PROJECT REPORT-EMOTION IDENTIFICATION

INTRODUCTION:

The goal of this project is to produce a system for emotion recognition from audio records. Audio emotion recognition has gradually become an extremely important tool for a host of applications, such as mental health diagnosis, education, customer service, and entertainment—applications increasingly dependent on understanding human emotions for better user experience and outcomes. Emotions are highly critical to human communication, and recognition through audio cues may unlock much of the insights for user sentiments and psychological states. Therefore, we utilize advances in machine learning and audio processing to classify emotions like "happy," "sad," "angry," "fearful," and "neutral" based on the audio characteristics within voice recordings using neural networks and feature extraction techniques.

Emotion recognition from audio data has more practical applications in many industrial fields. For instance, this system can help in mental health evaluation by detecting distress signs or depression signs appearing in a patient's voice. Companies may use customer service to detect frustration or satisfaction for timely responses to customer needs. This project may also have applications in educational technology, where it would be highly relevant to understand the emotional state of a student and thereby make the learning experience more personalized. Such systems are important for developing emotional understanding and engagement across various sectors in today's increasingly virtual world.

DATASET:

DATASET USED : CREMA-D

Dataset: https://github.com/CheyneyComputerScience/CREMA-D

The primary dataset used in this project is the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D), specifically designed to support research in emotion recognition through audio. CREMA-D captures a broad variety of expressions and is well-suited for creating models that generalize across diverse emotions and speakers. The dataset includes 91 actors—48 males and 43 females—ranging in age from 20 to 74, and representing African American, Asian, Caucasian, Hispanic, and other unspecified backgrounds. This wide demographic variety ensures that the dataset captures a realistic and nuanced range of emotions, helping to reduce overfitting risks in machine learning models trained on it.

CREMA-D consists of 7,442 audio clips, with each clip labeled with one of six unique emotions: Anger, Disgust, Fear, Happy, Neutral, and Sad. To capture different intensities of emotional expressions, these emotions are further classified into four levels of intensity: Low, Medium, High, and Unspecified. This nuanced categorization enables the dataset to capture subtle variations in emotion, allowing the model to detect not only the type of emotion but also its intensity, adding depth and granularity to emotion recognition tasks. Each audio clip is generated from one of 12 standardized sentences spoken with varied emotional tones, creating approximately 1,240 distinct utterances per emotion across the intensity levels.

The diversity in speakers, emotions, and expression intensities makes CREMA-D a robust dataset for training well-generalized emotion recognition models. The high level of variation in both speaker demographics and emotional intensity helps prevent models from overfitting, allowing them to perform well on new, unseen data.

The following table outlines the key statistics of the dataset,providing a clear view of its structure and variety for better understanding and reference in the context of this project

Attribute	Value
Total clips	7442
No.of Speakers	91(48male,43female)
Age Range	20-74years
Emotions	6(Anger, Disgust, Fear, Happy, Neutral, Sad)
Emotion Intensity Levels	4(Low,Medium,High,Unspecified)
Sentences	12

EXPERIMENTAL SETUP:

Our developed method includes several important stages: feature extraction, model selection, and training configuration. The raw audio files of the CREMA-D dataset were primarily used to extract meaningful features that capture subtlety in emotional expressions. This system employs MFCCs widely recognized in speech processing due to their capability to capture the timbre and texture of audio. Along with that, we extract features like chroma, spectral contrast, and tonnetz, enhancing the accuracy of the model with a good, rich audio data representation. These features combined together capture emotion in speech much more effectively than just raw audio alone.

We selected a CNN combined with an LSTM layer to capture the spatial and temporal nature of audio features. CNN is used for spatial representation learning, and the sequential aspects are captured with the LSTM. Thus, we use a model with three convolutional layers each followed by batch normalization and max pooling that progressively extract higher-level audio features, feeding the output of this model into an LSTM layer where the sequential relationships must be learned to understand the progression of emotions in the audio clip. Finally, the fully connected dense layer uses softmax activation to yield the output as a probability distribution across the six emotion classes.

