Journal of Advanced Research in Intelligence Systems and Robotics
Volume 5, Print Issue 1 - 2023, Pg. No. 1-5
Peer Reviewed Journal

Research Article

A New Approach to Speech-to-Text Conversion Using Deep Learning

Gaurav Kumar Jha

Department Of Computer Science & Engineering Chandigarh University, Gharuan, India.

I N F O

E-mail Id:
19BCS1469@gmail.com
Orcid Id:
https://orcid.org/0009-0003-9327-8641
How to cite this article:
Jha GK. A New Approach to Speech-to-Text Conversion Using Deep Learning. *J Adv Res Intel Sys Robot* 2022; 5(1). 1-5.

Date of Submission: 2023-03-05
Date of Acceptance: 2023-04-25

A B S T R A C T

In the current business environment, progress depends on effective communication. Not just on a corporate level, but also on a personal level, it is crucial to convey information to the appropriate person and in the appropriate way. The globe and communication methods are both going towards digitization. In today's technologically advanced society, communication by phone conversations, emails, text messages, other means has become essential. Many applications that function as a mediator and aid in efficiently transmitting messages in the form of text or audio signals over miles of networks have emerged in order to fulfil the objective of effective communication between two parties without obstacles.

Keywords: Deep Learning, Speech-To-Text, Conversations, Emails, Text Messages

Introduction

Cell phones have developed into a necessary form of communication for contemporary culture over the past several years. We can quickly send texts and phone calls from one place to another. It is well recognised that verbal communication is the best method for conveying accurate information and preventing misquotations. On phone calls, verbal communication can readily bridge the gap over a long distance. Speech recognition technology has recently made a ground-breaking contribution to SMS technology, allowing voice conversations to be transformed to text messages. TTS, STT, translation are utilised in many programmes designed to help the disabled. They can also be applied in other ways, as shown by the following examples: Siri is an intelligent automated assistant that is installed on electronic devices to improve user engagement with those devices and their local and/or remote services. Uses Text-To-Speech (TTS) and voice recognition technology from Nuance Communications. The different types of speech, voice recognition, speech to text conversion, text to speech conversion, speech translation will all be discussed in this essay. Pre-emphasis of signals, feature extraction, signal recognition are the steps we'll take to recognise speech, which will aid in our training and testing processes. While there are several models that may be used for this, the Hidden Markov Model and Dynamic Time Warp are the two that are most frequently employed. The Hidden Markov Model is a stochastic model that is used to connect different phases of transition with one another. . Similar to how we employ DTW and HMM models for audio to text conversion, we also use a variety of Neural Network models since they are effective at speaker adaption, isolated word recognition, phoneme categorization. As of late 2014, end-to-end ASR is also being tested to produce comparable outcomes. Speech synthesis is effective at producing synthetic human speech from tokenized words. In this study, various machine translation techniques and engines will also be evaluated and contrasted. The elements of voice creation listed below are what applications refer to while using various speech-related functionalities.[15]

Phonation (Producing Sound)

- Fluency
- Intonation
- Pitch variance

Literature Review

We have examined the current systems for speech recognition, speech to text conversion, speech to text conversion, machine learning techniques in this review study.

Speech Recognition

The ability of a machine or programme to recognise words and phrases in spoken language and translate them into machine-readable format is known as speech recognition. The following criteria can be used to classify speech recognition systems:

Speaker: Every speaker has a distinctive voice. Therefore, the models are either made for a particular speaker or a different speaker.

Vocal Sound

Some models are capable of distinguishing between discrete utterances and utterances that are separated by pauses.

Speech Recognition

Vocabulary: The system's complexity, effectiveness, precision are significantly influenced by the size of the vocabulary.

