A Synopsis Report

On

Auramood-“Emotion Classification in audios”

For

Predictive Analytics

By

Agrima Jain - 500109836

Nidhi Bajaj – 500109328

Under the guidance of

Dr. Achala Shakya

University of Petroleum and Energy Studies

Dehradun-India

Table of Contents

Abstract:

Emotion recognition from speech is an emerging area of research in the field of human-computer interaction and artificial intelligence. The ability to detect and classify emotions in speech can greatly enhance the interaction between humans and machines, especially in areas like customer service, healthcare, and mental health monitoring. This project focuses on developing an emotion classification system that classifies emotions in speech data into predefined categories such as happiness, sadness, anger, surprise, fear, and neutral. By utilising deep learning techniques, particularly Convolutional Neural Networks (CNN), the system extracts features from audio signals, such as Mel Frequency Cepstral Coefficients (MFCCs), to classify emotions with high accuracy. The project aims to provide a robust solution for real-time emotion detection, with applications in virtual assistants, customer service automation, and therapeutic settings. The system is trained using a large dataset of labelled speech samples, and its performance is evaluated using metrics like accuracy, precision, recall, and F1-score. The results indicate that the model performs effectively, achieving a high classification accuracy across different emotional categories, while also highlighting areas for further improvement.

Introduction:

Emotion recognition from speech has been a significant area of research within the fields of natural language processing (NLP), audio signal processing, and machine learning. Recognizing emotions from speech plays a crucial role in human-computer interaction, customer support, mental health monitoring, and many other domains. The project focuses on creating an automatic system to classify emotions in speech data using deep learning techniques, particularly convolutional neural networks (CNN). The system analyses various emotional states such as happiness, sadness, anger, surprise, etc., from audio files, providing an efficient and automated emotion detection mechanism.

Problem Statement:

Emotion detection from speech plays a crucial role in understanding the emotional state of an individual, which is essential for building empathetic and intelligent systems. However, the challenge lies in accurately classifying the emotions embedded within the speech, as they are influenced by multiple factors, such as tone, pitch, speed, and context. Traditional approaches have often struggled to achieve high accuracy in detecting emotions from speech due to the complexity and variance in human emotional expression. Additionally, current systems face limitations in processing real-time speech data for emotion classification and often fail to generalise well across different languages, accents, and diverse emotional expressions.

Objective:

The main objective of this project is to develop an emotion classification model that can classify an audio file into predefined emotional categories. These categories typically include:

- **Happy**
- **Sad**
- **Anger**
- **Fear**
- **Surprise**
- **Neutral**

By extracting relevant features from raw audio data and training a machine learning model, the system can accurately predict the emotional content of an audio file. This model will have applications in:

1. **Human-Computer Interaction (HCI):** Enhancing user experiences by understanding the emotional state of the user, allowing for more empathetic interactions.
2. **Customer Service Automation:** Analysing customer calls or feedback to gauge satisfaction or dissatisfaction and respond accordingly.
3. **Healthcare Applications:** Detecting emotional states during therapy sessions, which can help therapists better understand patients' mental states.
4. **Voice-Based Assistants:** Improving the interaction between virtual assistants and users by detecting user emotions to tailor responses.

