Speech Emotion Recognition Using CNN-BiLSTM on Multi-Corpus Data

Le Phuoc Loc, Nong Quoc An, Nguyen Hoang Gia Huy,
Le Nhut Tien, Doan Minh Duc, Phan Nguyet Minh, Ho Ngoc Hai Nhan
FPT University, Ho Chi Minh Campus, Vietnam

Abstract

Speech Emotion Recognition (SER) is a crucial area in the intersection of audio signal processing and affective computing. This project aims to build a well-generalized emotion recognition model using audio data from multiple public datasets. The proposed approach investigates Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) hybrid architectures for classifying emotional states from speech. We use a unified dataset combining RAVDESS, TESS, SAVEE, and CREMA-D, encompassing a wide range of speakers and emotional expressions. To enhance generalization and improve classification performance, we incorporate gender into the emotion labels (e.g., "female_happy"). Audio features include MFCC, Mel spectrogram, RMS energy, and zero-crossing rate, with extensive data augmentation applied (noise, pitch shift, time shift, speed variations). Among the architectures tested, the model with four CNN layers and one Bi-LSTM achieved the best accuracy of 89.81%. Results show that incorporating gender information improves model performance significantly.

1 Introduction

Speech Emotion Recognition (SER) is a rapidly evolving interdisciplinary field that seeks to identify and interpret human emotional states from speech signals. As voice-based interfaces become ubiquitous in human-computer interaction—powering devices like smart assistants, automotive systems, and telehealth platforms—the ability to accurately detect emotions from vocal cues has emerged as a critical capability. SER enables systems to move beyond mere speech recognition, allowing them to understand the emotional context of a speaker's voice, which is essential for creating empathetic, adaptive, and context-aware interactions. By decoding paralinguistic features such as pitch, intonation, rhythm, energy, and prosody, SER systems can infer emotions like happiness, sadness, anger, or fear, even when the spoken words themselves are emotionally neutral.

The significance of SER lies in its ability to enhance user experiences across diverse domains. For instance, emotionally intelligent virtual assistants can tailor their responses to a user's mood, while mental health monitoring systems can detect early signs of emotional distress. Despite its potential, SER remains a formidable challenge due to the complexity of human emotional expression. Unlike text-based sentiment analysis, which relies heavily on lexical content, SER must interpret subtle, non-linguistic vocal characteristics that vary significantly across individuals, cultures, genders, and recording conditions. These

variations, combined with the scarcity of large, diverse, and high-quality labeled datasets, make it difficult to develop models that generalize effectively to real-world scenarios.

1.1 Project Goal and Motivation

The primary objective of this project is to develop a robust and generalizable Speech Emotion Recognition (SER) model capable of performing reliably across varied speakers, genders, and acoustic environments. To this end, we adopt a multi-corpus strategy by integrating four publicly available SER datasets: RAVDESS, TESS, SAVEE, and CREMA-D. While these datasets are all composed of scripted, acted speech in English and are recorded in controlled or semi-controlled settings, they still provide moderate variability in speaker identity, vocal timbre, emotional expression, and background conditions. Though limited in spontaneous or real-world variability, combining these datasets serves as a meaningful step toward broader generalization by exposing the model to a wider distribution of voices, affective cues, and recording setups than any single dataset could provide in isolation.

A key innovation in this project is our novel label engineering strategy, which integrates gender and emotion into a single classification target (e.g., "male_angry," "female_sad"). This approach explicitly accounts for gender-specific differences in emotional expression, such as variations in pitch range, which are often overlooked in traditional SER models. By reformulating the classification task in this way, we aim to reduce intra-class variability caused by speaker gender while enabling the model to learn nuanced, gender-specific vocal patterns associated with different emotions.

To extract and model the complex temporal and spectral patterns in speech signals, we propose a hybrid CNN-BiLSTM architecture. Convolutional Neural Networks (CNNs) serve as effective feature extractors, capturing local patterns in mel-spectrograms, which represent the frequency content of audio signals over time. Bidirectional Long Short-Term Memory (BiLSTM) layers complement this by modeling long-term temporal dependencies, allowing the model to account for the sequential nature of emotional cues in speech. This hybrid approach leverages both spatial and temporal information, making it well-suited for the dynamic and context-dependent nature of emotional speech.

1.2 Real-World Applications of SER

The applications of SER are vast and span multiple industries, reflecting the growing demand for emotionally intelligent systems. Below are some key use cases:

• Smart Assistants and Dialog Systems: Virtual assistants like Amazon Alexa, Google Assistant, or Siri can use SER to detect user emotions such as frustration, excitement, or confusion, enabling them to adjust their tone, pacing, or response content. For example, a frustrated user might receive a calmer, more empathetic response to de-escalate their emotional state.

• Mental Health Monitoring: SER systems can analyze vocal patterns over time to detect signs of depression, anxiety, or emotional distress, offering a non-invasive tool for telehealth platforms or wearable devices. For instance, changes in pitch or speaking rate could indicate worsening mental health, prompting timely interventions.

• Customer Service Analytics: In call centers, SER can identify angry or distressed callers, allowing companies to route them to experienced agents or flag interactions for

quality assurance. This can improve customer satisfaction and streamline service operations.

• Driver Monitoring Systems: In automotive applications, SER can detect emotions like anger, fatigue, or stress in drivers, enabling real-time alerts to prevent accidents. For example, a system might suggest a break if it detects signs of frustration or drowsiness.

