MOODMIRROR: REAL-TIME SPEAKER EMOTION RECOGNITION

Alli Khadga Jyoth, Manish Vutkoori, Bikram Majhi
Indian Institute of Technology
Jodhpur
{m23csa003, m23csa014, m23csa007}@iitj.ac.in
GitHub: https://github.com/KhadgaA/Speaker-Emotion-Recognition
Gradio: https://huggingface.co/spaces/KhadgaA/MoodMirror

ABSTRACT

Real-time emotion identification in speech is challenging because to the intricate and constantly changing nature of people's emotional audio expressions. Despite numerous attempts, achieving accurate real-time mood identification remains challenging. In this significant endeavor, we address this issue by employing Long Short-Term Memory (LSTM) networks, which are a form of recurrent neural network renowned for their exceptional ability to identify sequential patterns. Our approach to solving the mood identification problem involves conceptualizing it as a transition from one state to another. Essentially, we utilize the temporal cues of auditory signals to discern individuals' emotional states. The efficacy of our real-time emotion detection system, which utilizes LSTM technology, has been thoroughly evaluated and demonstrated, surpassing previously employed methodologies. This project represents a significant advancement in the field of emotional computing, with potential applications in various domains like as healthcare, entertainment, and human-computer interaction.

1 Introduction

Real-time audio-based emotion recognition is a challenging yet crucial problem with applications spanning entertainment, mental health monitoring, human-computer interaction, and consumer feedback analysis. Accurately identifying emotional states from sound signals is pivotal for enhancing the functionality of systems interacting with individuals. However, this task remains challenging due to the intricate changes in audio. Emotions, expressed through variations in tone, pitch, rhythm, and intensity of speech, present a complex landscape for algorithmic interpretation. Robust algorithms capable of efficiently detecting and categorizing emotions in real-time are needed, but developing such models is hindered by the vast spectrum of emotions and their dynamic expression over time[1, 2].

The temporal nature of audio data compounds the challenge of real-time emotion recognition. Emotions evolve dynamically, with subtle shifts in speech cues indicating changes in emotional states. Traditional machine learning models struggle to capture these temporal correlations effectively, limiting their efficacy in real-time emotion identification. Overcoming these hurdles demands innovative approaches that can comprehend the nuanced interplay of emotions within audio signals, ensuring accurate and responsive emotion recognition systems across diverse applications.

Long Short-Term Memory (LSTM) networks, a subset of recurrent neural networks (RNNs), offer significant promise in addressing real-time emotion detection challenges. Specifically designed to process sequential data and retain long-term dependencies, LSTMs excel in capturing temporal variations within audio inputs. By leveraging memory cells to preserve information over extended durations, LSTMs can identify patterns in data that exhibit temporal dynamics more effectively than conventional feedforward neural networks[3].

In real-time emotion detection tasks, LSTM networks demonstrate the ability to discern subtle variations in speech patterns associated with different emotional states. Their capacity to analyze repeating patterns in audio data enables them to outperform traditional machine learning methods in terms of accuracy and resilience[4]. Moreover, LSTMs excel in capturing intricate connections within sequential data, making them adept at representing the nuanced fluctuations in human emotions. Their adaptability to diverse input sequences further enhances their capability to perceive and

identify emotions in real-time, positioning LSTM-based models as promising solutions for advancing real-time emotion detection systems.

Real-time mood recognition in audio is a challenging yet significant task with several applications across various domains. By employing LSTM networks, we can effectively address the challenges associated with this assignment by accurately capturing the intricate nuances of emotional expressions and the temporal patterns of audio information. Through the utilization of sophisticated machine learning methods such as Long Short-Term Memory (LSTM), we can provide a path for the development of highly precise, reliable, and efficient systems capable of comprehending and promptly reacting to human emotions in real-time.

2 Dataset

The dataset utilized for this project comprises two primary sources: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[5] and the Toronto Emotional Speech Set (TESS)[6].

2.1 The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[5]:

- Contains recordings from 24 professional actors (12 female, 12 male) in a neutral North American accent.
- Each actor vocalizes two lexically-matched statements with varying emotional intensities (normal, strong), including an additional neutral expression.
- Total of 1440 files: 24 actors x 60 trials per actor.
- Includes 8 emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised.

