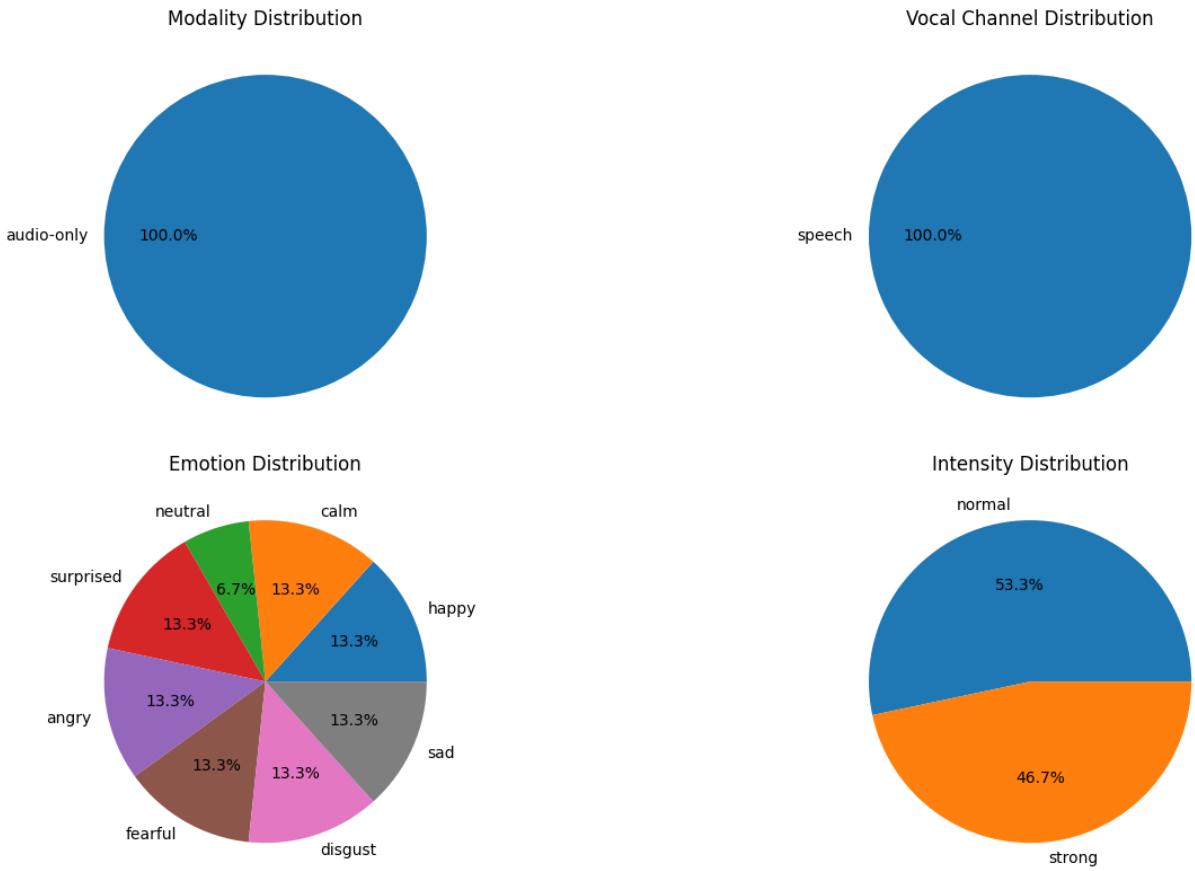
Audio Emotion Recognition Project Report

1. Introduction

This project focuses on audio-based emotion recognition using a dataset comprising various emotional vocal expressions. We utilized mel spectrograms for visual representation and applied data augmentation to enhance model generalization.