• E-Learning Platforms: Adaptive e-learning systems can use SER to gauge learner emotions, such as confusion or boredom, and adjust the pace, difficulty, or presentation of instructional content accordingly.

• Gaming and Virtual Reality (VR): In interactive gaming or VR environments, SER can guide narrative branching, adjust game difficulty, or create more immersive experiences by responding to a player's emotional state.

• Market Research and Advertising: By analyzing emotional responses to advertisements or product pitches, companies can better understand consumer preferences and tailor their strategies accordingly.

1.3 Challenges in SER and Research Gaps

Despite significant advancements, several challenges continue to hinder the widespread adoption of SER systems in real-world applications:

• Generalization Across Speakers and Contexts: Most SER models perform well on training data but struggle to generalize to unseen speakers, particularly when they differ in gender, age, accent, or cultural background. This lack of robustness limits their practical utility in diverse settings.

• Dataset Heterogeneity: Publicly available SER datasets vary widely in terms of emotion categories, recording conditions, speaker demographics, and annotation quality. For example, RAVDESS includes acted emotions with high-quality recordings, while CREMA-D relies on crowd-sourced annotations, introducing inconsistencies. Merging such datasets for unified training is non-trivial and requires careful preprocessing.

• Ambiguity and Subjectivity of Emotions: Emotions are inherently subjective and often ambiguous, even for human annotators (Schuller et al., 2013). Distinguishing between closely related emotions, such as fear and surprise or sadness and neutral, is challenging, leading to noisy labels and reduced model performance.

• Lack of Spontaneous Speech Data: Most SER datasets, including those used in this project, contain acted or scripted emotional speech, which differs significantly from spontaneous speech in natural settings. The absence of large-scale, real-world corpora limits the development of models that can handle authentic emotional expressions.

• Environmental and Recording Variability: Background noise, microphone quality, and recording environments can significantly affect the quality of speech signals, making it difficult for models to extract reliable emotional features in uncontrolled settings.

• Low-Resource Settings: Unlike automatic speech recognition (ASR), which benefits from vast labeled corpora, SER suffers from a scarcity of diverse, high-quality labeled data, particularly for spontaneous emotional speech. This project addresses these challenges through a multi-faceted approach: (1) combining multiple datasets to increase demographic and emotional diversity, (2) introducing gender-aware label engineering to improve generalization across speakers, (3) employing robust data augmentation techniques to simulate real-world variability, and (4) leveraging a hybrid CNN-BiLSTM architecture to capture both local and sequential patterns in speech.

This project addresses these challenges through a multi-faceted approach: (1) combining multiple datasets to increase demographic and emotional diversity, (2) introducing gender-aware label engineering to improve generalization across speakers, (3) employing robust data augmentation techniques to simulate real-world variability, and (4) leveraging a hybrid CNN-BiLSTM architecture to capture both local and sequential patterns in speech.

1.4 Related Work

The field of SER has seen significant research efforts aimed at addressing generalization, feature extraction, and model robustness. Below, we summarize key contributions and highlight how our work builds upon them:

• Satt et al. (2017): Proposed a CNN+LSTM framework that used mel-spectrograms as input for SER. Their model achieved strong performance on the IEMOCAP dataset but struggled with cross-corpus generalization due to dataset-specific biases. Our approach extends this by training on multiple datasets and incorporating gender-aware labels to enhance robustness.

• Gideon et al. (2019): Investigated domain adaptation techniques to address dataset mismatch in SER, focusing on datasets like IEMOCAP and MSP-Improv. They employed adversarial learning to align feature distributions across datasets, improving generalization. However, their method required complex training procedures. Our label engineering approach offers a simpler, more scalable solution by embedding gender information directly into the classification target.

• Zhao et al. (2019): Introduced a multi-task learning framework that jointly modeled emotion and speaker identity. Their results demonstrated that explicitly accounting for speaker-specific features could improve emotion recognition. Our gender-aware label strategy aligns with this idea but simplifies the process by avoiding the need for auxiliary speaker information during inference.

• Pepino et al. (2020): Investigated the role of data augmentation in SER, showing that techniques like noise injection and pitch shifting could improve model robustness. We build on this by incorporating extensive augmentation strategies tailored to our multi-corpus dataset.

While these studies provide valuable insights, they often rely on complex architectures, additional supervision, or dataset-specific assumptions. Our approach is designed to be both effective and practical, combining multi-corpus training, gender-aware label engineering, and a hybrid CNN-BiLSTM architecture to achieve robust performance without requiring auxiliary inputs at inference time. By addressing key challenges like generalization and dataset heterogeneity, this project aims to advance the state-of-the-art in SER and pave the way for more reliable, real-world applications.

2 Methodology

The Speech Emotion Recognition (SER) pipeline is designed to create a robust and generalizable model capable of detecting emotions from speech across diverse speakers, genders, and recording conditions. The pipeline encompasses several critical stages: dataset preparation, audio preprocessing, feature extraction, data augmentation, and model training. By combining four publicly available datasets—RAVDESS, TESS, SAVEE, and CREMA-D—we create a unified, diverse corpus. To enhance generalization, we introduce a novel

label engineering approach that combines gender and emotion into a single classification target (e.g., "male_angry," "female_sad"), resulting in 14 distinct classes (7 emotions $\times$ 2 genders). This section provides a detailed overview of each stage, supported by equations, figures, and dataset characteristics to ensure clarity and reproducibility.

