Gender and Emotion Recognition using Voice

(Project Report)

A.G. J. Thenura	AS2019554 (Group Leader)
G. L. H. B. Gaweshika	AS2019358
S. V. V. Vibudena	AS2019560
N. Hirusha Wanigasingha	AS2019566
W. A. P. T. Weragoda	AS2019574
I.U.Wickramasinghe	AS2019576

A proposal submitted in fulfillment of requirements for the CSC 364 1.5 course unit

Degree of Bachelor of Science in Computer Science

Department of the Computer Science

University of Sri Jayewardenepura

Content

1. Introduction 3
2. Significance of the Research 5
3. Literature Review 6
4. Research Questions 8
   4.1 Aim 8
   4.2 Objectives 8
5. Methodology 10
   5.a Model 01 (Emotion Detection) 11
   5.a.1 Data acquisition 11
      5.a.2 Voice extraction and Cleanup 13
      5.a.3 Preprocessing 14
      5.a.4 Feature Engineering 17
      5.a.5 Modeling 18
      5.a.6 Evaluation 20
   5.b. Model 02 (Gender Detection) 24
      5.b.1 Data acquisition 24
      5.b.2 Voice extraction and Cleanup 24
      5.b.3 Preprocessing 24
      5.b.4 Feature Engineering 25
      5.b.5 Modeling 26
      5.b.6 Evaluation 28
6. Results and discussion 29
   6.1 Results and Discussions (Model 1) 29
   6.2 Results and Discussions (Model 2) 33
7. Conclusion 35
8. Acknowledgement 37

Appendix 38

References 39

1. Introduction

Human speech expresses emotional meaning through semantics and specific attributes of the voice. Due to the ability of the human brain to recognize extra information in voices, humans can quickly determine the gender and emotions of people with the help of their body language, facial expressions, and voices. Building a computer program to identify gender and emotions can be used in various technologies to create great user experiences.

Due to the complexity that arises with technological development, the involvement of an intelligent machine was a massive contribution to the fields such as Financial, Education, Health, and Legal. In the Banking and Financial sector, fraud detection is made easier with the usage of systems because banks could use an emotion recognition system to analyze the customer's voice during the call and detect any signs of stress, anxiety, or nervousness that could be associated with fraudulent activity. Remote education applications are another usage of emotion-recognizing systems that could enable the system to provide personalized feedback or recommendations about the student by analyzing their speech during presentations or any other engaging activity. In the health sector, mental health is considered a susceptible area that is hard to predict and clarify. These healthcare system applications combined with natural language processing systems help to identify when a patient is experiencing symptoms of depression or anxiety. This can enable the system to provide timely interventions or alerts to the patient's care team.

Further, law enforcement agencies can use emotion detection systems to investigate a case of criminal activity involving a suspect making phone calls to victims. The law enforcement agency could use an emotion recognition system to analyze the suspect's voice during the calls and detect any signs of stress, anxiety, anger, or nervousness that could be associated with the criminal activity. Also, these systems can be efficiently utilized in the crime detection process to analyze the suspect's

emotions based on their speech and validate the accuracy of their information.

Furthermore, apart from emotion analysis, Natural language processing (NLP) is now used to teach computers to discern gender in human speech. This also significantly contributes to many sections, such as customer service, Education, and healthcare. Gender detection systems can be used in customer service and call center optimization to identify the gender of the customer and provide them with more customized feedback. In social media advertising, gender detection systems can be effectively utilized to send more customized messages and product recommendations to the target audience. Language learning apps could provide feedback on gender-specific grammar rules in language learning apps to provide more specific feedback on language use and pronunciation.

The model is constructed using recorded samples of male and female voices with six emotions such as happy, sad, neutral, angry, excited, and frustrated. This study outlines the construction of computer software to simulate voice and speech acoustic analysis for identifying the gender and emotions of a real-time input. This task can be challenging due to variations in voice quality, accent, and speaking style. Many efficient approaches have been adopted throughout the development of the aforementioned computer system to overcome those challenges.

2. Significance of the Research

The ability to detect emotions and gender from a person's voice can enhance the personalization and customization of digital assistants, virtual reality experiences, and other interactive technologies. Although many applications have been developed in the area of emotion and gender detection using natural language, some significance highlights this developed system when compared with similar alternatives.

In this study, two datasets have been used in the process of training and testing the emotion recognition model. They are namely; Ryerson Audio-Visual Database of emotional speech and song(RAVDEES), and Toronto Emotional speech sets(TESS). In combination, these data sets have more than 4,000 recorded samples of male and female voices in six different emotions such as happy, sad, neutral, angry, excited, and frustrated. The use of a more extensive set of sample voices for training the model has created a positive impact on the performance of the model.

Another significance of this study can be identified as its high accuracy. The machine learning model is trained and validated using more than 4,000 samples of data. On completing the required number of epochs, the system outputs a predicted accuracy as high as “84%”. This is a considerably high accuracy level compared to similar research models developed to perform the same set of tasks.

Moreover, this system can identify gender and emotions using only voice commands. No other interactive body part is used when predicting the speaker’s emotion. Similarly, no biometric identification is used when predicting the speaker’s gender through the analyzed model.

3. Literature Review

One common approach is to use machine learning algorithms to analyze the acoustic properties of speech, such as pitch, loudness, and rhythm. These algorithms can be trained on a dataset of labeled speech samples to learn the patterns that distinguish different emotions and genders.[1] Several studies have been conducted to extract the spectral and prosodic features which would result in the correct determination of emotions. The emotion classification using human speech utterance can be based on calculated bytes, while the gender identification can be made using calculated pitch. [2].

Some studies have focused on using traditional machine learning methods like Decision trees and Support Vector Machines in emotion recognition. Many researchers have applied SVMs to the problem of emotion classification in an attempt to increase accuracy. In [3], the authors have developed a model for recognizing emotions using Support vector machines in neural networks, giving more than 60% accuracy. In [4], the authors have researched the Hidden Markov Model and Support Vector Machine feature regarding speech emotion recognition. More recent techniques such as deep neural networks can also be utilized in emotion and gender recognition. Many studies have found that deep learning approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can achieve high accuracy rates for gender recognition.

Another approach is to extract Mel-Frequency Cepstral Coefficients (MFCCs) or Linear Predictive Coding (LPC) coefficients and use these as input features to a machine learning model. Some studies have also used other features such as pitch, formants, and prosody and have found that they can provide valuable information for emotion recognition. In the work developed in [5], authors proposed a speech emotion recognition model for Thai subjects. Mel spectrogram and mel-frequency cepstrum coefficient (MFCC) is used for feature extraction, and emotions are classified by combining a one-dimensional convolutional neural network

(Conv1D) and a two-dimensional convolutional neural network (Conv2D). This study utilizes a dataset from the VISTEC-depa AI Research Institute of Thailand, which includes 21,562 sound samples. The results show that Conv2D with MFCC achieves the highest accuracy rate of 80.59%.

In recent years, much work has been carried out to recognize emotional information in speech automatically, but the lack of significant improvement in recognition accuracy is still a major problem in the field of speech emotion recognition. It is important to note that the accuracy of the emotion and gender detection using voice can vary depending on the dataset. The performance of the model can also be affected by the quality of the audio data, the presence of noise, and the context of the speech. Also, it should be noted that most of the studies in this field have focused on English, and the results may vary for other languages and accents.

