Fully-connected and dropout layers in Speech Emotion Recognition

Ryan Pankratz
Department of Computer Science
University of Toronto Mississauga
Mississauga, ON L5L 1C6

Srisaran Chandramouli
Department of Computer Science
University of Toronto Mississauga
Mississauga, ON L5L 1C6

Dishant Tawade
Department of Computer Science
University of Toronto Mississauga
Mississauga, ON L5L 1C6

Nicholas Ospina
Department of Computer Science
University of Toronto Mississauga
Mississauga, ON L5L 1C6

Abstract

This project aims to develop a Speech Emotion Recognition (SER) system capable of accurately classifying emotional states from spoken audio. These models have many real world uses, such as interpreting emotions through call audio for call centers and virtual assistant, as well as improved therapy machines to help peoples mental health [8][9]. The task involves processing raw audio inputs to extract meaningful features and applying deep learning models to classify emotions such as neutral, calm, happy, sad, angry, fearful, disgust, and surprised. We make use of the RAVDESS dataset to train and test our model on emotional speech [6]. Our approach involves using a deep learning model that combines Convolutional Neural Networks (CNNs) for spatial feature extraction from spectrograms and Bi-Directional Long Short-Term Memory (Bi-LSTM) networks for capturing important features within the audio sequence. These are then fed into a multi-layer perceptron that classifies these extracted features to one of the emotions. Because of the increase in popularity of ensemble models to combat overfitting [8], our model uses dropout layers in the classifier to simulate an ensemble architecture neural network. We then tested the effectiveness of this approach against an identical model without dropout layers. Our results show that there is little performance to be gained from utilizing dropout layers in this way, and that alternate techniques, such as advanced data augmentation, should be relied on to better generalize the model.

Our Github repository can be found here: https://github.com/ryan-pankratz/CSC413-SER-Model

1 Introduction

Over the last decade, the SER field has become an interesting and challenging topic in human behaviour analysis research [9]. The applications of SER models are vast, including interpreting emotions through call audio to enhance customer service experience during virtual assistance and improved mental health interactions and serve as therapy tools [8][9].

The motivation for this work stems from the growing demand for emotionally intelligent systems capable of understanding users' needs and enhancing their experience. Despite its potential, SER presents several challenges, including variability in speech patterns, background noise, and subjective

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

nature of emotions. By addressing these challenges, this project aims to contribute to the development of existing SER systems. This paper addresses these challenges through innovative approaches that leverage advancements in signal processing, feature extraction, and deep learning techniques.

The problem we aim to solve is the accurate classification of emotions from raw audio data. The input to our model is audio recordings containing speech from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which are then pre-processed and transformed into feature representations such as Mel-Frequency Cepstral Coefficients (MFCCs), Root Mean Squared Error (RMSE) and spectral features such as Zero Crossing Rate, Mel Spectrogram, and Chroma Short Time Fourier Transform (STFT).

The proposed model uses neural network architectures, specifically CNNs and Bidirectional Long Short-Term Memory (Bi-LSTM) networks to extract spatial features and capture temporal dependencies, making the hybrid approach reasonable for this problem. These features are then fed to the classifier portion of the model. In this paper, we test the effectiveness of dropout layers in SER to reduce overfitting by testing against an identical model without them. Our results show that there is little to no performance gain between the two models, and we discuss alternative methods to reduce overfitting besides dropout layers. From our analysis, the best way to generalize the model is to include multiple datasets, and to use advanced data augmentation techniques to greatly increase the variability of the training set.

2 Backround and related work

In SER problems, it is difficult to derive generalized features from audio that help discriminate between different emotions for different speakers. This is because speakers' vocal cues often vary by gender, cultural linguistics, and age, making it especially difficult to determine what features are correlated with emotion [8]. Because of this, much research has gone into different ways of processing the audio for emotion classification. Much of the recent success in SER classification has used speech features and spectrogram analysis [8]. This is done by framing the audio into a sequence of fixed-length sections. As the speech signals stay invariant briefly, a frame size of 10 to 20 ms is recommended [13]. After framing, the different frames could have discontinuity between each other. Here, many researchers use a Windowing function to minimize the discontinuities between each frame by adding an overlapped section [8]. Among 32 studies that mentioned their windowing specifics, 68.7% of them use a window size of 25ms with an overlap of (1/3)rd of frame size[8]. For SER models, the most effective windowing function is the Hamming Window [2], which is given by the formula: $w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1})$ [7][13], where $0 \leq n \leq N$ is a frame number, and N is the number of frames in an audio segment.