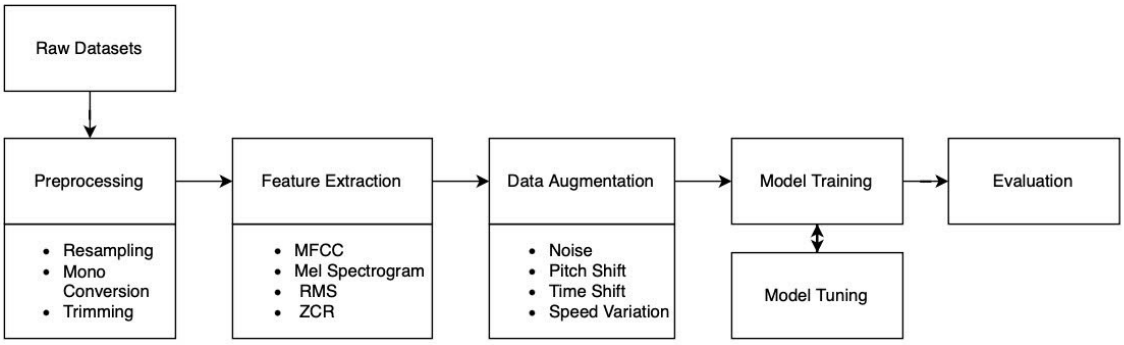


Figure 1: Pipeline Overview

2.1 Dataset Description

To address the challenge of generalization, we combine four diverse, publicly available SER datasets: RAVDESS, TESS, SAVEE, and CREMA-D. These datasets were selected for their complementary characteristics, including differences in speaker gender, moderate variation in age and English accents, diverse emotion categories, recording quality, and utterance types. Below is a detailed breakdown of each dataset:

• RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) - Speakers: 24 (12 male, 12 female, North American English speakers).
- Emotions: Neutral, calm, happy, sad, angry, fearful, disgust, surprised (8 emotions).
- Format: 48 kHz WAV files, high-quality audio.
- Utterances: Scripted phrases (e.g., "Kids are talking by the door," "Dogs are sitting by the door") spoken in various emotional tones.
- Recording Setup: Controlled studio environment with minimal background noise, recorded using professional microphones.
• TESS (Toronto Emotional Speech Set) - Speakers: 2 female (Canadian English speakers, older adults).
- Emotions: Angry, disgust, fear, happy, neutral, pleasant surprise, sad (7 emotions).
- Format: 16 kHz WAV files.
- Utterances: 200 target words (e.g., "say," "run") spoken in different emotional tones, resulting in 2800 utterances.
- Recording Setup: Clean studio environment, recorded with high-fidelity equipment.
• SAVEE (Surrey Audio-Visual Expressed Emotion) - Speakers: 4 male (British English speakers, academic professionals).
- Emotions: Angry, disgust, fear, happy, neutral, sad, surprised (7 emotions).
- Format: 44.1 kHz WAV files.
- Utterances: Scripted sentences (e.g., "She had your dark suit in greasy wash water all year") designed to evoke specific emotions.
- Recording Setup: Professional studio with consistent recording conditions.

• CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) - Speakers: 91 (48 male, 43 female, diverse ethnicities, primarily American English).
- Emotions: Anger, disgust, fear, happy, neutral, sad (6 emotions).
- Format: 44.1 kHz WAV files.
- Utterances: Multiple scripted lines (e.g., "I'm on my way to the meeting") performed at varying intensity levels.
- Recording Setup: Semi-controlled environment with some variability in background noise and microphone quality.

These datasets collectively provide a rich and diverse corpus, covering a wide range of speakers (121 total, 64 male, 57 female), emotions, accents (North American, Canadian, and British English), and recording conditions. To harmonize the datasets, we mapped their emotion labels to a common set of seven emotions (angry, disgust, fear, happy, neutral, sad, surprised), to ensure consistency across datasets, we harmonized emotion labels by merging semantically similar categories — for example, 'calm' from RAVDESS was treated as 'neutral', and 'pleasant surprise' from TESS was considered equivalent to 'surprise'. The final dataset comprises over 12,162 original samples, which are expanded through augmentation.

