International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | Jun 2022 www.irjet.net p-ISSN: 2395-0072

A survey on Enhancements in Speech Recognition

Nitin Kendre[1], Vivek Parit[2], Asazad Pathan[3], Akash Kamble[4], Prof. T.V. Deokar[5]

[1,2,3,4] *Students, Department of Computer Science & Engineering, Sanjeevan Engineering and Technology Institute Panhala (Maharashtra)*
[5] *Assistant Professor, Department of Computer Science & Engineering, Sanjeevan Engineering and Technology Institute Panhala (Maharashtra)*

Abstract – *Purpose of This study is to know the enhancements in speech recognition field. From starting of the 21st century people are working and researching in the area of voice recognition. Researchers have contributed many things in this area. Normal speech without any noise is easy to understand by computers, but if the speech includes noise, then it is very difficult to understand by computer and separate the noise from the speech. There are various reasons to have noise in speech like background noise, environmental noise, signal noise, crowded places, etc. In this paper, we are going to present various techniques to enhance the speech recognition system to work in any environment by researchers. Also, some advanced enhancements in speech recognition to use this system in other situations like emotion recognition.*

Key Words: **Robust Speech Recognition, Artificial Intelligence, Feature Extraction, Noise Reduction, Deep Learning.**

1. INTRODUCTION

The technique through which a computer (or another sort of machine) recognizes spoken words is known as speech recognition. Essentially, it is conversing with your computer and having it accurately recognize your words. Simply, it means talking to the computer and having it correctly recognize what you are saying.

Voice is the most common and fastest mode of communication, and each human voice has a distinct quality that distinguishes it from the others. As a result, not only for humans but also for automated machines, voice recognition is required for easy and natural interaction [12].

Speech recognition has applications in a variety of sectors, including medicine, engineering, and business. The general problem with speech recognition is speaking rate, gender, age, and the environment in which the discussion is taking place, and the second issue is speech noise [12]. If we can solve these issues with speech recognition, it will be much easier to create goods or systems that people can use everywhere, even in crowded areas or in noisy environments.

Therefore, it is necessary to remove or reduce the amount of noise in a speech to do effective recognition of speech or voice. And to reduce or remove the noise from speech we have to know the basics of recognition. The basic model of speech recognition or speech-to-text model is shown in figure 1. Figure 1 depicts the basic model of speech recognition, also known as the speech-to-text paradigm.

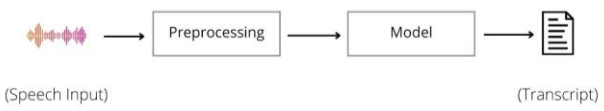

Fig-1: The basic model of Speech Recognition.

1.1 Speech Input

A human voice is captured or recorded using a microphone and sound card connected to the computer as speech input. Modern sound cards can record audio at sampling rates ranging from 16 kHz to 48 kHz, with bit rates ranging from 8 to 16 bits per sample, and playback speeds of up to 96 kHz [12].

1.2 Preprocessing

Signal processing takes place in this step. This process converts an analog signal to a digital signal and does noise reduction, as well as changes audio frequencies to make it machine-ready [12][13].

1.2.1 Feature Extraction

The next step in pre-processing is to choose which features will be valuable and which will be unnecessary. We need to understand MFCCs (Mel Frequency Cepstral Coefficients) in order to extract features.

1.2.2 MFCCs

MFCCs is a method for extracting features from audio signals. It divides the audio signal's frequency bands using the MEL scale, then extracts coefficients from each frequency band to create a frequency separation. The Discrete Cosine Transform is used by MFCC to conduct this operation. The MEL scale is based on human sound

perception, or how the brain analyses audio impulses and distinguishes between different frequencies.

Stevens, Volkmann, and Newmann proposed a pitch in 1937 that gave birth to the MEL scale. It's a scale of audio signals with varying pitch levels that people judge based on their distances being equal. It's a scale that's based on how people perceive things. For example, if we hear two sound sources that are far apart, our brain will be able to determine the distance between them without having to see them. This scale is based on how we humans use our sense of hearing to determine the distance between audio signals. The distances on this scale increase with frequency because human perception is non-linear.

This MEL scale is used to separate the frequency bands in audio, after which it extracts the needed information for recognition.