Because of the difficulty in trying to derive features from audio input, there have been many audio transformations that have become popular choices for SER [8]. The first of these features is a Fourier transformation of the audio. Researchers use Fourier transformations to read the frequency information from the audio frames [1][8]. However, since traditional Fourier transformations lose temporal frequency information, some researchers have instead used a more advanced Fourier transformation called a Short-Time Fourier Transformation (STFT) to maintain the temporal-frequency information [4]. The STFT is also used to calculate many of the other more advanced features. This includes the mel-spectrogram, which is an image that captures the changes in amplitude with respect to time. Recent researchers make use of the Mel-spectrogram to read changes in the amplitude of speech over time, getting very strong results with a variety of models [12][17]. In addition to these, the Mel-Frequency-Cepstral coefficients (MFCCs) are one of the most widely used features in SER models [8]. They are computed by taking the Hamming-windowed audio, mapping it to the mel-scale, calculating the logarithms of the energy in the chosen frequency band, and finally applying inverse Fourier transform to transform to the cepstral domain to get MFCCs [2]. Additionally, research has shown that zero-crossing rate is useful feature for endpoint detection and mute detection in the audio. It is a measure of how often the audio waveform crosses the zero-axis in each frame [8][17]. In addition to these features, the RMSE is often used

In addition to the features, much work has been done showing the improvements of data augmentation in SER. One such improvement is the DAARIP (Data Augmentation Algorithm based on the Retinal Imaging Principle) algorithm, which generates the spectrogram of the audio file using STFT, then generates a set of images of different sizes using the retinal imaging principle and convex lens, and finally converts them to the required size [11]. Using this augmentation technique, SER researchers have seen incredible performance increases on their model.

There are many different model architectures that researchers use for computing SER classification. Early SER models were often made using SVM classifiers, Gaussian Mixture models, or CNNs [8]. More recently, researchers have found success through using a combination of CNNs and Bi-directional LSTMs, or CNNs with attention in order to read features from the mel-spectrogram while also taking into account the sequence nature of audio and changes in pitch over time [3][8][10][16]. Much of the more advanced SER success has been through deep neural networks (DNNs), where they extract very deep speech features from the mel-spectrogram, and use them as input to the classifier part of the model [8][11].

3 Data

The data required to create an SER model consists of audio recordings labeled with emotional states. The model uses the data from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This dataset contains 1440 unique files, where 24 actors (12 male, 12 female) perform two lexically matched statements in a neutral North American accent with 60 trials each [6]. Audio files come with 8 different expressions which are neutral, calm, happy, sad, angry, fearful, surprise, and disgust. Each of these expressions are produced at two levels of emotional intensity: normal and strong except the neutral class [6].

These audio files are pre-processed with data cleaning to fine-tune the data before it is fed into the classification model. RAVDESS is initially stored in 48 kHz sampling rate which is downsampled to a rate of 16kHz stored in PCM format. This sampling rate has been shown to provide an optimal emotion classification [13]. A frame size of 20ms is used based on an optimal size between 10-20ms [13]. Furthermore, the WebRTC Voice Activity Detector (VAD) library developed by Google, is used to reduce noise and silence within the data and avoid unpredictable fluctuation to improve model performance. This way, the data was focused solely on the speech and not the silence.

This data is split into training, validation, and test sets. A random seed is used to generate a reproducible split. The split is 60% training, 15% validation, and the remaining 25% testing. The data within the training set is then augmented to avoid overfitting. This is done by pitching the original audio both up and down, and adding a small amount of Gaussian noise. The final training set includes 3456 data points.

For the feature extraction, we made use of the Librosa library to derive the Chroma-STFTs, MFCCs, Zero Crossing Rate, Root Mean Square Error, and the mel-spectrogram. These features are chosen due to their proven success in many other SER models [8]. Since our input is a sequence, the second dimension of the output is of variable length K. These features all have the same variable length K, and are stacked vertically into a matrix with a final size of $162 \times K$. We do this in order to keep the information from each time step. Because the dimension K is different for each input audio file, each vector in the batch is padded with zeros to make it of shape $B \times 162 \times K_{max}$, where K_{max} is the maximum length of an input sequence in the batch, and B is the batch size.