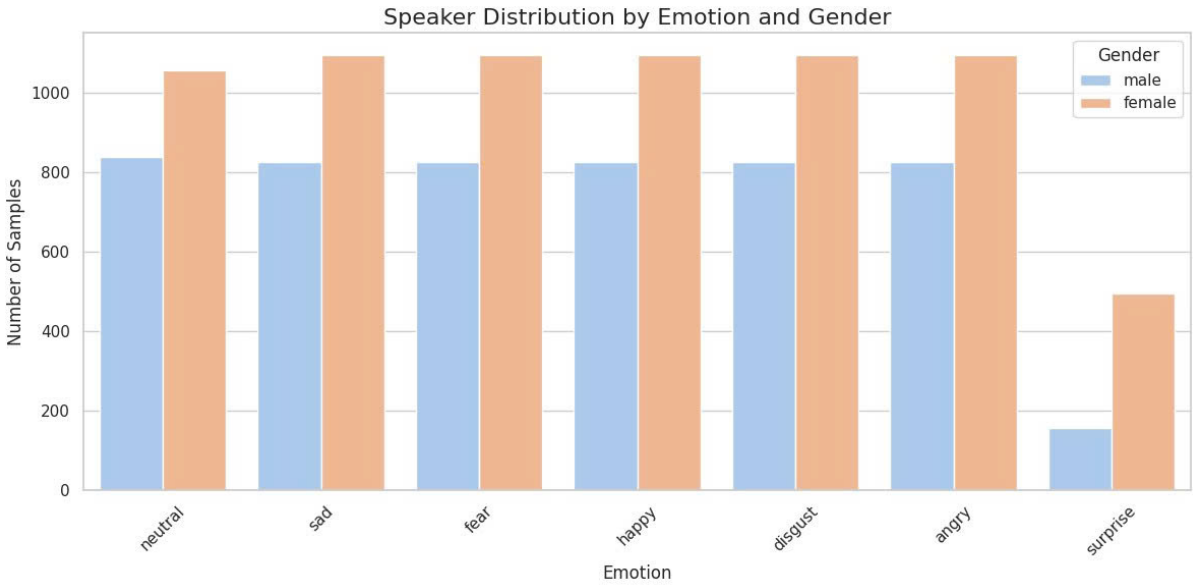


Figure 2: Speaker Distribution

2.2 Audio Preprocessing and Feature Extraction

To ensure consistency across datasets, we applied standardized preprocessing steps to all audio files. These steps address variations in sampling rate, channel configuration, and duration, preparing the data for feature extraction.

Feature Extraction Overview

We extracted a combination of acoustic features that capture both spectral and temporal characteristics of emotional speech. These features were selected for their proven effectiveness in speech emotion recognition (SER) tasks and their ability to represent paralinguistic cues. The extracted features include:

- **Mel-Frequency Cepstral Coefficients (MFCC)**
  MFCCs are widely used in speech processing as they approximate the human auditory system's response to sound. They represent the short-term power spectrum on a mel scale:

$$c_n = \sum_{k=1}^{M} \log(S_k) \cos\left(n\,(k-0.5)\,\frac{\pi}{M}\right)$$

  where:

  - S_k: Power of the k-th mel-filter bank
  - M: Number of mel filters (set to 40)
  - n: MFCC index

  We extracted 40 MFCCs using a 25 ms window and a 10 ms hop length. Each MFCC vector was averaged over time, resulting in a 40-dimensional feature vector per audio sample.

- **Mel Spectrogram**
  The mel spectrogram provides a time-frequency representation of the signal, capturing energy distribution across mel-scaled frequency bands:

$$S_{\text{mel}}(f,t) = \sum_{k} |X(f,t)|^2 H_k(f)$$

  where $H_k(f)$ is the k-th mel filter. To align with human auditory perception, we convert the amplitude to decibels:

$$S_{\text{dB}}(f,t) = 10 \log_{10}(S_{\text{mel}}(f,t))$$

  We extracted 128 mel-frequency bands using a 25 ms window and 10 ms hop length. These were averaged over time, resulting in a 128-dimensional feature vector per sample.

- **Root Mean Square (RMS) Energy**
  RMS energy measures the loudness of a signal, which often correlates with emotional intensity:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$$

  where x_i is the amplitude of the i-th sample in a frame of length N.

- **Zero Crossing Rate (ZCR)**
  ZCR quantifies the rate of sign changes in the waveform, reflecting the presence of high-frequency components:

$$\text{ZCR} = \frac{1}{N-1}\sum_{i=1}^{N-1} \mathbb{1}\{x_i x_{i+1} < 0\}$$

  where $\mathbb{1}\{\cdot\}$ is the indicator function.

All features were extracted and stored as NumPy arrays for efficient model training. The combination of MFCCs, mel spectrograms, RMS, and ZCR was chosen over alternatives such as Chroma features or spectral contrast. MFCCs and mel spectrograms capture the core spectral-temporal properties of emotional speech, while RMS and ZCR add complementary information related to signal dynamics.

An ablation study conducted during model development confirmed the relevance of RMS and ZCR—excluding either resulted in a 3–5% drop in validation accuracy, affirming their contribution to emotion discrimination.

2.3 Data Augmentation

To enhance model robustness and simulate real-world variability, we applied a suite of data augmentation techniques. These augmentations address challenges like background noise, vocal tone differences, and speaking rate variations, which are common in practical SER scenarios (Latif et al., 2020). Each original sample was augmented to produce six additional versions, significantly expanding the dataset. The augmentation techniques include:

1. **Additive White Gaussian Noise**
    - Gaussian noise was synthetically generated and added to the waveform to simulate background noise conditions.
    - Purpose: Simulates noisy environments, such as phone calls or public spaces.
2. **Pitch Shifting**
    - Applied ± 2 semitones using `torchaudio.functional.pitch_shift`.
    - Purpose: Mimics natural variations in vocal pitch across speakers or emotional states.
3. **Time Shifting**
    - The waveform is rolled (circular shift) by 20% of its length.
    - Purpose: Simulates minor misalignments or pauses in recordings, improving robustness to timing errors.
4. **Speed Variation**
    - Changes the playback speed using resampling to $0.8\times$ (slower) and $1.25\times$ (faster) of the original rate.
    - Purpose: Introduces variability in speaking rate, which varies across speakers and emotions.

The augmentation process increased the dataset size from approximately 12,162 original samples to over 85,134 samples ($12,162 \times 7$, including the original). This expansion improved model generalization. Training time increased significantly due to the larger dataset, but this was offset by the improved model performance.

