DEVELOPING A MULTIMODAL EMOTION RECOGNITION SYSTEM SOLUTION

by

Aakanksha Chouhan
B.Tech Computer Science & Engineering
2017-2021
SRM University AP, Amaravati, India

Vision and Interaction Group
Supervised by :
Dr Stefan Winkler

Dec 2020 - May 2021

CONTENTS

Abstract

The speech emotion recognition solution was created using traditional machine learning methods. A unique approach was taken, which includes joining numerous machine learning algorithms to classify speech samples progressively. A support vector machine (SVM), a multi-layer perceptron (MLP) were prepared on the Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS), the Toronto Emotional Speech Set (TESS), the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D). Custom feature set is utilized, and its performance is analyzed and hence compared for both the models. Besides, the speech emotion recognizer created, could be integrated with a facial expression recognition model to make a robust, multimodal emotion recognition framework. The aim was to get more precise predictions of emotions by processing information from the sound and video signals.

CHAPTER 1

INTRODUCTION

Literature Review

Speech signal processing has made significant advancements. A portion of its applications incorporate speaker recognition ,automatic speech recognition, language recognition, and mental stress detection. Another colossal space of use is speech emotion recognition (SER). A speech emotion recognition framework can be utilized in call centers to evaluate consumer satisfaction, or it tends to be utilized to improve the learning experience of clients in e-learning platforms, or it can even be utilized in assistive robots to make them more empathetic towards people.

Machine learning strategies in speech processing have become progressively normal, on account of massive enhancements in computational power in the course of recent years. A couple of various machine learning and deep learning classifiers have demonstrated to yield incredible outcomes for emotion speech recognition. The three most basic procedures incorporate the support vector machine (SVM), the multi-layer perceptron (MLP), and the recurrent neural organization (RNN). The authors in [1] utilized three layers of parallel SVM models for the multi-class emotion recognition task. Each model was prepared on one emotion and characterized that emotion against different emotions in a one-versus-all (OVA) style. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) was utilized as the dataset, and attributes like energy, pitch, mel-recurrence cepstral coefficients (MFCC), perceptual linear predictive (PLP), filter bank, and first and second derivatives of all features were extricated as a frame based feature (set) utilizing the Kaldi tool.

In [2], a multilayer perceptron (MLP) was trained on the Emotional Prosody Speech and Transcripts (EPST), an English speech emotion corpus, and KSUEmotions, an Arabic speech emotion corpus, to make a multilingual speech emotion classifier. Audio features utilized incorporate pitch, power, formants, jitter, shimmer, and speech rate. These features were extracted utilizing the PRAAT package, and different combinations of these audio features were tested and compared.

The work in [3] utilized RICOLA, a speech emotion dataset in the French language, alongside a cascaded deep learning architecture that consisted of a convolutional neural network (CNN) trailed by recurrent long short-term memory (LSTM) layers. The CNN took in the sound attributes from raw utterances, which avoided the requirement for conventional hand-picked feature extraction – a process named "end-to-end speech emotion recognition."

The authors in [4] also implemented a cascaded system by studying various combinations of a support vector regression (SVR) model and a bidirectional long short- term memory deep recurrent neural network (BLSTM-DRNN). One implementation, dubbed “dependent training,” used the first model's prediction output to be fed to the second model’s features, along with the other audio features. The other implementation, “independent training,” involved training the models separately but adding Gaussian white noise to the data used for training the first model to modify the true labels and create pseudo predictions. These pseudo predictions were then used as features for the second model along with the other audio features.

In [5], the authors used six SVM classifiers in an OVA binary classification method. All SVM classifiers used the radial basis function (RBF) kernel, and each one gave confidence of an input utterance being the emotion it was trained upon with the input samples. The final prediction came from the SVM classifier that gave the highest confidence. The LDC speech emotion corpus was used, and the final model performance was compared with naïve human coders.

The authors in [6] created a virtual game for children with ASD. The goal was to teach these children how to recognize and express emotions in a game scenario through facial expressions, tone-of-voice and body gestures. The results of the study indicated that there was an improvement in the emotion recognition and socialization skills of the participating children.

In this work, a speech emotion recognition system was developed using two machine learning algorithms – a SVM, an MLP were trained separately and combined using majority voting to give the final emotion class prediction. The datasets they were trained on include the Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS), the Toronto Emotional Speech Set (TESS), and the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D). Additionally, a noise file was generated and added to the final pool of data samples used to train and evaluate the speech emotion recognition system for each clean speech utterance of the three datasets.

CHAPTER II.

AUDIO FEATURES

Recording Audio

Any sound is created by an object's vibration, which causes local air molecules to oscillate and produce a sound wave. Sound waves are a type of mechanical wave that requires a medium to transfer energy from one point to another. For sound waves, this medium is air. Sound waves that exist in nature are analog, continuous-time signals. It needs to be converted into a digital, discrete-time signal to record and store sound. The conversion is done by sampling the audio amplitudes at discrete points in time. The number of audio samples taken per second is defined as the sampling rate. Figure 1 shows the result of using different sampling rates.

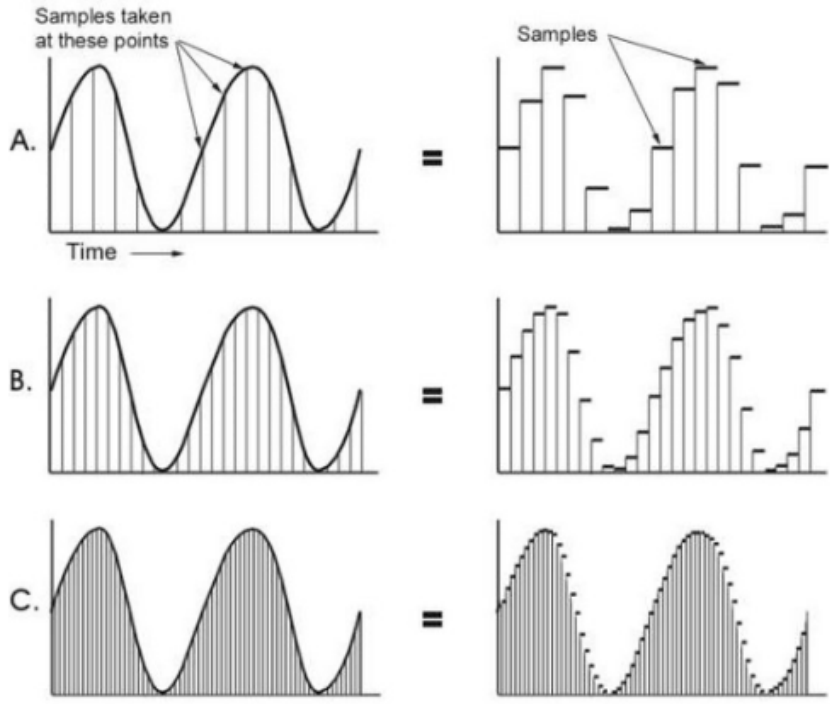

Figure 1 Increasing the sampling rate of an audio signal, where the analog continuous-time signal is shown on the left-hand side of the equal sign [7].

