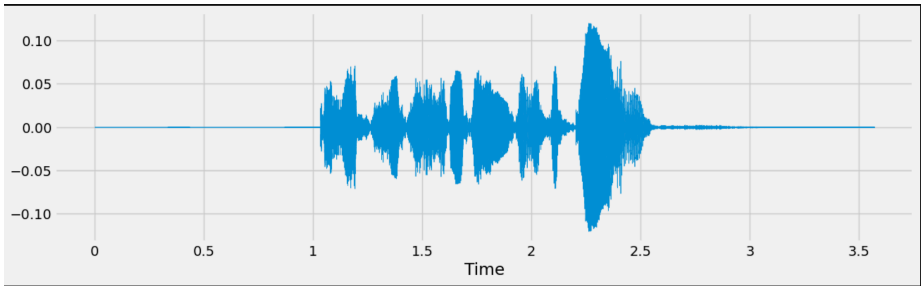
Speech Emotion Recognition using Spectrograms

COURSE PROJECT
DA221M



Yesh Lohchab*
230102115
yesh.3119@iitg.ac.in

Vanshika Mittal*
230102109
m.vanshika@iitg.ac.in

* Equal Contribution. Listing Order is Random

Abstract

Emotion recognition is a strategy for social robots used to implement better Human-Robot Interaction and model their social behaviour. Since human emotions can be expressed in different ways (e.g., face, gesture, voice), multimodal approaches are useful to support the recognition process. However, although there exist studies dealing with multimodal emotion recognition for social robots, they still present limitations in the fusion process, dropping their performance if one or more modalities are not present or if modalities have different qualities.

1 Introduction

The voice is a primary resource that is widely used and extremely important for communication among human beings; it transmits ‘‘more than simple words’’. Thus, the use of speech to explore emotions is significant due the rich information that can be extracted. To perform the recognition of emotions through speech, it is first necessary to extract and classify the resources. Thus, it is common to use resource extraction processes in audio files.

2 Related Works

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9705576

3 Theoretical Background

3.1 Convolutional Neural Networks

CNNs are deep learning algorithms that help computers interpret images similarly to humans. They are widely used for image classification and computer vision tasks, learning image features to make predictions. CNNs consist of:

- **Convolutional Layer**: Uses filters (kernels) to extract features from the image through matrix operations, producing a feature map.
- **Activation Layer**: Adds non-linearity with functions like ReLU, Softmax (for multiclass classification), or Sigmoid (for binary classification).
- **Pooling Layer**: Reduces dimensionality and minimizes overfitting by downsampling with methods like max or average pooling.
- **Fully Connected Layer**: Connects each node to previous nodes, classifying based on extracted features.

3.2 Transfer Learning for CNN

Building a model from scratch can be time-consuming. Transfer learning addresses this by using pre-trained CNN models as a starting point for related tasks. It leverages knowledge from the original model, adjusting weights for the new dataset. This works because early layers learn basic features like color and edges that are common across tasks. Fine-tuning is essential to adapt the model to the new task.

3.5 Class Weights

Class imbalance occurs when some classes have many samples while others have few. This makes it challenging for the model to learn minority class features, leading to bias toward the majority class. One solution is to assign class weights, giving more weight to minority classes and less to majority classes. This way, the model penalizes misclassification of minority classes more, reducing bias toward certain classes.

$$W_j = \frac{number\ of\ samples}{(samples\ of\ class)_j}$$

$$L = \sum_i -\alpha_i y_i \log(p_i)$$

3.6 Loss Function

The loss function measures how well the model predicts the expected outcome, guiding weight adjustments during training to minimize this value. Common loss functions include mean squared error, cross-entropy loss, hinge loss, and mean absolute error.

We used class-weighted cross-entropy loss, which factors in class weights to handle imbalanced data. Cross-entropy loss, or log loss, is commonly used for classification models with predictions between 0 and 1.

- For **binary classification**:
- For **multi-class classification**:

(α stands for the class-weight

3.7 Models Used

In recent years, deep learning methods have been successfully applied to image classification tasks. Certain convolutional neural network (CNN) architectures are renowned for their remarkable performance across various benchmarks. The CNN architectures used in our approach are as follows:

ResNet - ResNet50's[1] architecture solves the issue of multiple nonlinear layers that fail to learn identity maps and suffer from degradation issues. It does this by using stacked residual units that incorporate convolution and pooling layers.

EfficientNet -EfficientNet[2] architecture scales neural network models by uniformly enhancing depth, width and resolution dimensions, justified by the need for increased layers and channels in larger input images.