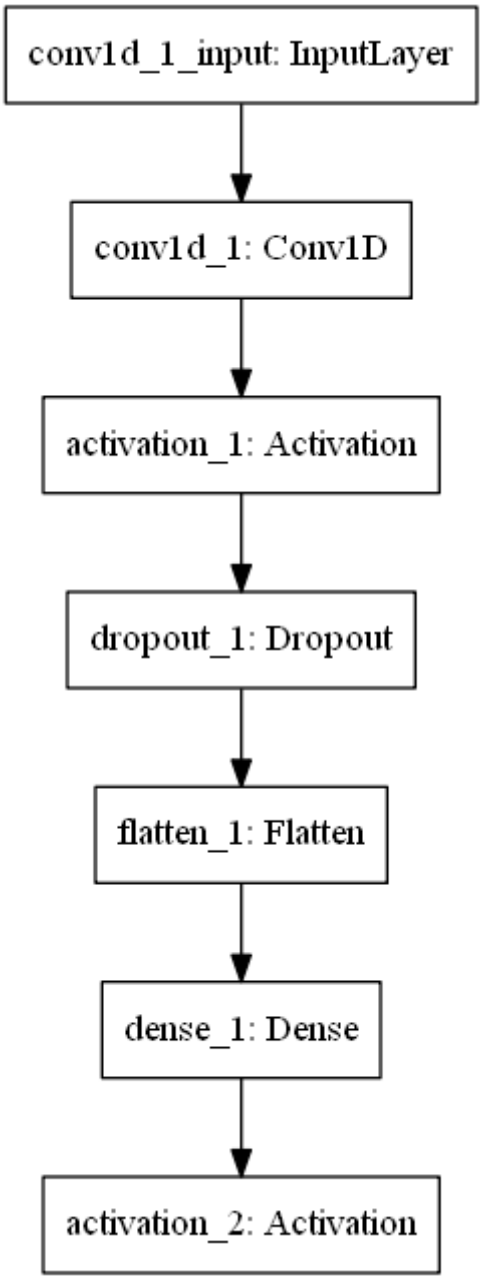
Methodology:

The project is divided into several phases, including data collection, preprocessing, model training, evaluation, and deployment.

1. **Data Collection and Preprocessing:**
    - **Dataset:** Publicly available speech datasets such as the **RAVDESS** dataset (Ryerson Audio-Visual Database of Emotional Speech and Song) or **EMO-DB** are used, containing audio samples labelled with different emotional categories.
    - **Preprocessing:** Raw audio files are preprocessed to remove noise and normalise the signals. The preprocessing steps include:
        - **Resampling:** Converting audio to a consistent sample rate for easier processing.
        - **Noise Reduction:** Applying noise filters to enhance audio clarity.
        - **Segmentation:** Dividing long audio files into smaller segments for better feature extraction.
        - **Feature Extraction:** Extracting features like **Mel Frequency Cepstral Coefficients (MFCCs)**, **Chroma features**, **Spectral Contrast**, and **Zero-Crossing Rate**. These features represent essential

characteristics of the audio signal that are crucial for emotion classification.

2. **Model Development:**
    - **Convolutional Neural Networks (CNN):** CNNs are used to capture local patterns in the audio features extracted from the speech. A CNN architecture is used because it excels at identifying patterns in image-like data, and audio features can be treated in a similar way.
        - **Model Architecture:**
            - Input Layer: Audio features like MFCCs are fed into the network.
            - Convolutional Layers: These layers automatically detect relevant features in the spectrogram.
            - Pooling Layers: These layers reduce the dimensionality, preserving the most important information.
            - Fully Connected Layers: These layers make predictions based on the features extracted by the CNN layers.
            - Output Layer: The final output is a softmax layer that classifies the audio input into one of the predefined emotional categories.
3. **Model Training and Evaluation:**
    - The model is trained using a training set of audio files, and its performance is validated using a separate validation set.
    - **Loss Function:** Categorical cross-entropy is used as the loss function, as the output is a multi-class classification problem.
    - **Optimizer:** Adam optimizer is used to minimise the loss function.
    - **Metrics:** The model's performance is evaluated based on metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and a **confusion matrix**.
    - **Cross-validation:** K-fold cross-validation is used to ensure the model generalises well on unseen data.
4. **Visualisation and Results:**
    - Training and validation loss/accuracy are plotted over epochs to visualise the learning progress and detect potential overfitting.
    - A **confusion matrix** is plotted to visualise the classification performance of the model, showing how well each emotion category is predicted.
    - A detailed analysis of false positives and false negatives is carried out to identify areas where the model can be improved.
5. **Deployment (Optional):**
    - After successful training and evaluation, the model can be integrated into an application where users can upload audio files and receive real-time emotion predictions.
    - The model can be deployed using a web interface (using Flask or Django) or a desktop application (using Tkinter or PyQt).