Even though increasing the sampling rate allows for a better approximation of the actual analog signal, it also increases data being recorded. A sampling rate of 16 kHz (16,000 samples/second) is usually used for most audio signal processing applications. Furthermore, since analog audio signals can have an infinite number of possible amplitude values, the amplitude needs to be discretized when converting into a digital signal. The bit depth is the number of possible amplitude values for a single sample of a discrete-time audio signal. Figure 2 shows an analog signal sampled at a bit depth of 4 bits per sample, which gives a total of 2_4 or 16 possible amplitude values for each audio sample. Like the sampling rate, the higher the bit depth, the higher the discrete-time audio signal's resolution. Most audio recordings today are 16-bit audio, with 2_{16} = 65,536 possible amplitude values.

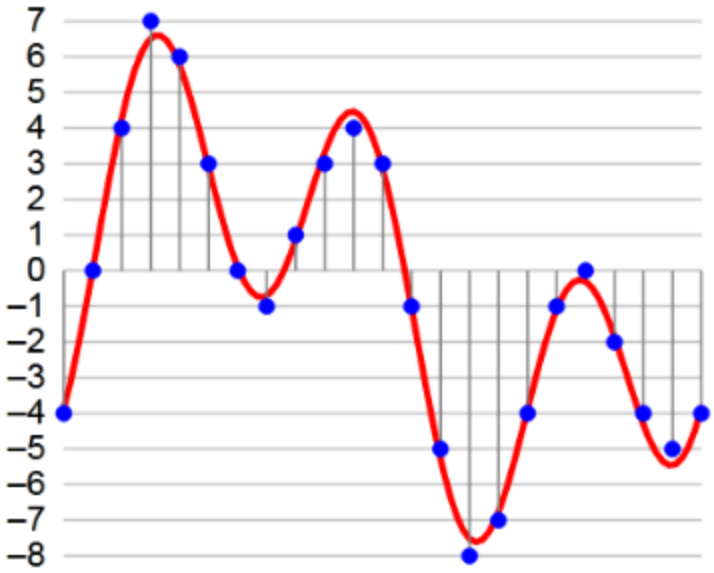

Figure 2 An analog signal (shown in red) is sampled at a bit depth of 4 bits per sample [8]

Audio Preprocessing

A speech signal is a result of a non-stationary process and therefore creates non- stationary data. This means that the statistical properties of the data, such as mean amplitude, standard deviation, and other metrics, change over time. The speech signal is divided into multiple overlapping audio segments called frames to make the data statistically stationary in each frame so that fast Fourier transform (FFT) can be applied for spectral analysis. A typical audio frame is somewhere between 20 to 40 milliseconds long. Figure 3 shows an example of framing a continuous-time signal.

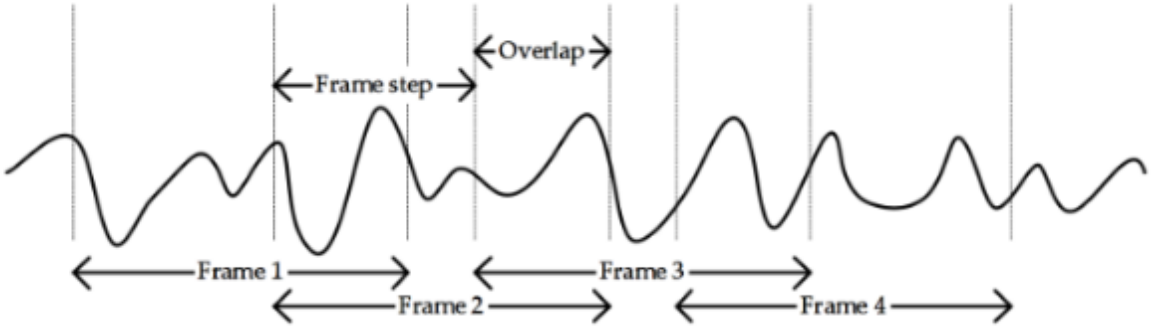

Figure 3 Framing an audio signal [9].

When FFT is applied to any signal, it is assumed that the signal is periodic. Speech signals are non-periodic by nature, and since they do not drop to zero amplitude at the end of each audio frame, the FFT will create high-frequency artifacts at these places. A window function is applied to the audio frames to avoid this issue. This technique is called windowing. A window function is a mathematical function that has an amplitude of zero outside some defined interval. When a window function is multiplied with the speech segment in an audio frame, the resulting speech segment will have an amplitude of zero outside the interval defined in the window function. This essentially smoothens out the edges of the signal in each audio frame. The overlapping regions in audio frames, which are around ten milliseconds long, ensure no audio segment is lost during preprocessing. Figure 4 shows an example of windowing.

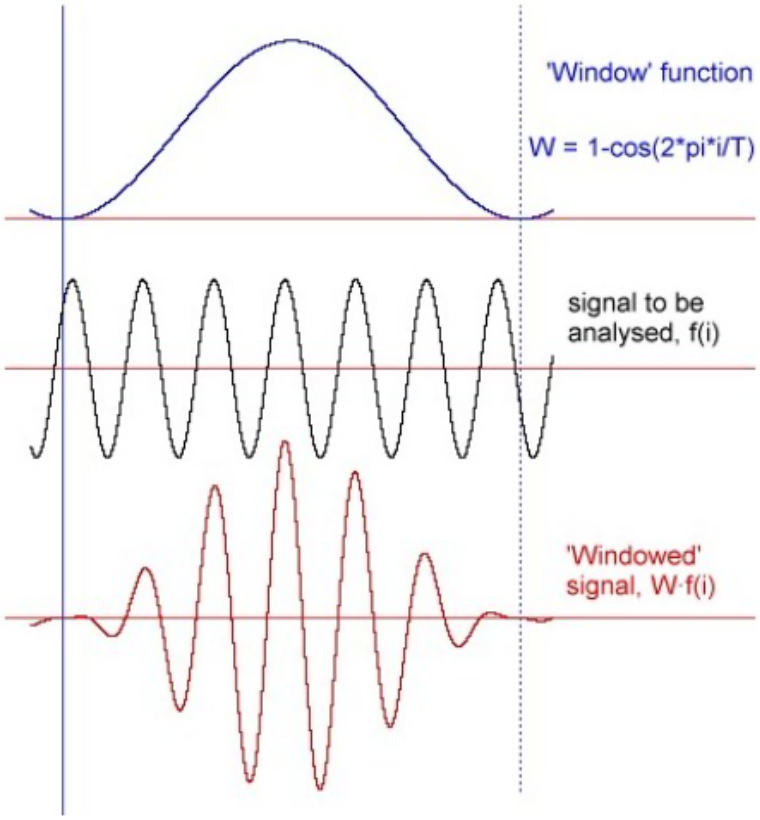

Figure 4 Using a window function on a sinusoid [10].

