DISCOURSE RECOGNITION WITH DEEP LEARNING

A V ALLIN GEO¹, Dr.K.P.Kaliyamurthie²
¹Assistant Professor ²Dean, Dept. of CSE,
BIST,BIHER, BharathUniversity ,
Chennai, Tamil Nadu, India
¹seemeallin@gmail.com, ²kpkaliyamurthie@gmail.com

Abstract: This paper discusses speech recognition and how it can be made utilizing machine learning. This paper also discusses how errors can be figured out to execute the program decreasing the possibility of errors. Additionally examined in the paper is the means by which the program translates the information and predicts the speech. The paper additionally talks about the utilization of Nyquist theorem and Fourier Transformation to make speech recognition conceivable.

Keywords: Deep Learning, Speech Recognition, Neural Networks, algorithms, Representation, Nyquist theorem, Fourier Transformation, Classification, Error

1. Introduction

Speech recognition (SR) is the intermediate sub-field of computational etymology that creates approaches and advances that empowers the recognition and interpretation of talked dialect into text by computers. It is otherwise called "automatic speech recognition" (ASR), "computer speech recognition", or only "speech to text" (STT). It fuses learning and research in the semantics, software engineering, and electrical designing fields[1,2].

Some SR frameworks utilize "training" (additionally called "enlistment") where an individual speaker peruses text or detached vocabulary into the framework. The framework examines the individual's particular voice and uses it to adjust the recognition of that individual's speech, bringing about expanded precision[3].

The term speech recognition or speaker distinguishing proof alludes to recognizing the speaker, as opposed to what they are stating. Perceiving the speaker can streamline the assignment of deciphering speech in frameworks that have been prepared on a particular individual's voice or it can be utilized to validate or check the personality of a speaker as a major aspect of security procedure[4].

2. Literature survey

1. As indicated by MehryarMohri, Pedro Moreno, and Eugene Weinstein, they investigate computerized revelation of topically-rational portions in speech or text arrangements. They give two new discriminative topic segmentation algorithms which utilize another measure of text comparability in light of word co-event[5,6]. Both algorithms work by discovering extrema in the likeness motion over the text, with the last algorithm utilizing a minimal support-vector based portrayal of a window of text or speech perceptions in word comparability space to battle clamor presented by speech recognition errors and off-topic content. In analyses over speech and text news streams, they demonstrate that these algorithms outflank past techniques. They watch that topic segmentation of speech recognizer yield is a more troublesome issue than that of text streams; in any case, they show that by utilizing a grid of contending hypotheses as opposed to only the one best speculation as a contribution to the segmentation algorithm, the execution of the algorithm can be progressed.[7-8]

2. As per Baidu Silicon Valley AI Lab, They demonstrate that an end-to-end deep learning methodology can be utilized to perceive either English or Mandarin Chinese speech–two immeasurably extraordinary dialects. Since it replaces whole pipelines of hand-designed segments with neural networks, end-to-end learning permits us to deal with a various assortment of speech including noisy situations, accents and distinctive dialects. The key to their approach is their utilization of HPC techniques, empowering tests that already took weeks to now keep running in days. This permits us to emphasize all the more rapidly to distinguish unrivaled structures and algorithms. Therefore, in a few cases, their framework is aggressive with the interpretation of human specialists when benchmarked on standard datasets. At last, utilizing a method called Batch Dispatch with GPUs in the data center, they demonstrate that their framework can be reasonably conveyed in an online setting, conveying
According to Alex Graves and Navdeep Jaitly, this paper displays a speech recognition framework that specifically interprets audio data, without requiring a transitional phonetic representation. The framework depends on a mix of the deep bidirectional LSTM recurrent neural network design and the Connectionist Temporal Classification objective function. An alteration to the objective function is acquainted that prepares the network with limit the desire of an arbitrary transcription loss function. This permits an immediate enhancement of the word error rate, even without a lexicon or language model. The framework accomplishes a word error rate of 27.3% on the Wall Street Journal corpus with no earlier linguistic data, 21.9% with just a lexicon of permitted words, and 8.2% with a trigram language model. Consolidating the network with a baseline system additionally decreases the error rate to 6.7%.