Overall the literature suggests that emotion and gender recognition using voice and NLP is a challenging task, but promising results have been achieved using various techniques. However, more research is needed to address issues such as variability in voice quality, accent, and speaking style and ethical and privacy concerns.

4. Research Questions

4.1 Aim

This research project aims to develop a system that can accurately recognize gender and emotional aspects, regardless of the semantic content of the speech. The ultimate goal is to apply this system in various fields such as speech recognition, natural language processing, and affective computing to improve human-like interaction.

4.2 Objectives

1. To identify the type of emotion behind the speech, to better interpret those, and to develop such an application to successfully implement the idea.

2. To Develop machine learning algorithms that can accurately classify the gender of a speaker by analyzing the differences in voice characteristics between male and female voices.

3. Improve the accuracy and reliability of emotion recognition in speech, through the use of advanced machine learning, deep learning algorithms, and a variety of large data sets.

4. Identify the impact of various demographic, linguistic, and environmental factors on gender and emotion recognition performance.

5. To critically review the existing systems, concepts, and tools. Similar systems have been developed to recognize emotion and gender separately.

6. To develop and compare multiple machine learning and deep learning models to identify the best approach for speech emotion detection.

7. Optimize the model for real-time use by reducing its computational complexity and latency.

5. Methodology

This project work is based on Natural Language Processing. There is a collection of fundamental tasks that appear frequently across various NLP projects. Among them, in this project voice classification has been applied. Below figure1 shows the NLP pipeline.

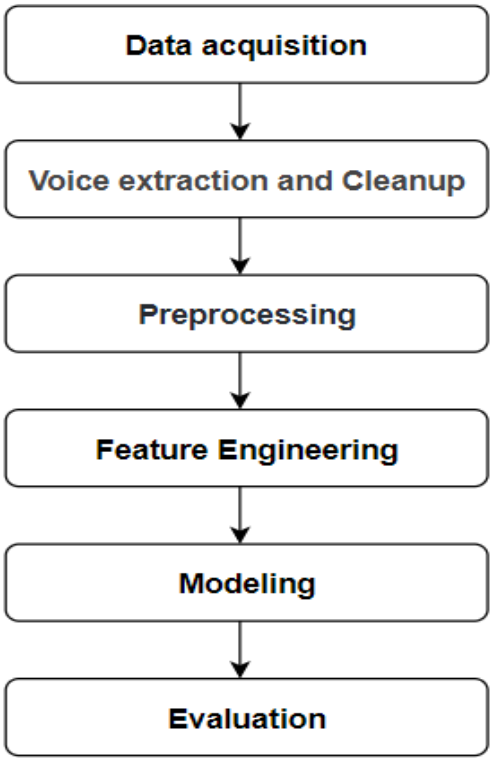

Figure 1 - The Methodology Pipeline

5.a Model 01 (Emotion Detection)

5.a.1 Data acquisition

In this research project on gender and emotion recognition using voice, two secondary datasets have been utilized. The used datasets are as follows.

1. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
   Link : https://zenodo.org/record/1188976#.X4sE0tDXKUl

   The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was recorded with 24 professional actors which includes 12 females and 12 males. The dataset is vocalized using two lexically-matched statements in a neutral North American accent. Each expression is produced at two levels of emotional intensity namely; normal and strong, with an additional neutral expression.

   - Total number of files: 1440 files = 24 actors x 60 trials per actor
   - Captured emotions: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised
   - File naming convention: Each of the 1440 files have a unique filename.

   The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav) The 3rd identifier represents the emotion 01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).

2. Toronto Emotional Speech Set (TESS)
   Link: https://tspace.library.utoronto.ca/handle/1807/24487

   TESS modeled by using a set of 200 target words, which were spoken in the carrier phrase “Say the word” followed by a single linguistic expression. The dataset is recorded by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of the seven emotions described below. Both actresses speak English as their first language. Audiometric testing indicated that both actresses have thresholds within the normal range.

   - Total Number of files: 2800 files = 2 actors x 200 phrases x 7 emotions
   - Captured Emotions: Neutral, Happiness, Sadness, Anger, Fear, Disgust, Surprise

5.a.2 Voice extraction and Cleanup

The following data was extracted from both RAVDESS and TESS

1. Emotion Representation

   RAVDESS: The filename contains a fixed placed int (3rd numerical identifier) that represents an emotion. (e.g. 03-01-**03**-01-02-01-12.wav represents ‘happy’)

   TESS: The filename contains a string representation of an emotion, e.g 'happy'.

2. Sample Rate

   The number of audio samples per second is extracted as the sample rate.

   Sample rate of RAVDESS dataset: 48kHz

   Sample rate of TESS dataset: 24.414kHz.

Once the data is extracted from the above datasets. The data is cleaned through numerous approaches to silence, mic breathing, and other miscellaneous sounds from the audio file. The data cleaning was performed in the following steps in the model.

1. The silence at the beginning and the end of the audio data is trimmed to remove unwanted data, by using Librosa.
2. Noise reduction is being performed by the noisereduce and pydub libraries.

5.a.3 Preprocessing

Preprocessing involves various operations such as Normalization, Padding, Feature Extraction, Data Labeling, Data splitting, and One-Hot encoding.

