Emotion Recognition from Speech

A classification model utilizing the RAVDESS data set

Samantha Gonzales, Savannah Stewart, Bhagavan Bejjipurapu, Aditya Naredla

Processing

Data Set

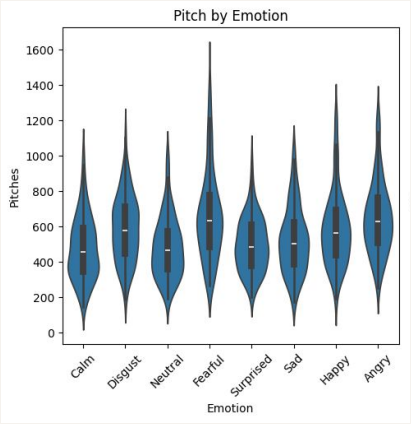
- Data set: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
- Available through: Zenodo
- The data set provided the following:
  - Emotion
  - Emotional Intensity
  - Statement
  - Repetition
  - Actor ID
  - Gender
  - Modality
  - Voice Channel

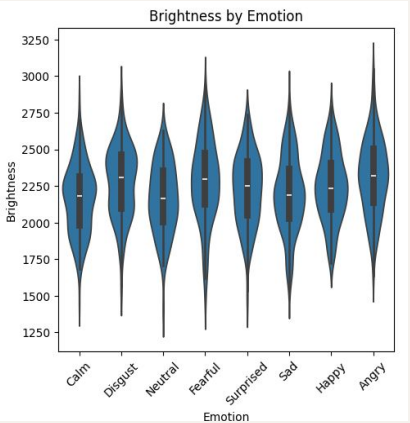
	Modality	Vocal Channel	Emotion	Intensity	Statement	Repetition	Actor ID	Gender	File Path
0	Audio-only	Speech	Calm	Strong	Kids are talking by the door	1	7	Male	./ravdess_data/Actor_07/03-01-02-02-01-01-07.wav
1	Audio-only	Speech	Disgust	Normal	Dogs are sitting by the door	2	7	Male	./ravdess_data/Actor_07/03-01-07-01-02-02-07.wav
2	Audio-only	Speech	Neutral	Normal	Dogs are sitting by the door	2	7	Male	./ravdess_data/Actor_07/03-01-01-01-02-02-07.wav
3	Audio-only	Speech	Fearful	Normal	Kids are talking by the door	2	7	Male	./ravdess_data/Actor_07/03-01-06-01-01-02-07.wav
4	Audio-only	Speech	Surprised	Strong	Kids are talking by the door	1	7	Male	./ravdess_data/Actor_07/03-01-08-02-01-01-07.wav
5	Audio-only	Speech	Calm	Normal	Kids are talking by the door	1	7	Male	./ravdess_data/Actor_07/03-01-02-01-01-01-07.wav
6	Audio-only	Speech	Calm	Normal	Dogs are sitting by the door	2	7	Male	./ravdess_data/Actor_07/03-01-02-01-02-02-07.wav
7	Audio-only	Speech	Sad	Normal	Dogs are sitting by the door	2	7	Male	./ravdess_data/Actor_07/03-01-04-01-02-02-07.wav
8	Audio-only	Speech	Disgust	Normal	Dogs are sitting by the door	1	7	Male	./ravdess_data/Actor_07/03-01-07-01-02-01-07.wav
9	Audio-only	Speech	Fearful	Normal	Kids are talking by the door	1	7	Male	./ravdess_data/Actor_07/03-01-06-01-01-01-07.wav

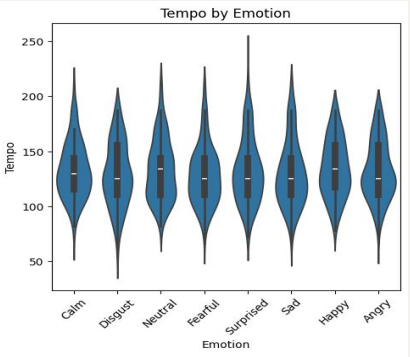
Rows 0 - 9 of dataframe from original data set - no feature extraction

Feature Extraction

- **Mel-Frequency Cepstral Coefficients (MFCC)**: Captures timbral features
- **Mel-Spectrogram**: Captures frequency content of audio signals over time
- **Tempo**: Estimates the tempo by beats per minute
- **Spectral Centroid (Brightness)**: Captures where the energy of sound is concentrated in a frequency
- **Pitch**: Captures pitch and its magnitude in audio signals over time