4 Model architecture

Similar to other SER models, our model has two parts. The first part of the model extracts features from the raw audio, and the second part uses these features to classify the speakers emotion. Our model derives the important features processed from the mel-spectrograms, MFCCs, RMSEs, Zero Crossing rates, and STFTs for input into the classifier. For the deep feature extraction, many researchers have reported success using convolution layers with max-pooling like what is used in image processing [5]. Furthermore, since each frame from the dataset is quite large, using fully connected layers would likely be too large of a model. Therefore, our model uses two convolution layers to capture the deep features from the input spectrogram frames. Since our input data is of size 162 X K, where K is a variable length of the sequence, the first convolution layer has 162 input channels. From this, we wanted to reduce the dimensionality of the output channels, so we used 60 output channels for the first layer, with a kernel of size 5, and a stride of one. For our max pool layers, our model got the best validation accuracy with a pooling size of 5. For the second and final convolution layer, our tuning showed that the best kernel size was 3, and we used 40 output channels to reduce the dimensionality even further.

In addition to these initial convolution layers, research has shown that concatenating recurrent Bi-directional LSTM layers after the convolution layers outperforms other models that only use convolution [5]. This is further supported by the fact that the input data is a sequential audio signal and therefore lends itself well to a recurrent architecture. Thus, we used an LSTM layer to consolidate the important features from each frame into a flattened vector [5]. For our LSTM, we use have the input size of 40 from our output channels, and the output size is set to 80. The consolidated the variable sequence lengths to one dimension. This was done by concatenating both the average of each output of the LSTM, and the maximum of each output of all the outputs at each time-step (This gives us a vector of size 160).

For the the classifier part of our SER model, there has been an increase in reported success on using ensemble architecture models [8]. Therefore, we tested utilizing fully connected layers with dropout layers in-between to simulate having an ensemble-like model to reduce overfitting. However, it has been shown that utilizing dropout layers with convolution layers only, tends to not perform well [15]. As a result, the last three layers of our model are fully connected layers to test our hypothesis. For the first part of our classifier model, we use a batch normalization layer in order to normalize the inputs from the LSTM. After this, we use 3 fully connected layers, with a ReLU activation and dropout layer between them (except for the last fully connected layer, which is our output). From our parameter tuning, we found that a dropout percentage of 0.25 gave the best for validation accuracy. In addition to this, we tested using both ReLU and LeakyReLU activations between the layers, and found that ReLU layers gave better validation performance and more stable gradient updates. The first fully connected layer of our classifier has an output size of 60, the second has an ouput size of 50, and our final layer has an output size of 8 (the number of classes). We chose the sizes in order to over time reduce the dimensionality of the input from the LSTM before finally classifying. In order to test our hypothesis, we also created a model that uses the exact same architecture, except that it does not use dropout layers. This way, the dropout layers act as our independant variables to verify whether they were successful at reducing overfitting.

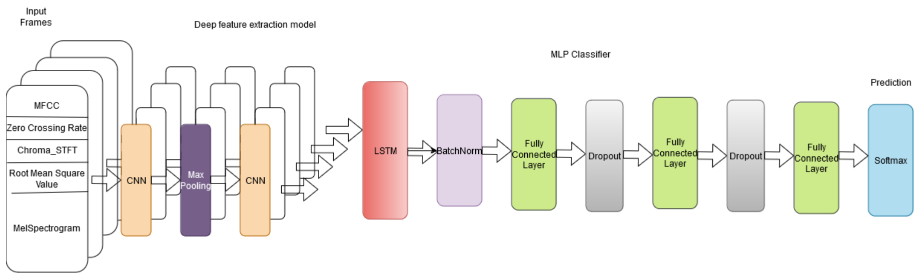

Figure 1: Here is our model architecture with the input spectrogram frames from one audio sample. The frames are inputted to the CNNs to extract the deep features of the spectrograms, MFCCs, zero crossing rates, etc, and then consolidated by the following LSTM. After this, the max and average values from each computation of the LSTM are used as input to the classifier.