Features in Speech Emotion Recognition

Audio features are essential in any audio classification task, whether it be speech emotion recognition, speaker recognition, automatic speech recognition, or mental stress detection. In machine learning, features are properties of data that help a machine learn to differentiate between data classes. Researchers have used a wide variety of audio features to classify emotions in speech. In this thesis, two feature sets have been studied. Some of the common low-level descriptors used in emotion speech recognition will be briefly described in the following paragraphs.

The mel-frequency cepstral coefficients (MFCCs) were first introduced in [11]. The first step in calculating the MFCCs is to frame the audio signal into small overlapping audio frames of length 25-40 ms. For a sampling rate of 16 kHz and a frame length of 32 ms, this results in a total of 16x32 = 512 audio samples per frame. Moreover, if the frame step size (hop length) is 16 ms, it results in 16x16 = 256 audio samples in the overlapping regions. Next, the periodogram estimate of the power spectrum is calculated for each audio frame. Then, the spectrum's powers are mapped onto the mel-scale using a mel-filterbank that contains a set of 20-40 triangular overlapping filters, as shown in Figure 5. These filters are spaced according to the mel-scale and give the filterbank energies when applied to the power spectrum. After getting the filterbank energies, the logarithm function is applied to them. This is done because human beings do not hear loudness on a linear scale. The log operation compresses the features so that they match more closely to what humans hear. The final step is to perform a discrete cosine transform (DCT) on the log mel-filterbank energies, which gives the mel-frequency cepstral coefficients.

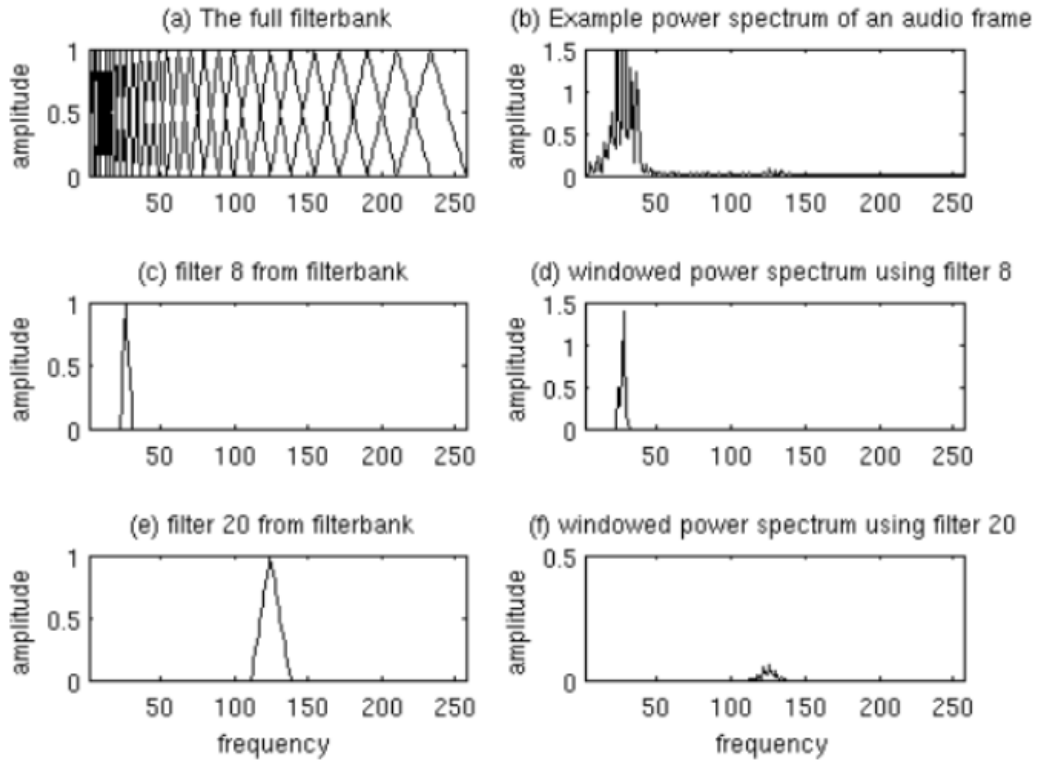

***Figure 5** Applying a mel-filterbank to the power spectrum of an audio frame [12].*

The pitch of a sound is the perceived fundamental frequency F_0, the frequency at which vocal cords vibrate in voiced sounds. Even though the pitch is a qualitative measure, for speech analysis purposes, it is considered to be equal to the logarithmic F0. [13] Just like pitch, loudness is another qualitative measure. Loudness is the perceived intensity of any sound [14].

Formants are specific peak frequencies of vocal tract resonance. They determine the quality of vowels in speech. The first three formant frequencies are labeled F_1, F_2, and F_3. Formants usually occur at 1,000 Hz intervals [15]. Harmonics-to-noise ratio (HNR) is a measure that relates the energy in the periodic part of speech (harmonics) to the energy in the noise section measured in decibels (dB) [16].

Jitter and shimmer are standard perturbation measures in speech analysis. Jitter is a measure of the instability of the fundamental frequency, while shimmer is a measure of amplitude instability in dB [17].

Feature Set

The feature set used in this work is a custom feature set which was created specifically for this work. A trial-and-error method was used on a group of unconventional audio features. The result was a collection of 36 low-level descriptors. They are:

- MFCCs: The first 26 MFCCs were extracted for each audio frame using the HTK implementation [18].
- Spectral contrast: It represents the relative spectral distribution [19]. Seven spectral contrast values were extracted per audio frame.
- Polynomial coefficients: Coefficients of fitting an n_{th}-order polynomial to the columns of a spectrogram. Two polynomial coefficients were extracted per audio frame for a polynomial of order one [20]. .
- RMS energy: The root-mean-square energy of each audio frame. One RMS energy was extracted per audio frame.

All 36 low-level descriptors mentioned above were extracted from speech data using Librosa, a Python library for music and audio analysis [21]. The Python library does all of the audio processing, including framing and windowing. For the final speech emotion recognition model, a sampling rate of 16 kHz was used along with a frame length of 32 ms (512 samples) and a step size of 16 ms (256 samples). For performing FFT, 512 samples were considered per audio frame. Librosa is a reliable tool for audio feature extraction and has been used by researchers for various audio classification tasks. [22][23][24].

Functionals are functions that are applied to a vector. Examples include the mean, standard deviation, maximum, minimum, median, mode, and other metrics. Since each low-level descriptor is extracted for each audio frame, applying functionals provides feature values for the entire audio signal. For the MFCCs, the mean and the standard deviation functionals were used. The mean functional was used for the spectral contrast, polynomial coefficients, and RMS energy. This resulted in a total of 62 audio features in the custom feature set, as listed in Table 1.

Table 1 List of audio features used in the custom feature set.