1. **Basic Speech Recognition Model:** As seen in figure 1[1], all speech recognition systems adhere to a set of standard procedures.
2. **Pre-Processing:** In order to be processed subsequently, the analogue voice signal is converted into digital signals. To spectrally flatten the signals, this digital signal is passed via first order filters. As a result, the energy of the signal at higher frequencies is increased.
3. **Feature Extraction:** This procedure identifies a set of utterance parameters that are related to speech signals. The acoustic waveform is processed to compute these parameters, also referred to as features. The primary goal is to compute a series of feature vectors (relevant data) that provides a condensed representation of the input signal as it is. The following is a discussion of frequently used feature extraction techniques:

Linear Predictive Coding (LPC): The fundamental concept is that a voice sample can be roughly represented as a linear combination of prior speech samples. A frame of N samples is blocked from the digitised signal. Then, to reduce signal discontinuities, each sample frame is windowed. Then, each framed window is automatically associated. The LPC analysis, which transforms each frame of autocorrelations into an LPC parameter set, is the final stage.

Mel-Frequency Cestrum Coefficient (MFCC): This method, which makes advantage of the human auditory perception system, is particularly effective. MFCC processes the incoming signal in the following ways: Framing: If there is interference, the speech waveform is chopped to remove it;Windowing: reduces the signal's discontinuities; Simple Fourier Transform: changes the frequency domain of each frame from time domain; Mel Filter Bank Algorithm: To simulate human hearing, the signal is displayed against the Mel spectrum.[2]

Dynamic Time Warping: Based on dynamic programming, this approach compares two time series that could have different speeds to determine how comparable they are. It seeks to iteratively align two feature vector sequences, one from each series, until an ideal match (based on the right metrics) between them is discovered.

Acoustic Models: The core component of Automated Speech Recognition (ASR) systems, Acoustic Models provide a relationship between acoustic data and phonetics. Training creates a link between the fundamental speech components and acoustic observations.

4. **Language Models:** This model predicts the likelihood that a word will appear following a word sequence. It includes the structural restrictions built into the language to produce occurrence probability. The language model distinguishes between words and phrases that sound alike.
5. **Pattern Classification:** Pattern classification is the process of assessing similarities between an unknown pattern and an established, reliable reference pattern. At the time of testing, the system has been trained, patterns are categorised to identify speech. For pattern matching, other methods include[1]

Template-Based Approach: This strategy utilises a database of speech patterns that serve as a reference for dictionary words. By comparing the spoken word to the reference template, speech can be recognised.[14]

- **Knowledge-based Approach:** In this method, a system is trained to automatically construct a set of production rules from samples using a collection of features from speech.
- **Approach Based on Neural Networks:** This method can handle increasingly difficult recognition tasks. The fundamental concept is to gather and integrate knowledge from many knowledge sources with the current issue.[8]
- **Statistical Based Approach:** In this strategy, training techniques are used to statistically model differences in speech.

6. **Techniques for Converting Speech to Text:** The process of translating spoken words into written texts is known as speech to text conversion. It is the same as speech

recognition, however the latter term is used to refer to the broader speech comprehension process. The ideas and procedures for speech recognition are followed by STT, although each phase uses a different combination of approaches. some prevalent the techniques of conversion are covered below.

7. **Hidden Markov Model (HMM):** A speech signal can be considered as a piecewise stationary signal or a short-time stationary signal, hence the Hidden Markov Model (HMM) is a statistical model used in speech detection. HMM models are helpful for mobile users' real-time speech to text conversion.[10]

The Following Factors Affect It

- **Recognition Accuracy:** Accurate recognition involves comparing an unknown test pattern to each reference pattern in a sound class and calculating how similar each test pattern and each reference pattern are to one another. In a perfect world, it would be 100% accurate and independent of the speaker, making it the most crucial component of any recognition system
- **Recognition Speed:** Users will get impatient and the system would lose its relevance if it takes a long time to recognise the speech. The following processes are applied to the signals[3]
- **Pre-Processing**: The incoming voice signals are divided into speech frames, each of which provides a distinct sample while lowering noise
- **HMM Training:** Using one or more test patterns that match to speech sounds from the same class, training entails constructing a pattern indicative of the attributes of a class
- **HMM Recognition:** This method involves calculating a measure of similarity (distance) between the unknown test pattern and each reference pattern from the sound class.For recognition, maximum likelihood is employed

Based on Artificial Neural Network Classier

Cuckoo Search Enhancement: ASR with Cuckoo Search Optimisation is used to improve recognition, communication, to filter out undesired noise. ASR was designed to improve the interface between humans and machines. A three-step procedure is used for the same[4]

The most crucial step in speech recognition is the pre-processing of speech signals, which is done to eliminate unnecessary waveforms from the signal. To eliminate background noise, the signals are passed via high-pass filters.