2.4 Dataset Splitting

To ensure a fair and unbiased evaluation, the augmented dataset was split into training and testing subsets using an 80:20 ratio. The split was stratified to maintain the following:

- **Balanced Gender-Emotion Classes:** Equal representation of all 14 classes (7 emotions $\times$ 2 genders).
- **No Speaker Overlap:** Speakers in the training set were excluded from the test set to prevent leakage of speaker-specific traits.

The final training set contained approximately 67,200 samples, and the test set contained 16,800 samples. This large dataset size, combined with the diversity of speakers and emotions, supports the training of deep learning models without overfitting. The stratified split ensured that the model was evaluated on its ability to generalize to unseen speakers, a critical requirement for real-world SER applications.

3 Model Architecture

The design of a high-performing Speech Emotion Recognition (SER) model requires an architecture that effectively captures both local acoustic features and long-term temporal dynamics in speech signals. To achieve this, we developed a hybrid architecture combining Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (Bi-LSTM) layers. CNNs excel at extracting localized patterns from spectrogram-like inputs (e.g., Mel-Frequency Cepstral Coefficients (MFCCs), mel spectrograms), while Bi-LSTMs model the sequential evolution of emotional cues over time. This section details the architectural rationale, model variants, regularization techniques, training procedures, and key observations, supported by diagrams, tables, and training curves for clarity and reproducibility.

3.1 Architectural Rationale

The choice of a CNN-BiLSTM hybrid architecture is motivated by the complementary strengths of its components in processing speech data:

- **Role of CNNs:** CNNs are highly effective for processing two-dimensional feature maps, such as MFCCs or mel spectrograms, which represent the time-frequency structure of audio signals (Trigeorgis et al., 2016). Each convolutional layer applies learnable filters to detect local patterns, such as variations in pitch, energy, or timbre, which are critical for distinguishing emotional states. For example, a high-pitched, rapidly changing energy pattern may indicate surprise, while a low-pitched, steady pattern may signal sadness. By stacking multiple convolutional layers with ReLU activations and max-pooling, the model progressively abstracts these local features into higher-level representations, reducing spatial dimensions while preserving emotional cues. This hierarchical feature extraction is particularly suited for spectrogram inputs, as it captures frequency-specific patterns that are invariant to small temporal shifts.
- **Role of Bi-LSTMs:** Emotional expression in speech is inherently sequential, with cues like rising pitch or fluctuating energy unfolding over time. LSTMs are recurrent

neural networks designed to model long-term dependencies, making them ideal for capturing the temporal dynamics of emotions. Unlike unidirectional LSTMs, which process sequences in one direction (past to future), Bi-LSTMs process data in both forward and backward directions, providing access to the full temporal context of an utterance. This is particularly important for emotions like surprise, where a sudden pitch rise may only be meaningful when viewed in the context of the preceding and following segments. Bi-LSTMs were chosen over Gated Recurrent Units (GRUs) due to their superior ability to handle longer sequences.

The synergy of CNNs and Bi-LSTMs allows the model to first extract robust local features and then model their temporal evolution, enabling accurate discrimination of subtle emotional differences across gender-emotion classes (Neumann & Vu, 2018)

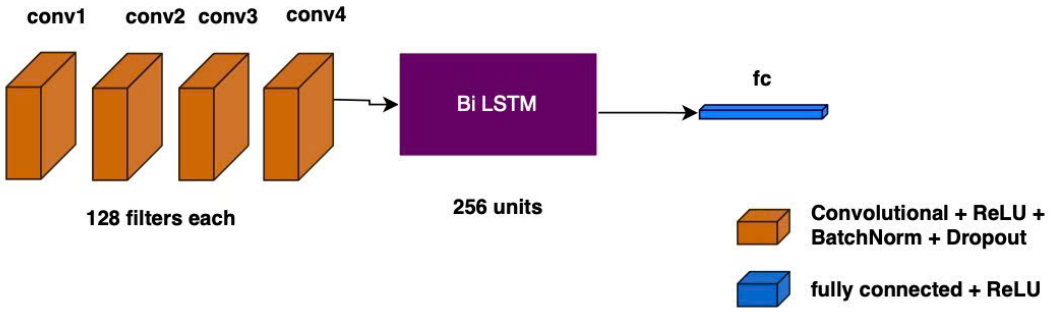


Figure 3: Model A Architecture Diagram

3.2 Model Variants

We designed and evaluated three model variants to explore the impact of architectural depth and label engineering on SER performance:

- **Model A:**
  - **Configuration:** 4 CNN layers (128 filters, 3×3 kernels, ReLU activation) + 1 Bi-LSTM layer (256 units) + 1 Dense layers (14 units).
  - **Output:** 14 gender-emotion classes (7 emotions × 2 genders, e.g., "male_angry," "female_sad").
  - **Accuracy:** 85.5% (validation set).
  - **Notes:** This model balances depth and generalization, leveraging gender-emotion labels to disambiguate speaker variability.
- **Model B:**
  - **Configuration:** 5 CNN layers (128 filters, 3×3 kernels, ReLU activation) + 1 Bi-LSTM layer (256 units) + 1 Dense layers (14 units).
  - **Output:** 14 gender-emotion classes.
  - **Accuracy:** 71.7% (validation set).

- **Notes:** The additional CNN layer increased model capacity but led to overfitting, as the deeper architecture required more data to generalize effectively.