Feature extraction by Emotion for: Pitch, Brightness, & Tempo

Classification Models

Random Forest

```
forest_classifier = RandomForestClassifier(max_features=6,
                                           random_state=42,
                                           n_estimators=700,
                                           criterion='gini',
                                           max_depth=None)


forest_classifier.fit(X_train, y_train) # train the model
y_pred = forest_classifier.predict(X_test) # test the model

y_test_labels = emotion_encoder.inverse_transform(y_test)
y_pred_labels = emotion_encoder.inverse_transform(y_pred)
```

- Step 1
  - Encode categorical variables into numerical variables for use in model
- Step 2
  - Split into train and test groups
  - Used 80/20 ratio for train and test
- Step 3
  - Define hyperparameters for model
  - Used RandomizedSearchCV to find optimal parameters with our training data
- Step 4
  - Train the model
  - Test the model

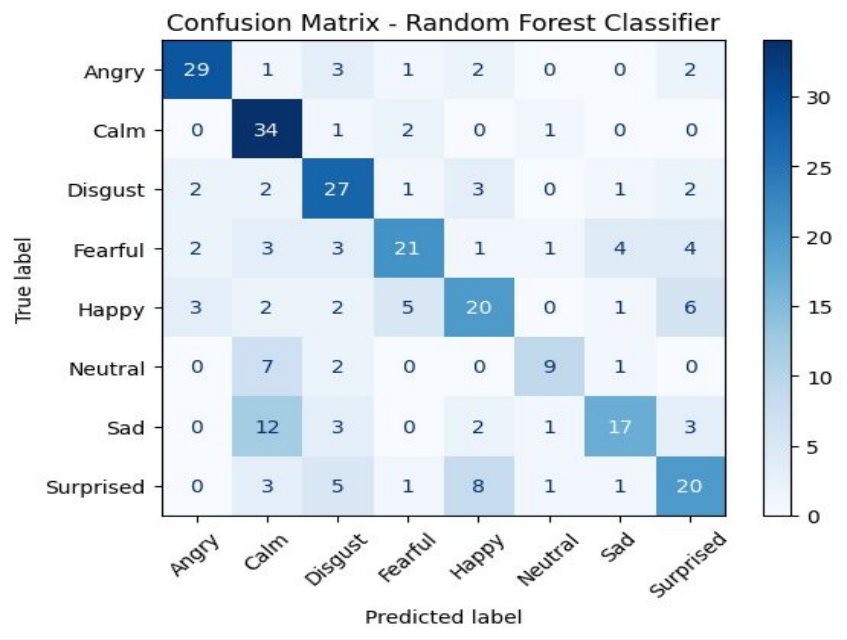
Results

Accuracy: 0.6145833333333334

Classification Report:

	precision	recall	f1-score	support
Angry	0.81	0.76	0.78	38
Calm	0.53	0.89	0.67	38
Disgust	0.59	0.71	0.64	38
Fearful	0.68	0.54	0.60	39
Happy	0.56	0.51	0.53	39
Neutral	0.69	0.47	0.56	19
Sad	0.68	0.45	0.54	38
Surprised	0.54	0.51	0.53	39
accuracy			0.61	288
macro avg	0.63	0.61	0.61	288
weighted avg	0.63	0.61	0.61	288

Accuracy Score



Confusion Matrix

XGB

```
xgb_model = xgb.XGBClassifier(
    objective='multi:softmax',
    num_class=len(emotion_encoder.classes_),
    max_depth=None,
    learning_rate=0.1,
    n_estimators=700,
    subsample=0.9,
    colsample_bytree=0.9,
    random_state=42,
    use_label_encoder=False,
    eval_metric='mlogloss'
)

xgb_model.fit(X_train, y_train)

y_pred = xgb_model.predict(X_test)

y_test_labels = emotion_encoder.inverse_transform(y_test)
y_pred_labels = emotion_encoder.inverse_transform(y_pred)
```

- Step 1
  - Encode categorical variables into numerical variables for use in model
- Step 2
  - Split into train and test groups
  - Used 80/20 ratio for train and test
- Step 3
  - Define hyperparameters for model
  - Used RandomizedSearchCV to find optimal parameters with our training data
- Step 4
  - Train the model
  - Test the model

