# Audio Emotion Classification using Deep-learning

**Vaibhav Sundharam**
Department of Computer Engineering
Virginia Polytechnic Institute and State University
vaibhavsundharam@vt.edu

**Shagun Johari**
Department of Computer Engineering
Virginia Polytechnic Institute and State University
shagunjohari@vt.edu

## Abstract

Human emotions are one of the strongest ways of communication. Even if a person doesn't understand a language, he or she can very well understand the emotions delivered by an individual. In other words, emotions are universal. The idea behind the project is to develop a Speech Emotion Classifier using deep-learning to correctly classify a human's different emotions, such as, neutral speech, angry speech, surprised speech, etc. We have deployed three different network architectures namely 1-D CNN, LSTMs and Transformers[1] to carryout the classification task. Also, we have used two different feature extraction methodologies (MFCC & Mel Spectrograms) to capture the features in a given voice signal and compared the two in their ability to produce high quality results in deep-learning models.

## 1 Introduction

Humans communicate not only through their words but also their body language, facial expressions, and cadence. Psychologists have been studying human emotions for many decades now. Analyzing emotions can help in the treatment of mental disorders and can play a major role in psychotherapy and the medical industry [1]. Apart from this, emotion recognition can help in various industries, such as marketing and advertising. For example, customer service departments can use emotion recognition to gauge the feelings of an individual and use that to improve their services [2, 3]. Emotion recognition has many applications in software engineering processes as well. One simple example is using emotion recognition on a smart-device, where the device could play uplifting songs on detecting a sad mood. It is also playing a vital role in human-machine interfaces (HMI) [4]. Speech is already being used in HMI and adding emotion recognition to it will help to bridge the gap between humans and machines. Furthermore, using emotion recognition can produce better quality systems with higher usability [5]. Hence, this can be the next step in making an *aware* AI that understands the emotional states expressed by the human subject and respond accordingly [6].

Emotions can be extracted by an individual's different modalities like facial expressions, body language, voice, and EEG signals [7]. In this project, our objective was to create an audio emotion classifier based on artificial neural networks capable of differentiating between eight different human emotions such as happy, sad, angry, fearful, disgust, surprise, calm and neutral by using an audio signal as the input. We have used Mel Frequency Cepstral Coefficients (MFCC) and Mel Spectrograms (MS) for feature extraction from the audio signal and have created three different models to perform the classification. The first model is a simple 1D convolution neural network whereas, the other two

[1] https://github.com/vaibhavsundharam/Speech-Emotion-Analysis

networks are based on LSTMs and Transformers. The structure of the report is as follows. Section II presents a brief description on some of the previous research work done in this domain followed by section III that details the approach we took to develop the classifier. In section IV we discuss the results obtained from our experiments followed by the conclusion.

## 2 Related Work

An extensive amount of work has been done in emotion recognition. El Ayadi et al.[8] and Anagnostopoulos et al.[9] are two very detailed survey papers talking about the study of speech recognition and their various applications. They have discussed, in detail, what should be the correct choice of features for representing speech, what the proper design of the system should be, and how to correctly prepare the database for speech recognition. Zhang et al.[10] have made a shared audio recognition system for speech and singing (rather than making two different systems) adopting the RAVDESS dataset. They made three different models using support vector machines (SVM): a simple model, and two hierarchical models. In [11], the authors have also made an audio recognition system using the RAVDESS dataset to differentiate between singing and speech but the difference is that they have used multi-task learning and trained four classification tasks, namely, speech-male, speech-female, song-male, and song-female. Z. Zhao et al.[12] have used a combination of attention-based BLSTM and fully convolutional networks in their speech emotion recognition system using CHEAVD and IEMOCAP datasets. They proved that using the attention mechanism, with BLSTM and a deep neural network structure, can help to boost the performance of an emotion recognition system. Lastly, in [13] the authors have used IEMOCAP and MSP-IMPROV datasets for implementing emotion detection along with speaker and gender detection. They have used progressive neural networks and compared their model to the traditional deep learning methods of pre-training and fine-tuning (PT/FT) as well as observed transfer learning between emotion, gender, and speech recognition.