3.8 Evaluation Metrics

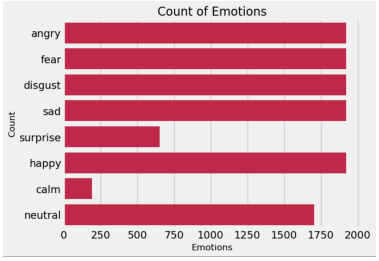
Accuracy: The ratio of correct predictions to the total samples.

- **TP**: True Positive
- **TN**: True Negative
- **FP**: False Positive
- **FN**: False Negative

Balanced Accuracy: Useful when the dataset is imbalanced, as standard accuracy may be biased toward the majority class.

ROC: A curve plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds, showing the classifier's performance.

AUC Score: The area under the ROC curve, indicating the model's ability to distinguish between classes. AUC values range from 0 to 1, with higher values indicating better separability.



4 Materials and Methods

4.1 Dataset

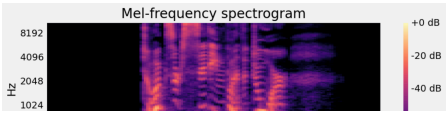
1) **Savee**
2) **Crema-D**
3) **Tess**
4) **Ravdess**

4.2 Data Preprocessing:

4.3 Proposed Framework

---

[2] Mingxing Tan, & Quoc V. Le. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.



4.3.1

Proposed Approach Outline

Attached table shows the flow of project from the dataset Loading to fine tuning and testing.

4.3.2 Fine Tuning Models

For our classification tasks (multiclass), we employed State-of-the-art architectures, namely EfficientNet, and ResNet pretrained on ImageNet1k dataset. This dataset consists of 1000 classes with over a million image samples.

For our models we introduced a dense layer with ReLU activation. Finally, a dense output layer with 5 neurons is added with softmax activation function.

In both architecture, an additional global average pooling 2D layer was added before the newly added layers to enhance feature extraction.

Metalearner	Hyperparameters
EfficientNetB7	min_samples_split=16, n_epochs=5, batch_per_epoch=224

Table 1: Multiclass Classification Ensemble Meta-learners' best hyperparameters

After hyperparameter tuning, the best classifier was trained on the train set and used to generate the accuracy score, balanced accuracy score and AUC values on the train, test and val sets.

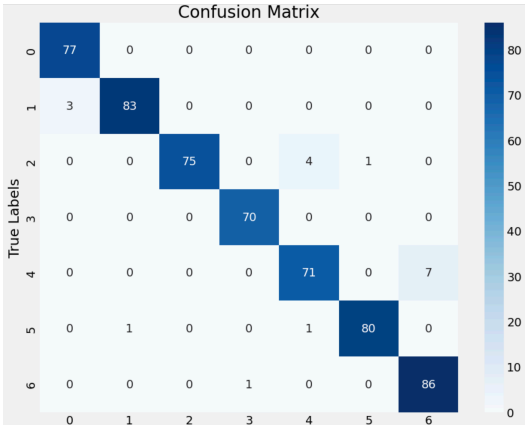
5 Experimental Results

This section summarizes the results obtained by various CNN and meta learner models used in this paper.

5.1 Fine Tuning

Table 6: Final Fine Tuning Results for Multiclass classification

EfficientNetB7 was chosen for fine tuning the model on the training dataset. This model was chosen because of its excellent score in imagenet based image muticlass classification tasks and the Randomised Search CV probable results. Various hyper parameters were chosen to train this model (the best ones have been mentioned above). Some computational limitations were a hinderance to this fine tuning. Overall this model was the best fit for this classification task.



```
Classification Report:
              precision    recall  f1-score   support

       angry       0.96      1.00      0.98        77
        fear       0.99      0.97      0.98        86
     disgust       1.00      0.94      0.97        80
         sad       0.99      1.00      0.99        70
    surprise       0.93      0.91      0.92        78
       happy       0.99      0.98      0.98        82
        calm       0.92      0.99      0.96        87

    accuracy                           0.97       560
   macro avg       0.97      0.97      0.97       560
weighted avg       0.97      0.97      0.97       560
```

concludes this document.

6 Conclusion

In this paper, we introduced a spectrogram learning fine tuning model using EfficientNetB7 for Constructing a speech emotion recognition model which predicts mood by just listening to the speech of the individual. This model is based on an innovative idea which was not mentioned in the Research Paper and has been adopted and thought by us. This model performed exceptionally with 97% precision, recall and F1-score. This