IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Real Time Speech to Text Using Automatic Speech Recognition

Rakesh Choudhary[1], Dr M.N. Nachappa[2]

[1]Post Graduate Student, Department of Master of Computer Applications

School of CS & IT

JAIN(Deemed-to-be-University)

Bangalore, India

[2]Professor and Head,

School of Computer Science and Information Technology,

JAIN (Deemed to be University)

Bangalore, India

ABSTRACT

Deaf individuals face a lot of inconvenience in communicating due to various type of hearing loss, and their hearing loss may affect their interpersonal communication and ability to recognize facial emotions. Little research has been done on how new technology could assist and solve hearing impairments of deaf people. This paper examines the methods created by numerous researchers to convert speech to text in real time, which would help deaf individuals communicate. We examine these approaches and list their benefits and drawbacks in order to see where this domain might be improved. This publication intends to provide young scholars interested in contributing to this topic with an overview of the numerous options available.

Keywords Automatic Speech Recognition, Neural Network, Real time speech to text.

INTRODUCTION

Hearing loss and deafness affect 5% of the world's population, according to the World Health Organization (2020). There are 432 million adults and 34 million children among them. One of the most serious repercussions of not being able to hear is how a person can become isolated from his or her community. Deaf people are unable to socialise and work in the same ways that hearing people do. The communicational disconnect between hearing and non-hearing prevents people from recognizing the abilities and soft skills, including academic and interpersonal skills, possessed by the Deaf. This emphasises the major challenges that Deaf individuals experience, such as isolation and communication barriers.

Researchers and engineers have devised a method that employs Automatic Speech Recognition (ASR) in which the system accepts audio input from a speaker, processes it by breaking down the various components of speech, and then converts it to text. In order to decode an audio signal and then offer the best transcription possible these models are trained to recognise specific words and speech patterns. The technology generates a wave file from voice, which is then cleaned to remove background noise and adjust the level. This filtered waveform is disassembled and analysed in order. The ASR examines these sequences and uses statistical probability to determine whole words and phrases.

Automatic Speech Recognition(ASR) has become a widely used tool in all aspects of life. Google Cloud Speech-to-text API, Rev.ai Speech-to-text API, IBM Watson Speech-to-text cloud services are few of the available ASR services. This services can be used by anyone for any type of language. For example, on the basis of the Kaldi toolbox they developed and evaluated a stand-alone, low-latency, and relatively high-accuracy real-time system for Japanese Speech to text system [1].

A speech to-text application in which a user enters a text transliterated into Sinhala Unicode characters, and the system converts the text to a Sinhala male or female synthesized voice clip and transfers it to the active call [2]. There are different approaches for implementing ASR the most common one being on noise speech and cleaned speech.

APPLICATIONS

In today's world, any type of voice recognition technology will always be relevant. These systems can be used in real-time research, investigation, and identification in a variety of domains. However, this is a developing subject, and as with any developing field, there will be new technology, new software, and new techniques to implement old projects in order to improve them, and the field of computer speech is no exception.

A few applications for a system that can perform speech to text are listed below:

- Cochlear Implants: Automatic Speech Recognition can enable communication between human and machine which can be used in cochlear implants to help individuals with hearing loss to restore hearing under environment with noise.

- Biometric Lock: Researcher's can make use of ASR to develop biometric voice recognition which would just require speech of the individual rather than physical input or token. This can later be used in various field of market like fraud detection, Financial services etc.

- Log Filing: Automatic Speech Recognition can be used in Log filing where in conversion of speech to text can be done on real time. This can be very useful while doing interrogation, court dealings or making note.

AVAILABLE DATSETS

Speech to text translation datasets come in a wide range and quantity. It's tough to find the correct dataset because many of the datasets available on Google aren't up to the standard required to train a speech recognition model. The following are some examples of useful datasets:

- Google AudioSet: Google AudioSet is a growing ontology that includes 635 audio event classes and 2,084,320 human-labelled 10-second sound snippets culled from YouTube videos. Google gathered information from human labellers in order to investigate the occurrence of various audio classes in 10-second YouTube video clips. Searches based on metadata, context, and content analysis are used to propose segments for labelling. This dataset intends to provide a standard, realistic-scale evaluation job for audio event identification as well as a starting point for a comprehensive vocabulary of sound events by releasing AudioSet.