1.3 Model

Classification models are mostly used in the model section to recognize or search for words detected in audio. It is now simple to recognize speech using neural networks, and employing neural networks for speech recognition has proven to be quite beneficial. For speech recognition, RNNs (Recurrent Neural Networks) are commonly used. Speech recognition can also be done without the use of neural networks, although it will be less accurate.

Below are some models used for speech recognition, like recurrent neural network, Hidden Markov Model etc.

1.3.1 Recurrent Neural Network

Recurrent Neural Networks are a type of neural network that is used mostly in NLP. RNNs are a type of neural network that uses past outputs as inputs while maintaining hidden states. RNN features a memory notion that preserves all data about what was calculated up to that time step. RNNs are referred to as recurrent because they do the same task for each combination of inputs, with the result dependent on past calculations. Fig-2 shows the simple illustration of recurrent network [15].

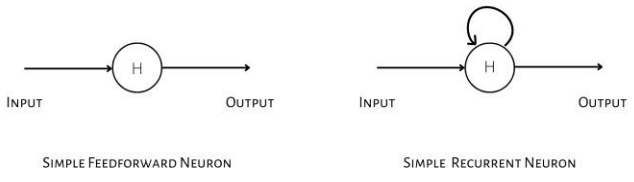

Fig-2: Simple Architecture of RNN.

Recurrent neural network is used for voice classification in speech recognition. Also, it is used for language translation.

1.3.2 Hidden Morkov Model

When any person speaks, basically it creates the vibrations. To convert this speech into text computer uses several complex steps. First it converts our analog speech signals into digital signals. Then it does pre-process on that signal, like it removes non needed part from signals, it removes noise and do feature extraction. That is, it selects required features from signals and pass it to the Hidden Markov Model, now HMM do the search for word in Database. Then after word searching it sent to the users [14].

Fig-3 shows the simple architecture of Hidden Markov Model for Speech Recognition.

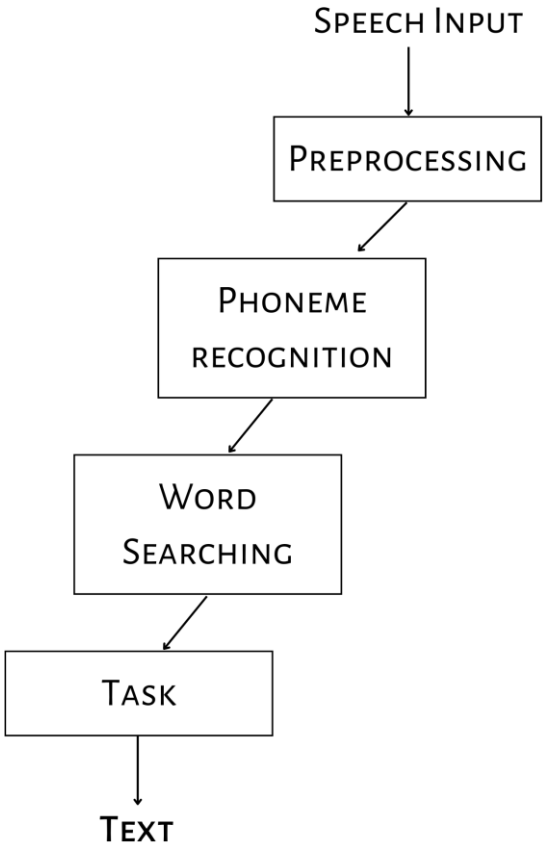

Fig-3: HMM for Speech Recognition

1.4 Transcript

The transcript is the text obtained after the user's audio has been converted to text. The Speech to Text method is the name given to this procedure. The transcript is then sent into a text-to-speech model or system, which allows the computer to speak.

2. HISTORY OF SPEECH RECOGNITION

"Audrey," a system built by Bell Laboratories in the 1950s, was the first authorized example of our contemporary speech recognition technology. However, audrey takes a full room to be set up and can only identify 9 numbers uttered by its inventor with 90% accuracy. Its goal was to assist toll operators in taking more telephone call over the line, but its high price and inability to detect a wide range of voices rendered it unfeasible.

