Speech to text Conversion using Deep Learning Neural Net Methods

BABU PANDIPATI [1], Dr. R.PRAVEEN SAM [2]

[1] Research Scholar, Research and Development Center, Bharathiyar University, Coimbatore, India.
babuphd2015@gmail.com
[2] Professor, Department of CSE, G.Pulla Reddy Engineering College, Kurnool, India.
rpraveensam.cse@gprec.ac.in

Abstract

Internet has grown in the past and has transformed several fields and changed numerous lives. Internet can be a blessing for humanity. The primary field that has been transformed by internet technology is communication. Internet has allowed speedier and simpler communication. In this paper, we intend to explore the various methods for conversion of speech-to-text that can be utilized in an email system that is based on voice. This method is built on the interactive voice response. The goal is to research and evaluate the different methods that are used in STT conversions, and find the most efficient method that is able to be adapted to both conversion processes. In the end, based on a review study, it has been discovered that HMM using deep neural networks is the most effective statistical model , and therefore the best for STT. In the end, a model that uses HMM and ANN techniques to convert STT conversion is suggested.

Keywords: speech translation, speaker, steps for speech recognition and classification

Introduction

In contemporary civilized societies, speaking between humans is one of the most popular methods. Different thoughts that arise in the brain of the speaker can be communicated through the use of speech using phrases, words, or sentences, by applying certain grammar rules.

Speech is the primary method of human communication and is one of the naturalest and effective method of sharing information between humans through speech. When separating speech into voiced, unvoiced , and silence (VAS/S)an fundamental acoustic segmentation in speech, which is vital to speech could be regarded. When a series of sounds are called phonemes , this technique could very closely resemble the sounds of every letter of the alphabet that creates the human speech.

The majority of Information on the internet is available to those who are able to read or comprehend the language with a high level of accuracy. Technologies for language can offer solutions through standard interfaces, so that digital content is accessible to all people and allow for communication among various people who speak different languages. These

technologies play an essential function in multi-lingual societies like India with a total of 1652 dialects or native languages. Speech to Text conversion receives input from a microphone as speech, and later transformed into text which is displayed on the desktop. Speech processing is the research of signals and the different methods utilized for processing them.In this process, a variety of applications like speech coding, speech synthesizer, speech recognition and technology for recognition of speakers that use speech processing are employed.Among the mentioned above, speech recognition may be the most crucial one.

The primary function in speech recognition is transform the acoustic signal received from the microphone or telephone into a set words. To extract and analyze the linguistic information that is transmitted by a speech signal we must use electronic circuits or computers. This procedure is used in a variety of applications, including security devices, home hold devices, cellular phones, ATM machines as well as computers..

Speech to Text is a software that recognizes speech that allows speech recognition as well as translation from spoken languages into text using computational language. It's also referred to as computer-based speech recognition. Certain tools, applications and devices are able to transcribe audio streams in real time to display text and respond to it.

Literature Survey

1.Yee-Ling Lu, Manwai and Wan-Chi siu explain text-to-phoneme conversion making use of recurrent neural networks that have been developed using the real time Recurrent Learning (RTRL) algorithm[3].

2.Penagarikano, M.; Bordel, G This article describes a method to perform the conversion of text to speech and an investigational test that is conducted using a task-based Spanish corpus. It also reports the results of an analysis are also reported..

3.Sultana, S.; Akhand, M. A H; Das, P.K.; Hafizur Rahman, M.M. Explore Speech-to-Text(STT) conversion with SAPI to convert speech into text for Bangla language. Although the results are promising for STT related research the researchers identified several factors that could improve performance and may provide greater accuracy . They also ensure that the concept of this study can help other languages that perform Speech-to-Text conversion and similar tasks.[3].

4.Moulines, E., in his paper "Text-to-speech algorithmic systems are based on FFT synthesizing," present FFT synthesis techniques to create the French text-to-speech software system that is built in diaphone concatenation. FFT synthesis techniques are capable of producing top quality prosodic modifications of natural speech. Different approaches are designed to minimize distortions that result from diaphone concatenation..

5.Decadt, Jacques,Daelemans, Walter and Wambacq provides a way to improve the ability to read the output of text in a vast vocabulary Continuous speech recognition, when out-of-vocabulary terms occur. The principle is to replace undefined terms in transcriptions using

the result of a phoneme recognition process which is then processed using a phonemeto-grapheme converter. This method makes use of machine learning techniques.

Methodology

A.Speech Analysis

Speech analysis uses a an appropriate frame size for analysis and extraction of speech signals. The speech analysis process is carried out using two methods.