- VoxCeleb: This dataset is a large-scale speaker recognition dataset. This dataset includes around 100,000 sentences from 1,251 celebrities culled from YouTube videos, representing a wide range of accents, occupations, and ages. This dataset has intriguing use case for distinguishing and identifying whose superstar the voice belongs to.

- 2000 HUB5 English Evaluation Transcripts: This dataset was developed by Linguistic Data Consortium is a collection of transcripts from 40 English telephone calls. Conversational speech over the phone was the focus of this dataset with the task of transcribing it into text. The objective was to investigate interesting new areas in conversational speech recognition, develop improved technology that incorporates those concepts, and assess the performance of new technology. The transcripts for the 40 source speech data files utilised in the evaluation are available in.txt format in this dataset.

- LibriSpeech ASR Corpus: Approximately 1,000 hours of 16kHz read English speech make up LibriSpeech dataset which is derived from audiobook. This dataset is a volunteer initiative that has produced over 8,000 public domain audio books, the majority of which are in English. The majority of the recordings are based on Project Gutenberg texts, which are also in the public domain.

- TED-LIUM Corpus: This dataset was compiled using audio speeches and transcripts from the TED website. This dataset includes 2,351 audio talks, 452 hours of audio, and 2,351 STM-formatted aligned automatic transcripts.

RELATED WORK

The use of neural networks in a variety of computer voice recognition subjects has clearly demonstrated that speech recognition can convert almost anything given the correct resources. Obtaining the appropriate datasets for the audio clip project is arguably the most difficult component of a process like this. The dataset will serve as the project's foundation, as neural networks will use it to train and learn how to recognise a certain voice. The effort of figuring out and putting up networks, as well as strategies to improve them, will be easier to figure out and implement with the help of other publications; datasets are irrelevant. What matters are the networks that have been built, the outcomes that have been achieved, and how it was created.

During the survey we found many different approaches taken for this application.

In automatic speech recognition systems, background noise is the most common cause of performance loss. During a voice call, the microphone catches ambient disturbances such as sound from the environment or appliances, street and traffic noises, in addition to the user's speech. At the receiver end, all of these noise interferences dramatically reduce speech quality and intelligibility. Remote users are frequently asked to relocate away from the noise source to remedy this problem. However, because noise sources are sometimes inevitable, this technique does not always work. Thus, noise suppression is an important feature that users demand during a phone call to remove background disturbances and maintain the intelligibility and quality of the received speech signal.

In [3] they proposed an Integrated Noise Detector and Denoiser (INDDICA) for audio-based calls using a lightweight deep learning method. The model was created using the CR-CED architecture, with a modified training phase to accommodate the speech denoiser. For noised speech, they used the UrbanSound8k dataset, whereas for clean voice, they used the Mozilla Common Voice dataset.

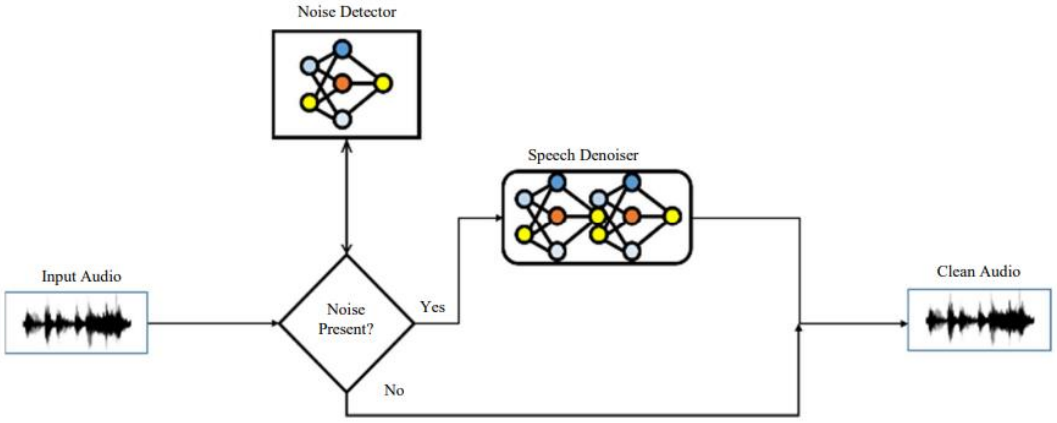

Fig 1. Architecture of the INDDICA model[3]