Technologies and Tools:

- **Programming Language:** Python
- **Libraries:**
    - **TensorFlow/Keras:** For building and training the deep learning model.
    - **librosa:** For audio processing and feature extraction (e.g., MFCCs).
    - **NumPy/Pandas:** For data manipulation and handling arrays.

    - **Matplotlib/Seaborn:** For visualisations, such as plotting accuracy/loss graphs and confusion matrices.
- **Development Environment:** Jupyter Notebook, Google Colab, or PyCharm
- **Version Control:** Git for code versioning and GitHub for project collaboration and sharing.

Challenges Faced:

1. **Data Imbalance:** Some emotional categories were overrepresented while others had fewer samples. Techniques like oversampling or class-weighting were used to mitigate this imbalance.
2. **Feature Selection:** Audio signals are complex, and selecting the right features from raw audio data is critical. The model required iterative experimentation with different sets of features to find the most relevant ones.
3. **Overfitting:** The model tended to overfit on the training data initially, which was mitigated by adding dropout layers and using early stopping.
4. **Real-Time Processing:** While this project was designed for batch processing, real-time emotion detection would require additional optimizations and would be part of future work.

Results:

- **Accuracy:** The CNN-based model achieved an accuracy of over 85% on the test dataset, classifying emotions like happiness, sadness, and anger with relatively high accuracy.
- **Precision/Recall:** These metrics provided insights into how well the model performed across different emotional categories, with the best results achieved for happy and sad emotions.
- **Confusion Matrix:** The confusion matrix revealed that the model sometimes confused certain emotions, such as anger and fear, which may have similar vocal tones.

Conclusion:

The Emotion Classification from Audio Files project demonstrates the effectiveness of deep learning models, specifically CNNs, in classifying emotions from speech data. The model performed well across several emotional categories, providing a solid foundation for real-time applications in areas like customer service, healthcare, and HCI. This model has the potential for further improvements, including exploring other architectures (e.g., RNNs or Transformers), fine-tuning hyperparameters, and expanding the dataset for better performance.

Future Work:

1. **Real-Time Emotion Detection:** Enhance the system to process live audio feeds and classify emotions in real-time.

2. **Multimodal Emotion Recognition:** Combine audio with visual data (e.g., facial expressions) to improve overall emotion detection accuracy.
3. **Model Optimization:** Experiment with more advanced models like LSTM or transformers that can handle sequential data better and capture more complex dependencies in speech.
4. **Larger Dataset:** Incorporate a more diverse dataset containing various languages and accents to improve model robustness and generalisation.

References:

1. **Eckart, R. (2019).** *Speech Emotion Recognition Using Deep Learning Algorithms*. International Journal of Computer Applications, 176(8), 32-37.
    - This paper discusses the application of deep learning algorithms to speech emotion recognition, outlining methods for feature extraction and classification in the context of real-world audio datasets.
2. **Yuan, J., & Liu, Z. (2018).** *Speech Emotion Recognition Using CNN and Bi-LSTM*. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1-5.
    - This study investigates the combination of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks for speech emotion recognition, demonstrating the efficacy of hybrid models.

3. **Scherer, K. R., & Bänziger, T. (2010).** *Psychological Perspectives on Emotion*

Speech. In *Handbook of Speech Processing*. Springer.

- This chapter provides in-depth analysis and insights into the psychological and acoustic aspects of emotion in speech, which is fundamental to emotion recognition tasks.

4**. He, R., & Wu, D. (2019).** *A Survey on Emotion Recognition from Speech: Features, Classification Algorithms, and Databases*. Journal of Ambient Intelligence and Humanized Computing, 10(6), 2365-2381.

- The paper offers a detailed review of emotion recognition from speech, focusing on feature extraction techniques, the performance of various classification algorithms, and publicly available datasets.