5 Results

We trained both our models for 80 epochs with a batch size of 256. First, we will discuss the results of training our model with dropout layers, and then compare it to our model without them. Our model with dropout layers ended with a training accuracy of 94% and a validation accuracy of 48%. When testing this model on our test set, it achieved an accuracy of 58%. Observing the training curve, it is clear that the model became overfitted to the training set even when using dropout layers. This can be seen in how our training accuracy increases steadily, while the validation curve levels out around 40-50% accuracy. Though, looking at the training curve, the curve is quite smooth while moving upwards for the training set, which suggests that the chosen parameters and learning rate worked well with our features.

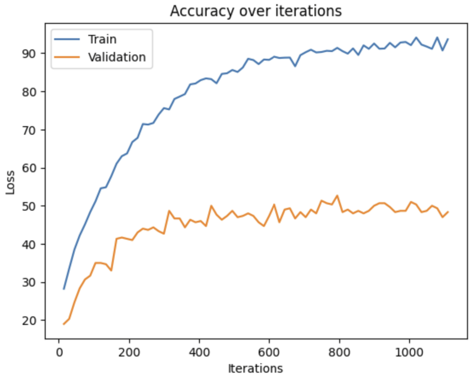

Figure 2: This figure shows the training curve from our model with dropout layers. We can see how the model's training accuracy steadily increases to 94%, but the validation accuracy plateaus around 40-50%.

In addition to the general testing accuracy, we also want to investigate the false classifications and classification rates of the different emotions in our model. In the following figure, we have the confusion matrix showing how the different emotions were classified:

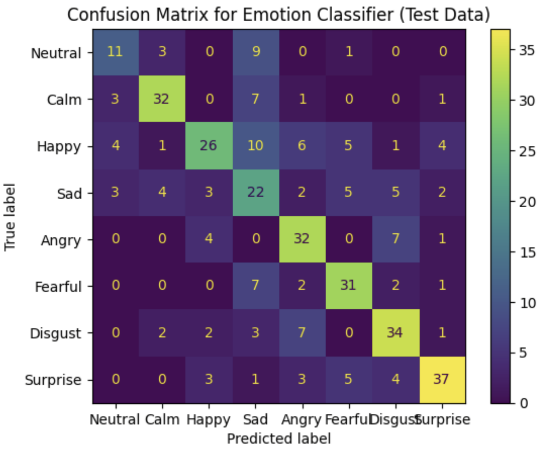

Figure 3: The confusion matrix of our model on the testing set. Most of the emotions are classified correctly, but some pairs often are mistaken for another (as seen with sadness and happiness, or anger and disgust)

Interestingly, we can see that the most confused emotions were happiness and sadness, and anger and disgust. This is likely due to similarities in pitch and/or amplitude that our model is having trouble differentiating. In addition, the neutral emotion class often seems to be mistaken for sadness or calmness as well. This is likely because the RAVDESS dataset has less audio files with the neutral emotion than any other emotion. Therefore, it makes sense that our model has more difficulty classifying it than other models.

In addition to testing our model that includes dropout layers, we also tested it against our model with the same architecture and hyper parameters, but without the dropout layers. We did this to show whether the dropout layers were successful in minimizing overfitting. Again, we trained it by running 80 epochs with a batch size of 256. From this, our model obtained a training accuracy of 96%, a validation accuracy of 47%, and a test accuracy of 57%. This shows that, while our model that included dropout layers may have helped slightly with overfitting, it did not make enough of a difference to allow the model to become more generalized.

Here is the confusion matrix of our alternate model:

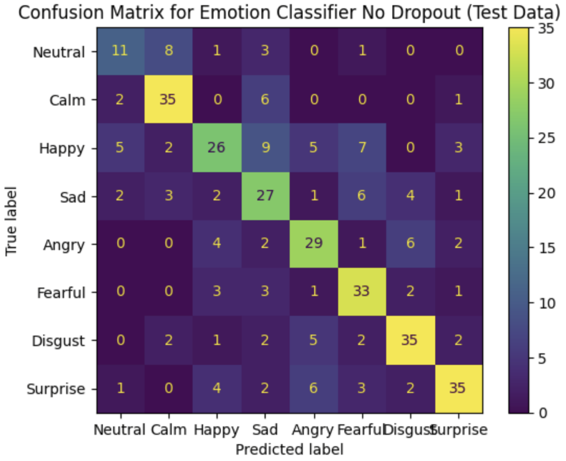

Figure 4: The confusion matrix of our alternate model on the testing set. Again, most of the emotions are classified correctly, but the same sets of pairs often are mistaken for another

The confusion matrices for our models are very similar, showing that similar miss-classifications are done both with and without dropout layers. While the dropout layers may have helped slightly with the testing accuracy of the model, the general performance of the model is not widely affected by the use of dropout layers for regularization.

