Empowering Communication: The Evolution and Potential of Speech Recognition System

Vaibhav Tiwari
MSc Data Analytics
Dublin Business School
Dublin, Ireland
10621051@mydbs.ie

ABSTRACT

Speech is the primary and most essential part of human communication. Speech recognition systems are an important part of human-computer interaction. The primary aim of the speech recognition system is to achieve high accuracy and low latency in a real-time environment. Applications of speech recognition technology are voice assistants, transcripts generation systems, IoT devices, and more. This paper provides a detailed overview of the major technological perspective and fundamental progress made in each stage of building the speech recognition system. In this system, we will extract the features of the audio and make necessary predictions of emotions out of it. A comparative study of each of these phases is discussed in detail at each stage. In this article, we will discuss a few successfully applied applications of NLP to design an emotion recognition speech; we seek to find the most efficient deep neural network architecture and machine learning algorithm with the help of the RAVDESS dataset.

KEYWORDS

Speech Recognition, RAVDESS, Deep Learning, Neural Network, RAVDESS

1 Introduction

This section contains a brief overview of the project's background, motivation, problem statement, challenges, goals, and report structure.

1.1 Background

Humans' speech consists of multiple emotions and each emotion has different features associated with it. Each of these features has totally different statistical measures which help in determining the type of emotion associated with it. The primary aim of this project is to develop an accurate and efficient speech recognition speech that is evaluated based on the accuracy i.e., its ability to recognize words and emotions of the speakers. It is processed under machine learning models, and deep neural networks in order to successfully extract the relevant features from the speech and train model in its relevance.

1.2 Motivation

This project allowed me to study the topics of my interest in machine learning, which was always a fascination for me. The widespread and constant use of such systems as Alexa, Siri, etc… made me even for interested in this domain of work. Machine Learning is so widely used that it inspired me to adopt it as the foundation for my idea.

1.3 Dataset

RAVDESS dataset which we used in building this emotion recognition system comprises 1440 speech files and 1012 song files. This dataset contains the recording of 24 professional actors (12 males & 12 females), vocalizing 2 lexically-matched statements in a neutral North American accent. Speech files in RAVDESS contain calm, happy, sad, angry, fearful, surprised, and disgusted expressions; on the other hand, song files include calm, happy, sad, angry, angry, and fearful emotions. A group of 247 individuals people provided a rating to these files; untrained research participants from North America. Further, 72 more participants provided test-retest data.

2 Literature Review

The last few years have been crucial in the field of artificial intelligence and a lot of development has taken place in this field. A lot of literature material is available on the internet which helps us in the current stage of this field. Many pieces of research have been done by experts to gain more insights into sentiment analysis through the recognition of the speech. So, let's deep dive into the world of speech recognition.

A typical speech recognition speech consists of the following procedure:

1. Data Collection
2. Data Pre-processing
3. Feature Extraction
4. Classification
5. ML Modeling

We will quickly discuss the various structures and learnings employed in our models. These algorithms and structures can be used in various fields of work, including text analysis, NLP problems, time-series analysis, and many more.

2.1 Supervised Learning

This learning technique is one of the most common approaches used for speech recognition tasks. In the supervised learning case, algorithms learn to recognize the appropriate emotion as per the labeled dataset, in which the input variable (audio signal feature) and output (emotion label) corresponding to the inputs are already provided. Supervised learning results in the case of speech recognition have achieved impressive results especially due to the presence of CNNs and RNNs.

2.1 Convolutional Neural Networks (CNN)

Neural Networks have shown significant improvement in the past in speech recognition tasks. CNN has been applied widely in the field of NLP because of its better accuracy and its ability to handle complex data.

A typical CNN architecture comprises of an input layer, a convolution layer, a pooling layer, a fully connected layer, and an output layer that are deeply connected with each other. 2D-CNNs construct 2D feature maps corresponding to each input. In 2-dimensional CNNs, one axis represents the frequency domain and the other axis represents the time domain. In contrast, 1-dimensional CNNs accept vectors directly rather than matrices as in the case of 2-dimensional CNNs.