- **Model C:**
  - **Configuration:** 4 CNN layers (128 filters, 3×3 kernels, ReLU activation) + 1 Bi-LSTM layer (256 units) + 1 Dense layers (7 units).
  - **Output:** 7 emotion-only classes (no gender information).
  - **Accuracy:** 68.7% (validation set).
  - **Notes:** Omitting gender information reduced performance, particularly for emotions with overlapping acoustic features (e.g., female happy vs. male sad).

Model	CNN Layers	Bi-LSTM Units	Output Classes	Validation Accuracy
A	4	256	14 (Gender-Emotion)	85.5%
B	5	256	14 (Gender-Emotion)	71.7%
C	4	256	7 (Gender-Emotion)	68.7%

Table 1: Model Comparison

3.3 Regularization Techniques

To prevent overfitting and enhance generalization, we incorporated several regularization techniques, with their rationale and impact detailed below:

- **Dropout:**
  - **Configuration:** Applied with a rate of 0.3 between the final CNN layer and the Bi-LSTM layer.
  - **Rationale:** Dropout randomly deactivates 30% of neurons during training, forcing the network to learn redundant and robust representations. This reduces reliance on specific neurons and mitigates overfitting, especially given the large dataset size (over 84,000 samples).

- **L2 Regularization:**
  - **Configuration:** Applied an L2 penalty of (1×10^{-5}) to all convolutional and dense layer weights.
  - **Rationale:** L2 regularization penalizes large weights, encouraging simpler models that generalize better. This is particularly important for deeper architectures like Model B, which showed signs of overfitting without regularization.

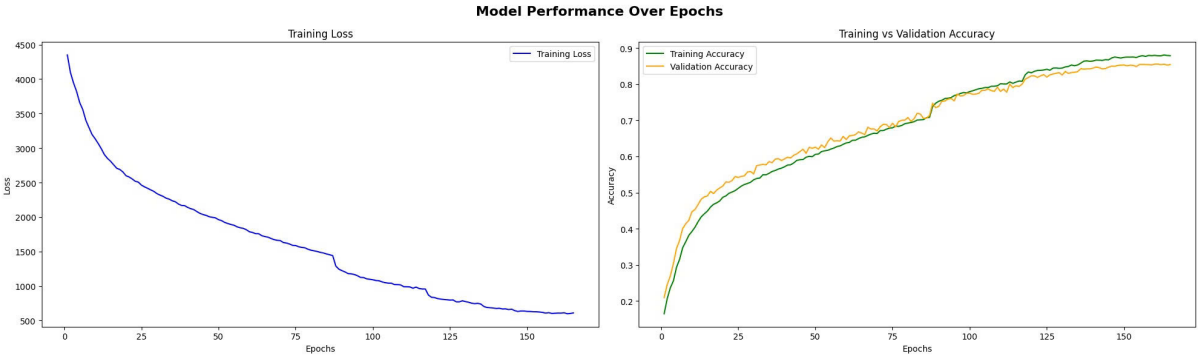
- **Batch Normalization:**
  - **Configuration:** Applied after each CNN layer before max-pooling.
  - **Rationale:** Batch normalization normalizes the output of each convolutional layer, reducing internal covariate shift and making training less sensitive to weight initialization. This accelerates convergence and stabilizes deeper networks.

3.4 Training Procedure

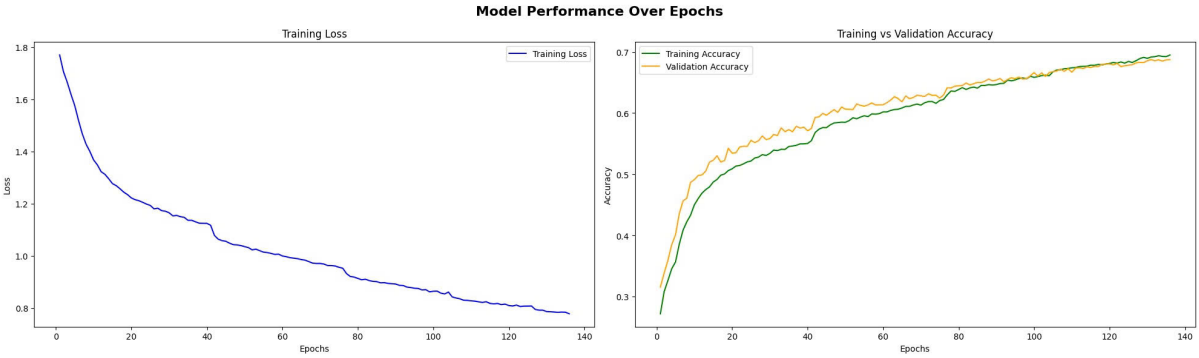
All models were trained with a consistent configuration to ensure fair comparison:

- **Dropout:** Categorical Cross-Entropy, suitable for multi-class classification.
- **Optimizer:** Adam with an initial learning rate of 0.001, paired with a ReduceLROnPlateau scheduler (factor = 0.5, patience = 3) to reduce the learning rate when validation loss plateaus.
- **Batch Size:** 32, balancing memory efficiency and gradient stability.
- **Max Epochs:** 200, with early stopping (patience = 5) to halt training if validation loss did not improve.
- **Validation Metric:** Accuracy, monitored on the validation set (20% of data).
- **Hardware:** NVIDIA T4 GPU (16 GB), enabling efficient training of deep models.