6 Discussion

Even when using dropout layers, our model still overfits our training set. In this section, we will discuss how our model compares to another model trained on only on the RAVDESS dataset, models trained on other emotional speech datasets, and a cutting-edge SER model which successfully avoids overfitting and obtains a very high training accuracy. Firstly, we will discuss how our model compares to models trained on the RAVDESS dataset. We will compare with the model uploaded to kaggle by user aditya1220 [14]. This user used the same input features as ours, except they computed the average of each one. This caused the input to be a 1-dimensional vector, which loses some of the temporal information from the sequential nature of speech. Comparatively, our model kept the variable sequence length and instead using a bi-directional LSTM to consolidate the information across each time step. In addition, their model uses similar techniques of data augmentation, where they also change the pitch and add Gaussian noise to the input data. In addition to this, they also augmented the data by shifting the audio by several milliseconds. Their model achieved a very comparable performance of 59% testing accuracy and 50% validation, showing that their model also suffers from overfitting. From this, we gather that that using only the RAVDESS dataset is not diverse enough speech data to generalize well, and is likely the main reason our model overfits.

Next, we will discuss how our model performs compared to other general SER models from researchers. Compared to a slightly earlier model, Lim, W., Jang, D., & Lee, T used a very similar architecture model to ours. They use CNNs to read features from the STFT, and then consolidate it using a LSTM, before classifying it using a fully connected layer [5]. Using this method on the Berlin database, which is a databse of 535 utterances, they were able to achieve an average accuracy of 87% on the dataset. Which is quite a performance improvement on our model. This is likely because of the differences in speech style in the audio dataset [5]. Finally, we compare our model with a cutting edge model proposed by Niu et al. Using deep retinal convolutional neural networks, and DAARIP to generate different spectrograms in the training data, they were able to obtain a testing accuracy close to 99% [11]. This is because of both the advanced model they used, as well as because DAARIP was able to generate around 40 times as much training data than was originally present in the dataset. In their paper, they show their architecture without using DAARIP only achieved a validation accuracy of 42% and a testing accuracy of 41% [11]. In addition, much of the research they reference discusses how other models achieve around 50-70% testing accuracy. This shows that the best way to reduce overfitting is through advanced data augmentation on a diverse dataset.

Compared to these other models, we conclude that our model's overfitting is most likely due to the smaller amount of training examples combined with complexity of our model, which caused the parameters to become over-tuned to patterns only found in the training set. Besides using dropout, which is a form of regularization, this can likely be improved by adding additional datasets to train our model, and through more advanced data augmentation techniques such as DAARIP in order to generate a more robust training set for our model.

7 Limitations

A limitation of our model lies in the chosen data. The "Neutral" class is underrepresented in the dataset, with fewer neutral emotion training examples compared to the other emotions. This means that the neutral class is expected to have a lower classification rate than other classes, which is what we see when analyzing the confusion matrix. Another limitation is that the high variance of a real audio sample is very difficult to generalize and predict because there are so many different rates of speech, tone and pitch. Even though we used data augmentation to change the pitch and add some noise, large amounts of pitch variance can still cause worse accuracies.

Additionally, the audio recordings from each actor in the RAVDESS dataset have a relatively consistent rate of speech. This means that any audio recording input that has a reasonably slower or faster rate of speech would likely result in a much lower accuracy for our model because it does not have training data to generalize for different rates of speech.

As shown in our results and discussion sections, our model is limited in that it still overfits the RAVDESS dataset, even when using dropout layers, and because of this, only manages to obtain a test accuracy of 58% while the training accuracy is close to 94%. Since our model was trained on only one dataset, which is a comparatively smaller amount of data, our model is more prone to overfitting. This could be because our model is too complex to learn from this size of training data. This would result in a lower accuracy for the validation/testing sets, and shows that testing on audio outside the dataset may be more likely to be misclassified.