## 3 Approach

### 3.1 Datasets

There are many datasets available for audio emotion recognition. Many of them consist of trained actors who enunciate short phrases in various emotions. Some commonly used datasets for emotion classification are SAVEE (Surrey Audio-Visual Expressed Emotion) [14] and TESS (Toronto Emotional Speech) [15]. SAVEE dataset consists of only four male speakers whereas, the TESS dataset consists of only two female speakers. It is important to note that both male and female voices are required to design a novel emotion classification neural network as we do not want the network to be biased towards only one gender. The issue with these two datasets is that even though they have good quality audio files, both have very limited in terms of variety of voice actors. We wanted a dataset that had a larger number of voice actors as compared to the aforementioned datasets. Hence, we decided to use the RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) [16] dataset. RAVDESS is one of the most frequently used datasets for this task as it consists of 12 female and 12 male speakers who are showcasing eight different emotions, namely, *neutral, calm, happy, sad, angry, fearful, disgust*, and *surprised*. Each speaker is enunciating two sentences ("Kids are talking by the door" and "Dogs are sitting by the door") in eight different emotions and with two different intensities (strong and normal) except for the *neutral* emotion, which only has a *normal* intensity. Hence, each actor has 60 enunciations making a total of 1,440 audio files.

### 3.2 Data pre-processing

The classification network should be trained separately on male and female voices. This is because males and females have very different physiologic and acoustic features [17, 18]. For example, a female voice has a higher pitch as compared to a male as depicted in Figure 1. Hence, we separated male and female voices in the dataset. Each subset is then split into training, validation, and testing set in the ratio of 8:1:1.

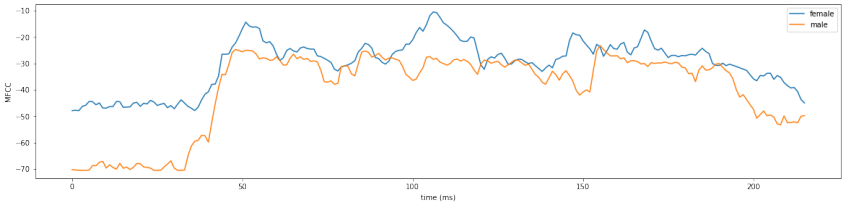
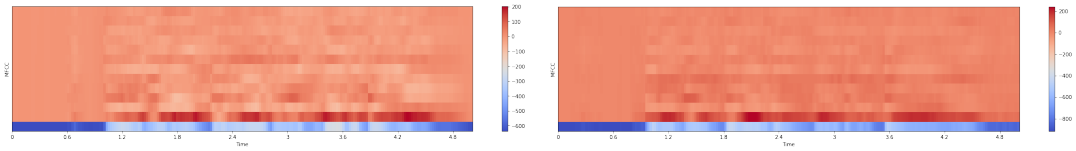



Figure (1) Difference between the pitch of a male and a female voice actor enunciating the same dialog with the same emotion (*angry*)

### 3.2.1 Feature Extraction

The first approach for extracting features from the voice signal was based on Fourier transforms. Unfortunately, this approach was flawed as Fourier transforms are not useful in showcasing how a human individual perceives sound. Hence, we decided to use Mel Frequency Cepstral Coefficients (MFCC) [19] and Mel Spectrograms [20]. Generally, we take the Fourier transform to decompose a time-domain signal into its frequency components. But, for calculating MFCC, the cosine transformation of the log of the magnitude of the Fourier spectrum is taken. Hence, the signal is neither in the time domain nor in the frequency domain, rather, it is in the quefrency domain [21] with its spectrum termed as cepstrum. Cepstrum gives us information about the rate of change of spectral bands. The vocal tract (including tongue, teeth, etc.) determines the sound generated and, the MFCC accurately represents the envelope of the time power spectrum of a speech signal which is a representation of the vocal tract. Figure 2a and Figure 2b represent the MFCC plots of a male and a female speaker respectively showcasing *angry* emotion. We can see that both plots look different hence affirming that males and females have different acoustic features [17, 18].