Low-level descriptors	*Functionals*	*Audio Features*
26 MFCCs	Mean, standard deviation	**52**
7 Spectral Contrasts	Mean	**7**
2 Polynomial Coefficients	Mean	**3**
1 RMS Energy	Mean	**1**

Experiments with (Custom) Feature Set

Python's Librosa library was used to extract the custom feature set from all the data samples. The customization included 36 low-level audio descriptors - the Mel-frequency cepstral coefficients (MFCCs), the root-mean-square (RMS) energy, the spectral contrast, and the polynomial coefficients. Among these low-level descriptors, the MFCCs and the RMS energy were used in most of the prior speech emotion recognition related work. The other descriptors are mainly used in music classification tasks. However, they have shown to yield good classification accuracy when applied to emotion classification tasks in this work. A total of 62 audio features were created using the four low-level audio descriptors of the custom feature set for the SVM and MLP models. They are 26 mean values of first 26 MFCCs across all audio frames, 26 standard deviations of first 26 MFCCs across all audio frames, one mean RMS energy across all audio frames, seven mean values of spectral contrast across all audio frames, and two mean values of polynomial coefficients across all audio frames.

CHAPTER III.

DATASETS

For this research work, three separate speech corpora were selected. They are the Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS), the Toronto Emotional Speech Set (TESS), and the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D). All three datasets are available online to be used by researchers, free of cost. These datasets were designed and created specifically for speech emotion recognition and were evaluated and validated by multiple individuals. There are two main types of speech datasets that are used by researchers in this field. The first type contains recordings of people who express genuine emotions by being subjected to external influence, such as image, video, and audio; the second type contains recordings of professional actors reading outlines from a script while acting out the emotions. The former type of dataset is known in the literature as a spontaneous dataset, and the latter is known as a simulated dataset. Researchers usually use simulated speech data for speech emotion recognition tasks because they are accurately labeled, and actors are very good at expressing each emotion with reasonable accuracy. Also, most spontaneous datasets are limited in terms of the number of emotion classes.

Out of the three speech corpora used, the RAVDESS and CREMA-D corpora contain multi-modal audio and video data.

Ryerson Audio-Visual Database of Emotional Speech and Songs

The Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS) was released in 2018 by researchers of the SMART lab at Ryerson University in Toronto, Ontario, Canada. It is a simulated, multi-modal dataset that contains both video data and audio data. For the audio data, the actors recorded the sentences both as normal speech and as songs. The song data was not considered for this thesis. The audio files were recorded at 16 bits per sample, at a sampling rate of 48 kHz, and in WAV audio format. A total of 24 actors of age range 21-33 years had taken part in creating this dataset, where half of the samples contain male actors and the other half are female actors. Each actor recorded two lexically matched sentences in eight different emotions. The two sentences are "Kids are talking by the door" and "Dogs are sitting by the door." Moreover, the eight emotions are neutral, calm, happy, sad, anger, fear, disgust, and surprise. The emotion labels are included in the WAV audio file names. Out of the eight emotions, seven of them were recorded twice per sentence – once with normal intensity and the other time with stronger intensity. There was only one recording per sentence for the neutral emotion since there is no strong intensity for this emotion. Since calm emotion data was imbalanced, this class was excluded from the study. This gives a total of 1,440 recording samples, with 24 actors x 2 sentences x 8 emotions x 2 repetitions x 2 emotional intensity with the exception of the neutral emotion. Thus, seven emotions have 192 data samples, and the neutral class has 96 data samples. Data resampling was used to match the neutral class count to the rest of the classes . The RAVDESS speech corpus can be downloaded from [25][26].

https://zenodo.org/record/1188976

Toronto Emotional Speech Set

The second speech corpus used for this study was the Toronto Emotional Speech Set (TESS). This simulated dataset was created in 2010 by researchers from the University of Toronto Psychology Department. It contains recordings from two actors, both females. The younger actor was 26 years old at the time of recording, while the older actor was 64 years old. They have recorded 2,800 sentences in seven different emotions – anger, disgust, fear, happy, surprise, sad, and neutral. Unlike RAVDESS, this is a balanced dataset with each emotion class having 400 data samples. However, just like RAVDESS, the sentences spoken by the actors are lexically similar. Each actor recorded the phrase "Say the word___" followed by one of 200 different target words in each affective state. All audio files were recorded at 16-bits per sample, at a sampling rate of 24,414 Hz, and saved in WAV audio file format. The emotion labels were extracted from the file names [27]. The TESS dataset can be downloaded from [28].

https://tspace.library.utoronto.ca/handle/1807/24487

Crowd-sourced Emotional Multimodal Actors Dataset

The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) was the third speech emotion dataset used to develop the speech model in this thesis. Just like RAVDESS and TESS, this is also a simulated speech corpus. It was released in 2014 as a collaboration between researchers from the University of Pennsylvania, Ursinus College and the University of Illinois at Chicago. Actors who participated had ages ranging from 20 to 74 years old. There were 48 male actors and 43 female actors with a total of 91, and even though they came from different ethnic backgrounds, they were all English speakers. Like RAVDESS, it is also a multimodal dataset. The speech recordings were done at 16-bits per sample, at a sampling rate of 16 kHz, and saved in the WAV audio format. The actors recorded twelve different emotionally neutral sentences in six different emotions of anger, disgust, fear, happy, neutral, and sad, at four different intensity levels of low, medium, high, and unspecified. Example phrases include “Don’t forget a jacket,” “I think I’ve seen this before,” “I think I have a doctor’s appointment.” Each emotion class has 1,271 data samples, except for the neutral class, which has 1,087 data samples. Resampling was used to create a balanced dataset [29]. The CREMA-D dataset can be accessed from [30]. Table 2 gives a summary of the three datasets used for this work.

https://github.com/CheyneyComputerScience/CREMA-D

Table 2 Summary of all three speech emotion corpora

Corpus	Age of participants	No. of sentences	No. of participants	Emotions	Samples per class (balanced)	Total (balanced)
RAVDESS	21-33 y. o	2 (with two repetitions and intensities)	24 (12 males, 12 females)	8 (calm excluded)	192	1,344 (192 x 7)
TESS	26-64 y.o	200 (three common words, one changing)	2 (0 males, 2 females)	7	400	2,800 (400 x 7)
CREMA - D	20-74 y.o	12 (four intensities)	91 (48 males, 43 females)	6 (surprise missing)	1,271	7,626 (1,271 x 6)

CHAPTER IV.

RESEARCH METHODOLOGY

Experimental Procedure

Three speech emotion corpora were gathered from the web - the Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS), the Toronto Emotional Speech Set (TESS), the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) to make the speech emotion recognition framework. First, the RAVDESS corpus was chosen for training and optimizing the support vector machine (SVM), the multi-layer perceptron (MLP) model. The RAVDESS dataset contains information from all the seven emotion classes utilized in this work, and there is an equivalent number of male and female actors. Likewise, the recording quality in RAVDESS is better contrasted with CREMA-D. Besides, training the models on a single dataset was quicker than training them on each of the three. Every one of these factors settled on RAVDESS as the best first choice. However, subsequent to utilizing the RAVDESS hyperparameter settings on the other two datasets, the outcomes were substandard. This was on the grounds that the RAVDESS dataset had not many data points, thus the machine had low generalising capability when tried on other datasets.