The training process was monitored for validation loss to prevent overfitting, with the learning rate reduced dynamically to fine-tune convergence.



(a) Model A



(b) Model C

Figure 4: Training Curves

3.5 Observations

Key findings from the model training and evaluation include:

- **Model A Superiority:** Model A achieved the highest validation accuracy (85.5%), supporting the hypothesis that gender-emotion label engineering reduces intra-class variability. The model effectively distinguished emotions with overlapping acoustic features (e.g., female happy vs. male sad) by leveraging gender context.
- **Model B Overfitting:** Model B, despite its deeper architecture, exhibited signs of overfitting—achieving a high training accuracy (85.5%) but significantly lower validation accuracy (71.7%). The widening gap between training and validation performance after epoch 10 indicates that the model's added complexity, particularly the fifth CNN layer, was not adequately supported by
- **Model C Limitations:** Model C's lower accuracy (68.7%) highlights the importance of gender information in SER. Without gender context, the model struggled with emotions exhibiting similar acoustic profiles, such as high-pitched female happy and male sad utterances.
- **Regularization Impact:** Dropout and L2 regularization were critical for Model A's generalization, while batch normalization ensured stable training, particularly for the deeper Model B.

These observations underscore the effectiveness of the CNN-BiLSTM architecture with gender-emotion labels and highlight the need to balance model complexity with dataset size to avoid overfitting

4 Result

We evaluated model performance on a held-out 20% test set. The key metric was classification accuracy, supported by precision, recall, and F1-score from detailed classification reports. Accuracy was chosen as the main metric due to a balanced class distribution created by data augmentation and stratified splitting. Below, we compare three models—focusing on the impact of using gender-emotion labels and architectural complexity.

Model	CNN Layers	Bi-LSTM Layers	Target Variable	Accuracy
A	4	1	Gender + Emotion	85.5%
B	5	1	Gender + Emotion	71.7%
C	4	1	Emotion Only	68.7%

Table 2: Comparison of model architectures and their performance.

4.1 Model Comparison

Model A, which includes gender-emotion combined targets and a moderate CNN depth, achieved the highest accuracy (85.5%). It significantly outperformed Model C, which used emotion-only labels and achieved 68.7%. This supports our hypothesis: incorporating speaker gender into emotion labeling enhances discriminability and enables better feature learning.

Model B, which increased CNN depth, surprisingly underperformed (71.7%). This indicates diminishing returns or overfitting with deeper convolutional stacks, especially when batch size is limited and input variation is already high due to augmentation.

4.2 Class-Level Performance

Model A (Gender + Emotion)

Class	Precision	Recall	F1-score	Support
Angry	0.90	0.89	0.90	2281
Disgust	0.88	0.83	0.86	2324
Fear	0.86	0.82	0.84	2309
Happy	0.86	0.85	0.86	2317
Neutral	0.82	0.90	0.86	2240
Sad	0.84	0.86	0.85	2334
Surprise	0.94	0.94	0.94	790
Accuracy	–	–	0.86	14595
Macro Avg	0.87	0.87	0.87	14595
Weighted Avg	0.86	0.86	0.86	14595

Table 3: Model A (Gender + Emotion)

- Strongest classes: Angry and Happy were classified with the highest precision and recall.
- Most confusable: Neutral had high recall (0.94) but lower precision (0.85), suggesting it's often predicted correctly, but also over-predicted.
- Balanced performance across all classes, indicating high model generalization

Model C (Emotion Only)

Class	Precision	Recall	F1-score	Support
Angry	0.76	0.79	0.77	2281
Disgust	0.67	0.60	0.63	2324
Fear	0.72	0.59	0.65	2309
Happy	0.67	0.65	0.66	2317
Neutral	0.60	0.73	0.66	2240
Sad	0.67	0.70	0.68	2334
Surprise	0.84	0.87	0.86	790
Accuracy	–	–	0.69	14595
Macro Avg	0.70	0.70	0.70	14595
Weighted Avg	0.69	0.69	0.69	14595

Table 4: Model C (Emotion Only)

- Confusion hotspots: Disgust is often misclassified as Neutral or Sad.

- Neutral and Surprise have high recall but low precision—suggesting they're frequently predicted, often wrongly.
- Performance drop across all metrics reflects the model's struggle without gender differentiation.