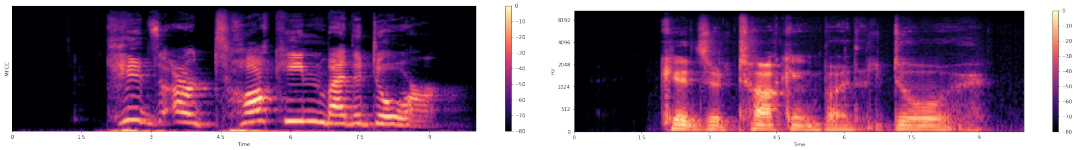
$$Mel(f) = 2595 \times \log(1 + \frac{f}{700}) \tag{1}$$



(a) Female (b) Male

Figure (2) MFCC plots of a male and a female speaker showcasing *angry* emotion

Mel spectrograms are calculated by taking the log of the frequency components obtained after the FFT of an audio signal and converting the amplitude of the audio signal to decibels. The frequency is mapped to the Mel scale using Eq. 1 [21]. Figure 3 shows the Mel spectrograms for a male and female audio showcasing the *angry* emotion. The difference between the acoustic features of a male and a female can also be seen from the aforementioned figure.



(a) Female (b) Male

Figure (3) Spectrograms of a male and female speaker showcasing 'angry' emotion

### 3.2.2 Data Augmentation

Training an artificial neural network usually requires a sufficiently large amount of data. To circumvent this issue, various augmentations can be applied to synthetically generate new data points. For

example, for audio data, augmentations include adding noise, shifting the audio, changing the pitch, stretching the audio, etc. In our implementation, we have added white Gaussian noise to the voice signals. We have synthesized two noise signals sampled from a normal distribution and added them to the voice signal thereby, increasing the number of training examples from 576 to 1728 for both male and female data subsets. This augmentation not only increased the number of data points but also helped in making the model much more robust.

### 3.3 Model Architectures

We implemented three different neural network architectures to perform the classification task. The first architecture is the baseline model and is built using 1-D convolution neural networks. The second and the third model that we implemented use Transformers and Long Short Term Memory (LSTM) architectures at their core. In the following section, we will deep dive into each of the network architectures.

#### 3.3.1 Baseline model

Convolution neural networks (CNNs) have shown their prowess in various machine learning applications ranging from computer vision tasks to natural language processing (NLP). In today's world, there is no domain left untouched by CNNs. Hence, our first obvious choice was to use 1-D CNNs to build the baseline model. Figure 4 depicts the block diagram of the baseline model we implemented. Each *Conv 1D Block* consists of a conv1d layer followed by batch normalization and relu activation. Dropout layers were introduced at various locations throughout the model to circumvent the issue of overfitting.

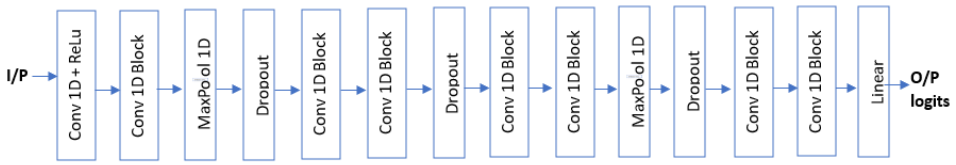



Figure (4) Block diagram of the baseline model

#### 3.3.2 LSTM Model

Recurrent neural networks (RNNs) were one of the first successful models deployed for natural language processing applications as they allow information to persist. This is useful, especially for sequential data like speech and language, where the prediction of the next word in the sentence/speech depends upon the phrases that came before. At each level inside an RNN, information from the previous layers gets transferred to the next layer. This information transfer creates a dependence between subsequent layers inside the network. But, RNNs fail to learn long-term dependencies, i.e. they perform poorly when the gap between relevant information and the place where it is required becomes large. Long-Short Term Memory (LSTM) [25, 26] networks with attention mechanisms were introduced to address some of the drawbacks of RNNs. LSTMs are also sequential models. But, unlike RNNs, they can selectively store and discard information that is important and not so important. In our implementation, we have used a network architecture inspired by bidirectional LSTMs[2] to perform the classification task. The network architecture of the model is depicted in Figure 5.

#### 3.3.3 Transformer Model

Despite the success of LSTM in NLP tasks, it has its own set of problems. Firstly, LSTM networks are plagued by long-term dependence problem, similar to RNNs. Secondly, parallelization is not possible in RNNs and LSTMs because to process sequential data such as a sentence, word by word processing is required. To address the problem of parallelization, Transformers [22] were introduced. Transformers consist of an encoder, a decoder and have made use of multi-head attention layers and

[2]Ref: https://github.com/Data-Science-kosta/Speech-Emotion-Classification-with-PyTorch