The next innovation was IBM's Shoebox, which debuted at the World's Fair in 1962, took nearly 12 years to build and was able to detect and discriminate between 16 words. However, person had to pause and talk slowly to ensure that the system picked up on what they were saying.

Then, The Defence department began to appreciate the significance of voice recognition technology in the early 1970s. The capacity of a computer to interpret genuine human language might be extremely useful in a variety of military and national defence applications.

As a result, they invested five years on DARPA's Speech Understanding Research programme, one of the largest initiatives of its sort in speech recognition history. One of the most notable discoveries to emerge from this research effort was "Harpy," a system capable of recognising over 1000 words.

Speech recognition systems were so common in the late 1970s and early 1980s that they found their way into children's toys. The Speak & Spell, which used a voice chip, was created in 1978 to assist kids spell words. The voice chip included therein would prove to be a significant instrument for the next stage of speech recognition software development.

Then the "Julie" doll was released in 1987. Julie was capable to answer to a person and identify between speaker's voices in an amazing (if not disturbing) exhibition.

Then after Three years AT&T was experimenting with over-the-phone voice recognition technologies to help answer customer support calls.

Then Dragon launched "Naturally Speaking" in 1997, which allowed natural voice to be interpreted without the need of pauses.

And after that much innovations and research in this field today we have digital assistants like Google Assistant, Alexa, Siri, Cortana which can process natural speech without pauses and can recognize thousands of words in any language.

3. RELATED WORK

Many researchers have worked on various techniques to enhance the speech recognition system in any environment. Given enhancements in various papers proved helpful in speech recognition systems.

3.1 Automatic Recognition of Noisy Speech

Jisi Zhang et.al. [1], have explored A framework for training a monoaural neural augmentation model for robust voice recognition. To train the model, they used unpaired speech data and noisy speech data. They conducted trials using an end-to-end acoustic model and found that the WER was lowered by 16 to 20%. In the real development and evaluation sets, using a more powerful hybrid acoustic model, the processed speech reduced WER by 28% and 39%, accordingly, when compared to the unprocessed signal.

3.2 Integration of Speech Recognition and Self-Supervised learning

Xuankai Chang et.al. [2], enhanced the end-to-end ASR (Automatic Speech Recognition) model for powerful voice recognition. When comparing their suggested model to a traditional end-to-end ASR model, we can see that theirs incorporates a speech augmentation module and a self-supervised learning representation module. The voice enhancement module improves the sound quality of loud speech. The self-supervised learning representation module, on the other hand, performs the feature extraction and picks useful features for voice recognition.

3.3 Noisy Robust Speech Recognition

Developing a speech recognition system that works in noisy surroundings necessitates a vast amount of data, including noisy speech and transcripts, to train the system, yet such large data is not readily available. Chen Chen et al. have suggested a generative adversarial network to recreate a noisy spectrum from a clean spectrum with only 10 minutes of in-domain noisy voice data as labels. They've also proposed a dual-path speech recognition method to boost the system's robustness in noisy situations. The proposed technique improves the absolute WER on the dual-path ASR system by 7.3% to the best baseline, according to the experimental data [3].

3.4 Speech Emotion Recognition

Surekha Reddy Bandela et.al. [4], have presented a strategy that combines unsupervised feature selection in combination with INTERSPEECH 2010 paralinguistic characteristics, Gammatone Cepstral Coefficients (GTCC), and Power Normalized Cepstral Coefficients (PNCC) to improve speech recognition for emotion recognition. In both clean and noisy situations, the proposed system is tested. For de-noising the noisy voice signal, they adopted a dense Non-

Negative Matrix Factorization (denseNMF) approach. For Speech Emotion Recognition analysis, they employed the EMO-DB and IEMOCAP databases. They were able to achieve an accuracy of 85 % using the EMO-DB database and 77 % using the IEMOCAP database. With cross-corpus analysis, the proposed system can be improved for language independence [4].

3.5 Speech Recognition based on multiple Deep Learning Features

Because the English language has the advantage of huge datasets in speech recognition, everyone loves it for speech recognition systems. Speech recognition in the English language is a simple task thanks to big datasets and a wide community. As a result, Zhaojuan Song et al. [5], chose English Speech as their study topic and created a deep learning speech recognition method that incorporates both speech features and speech attributes. They used the deep neural network supervised learning approach to extract the high-level qualities of the voice and train the GMM-HMM ANN model with additional speech features. They also used a speech attribute extractor based on a deep neural network trained for several speech attributes, with the retrieved features being categorized into phonemes.