8 Ethical Considerations

In this dataset, the voice actors consented for their data to be used in machine learning models [11]. Therefore, the use of this dataset is not an ethical concern. However, our model could be used for malicious actions. As mentioned above, mental health interaction systems are an important use of SER models [9]. These SER models are vital for this type of service in ensuring that proper therapy is and create more harm than good. For this reason, many precautions need to be taken regarding these emotion recognition models. However, malicious individuals could take this proposed model and create a language model trained to subtly manipulate a person depending on their distress, sadness, etc. This could cause further distress and lead to devastating outcomes for the victim's mental and physical health. This is one of many ways that SER models could be used maliciously, and create more harm than good. For this reason, many precautions need to be taken regarding these emotion recognition models.

9 Conclusion

We obtained a testing accuracy of 58%, and a training accuracy of 94% using our proposed model, showing that our model was still overfitting our training data even with the inclusion of dropout layers. While dropout layers add a small advantage against overfitting, there are several other techniques to reduce overfitting that may be more beneficial to pursue for future SER models. From our investigation, the best performance increases come from advanced data augmentation techniques like DAARIP in order to generate a large increase in training examples, and using a deep model to learn the deep features from the spectrograms. This way, the model will generalize by learning from many different speeds, pitches, and loudness of speeches. If we were to train our more complex model on a larger dataset, our model would likely have been much more generalized overall.

References

[1] Basu, Saikat, et al. “A Review on Emotion Recognition Using Speech.” *2017 International Conference on Inventive Communication and Computational Technologies* (ICICCT), IEEE, 2017.

[2] Bidgoli, Hossein. *Encyclopedia of Information Systems*. Academic Press, 2003.

[3] Jalal, M. A., Moore, R. K., & Hain, T. (2019). Spatio-Temporal Context Modelling for Speech Emotion Classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, pp. 853-859. https://doi.org/10.1109/ASRU46091.2019.9004037.

[4] Kerkeni, Leila, et al. “Automatic Speech Emotion Recognition Using an Optimal Combination of Features Based on EMD-TKEO.” *Speech Communication*, vol. 114, 2019, pp. 22–35.

[5] Lim, W., Jang, D., & Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, Korea (South), pp. 1-4. https://doi.org/10.1109/APSIPA.2016.7820699.

[6] Livingstone, S. R., & Russo, F. A. (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. https://doi.org/10.1371/journal.pone.0196391.

[7] Lokesh, S., and M. Ramya Devi. “Speech Recognition System Using Enhanced Mel Frequency Cepstral Coefficient with Windowing and Framing Method.” *Cluster Computing*, Springer US, 4 Dec. 2017, https://link.springer.com/article/10.1007/s10586-017-1447-6.

[8] Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning — A systematic review. *Intelligent Systems with Applications*, 20, 200266. https://doi.org/10.1016/j.iswa.2023.200266.

[9] Mekruksavanich, S., Jitpattanakul, A., & Hnoohom, N. (2020). Negative Emotion Recognition using Deep Learning for Thai Language. In *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, Pattaya, Thailand, pp. 71-74. https://doi.org/10.1109/ECTIDAMTNCON48261.2020.9090768.

[10] Mu, Y., Gómez, L. A. H., Montes, A. C., Martínez, C. A., Wang, X., & Gao, H. (2017). Speech emotion recognition using convolutional-recurrent neural networks with attention model. *DEStech Transactions on Computer Science and Engineering*, 341-350.

[11] Niu, Y., Zou, D., Niu, Y., He, Z., & Tan, H. (2017). A breakthrough in speech emotion recognition using deep retinal convolution neural networks. *arXiv preprint arXiv:1707.09917*.

[12] Pereira, Mildred, et al. “Analysis of Windowing Techniques for Speech Emotion Recognition.” *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, Feb. 2016, pp. 1–6, https://doi.org/10.1109/icices.2016.7518859.

[13] Rabiner, L. R. (1978). *Digital Processing of Speech Signals*. Pearson Education India.

[14] Singh, Aditya Kumar. April 2024. Speech Emotion Recognition. https://www.kaggle.com/code/aditya1220/speech-emotion-recognition/notebook

[15] Wu, H., & Gu, X. (2015). Towards dropout training for convolutional neural networks. *Neural Networks*, 71, 1-10.

[16] Zhang, Yuanyuan, et al. “Attention Based Fully Convolutional Network for Speech Emotion Recognition.” *IEEE Conference Publication*, IEEE Xplore, 2018, https://ieeexplore.ieee.org/document/8659587/.

[17] Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN. *Sensors*, 17(7), 1694. https://doi.org/10.3390/s17071694.