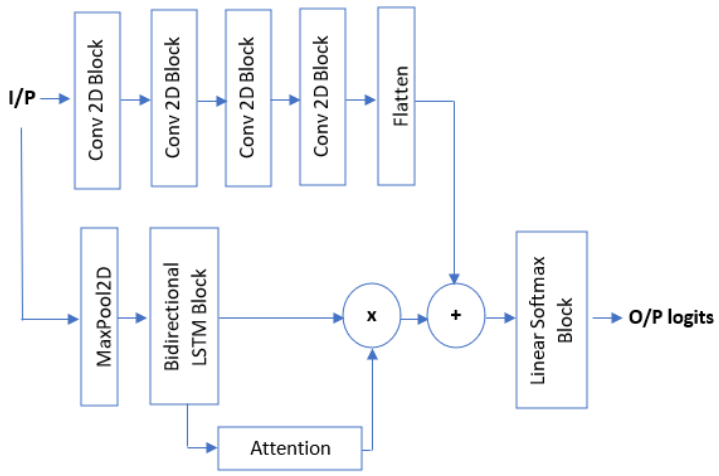



Figure (5) Block diagram of the LSTM based model

rely entirely on attention mechanism to extract global dependencies between input and outputs. Thus, our third implementation is based on Transformers[3] the block diagram of which is shown in Figure 6.

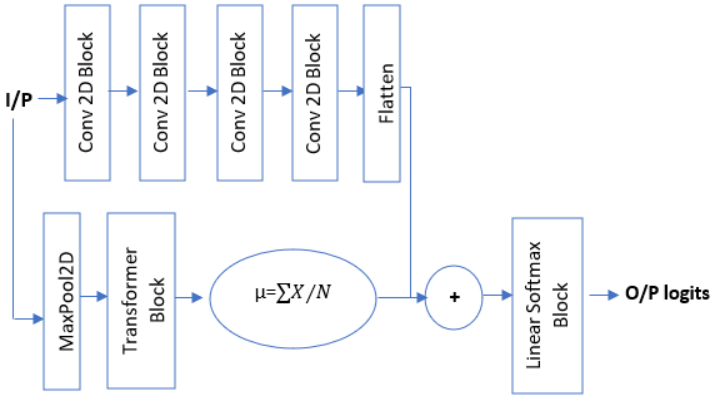



Figure (6) Block diagram of the Transformer based model

### 3.4 Model Parameters

In this section, we will discuss the performance metrics and the loss function used in the implementation, as well as the set of hyper parameters used to train our models. Additionally, we will also discuss parameters chosen for extracting features using MFCC and Mel spectrogram.

#### 3.4.1 Cost function, Performance Metric and Optimizer

In the project, we have employed cross-entropy loss. This criterion combines LogSoftmax and NLLLoss. Cross entropy loss is a parameter that is used in classification models whose output is a probability ranging from zero to one. As the predicted probability of a class diverges from the ground truth, cross-entropy loss increases. According to [23], the loss can be defined as in Eq. 2.

$$L(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right) = -x[class] + \log(\sum_j \exp(x[j]) \quad (2)$$

[3]Ref: https://github.com/Data-Science-kosta/Speech-Emotion-Classification-with-PyTorch

Table (1) Model vs. loss and accuracy

| Architecture | Feature | Male | | Female | |
|---|---|---|---|---|---|
| | | Loss | Accuracy (%) | Loss | Accuracy (%) |
| Baseline | MFCC | 1.347 | 53.0 | 1.320 | 54.0 |
| LSTM | MFCC | 1.868 | 36.0 | 1.808 | 29.0 |
| | Mel Spectrogram | 1.082 | 60.0 | 1.361 | 44.0 |
| Transformers | MFCC | 2.159 | 19.0 | 1.194 | 19.0 |
| | Mel Spectrogram | 0.961 | 74.0 | 0.902 | 76.0 |