1. Segment Analysis

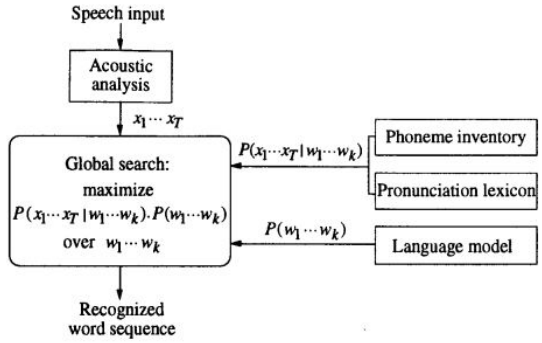


Fig 1 : Mechanism of state-of-the-art speech to Text

Speech is a dazzling method of human computer interaction It is "hands free"; it is a simple way to learn. With modern technology and flow diagrams, algorithms and techniques we can interpret speech signals quickly and detect the text that is being spoken by the speaker. For this method, we're planning to create an on-line speech-to-text engine[4]. The system captures speech in real-time through a microphone . It then process the spoken speech to recognize the spoken text.

The concept aims at developing an application that can provide the user with two functionalities that include: First, sending an Email using making the voice input of the user to text and sending it. Second, to convert to convert the Text at the receiver's end to voice output , and then narration of the message to the user. The STT conversion model STT conversion is executed using HMM along with Neural Network as it gives the best precision for STT.

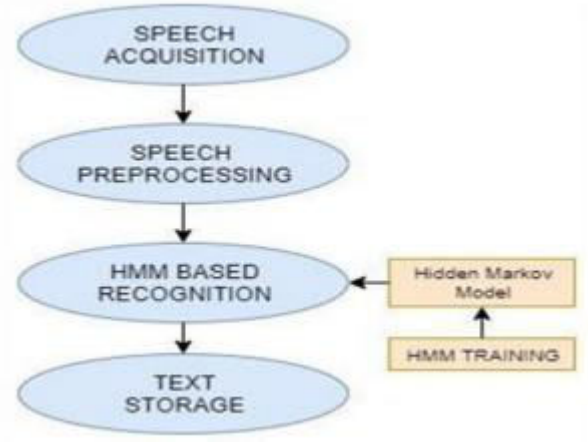


Fig 2 : STT using HMM

Figure 2 represents basic steps for STT conversion using HMM. It consists of 4 steps:

1. Speech Acquisition: Receiving the Speech Data as an input. Speech is recorded by the microphone and then stored in memory.

2. Speech preprocessing: The noise and interference in speech signals are eliminated and the result is speech that is converted. Through the detection of voice activity the pauses are removed out of voice output.

3. HMM based recognition: The HMM is created by the system to represent each word that is in the vocabulary, and the models are trained during the training. From speech processing to the HMM model's construction the steps of learning are carried out and then the generated HMMs are loaded..

4. Text Storage: The text that is matched is stored as output, and then compiled to produce the output text. The model used for TTS convert is further implemented using HMM training for efficient results..

Convolution Neural Network

Convolution Neural Network is the most well-known variant of deep learning that is used for speech recognition. According to Aggarwal and Passricha [1414] Mitra V. and others. 12] Convolution neural network is a supervised learning algorithm. CNN's have shown impressive performance thanks to its innovative features, such as sharing of weight convolution, convolution and pooling. CNN's are often employed to learn high-level functions and the pooling process can be used to reduce dimensionality.

CNN is a type of deep-learning algorithms. It is a feed forward, deep learning model that is supervised. It is comprised of a convolution layers, pooling, and a fully connected layers. It is composed of sparse interaction with parameter sharing, as well as Equivariant representation as the basic concepts [1414. Convolution layer and the pooling layer form the foundation of CNN. Locality, weight sharing , and pooling are the three main components of CNN.

All of these factors play a important factors in the improving the quality for speech recognition. The amount of weight distributed across networks is decreased due to localization. Weight sharing can be used to limit overfitting and increases the robustness that the algorithm can provide. The characteristics extracted from the weight sharing and locality models combined in pooling.

Results and Discussion:

Automatic speech recognition (ASR) is the process of converting speech into text using a machine or computer. ASR software recognizes the words and sub-words contained within the signal. and studies audio characteristics of speech which includes examination and analysis of the speech's physical properties including frequency, intensity and duration.

