Speech Emotion Recognition

Machine Learning in Statistics: Group 3
Dylan Arndt, Maddi Lynch, Zhiying Piao, and Gulnaz Yerdenova

Introduction

In the realm of human-computer interaction, our project focuses on using machine learning to recognize three emotions (angry, sad, neutral) conveyed through verbal speech. By analyzing acoustic features like pitch, Chroma, MFCC, and ZCR, our aim is to develop a model capable of accurately classifying emotions. The potential applications of this technology are diverse, spanning across healthcare, education, and beyond, where empathetic and responsive systems can enhance user experiences.

Data Sources

The data for this project was sourced from two established emotional speech datasets:

1. Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) is a set of 7,442 audio clips recorded from 91 actors expressing emotions.
2. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a set of 1,440 audio clips from 24 actors expressing emotion through speech and song.

These datasets include both male and female actors, which is significant in developing a model that can classify emotion from a variety of speakers.

Software/Hardware

We used Google Colaboratory in order to collaboratively develop the code and use a shared dataset that was certain to be indexed consistently. While this approach helped multiple people work simultaneously, it did limit our computing power to run more advanced models or feature extraction. For example, attempting to run a SVC model with more than 100 extracted features from the spectrograms of each audio file exceeded the computing limit.

Feature Extraction

We conducted a thorough research of prior SER algorithms to identify which features of the speech were the most important in classifying emotions. This was necessary because Google

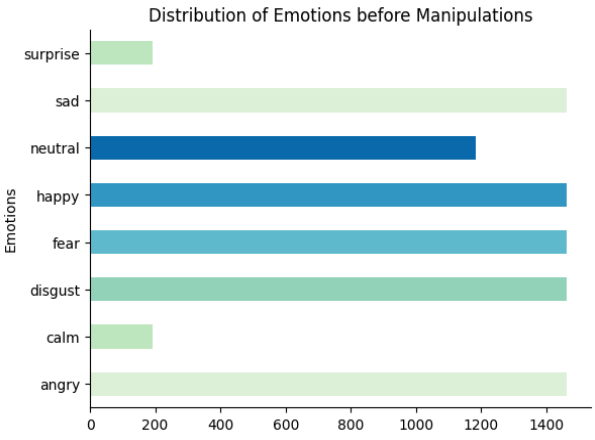
Colab's computational capacity couldn't handle utilizing all available features and allowing the model to determine their importance. Hence, we focused on hand picking the essential features.

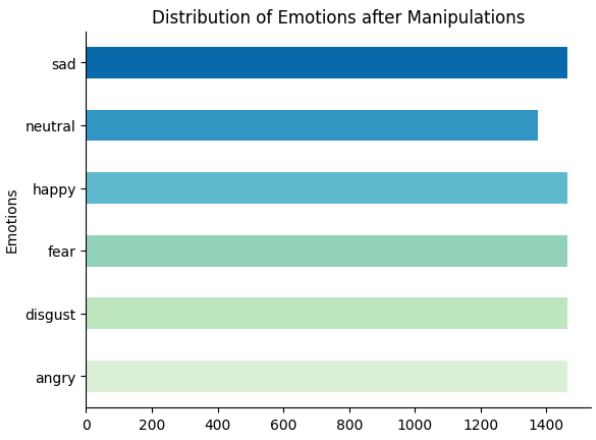
- **Chroma Features** were extracted to capture the tonal nature of speech. Tonality refers to the arrangement of sounds "according to pitch relationships in interdependent spatial and temporal structures" (Bello, 2017). This is composed of 12 pitches related through structures of melody (horizontal, sequential) and harmony (vertical, synchronous).
- **Zero-Crossing Rate (ZCR)** was extracted to quantify the noise level and frequency of the audio.
- **Pitch** was extracted to include the fundamental frequency of the audio. This fundamental frequency is often related to emotional intensity, and can help refine an SER model.
- **Mel-Frequency Cepstral Coefficients (MFCCs)** is a key component of SER models and is comprised of a set of features that describe multiple dimensions of an audio, including shape along time and frequency domains. These features indicate the short-term power spectrum and timbral aspects of audio.

Data Preparation

Addressing Imbalance in the Dataset and Emotion Selection

Our original dataset displayed an imbalance, particularly with fewer samples representing surprise and calm emotions. After careful examination, we found minimal differentiation between calm and neutral audio samples. As a result, we opted to combine the calm and neutral categories. After manipulations, there was a balanced number of audio files for each of the 6 emotions across the two datasets, with slightly fewer clips using a neutral emotion.





Data cleaning and processing

We conducted thorough data cleaning which included proper labeling, identifying missing values, and removing redundant columns. Additionally, because our data contained categorical labels, we employed label encoding to prepare it for model usage. Lastly, we scaled the data using StandardScaler(), which improved accuracy scores and decreased training time.