4. Antonio Lagarda, Jorge Civerajcivera, Alfons Juan ajuan, and Francisco Casacuberta, this paper portrays ongoing research work by the Pattern Recognition and Human Language Technology (PRHLT) group (UPV PASCAL2 hub) in two important technology transfer projects: i3media and erudito.com. From one viewpoint, i3media (2007-2010) is a 35 MC "tractor" technology project extend inside the Spanish Programa CENIT-Ingenio 2010, gone through a consortium of 12 fundamental endeavors of the media sector, which likewise include 19 research groups, including PRHLT. i3media concentrates on the creation and automated management of intelligent audiovisual content, in order to encourage both, content personalisation and interaction with clients (i3media.barcelonamedia.org). Our cooperation in i3media is focused on interactive machine translation, to exchange and adjust our experience on this project to i3media-particular needs. Then again, erudito.com (2010-2012) is a 1.4 MC trial configuration extend, bolstered by the Spanish Ministry of Industry, Tourism and Trade under the Avanza I+D program, went for building up a tool to embody, disperse and wisely utilize advanced substance, for example, that appeared on topical TV channels. In this project, PRHLT adds to the advancement of interactive closed captioning (speech transcription) and machine interpretation tools.

5. Felix Weninger, In this article, presents CURRENTNT, an open-source parallel implementation of deep repetitive neural networks (RNNs) supporting representation handling units (GPUs) through NVIDIA's Computed Unified Device Architecture (CUDA). CURRENTNT underpins uni and bi-directional RNNs with Long Short-Term Memory (LSTM) memory cells which conquer the vanishing gradient problem. As far as anyone is concerned, CURRENTNT is the main openly accessible parallel implementation of deep LSTM-RNNs. Benchmarks are given on an uproarious speech recognition errand from the 2013 second CHiME Speech Separation and Recognition Challenge, where LSTM-RNNs have been appeared to convey best execution. In the outcome, double-digit speedups in bidirectional LSTM training are accomplished concerning a reference single-threaded CPU implementation.

3. Present system

To change over speech to on-screen content or a PC summon, a PC needs to experience a few complex strides. When you talk, you make vibrations noticeable all around. The analog-to-digital converter (ADC) makes an interpretation of this analog wave into digital information that the PC can get it. To do this, it tests, or digitizes, the sound by taking exact estimations of the wave at continuous interims. The framework filters the digitized sound to evacuate undesirable commotion, and in some cases to separate it into various groups of (recurrence is the wavelength of the sound waves, heard by people as contrasts in pitch). It additionally standardizes the sound, or alters it to a consistent volume level. It might likewise must be transiently adjusted. Individuals don't generally talk at a similar speed, so the sound must be changed in accordance with match the speed of the layout sound examples as of now put away in the framework's memory.

Next the flag is separated into little fragments as short as a couple of hundredths of a moment, or even thousandths on account of plosive consonant sounds - consonant stops delivered by blocking wind stream in the vocal tract - like "p" or "t." The program then matches these sections to known phonemes in the proper dialect. A phoneme is the littlest component of a dialect - a portrayal of the sounds we make and set up together to frame important expressions. There are around 40 phonemes in the English dialect (diverse etymologists have distinctive suppositions on the correct number), while different dialects have increasingly or less phonemes.

4. Proposed system

My work in deep learning will incorporate better comprehension of the words perceived by the program to decrease error event in the deciphered sound. The research will likewise incorporate the better comprehension of the sound by the program in loud and echo circumstances. My work will also include the change to text from sound directly without the transformation to phonetic speech in between.
additionally propose the consideration of linguistic grammar use in the program to help the machine make probabilities of combination of words that make up a sentence[17].

Speech recognition is attacking our lives. It’s incorporated with our telephones, our diversion comforts and our keen watches. It’s notwithstanding robotizing our homes. For just $50, you can get an Amazon Echo Dot—a enchantment box that permits you to request pizza, get a climate report or even purchase waste bags—just by speaking out loud[18].

The big problem is that speech varies in speed. One person might say a word very quickly while other person will say the word slowly. Both both sound files should be recognized.

The first step in speech recognition is the need to feed sound waves into a computer.

Sound waves are one-dimensional. At every moment in time, they have a single value based on the height of the wave. To turn sound wave into numbers, the height of the wave is recorded at equally-spaced points. This process is called sampling.[19]

For example I will take the sound “hello” sound wave 16,000 times per second.

![Image of sound waves](image)

By sampling the sound wave, one is just creating a rough estimation of the original sound wave because the program is only making occasional readings. This way one is losing some sound in between every reading.