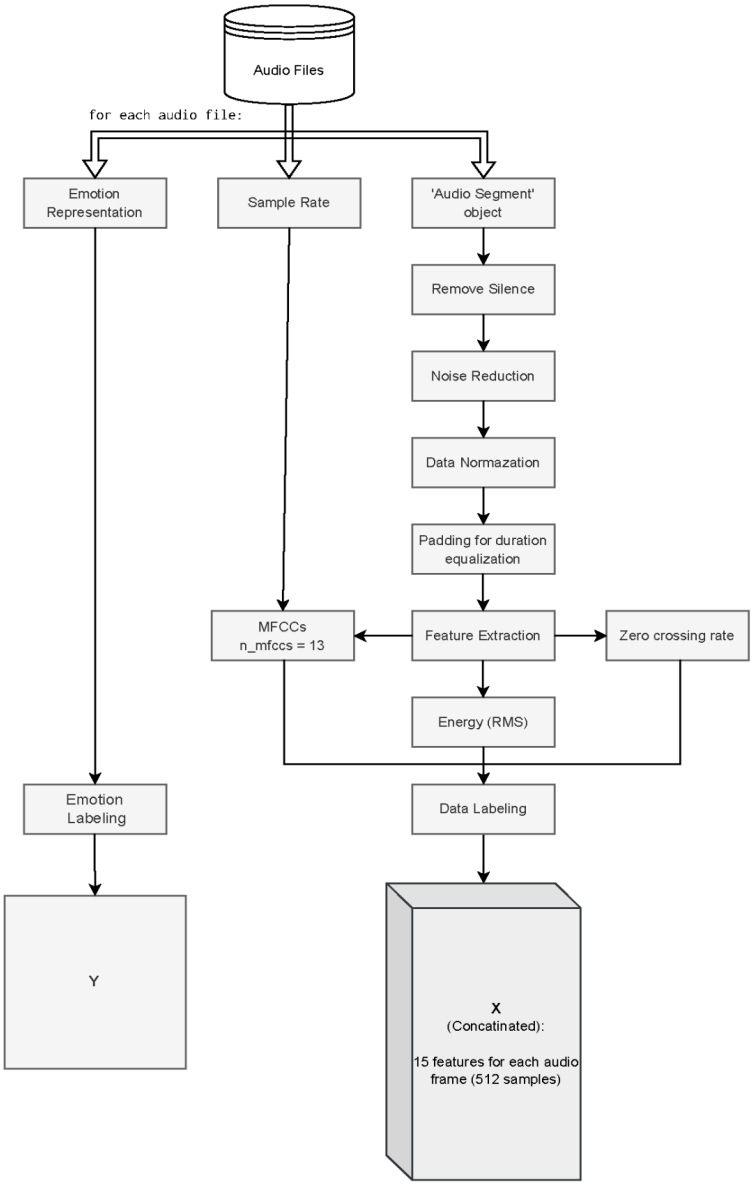

Figure 2 - The preprocessing model

1. Normalization

   Normalization was mainly used to ensure that audio is at an optimal level for playback or for further processing. In this model, an 'AudioSegment' object is normalized to + 5.0 dBFS, by the effects module of the pydub library.

2. Padding

   Padding is done for duration equalization since the model requires that all input audio samples have the same length.

3. Feature Extraction

   This phase converts all the audio files into 15 features that can be input into a model. These 15 features were extracted using the following 3 techniques.

   a. Root Mean Square (RMS)
   b. Zero Crossed Rate (ZCR)
   c. Mel-Frequency Cepstral Coefficients (MFCCs)

4. Data Labeling

   In this phase, categorical labels are encoded to integer values & fit the encoder to the labels, and transform the labels into integer values.

5. Data Splitting

   Split of X and Y to train, validate, and test sets. X denotes all the features in the model and Y denotes the 8 emotions. Out of 4240 data samples, 3707 were used to train, 368 for validation & 162 for testing.

6. One-Hot Encoding

   In audio classification using LSTM, one-hot encoding can be used to convert the class of labels of the audio files into a binary vector representation. This conversion was mainly performed through one hot encoding. It also avoids the problem of assigning arbitrary values to the categories, which can introduce unintended biases or relationships. To perform this y_train and y_val were converted

to ‘One-Hot’ vectors. In this model, we created a dictionary that maps each label to a unique integer using a dictionary comprehension and then create an array of zeros with the shape (num_samples, num_classes) and perform one-hot encoding by setting the corresponding element to 1 in each row of the array.

5.a.4 Feature Engineering

This involved selecting relevant voice features that are indicative of different emotions. The selected features being extracted with librosa for the speech emotion recognition model are:

1. Root Mean Energy(RME) : In this work, RMS measures the overall energy level of the signal. It distinguished between high-energy like excitement and low-energy emotions like sadness.
2. Zero Crossed Rate(ZCR) : This measures how often the signal crosses the zero axis. It distinguished between voiced sounds and unvoiced sounds.
3. Mel-Frequency Cepstral Coefficients (MFCCs) : MFCCs are commonly used for speech recognition tasks and they capture information about different speakers in the model.

Lastly, the features are extracted using a frame length (frame_length = 2048) of 2048 samples and a hop length (hop_lentgh = 512) of 512 samples, which means that the signal is divided into frames of 2048 samples with a 512-sample overlap between adjacent frames. This ensures that the feature values are equally spaced in time and can be compared across different segments of the signal.

$$\text{The number of sequential Features} = \frac{2048}{512}$$

$$= 4$$

5.a.5 Modeling

The model is executed with the Keras library. The model is comprised of 2 hidden LSTM layers with 64 nodes each and an output (dense) layer with 8 nodes, each for one emotion using the 'softmax' activation. The optimizer that led to the best results was 'RMSProp' with default parameters. The batch size of the model is 23. The model summary is as follows.

