Deep Learning Project Report

- Tanmay Agrawal M22MA011

- Aviral Triphati M22MA012

- Pothula Akash M22MA007

INTRODUCTION:

This project is about recognizing a person's emotion from his/her voice, our program is capable of **recognizing both male as well as female** voices, and can recognize **8 different emotional states**.

MOTIVATION:

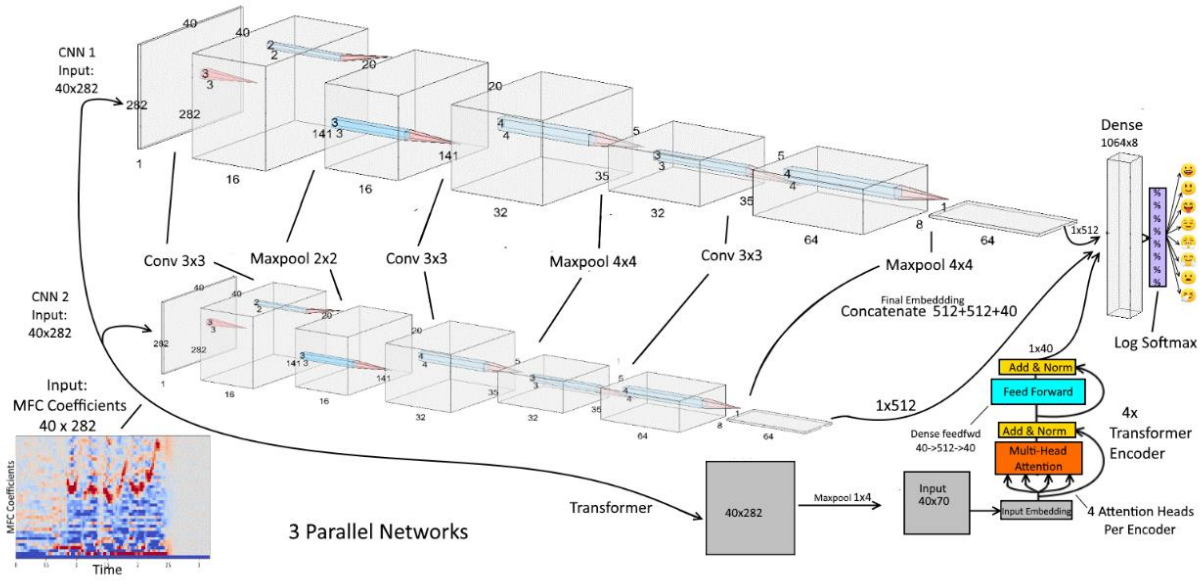
The global voice assistant market size was valued at **USD 2.48 billion** in 2020 and is expected to **grow at a CAGR of 32.7%**. This shows the increased amount of Human-Machine interaction through voice, but still many of the voice assistants lack "emotional recognition", we hope that our project can get AI a step closer in understanding our emotions.

DATASET:

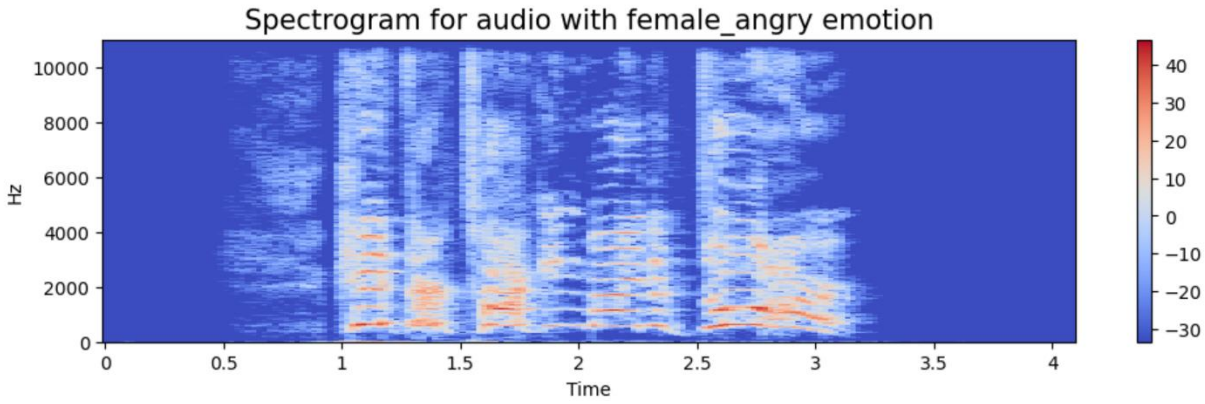
Ryerson Audio-Visual Database of Emotional Speech and Song.The audio files are in the WAV format and have a sampling rate of 48 kHz and a bit depth of 16 bits. The dataset also includes metadata such as the gender and age of the actors, the emotion expressed, and the intensity of the emotion.