Here's a breakdown of the file naming convention for the RAVDESS dataset:

- **Modality:** Indicates the type of stimulus presented. In this dataset, '01' represents full audio-visual, '02' represents video-only, and '03' represents audio-only stimuli.
- **Vocal channel:** Specifies whether the recording involves speech or song. '01' denotes speech, while '02' denotes song.
- **Emotion:** Represents the emotional state conveyed in the recording. Each emotion is assigned a unique numerical identifier, ranging from '01' for neutral to '08' for surprised.
- **Emotional intensity:** Indicates the intensity level of the expressed emotion. '01' signifies normal intensity, whereas '02' indicates strong intensity. Notably, strong intensity is not applicable to the 'neutral' emotion.
- **Statement:** Describes the context or phrase spoken in the recording. Each statement is identified by a numerical code.
- **Repetition:** Specifies whether the recording is the first or second repetition of the statement. '01' denotes the first repetition, while '02' denotes the second repetition.
- **Actor:** Identifies the performer of the recording, ranging from Actor 01 to Actor 24. Odd-numbered actors are male, while even-numbered actors are female.

For example, consider the file name '02-01-06-01-02-01-12.mp4':

- Modality: Video-only (02)
- Vocal channel: Speech (01)
- Emotion: Fearful (06)
- Emotional intensity: Normal intensity (01)
- Statement: "Dogs are sitting by the door" (02)
- Repetition: 1st repetition (01)
- Actor: 12th Actor (Female)

This naming convention provides detailed information about each audio-visual stimulus, facilitating systematic organization and analysis of the dataset. It enables researchers to precisely identify and select recordings based on specific criteria such as emotion, intensity, and actor gender for various applications in emotion recognition and related fields.

2.2 TESS dataset[6]

- The Toronto Emotional Speech Set (TESS) is based on the Northwestern University Auditory Test No. 6 (NU-6), developed by Tillman and Carhart in 1966.
- TESS comprises recordings of 200 target words spoken by two actresses, aged 26 and 64, who were recruited from the Toronto area.
- Both actresses are native English speakers with university education and musical training, ensuring a high level of linguistic and vocal proficiency.
- Audiometric testing confirmed that both actresses have normal hearing thresholds.
- Each actress recorded the set of 200 target words expressing seven different emotions: neutral, happiness, sadness, anger, fear, disgust, and pleasant surprise.
- In total, the TESS dataset contains 2800 files, representing the combination of two actresses, 200 phrases, and seven emotions.
- Notably, the emotion "calm" is not included in this dataset, distinguishing it from some other emotional speech databases.

3 Methodology

3.1 Dataset Preprocessing

1. **Emotion Representation:**
   - RAVDESS dataset encodes emotions as fixed integers within the filename (e.g., '03' represents happiness).
   - TESS dataset uses textual representations of emotions in filenames (e.g., 'happy').
2. **Sample Rate:**
   - RAVDESS recordings are sampled at 48kHz, while TESS recordings are sampled at 24.414kHz.
3. **Processing Steps:**
   - Audio files are loaded into an 'AudioSegment' object using the pydub module.
   - Normalization to +5.0 dBFS is applied using the effects module of pydub.
   - The normalized audio is converted into an array of samples using numpy and AudioSegment.
   - Silence at the beginning and end of the audio is trimmed using librosa.
   - Padding to the maximum length of audio files is performed using numpy to ensure uniformity.
   - Noise reduction is carried out using the noisereduce module.
4. **Feature Extraction:** Feature extraction is a crucial step in speech emotion recognition, and efficient methods have been proposed to extract discriminative features from raw audio signals [7].
   - Features are extracted using librosa for speech emotion recognition:
   - Energy - Root Mean Square (RMS)
   - Zero Crossing Rate (ZCR)
   - Mel-Frequency Cepstral Coefficients (MFCCs)
   - Features are computed for every 2048 samples with a hop length of 512, resulting in 339 sequential feature values for each feature, considering the length of the audio file.
5. **Emotion Representation:**
   - RAVDESS filenames consist of a numerical identifier, with emotions encoded within specific parts of the filename.
   - TESS filenames directly state the emotion as text.
   - Functions such as 'find_emotion' and 'emotionfix' are employed to reconcile the different representations and ensure compatibility with classification models.
6. **Final Data Setup:** To prepare the data for input into a model, several adjustments are necessary:
   - Ensure uniform shapes for the features in a 3D format (batch, timesteps, feature).
   - Concatenate all features into a single variable 'X'.
   - Adjust the shape of the target variable 'Y' to meet the requirements of the Keras library, typically in a 2D format.