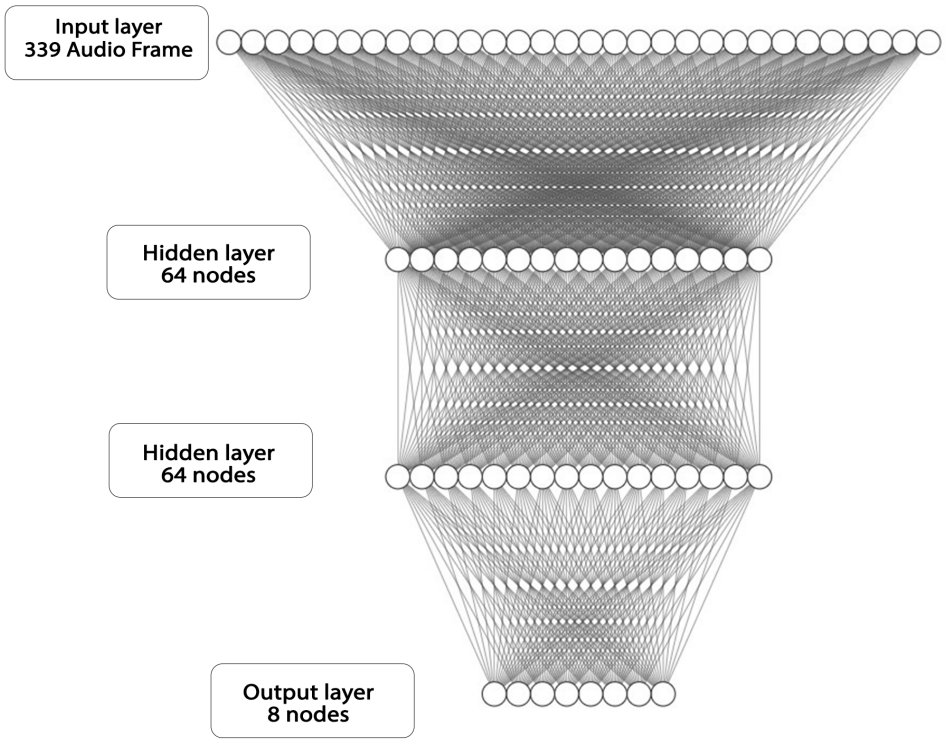

Figure3 - The Model Visualization

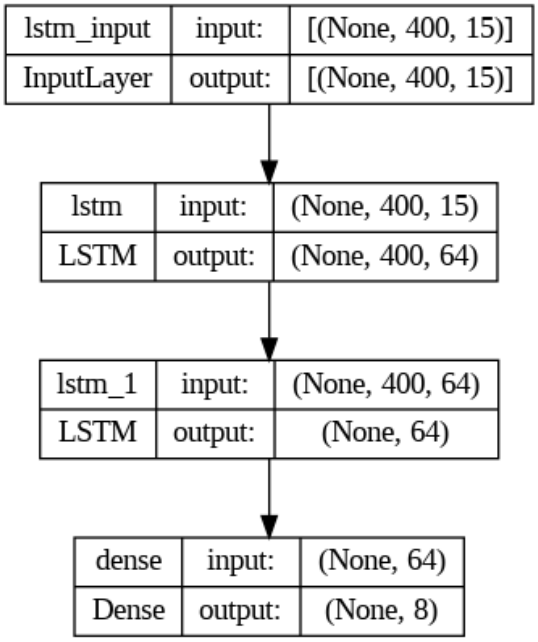

Figure 4 - The Layer diagram

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 400, 64)           20480

 lstm_1 (LSTM)               (None, 64)                33024

 dense (Dense)               (None, 8)                 520

=================================================================
Total params: 54,024
Trainable params: 54,024
Non-trainable params: 0
```

Figure 5 - Layers of the model

5.a.6 Evaluation

The following 3 metrics were used to evaluate the accuracy and the performance of the model.

1. Train (fit) Visualization

   This visualizes the variation of loss values with the number of epochs during the training process.

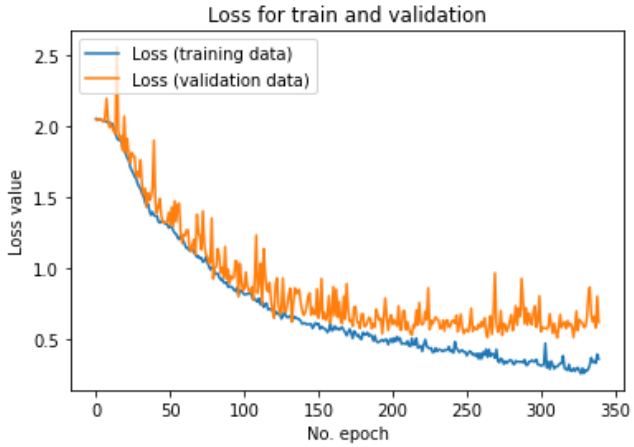

Figure 6

This visualizes the variation of the model's categorical accuracy value with the number of epochs during the training

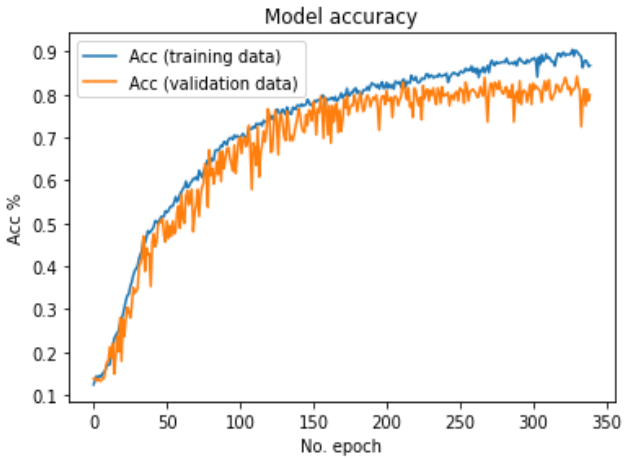

Figure 7

2. Validation Set Evaluation

- Validation set score

```
12/12 - 1s - loss: 0.5901 - categorical_accuracy: 0.8424 - 1s/epoch - 96ms/step
```

Figure 8

- Validation set confusion matrix

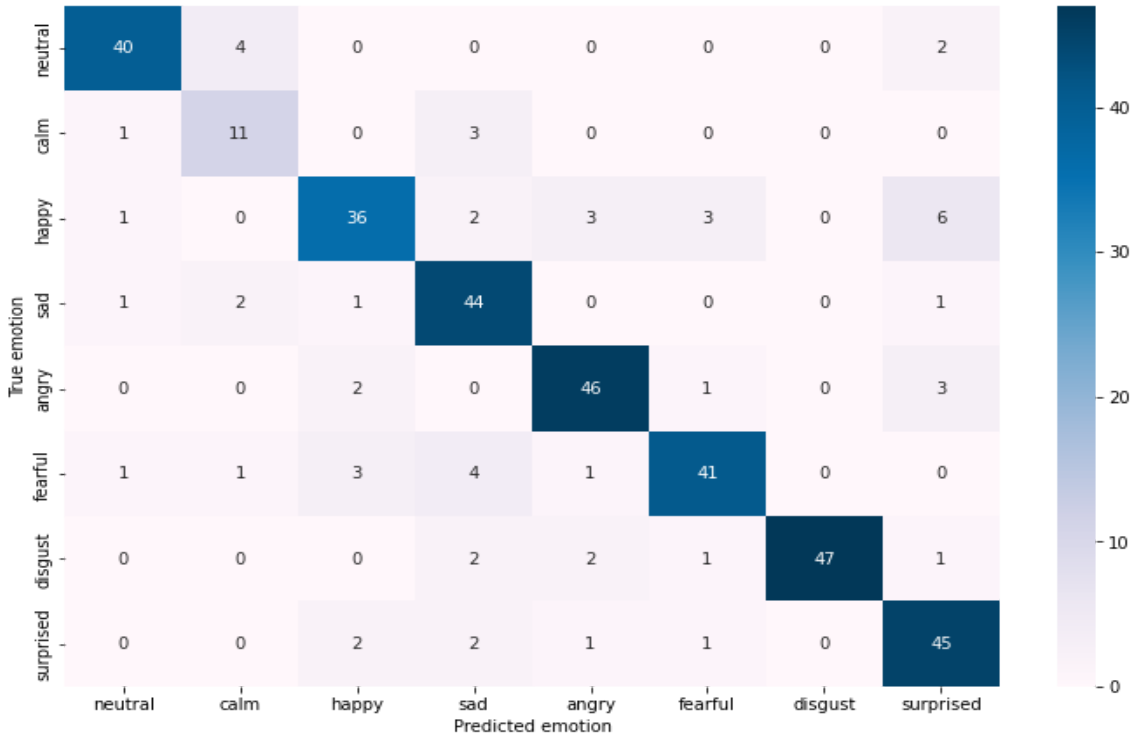

Figure 9

- Validation set predicted accuracy rates

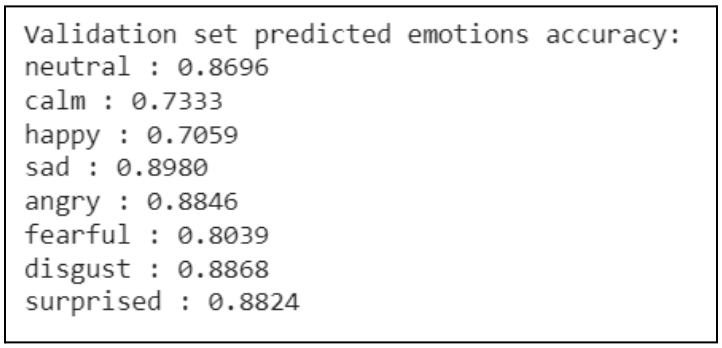

Figure 10

3. Test Set Evaluation

    - Test set score

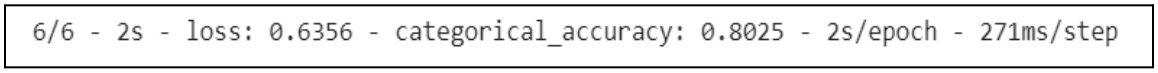

Figure 11

- Test set confusion matrix

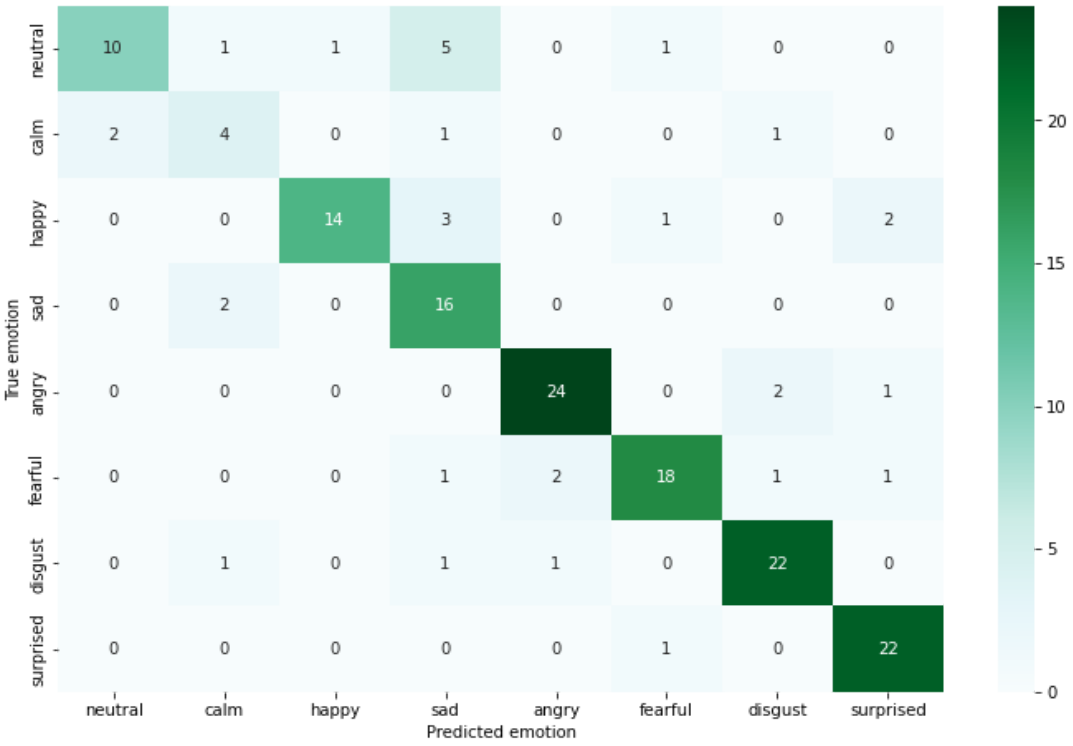

Figure 12

- Test set predicted accuracy rates