ALGORITHMS / ARCHITECTURES:

Two parallel convolutional neural networks (CNN) in parallel with a Transformer encoder network.



Feature used for training is MFCC(Mel Frequency Cepstral Coefficients) . Mel Spectrograms are used in calculating MFCCs, which are a higher-level representation of pitch transition

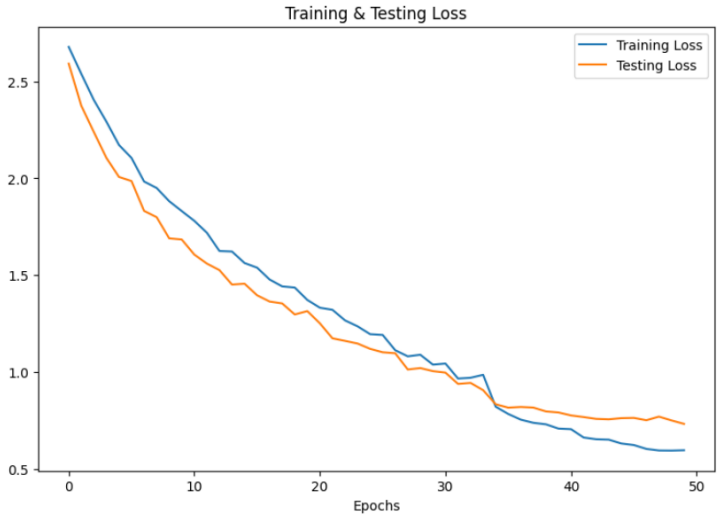


Since , our model was highly parameterised it was prone to overfitting. To overcome this we used augmentation.

Augmentation techniques used:

- Random Noise
- Stretch
- Gaussian Noise

Results:



Testing accuracy : 51%

Analysis:

- The unconventional architecture of **CNN + Transformer [5]**, **[6]** didn't work out very well, mainly because **transformers require a large amount of data** to show good results, but **RAVDESS** has only **7356 files [2], [3]**

- **Facebook's pre-trained** (Hubert-large-superb-er) is performing the worst, because it was **trained** on audio in the range of **16kHz**, whereas the audio range of **RAVDESS dataset is 24kHz.**

Room for improvement:

➔ Instead of using normal transformers we can use **Vision Transformers** (ViT) for small data[4]

➔ We can get a better accuracy using a model called as **XLSR-Wav2Vec2,** because it is trained on a wide audio **range of 20Hz-20kHz**, this range **includes the range of RAVDESS dataset.**

➔ Incorporating different datasets to make our model more robust such as SAVEE and CREMA-D . We could be also use multi modal dataset like IEMOCAP.

KEY DEEP LEARNING (DL) USED:

The major DL part used in the project was:

1- **Data pre-processing:** from sound waves to arrays (tensors)
2- **Hyper-parameter Tuning:** selecting the best values for model training
3- **Training the model:** getting the model learn and train on RAVDESS dataset
4- **Using a pre-trained model:** making use of facebook's model
5- **Experimenting with a new and unconventional architecture (CNN + Transformer):** As transformers are not preferred for this task, this was a new experiment on our part

REFERENCES:

[1] Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network (hindawi.com)
[2] RAVDESS Dataset | Machine Learning Datasets (activeloop.ai)
[3] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English | PLOS ONE
[4] Optimizing Deeper Transformers on Small Datasets (aclanthology.org)
[5] https://github.com/IliaZenkov/transformer-cnn-emotion-recognition
[6] https://www.mdpi.com/2079-9292/11/23/3935