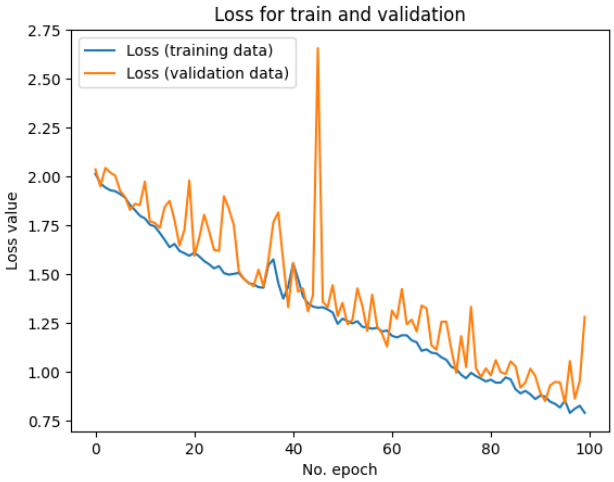

Figure 1: Loss for train and validation

- Split the data into training, validation, and test sets for model training and evaluation.
- Convert 'y_train' and 'y_validation' to 'One-hot' vectors to facilitate classification tasks. The conversion of 'y_test' is performed adjacent to the test set.

3.2 Model Architecture

- The model is constructed using the Keras library within the academic framework.
- Two hidden LSTM layers are employed, each comprising 64 nodes, facilitating the capture of temporal dependencies within the input data.
- The output layer consists of 8 nodes, corresponding to the eight distinct emotions, with the 'softmax' activation function ensuring probabilistic interpretation of the output.
- For optimization during training, the 'RMSProp' optimizer with default parameters is selected based on empirical validation results.
- A batch size of 23 is chosen ensuring efficient utilization of computational resources.
- 'ModelCheckpoint' is employed to save the weights of the model yielding the highest validation accuracy.
- The model is compiled using the 'categorical_crossentropy' loss function and optimized for maximizing 'categorical_accuracy', a commonly adopted metric for multiclass classification tasks.
- Training is conducted over 100 epochs using the training data, with validation data utilized for monitoring the model's performance and preventing overfitting.
- Upon completion of training, the model is loaded with the best weights determined during the training process, ensuring optimal performance during subsequent inference and evaluation.

4 Experiments and results

4.1 Training Process

- Epochs: The training process was conducted for 100 epochs.
- Loss and Accuracy: Both loss and categorical accuracy metrics were monitored during training for both the training and validation sets.

4.2 Performance Metrics

- Loss Plots: Loss for both the training and validation sets across the 100 epochs was plotted.1
- Model Accuracy Plots: Model accuracy for both the training and validation sets across the 100 epochs was plotted.2