```
Test set predicted emotions accuracy:
neutral : 0.5556
calm : 0.5000
happy : 0.7000
sad : 0.8889
angry : 0.8889
fearful : 0.7826
disgust : 0.8800
surprised : 0.9565
```

Figure 13

5.b Model 02 (Gender Detection)

5.b.1 Data acquisition

Common Voice
Link: https://www.kaggle.com/datasets/mozillaorg/common-voice

This Common Voice dataset consists of unique MP3 audio clips and corresponding text files. The dataset has 26119 recorded hours, 17127 validated hours, and more than 104 languages. Further, it includes demographic metadata like age, sex, and accent that can help train the accuracy of speech recognition engines.

- Number of actors: more than 60,000 actors.
- Total Number of files: more than 9 million clips over 104 languages.
- Captured gender: Male, Female, Other.

5.b.2 Voice extraction and Cleanup

The voice paths are their respective features are mapped using the CSV file balanced-all.

5.b.3 Preprocessing

1. Feature Extraction
   This phase converts all the audio files into features that can be input into a model. The data path and the CSV file were also used during the feature extraction for maximum utilization. The extracted features are MFCC, Chroma, MELintopectrogram Frequency (mel), Contrast, and Tonnetz.

2. Data Splitting
   Split of X and Y to train, validate, and test sets. X denotes all the features in the model and Y denotes the gender. From the used data sample 80% of the data was used to train the model while

20% of the data was used to validate and test the model. (10% each)

5.b.4 Feature Engineering

In this phase different types of audio features are extracted from the audio files. The features that are extracted include :

1. MFCC (mfcc) : Used to capture variations in frequency components of a voice signal to distinguish between male and female speakers.
2. Chroma (chroma) : Used for capture aspects of the spectral content of a signal that are related to the speaker's voice.
3. MEL Spectrogram Frequency (mel) : This maps the frequency spectrum of an audio signal to the mel scale to capture the variations in frequency components.
4. Contrast (contrast) : Captured the maximum and minimum magnitudes of adjacent frequency bands.
5. Tonnetz (tonnetz) : Captured the tonal characteristics of an audio signal

For each feature, the function computes the feature value by calling the appropriate librosa function and computing the mean value of the resulting feature matrix along the time axis. The resulting feature values are concatenated to form a single feature vector and extracted features are then saved as numpy arrays. These features are used as input to train a machine-learning model for audio classification.

5.b.5 Modeling

The model is executed with the Keras library. The model consists of 5 hidden layers. A 30% dropout rate after each fully connected layer is adopted to prevent overfitting. The activation function used is sigmoid and binary cross-entropy is the loss function used. The batch size of the model is 64. The adopted neural networking model can be summarized as follows.

```
Model: "sequential_12"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_72 (Dense)            (None, 256)               33024