Performance Metrics

Individual performance metrics have been selected to assess the results of the experiments conducted in this research. They are listed below.

1. Training, Validation, and Test Accuracy
2. Learning curves
3. Accuracy curves
4. Loss curves
5. Precision and Recall
6. Confusion Matrix
7. K fold cross validation

Methodology

The flow diagram shown in Figure * explains how the model hyperparameters were tuned during the training phase of a model. At first, a speech emotion corpus was selected. Then framing and windowing were applied as a part of data preprocessing workflow. After that, the low-level descriptors were extracted from the audio samples, and functionals, such as mean and standard deviations were computed across all audio frames. This resulted in the audio features that were then scaled using standardization. Standardization is performed by subtracting the mean of a feature from that feature value and then dividing it by the feature's standard deviation. This ensures that the feature values have a mean of zero and has a standard deviation of one. Machine learning algorithms like SVM are sensitive to unscaled data. The next step was to partition the data into the training set, the validation set, and the test set. After that, the initial hyperparameter values were set, and then the machine learning algorithm was trained on the training set. Once training was done, the model performance was evaluated on validation data. If the performance metrics indicated a case of either overfitting or underfitting, the hyperparameters were re-tuned, and therefore the machine learning algorithm was trained again using the new hyperparameter values. This process was repeated until a considerably fair bias-variance tradeoff was achieved. Once the model was finalized, an unbiased performance evaluation was obtained using the test set. The performance metrics from this final evaluation indicate how the model will perform when exposed to previously unseen data.

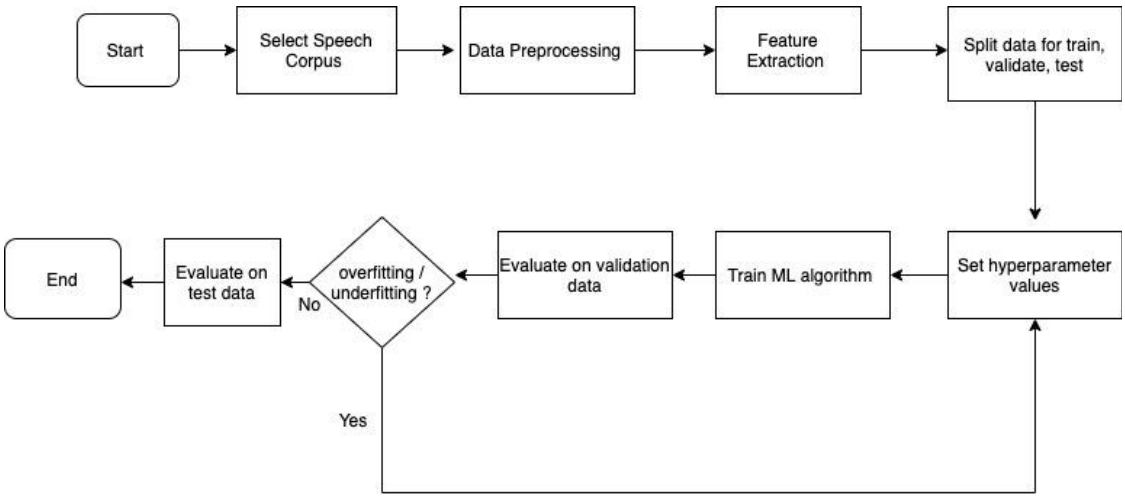

Figure * Flowchart depicting of the workflow for model training and hyperparameter tuning

The hyperparameter tuning experiments were conducted on the complete dataset using the custom feature set. These hyperparameter values were also used when the algorithms were trained on all datasets. Once all 48 experiments were completed, one model was selected from each type of machine learning classifier.

Computational Resources

For this work, the Python version 3.7.4 was used for the development of the machine learning models. Table 3 lists all the Python libraries used, along with their versions. The Python projects are susceptible to package versions due to the dependencies among packages. One way to avoid this is to create separate Python environments, i.e., install separate Python versions, for each project.

Library	Version
scikit - learn	0.24.2
joblib	1.0.1
tensorflow	2.4.1
tensorflow - estimator	2.4.1
h5py	3.2.1
numpy	1.19.2
pandas	1.2.4
matplotlib	3.1.3
praat -parselmouth	0.3.3
librosa	0.8.0
numba	0.48.0

CHAPTER V.

RESULTS AND DISCUSSION

The complete dataset is comprised of all the original clean speech utterances of the Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS), the Toronto Emotional Speech Set (TESS), and the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D). The minority classes were resampled (with replacement) to match the sample count of the majority classes. The neutral class samples were lower in both RAVDESS and CREMA-D, and the surprise class was missing from CREMA-D. The hyperparameters of all the models discussed in this section were tuned while being trained on the complete dataset. The data split of 80:10:10 was used, where 80% of the dataset was used in training the models, while 10% was used for validation and the other 10% was used for testing. Each data split was stratified, meaning that there were an equal number of data samples per emotion class in each of the three data splits (training, validation, and test).

SVM Model

In Figure 6, the learning curves were plotted for the SVM model trained on the Complete Clean dataset using the custom feature set. Figure 7 shows the confusion matrix for this model, and Table 4 gives a summary of the results. For this model, the radial basis function (RBF) kernel was used, with C=10.0 and γ=0.01. The Scikit-learn library was utilized to develop the SVM model in Python.

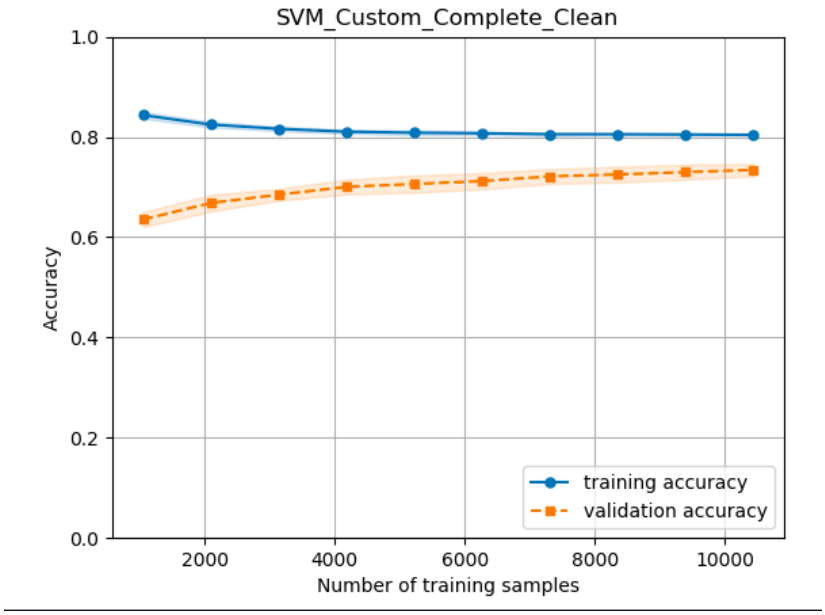

Figure 6 Learning curves for the SVM model trained on the Complete Clean corpus, using the custom feature set.