Then, using a neural network based on the linear feature fusion technique, speech characteristics and speech attributes are integrated into the same CNN framework, considerably improving the performance of the English voice recognition system [5].

3.6 Robust Speech Recognition

A.I. Alhamada, et al. al[6], have looked into previous work and discovered a suitable deep learning architecture. To improve the performance of speech recognition systems, they used a convolutional neural network. They discovered that a CNN-based approach had a validated accuracy of 94.3%.

3.7 Hybrid-Task Learning for ASR

Gueorgui Pironkov, et.al. [7]. presented a new technique named Hybrid-Task learning. This method is based on a mix of Multi-Task and Single-Task learning architectures, resulting in a dynamic hybrid system that switches between single and multi-task learning depending on the type of input feature. This hybrid task learning system is especially well suited to robust automatic speech recognition in situations where there are two types of data available: real noise and data recorded in real-life conditions, and simulated data obtained by artificially adding noise to clean speech recorded in noiseless conditions [7].

As a result of Hybrid Task Learning, ASR performance on actual and simulated data beats Multi-Task Learning and Single Task Learning, with Hybrid Task Learning bringing up to a 4.4% relative improvement over Single Task Learning. This Hybrid Task Learning technique can be tested on different datasets in the future [7].

3.8 Audio Visual Speech Recognition

An audio-visual speech recognition system is believed to be one of the most promising technologies for robust voice recognition in a noisy environment. As a result, Pan Zhou et al. [9] proposed a multimodal attention-based approach for audio-visual speech recognition, which could automatically learn the merged representation from both modes depending on their importance. The suggested method achieves a 36% relative improvement and beats previous feature concatenation-based AVSR systems, according to experimental results. This strategy can be tested with a larger AVSR dataset in the future [9].

3.9 Audio Visual Speech Recognition for Hearing Impaired

Hearing-impaired persons would benefit greatly from artificial intelligence; approximately 466 million people worldwide suffer from hearing loss. Students who are deaf or hard of hearing depend on lip-reading to grasp what is being said. However, hearing-impaired students encounter other hurdles, including a shortage of skilled sign language facilitators and the expensive cost of assistive technology. As a result, L Ashok Kumar et al. [8], proposed an approach for visual speech recognition built on cutting-edge deep learning methods They merged the results of audio and visual. They also suggested a novel audio-visual speech recognition algorithm based on deep learning for efficient lip reading. They reduced the word error rate in the Automatic Speech Recognition system to about 6.59 % and established a lip-reading model accuracy of nearly 95 % using this model. In the future, a BERT-based language model could improve the suggested model's Word Error Rate [8].

3.10 Robust Speech Recognition Using Reinforcement Learning.

The goal of Deep Neural Network-based voice enhancement is to reduce the mean square error among improved speech and a clean reference. As a result, Yih-Liang Shen et al. [10] suggested a reinforcement learning approach to improve the voice enhancement model based on recognition results. They used the Chinese broadcast news dataset to test the proposed voice enhancement system (MATBN). The RL-Based voice enhancement may effectively reduce character mistake rates by 12.40% and 19.23% by using recognition errors as the goal function. More noise kinds will be investigated in the future when developing the RL-based system [10].

3.11 Noise Reduction in Speech Recognition

We've seen that reducing speech noise is essential for constructing reliable speech recognition systems in noisy contexts. As a result, Imad Qasim Habeeb et al. [11] suggested an ensemble approach that uses numerous noise-reducing filters to increase the accuracy of Automatic Speech Recognition. To boost Automatic Speech Recognition Accuracy, they applied three noise reduction filters. These three filters generate numerous copies of the speech signal, with the best copy chosen for the voice recognition system's final output. They studied this model and found promising results, such as a 16.61% drop-in character error rate and an 11.54% reduction in word error rate when compared to previous strategies [11].

The results of the experiments showed that alternative noise reduction filters could potentially provide additional information about the phonemes to be categorized, which could be used to improve the Automatic Speech Recognition System's accuracy. However, because a human can guess words or phrases even if he doesn't hear them completely, a technique based on the n-gram model can be developed to give a mechanism similar to the human hearing system [11].