The INDDICA model consists of the two modules:

- Noise Detector: Noise is almost always present at some point during a call, thus running it continuously is a waste of time. This model is based on a CNN binary classifier, which analyses raw audio and determines whether or not there is noise present. When there is noise, the voice denoiser model is enabled or deactivated.
- Speech Denoiser: During this phase, the need of providing clear voice quality in any situation is stressed, and Short Time Fourier Transform (STFT) with magnitude for speech enhancement is used.

Technique	Matching
Fully matched training (FMT)	$SNR_t = SNR_r$, $n_t(t) = n_r(t)$
Noise matched training (NMT)	$SNR_t \neq SNR_r$, $n_t(t) = n_r(t)$
SNR matched training (SNRMT)	$SNR_t = SNR_r$, $n_t(t) \neq n_r(t)$
Multi-style training (MST)	$SNR_t \neq SNR_r$, $n_t(t) \neq n_r(t)$

Table 1. Noised Speech Training Techniques [4]

Researchers in [4] used a noisy dataset in which the ASR system was trained on the signal-to-noise ratio (SNR). In the Table 2, SNRt demotes SNR in the training mode and SNRr denotes SNR in recognition mode, nt(t) and nr(t) denote background noise respectively. Table 2 shows four distinct types of techniques that can be utilised to work with noisy speech. Three techniques were used in this system

- Fully Matched Training: Speech with the same SNR and noise type is used to train and test ASR systems.
- Noised Mixed Training: ASR systems are taught and tested on speech with varying SNR and noise types.
- Multi-Style Training: ASR systems are taught and evaluated on speech with a variety of SNRs and types of noise.

Additive mixture of signal and noise was calculated based on

$$s(t) = k.x(t) + n(t),\ k = 10^{0.05(SNR_0 - SNR)}$$

SNR0 is the intended signal to noise ratio, and SNR corresponds to initial speech and noise signals, with x(t) representing clear speech and n(t) representing noise speech. The dataset utilised in this study consisted of 10 names of numerals in Russian as speech signals with fourteen different types of noise.

The accuracy percentage of the results was calculated using SNR in decibels of noise in various types of noise.

In [5] the paper aimed to train an automatic speech recognition system based on the dataset which consists of 10 names of numerals in Russian as speech signals with fourteen different types of noise with the modification of using noise spectrum in training and recognition.

Six samples of loud speech were used to test the ASR system. All ten words were used in the test sentences, with 0.3–0.5 second intervals between them. The recognition accuracy is

$$Acc\% = \frac{N - D - S - 1}{N}\ x\ 100\%$$

Where N is the total number of labels in the reference transcriptions, D is the number of deletion mistakes, S is the number of substitution errors, I is the number of insertion errors, as determined by the test findings.

