
[2] *MCA Student, AIMIT, Mangaluru, 2117117vishwas@staloysius.ac.in*
[3] *MCA Student, AIMIT, Mangaluru, 2117044sahir@staloysius.ac.in*

Abstract: Speech is the most effective means for humans to communicate their ideas and emotions across a variety of languages. Every language has a different set of speech characteristics. The tempo and dialect vary from person to person even when speaking the same language. For some folks, this makes it difficult to understand the message being delivered. Long speeches can be challenging to follow at times because of things like inconsistent pronunciation, tempo, and other factors. The development of technology that enables the recognition and transcription of voice into text is aided by speech recognition, an interdisciplinary area of computational linguistics. The most crucial information is taken from a text source and adequately summarized by text summarization.

Key words: Speech recognition, Deep learning, computational linguistics, feature extraction, feature vectors.

1. INTRODUCTION

To select the proper output, some Voice is frequently used and regarded as important information while engaging with others. Through comprehension and recognition, voice recognition technology enables machines to convert human vocal signals into equivalent commands. Speech is the most effective form of expression for thoughts and feelings when learning new languages. In the survey we conducted this is very useful when we want to communicate with others. This project will convert speech to text or text to speech using deep learning technique using CNN (conventional neural networking), Just like Google's google Assistant ,Apple's SIRI, Samsung's Bixby. A combination of speech to text conversion and text summarization is used in the suggested work. Applications that call for concise summaries of lengthy talks will benefit from this hybrid approach, which is quite helpful for documentation. Deep learning is a sort of AI and machine learning that mimics how people learn specific types of information. Nowadays, numerous applications use human-machine interaction [1]. Speech is one of the interactional media. The primary difficulty in human-machine interaction is identifying emotions in speech.

2. OBJECTIVES

The objective of voice recognition is to use linguistic and phonetic data to convert the input speech feature vector series into a sequence of words. A full voice recognition system, according to the system's structure, consists of a feature extraction algorithm, acoustic model, language model, and search algorithm. A multidimensional pattern recognition system is essentially what the speech recognition system does.

Speech recognition provides input for automatic translation, generates print-ready dictation, and allows hands-free operation of various devices and equipment—all of which are especially helpful to many disabled people. Medical dictation software and automated telephone systems were some of the first speech recognition applications [2].

Speech recognizers are made up of a few components, such as the speech input, feature extraction, feature vectors, a decoder, and a word output. The decoder leverages acoustic models, a pronunciation dictionary, and language models Benefits:

- It can help to increase productivity in many businesses, such as in healthcare industries.
- It can capture speech much faster than you can type.
- You can use text-to-speech in real-time.
- The software can spell with the same ability as any other writing tool.

Helps those who have problems with speech or sight.

3. LITERATURE REVIEW

The most crucial component of human communication is speech. Although there are many ways to express what we think and feel, speaking is often regarded as the primary form of communication. The Google API can be used to convert recorded speech to text. Because the retrieved text does not contain a period, it is challenging to split the content into sentences that were created using the Google API. In the suggested model, a period is added at the end of each phrase to distinguish them from one another.

The theoretical algorithms used to construct voice recognition were explained in this study. The precise steps involved in voice recognition, such as biometrics acquisition, preprocessing, feature extraction, biometrics pattern matching, and recognition outcomes, are first described. The detailed introduction of speech recognition in biological features [3]. The primary procedures, recognition strategies,

and application situations for voice recognition are outlined in this paper. The secret to ensuring recognition effectiveness is learning how to extract feature information sensibly.

The voice-input voice-output communication aid (VIVOCA), a novel type of augmentative and alternative communication (AAC) technology for people with severe speech impairment, is described. The VIVOCA creates messages from the user's disordered speech and transforms them into synthetic speech. The findings demonstrate that the VIVOCA device performed better when phrase construction was the goal rather than spelling [4]. This is because there are normally 3–10 competing words in these trials, which makes the ambiguity relatively low in the phrase building mode.