Muti Channel Speech Separation With Soft Time-Frequency Masking

The primary goal of monaural speech separation technique is to determine the sources of a linearly mixed single microphone signal, based on the signal that is observed. The solution addresses the issue of segregating simultaneous speech using the use of a spatial filtering stage as well as the subsequent time-frequency masking phase. The focus of research shifted to matching recognition, however, it was found out that the overlapping of speech is more challenging to manage. This led to the use of beamforming techniques based on classical techniques and blind source separation techniques like time-frequency masking, and reduction of mutual information (MMI)

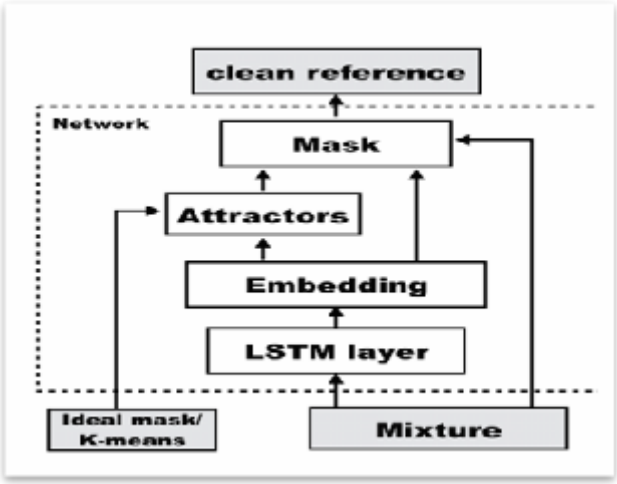
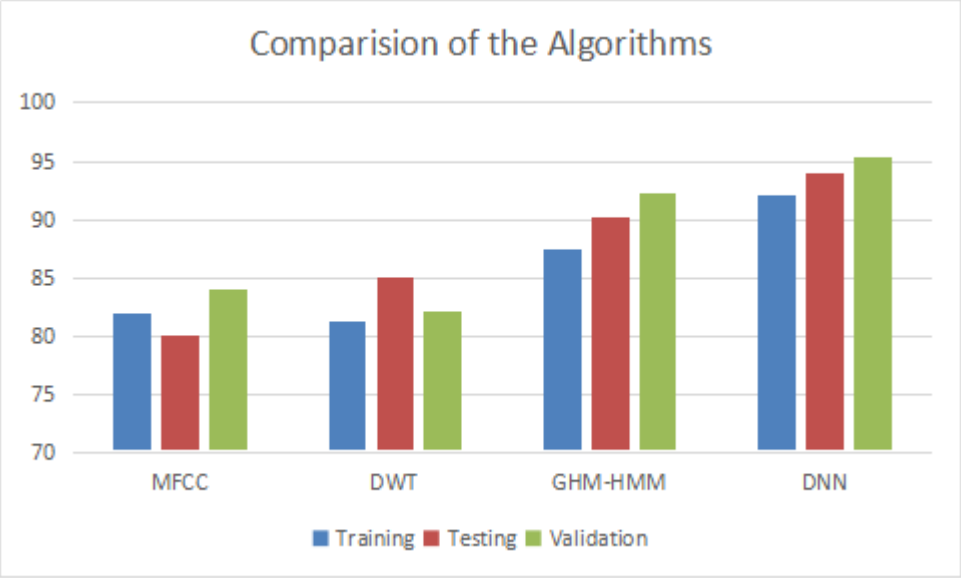


Figure 9: LSTM and Neural implementation of the speech methods

This model for speech recognition as well as the Speech to Text conversion Behavior can be determined based on the speech by implementing an in-depth analysis of the data. A lot of speech data can be easily analysed and Deep architectures encompass a range of variations of the same basic approach. Each has had success in specific areas. Mutitalker Speech Separation with Utterance Level Permutation Training of Deep Neural Network. The Speech recognition was trained, tested and evaluated using MFCC, DWT, GMM-HMM and DNN. Overall, the results are satisfactory and the findings are DNN has been able to perform with good precision when compared with other models..

Algorithms	Training Accuracy	Testing Accuracy	Validation Accuracy
MFCC	82	80	84
DWT	81.3	85	82.1
GHM-HMM	87.4	90.2	92.3
Deep Neural Network	92	94	95.3



7 CONCLUSION

The overall effectiveness of the algorithm has been improved through the use of DNN models in the context for speech recognition. The speech mix that is heard in audio can be separated using the DNN that performs deep clustering using attracted points in the embedding space that bring together the embedded that belong to a particular speaker. Each attract within this space can be used to build the time frequency mask for each speaker of the mix. Based on the mix of audio signal after the application of DNN Speech is isolated from the speaker mix and the signal of every speech is recorded. The model's performance is reliable in the sense of testing and validating the speaker's audio in terms of registration and identifying. The algorithm that is currently used been outnumbered when compared to the deeper neural network. The recognition of speech is possible with more enhancements from the DNN.