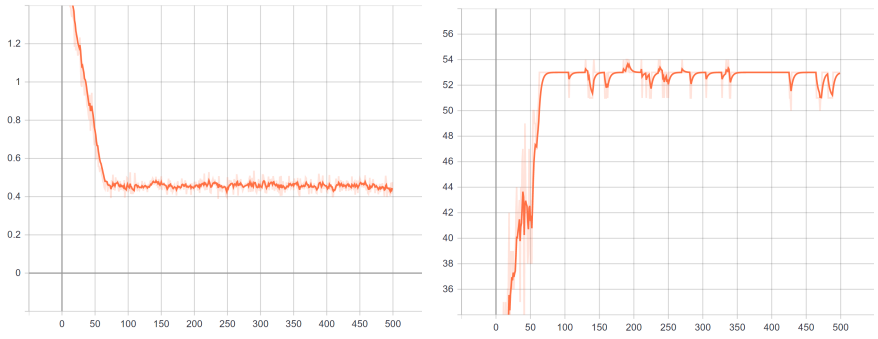
Accuracy is chosen as the performance metric that measures how close a predicted value is to the ground truth. In this classification problem, it was imperative to classify the emotion in the voice with confidence and not just the accuracy of our predictions. That is why we chose cross-entropy loss as our evaluation matrix. The neural networks that are employed in this project try to minimize this loss while training. Adam [24] was used as the optimizer with a weight decay of 1e-4 and a learning rate of 1e-3 with the default values of $\alpha$ and $\beta$. Also, the batch size was set to 150.

### 3.4.2 Other Parameters

The sampling rate for all the audio files was 48000 Hz with a duration set to 3 seconds. Voice signals with a length shorter than 3 seconds were zero-padded to make the length of each voice signal the same. For calculating the Mel spectrogram, an FFT window of 1024 and a hop length of 256 was taken. On the other hand, 13 MFCC values were extracted for a given voice signal with type-2 discrete cosine transform. For calculating the noise signal, we have taken the target SNR values in the range of 15dB - 30 dB.

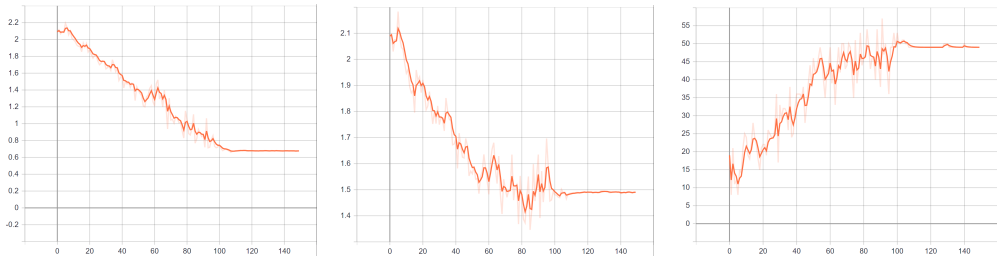
## 4 Experiments and Results

In the following section, we will discuss in detail the experiments performed, as well as the results observed. Table. 1 presents the testing accuracy and loss for the three implemented models rounded to the nearest real number. From the table, it is evident that the 1-D baseline model, trained on normalized MFCC feature set for 500 epochs, gives a decent performance with accuracy in the ballpark of 54% for both male and female subjects. Even though the results are not highly impressive, they are still better than the naive approach that yields an accuracy of only 12.5%. Also, the same neural network is giving a good performance for both male and female data subsets. Figure 7a and Figure 7b represents the training loss vs. epoch and validation accuracy vs. epoch graphs for the baseline model trained on male voice data. Figure 8a, 8b and 8c represent the training loss, validation loss and accuracy plots for the LSTM model trained on male MFCC features. Also, from Table 1 it can be seen that the performance is subpar with the highest accuracy of 36% achieved in the case of male subjects whereas, for female subjects, a maximum accuracy of only 29% was reached. With MFCC as the feature extractor, the LSTM model is under performing as compared to our baseline model. Performance seems to improve when we use LSTM with MS features (Figure. 8d, 8e and 8f). But, this is only the case with male voice data (test accuracy: 60%) as there is a substantial decrease in the model's performance with an accuracy of only 44% when female voices are tested on the trained model. It should be noted that the same model is used for the training of male and female voices (i.e. model architecture and parameters are kept constant for both datasets). From the experiment, it is observed that the same LSTM models cannot be used for both male and female voice data. Hence, different models tuned on different hyperparameters should be used to improve performance. Figure 8g to 8l represent the plots obtained for the Transformers based networks trained on MFCC and Mel Spectrogram features. From Table 1 it can be seen that Transformers, with Mel Spectrogram features as inputs, have outperformed all the aforementioned architectures with a testing accuracy of 74% on male voices and 76% on female voices. On the other hand, the lowest performance on male and female test data was achieved when Transformers were used alongside MFCC features.