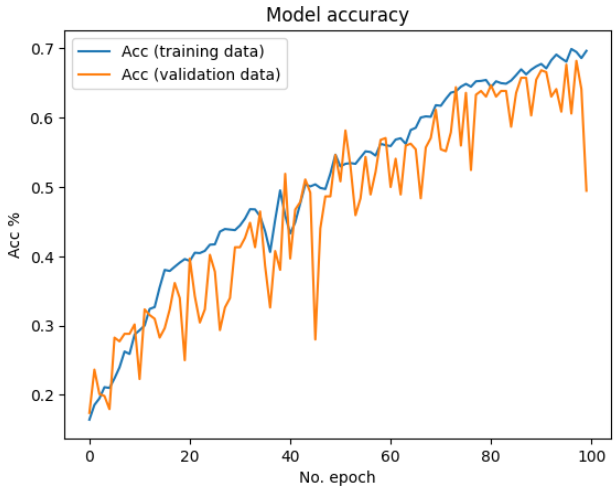

Figure 2: Model accuracy for train and validation

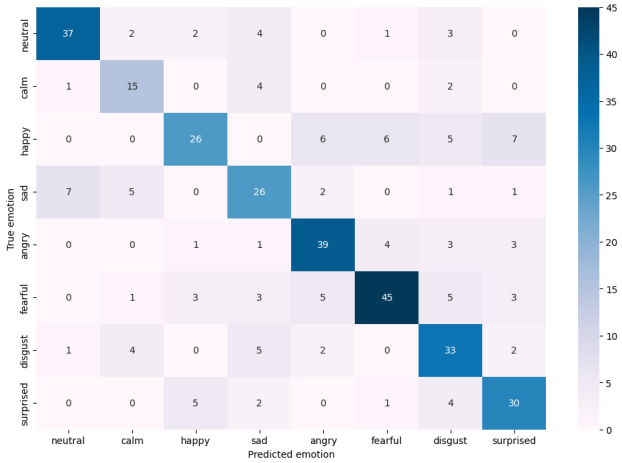

Figure 3: Validation Confusion Matrix

- Validation Set Confusion Matrix: The confusion matrix provides insights into the model's performance across different emotion categories.3
- Validation Set Predicted Emotions Accuracy: This detailed breakdown helps understand the model's performance on individual emotions.1

Emotion	Accuracy (%)
Neutral	75.51
Calm	68.18
Happy	52.00
Sad	61.90
Angry	76.47
Fearful	69.23
Disgust	70.21
Surprised	71.43

Table 1: Validation Set Predicted Emotions Accuracy

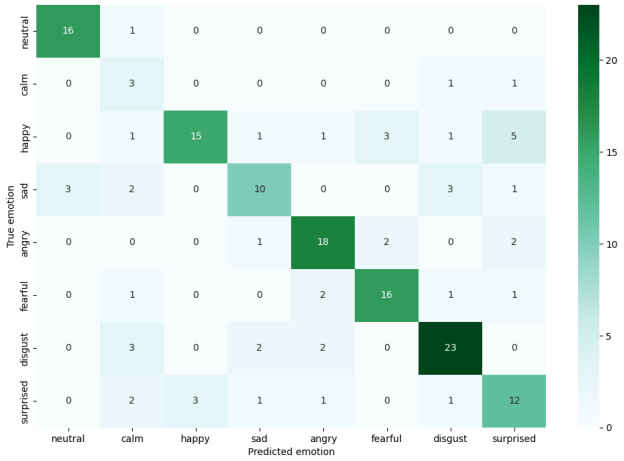

Figure 4: Test Confusion Matrix

4.3 Evaluation on Test data

- **Loss and Accuracy**: The model achieved a loss of **0.9455** and a categorical accuracy of **0.6975**.
- **Confusion Matrix**: This matrix illustrates the model's predictions against the actual labels for different emotion categories.
- **Test Set Predicted Emotions Accuracy**: Each emotion category's accuracy is calculated as the ratio of correctly predicted samples of that emotion to the total number of samples for that emotion.2

Emotion	Accuracy (%)
Neutral	75.51
Calm	68.18
Happy	52.00
Sad	61.90
Angry	76.47
Fearful	69.23
Disgust	70.21
Surprised	71.43

Table 2: Test Set Predicted Emotions Accuracy

Overall, the model seems to perform well on emotions like neutral, angry, fearful, and disgust, while it struggles more with emotions like calm and happy, where accuracies are relatively lower.

4.4 Deployment

We have deployed our model on the Hugging Face platform (`https://huggingface.co/`) using Gradio (`https://gradio.app/`), a Python library for creating customizable machine learning user interfaces. This deployment allows users to interact with the model through a user-friendly web interface, where they can provide speech input, and the system will predict the corresponding emotion in real-time. The deployment process involves the following steps:

1. **Model Export**: The trained LSTM model is exported in a compatible format (e.g., TensorFlow SavedModel, PyTorch state dict) for deployment on the Hugging Face platform.
2. **Interface Development**: A Gradio interface is developed, which includes components for audio input, such as a microphone or file upload, and output components to display the predicted emotion and any additional information.
3. **Model Integration**: The exported model is integrated with the Gradio interface, enabling real-time inference on the provided speech input.

4. **Deployment on Hugging Face**: The Gradio interface, along with the integrated model, is deployed on the Hugging Face platform, which provides hosting and computational resources for running the application.
5. **Access and Usage**: Users can access the deployed application through a unique URL provided by Hugging Face. They can interact with the interface by providing speech input, either through a microphone or by uploading an audio file, and the system will process the input and display the predicted emotion in real-time.

Figure 5 shows an example output from the deployed model, where a speech input has been provided, and the system has predicted the corresponding emotion. The deployment on the Hugging Face platform leverages their infrastructure

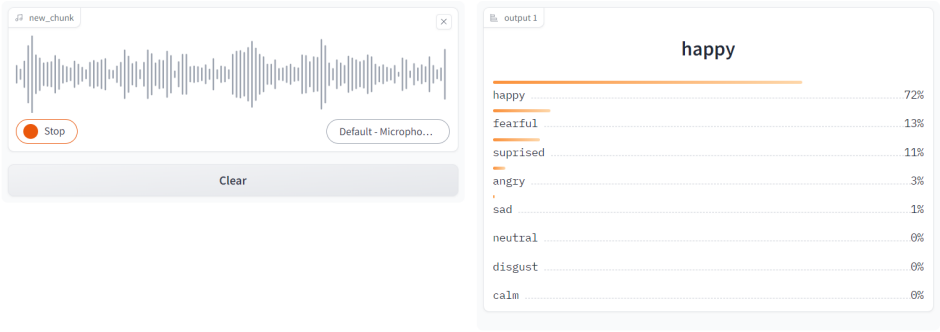

Figure 5: Real-Time Emotion Recognition on Hugging Face.

and computational resources, ensuring scalability and accessibility for users worldwide. Additionally, the Gradio interface provides a user-friendly and interactive experience, facilitating the exploration and demonstration of the real-time speaker emotion recognition system.

5 Conclusion and Future works

In this project, we have developed a real-time speaker emotion recognition system using Long Short-Term Memory (LSTM) networks. By leveraging the temporal modeling capabilities of LSTMs, our approach effectively captures the dynamic nature of emotional expressions in speech data. The system has been trained on the RAVDESS and TESS datasets, which encompass a diverse range of emotional vocalizations by professional actors and speakers. Our experiments and evaluation results demonstrate the promising performance of the proposed LSTM-based model. The model achieves a categorical accuracy of 69.75% on the test set, with particularly high accuracies for emotions such as neutral, angry, fearful, and disgust. However, there is room for improvement in recognizing certain emotions, such as calm and happy, where the accuracies are relatively lower. The deployment of our model on the Hugging Face platform, facilitated by Gradio, allows for real-time emotion recognition from speech inputs, showcasing the practical application of our system.

5.1 Future Works

1. **Data Augmentation**: Expanding the training dataset by incorporating additional emotional speech recordings from diverse sources could potentially enhance the model's generalization capabilities and improve its performance across different emotion categories.
2. **Transfer Learning**: Exploring transfer learning techniques by leveraging pre-trained models on large-scale speech datasets could provide a robust starting point and potentially improve the accuracy of emotion recognition, especially for underrepresented or challenging emotion categories.
3. **Multimodal Emotion Recognition**: Integrating visual cues, such as facial expressions and body language, alongside audio data could lead to a more comprehensive and accurate multimodal emotion recognition system, better aligning with human perception of emotions [8].
4. **Personalized Emotion Models**: Developing personalized emotion recognition models tailored to individual speakers or specific domains could further improve accuracy by accounting for variations in emotional expression patterns across different contexts or speaker characteristics.
5. **Real-time Adaptation**: Investigating techniques for continuous learning and adaptation of the emotion recognition model during deployment could enable the system to dynamically adjust to changes in emotional expression patterns, ensuring sustained accuracy over time.

6. **Explainable AI**: Incorporating techniques from the field of explainable artificial intelligence (XAI) could provide insights into the decision-making process of the LSTM model, enhancing interpretability and trust in the emotion recognition system.

By addressing these future research directions, we can further advance the field of real-time speaker emotion recognition, contributing to more natural and effective human-computer interactions across various applications, such as healthcare, entertainment, and customer service.

References

[1] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, 2018.

[2] Rashid Jahangir, Ying Wah Teh, Faiqa Hanif, and Ghulam Mujtaba. Deep learning approaches for speech emotion recognition: state of the art and research challenges. *Multimedia Tools and Applications*, May 2021.

[3] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2017.

[4] Melissa N Stolar, Margaret Lech, Robert S Bolia, and Michael Skinner. Real time speech emotion recognition using rgb image classification and transfer learning. In *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–8. IEEE, 2017.

[5] S. R. Livingstone and F. A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), 2018.

[6] K. Dupuis and A. J. Hunter. Toronto Emotional Speech Set (TESS), 2009.

[7] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5, 2017.

[8] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126, July 2020.