 dropout_60 (Dropout)        (None, 256)               0

 dense_73 (Dense)            (None, 256)               65792

 dropout_61 (Dropout)        (None, 256)               0

 dense_74 (Dense)            (None, 128)               32896

 dropout_62 (Dropout)        (None, 128)               0

 dense_75 (Dense)            (None, 128)               16512

 dropout_63 (Dropout)        (None, 128)               0

 dense_76 (Dense)            (None, 64)                8256

 dropout_64 (Dropout)        (None, 64)                0

 dense_77 (Dense)            (None, 1)                 65

=================================================================
Total params: 156,545
Trainable params: 156,545
Non-trainable params: 0
```

Figure 14

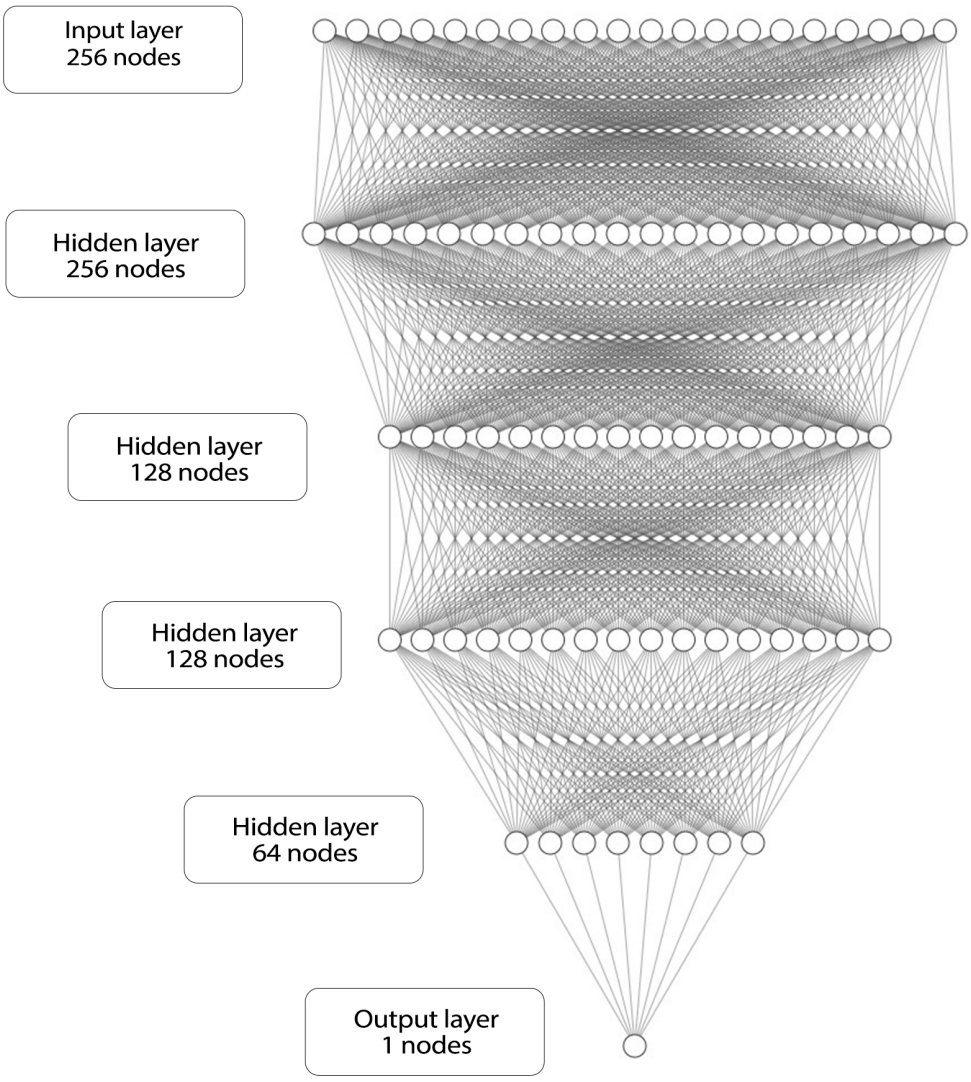

Figure 15 - The Model Visualization

5.b.6 Evaluation

1. Train (fit) Visualization

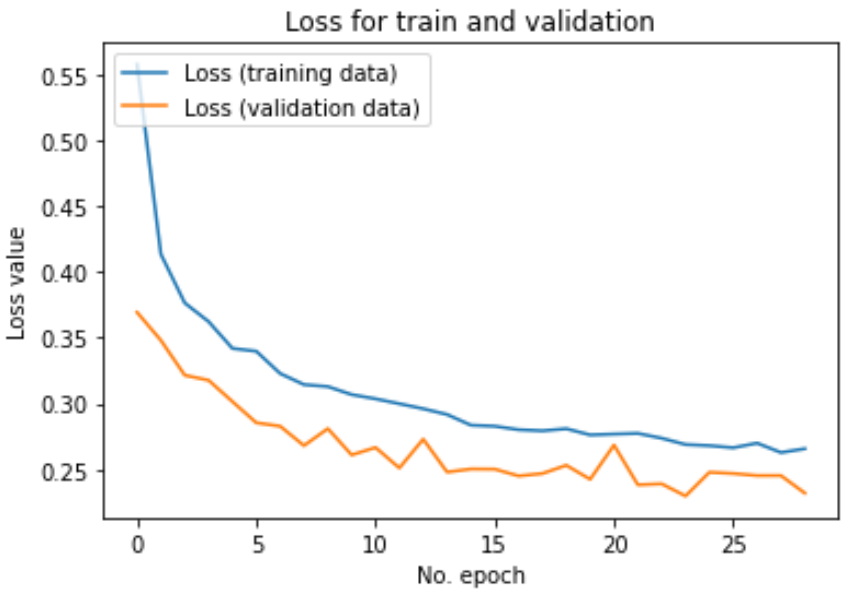

Figure 16

2. Validation Set Evaluation (Validation set score)