Computer models can be used to predict voice recognition. There are numerous files with a variety of audio and audio files in huge audio or video files with many minutes in length. This researcher selected the appropriate sound from a sizable file to listen to. Deep learning was employed in this study to categorize speech. The model was trained using the Google corpus. We had accuracy of 66.22%. The detailed introduction of speech recognition in biological features [5]. The primary procedures, recognition strategies, and application situations for voice recognition are outlined in this paper. The secret to ensuring recognition effectiveness is learning how to extract feature information sensibly.

When the audio is distorted by noise, the audio-visual speech recognition (AVSR) system is regarded as one of the most promising options for accurate speech recognition. To achieve good recognition performance, however, careful sensory feature selection is essential. In this study, they suggested an AVSR system based on MSHMMs for multimodal feature integration and isolated word recognition in addition to deep learning architectures for audio and visual feature extraction [6]. Our test findings showed that the deep denoising auto encoder can efficiently remove the effect of noise superimposed on original clean audio inputs when compared to the original MFCCs.

They discuss our research on the usefulness of using representation learning on sizable unlabeled speech corpora for speech emotion recognition (SER). The relatively small emotional speech datasets were the main focus of earlier work on representation learning for SER, and no further unlabeled speech data were utilized. They have demonstrated in this study that adding representations produced by an auto encoder that was trained on a sizable dataset consistently increases the provided SER model's recognition accuracy [7]. Additionally, we provided t-SNE visualizations that demonstrate the representations' ability to discriminate between low and high levels of arousal.

4. CHALLENGES AND ISSUES

During the last few years, speech recognition has improved a lot. This can mainly be attributed to the rise of graphics processing and cloud computing, as these have made large data sets widely distributable.

Some Challenges are:

- Audio / Video Conferencing with Background Noise.
- Speech Recognition and Voice Assistant Devices.
- Lack of Trust and Privacy Issues.
- Touch less Screens.
- The Future of Voice Recognition Technology.
- In Summary.

With recent developments, it's going to be interesting to see how the momentum of rapid growth can be maintained and how the current challenges of speech recognition will be dealt with [8].

The goal of Automatic Speech Recognition (ASR) for the past five to ten years has been to decode voice inputs as accurately as possible. Systems like Siri, Alexa, and Google Assistant, which are well-known, were made feasible by this. Voice recognition has entered our daily lives thanks to these well-known voice assistants. In this essay, we'll examine the speech recognition industry's existing difficulties and potential future advances. Reach and loud settings are the two main causes of the current difficulties in voice detection. This necessitates even more accurate systems that are capable of handling the most challenging ASR use-cases. Consider speech recognition during a boisterous family dinner, live interviews, or group meetings [9]. These are the upcoming difficulties for next-generation voice recognition.

Beyond this, voice recognition needs to support additional languages and a larger range of subjects. A lot of the data that ASR currently needs to operate well has simply not been obtained for certain languages and topics. Without them, ASR systems will remain quite constrained. Voice assistants and Voice Powered User Interfaces (VUIs) have a straightforward use-case. They enable spoken commands from people to be translated into actions by machines. Even if the use-case seems to be crystal evident, the ideal approach to human-machine interactions is still being developed [10]. Naturally, speech recognition will have difficulties as a result.

5. CONCLUSSION

Here we have some dictating tips to consider for improved results:

- Speak in an even tone and clarity. If you are whispering, then words could not be interpreted in the correct way.

- It is always better to pause before and after a command while avoid taking a pause in the midst of command issuance. So that it could not be interpreted as a dictation.

- Prefer to speak in complete sentences even including punctuation, to give proper context.

Here are the ways that we can consider bringing improvement in our text to speech technology in the best possible way:

1) Understand the type of errors

A text to speech tool more often comes up with the array of words based on what it has heard. This is what the tools have been designed to do. However, deciding which words string it has heard can become a bit tricky for it. Therefore, there can occur errors that can throw users off.