Results

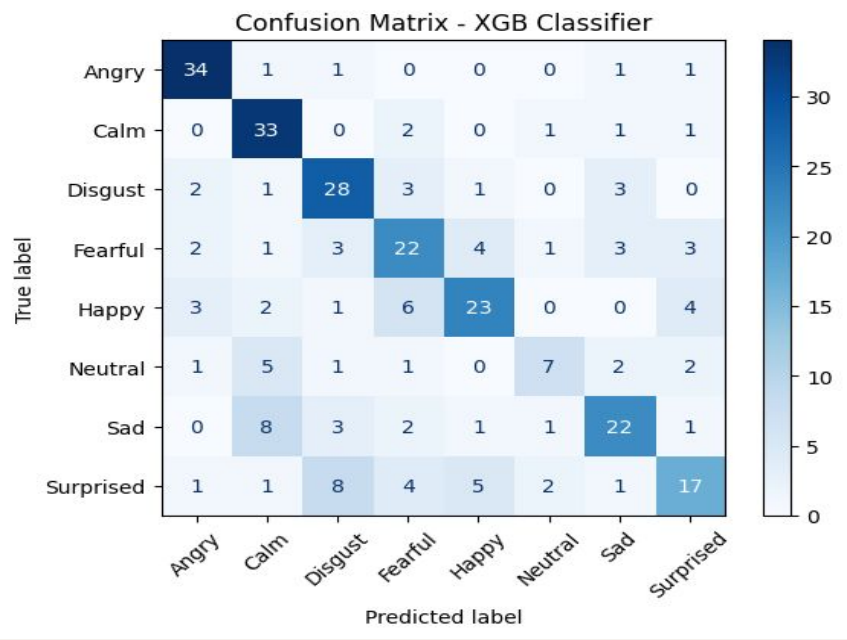
```
Accuracy: 0.6458333333333334

Classification Report:
              precision    recall  f1-score   support

       Angry       0.79      0.89      0.84        38
        Calm       0.63      0.87      0.73        38
     Disgust       0.62      0.74      0.67        38
     Fearful       0.55      0.56      0.56        39
       Happy       0.68      0.59      0.63        39
     Neutral       0.58      0.37      0.45        19
         Sad       0.67      0.58      0.62        38
   Surprised       0.59      0.44      0.50        39

    accuracy                           0.65       288
   macro avg       0.64      0.63      0.63       288
weighted avg       0.64      0.65      0.64       288
```

Accuracy Score



Confusion Matrix

Contributions

Samantha

- Data Processing
- Extract Labels from filename identifiers
- Dataframe creation
- Data Visualizations
- MCFF feature extraction function
- Random Forest Classifier Model
- Confusion matrices
- Feature Importance visualizations

Savannah

- Mel-Spectrogram feature extraction
- Spectral Centroid feature extraction
- Tempo feature extraction
- Pitch feature extraction
- Data Visualizations
- Random Forest Classifier Fine Tuning

Bhagavan

- Encode Categorical Labels into Numerical Labels
- Classification Report for Models
- XGB Classifier Model
- Attempted CNN Modeling for predicting labels

Aditya

- Presented own work

References

- [1]A. S. M and A. U. M, "Enhancing Speech Emotion Recognition Through Advanced Feature Extraction and Deep Learning: A Study Using the RAVDESS Dataset," 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-7, doi: 10.1109/ICITEICS61368.2024.10624849. keywords: {Deep learning;Emotion recognition;Sentiment analysis;Accuracy;Computational modeling;Speech recognition;Feature extraction;Speech Emotion Recognition;Deep Learning;Sentiment Analysis;Mel-frequency cepstral coefficient;Librosa},
- [2]M. M. Rezapour Mashhadi and K. Osei-Bonsu, “Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest,” PloS one, vol. 18, no. 11, p. e0291500, 2023, doi: https://doi.org/10.1371/journal.pone.0291500.
- [3]S. Hamsa, I. Shahin, Y. Iraqi and N. Werghi, "Emotion Recognition From Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier," in IEEE Access, vol. 8, pp. 96994-97006, 2020, doi: 10.1109/ACCESS.2020.2991811. keywords: {Emotion recognition;Speech recognition;Hidden Markov models;Feature extraction;Transforms;Modulation;Computational modeling;Emotion recognition;noise reduction;cochlear filterbank;feature extraction}
- [4]Y. R. Rochlani and A. B. Raut, "Machine Learning Approach for Detection of Speech Emotions for RAVDESS Audio Dataset," 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2024, pp. 1-7, doi: 10.1109/ICAECT60202.2024.10468810. keywords: {Support vector machines;Dimensionality reduction;Emotion recognition;Time-frequency analysis;Speech recognition;Forestry;Feature extraction;Classification;Emotion recognition;Mel frequency cepstral coefficients;Pitch;Short term energy;Support Vector Machine;Decision Tree;Random Forest Classifier}
- [5]”Mel-frequency Cepstral Coefficients (MFCC) for Speech Recognition.” GeeksforGeeks. https://www.geeksforgeeks.org/mel-frequency-cepstral-coefficients-mfcc-for-speech-recognition/ (accessed May 6, 2025)
- [6]”Emotional prosody.” Wikipedia: The Free Encyclopedia. https://en.wikipedia.org/wiki/Emotional_prosody (accessed March 31, 2025)

Thanks