```
189/189 - 1s - loss: 0.2297 - accuracy: 0.9125 - 976ms/epoch - 5ms/step
```

Figure 17

6. Results and discussion

6.1 Results and Discussions (Model 1)

1. The Emotion Recognition model gives 84% categorical accuracy. Categorical accuracy, also known as classification accuracy, is a metric used to evaluate the performance of a classification model and is an intuitive way to measure the model's overall accuracy. It is calculated as the number of correctly classified audio samples divided by the total number of samples in the dataset. According to *Figure 8,* that result is a piece of evidence to prove that model 1 performs well. In the first stage of this project, we trained a model that only used a CREMA-D data set and it gave 44% less accuracy because of the limited data size and limited diversity of the dataset. So the team decided to combine the dataset (RAVDEES and TESS) to overcome that problem.

2. According to the following bar charts, it can be concluded that most of the time it gives mixed emotions when considering the probability of each emotion. In such situations, we have to consider the highest probability, but in most cases, the highest probability goes to disgust emotion. And the other obvious result is normally emotions such as neutral, calm, and happy are difficult to classify. This happens because emotions are complex and can be influenced by a variety of factors, including cultural and social contexts. While NLP models can detect certain cues in a person's voice, such as tone and pitch, they cannot fully capture the nuances and complexities of human emotions.

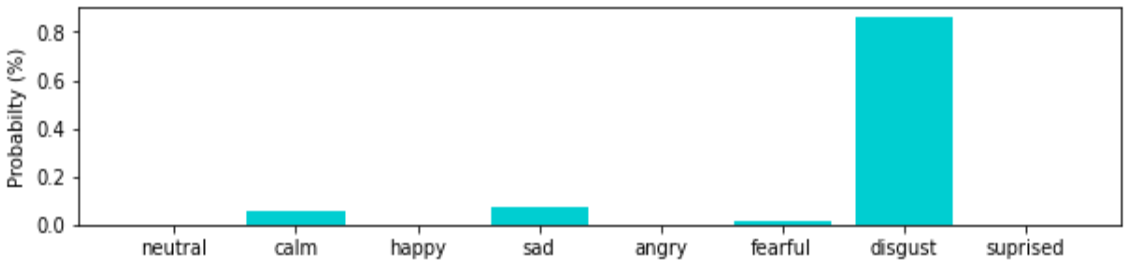

Figure 18

Tested emotion: Happy

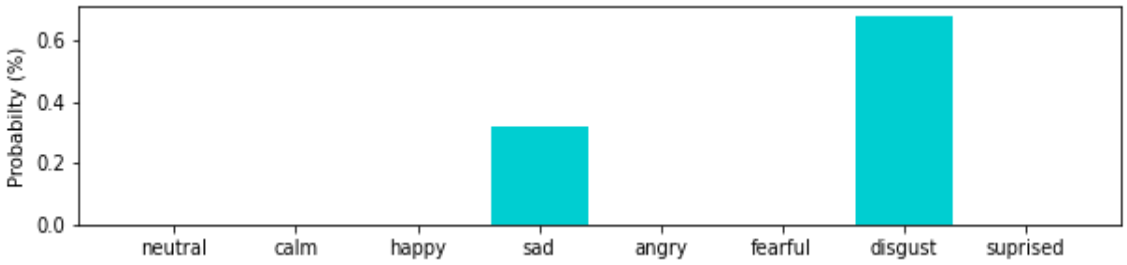

Figure 19

Tested emotion: Sad

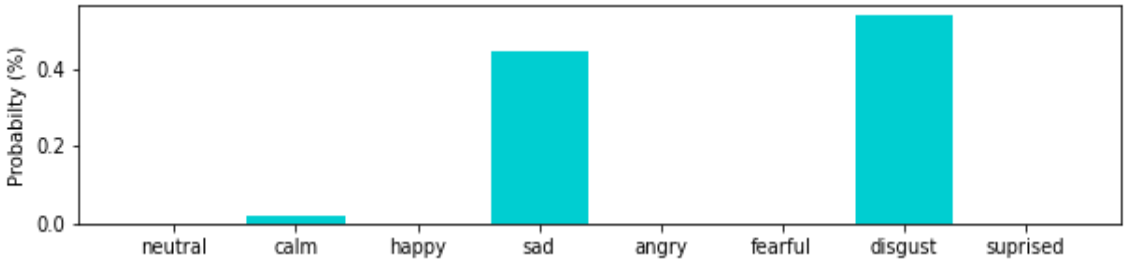

Figure 20

Tested emotion: Neutral

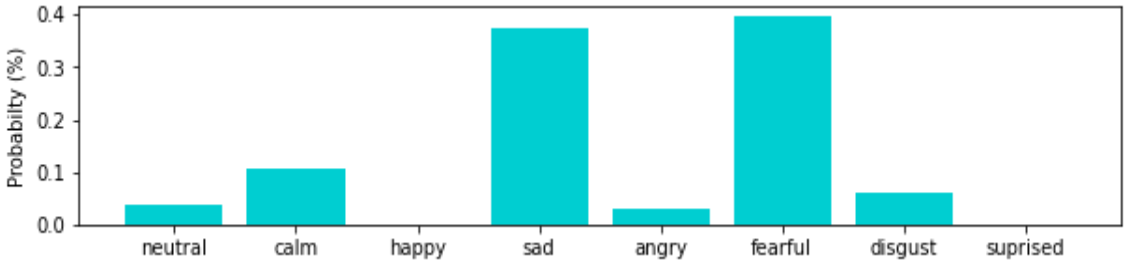

Figure 21

Tested emotion: Fear

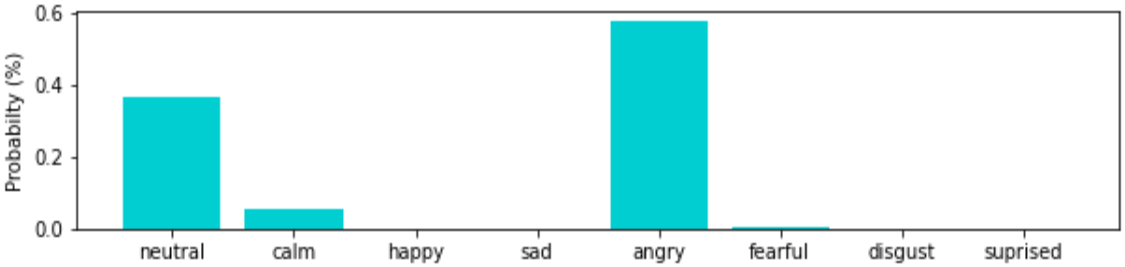

Figure 22

Tested emotion: Angry

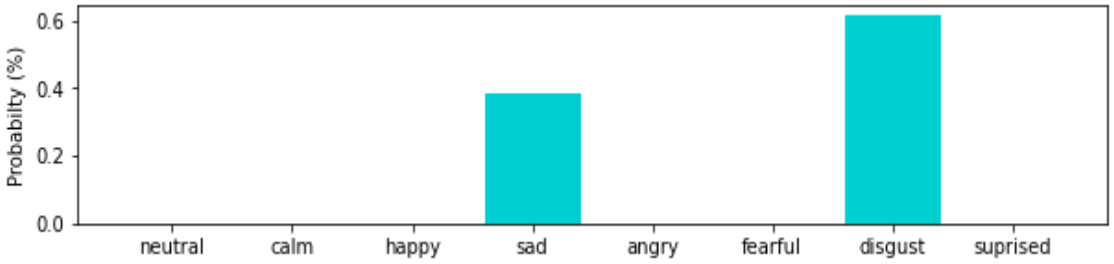

Figure 23

Tested emotion: Disgust

3. Model 1 maintains good value accuracy and it proves that overfitting doesn't occur in model 1. To prevent overfitting an early stopping was added to pause the training phase before the machine learning model learns the noise in the data. Further adding dropouts will also reduce overfitting. Dropouts is a regularization technique that helps to reduce model complexity, generalize the learned features, smooth decision boundaries, etc.

4. At the beginning model 01 was trained using CNN and it also gave less accuracy and wrong predictions during training the model. The next choice was to use LSTM which is Deep Neural Network. LSTM is well-suited for processing sequential data such as speech or audio and it is good at capturing changes in pitch, tone, and other temporal dynamics in speech. CNN only captures a limited amount of context around each input data point and it causes fewer accuracies in models. So from this project, it was

highlighted that LSTM is generally a better choice due to its ability to process sequential data, capture temporal dynamics, handle variable-length input sequences, and capture more context around each input data point.

5. Initially audio files were converted into spectrograms and that was not very practicable due to the large size of the combined dataset. So we focused on the MFCC which is very useful for training models on large datasets. MFCCs also have been shown to be effective in capturing important acoustic features that are relevant to emotions. MFCCs can be computed directly from raw speech signals and typically require fewer data and computing resources compared to spectrograms. Therefore, we used MFCCs directly as input features for our machine learning model, which helped us effectively recognize different emotions from voice signals, especially when dealing with our large datasets.

6. In our Model 1 we used ‘RMSProp’ as our optimizer and it had a significant impact on the performance rather than using optimizers ‘adam’ or ‘SGD’. RMSProp causes increased convergence speed, Sensitivity to the learning rate and prevents overfitting.

6.2 Results and Discussions (Model 2)

1. Gender Recognition model gives 91.46% accuracy.