The voice signal is used to extract two different types of acoustic characteristics. They are Linear Predictive Coding Coefficients (LPCC) and Mel Frequency Cepstrum Coefficients (MFCC).

Artificial neural networks are utilised in this instance to classify data. A three-layered classifier, the neural network has n input nodes, l hidden nodes, k output nodes. A two-layered Feed Forward Backpropagation Neural Network (FFBNN) of three units-two input units, three Hidden units, one output unit-is used to implement ANN in CSO (Cuckoo Search Optimisation). The input layer in this case comprises of two inputs from which two features-MFCC and LPCC features-have been retrieved. These properties are provided as input to train networks, which then generate outputs that correlate to the input.

Text to Speech Conversion

Text-To-Speech is a process that turns input text into digital audio, which is then spoken after being processed, comprehended, assessed. The image depicts every step in the text to speech conversion process, although the key stages of TTS systems are shown in.[5]

Text processing involves analysing, normalising (which handles acronyms and abbreviations and matches the text), transcribing the input text into phonetic or linguistic representation.

Speech Synthesis: A few of the approaches used for speech synthesis are[5]

1. **Articulator Synthesis:** This technique generates speech using mechanical and acoustic models. It generates understandable synthetic speech, but because it lacks natural sound, it is not frequently utilised.
2. **Formant Synthesis:** In this system, a parametric representation of each unique speech segment is saved. In formant synthesis, parallel and cascade are the two fundamental structures. However, some mixture of these two structures is employed for better performance. Band-pass resonators that are coupled in sequence make up a cascade formant synthesiser. The input of the following formant resonator is affected by the output of the previous one. Only formant frequencies are required for the cascade structure's control information. All formants get the excitation signal at once, the outputs are then added.[5]
3. **Concatenative Synthesis:** With this method, sound is created by concatenating small sound samples, or units. A database created from the recording of previous sequences is utilised in voice synthesis to create user-specific sound sequences. Concatenative synthesis units are:[5] phone: a single sound unit; Diphone is defined as the signal from a phone's midpoint or point of least change to a similar position in the next phone; triphone is a part of the signal taking in a sequence from a phone's midpoint all the way through the next one to the midpoint of a third.

Language Translation

There are many different languages spoken in India. According to the 2001 Census, 122 languages were spoken by more than 10,000 people, 30 languages were spoken by more than a million native speakers. For this reason, it is crucial to have programmes and procedures that can translate text between languages while maintaining the integrity of the message. The study of translation from one language to another using machine translation systems is known as Machine Translation (MT), a branch of artificial intelligence and natural language processing.[6] Decoding the meaning of the source text and reencoding it in the target language are two ways to summarise the human translation process. Below is a discussion of a few machine translation models:

Rule-Based Machine Translation (RBMT): Rules-Based Machine Translation (RBMT) Morphological, syntactic, semantic analysis of the source and destination languages is used to construct the translation. Such a system is composed of a number of rules: Grammar rules are essentially an analysis of grammatical structures in languages (syntax, semantic, morphology, part-of-speech tagging, orthographic features); a bilingual or multilingual lexicon dictionary for looking up words during translation while the software programme allows effective and efficient interaction of components; and software programmes to understand and process those rules. Three categories of rule-based models exist:

- **Straight Forward:** It uses a dictionary
- **Transfer:** Each SL input text is processed through an intermediate representation using lexicons and structural analysis
- **Interlingual:** The translation process transforms the source language into an intermediate language that is unrelated to any of the target languages

Statistical Machine Translation (SMT): The employment of machine learning techniques characterises Statistical Machine Translation (SMT). SMT is a data-driven methodology that use parallel aligned corpora and views translation as a problem of mathematical inference. Every sentence in the target language is, in this case, a probabilistic translation from the original language. The accuracy of translation increases with probability, vice versa. The foundation of SMT architecture consists of two models: a translation model and a language model. The translation model calculates the conditional probability of the target language output given source language input.