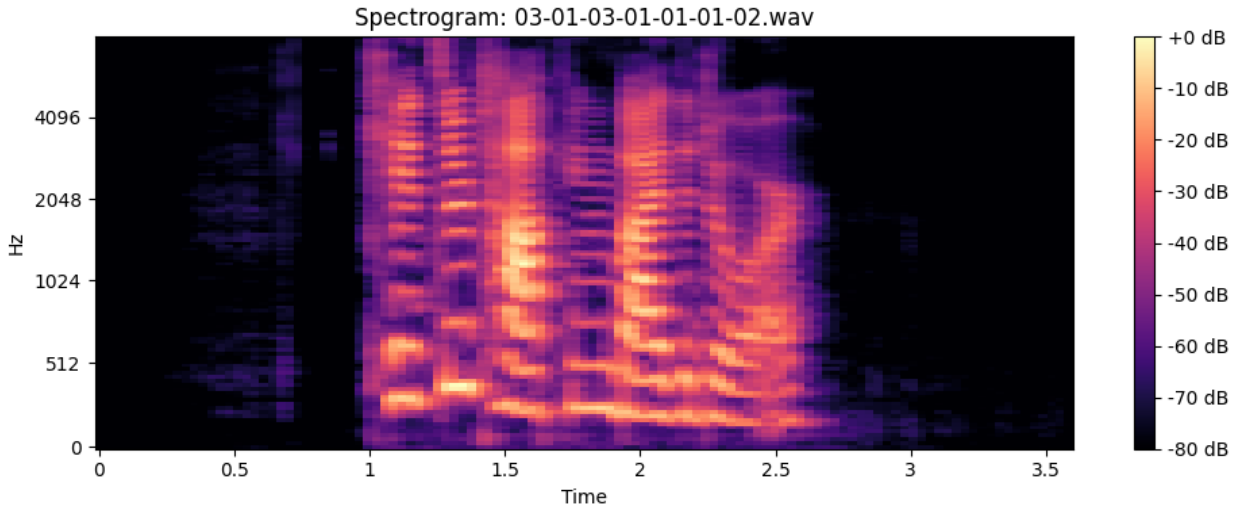
2. Dataset Overview

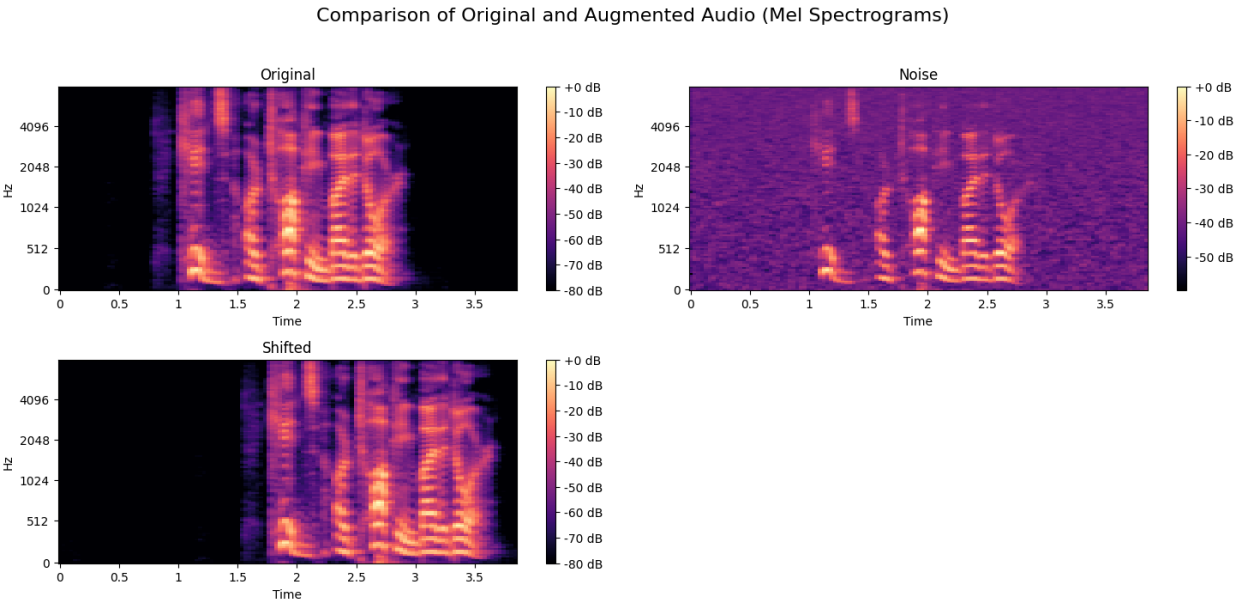
The dataset consists solely of audio-only samples in the speech vocal channel. Emotions are evenly distributed across categories, and both genders are equally represented. Two levels of intensity—normal and strong—are present.



3. Spectrogram and Data Augmentation

Below is an example mel spectrogram for an audio file, followed by augmented versions using noise, pitch shift, and time stretching. These transformations help build a more robust model.





4. Feature Extraction

Features were extracted using audio processing techniques, resulting in a high-dimensional matrix representing different audio variants.

Extracted feature matrix shape: (32, 334)

	emotion	variant	f1	f2	f3	f4	f5	f6	f7	f8	...	f323	f324	f325	f326	f327	f328	f329	f330	f331	f332
0	happy	original	-495.576782	11.650211	-21.840162	-6.352988	-20.073908	-14.741199	-13.168602	-10.453532	...	-0.027645	0.037898	-0.001897	-0.017497	0.094553	0.133020	0.221166	0.142006	0.070276	0.062787
1	happy	noise	-266.983143	8.371705	-5.938495	-5.554313	-5.644103	-4.487367	-4.978689	-3.386167	...	-0.019674	0.013542	-0.010378	-0.018521	0.082208	0.127709	0.171561	0.112956	0.055743	0.046083
2	happy	stretch	-509.381439	12.584217	-24.613226	-9.105054	-21.186100	-15.680207	-14.990541	-11.168181	...	-0.018180	0.050587	-0.001922	-0.023894	0.116464	0.156152	0.233462	0.134058	0.081562	0.066859
3	happy	pitch	-525.867249	2.509046	-29.765924	-12.621187	-27.125391	-12.187161	-15.910378	-4.748474	...	-0.024689	0.017893	-0.019326	0.031692	0.159656	0.097983	0.226381	0.124871	0.064375	0.066734
4	calm	original	-684.253906	37.808414	-0.393736	-0.190155	-2.567374	-10.388621	-13.640526	-9.459058	...	-0.008755	0.064984	-0.014269	-0.012262	0.128193	0.097874	0.221610	0.181238	0.069680	0.067445