Model and Algorithm Selection

	Accuracy Score		
Model	6 Emotions	3 Emotions	Binary
ExraTreesClassifier	0.46	0.77	0.87
SVC	0.49	0.77	0.88
Gradient Boosting	0.48	0.76	0.88
Random Forest	0.48	0.76	0.87

Before starting with model training, we partitioned our dataset into two subsets: 20% for testing and 80% for training.

Using all 6 emotions

We began with an attempt to use Extra Trees, Random Forest, Gradient Boost, and SVC model to classify all 6 emotions. We quickly realized this was too big of a problem to solve with the number of features we extracted. In fact, the best model accuracy we were able to obtain was only 0.49 using a SVC algorithm.

Binary model

The four algorithms were used to answer a simpler, binary classification problem - is the speaker angry (1) or not angry (0)? We were able to achieve a very high accuracy, reaching 0.88 with SVM and Gradient Boosting algorithms.

Using 3 emotions

In order to balance the effectiveness of the model with the business use case, and given our successful accuracy scores, we expanded the binary classifications to include neutral, sad, and angry. This approach introduced more complexity compared to binary models but struck a better balance compared to attempting to include all six emotions.

Final Model

After assessing four algorithms, it became clear that SVC and ExtraTreesClassifier obtain better accuracy scores. We decided to use Support Vector Machine (SVM) for several reasons:

1. **Simplicity and Interpretability**: SVM models are easy to understand for both data scientists and stakeholders, unlike more complex alternatives such as neural networks or Random Forests.
2. **Ease of Implementation and Training**: SVMs are straightforward to implement and train. We also achieved a shorter training time compared to some other algorithms.
3. **Robustness to Overfitting**: SVMs are less prone to overfitting, especially when compared to ensemble models with decision trees.
4. **Suitability for Small Datasets**: SVMs perform well with small datasets, making them a suitable choice for our scenario.

SVM Model Training

Hyper parameter optimization

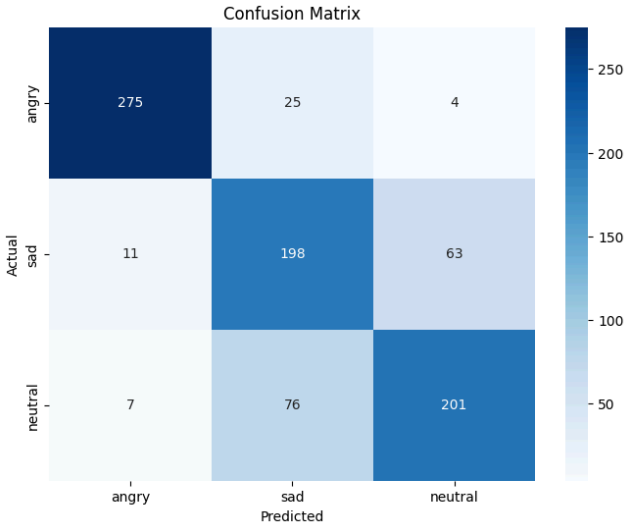
```
param_grid = {
        'C': [0.1, 1, 10, 100],  # Regularization parameter
        'kernel': ['linear', 'rbf', 'poly'],  # Kernel type
        'gamma': [0.001, 0.01, 0.1, 1, 10]
 }
```

Using grid search, we identified the best hyperparameters:

- 'C' : 10, which is a relatively high C and minimizes misclassification.
- 'kernel': ‘rbf’, with the complex decision boundaries for emotional classification, it makes sense that ‘rbf’ was chosen because it is more flexible than poly or linear.
- 'gamma': 0.01, instances have a wider influence, allowing for smother decision boundaries.

Model Performance

Tuned hyperparameters improved the model accuracy to 0.78. We achieved the confusion matrix presented below:



From the confusion matrix we can conclude that angry had the highest accuracy, followed by neutral. Sad emotions were misclassified as neutral 23% of the time and neutral was misclassified as sad 27% of the time. It is not surprising since sad and neutral could sound similar even to the human ear.

	Precision	Recall	F1-Score
Angry	0.94	0.90	0.92
Sad	0.66	0.73	0.69
Neutral	0.75	0.71	0.73

Precision, recall and f1-score metrics align with the findings from the confusion matrix, highlighting that angry emotions were the easiest to identify. This consistency in performance metrics supports the findings of the binary model (angry vs. not angry), which exhibited slightly better performance.

Conclusion

We are satisfied with the performance of our model given the limited number of features extracted. If given additional time and computing power, we would update the model in order to enhance performance and usability. First, we would extract far more features, starting with spectrograms. With additional features and emotions to classify, we would explore the existing four algorithms as well as additional neural network models. Finally, we would add noise to the audio data in order to better train the model for real-world audio quality.

References

Bello, Juan Pablo. (2017, September). *Chroma and tonality*. https://wp.nyu.edu/jpbello/wp-content/uploads/sites/1691/2017/09/6-tonality.pdf

Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech emotion recognition through hybrid features and convolutional neural network. *Applied Sciences*, 13(8), 4750. https://doi.org/10.3390/app13084750