4. DISCUSSION AND CONCLUSIONS

This study discusses strong voice recognition and critical factors in speech recognition, as well as diverse methodologies used by different researchers. We've looked at how speech recognition works, covering audio recording, voice pre-processing, classification models, and neural networks for speech recognition. We also spoke about how to extract features from speech signals. An overview of tests conducted by authors with their work has also been included.

Robust speech recognition is still a long way off for humans, and developing such a system with good performance is a difficult undertaking. In this paper, we have tried to provide a comprehensive overview of key advancements in the field of Automatic Speech Recognition. We've also attempted to list some research that could be useful in developing systems for physically challenged people.

REFERENCES

[1] Jisi Zhang, Catalin Zoril a, Rama Doddipatla and Jon Barker "On monoaural speech enhancement for automatic recognition of real noisy speech using mixture invariant training." arXiv preprint arXiv:2205.01751, 2022.

[2] Xuankai Chang, Takashi Maekaku, Yuya Fujita, Shinji Watanabe "End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation." arXiv preprint arXiv:2204.00540, 2022.

[3] Chen Chen, Nana Hou, Yuchen Hu, Shashank Shirol, Eng Siong Chng "NOISE-ROBUST SPEECH RECOGNITION WITH 10 MINUTES UNPARALLELED IN-DOMAIN DATA." arXiv preprint arXiv:2203.15321, 2022.

[4] Surekha Reddy Bandela, T. Kishore Kumar "Unsupervised feature selection and NMF de-noising for robust Speech Emotion Recognition", Applied Acoustics, Elsevier, Vol 172, January 2021.

[5] Zhaojuan Song "English speech recognition based on deep learning with multiple features." Springer, 2019.

[6] A. I. Alhamada, O. O. Khalifa, and A. H. Abdalla "Deep Learning for Environmentally Robust Speech Recognition", AIP Conference Proceedings, 2020

[7] Gueorgui Pironkov, Sean UN Wood, Stephane Dupont "Hybrid-task learning for robust automatic speech recognition." Elsevier, Computer Speech & Language, Volume 64, 2020.

[8] L Ashok Kumar, D Karthika Renuka, S Lovelyn Rose, M C Shunmuga Priya, I Made Wartana "Deep learning based assistive technology on audio visual speech recognition for hearing impaired.", International Journal of Cognitive Computing in Engineering, Volume 3, 2022.

[9] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, Jia Jia "MODALITY ATTENTION FOR END-TO-END AUDIO-VISUAL SPEECH RECOGNITION", arXiv preprint arXiv:1811.05250, 2019.

[10] Yih-Liang Shen, Chao-Yuan Huang, Syu-Siang Wang, Yu Tsao, Hsin-Min Wang, and Tai-Shih Chi, "REINFORCEMENT LEARNING BASED SPEECH ENHANCEMENT FOR ROBUST SPEECH RECOGNITION ", arXiv preprint arXiv:1811.04224, 2018.

[11] Imad Qasim Habeeb, Tamara Z. Fadhil, Yaseen Naser Jurn, Zeyad Qasim Habeeb, Hanan Najm Abdulkhudhur, "An ensemble technique for speech recognition in noisy environments.", Indonesian Journal of Electrical Engineering and Computer Science Vol. 18, No. 2, May 2020

[12] Pranjal Maurya, DayasankarSingh "A Survey – Robust Speech Recognition", International Journal of Advance Research in Science and Engineering, Volume 7, 2018.

[13] Atma Prakash Singh, Ravindra Nath, Santosh Kumar, "A Survey: Speech Recognition Approaches and Techniques", IEEE Xplore, 2019.

[14] Neha Jain, Somya Rastogi "SPEECH RECOGNITION SYSTEMS - A COMPREHENSIVE STUDY OF CONCEPTS AND MECHANISM.", Acta Informatica Malaysia, 2019

[15] Dr.R.L.K.Venkateswarlu, Dr. R. Vasantha Kumari, G.Vani JayaSri "Speech Recognition By Using Recurrent Neural Networks.", International Journal of Scientific & Engineering Research Volume 2, 2011.