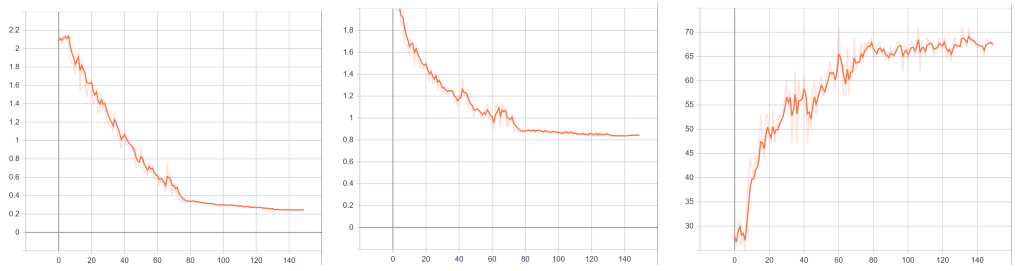


(a) Training Loss Plot (male)- x: epochs, y: loss (b) Val. Accuracy Plot (male)- x: epochs, y: accuracy

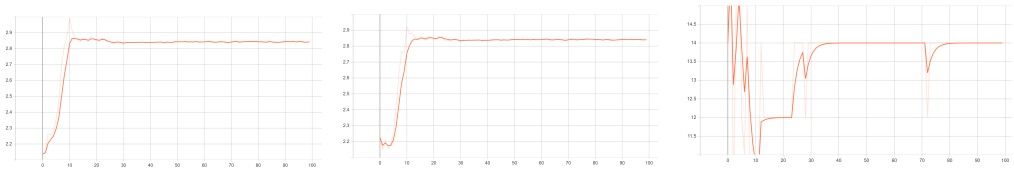
Figure (7) Training Loss and Accuracy Plots (male) for the baseline model



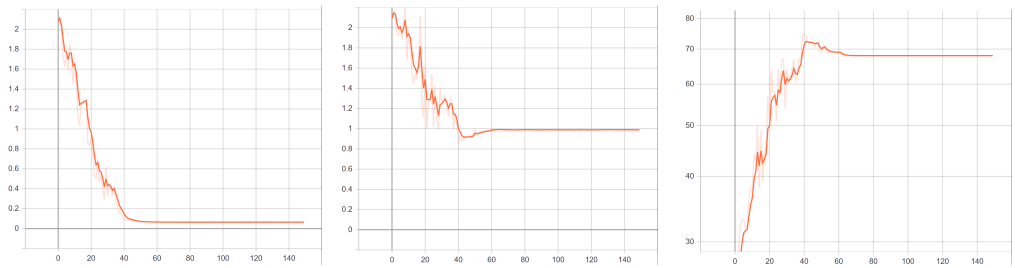
(a) Training Loss: LSTM-MFCC (b) Validation Loss: LSTM-MFCC (c) Validation Accuracy: LSTM-MFCC



(d) Training Loss: LSTM-MS (e) Validation Loss: LSTM-MS (f) Validation Accuracy: LSTM-MS



(g) Training Loss: Transformer-MFCC (h) Validation Loss: Transformer-MFCC (i) Validation Acc.: Transformer-MFCC



(j) Training Loss: Transformer-MS (k) Validation Loss: Transformer-MS (l) Validation Acc.: Transformer-MS

Figure (8) Training/Validation loss (male) and accuracy vs. epochs (male) for various NN architectures- x: epochs, y: loss/accuracy

## 5 Conclusion

In this project, we attempted different deep neural network architectures to implement audio emotion recognition using cepstral based features, namely, Mel spectrograms and Mel frequency cepstral coefficients. From our experiments, we observed that when Mel Spectrograms were used as the feature extractors, there was an overall increase in the performance of the models we implemented. Out of the three models, the Transformer based model had the best performance giving 74% and 76% accuracy for the male and female test sets respectively. On the other hand, the LSTM based model gave results of only 60% on the male dataset. The same model trained on the same hyper parameters, though, gave only 44% accuracy for the female dataset. This high variance in the testing accuracies is not observed for the Transformer based model. For future works it is important that apart from gender, other factors, such as age, are also considered when implementing emotion recognition as age also affects the acoustic and physiologic features of a human's voice. Creating a separate model for different gender and age categories can lead to a better performance as compared to using only one classifier. Furthermore, certain scenarios could exist where a mixture of voices are present or noise is present in the background. In such cases, Independent Component Analysis (ICA) can be employed to separate out the speech signals before sending them into the speech emotion recognition system. Hence, we have also implemented a simple ICA system[4].