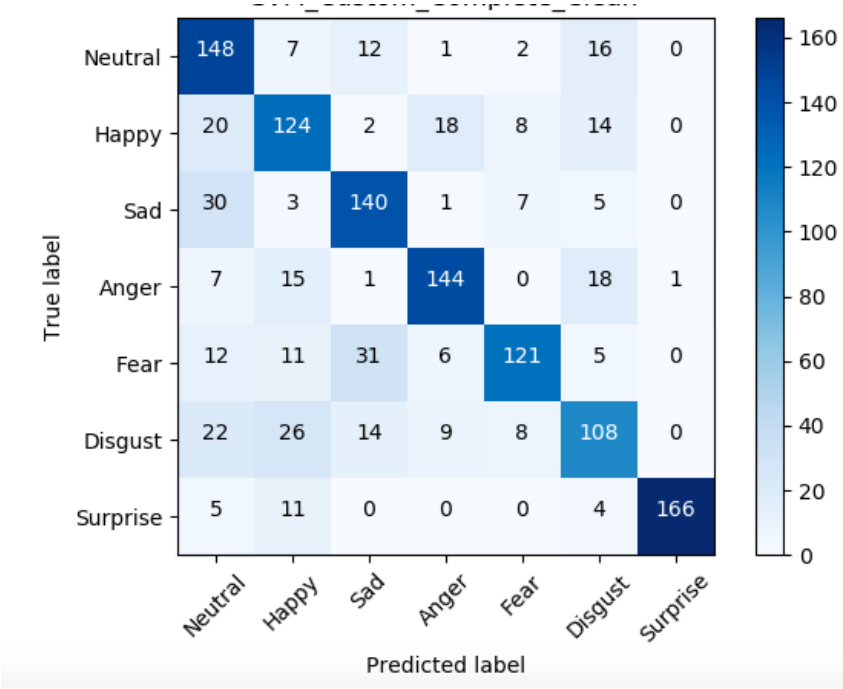

FIgure 7 Confusion matrix for the SVM model trained on the Complete Clean corpus, using the custom feature set.

Table 4 Result summary for the SVM model trained on the Complete Clean corpus, using the custom feature set.

Data samples in corpus	10,435
Training : validation : test	80:10:10
Training accuracy	81.0 %
Validation accuracy	72.0 %
Test accuracy	73.0 %
Precision	74.0 %
Recall	73.0 %

The learning curves of this model show that the validation accuracy closely follows the training accuracy. This is due to the values picked for the C and γ parameters, which ensured that the model did not overfit the training data. In the confusion matrix of the experiment, if all numbers across a row are added, it gives the total number of data samples in the test set for the emotion label mentioned in that row's name. Since the test set was equal to 10 % of the entire dataset, it contained a total of 2,604 data samples. For the stratified test set with seven emotion classes, this resulted in 2,604/7 = 372 data samples per emotion class, each is equal to the sum of the numbers in each row of the confusion matrix. From the confusion matrix, it can be seen that the surprise emotion was the most accurately detected emotion, followed by anger, sadness, and neutral.

MLP Model

The number of artificial neuron units used in an artificial neural network and the number of layers are hyperparameters that can be tuned for getting high accuracies. There is no golden rule for selecting the number of neurons or layers. Researchers usually experiment with these parameters and select values that provide the highest performance. A common convention among computer scientists is to use log BASE-2 numbers, like 64, 128, and 256 [31]. Another convention is to use increments of 50 or hundred, like 50, 100, and 200 [32]. The number of input-layer neurons is equal to the number of input features, and the number of output-layer neurons is equal to the number of classes in the dataset. Even though there is no rule for selecting the number of neurons in the hidden layer(s), there are some rules-of-thumb that can be followed, according to [33]. To design the MLP architecture of this model, a number of neurons, such as 10, 50, 100, 200, and 500, were selected for each layer. The rules-of-thumb described in [33] were used to select the final number of neurons and layers for the high-performing architectures. The architecture for the MLP model that used the custom feature set is shown in Figure 8 There are 62 units in the input layer, which correspond to the number of input features in the custom feature set. The first hidden layer has 105 units, which is 170% of the number of input units used. The second hidden layer has 62 units, which is equal to the number of input neurons. Finally, the output layer has seven units, corresponding to the seven emotion classes used in this work.

Input layer
62 units

Hidden layer 1
105 units

Hidden layer 2
62 units

Output layer
7 units

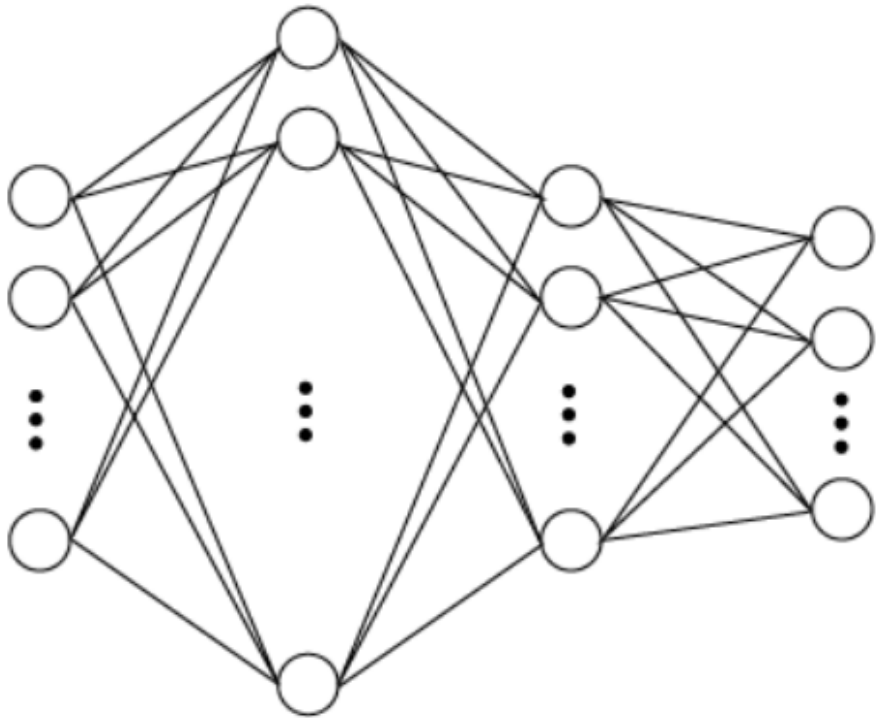

Figure 8 Architecture of the MLP used with the custom feature set.