While guessing the wrong word is one of the classic problems in speech technology because they present all kinds of potential mishearing that sound similar. But a whole sentence might make some sense.

Use high-quality headset microphone

Using a high-quality headset microphone is one of the most important factors to improve voice recognition. It is because these are not only capable of catching the right words, but also have the ability to hold a microphone in front of your mouth at a consistent position directly. Therefore, these can offer more desirable. This can help you in getting more desirable speech recognition results by remaining positioned consistently.

2) Make corrections

Most commonly speech technology learns from the corrections that are being made by you. It is because most of these tools are based on artificial intelligence and deep learning technology. Therefore, these will be going to learn your corrected words and will use those for the next time.

3) Use automatic formatting

There are some tools available in speech recognition technology that offer automatic formatting solutions. These can help in formatting various types of text automatically. It can also help your text to speech solutions to format specific phrases and words as per your preferences.

There is a continuous improvement in text to speech technology, but speech recognition systems are having great difficulty in attaining 99% accuracy. However, considering these some effective tips can help you in getting better results.

recognition. As computer information technology develops, speech recognition technology will advance considerably. Several businesses, including public security, mobile Internet security, and automotive network security, are expected to employ this technology. Speech recognition research consequently has two main objectives: improving the information society and boosting living standards. Speech recognition is a key man-machine interface tool in information technology with strong scientific importance and vast application usefulness. In this study, biological features are completely introduced for speech recognition. This article describes the key steps, recognition tactics, and application scenarios for voice recognition. Ability to reasonably extract feature information is necessary for effective

REFERENCES

[1] Xinman Zhang School of Electronics and Information Engineering, MOE Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University Xi'an, China e-mail: zhangxinman@xjtu.edu.cn "An Overview of Speech Recognition Technology" 2019 4th International Conference on Control, Robotics and Cybernetics (CRC)

[2] Mark S. Hawley, Stuart P. Cunningham, Phil D. Green, Pam Enderby, Rebecca Palmer, Siddharth Sehgal, and Peter O'Neill" A Voice-Input Voice-Output Communication Aid for People With Severe Speech Impairment" IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, VOL. 21, NO. 1, JANUARY 2013

[3] Phoemporn Lakkhanawannakun, Chaluemwut Noyunsan Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Isan, Khon Kaen Campus, Khon Kaen, 40000, Thailand" Speech Recognition using Deep Learning"

[4] Kuniaki Noda · Yuki Yamaguchi · Kazuhiro Nakadai · Hiroshi G. Okuno · Tetsuya Ogata" Audio-visual speech recognition using deep learning" Published online: 20 December 2014 © Springer Science+Business Media New York 2014

[5] Michael Neumann, Ngoc Thang Vu University of Stuttgart, Germany {michael.neumann|thang.vu}@ims.uni-stuttgart.de" Published online: 20 December 2014 © Springer Science+Business Media New York 2014".

[6] mY. H. Ghadage and S. D. Shelke, "Speech to text conversion for multilingual languages," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, pp. 0236-0240, 2016.

[7]] F. Zheng, L. T. Li, M. Z. Escale, and H. Zhang, "Voice print recognition technology and its application status,"Research on Information Security, vol. 2, no. 1, Jan. 2016, pp. 44–57.

[8] Conference on Acoustics, vol. 24, no. 7, July. 2012, pp. 1315–1329. [9] C. H. Zhou, "Research on Speaker recognition system based on MFCC feature and GMM Model," Ph.D. dissertation, Dept. Electron. Eng., Lanzhou University of Technology, Lanzhou, China, 2013

[9] Jose D V, Alfateh Mustafa, Sharan R, "A Novel Model for Speech to Text Conversion," International Refereed Journal of Engineering and Science (IRJES), vol 3, no. 1, 2014.

[10] L. Liu, "Research on Fusion and Recognition Methods on Multimode Biometrics," Ph.D. dissertation, Dept. Electron. Eng., University of Electronic Science and Technology, Chengdu, China, 2010.