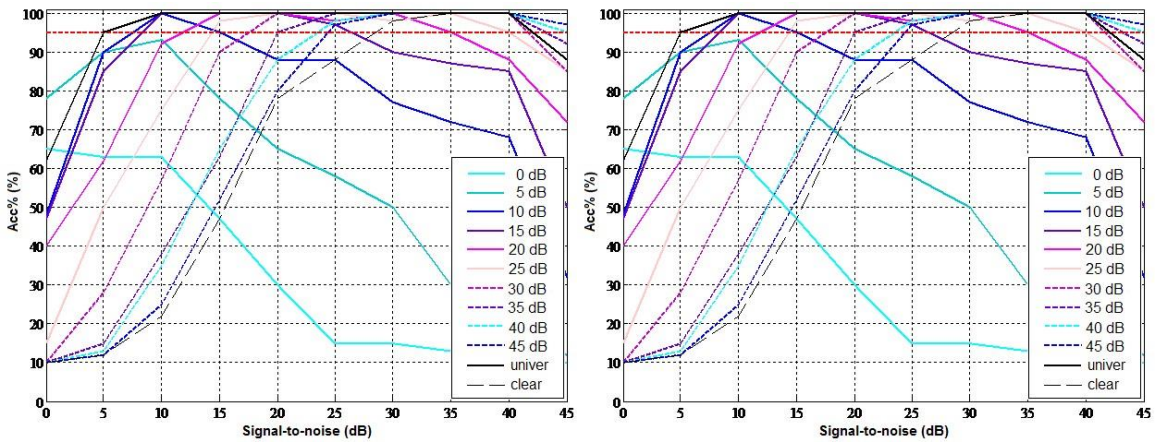

Fig. 2.Acc% for ASR system trained on clean speech [5] Fig. 3. Acc% for ASR system trained on clean speech[5].

Hearing recovered with cochlear implants (CI) for profoundly deafened individuals has difficulties in challenging listening contexts with background noise or reverberation, according to study [6]. For CI users, ASR systems are often trained with clean speech. This paper employed the TIMIT Dataset, which contains 4158 train sentences said by 326 males and 136 females, as well as 1512 test sentences spoken by 112 males and 56 females, to address these challenges.

During the Speech Synthesizer phase, a text to speech system was developed that could generate audio speech files with a sample rate of 16kHz and two default speakers. To create the reverberant data, four room impulse responses (RIR) with reverberation times of 0.3, 0.6, 0.8, and 1.0s were convolved with the anechoic clean train and test phrases from the TIMIT corpus. The ASR system is trained on the TIMIT anechoic silent training set as well as different chunks of the TIMIT reverberant train corpus. Under the evaluation of the ASR system performance in anechoic quiet and four reverberant settings, all TIMIT test material was employed.

During the speech intelligibility testing, each sentence was delivered twice to the CI listener, and the listener was asked to repeat the words he or she understood. To calculate the speech intelligibility score, the number of right words identified by the listener in each condition was divided by the total number of test words utilised in that condition. The order in which the conditions and texts were given to the CI users varied between subjects.

In [7] researches aim to apply neural network model for identifying continuous speech recognition. The Hybrid technique employing Voice Activity Detection (VAD) and Speech Enhancement Algorithm is used to evaluate the performance of objective measurements of noisy input of continuous speech signal (SEA).

During the training phase, the system creates a Hidden Markov Model (HMM) model and trains it. The training processes, from VAD through Speech Enhancement to HMM model construction, are carried out using Matlab scripts on a PC. This process is divided into several steps:

- Speech Acquisition: A single channel microphone is used to capture speech, which is then transformed to 8 bit PCM digital samples at an 8 kHz sampling rate.
- Signal preprocessing: Signal Preprocessing entails taking speech samples as input, blocking them into frames, and returning a distinct pattern for each sample.
- Training method: The system creates the codebook after preprocessing the input speech samples to extract feature vectors. The codebook and the weighted cepstrum matrices for various users and digits are compared. The Baum-Welch algorithm is used to train the closest vector indices. For this, a six-state HMM is utilised.

- Recognition method: The process of comparing the unknown test pattern to each sound reference pattern and determining a measure of similarity between the test pattern and each reference pattern is known as recognition. A maximum probability estimate is used to recognise the digit.

This study [8] tries to build a real time recognition algorithm of English speech based on Hidden Markov Model and Edge Computing. The model examines the speech signal's preprocessing process before proposing the use of a dual-threshold method to detect the voice signal's endpoint. Then, two approaches for extracting distinctive parameters, linear prediction cepstral coefficient and Mel cepstral coefficient, are examined in depth.

With an immature pretrained model and script, this research [9] presents a strategy for improving automatic voice recognition datasets. The dataset used in this model is made up of Korean news videos and scripts. The videos range in length from 20 minutes to 60 minutes. The dataset contains 64 videos. They create pair of an audio and its script by comparing the chunks produced from the pre-trained model with the ground truth script. The audio in each pair has the exact start and end of each phrase, and the script is clear because we utilise a human-written script. This method extracts automatic speech recognition dataset in a precise and effective manner in trials using news videos and scripts. A speech synthesising model can also be trained using the new dataset.