In building a successful CNN, weights, and biases play an important role, as they develop a neural network that propels data flow through the network, via a method called forward propagation. Once forward propagation is complete, the direction of flow reverses, and the identification of nodes and connection point is done, via a method called reverse propagation.

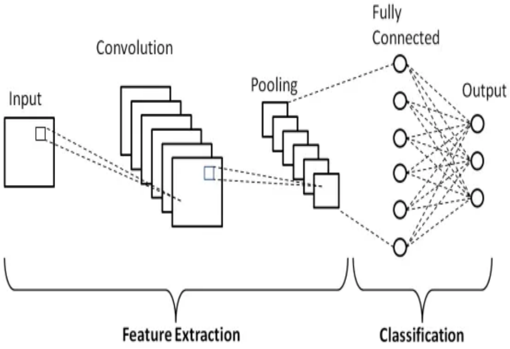

Figure 1: Architecture of CNN

2.2 Weight Sharing in CNNs

Weights are used in connection management between the connected units in a neural network. While building a neural network, these weights are increased or decreased in order to obtain the best-performing model. Weights are the deciding components in the neural network that determines the influence input data has over generated output.

Weight sharing is a technique for reducing the number of weights to be trained in CNNs i.e., it simply uses similar weights for the nodes that have symmetry to other nodes.

Below is a typical application of weight sharing in CNNs, as they work by passing a filter over the signal input. For example, a 7x7 signal and a 3x3 filter with a stride size of 1; and return a feature map as an output of size 5x5.

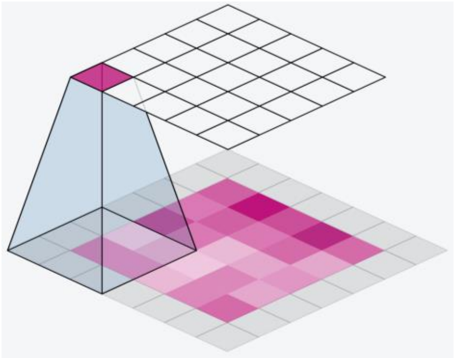

Figure 2: Weight Sharing in CNN

The above method reduces the number of weights resulting in the reduced computational cost for model training and making the model insensitive to the feature search in the input signal.

In the case of the traditional CNN, each filter scans the entire signal to detect its features. As a result, it might cause OVERFITTING, more resource utilization, and longer training times. Therefore, to resolve this issue weight sharing technique is used.

There are two types of weight-sharing methods to adopt while working with CNNs are:

1. Full Weight Sharing (FWS)
2. Limited Weight Sharing (LWS)

In the case of the speech recognition system, the LWS method of working is used. In this method, only certain weights are shared across multiple layers, but all weights are not shared as in the traditional CNN. Furthermore, one of the ways to implement LWS is to group filters into multiple clusters and share the weights among them.

2.3 Transfer Learning Models

Transfer learning is a machine-learning technique in which a pre-trained model is used as an initial point to accomplish new tasks. Rather than building a new model from scratch, transfer learning alleviates this load by customizing the pre-existing models in two different ways: Feature Extraction and Fine-Tuning. And utilizing the knowledge and experience gained from training the model on a large dataset, and applying it with necessary modifications to the required scenario.

In the present scenario, DeepSpectrum and PANNs are the most influential used for solving audio tasks. For example, ImageNet is one of those pre-trained models which is trained on a large dataset, which after necessary modifications can be re-trained for the new task. After doing this, a more generalized and more accurate model can be obtained with fewer training resources required for computation.

In our project, our RAVDESS dataset has multiple audio files, and these files in this dataset follow a certain naming convention that tells us about the emotion associated with them. Thus while performing feature extraction of the audio file and finding the patterns associated with those values, we can generate machine learning models and convolutional neural networks which will help in predicting the correct emotion for the model, and as the file name convention tells the emotion associated with it; our problem is a case of supervised learning.

3 Research Questions

The primary research questions to be addressed in this study are:
1. How can speech recognition systems achieve high accuracy and low latency in a real-time environment?
2. What features are extracted from audio to predict emotions in a speech recognition system?
3. What is the most efficient deep neural network architecture and machine learning algorithm for emotion recognition in a speech recognition system, as evaluated with the RAVDESS dataset?