The Adam optimizer was used in order to minimize the loss function, which in this case is the categorical cross-entropy loss for the MLP model. The rectified linear unit (ReLU) activation function was used and for the output units, the Softmax activation function was used, which provides the prediction accuracies for each class for the hidden layer units. Instead of using a fixed learning rate, a learning rate scheduler was used to change the learning rate as the training progressed. An inverse time decay function was used as the learning rate schedule, with an initial

learning rate of 0.01, 1000 decay steps, and a decay rate of 80%. The training, validation, and testing data were each divided into batches of size sixteen, and 50 epochs were used during training. Dropout is a regularization technique where a fraction of the connections between the hidden layer neurons are randomly dropped during training. This ensures that the model is not overfitting to the training data. During the validation and testing phases, however, all the neurons are connected, i.e., no dropout is used. After trying out different dropouts at different positions, the best combination was used. A dropout of 30% was used between the two hidden layers, and a 10 % dropout was used between the second hidden layer and the output layer. The accuracy curves for the MLP model are plotted in Figure 9 while the loss curves are plotted in Figure 9 showing the confusion matrix of this model. The result summary is given in Table 5 The model was created using the Keras API from the TensorFlow library in Python.

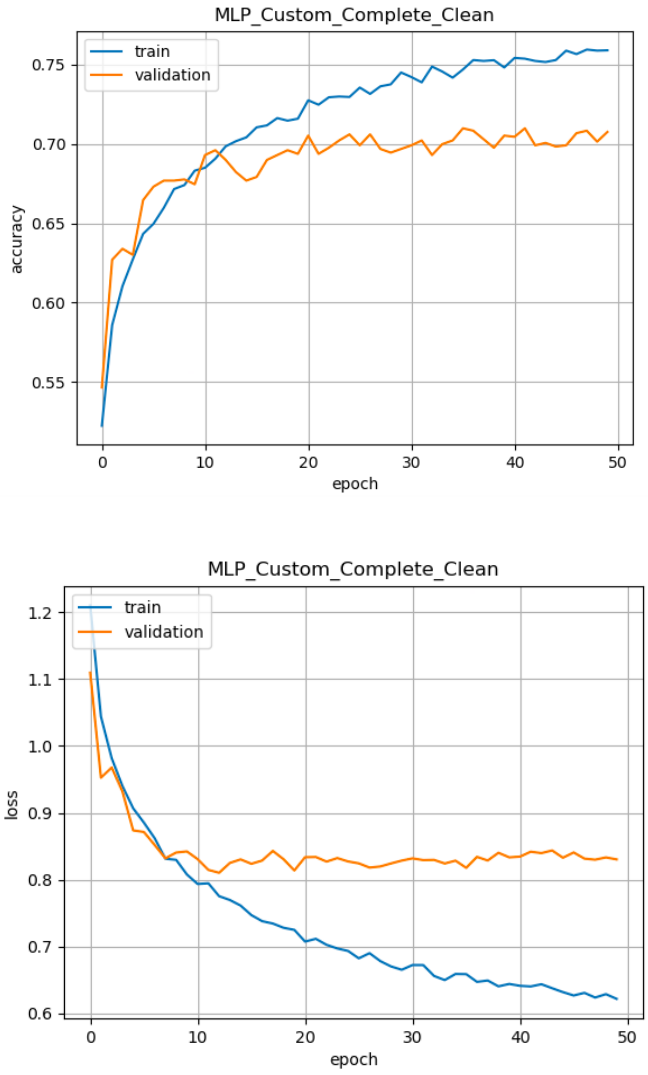

Figure 8 Accuracy curve and loss curve for the MLP model trained on the complete dataset using the feature set of 62 .

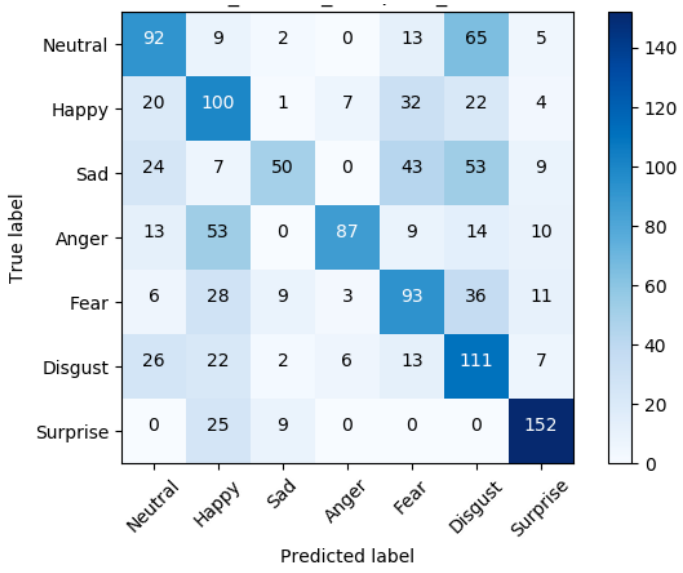

Figure 9 Confusion matrix for the MLP model trained on the Complete Clean corpus, using the custom feature set of 62.

Table 5 Result summary for the SVM model trained on the Complete Clean corpus, using the custom feature set.

Data samples in corpus	10,435
Training : validation : test	80:10:10
Training accuracy	76.0 %
Validation accuracy	71.0 %
Test accuracy	73.8 %
Precision	85.0%
Recall	67.0 %

For this MLP model, the accuracy curves are very close, and the same can be observed for the loss curves. This is a sign of a properly tuned neural network. During the first round of experiments, the model was trained without any regularization. The resulting curves and classification scores indicated a huge overfitting issue. After introducing dropout between the layers, the gap between the validation accuracy and the training accuracy was reduced. Also, the learning rate was kept constant during the first few experiments. The problem with that was the optimizer kept overshooting the minimum loss due to the fixed learning rate being unnecessarily high at that stage of the training. This was visible from the rising loss curves. After using a learning rate scheduler, which slowly reduced the learning rate as training progressed, the losses seemed to decrease consistently. The test accuracy of this model was about one percent less than that of the SVM model. However, the precision score was higher in this model. The training of this MLP model was halted after 50 epochs because the validation loss was almost steady after the 50th epoch. Again, the surprise emotion was the most accurately classified emotion, followed by anger and sadness, looking at the confusion matrix. The training time of this model was similar to that of the SVM.

CHAPTER VI.