## References


[1] Abu Shaqra, Ftoon & Duwairi, Rehab & Al-Ayyoub, Mahmoud. (2019). Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models. *Procedia Computer Science.* 151. 37-44. 10.1016/j.procs.2019.04.009.

[2] Petrushin, Valery. "Emotion in speech: Recognition and application to call centers." *Proceedings of Artificial Neural Networks in Engineering.*, Vol. 710. 1999.

[3] Petrushin, Valery A. "Emotion recognition in speech signal: experimental study, development, and application." *Sixth International Conference on Spoken Language Processing*, 2000.

[4] Meng, Hao & Yan, Tianhao & Yuan, Fei & Wei, Hongwei. (2019). Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. *IEEE Access.* PP. 1-1. 10.1109/ACCESS.2019.2938007

[5] Kolakowska, Agata, et al. "Emotion recognition and its application in software engineering." *Human System Interaction (HSI)*, 2013 *The 6th International Conference on. IEEE*, 2013.

[6] Cen, Ling & Wu, Fei & Yu, Zhu & Hu, Fengye. (2016). A Real-Time Speech Emotion Recognition System and its Application in Online Learning. 10.1016/B978-0-12-801856-9.00002-5.

[7] Brave, Scott, & Clifford Nass. "Emotion in human–computer interaction." *Human-Computer Interaction* **(2003)**: 53

[8] El Ayadi, Moataz, Mohamed S.Kamel, & Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition* **44.3** (2011): 572-587.

[9] Anagnostopoulos, Christos-Nikolaos, Theodoros Iliou, & Ioannis Giannoukos. "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011." *Artificial Intelligence Review* **43.2** (2015): 155-177.

[10] Zhang, Biqiao & Essl, Georg & Mower Provost, Emily. (2015). Recognizing Emotion from Singing and Speaking Using Shared Models. 10.1109/ACII.2015.7344563.

[11] Zhang, Biqiao & Mower Provost, Emily & Essi, Georg. (2016). Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. 5805-5809. 10.1109/ICASSP.2016.7472790.

[12] Z. Zhao, Y. Zheng, et al, "Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018.

[13] Gideon, John & Khorram, Soheil & Aldeneh, Zakaria & Dimitriadis, Dimitrios & Mower Provost, Emily. (2017). Progressive Neural Networks for Transfer Learning in Emotion Recognition. 1098-1102. 10.21437/Interspeech.2017-1637.

[14] Jackson, Philip & ul Haq, Sana. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database.

[15] Pichora-Fuller, M. Kathleen; Dupuis, Kate, 2020, "Toronto emotional speech set (TESS)" database.


[4]Ref: https://github.com/corymaklin/ida

[16] Livingstone SR, Russo FA (2018) 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** 13(5): e0196391.

[17] Titze, Ingo R. "Physiologic and acoustic differences between male and female voices." *The Journal of the Acoustical Society of America* **85.4** (1989): 1699-1707.

[18] Klatt, Dennis H. & Laura C. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *The Journal of the Acoustical Society of America* **87.2** (1990): 820-857.

[19] Sreeram, Lalitha & Geyasruti, D. & Narayanan, Ramachandran & M, Shravani. (2015). Emotion Detection Using MFCC and Cepstrum Features. *Procedia Computer Science* 70. 29-35. 10.1016/j.procs.2015.10.020.

[20] Meng, Hao & Yan, Tianhao & Yuan, Fei & Wei, Hongwei. (2019). Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. *IEEE Access* PP. 1-1. 10.1109/ACCESS.2019.2938007.

[21] B.P. Bogert, M.J.R. Healy, and J.W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Time Series Analysis*, M. Rosenblatt, Ed., 1963, ch. 15,pp. 209–243

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *In Advances in Neural Information Processing Systems*, pages 6000-6010

[23] *PyTorch Documentation* TORCH.NN url:https://pytorch.org/docs/stable/nn.html #torch.nn.CrossEntropyLoss

[24] Kingma, D. P. & Ba, J. (2014), 'Adam: A Method for Stochastic Optimization' , cite arxiv:1412.6980 Comment: *Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego*, 2015 .

[25] Schuster, Mike & Paliwal, Kuldip. (1997). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions* on. 45. 2673 - 2681. 10.1109/78.650093.

[26] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. *Neural computation.* 9. 1735-80. 10.1162/neco.1997.9.8.1735.