5 rows × 334 columns

5. Train-Test Split

A total of 5760 samples were used, split into 4320 training and 1440 testing samples. Each sample was represented by 29,184 features.

```
Maximum flattened feature length: 29184
Total samples: 5760
Train samples: 4320, Test samples: 1440
X_train shape: (4320, 29184), y_train shape: (4320,)
X_test  shape: (1440, 29184),  y_test  shape: (1440,)
```

6. Model Performance

6.1 Logistic Regression

The Logistic Regression model achieved an accuracy of approximately 72%, with balanced precision and recall across most emotions.

```
Logistic Regression
Accuracy: 0.7222222222222222
              precision    recall  f1-score   support

       angry       0.77      0.70      0.73       192
        calm       0.73      0.77      0.75       192
     disgust       0.75      0.69      0.72       192
     fearful       0.80      0.80      0.80       192
       happy       0.71      0.65      0.68       192
     neutral       0.66      0.66      0.66        96
         sad       0.65      0.71      0.68       192
   surprised       0.70      0.77      0.73       192

    accuracy                           0.72      1440
   macro avg       0.72      0.72      0.72      1440
weighted avg       0.72      0.72      0.72      1440
```

6.2 Random Forest

The Random Forest classifier performed with lower accuracy (about 62%) compared to Logistic Regression, showing weaker performance for certain emotions.

```
Random Forest
Accuracy: 0.61875
              precision    recall  f1-score   support

       angry       0.65      0.66      0.65       192
        calm       0.70      0.72      0.71       192
     disgust       0.58      0.68      0.62       192
     fearful       0.62      0.63      0.63       192
       happy       0.62      0.50      0.55       192
     neutral       0.75      0.47      0.58        96
         sad       0.70      0.55      0.61       192
   surprised       0.49      0.68      0.57       192

    accuracy                           0.62      1440
   macro avg       0.64      0.61      0.62      1440
weighted avg       0.63      0.62      0.62      1440
```

7. Random Forest Predictions

The following predictions were made on 30 original test samples, showing a mix of accurate and misclassified results.

```
Random Forest Predictions on 30 Original Samples
Index 4382 → Predicted: angry, Actual: angry
Index 5060 → Predicted: calm, Actual: calm
Index 2187 → Predicted: calm, Actual: calm
Index 4813 → Predicted: neutral, Actual: neutral
Index 5298 → Predicted: calm, Actual: calm
Index 1101 → Predicted: surprised, Actual: surprised
Index 2210 → Predicted: calm, Actual: calm
Index 3807 → Predicted: angry, Actual: angry
Index 3524 → Predicted: surprised, Actual: surprised
Index 5462 → Predicted: fearful, Actual: fearful
Index 2462 → Predicted: sad, Actual: disgust
Index 2364 → Predicted: angry, Actual: angry
Index 4622 → Predicted: happy, Actual: happy
Index 1499 → Predicted: sad, Actual: sad
Index 132 → Predicted: fearful, Actual: fearful
Index 1718 → Predicted: sad, Actual: sad
Index 1835 → Predicted: fearful, Actual: fearful
Index 530 → Predicted: calm, Actual: calm
Index 2596 → Predicted: angry, Actual: angry
Index 157 → Predicted: angry, Actual: angry
Index 5629 → Predicted: calm, Actual: calm
Index 1844 → Predicted: surprised, Actual: surprised
Index 177 → Predicted: happy, Actual: angry
Index 4506 → Predicted: fearful, Actual: fearful
Index 4358 → Predicted: sad, Actual: sad
Index 2611 → Predicted: angry, Actual: angry
Index 1871 → Predicted: fearful, Actual: fearful
Index 5046 → Predicted: calm, Actual: calm
Index 2909 → Predicted: disgust, Actual: calm
Index 101 → Predicted: disgust, Actual: disgust
```

8. Conclusion

This project demonstrated the effectiveness of audio-based emotion classification using spectrograms and traditional classifiers. Logistic Regression provided the best performance, and data augmentation contributed to better model robustness. Further improvements could involve using deep learning models and attention mechanisms.