```
Evaluating the model using 6694 samples...
Loss: 0.2350
Accuracy: 91.46%
```

Figure 24

This Gender recognition model performed well for that 'Common voice' data set. Here we added early stopping and it overcome the model overfitting. Figure 16 shows that the loss is decreasing over time, and the accuracy is increasing.

2. Here we only filtered the labeled samples in the genre field. Before adding to the model we balanced the dataset so that the number of female samples is equal to male samples which helped the neural network not to overfit on a particular gender and this directed us to better predictions as figures 25 and 26.

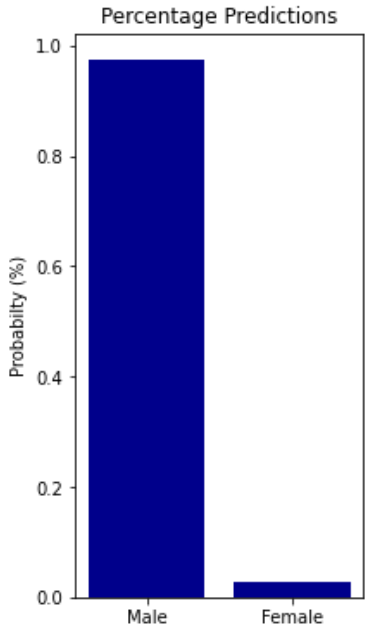

Figure 25

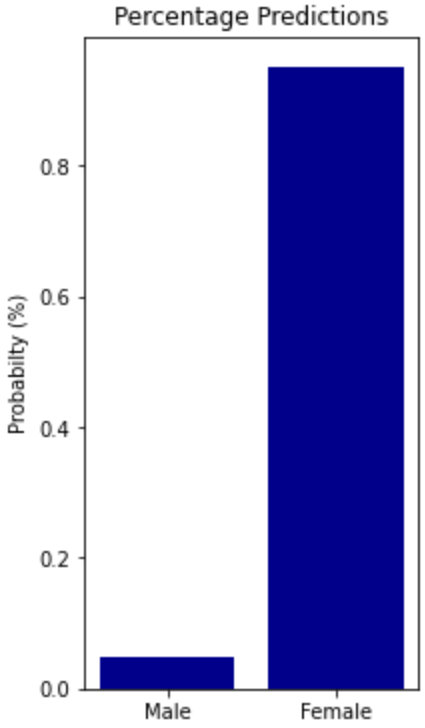

Figure 26

3. Mel Spectrogram extraction technique has been used to get a vector of the length 128 from each voice sample.

4. When model training stage, we used 30% as a dropout value after every fully connected layer. It causes to prevent regularizations such as overfitting.

5. Here we used LSTM as a deep neural network and it helps to increase the accuracy and also has previously mentioned advantages as well.

7. Conclusion

In this project, we developed a system for gender and voice emotion recognition using a Deep Neural Network. Specifically, These LSTM models were trained to classify emotions and gender using voice signals. Our results show that these approaches can accurately recognize and classify emotions and gender using the human voice.

In conclusion, the development of a gender and emotion recognition system using voice is a challenging task due to variations in voice quality, accent, and speaking style. However, this project has demonstrated that efficient approaches such as machine learning algorithms can be adopted to overcome these challenges and develop a reliable computer system.

Some datasets such as, RAVDEES and TESS gathered audio files from the same actor with different emotions, so it is possible to have inaccuracies or inconsistencies in the labeled emotions. Sometimes, the actual emotion conveyed by the actor was not accurately captured in the labeling process. Emotions are complex and multifaceted and can be influenced by a wide variety of factors.

Most of the time human voice is different from person to person because their pitch, tone, accent, and speech patterns are different. But NLP models can be trained on datasets that are biased towards certain voice types, such as male voices with lower frequencies and female voices with higher frequencies. But it is possible for men to have higher frequencies and for women to have lower frequencies. This can lead to errors in natural language processing (NLP) models that are designed to identify gender based on voice characteristics.

We can conclude that if we have a diverse and representative primary dataset is very important when training a model, as it can help reduce certain problems such as bias and overfitting. Moreover, we can reduce the wrong predictions that occurred due to the same actor uttering different phrases with different emotions.

As future work, we can implement this to get real-time inputs and can connect to a database to save them. The saved audio files can be utilized in emotion prediction and gender detection with required pre-processing and feature extractions. Hence enabling real-time output for the inputs. Once the system is successfully implemented with all the modifications the model can be used in numerous applications as mentioned earlier.

Overall, the project has successfully demonstrated the potential of using voice-based gender and emotion recognition systems in various industries and provided insights into the challenges and efficient approaches for developing a reliable computer program for this purpose. The findings of this project can pave the way for further research and development in this field and as well as the other fields to improve the accuracy and effectiveness of gender and emotion recognition systems using voice.

8. Acknowledgement

Natural Language Processing is another module that gives an excellent opportunity to grab knowledge about the fast-growing Artificial Intelligence field. This project requires a considerable amount of hard work, research,40 and dedication. This project would not have been a success without the support of many individuals. Therefore, we would like to extend our sincere gratitude to all of them.

First, We would like to extend our sincere thanks to all the lecturers in the Department of Computer Science, Faculty of Applied Sciences for their extended support and guidance throughout this project and for sharing their knowledge through these years of our university life.

We are highly indebted to Prof. T. G. I. Fernando for his constant supervision, time devoted to discussion, encouragement, and valuable guidance, and for sharing his pearls of wisdom with us during this project.

We are also grateful to all the other resource persons for their valuable time and guidance provided to complete this study successfully.

Nevertheless, we express our gratitude toward our families and colleagues for their kind cooperation and encouragement which help us in the completion of this project.

Appendix

Github page link:

https://github.com/janithTG/emotion-gender-recognition-using-audio

References

1. -Hyun Park and Kwee-Bo Sim. “Emotion Recognition and Acoustic Analysis from Speech Signal” 0-7803-7898-9/03 Q2003 IEEE, International Journal on 2003, volume 3.

2. T L Nwe'; S W FoChango L C De Silva, “Detection of Stress and Emotion in speech Using Traditional And FFT Based Log Energy Features” 0-7803-8185-8/03 2003 IEEE (2003)

3. Keshi Dai1, Harriet J. Fell1, and Joel MacAuslan2“Recognizing Emotion In Speech Using Neural Networks”, IEEE Conference on “Neural Networks and Emotion Recognition” in 2013.

4. AasthaJoshi “Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm”, National Conference on August 2013.

5. Prombut, N.; Waijanya, S.; Promri, N. Feature extraction technique based on Conv1D and Conv2D network for Thai speech emotion recognition. In Proceedings of the NLPIR, Sanya, China, 17–20 December 2021.

6. M. Alnahhas, T. W. Haw, and C. P. Ooi, “Speech Emotion Recognition of Intelligent Virtual Companion for Solitudinarian,” *ResearchGate*, Dec. 2022. https://www.researchgate.net/publication/362792021_Speech_Emotion_Recognition_of_Intelligent_Virtual_Companion_for_Solitudinarian

7. B. T. Atmaja and A. Sasou, “Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations,” *Sensors*, vol. 22, no. 17, p. 6369, Aug. 2022, doi: 10.3390/s22176369.

8. S. Chamishka *et al.*, “A voice-based real-time emotion detection technique using recurrent neural network empowered feature modeling,” *Multimedia Tools and Applications*, vol. 81, no. 24, pp. 35173–35194, Jun. 2022, doi: 10.1007/s11042-022-13363-4.

9. M. Vetráb and G. Gosztolya, “Using the Bag-of-Audio-Words approach for emotion recognition,” Acta Universitatis Sapientiae, Informatica, vol. 14, no. 1, pp. 1–21, Aug. 2022, doi: 10.2478/ausi-2022-0001.

10. P. Rani and Geeta, “Gender and Emotion Recognition Using Voice,” *csjournals*, vol. 10, no. 2, pp. 165–174, [Online]. Available: http://www.csjournals.com/IJEE/PDF10-2