The decoder model maximises the two aforementioned probabilities to provide the best translation feasible.

Example-Based Machine Translation (EBMT): Example-based machine translation (EBMT) (iii) It is predicated on the concept of analogy. This method uses a corpus that includes writings that have already been translated as its source material. Sentences from this corpus are chosen that have similar sub-sentential elements to the sentence that has to be translated. Then, the sub-sentential elements of the original sentence are translated into the target language using the analogous sentences, these phrases are then combined to create a complete translation. Three steps of the analogy translation are used: matching, adaptation, recombination.

Matching: The SL input text is broken up, then examples from the database are searched for that closely resembles the input SL fragment string, the pertinent fragments are chosen. The required TL fragments are retrieved along with their associated TL pieces.

Adaption: Find the TL component of the pertinent match that corresponds to a particular portion in SL and align them if the match is exact; otherwise, the fragments are recombined to create the TL output.

Recombination: Combining pertinent TL fragments to create grammatically correct target text.

Hybrid Machine Translation: The development of new automated translation programmes and the growth of approaches over the past ten years have brought to light the drawbacks of using a single methodology to solve all translation issues.

Hybrid machine translation (MT) is a type of machine translation that combines different machine translation techniques into a single machine translation system. The RMBT and SMT procedures are combined, it takes advantage of their respective benefits. Thus, the creation of lexicon and syntax makes use of statistical data. Many different applications make use of or operate independently of translation engines like Google Translate, IBM Watson Developer Cloud, Microsoft Translators to help successfully translate languages for improved understanding.

We have read about the many STT and TTS approaches that are employed, as well as how they are applied. We can make the following conclusions after carefully examining the various voice, speech recognition, speech to text, text to speech, speech translation systems: Because of their computational viability, HMM functions as a better speech signal to text converter in STT under the 2 examined, despite its disadvantages. Similar to how formant synthesis using parallel and cascade synthesis under TTS systems researched functions as the optimum converter. Due to its incorporation of benefits from both rule-based and statistical machine translation systems, hybrid machine translation is widely used.

The development of grammatically correct, syntactically connected text is ensured, while SMT-related issues like

text smoothness, quick learning, data gathering are also taken care of.

References

1. Suman K. Saksamudre,P.P. Shrishrimal, R.R. Deshmukh, A Review on Different Approaches for Speech Recognition System, International Journal of Computer Applications (0975 8887) Volume 115 No. 22, April 2015.
2. Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj
3. P. Shrishrimal, A Comparative Study of Feature Extraction Techniques for Speech Recognition System, International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 12, December2014.
4. Ms. Anuja Jadhav, Prof. Arvind Patil, Real-Time Speech to Text Converter for Mobile Users, National Conference on Innovative Paradigms in Engineering Technology (NCIPET-2012) Proceedings published by International Journal of Computer Applications (IJCA)
5. Sunanda Mendiratta, Dr. Neelam Turk, Dr. Dipali Bansal, Speech Recognition by Cuckoo Search Optimization based Artificial Neural Network Classifier, 2015 International Conference on Soft Computing Techniques and Implementations- (ICSCTI) Department of ECE, FET, MRIU, Faridabad, India, Oct 8-10, 2015.
6. Suhas R. Mache, Manasi R. Baheti, C. Namrata Mahender, Review on Text- To-Speech Synthesizer, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, August 2015.

Aditi Kalyani, Priti S. Sajja, A Review of Machine Translation Systems in India and different Translation Evaluation Methodologies, Interna- tional Journal of Computer Applications (0975-8887) Volume 121 No.23, July 2015