4 Related Work

Constant research has been in the process to enhance the process of human-computer interaction, but no significant achievement has been made in this field still there exist lots of problems left to be overcome which are left to be resolved. A key factor in determining the success of these researches includes the feature extraction process. Current research in this domain of work is emphasized dealing with the deficiencies naturally present in speech.

The notable research work in speech recognition systems includes:

- Gan, Liong, Yau, Huang, Tan (2019, Volume 74. Pages 129-139, ISSN 0923-5965). OFF-ApexNet on micro-expression recognition system, Signal Processing: Image Communication, in this paper a novel feature extraction approach, OFF-ApexNet combines handcrafted features and a fully data-driven architecture is used.
- Another research suggests us to automatically model the channels and spatial cues using spectrograms, utilizing MLP to extract channel info and dilated CNN to extract contextual information.
- Nogueiras, Moreno, Bonafonte, and Mariño et al. (2001); Research Center TALP, Universitat Politècnica de Catalunya, SPAIN, the content of this publication has an HMM-based approach to emotion recognition has also been represented in which an accuracy higher than 80% is obtained despite having a low level of feature used.
- Lee and Tashev's (2015) publication on High-level Feature Representation using Recurrent Neural Networks for Speech Emotion Recognition shows how a bidirectional LSTM model is also adopted to extract a high-level representation of emotional states regarding temporal dynamics. Uncertainty of emotional labels i.e., all frames in a single utterance getting mapped to the same emotional label was overcome by assuming that each frame's label is regarded as a sequence of random variables.
- Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlıcek, Qian, Schwarz, Y., Stemmer, Vesely et al. (2011, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, US), the content of the publication informs that KALDI is a free, open-source toolkit for speech recognition systems. It uses finite-state transducers, core libraries of C++(the programming language used) supports acoustic modeling with subspace Gaussian mixture models(SGMM) and standard Gaussian mixture models along with linear and affine transforms.
- Swamy, and Ramakrishnan (August 2013, CSEIJ, Vol. 3, No. 4) published a paper 'An Efficient Speech Recognition System' et al. in which they discussed how speech recognition systems have been developed by using different techniques like Mel Frequency Cepstrum Coefficient(MFCC), Vector Quantization(VQ), and Hidden Markov Model(HMM). In this publication, MFCC is used to extract the features from the input signal, while the HMM is used on vector quantization(VQ) to identify the word in the speech by evaluating the maximum likelihood value for each of those words.
- Gaikwad, Santosh & Bharti, Gawali & Yannawar, Pravin et al. (2010); A Review on Speech Recognition Technique. International Journal of Computer Applications. Pages 1462-1976., in this publication multiple speech signal feature extraction techniques are used like PCA, ICA, LDA, Ceptral analysis, Wavelets, Spectral subtraction, RASTA filtering, and more.

REFERENCES

[1] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño (2001). Speech Emotion Recognition Using Hidden Markov Models.

[2] Santosh K. Gaiwad, Bharti W. Gawali, & Pravin Yannawar (2010). A Review on Speech Recognition Technique, Volume 10 - No 3.

[3] Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlıcek, Qian, Schwarz, Y., Stemmer, Vesely. The Kaldi Speech Recognition Toolkit.

[4] Lee, and Tashev (2015). High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition. [3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[5] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..

[6] Swamy, and Ramakrishnan. An Efficient Speech Recognition System. Computer Science & Engineering: An International Journal (CSEIJ), Vol. 3, No. 4, August 2013.

[7] Y.S. Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, Lit-Ken Tan (2019). OFF-ApexNet on micro-expression recognition system, Signal Processing: Image Communication, Volume 74. Pages 129-139, ISSN 0923-5965.

[8] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391

[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7916477/

[10] https://theaisummer.com/speech-recognition/

[11] https://iopscience.iop.org/article/10.1088/1742-6596/1973/1/012166/pdf

[12] https://analyticsindiamag.com/comprehensive-guide-to-different-pooling-layers-in-deep-learning/

[13]https://medium.com/sfu-cspmp/an-introduction-to-convolutional-neural-network-cnn-207cdb53db97

[14] https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

[15] https://tspace.library.utoronto.ca/handle/1807/24487