**Dublin Business School/MSc Data Analytics**

**Empowering Communication: The Evolution and Potential of Speech Recognition System**

Vaibhav Tiwari (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 fficient deep neural network architecture and machine learning algorithm with the help of the RAVDESS dataset.

**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.

**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.

**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

# 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.

**Architecture of CNN**



**Weight Sharing in CNN**



# 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?

# Reference / Bibliography

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](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://analyticsindiamag.com/comprehensive-guide-to-different-pooling-layers-in-deep-learning/
14. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio