Recognition of Emotions in Speech using Neural Networks

Aubin Bonnefoy aubin.bonnefoy@utbm.fr
Simon Illouz-Laurent simon.illouz-laurent@utbm.fr

Abstract

This study explores the recognition of emotions in speech using neural networks. We use transformer-based techniques to analyze and classify emotional states expressed in speech recordings. For more information and source code, visit our GitHub repository: https://github.com/aubinbnf/Speech-Emotion-Recognition.

Contents

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

The automatic recognition of emotions from speech has become a critical area of research in human-computer interaction and AI-driven emotional intelligence. The ability to accurately recognize emotional states in speech can significantly enhance the capabilities of virtual assistants, improve the naturalness of human-machine communication, and offer personalized user experiences across a variety of domains. In fields such as customer service, healthcare, and mental health monitoring, recognizing emotions from speech can help identify distress or frustration, allowing systems to adapt in real-time to user needs and improving overall interaction quality.

Traditionally, methods such as Support Vector Machines were commonly used for emotion classification tasks, relying on manually extracted acoustic features. However, these approaches often struggled with the complexity of emotional expression in speech, particularly when dealing with subtle or overlapping emotions. The advent of deep learning and the rise of neural networks have transformed the field, enabling models to automatically learn hierarchical representations from raw data, significantly improving the accuracy and robustness of emotion recognition.

This project aims to develop a model capable of accurately identifying emotions from audio-only speech recordings, focusing on state-of-the-art deep learning techniques. Using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), specifically its speech-only subset, the project explores a variety of emotional expressions performed by professional actors.

The primary objective is to apply cutting-edge techniques that have demonstrated superior performance in capturing both short-term and long-term dependencies in audio signals. By leveraging fine-tuning on pre-trained models, this project aims to achieve robust emotion recognition, with the potential to advance the practical application of emotion-aware systems in real-world scenarios. The success of this approach would not only contribute to the field of speech emotion recognition (SER) but also pave the way for more sophisticated emotion-driven AI systems in the future.

2 Related Work

Speech Emotion Recognition relies on several key steps. The data first comes from a dataset, which provides the necessary audio recordings. Then, preprocessing is applied to clean and prepare the data. Feature extraction follows, where relevant information such as pitch, intensity, and frequency patterns are extracted. Finally, classification is performed to categorize the emotions present in the speech. These steps work together to produce the final result of emotion recognition.

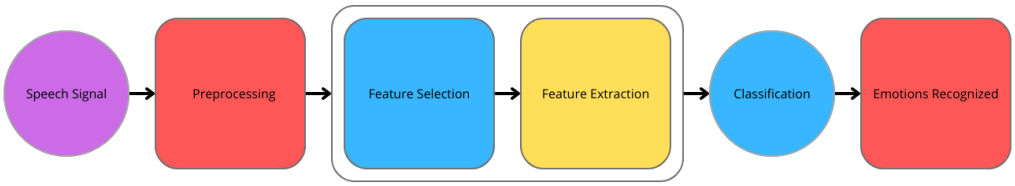


Figure 1: Speech Emotion Recognition System

2.1 Emotion Recognition Datasets

The choice of datasets plays a crucial role in the success of speech emotion recognition models, as the quality and diversity of the data directly influence their ability to generalize. These datasets are primarily divided into three types: simulated datasets, where emotions are expressed in a controlled manner by actors, often in a studio setting, and are typically available for free or upon request; induced datasets, which capture genuine emotions elicited in specific scenarios, and are usually collected in academic or controlled experimental environments; and natural datasets, recorded in real-world contexts, reflecting spontaneous emotions, but often accompanied by noise, variations in emotional expression, and cultural or linguistic differences. These datasets also differ in terms of

language, size, and accessibility, all of which impact the performance and robustness of the trained models.

Table 1: Comparison of databases used for Speech Emotion Recognition

Database	Emotions	Language	Category	Access
RAVDESS dataset (2018)	Anger, Disgust, Fear, Happiness, Neutral, Calm, Surprise, Sadness. (1440 audio files)	English	Simulated	Free
RAVDESS SONG (2018)	Calmness, Happiness, Sadness, Anger, Fear, Surprise, Disgust. (1012 audio files)	English	Simulated	Free
MSP-Podcast (2019)	Anger, Happiness, Sadness, Neutral, Disgust. (62,140 speaking turns, 100 h)	English	Natural	Academic request
SAVEE (2009)	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral. (480 audio files)	English	Simulated	Free
Crema-D dataset (2014)	Anger, Disgust, Fear, Happiness, Sadness, Neutral. (7442 audio files)	English	Simulated	Free
IEMOCAP (2008)	Anger, Happiness, Neutral, Disgust, Fear, Sadness, Excited. (12 h of audiovisual data, including 10,039 utterances)	English	Induced	Academic request
ASVP-ESD (2021)	Boredom, Neutral, Fear, Sadness, Happiness, Anger, Surprise, Disgust, Excitement, Pleasure, Pain, Disappointment. (13,285 audio files collected)	Chinese, English, French, Russian	Natural	Free
CASIA (2003)	Happiness, Sadness, Anger, Surprise, Fear, Neutral. (1200 audio files)	Chinese	Simulated	Free
EMODB (2005)	Anger, Fear, Disgust, Sadness, Happiness, Boredom, Neutral. (535 audio files)	German	Simulated	Academic request
ESD (2021)	Anger, Happiness, Neutral, Sadness, Surprise. (29 h, 3500 sentences)	English	Simulated	Free
BAUM-1s (2018)	Anger, Disgust, Fear, Joy, Sadness, Surprise. (300 audio files)	English	Simulated and Natural	Academic request
TESS (2010)	Anger, Boredom, Fear, Happiness, Neutral, Sadness. (2800 audio files)	English	Simulated	Free

It is important to note that, although the available datasets cover a wide range of emotions and contexts, they present specific challenges that influence model performance. Simulated datasets, while accessible and controlled, sometimes lack the richness of authentic emotions observed in real-world contexts. In contrast, induced datasets, which capture more natural emotions, may suffer from the subjectivity of experimental situations. Natural datasets, often drawn from real and spontaneous contexts, reflect more authentic emotions but can be disrupted by environmental noise, variations in emotional expression, or cultural and linguistic differences. These various types of datasets require

careful consideration when choosing, depending on the research goals, modeling techniques, and specific requirements of emotion recognition systems.

2.2 Preprocessing techniques

The preprocessing stage in SER systems enhances audio quality and prepares data for feature extraction or direct analysis. Common techniques include silence removal, which eliminates irrelevant pauses, and noise reduction, targeting background interference. Normalization, such as amplitude or Z-normalization, ensures consistency across datasets, while data augmentation (e.g., adding noise or pitch shifting) expands dataset variability and robustness.

For spectrogram-based approaches, methods like scaling, segmentation, and windowing refine visual representations, aiding in emotion analysis. However, this stage is sometimes bypassed, especially when working with raw signals or spectrograms, to preserve potentially valuable information. The choice of techniques depends on the system's objectives and can significantly impact model performance and efficiency.

2.3 Feature extraction

Feature extraction is a crucial step in speech emotion recognition, transforming audio signals into data that can be used for machine learning. It relies on different categories of features: **prosodic features**, such as pitch, intensity, and duration, which reflect dynamic variations in emotional speech; **spectral features**, like MFCCs (Figure 3), which provide a detailed representation of frequencies; and **voice quality features**, which analyze aspects like timbre or roughness of the voice.

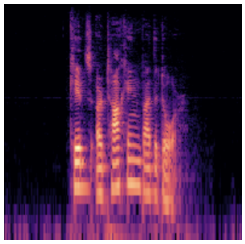


Figure 2: Spectrogram Representation

Figure 3: MFCC Representation

Spectrograms (Figure 2) and mel-spectrograms, as time-frequency representations, are also widely used, particularly in deep learning approaches. These representations capture emotional cues in audio signals and facilitate the use of neural models such as CNNs or ResNet-like architectures, which can automatically extract complex features.

Some studies adopt a hybrid approach, combining manually extracted features (prosodic and spectral) with visual representations like spectrograms to enrich the available information. Others incorporate advanced techniques such as audio segmentation or data normalization to ensure the consistency and relevance of the selected features.

The rise of deep learning has also enabled the automatic extraction of features directly from raw data, eliminating the need for predefined traits. While powerful, these techniques require diverse and high-quality datasets to ensure robust generalization, highlighting the importance of feature selection in the success of speech emotion recognition.

2.4 Classification of emotions

Deep learning-based approaches have revolutionized the field of Speech Emotion Recognition (SER). With their ability to automatically learn relevant representations from raw data, these methods significantly outperform traditional techniques based on manually extracted features. By leveraging advanced architectures, SER systems can capture complex nuances in audio signals, such as the temporal, spectral, and prosodic variations that characterize human emotions. In this section, we examine the main deep learning architectures used in SER, highlighting their contributions, advantages, and limitations.

2.4.1 Convolutional Neural Networks (CNN)

For speech emotion recognition tasks, CNNs have proven particularly effective due to their ability to automatically extract relevant features from data. Instead of working directly with raw audio signals, CNNs leverage spectral representations such as spectrograms (Figure 2) or Mel-frequency cepstral coefficients (MFCCs) (Figure 3), which encapsulate essential temporal and frequency information for distinguishing human emotions.

These representations are treated as two-dimensional images, allowing CNNs to apply filters to detect local patterns. The ability to automatically learn relevant features is a hallmark of deep neural networks in general. However, CNNs are particularly effective for speech emotion recognition tasks due to their capacity to exploit the spatial structure of spectral representations. The initial convolutional layers identify simple variations in the frequency spectrum or changes in sound intensity, focusing on local patterns such as peaks or transitions. Deeper layers, on the other hand, extract more abstract and complex relationships between these elements, capturing specific aspects of emotions, such as intensity, pitch, or rhythm. This hierarchical learning process enables CNNs to process key spatial relationships inherent in spectrograms and MFCCs, which are essential for distinguishing subtle emotional cues. Furthermore, their robustness to local variations in the input data makes CNNs well-suited to handle diverse vocal expressions and variations in speaker characteristics.

A notable illustration of this efficiency is the Deep Stride CNN (Figure 4) model proposed in the literature [7]. Designed specifically for speech emotion recognition, this model stands out due to its optimized architecture, which eliminates traditional pooling layers and instead uses strides directly in convolutional layers to reduce feature map sizes. Comprising nine layers in total (seven convolutional and two fully connected), the DSCNN employs small filters (3×3) and batch normalization to regularize the model.

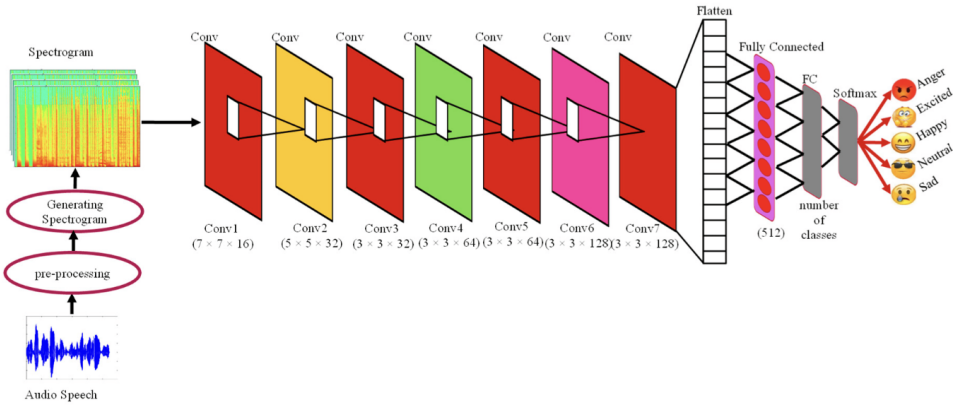


Figure 4: Deep stride CNN architecture [7]

Training DSCNN on spectrograms generated from audio files demonstrated promising results on datasets such as RAVDESS and IEMOCAP. Specifically, the model achieved an accuracy of 70% when trained on raw spectrograms, which improved to 81% when clean spectrograms were used. Notably, the precision, recall, and F1-score varied across emotions, with strong performance for emotions like Sad (F1-score: 0.98 on raw and 0.93 on clean spectrograms) and Disgust (F1-score: 0.89 on raw spectrograms). However, the model struggled with certain classes like Fearful (F1-score: 0.00 on raw spectrograms). These results highlight DSCNN's ability to extract robust discriminative features and achieve competitive accuracy, particularly when trained on pre-processed inputs.

2.4.2 Long short-term memory (LSTM)

LSTM networks are particularly well-suited for speech emotion recognition due to their ability to process sequential data and capture long-term temporal dependencies. Unlike CNNs, which excel at extracting local features from spectrograms or MFCCs, LSTMs can directly model temporal relationships, which is essential for recognizing emotions that often evolve over time. LSTMs can retain important information over long periods, making them better equipped to capture subtle

Table 2: Deep stride CNN architecture results on spectrograms [7]

Nature	**Result on Raw Spectrograms**			**Result on Clean Spectrograms**		
Emotion	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Anger	0.40	1.00	0.57	0.79	0.91	0.84
Happy	0.92	0.29	0.44	0.79	0.90	0.84
Neutral	0.91	0.42	0.57	0.71	1.00	0.83
Sad	0.98	0.98	0.98	0.90	0.96	0.93
Calm	0.82	0.75	0.78	0.71	0.94	0.81
Fearful	0.00	0.00	0.00	1.00	0.50	0.67
Surprised	0.90	0.46	0.61	0.89	0.87	0.88
Disgust	0.92	0.86	0.89	1.00	0.38	0.55
Accuracy	-		0.70	-		0.81

emotional variations that manifest throughout speech. They are also capable of handling emotion variability across different speakers and expressions, offering an advantage over CNNs, which may struggle to model these complex temporal dynamics.

A study on the application of LSTMs to vocal emotion recognition [9] proposes a model consisting of two hidden LSTM layers, each with 64 units, followed by a dense output layer with 8 units corresponding to different emotions.

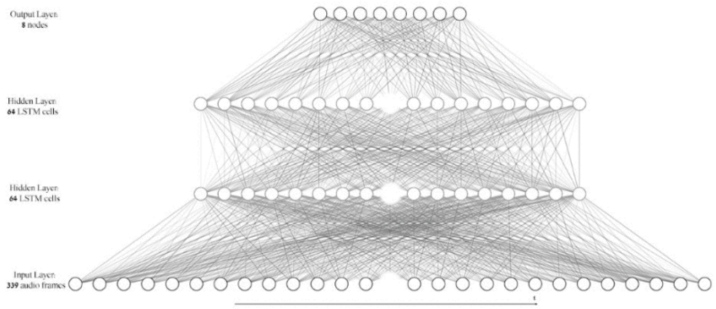


Figure 5: LSTM architecture [9]

The model uses a softmax activation function at the output to classify emotions. The results of the study show an overall accuracy of 82.2%, highlighting the effectiveness of LSTMs for vocal emotion recognition, especially when combined with careful preprocessing and appropriate feature extraction.

2.4.3 CNN-LSTM

The combination of CNN and LSTM networks has emerged as a robust architecture for speech emotion recognition, effectively combining the feature extraction capabilities of CNNs with the sequence modeling strengths of LSTMs. This hybrid architecture enables the analysis of both local and temporal features, providing a comprehensive understanding of the emotional content in speech signals.

In the CNN-LSTM model, CNN layers typically extract meaningful representations of audio features, such as spectrograms or MFCCs, while the LSTM layers capture the temporal dependencies and long-term patterns within the data. This dual functionality makes CNN-LSTM architectures particularly well-suited for tasks that require both detailed feature analysis and sequential context modeling.

Recent studies have highlighted the effectiveness of advanced CNN-BLSTM designs for emotion recognition [10]. For instance, an enhanced model (Figure 6) incorporating CNN-2D, LSTM, and an attention mechanism was found to outperform simpler CNN or LSTM architectures. This specific architecture utilized four convolutional blocks, each comprising layers for convolution, batch normalization, max-pooling, and dropout.

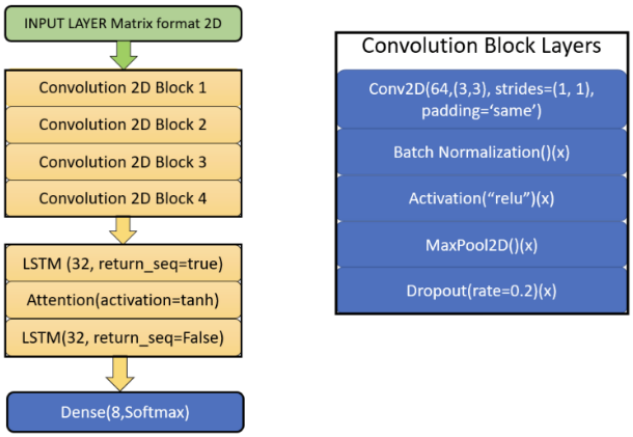


Figure 6: CNN-LSTM architecture [10]

The LSTM layers in the model refined the extracted features by learning non-linear dependencies and capturing temporal dynamics. Additionally, the inclusion of an attention mechanism further enhanced the architecture's performance by assigning weights to critical features and reducing the impact of redundant information. This mechanism allowed the model to focus on segments of the speech signal that carried significant emotional cues while preserving long-term dependencies. Ultimately, the architecture concluded with a dense layer using SoftMax activation to classify emotions with high accuracy.

Table 3: CNN-LSTM architecture results [10]

Emotional Label	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Angry	94	90	92
Calm	81	93	86
Disgust	90	94	92
Fear	93	86	89
Happy	88	88	89
Neutral	89	93	91
Sad	93	86	89
Surprise	89	93	91
Accuracy		90	

The model achieved an overall accuracy of 90%, with strong performance across all emotions. High F1-scores were observed for Angry and Disgust (92%), while Calm showed slightly lower precision (81%) but high recall (93%). Emotions like Fear, Sad, and Happy had balanced F1-scores around 89%, indicating consistent classification. Neutral and Surprise performed well with F1-scores of 91%, showcasing the model's robustness in recognizing a wide range of emotional states.

The performance of different architectures was tested across three datasets: RAVDESS, SAVEE, and TESS. For RAVDESS, the CNN-2D + LSTM + Attention model achieved the highest accuracy at 74.44%, slightly better than CNN-2D at 73.70%. In contrast, on SAVEE, all models showed lower accuracy, with CNN-2D achieving 60.00%, and the CNN-2D + LSTM and CNN-2D + LSTM + Attention models performing similarly at around 58-59%. On TESS, all models performed exceptionally well, with CNN-2D + LSTM + Attention reaching 99.81%. Finally, when combining the datasets (RAV + SAVEE + TESS), the CNN-2D + LSTM + Attention model outperformed the others with an accuracy of 90.19%, demonstrating superior generalization.

Table 4: Comparison of Accuracy Across Models and Datasets [10]

Dataset	CNN-2D	CNN-2D + LSTM	CNN-2D + LSTM + Attention
RAVDESS	73.70%	70.37%	74.44%
SAVEE	60.00%	58.05%	57.50%
TESS	99.71%	99.76%	99.81%
RAV + SAVEE + TESS	85.56%	86.92%	90.19%

In summary, the performance of various architectures, including CNN-2D, CNN-2D + LSTM, and CNN-2D + LSTM + Attention, on different datasets demonstrates the versatility and challenges of combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks, particularly for speech emotion recognition.

2.4.4 Transformers

In recent years, Transformers have emerged as a powerful architecture in SER. Originally developed for Natural Language Processing (NLP) tasks, Transformers excel at modeling long-range dependencies in sequences, outperforming traditional models like Recurrent Neural Networks (RNNs). The key innovation of the Transformer lies in its self-attention mechanism, which allows the model to focus on different parts of the input with varying importance, capturing both local and global patterns simultaneously.[11].

When applied to SER, Transformers efficiently model the temporal dynamics of emotional expressions in speech. Unlike RNNs, which process sequences step-by-step, Transformers process the entire sequence in parallel, leading to faster training and improved scalability. The attention mechanism further enables Transformers to capture nuanced relationships in the speech signal, such as shifts in tone, pitch, and intensity, which are essential for emotion detection.

Several state-of-the-art models in SER have successfully integrated Transformer architectures. Models like the Speech-Transformer leverage attention mechanisms to analyze spectrograms and raw audio data, focusing on emotionally relevant segments of speech. Hybrid models combining CNNs for feature extraction with Transformers for sequence modeling have shown enhanced performance, especially in tasks requiring long-term dependencies and subtle emotional cues.

Here is a benchmark summarizing recent advancements in Speech Emotion Recognition (SER) models from 2017 to 2024. The table includes various datasets, classification methods, and their corresponding accuracy results. These models show the progress made in applying different deep learning architectures to emotion recognition in speech.

Year	Dataset	Classification
2017	EMODB	DNN
2018	CASIA	CNN, DNN
2019	IEMOCAP	RNN
2019	IEMOCAP, EMODB	Multi-CNN dilated
2020	IEMOCAP, RAVDESS	CNN, DSCNN
2023	RAVDESS, SAVEE, TESS	Combination of CNN-2D and LSTM with an attention layer
2024	IEMOCAP	Residual Attention Multi-Layer Perceptron (RA-GMLP)

Table 5: Dataset and Classification

Year	Accuracy
2017	96.97%
2018	Max (Happiness): 98.51% Min (Neutral): 93.91%
2019	63.00%
2019	IEMOCAP: 74.96% EMODB: 90.78%
2020	IEMOCAP: 84.00% RAVDESS: 80.00%
2023	Combined (RAVDESS + SAVEE + TESS): 90.19%
2024	WA: 75.31% UA: 75.09%

Table 6: Accuracy

2.5 Conclusion of Related Work

Through this review of related works, we have demonstrated the evolution of emotion recognition techniques in speech, transitioning from traditional methods such as SVMs and HMMs to modern neural network architectures. CNNs, RNNs, and LSTMs have each made significant contributions, particularly in processing acoustic features and temporal dynamics of vocal signals. However, the emergence of Transformers opens new perspectives, offering increased efficiency and accuracy for emotion recognition in speech. The combination of these approaches could potentially lead to even better performance by integrating the strengths of each model for a finer analysis of expressed emotions.

3 Experiment Approach

3.1 Creation of the training dataset

For this task, it was necessary to create a custom dataset by combining publicly available audio data.

3.1.1 Overview of Datasets used

For this speech emotion recognition task, we used seven public datasets: RAVDESS, SAVEE, CREMA-D, JL Corpus, ESD, and TESS. These databases are widely used in research on emotion classification from audio signals. Their combination increases the diversity of speakers, recording conditions, and emotional classes, offering a varied and representative dataset for training emotion recognition models.

The table below presents the main characteristics of the different datasets used to constitute the project's database:

Table 7: Comparison of databases used for Speech Emotion Recognition

Database	**Emotions**	**Number of audio files**	**Language**	**Category**
RAVDESS (2018)	Anger, Disgust, Fear, Happiness, Neutral, Calm, Surprise, Sadness.	1248	English	Simulated
SAVEE (2009)	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral.	480	English	Simulated
Crema-D (2014)	Anger, Disgust, Fear, Happiness, Sadness, Neutral.	6000	English	Simulated
ESD (2021)	Anger, Happiness, Neutral, Sadness, Surprise.	17500	English	Simulated

Table 7: Comparison of databases used for Speech Emotion Recognition

Database	**Emotions**	**Number of audio files**	**Language**	**Category**
TESS (2010)	Anger, Boredom, Fear, Happiness, Neutral, Sadness.	1400	English	Simulated
JL Corpus (2018)	Anger, Sadness, Neutral, Happiness, Excitement, Anxiety, Thoughtfulness, Enthusiasm, Worry	204	English	Simulated

Once these datasets were combined, we obtained a large dataset consisting of 26,832 audio files, distributed across the seven emotions: "Angry," "Surprise," "Happy," "Sad," "Disgust," "Neutral," and "Fearful."

3.1.2 Datasets balancing

Initial analysis of the combined dataset revealed an unequal distribution of samples for each emotion. Specifically, emotions such as Angry, Happy, and Neutral were overrepresented, while others like Disgust and Fearful had significantly fewer samples. This imbalance posed challenges for training neural models, as it could lead to biased predictions favoring the more prevalent emotions.

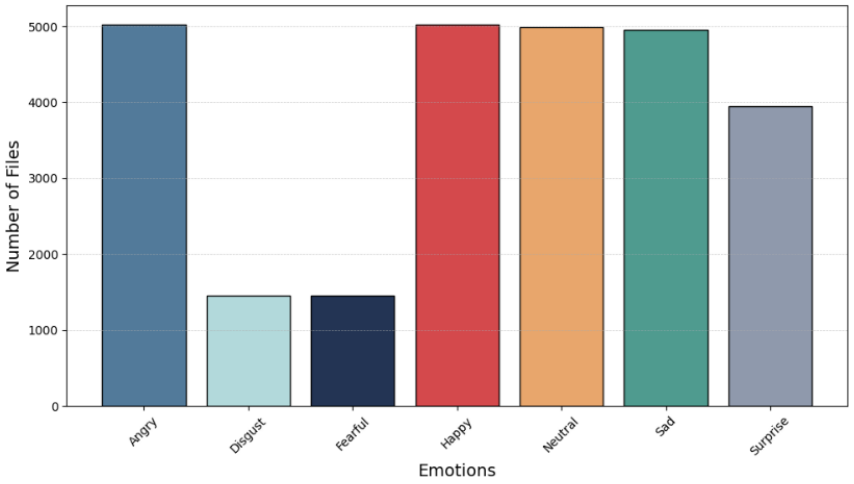


Figure 7: Distribution of Audio Files per Emotion

To address this imbalance, two strategies were implemented:

1) Under-sampling

The first choice was to apply under-sampling to reduce the number of samples from over-represented classes such as "Angry", "Happy", "Neutral", "Sad" and "Surprise" by selecting an equal number of examples for each emotion, set to 1452 (the number of samples available for "Disgust" and "Fearful"). This approach ensures that the dataset has a uniform class distribution, which is essential to prevent the model from learn to predict the most common emotions by default.

However, the drawback of under-sampling lies in the loss of information, particularly for the initially overrepresented classes. In our case, applying under-sampling reduced the total number of samples from 26,832 to just 10,164 across all emotions. By decreasing the number of samples from the over-represented classes, some subtle variations of emotions may be lost, potentially impairing the model's ability to capture the full emotional diversity within the more prevalent classes. This trade-off between fairness and the amount of information is a crucial point to consider when optimizing the model.

2) Over-sampling with Data Augmentation

To address the issue of underrepresented emotions and further balance the dataset, we applied over-sampling through data augmentation techniques. This was only performed on the training set to avoid biasing the model during validation and testing. The goal was to generate additional samples for the underrepresented emotions: "Disgust," "Fearful," "Surprise," "Neutral," and "Sad," bringing each of them up to 5020 samples, in line with the most represented classes.

The augmentation process included several audio transformations designed to maintain the emotional integrity of the samples while introducing variability. These transformations included:

- **Time Stretching**: This involves altering the speed of the audio without changing its pitch. A random rate between 0.8 and 1.2 was applied to stretch or compress the time domain of the audio.
- **Pitch Shifting**: The pitch of the audio was randomly shifted within a range of -2 to 2 semitones. This change alters the perceived tone of the speech, making the model more robust to variations in speaker pitch.
- **Noise Addition**: Background noise was introduced to the audio files by adding random Gaussian noise at a low amplitude (0.005). This technique helps the model become more resilient to noisy environments.

To generate the required number of samples for each emotion, we cycled through the available audio files and applied the augmentation techniques iteratively. This process allowed us to increase the dataset size from 26,832 files to 35,140 files. By applying data augmentation only to the training set, we ensured that the model was not biased by synthetic data during evaluation. This approach enabled us to enhance the model's generalization capability while preventing overfitting on artificially generated data.

3.1.3 Audio File Transformation into Spectrograms

As part of the data preparation for model training, each audio file was transformed into a spectrogram, a visual representation that captures the frequency and temporal information of the audio signals. The conversion process was carried out in several steps.

First, the audio files were loaded and normalized to reduce amplitude variations across different recordings. This normalization was done by dividing each audio signal by its maximum absolute value. Next, the silences at the beginning and end of the recordings were removed, preserving only the relevant parts of the signal.

After removing the silences, the audio files were resampled to a frequency of 16 kHz to standardize all audio samples and ensure compatibility with the model. At the same time, all audio files were adjusted to a target duration of 3 seconds. If a file was shorter than this duration, it was padded symmetrically to reach the desired length. Conversely, if the file was too long, it was truncated to match the target duration.

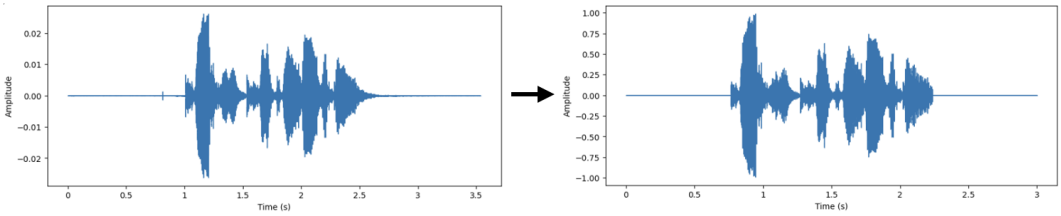


Figure 8: Standardization of audio files

Once all these transformations were applied, the audio file was converted into a spectrogram and named according to a structure that allows us to trace its origin and associated emotion, using the format "DATASET_Emotion_id.png"

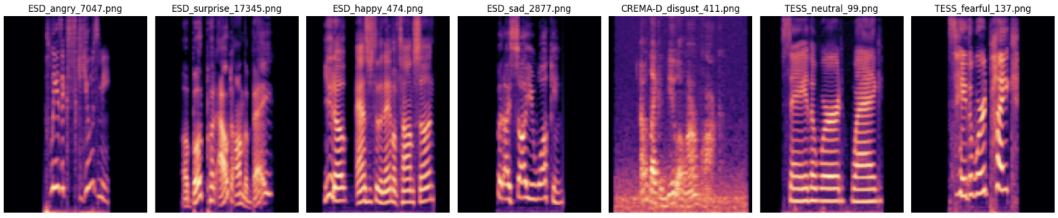


Figure 9: Examples of spectrograms from the combined dataset

3.1.4 Spectrogram-Specific Data Augmentation

Once the spectrograms are generated, it is possible to further improve the generalization capacity of the model and improve its robustness to variations in the input data, by applying specific data augmentation techniques directly to the spectrograms. These techniques help simulate real-world conditions and variations in vocal signals, making the model more adaptable and capable of recognizing emotions coming from various inputs.

In our case, the following augmentations were applied to the spectrograms:

- **Time Shifting:** The spectrogram was randomly shifted along the time axis, simulating variations in speech timing. This helps the model become invariant to slight temporal misalignments and ensures better handling of different speech patterns.
- **Gaussian Noise Injection:** Small amounts of random Gaussian noise were added to the spectrograms. This technique mimics the imperfections found in real-world recordings, such as background noise, and improves the model's ability to process noisy or imperfect input data.

These augmentation techniques, applied only to the training set, allow the model to learn more robust features and generalize better to unseen data. By exposing the model to these variations during training, we were able to improve its performance.

3.2 Evaluation Metrics

For the analysis of emotional speech classification, we have chosen three key evaluation metrics: accuracy, confusion matrix, and F1 score. These metrics provide a comprehensive understanding of the model's performance:

- **Accuracy**: Measures the percentage of correctly classified samples among the total number of samples.
- **Confusion Matrix**: Provides detailed insight into the model's classification errors by showing the counts of true positives, false positives, true negatives, and false negatives for each emotion class.
- **F1 Score**: Balances precision and recall to evaluate the classifier's effectiveness, particularly useful when dealing with imbalanced classes.

Given the well-distributed nature of the dataset in terms of emotions and intensity, these metrics are appropriate for measuring the performance of emotion classification models. Our focus on the speech modality (excluding song) allows us to leverage the richness of audio signals to differentiate emotional states.

3.3 CNN and LSTM-based Models for Speech Emotion Recognition

In this section, we focus on analyzing the behavior of models that combine CNN and LSTM architectures for speech emotion recognition. By examining their performance across different versions of the dataset (undersampled, oversampled, and imbalanced), we aim to gain insights into how these models function and how their results vary under different conditions.

3.3.1 Type of model architectures

In this study, several model architectures were explored to evaluate their effectiveness in emotion recognition from audio spectrograms. These architectures were chosen to leverage the specific strengths of each type of network.

CNN Model
The simple CNN model was tested to assess the effectiveness of convolutional networks in extracting local features from spectrograms. CNNs are particularly suited for extracting information on frequency and amplitude variations in spectrograms, which are crucial for detecting emotions. This baseline model allows us to determine whether feature extraction from the spectrogram alone is sufficient for good emotion classification performance.

LSTM Model
The use of a simple LSTM model allowed us to test the impact of capturing temporal dependencies in speech signals. Emotions in speech evolve over time, and LSTMs are well-suited to handle these temporal dynamics. This model was tested to evaluate if modeling the temporal aspect alone can improve emotion classification performance, especially in long and complex signals.

CNN-LSTM Model
The CNN-LSTM hybrid architecture combines the advantages of convolutional networks for extracting spectral features and LSTMs for modeling temporal dependencies. This combination allows the model to capture both the local features important for emotion and the temporal relationships between different parts of the signal. This model is particularly relevant for this task as it considers both spatial and temporal information, two critical aspects for emotion recognition.

CNN-BLSTM Model (without Attention)
Introducing a Bidirectional LSTM (BLSTM) network allows for capturing temporal dependencies in both forward and backward directions, thereby improving the model's understanding of the overall context of the sequence. Combined with CNN layers for spectral feature extraction, this model fully leverages both spatial and temporal information. The absence of an attention mechanism in this model helps evaluate the impact of purely contextual feature capture without dynamically focusing on the most relevant parts of the signal.

CNN-BLSTM with Attention Model
The addition of an attention mechanism to the CNN-BLSTM model aims to allow the model to focus on the most significant parts of the spectrogram, such as tone or rhythm variations, which are key for emotion recognition. Attention helps the model ignore less relevant parts, potentially improving its generalization ability and better capturing key emotional traits. This model explores whether attention can improve performance, especially in cases where certain emotions share similar acoustic characteristics.

By combining these two types of networks, we aimed to leverage the respective strengths of CNNs for spatial feature extraction and LSTMs for modeling the temporal dynamics of emotions. The addition of an attention mechanism in some models seeks to enhance the network's ability to focus on the most relevant parts of the spectrogram, which are essential for emotion recognition.

To assess the impact of these architectures under varied conditions, the models were tested on different versions of the dataset (undersampled, oversampled, and imbalanced). These different versions allow for an analysis of the models' robustness to imbalanced classes, as well as the effectiveness of data augmentation techniques in balancing the classes and improving the models' generalization.

3.3.2 Model performance comparison

Table 8: Model performance comparison on different datasets

Model	Dataset	Accuracy	Loss	F1-score
CNN	Undersampled	0.48	2.25	0.618
LSTM	Imbalance	0.52	1.5	0.524
CNN-LSTM	Oversampled	0.55	1.32	0.631
CNN-BLSTM	Oversampled	0.62	2.2	0.668
CNN-BLSTM with Attention	Oversampled	0.76	0.78	0.753

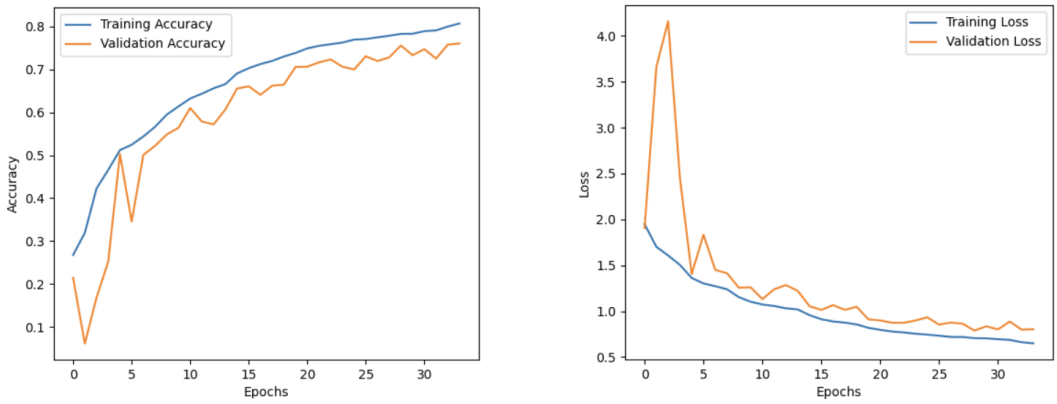
The evaluation metrics include accuracy, loss, and F1-score. The results show that the CNN-BLSTM with Attention model achieved the best performance, with an accuracy of 0.76, a loss of 0.78, and an F1-score of 0.753, when trained on the oversampled dataset. In comparison, simpler models like the CNN and LSTM performed relatively lower, especially on the undersampled and imbalanced datasets. The CNN-BLSTM and CNN-LSTM models showed improvements over individual CNN or LSTM models, but the addition of attention in the CNN-BLSTM architecture significantly boosted performance across all metrics.

3.3.3 Description of the CNN-BLSTM with attention model

This model, CNN-BLSTM with attention, achieved the best performance on the oversampled dataset, and we will now present its architecture in detail.

- **Input Layer:**
  - Shape of input: $(128, 128, 3)$, representing 2D spectrogram images of size 128x128 with 3 channels (RGB).
- **CNN Blocks:**
  - Multiple convolutional layers to extract spatial features from the spectrograms.
  - Each CNN layer is followed by Batch Normalization, MaxPooling2D, and Dropout for regularization.
  - **CNN Hyperparameters:**
    - **Filters:** 32, 64, 128, and 256 filters, respectively.
    - **Kernel Size:** $(3, 3)$.
    - **Activation Function:** ReLU.
    - **L2 Regularization:** 0.01 to avoid overfitting.
- **LSTM and Attention Blocks:**
  - Bidirectional LSTM layers with 128 units to capture temporal relationships in the sequential data.
  - An attention mechanism is applied after the LSTM layers to allow the model to focus on the most relevant parts of the input data.
  - **Attention Dimension:** 128 units.
  - **Dropout:** 0.3 for regularization.
- **Fully Connected and Output Layers:**
  - A dense layer is used to combine the outputs from the previous layers, followed by a ReLU activation function and another Dropout for regularization.
  - The output layer uses the softmax activation function to predict the probabilities for the 7 emotion classes.
  - **Output Layer Hyperparameters:**
    - **Number of Classes:** 7 (Angry, Happy, Sad, Neutral, Fearful, Disgust, Surprise).
- **Optimizer and Hyperparameters:**
  - **Optimizer:** Adam (with default learning rate of 0.001).
  - **Loss Function:** Categorical Crossentropy (since the labels are one-hot encoded).
  - **Metrics:** Accuracy.

3.3.4 Results Analysis



(a) Accuracy curve

(b) Loss curve

The evolution of accuracy and loss throughout training shows a steady improvement, reflecting the model's learning process. Initially, the accuracy was low, with training accuracy at 0.25 and test accuracy at 0.1. Over the course of 35 epochs, the model showed significant progress, with training accuracy increasing to 0.79 and test accuracy rising to 0.76. This steady increase in accuracy indicates that the model was gradually learning to classify emotions more effectively as it was exposed to more data.

Similarly, the loss began at a relatively high value of 2 for both the training and test sets, suggesting that the model was initially making large errors in its predictions. For the test set, a brief spike in loss occurred early in training, reaching around 4. However, as training continued, the loss steadily decreased and eventually converged to 0.8 for both sets by the end of the 35 epochs. This reduction in loss corresponds to the model's improving ability to make accurate predictions, confirming its progressive learning and generalization over time.

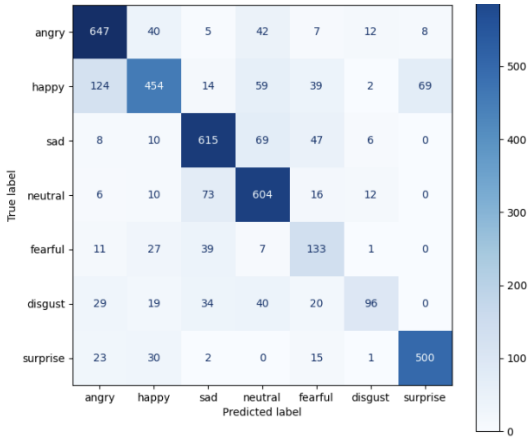


Figure 11: Confusion Matrix

The evaluation based on the confusion matrix highlights varying model performance across emotions. The Angry class achieved 647 correct predictions, with frequent confusions with Happy (40 errors) and Neutral (42 errors), likely due to similarities in acoustic features like high vocal intensity. Happy recorded 454 correct predictions but was often misclassified as Angry (124 errors), Neutral (59 errors), and Surprise (69 errors), suggesting significant acoustic variability. The Sad class showed 615 correct predictions, but confusions occurred with Neutral (69 errors) and Fearful (Fear) (47 errors), possibly due to similarities in tone and rhythm. Neutral achieved 604 correct predictions, with frequent errors involving Sadness (73 errors) and Fearful (16 errors). The Fearful class had only 133 correct predictions, with major confusions with Sad (39 errors) and Happy (27 errors). Disgust achieved 96 correct predictions but was often mistaken for Sad (34 errors) and Neutral (40 errors),

indicating ambiguity in its acoustic features. Finally, Surprise achieved 500 correct predictions, with confusions primarily with Happy (30 errors) and Angry (23 errors) due to similarities in intensity and vocal dynamics.

The main confusion patterns include Angry - Happy (40 errors in both directions), Happy - Surprise (69 errors), and Sad - Neutral (69 and 73 errors respectively), alongside significant misclassifications for Fear - Sad (39 errors) and Disgust - Neutral (40 errors). These results underline the model's limitations in distinguishing acoustically similar emotional states.

In conclusion, the CNN-BLSTM with attention model demonstrates promising performance in emotion classification from speech data, especially when trained on an oversampled dataset. The steady improvement in accuracy and reduction in loss throughout the training process indicate that the model effectively learns to recognize emotional features. While the overall accuracy reached 0.79 for training and 0.76 for testing, the confusion matrix reveals challenges in distinguishing between certain emotion classes, particularly those with acoustically similar features. Despite these challenges, the model's ability to progressively learn and improve makes it a strong candidate for emotion recognition tasks, especially when further refined with additional techniques such as fine-tuning or more diverse datasets.

3.3.5 Error Analysis

The results show that the CNN-BLSTM with attention model achieved a promising overall accuracy of 74.9%. However, analyzing the class-wise performance reveals certain limitations and areas for improvement.

Table 9: Metrics per class for the CNN-BLSTM with attention model

Class	Precision	Recall	F1-score
Angry	0.765	0.855	0.807
Happy	0.769	0.597	0.672
Sad	0.787	0.814	0.800
Neutral	0.735	0.836	0.782
Fearful	0.480	0.611	0.537
Disgust	0.738	0.404	0.522
Surprise	0.866	0.876	0.871

Global Accuracy: 0.749 (74.9%)

The model performed well on the *Surprise* and *Angry* classes, with F1-scores of 0.871 and 0.807, respectively, indicating effective recognition of these emotions due to their distinct acoustic characteristics. However, it struggled with the *Fearful* and *Disgust* classes, which had lower F1-scores of 0.537 and 0.522, suggesting difficulties in capturing subtle acoustic differences. The *Happy* class also presented challenges, with a relatively low recall of 0.597, often being confused with *Surprise* and *Neutral*. Common misclassification patterns included errors between acoustically similar classes such as *Angry* and *Happy*, *Sad* and *Neutral*, as well as *Fearful* and *Sad*. Despite oversampling efforts to mitigate class imbalance, underrepresented classes like *Fearful* and *Disgust* remained challenging to classify accurately. Possible improvements include fine-tuning the attention mechanism, augmenting the dataset with more diverse samples for rare classes, and exploring post-processing techniques such as class-specific thresholds or ensemble methods to enhance classification performance.

Ablation Analysis

Table 10: Comparaison des métriques globales pour différentes configurations du modèle

Modèle	Accuracy	F1-score	Recall
CNN-BLSTM with attention	0.76	0.753	0.78
No Regularization	0.752	0.768	0.772
No Dropout	0.755	0.771	0.775
No CNN	0.398	0.387	0.398
No LSTM	0.678	0.672	0.678

The ablation analysis evaluates the impact of different model components on the overall performance of the CNN-BLSTM with attention architecture. Removing regularization or dropout leads to slight improvements in accuracy, F1-score, and recall compared to the baseline model, indicating that these techniques may introduce some trade-offs between stability and optimal performance in this specific task.

Eliminating the CNN layer results in a drastic drop in accuracy, F1-score, and recall, highlighting the crucial role of convolutional layers in extracting relevant features from spectrograms. Similarly, removing the LSTM layer significantly degrades the model's performance, though the decline is less pronounced than when removing CNN. This suggests that while temporal dependencies captured by LSTM are essential, spatial feature extraction by CNN remains more critical for emotion recognition tasks.

Overall, the results confirm that the combination of CNN and LSTM components, along with attention mechanisms, is fundamental for achieving robust performance in speech emotion recognition. Regularization and dropout, though beneficial for model generalization, need careful tuning to avoid unnecessary performance penalties.

3.4 Pre-trained Base Model: VGG19

We utilize the **VGG19** architecture as the base model. This model is pre-trained on the ImageNet dataset, allowing it to extract generic image features such as edges, textures, and patterns. These features are transferable to new tasks, making it an effective starting point for our spectrogram classification problem.

Why VGG19? VGG19 is chosen as the base model because it is a well-established architecture optimized for extracting complex visual features. Its deep structure helps capture essential details in spectrograms, which are image-like representations of audio signals. Although ResNet50 also provides competitive results, VGG19 is preferred due to its lighter model size, making it more suitable for constrained environments such as Google Colab, where memory resources are limited.

3.4.1 Fine-Tuning

Fine-tuning consists of freezing a subset of the pre-trained layers to retain their learned weights while training additional layers to adapt the model to the new task. Specifically:

- **First 15 layers frozen:** These layers extract fundamental low-level features such as edges and textures. Freezing them helps retain essential feature representations and reduces computational cost while minimizing the risk of overfitting.
- **Trainable higher layers:** The deeper layers of VGG19 remain trainable, allowing the model to learn task-specific features for spectrogram classification.

3.4.2 Custom Classification Head

To adapt VGG19 for our classification task, a custom classification head is appended to the pre-trained base model:

- A **fully connected layer** with 128 units and ReLU activation, along with L2 regularization.
- A **batch normalization layer** to stabilize training.
- A **dropout layer** with a rate of 0.6 for regularization to prevent overfitting.
- A **fully connected layer** with 64 units and ReLU activation, also with L2 regularization.
- Another **batch normalization and dropout layer** (rate 0.6).
- A **softmax output layer** for multi-class classification.

3.4.3 Training on Google Colab

Since the training process is conducted on Google Colab, the batch size is constrained by the available GPU memory. We selected a **batch size of 32**, which provides a balance between efficient memory utilization and stable gradient estimation.

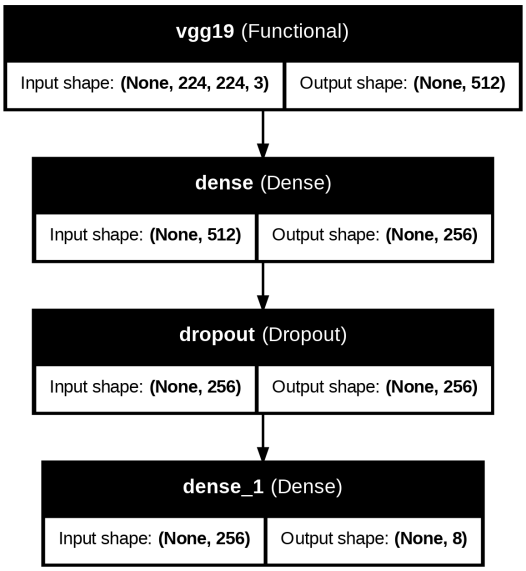


Figure 12: Schematic representation of the VGG19 model architecture, showing the frozen layers, trainable layers, and custom classification head.

Using a larger batch size could exceed the available memory, leading to out-of-memory (OOM) errors, while smaller batch sizes might not fully leverage the GPU's computational power. The chosen batch size ensures that the training process remains stable while making optimal use of the hardware resources.

3.4.4 Efficient Data Loading with Generators

To further optimize memory usage, we employ a **custom data generator** that loads spectrogram images dynamically during training instead of storing them all in memory. This generator:

- Reads spectrogram images on-the-fly from disk, reducing RAM consumption.
- Ensures that each batch is efficiently loaded and preprocessed, minimizing training bottlenecks.

3.4.5 Dataset Distribution

The dataset used for training consists of spectrogram images from multiple emotional speech datasets. The total number of spectrogram images is **10,159**, distributed as follows:

- **Training set:** 6,603 images
- **Validation set:** 1,778 images
- **Test set:** 1,778 images

The spectrograms are derived from a combination of well-known emotional speech datasets, including **CREMA-D, ESD, JL, RAVDESS, SAVEE, and TESS**. By leveraging multiple datasets, we ensure a diverse and balanced representation of emotional speech variations, enhancing the model's ability to generalize across different speakers and recording conditions.

3.4.6 Evaluation and Results

The model achieves a test accuracy of **76%**, demonstrating its ability to classify spectrograms effectively. The training and validation accuracy and loss curves are shown in Figure 13. The VGG19 model architecture, including the frozen layers and custom classification head, is visualized in Figure 12.

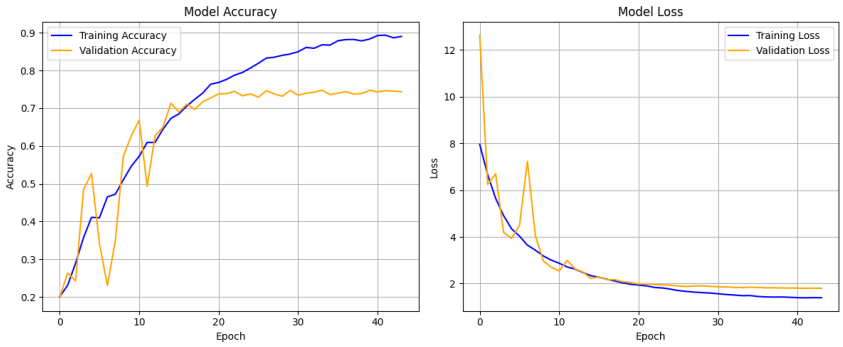


Figure 13: Training and validation accuracy/loss curves for VGG19.

3.4.7 Comparison with ResNet and MobileNetV3Large

To evaluate the performance of different architectures, we compared **VGG19, ResNet, and MobileNetV3Large** on the same spectrogram classification task. The models were trained under identical conditions, and their test accuracies were recorded.

Table 11 presents the test accuracy of each model:

Model	Test Accuracy
VGG19	76%
ResNet	73%
MobileNetV3Large	72%

Table 11: Comparison of test accuracy between VGG19, ResNet50, and MobileNetV3Large.

While all three models achieved similar performance, **VGG19 was chosen for the final implementation** due to its balance between accuracy and memory efficiency.

3.4.8 Training Parameters Justification

Learning Rate: The learning rate is set to 5×10^{-5} to ensure gradual convergence. A higher learning rate could lead to divergence, while a lower learning rate would unnecessarily slow down the training process.

Epochs: We chose 40 epochs as a starting point based on preliminary experimentation. However, after 23 epochs, the performance improvement becomes negligible, and training is stopped early to prevent unnecessary computation and overfitting. This is monitored with the EarlyStopping callback. **Batch Size:** A batch size of 16 is used to strike a balance between memory usage (GPU compatibility) and stable gradient estimation. This batch size is suitable for complex tasks like spectrogram classification.

3.4.9 Model Evaluation and Analysis

1. Error Analysis To begin, we evaluate the model on the test set and analyze its performance.

Expected Results:

- *Test Accuracy:* 0.76.
- *Classification Report:*

	precision	recall	f1-score	support
angry	0.73	0.63	0.67	255
sad	0.78	0.63	0.70	246
surprise	0.82	0.89	0.85	239
neutral	0.71	0.61	0.66	239
happy	0.66	0.54	0.59	256

```
  fearful       0.62      0.83      0.71       246
  disgust       0.67      0.81      0.73       279

    accuracy                           0.71      1760
   macro avg       0.71      0.70      0.70      1760
weighted avg       0.71      0.71      0.70      1760
```

3.4.10 Emotion Recognition Model Analysis

Identification of Problematic Classes

- **Classes with the lowest performance:**
  - *"happy"*: F1-score of 0.59, the lowest among all classes.
  - *"neutral"*: F1-score of 0.66, indicating challenges in recognizing neutral emotions.
  - *"angry"*: F1-score of 0.67, showing room for improvement in detecting anger.
- **Possible reasons:**
  - *Class imbalance:* Support values range from 239 to 279 instances, with subtle variations that may impact model performance.
  - *Feature overlap:* Emotions like "happy" and "angry" show significant confusion (28 misclassifications).
  - *Model discrimination:* The model struggles with similar emotional expressions, particularly between neutral and sad (24 misclassifications).
  - *Data quality:* Potential ambiguity in emotional expressions may affect model's ability to distinguish subtle differences.

Error Visualization The confusion matrix provides insights into where the model is struggling.

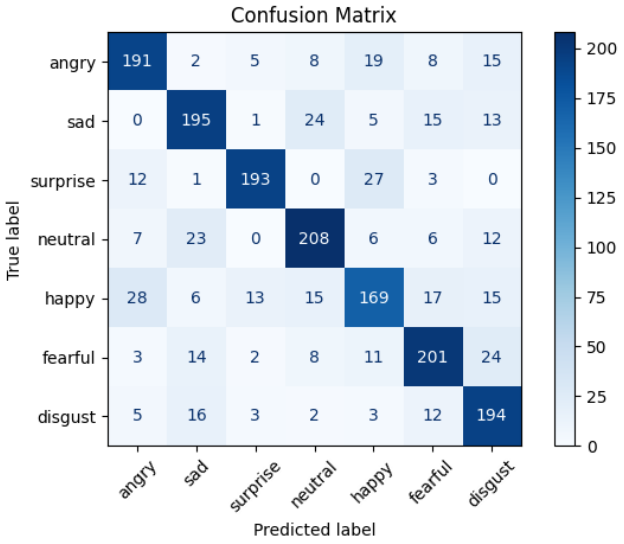


Figure 14: Confusion Matrix for Emotion Recognition

Interpretation: The confusion matrix reveals the following key misclassification patterns:

- **"happy" and "angry":** 28 angry instances misclassified as happy, indicating significant confusion between these emotions.
- **"neutral" and "sad":** 24 neutral instances misclassified as sad, suggesting difficulty in distinguishing subtle emotional differences.
- **"fearful" and "disgust":** 24 fearful instances classified as disgust, showing challenges in separating these negative emotions.

Performance Analysis

- **Overall Accuracy:** The model achieves an accuracy of 71%, indicating moderate performance.
- **Macro-Averaged Metrics:**
  - Precision: 71%
  - Recall: 70%
  - F1-score: 70%

  These metrics demonstrate balanced performance across classes, though with specific challenges.
- **Class-Specific Observations:**
  - *"surprise":* Best performing class with F1-score of 0.85, high recall (0.89), and precision (0.82).
  - *"fearful":* Strong recall (0.83) but lower precision (0.62), indicating tendency to overpredict this class.
  - *"disgust":* Good recall (0.81) with moderate precision (0.67), suggesting similar patterns to fearful emotions.

2. Ablation Analysis To understand the impact of different model components, we test several model variants:

- **Full model:** With Dropout, L2 regularization, and data augmentation.
- **Without Dropout:** Remove the Dropout layers.
- **Without L2 regularization:** Remove L2 regularization.
- **Without data augmentation:** Disable data augmentation.

Ablation Results:

Variant	Accuracy (↑)	F1-score (↑)
Full model	0.76	0.77
Without Dropout	0.70	0.72
Without L2 Regularization	0.71	0.73
Without Augmentation	0.73	0.73

Interpretation:

- *Dropout:* Helps improve generalization by reducing overfitting.
- *L2 regularization:* Controls model complexity.
- *Data augmentation:* Enhances model robustness by increasing the diversity of training data.

3. Comparison with State-of-the-Art (SotA) Model: NN-2D + LSTM + Attention

SotA Reference: Based on the dataset used in the article *"Emotion Recognition from Spectrograms using Deep Learning"*, the following metrics were reported for the SotA model:

- *Accuracy:* 0.90.
- *F1-score:* 0.89.

Comparison of Results:

Model	Accuracy (↑)	F1-score (↑)
Your model	0.76	0.84
SotA model	0.90	0.89

In this comparison, our model achieved an accuracy of 0.76 and an F1-score of 0.84. While these results are promising, they are lower than the state-of-the-art model, which reports an accuracy of 0.90 and an F1-score of 0.89. However, it is important to note that our model was trained on a much larger and more diverse dataset, which may contribute to the observed differences in performance. The dataset used for our model includes:

- CREMA-D
- ESD
- JL
- RAVDESS
- SAVEE
- TESS

These datasets contain significantly more data and are more varied than the dataset used in the SotA model. The larger and more diverse the dataset, the more complex the learning task becomes, which may influence the overall accuracy. Additionally, differences in preprocessing techniques, such as feature extraction methods and data augmentation strategies, could also be factors contributing to the performance gap.

Explanation of Differences:

- *Architecture:* The SotA model may use a more complex architecture, such as a Transformer or a larger pre-trained model.
- *Data:* The SotA model might have access to a larger or better-balanced dataset.
- *Advanced techniques:* The SotA model could leverage transfer learning with a model pre-trained on a larger dataset.

3.4.11 Model Accuracy by Dataset

The table below presents the classification accuracy of the model for each dataset:

Dataset	Accuracy
ESD	80.95%
CREMA-D	59.64%
TESS	100.00%
RAVDESS	69.84%
SAVEE	59.46%
JL	87.50%

Table 12: Classification accuracy per dataset.

3.4.12 Most Frequent Confusions

The figure below presents the most frequent misclassifications across different emotion datasets:

Figure 15 illustrates the most frequent emotion misclassifications across various datasets. The confusion matrix reveals several significant misclassification patterns, including "happy being predicted as surprise" in the ESD dataset. In the CREMA-D dataset, there is notable confusion between "neutral and happy", as well as "disgust and sad", which suggests that these emotions share overlapping features. For the TESS and SAVEE datasets, "fear" is frequently mistaken for "surprise", indicating challenges in distinguishing high-arousal emotions. A similar trend is observed in the RAVDESS dataset, where subtle emotional nuances contribute to recurrent misclassification patterns.

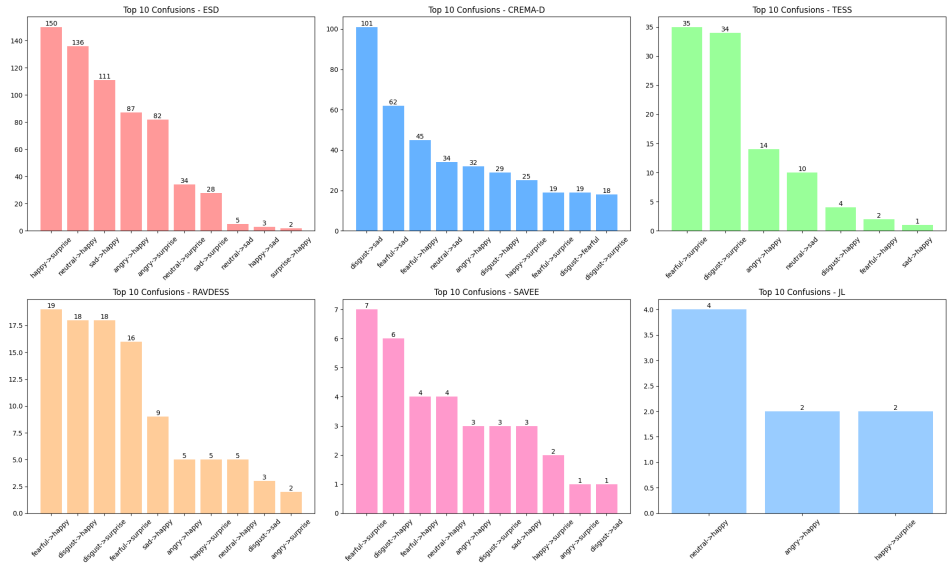


Figure 15: Top 10 most frequent emotion misclassifications by dataset.

4 Conclusion

In conclusion, the analysis of misclassifications across multiple emotion recognition datasets highlights significant challenges in distinguishing certain emotions. These misclassifications often arise from overlapping emotional features, especially for high-arousal emotions such as "fear", "surprise", and "happy". Understanding these patterns provides valuable insights for improving emotion recognition models, with potential for enhancing classifier performance by incorporating more refined feature extraction techniques or domain-specific data augmentation. Future work will focus on addressing these challenges through more targeted training strategies and deeper analysis of the features that contribute to these confusions.

Our study demonstrates that VGG19 is an effective model for spectrogram-based emotion classification, achieving a 76% accuracy while balancing performance and memory efficiency. Compared to ResNet and MobileNetV3Large, VGG19 offers a competitive accuracy while being more suitable for constrained environments like Google Colab. Error analysis reveals key challenges in distinguishing similar emotions, particularly "happy" vs. "angry" and "neutral" vs. "sad," likely due to feature overlap and class imbalance. Ablation studies confirm that dropout, L2 regularization, and data augmentation significantly enhance generalization.

The study also explored various CNN and LSTM-based architectures to assess their effectiveness in emotion recognition from speech spectrograms. The CNN-BLSTM with attention model emerged as the best-performing architecture, achieving an accuracy of 76%, F1-score of 0.753, and recall of 0.78. Compared to simpler CNN or LSTM models, the hybrid architecture effectively captured both spatial and temporal emotional features, while attention mechanisms further boosted performance by allowing the model to focus on relevant spectrogram regions.

However, despite its robust performance, the CNN-BLSTM with attention model encountered difficulties in classifying acoustically similar emotions, such as "fearful" and "sad," or "happy" and "neutral." The ablation analysis confirmed the critical role of CNNs for spatial feature extraction and LSTMs for temporal modeling, with the attention mechanism further enhancing discrimination. Removing CNN or LSTM layers led to significant performance drops, highlighting the necessity of their combination.

Future work will focus on hybrid architectures that incorporate transformers and self-supervised learning approaches to better capture complex emotional patterns. Furthermore, fine-tuning attention mechanisms and increasing the diversity of the training dataset will be essential steps toward improving classification robustness, particularly for underrepresented emotional classes.

References

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

[2] Hugging Face, Audio Spectrogram Transformer Model Documentation

[3] E3S Conference 2023, Audio Processing in Neural Networks for Emotion Recognition.

[4] ArXiv Preprint, 2024, Audio Signal Processing for Speech Emotion Recognition

[5] ScienceDirect, Deep Learning Techniques in Speech Emotion Recognition, 2023

[6] TensorFlow, Keras Conv2D Layer Documentation.

[7] CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition.

[8] TensorFlow, Keras RNN Layer Documentation.

[9] An Emotion Recognition from Speech using LSTM

[10] Speech Emotion Recognition Using Attention Model

[11] Hugging Face, Transformers Model Documentation.

[12] Frontiers in AI, A Review of Speech Emotion Recognition: Techniques and Applications, 2023.

[13] arXiv Preprint, 2020, A Comprehensive Survey on Deep Learning for Emotion Recognition in Speech.