CONCLUSION

Future Work

There are specific techniques that could be used to improve the speech emotion recognition system's performance. The most effective technique is to gather more data. The more data is used in training a machine learning model, the more variations in data samples are experienced and learned by the machine. Although it is not possible to always collect more data, the data at hand can be manipulated to improve the generalising capacity of the models. Data augmentation techniques would undoubtedly be at an added advantage. However, it is essential to collect properly labeled data, as data that are wrongly labeled can worsen the model performance. The models are as good as the data that has been fed to them. More data beats clever algorithms but better data beats more data. The three datasets used in this work were easily accessible to the public, free of cost. However, most speech emotion corpus is not easily accessible, as they require permission from the creators or some fee. Plus, there is a limited number of North American (English) speech emotion recognition datasets. Therefore, gathering more data is a challenging task. This work did not include any speech corpora containing recordings of children; such datasets could be used to train the speech emotion recognition system. Another way to get better prediction accuracies is to utilize more features for the design, development, and training of the deep learning model. Features are the diet to the algorithms. In the models used in this work, the classification accuracies did not exceed 80%. Using more features will increase the machine's learning capability and improve the classification accuracy, given that the current features are incapable of learning all the complexities of the data. However, there is a risk of overfitting the model to the training data when using more features. Thus, more features should be added with caution. Future work should be streamlined to strike a balance between the right kind of data and enough features with a simple model is expected to beat the sophisticated one trained on the poor quality and poor quantity of data.

REFERENCES

[1] N. Kurpukdee, S. Kasuriya, V. Chunwijitra, C. Wutiwiwatchai and P. Lamsrichan, “A study of support vector machines for emotional speech recognition,” 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), pp. 1-6, 2017.

[2] A. Meftah, Y. Alotaibi and S. Selouani, “Emotional speech recognition: A multilingual perspective,” 2016 International Conference on Bio-engineering for Smart Technologies (BioSMART), pp. 1-4, 2016.

[3] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller and S. Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200-5204, 2016.

[4] J. Han, Z. Zhang, F. Ringeval and B. Schuller, “Prediction-based learning for continuous emotion recognition in speech,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5005-5009, 2017.

[5] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan and W. Heinzelman, "Emotion classification: How does an automated system compare to Naive human coders?," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2274-2278, 2016.

[6] B. Schuller, E. Marchi, S. Baron-Cohen, A. Lassalle, H. O’Reilly, D. Pigat, P. Robinson, I. Davies, T. Baltrusaitis and M. Mahmoud, “Recent developments and results of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions,” IDGEI, 2015 (No pagination provided).

[7] "Digital Audio Basics: Sample Rate and Bit Depth," [Online] Available: https://www.izotope.com/en/learn/digital-audio-basics-sample-rate-and-bit- depth.html. [Accessed: 18-feb-2021].

[8] "Audio bit depth," [Online] Available: https://en.wikipedia.org/wiki/Audio_bit_depth. [Accessed: 18-feb-2021].

[9] "How do I calculate the number of overlapping frames an given audio file has?" [Online] Available: https://math.stackexchange.com/questions/2249977/how-do-i-compute-thenumber-of-overlapping-frames-an-given-audio-file-has. [Accessed: 18- feb-2021].

[10] "3. Signal Windowing," [Online] Available: http://www.atx7006.com/articles/dynamic_analysis/windowing. [Accessed: 18-feb- 2021].

[11] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 28, pp. 357–366, 1980.

[12] "Mel Frequency Cepstral Coefficient (MFCC) tutorial," [Online] Available: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/. [Accessed: 18-feb-2021].

[13] "Fundamental Frequency, Pitch, F0," [Online] Available: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-73003-5_775. [Accessed: 18-feb-2021].

[14] "Sound Intensity and Loudness," [Online] Available: https://www.nps.gov/teachers/classrooms/sound-intensity-and-loudness.htm. [Accessed: 18-feb-2021].

[15] "What are formants?" [Online] Available: https://person2.sol.lu.se/SidneyWood/praate/whatform.html. [Accessed: 02-mar- 2021].

[16] “Harmonicity,”, [Online] Available: https://www.fon.hum.uva.nl/praat/manual/Harmonicity.html. [Accessed: 02-mar- 2021].

[17] M. Farrús, J. Hernando and P. Ejarque, “Jitter and shimmer measurements for speaker recognition,” 8th Annual Conference of the International Speech Communication Association, Interspeech, vol. 2, pp. 778-781, 2007.

[18] “What is HTK?” [Online] Available: http://htk.eng.cam.ac.uk/ [Accessed: 09-Feb-2021].

[19] D. Jiang, L. Lu, H. Zhang, J. Tao and L. Cai, "Music type classification by spectral contrast feature," Proceedings. IEEE International Conference on Multimedia and Expo, vol. 1, pp. 113-116, 2002.

[20] O. Agcaoglu, B. Santhanam and M. Hayat, “Improved spectrograms using the discrete Fractional Fourier transform,” IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE), pp. 80-85, 2013.

[21] “librosa,”[Online]Available:https://librosa.org/doc/latest/index.html.[Accessed: 18-jan-2021].

[22] A. A. Bashit and D. Valles, "A mel-filterbank and MFCC-based neural network approach to train the Houston Toad call detection system design," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 438-443, 2018.

[23] A. A. Bashit and D. Valles, "MFCC-based Houston Toad call detection using LSTM," 2019 IEEE International Symposium on Measurement and Control in Robotics (ISMCR), pp. 1-6, 2019.

[24] A. A. Bashit and D. Valles, "A solar powered raspberry pi Houston Toad call detection system using neural network model," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1024-1027, 2018.

[25] S. R. Livingstone and F. A. Russo, “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, vol. 13, pp. 1-35, 2018.

[26] “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),” [Online] Available: https://zenodo.org/record/1188976. [Accessed: 18-jan-2021].

[27] “Toronto emotional speech set (TESS) A dataset for training emotion (7 cardinal emotions) classification in audio,” [Online] Available: https://www.kaggle.com/ejlok1/toronto-emotional-speech-set-tess. [Accessed: 08- mar-2021].

[28] “Toronto emotional speech set (TESS),” [Online] Available: https://tspace.library.utoronto.ca/handle/1807/24487. [Accessed: 17-feb-2021].

[29] C. Houwei, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, “CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset,” IEEE Transactions on Affective Computing, vol. 5, pp. 377-390, Jan 2014.

[30] “CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset),” [Online] Available: https://github.com/CheyneyComputerScience/CREMA-D. [Accessed: 18-Oct-2020].

[31] K. H. Lee, H. K. Choi, B. T. Jang and D. H. Kim, “A study on speech emotion recognition using a deep neural network,” International Conference on Information and Communication Technology Convergence (ICTC), pp. 1162-1165, 2019.

[32] F. A. Shaqra, R. Duwairi and M. Al-Ayyoub, “Recognizing emotion from speech based on age and gender using hierarchical models,” Procedia Computer Science, vol. 171, pp. 37-44, 2020.

[33] “The number of hidden layers,” [Online] Available: https://www.heatonresearch.com/2017/06/01/hidden-layers.html. [Accessed: 22- Feb-2021].

** **Link to the github repository** - **https://github.com/AAKANKSHACHOUHAN/Mulitmodal-ER**