The method in which model is developed as follows:

- Generation of Chunks: We obtain chunks predicted by the pretrained ASR model using audio data. Because the model is still in its infancy, some of the chunks are approximate and not exact. This chunks, have a precise start and end time that can't be retrieved from a script.
- Matching Chunks and Sentence script: In the human-written script, we identify a concatenation of chunks that are near to the sentence for each sentence. The optimal concatenation is found using a predetermined score algorithm that indicates the similarity between two sequences.
- Selecting Pairs: Using a predefined scoring function, we pick pairs with a score more than a threshold and dismiss pairs with a score less than the threshold. Then we create a new ASR dataset with the selected pairs.

CHALLENGES

- When using datasets, language hurdles might be a problem because some of the datasets are labelled in the creator's mother tongue rather than English. As a result, creating our own English labels for the dataset will take time.
- In model training, time is a resource that is heavily utilized. Most individuals strive for better models in the hopes of reducing training time and making the most of the time available to train the models. For training models that use enormous datasets, dealing with time consumption is a huge difficulty.
- Datasets are hard to come by. Particularly good ones that have all of the resources you'll require. Some datasets will have less than two category
- es, while others will have fewer audio samples to train models with. Larger datasets are more difficult to come by, but they also come with the challenge of working with big amounts of data, which can be problematic for those with data usage limits.

CONCLUSION

In the field of computer speech recognition, this experiment highlights the benefits and drawbacks of real-time speech to text conversion. The model that was constructed will produce excellent results and will function as intended for the audio sources that it will be analysing. The project's future lies in improving and advancing automatic speech recognition to produce a stronger and more accurate result for a larger dataset while also reducing the time it takes to train the network. The most difficult problem that all deep learning models face is resolving the time-consuming issue that they have. So far, no progress has been achieved in addressing this problem, but there is a lot of research being done in these areas to overcome these problems and make the usage of ASR more viable and easy for everyone who wants to utilise these systems in their job.

REFERENCES

1. Chee Siang Leow, Tomoaki Hayakawa, Hiromitsu Nishizaki, Norihide Kitaoka (2020). Development of a low-latency and real-time automatic speech recognition system. 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), 925–928.
2. M.S. Amarasekara, K.M.N.S. Bandara, B.V.A.I. Vithana, D.H. De Silva, A. Jayakody (2013). Real-time interactive voice communication - For a mute person in Sinhala (RTIVC). 2013 8th International Conference on Computer Science & Education, 671–675.

3. Vinayak Goyal, Siddhesh Chandrashekhar Gangan, Bhavin Shah, Prasenjit Chakraborty, Srinidhi N (2021). INDDICA: A on-device real time method for background noise removal from speech signal. 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 1–6.

4. Arkadiy Prodeus, Kateryna Kukharicheva (2017). Automatic speech recognition performance for training on noised speech. 2017 2nd International Conference on Advanced Information and Communication Technologies (AICT), 71–74.

5. Arkadiy Prodeus, Kateryna Kukharicheva (2016). Training of automatic speech recognition system on noised speech. 2016 4th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), 221–223.

6. Oldooz Hazrati, Shabnam Ghaffarzadegan, John H.L. Hansen(2015). Leveraging automatic speech recognition in cochlear implants for improved speech intelligibility under reverberation. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5093–5097.

7. C. Ganesh Babu, P. Sampath, S. Hariharan, S. Balakumar, Mohamed Noufal Babu (2017). Performance analysis of hybrid model of robust automatic continuous speech recognition system. 2017 International Conference on Inventive Computing and Informatics (ICICI), 303–306.

8. Juan Wu (2021). English real-time speech recognition based on hidden Markov and edge computing model. 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 376–379.

9. Minsu Kwon, Ho-Jin Choi (2019). Automatic speech recognition dataset augmentation with pre-trained model and script. 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), 1–3.