For the training of models, the number of epochs is set to 50. This allows enough iterations for the model to learn and shouldn't lead to overfitting. We use the learning rate as 0.001 with Adam optimizer since this approach allows one to start learning efficiently from the beginning, and this results in adaptive representations for unknown distributions of data. The size of the batch is set to 32: balancing requirements for memory with efficiency in training. To avoid overfitting, we apply dropout rate 0.3 during dense layers and consider only those epochs with validation accuracy via early stopping criteria. The splits of training-validation are 80–20 in order to get strong evaluations of our model performance.Besides, we applied data augmentation techniques such as pitch shifting and time stretching on the audio samples for better generalization of models.

RESULTS:

Our model of recognition based on CNN and LSTM layers labeled audio clips with good results and prominent outputs.The model reached an accuracy of 78% on validation over classical models in machine learning, such as Support Vector Machines (SVM)and k-Nearest Neighbors (k-NN).With 62% and 58% SVM and k-NN were not as accurate in well-capturing temporal and spatial features needed for appropriate emotion classification.Now,the improvement clearly points out that this CNN-LSTM model can better handle the complexities of features retrieved from audio.

What's more, the CNN-LSTM model managed high F1-scores for "Happy" and "Neutral." The scores were 0.83 for "Happy" and 0.85 for "Neutral." However, "Fear" and "Disgust" were problematic classes where the model scored poorly: 0.68 for "Fear" and 0.70 for "Disgust."Such overlapping acoustic characteristics in the dataset—similar tonal frequencies and intensity levels—might explain this. This implies that the addition of the LSTM layer enhances the model's accuracy in identifying nuanced differences within emotions. The LSTM layer allows the model to learn from sequential data, capturing subtle temporal cues that traditional classifiers and CNNs fail to recognize.

The table below summarizes the performance metrics across each emotion class:

Emotion	Precision	Recall	F1-Score
Happy	0.81	0.85	0.83
Neutral	0.87	0.84	0.85
Sad	0.77	0.73	0.75
Anger	0.74	0.78	0.76
Disgust	0.69	0.71	0.70
Fear	0.66	0.70	0.68

We tested a few optimizers and learning rates also and noticed that Adam is significantly much more stable than SGD, and it converges much faster. This is also reflected in the results through the loss curve of the model, where we saw the convergence to be faster and the optimization smoother. We also saw how different

batch sizes could affect the models. Again, a batch size of 32 gives a sweet balance between accuracy and the computational cost. The results will be summarized below for the different configurations:

Model Configuration	Optimizer	Batch Size	Accuracy(%)
CNN-LSTM, Adam,Batch=32	Adam	32	78
CNN-LSTM, SGD,Batch=32	SGD	32	72
CNN-LSTM, Adam,Batch=64	Adam	64	74
CNN-LSTM, Adam,Batch=16	Adam	16	76

These results indicate that our chosen CNN-LSTM configuration with Adam optimization provides the best performance. Future refinements could include addressing class imbalances within the dataset and experimenting with additional regularization techniques to further improve generalizability across all emotion classes.

CONCLUSION:

In this project, we developed an emotion recognition model using a CNN-LSTM architecture, trained and evaluated on the CREMA-D dataset. Our model achieved high accuracy, particularly in detecting "Happy" and "Neutral" emotions, demonstrating the effectiveness of combining convolutional and recurrent layers for capturing both spatial and temporal audio features.Despite this success, the model struggled with emotions like "Fear" and "Disgust," indicating potential areas for improvement. Future work could focus on increasing class balance, incorporating data augmentation techniques, and exploring more advanced neural architectures, such as transformers, to further enhance accuracy and generalizability across diverse datasets and real-world applications, such as emotion-driven recommendation systems or virtual assistants.

M.DEEKSHITHAREDDY

(VASAVI COLLEGE OF ENGINEERNG)