To overcome this problem Nyquist theorem is used. Using this theorem one can construct original sound waves from spaced out samples.

According to Wikipedia, Nyquist theorem is that in the field of digital signal processing, the sampling theorem is a fundamental bridge between continuous-time signals (often called “analog signals”) and discrete-time signals (often called “digital signals”). It establishes a sufficient condition for a sample rate that permits a discrete sequence of samples to capture all the information from a continuous-time signal of finite bandwidth.[10]

I have an array of numbers with each number representing the sound wave’s amplitude at 1/16,000th of a second interval.

Instead of feeding these numbers directly into a neural network I will preprocess the audio data to make the problem easier.

The sampled audio is grouped into 20 millisecond chunks. These numbers are then plotted on a simple line graph[20].

this recording is only 1/50th of a second long. To make this data easier for a neural network to process, the complex sound wave is broken up into component parts. This is done using Fourier Transformation.

As per Wikipedia, The Fourier change breaks down a component of time (a flag) into the frequencies that influence it to up, in a route like how a melodic harmony can be communicated as the frequencies (or pitches) of its constituent notes[21]. The Fourier change of an element of time itself is a complex-esteemed capacity of recurrence, whose outright esteem speaks to the measure of that recurrence exhibit in the first capacity, and whose intricate contention is the stage counterbalance of the major sinusoid in that recurrence. The Fourier change is known as the recurrence area portrayal of the first flag. The term Fourier change alludes to both the recurrence space portrayal and the scientific operation that partners the recurrence area portrayal to a component of time. The Fourier change is not restricted to elements of time, but rather with a specific end goal to have a bound together dialect, the space of the first capacity is regularly alluded to as the time area. For some elements of pragmatic intrigue, one can characterize an operation that turns around this: the backwards Fourier change, additionally called Fourier combination, of a recurrence area portrayal joins the commitments of all the diverse frequencies to recoup the first capacity of time.[11]

It breaks apart complex sound waves into simple sound waves. The energy of every simple sound wave is added. The result is a score of how important each frequency is in the 20millisecond audio clip. This is easier to represent in a chart than an array of numbers. If this process is repeated for every 20 millisecond chunk of audio I end up with a spectrogram.[22]
A neural network can find patterns in a spectrogram easily than from raw sound waves. This data representation is then fed to the neural network. The input to the neural network will be 20 millisecond audio chunks. For each chunk the network will try to figure out what the letter for the corresponding sound is. I will be using a recurrent neural network. A recurrent neural network has a memory that influences future predictions.[12]

After running the entire audio clip through the neural network we end up with a mapping of each audio chunk to the letter most likely spoken during that chunk. for example the mapping of the word hello spoken will look something like this[23].

In this case the neural network is predicting that the word fed could be “HHHEE_LL_LLLOOO” but it also predicts the possibility of the word to be “HHHUU_LL_LLLOOO” or even “AAAUU_LL_LLLOOO”. [13] to clean the predicted word particularly steps must be followed. first any repeated character is taken as a single character. hence HHHEE_LL_LLLOOO becomes HE_LL_LO
HHHUU_LL_LLLOOO becomes HU_LL_LO
AAAUU_LL_LLLOOO becomes AU_LL_LO

the next step is to remove the blanks. Then HE_LL_LO becomes HELLO
HU_LL_LO becomes HULLO
AU_LL_LO becomes AULLO

Of the possible predictions Hello is the most probable choice. This choice is done using large databases of text from which the neural network predicts possible words. These databases are created based on the region where the neural network is being used. In America the database would be created based on American way of speaking and otherwise.

5. Conclusion

By using Nyquist theorem and Fourier Transformation I have successfully created a program that can recognise words and convert that to text. The program converts the speech based on the dictionary that has been stored in the database that is accessed by the computer to predict the words. A recurrent neural network is used to influence future predictions.

References

[3] End-to-End Speech Recognition in English and Mandarin by Baidu Silicon Valley AI Lab
[4] Towards End-to-End Speech Recognition with Recurrent Neural Networks by Alex Graves and NaadeepJaitly
[5] Interactive Pattern Recognition and Human Language Technology for Digital Audiovisual Content Processing by Antonio Lagardaalagarda, Jorge Civerajcivera, Alfons Juan ajuan, and Francisco Casacuberta
[6] Introducing CURRENNT: The Munich Open-Source CUDA RecurReNT Neural Network Toolkit by Felix Weninger