4.3 Confusion Matrix Interpretation

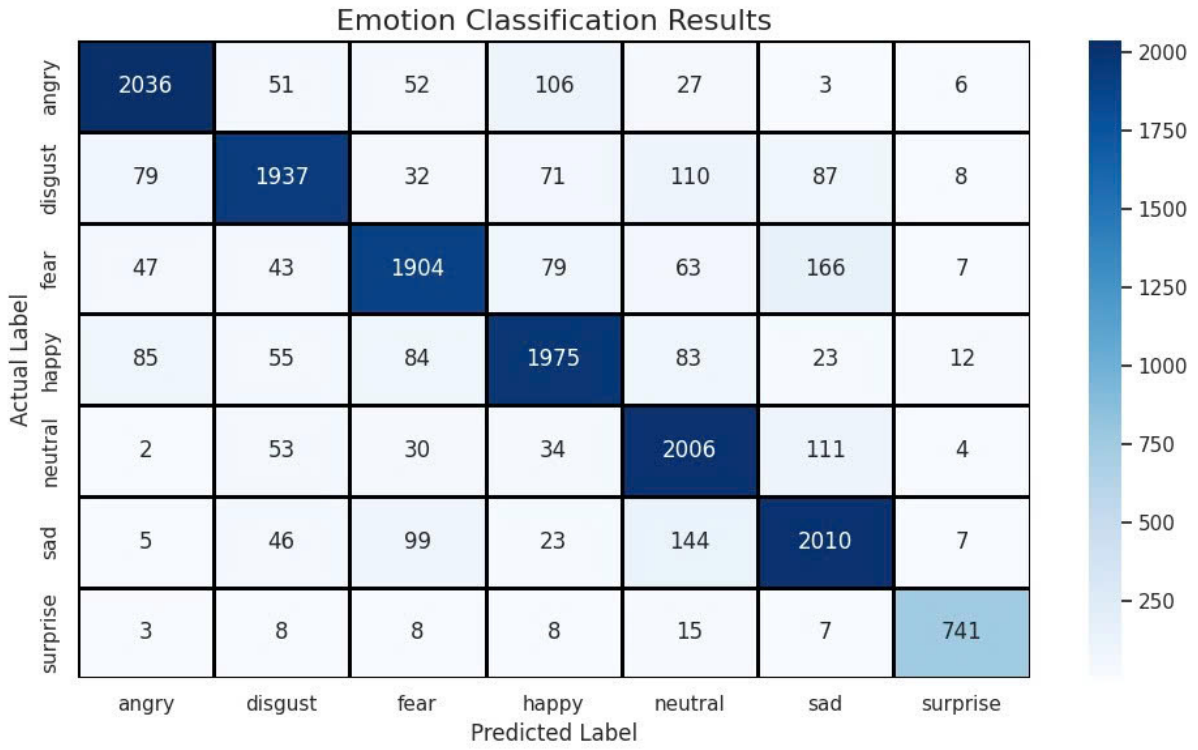


Figure 5: Confusion Matrix for Model A

- Angry and Disgust rarely confused (high diagonals, low off-diagonals).
- Misclassifications are mainly between Sad Neutral and Fear Happy, consistent with known acoustic overlaps.
- Surprise is most accurately classified despite its relatively low support (204 samples).

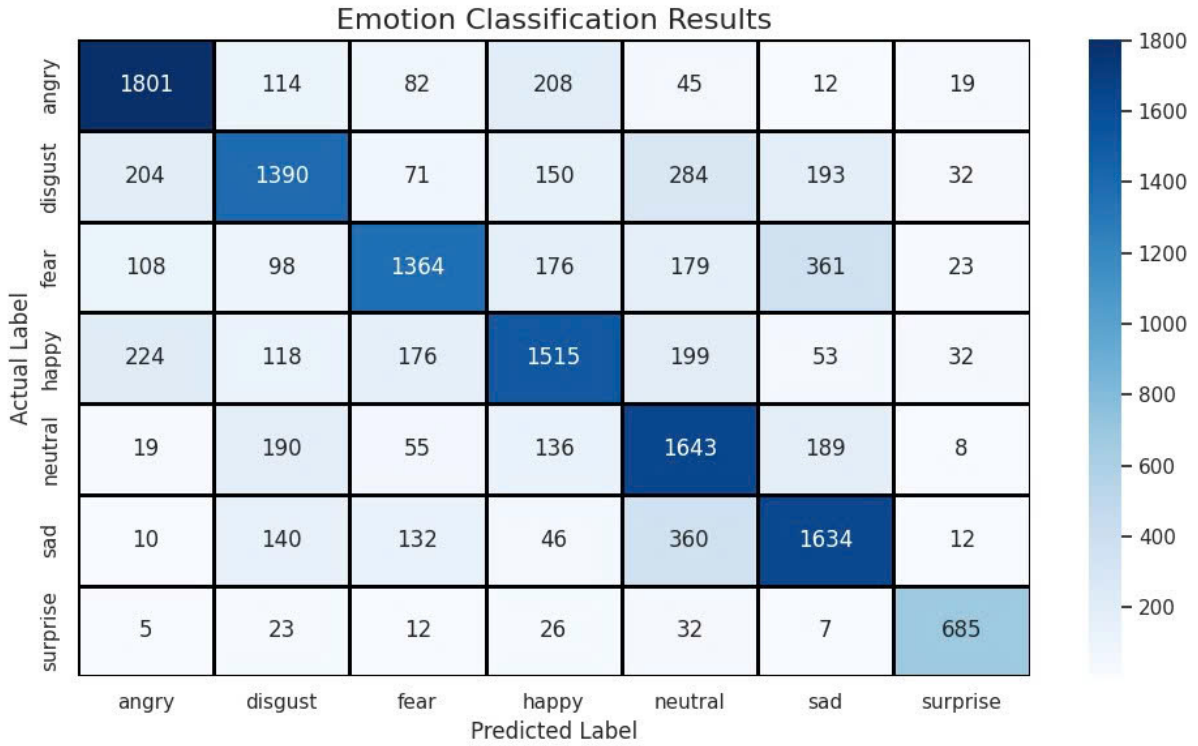


Figure 6: Confusion Matrix for Model C

- Model C frequently confuses Disgust with Neutral and Sad, and Happy with Fear.
- Poor separation between emotions with similar prosody but differing semantics—likely due to absence of speaker gender cues.
- General trend: diagonal dominance weakens, off-diagonal values rise → confirms label generality hurts precision.

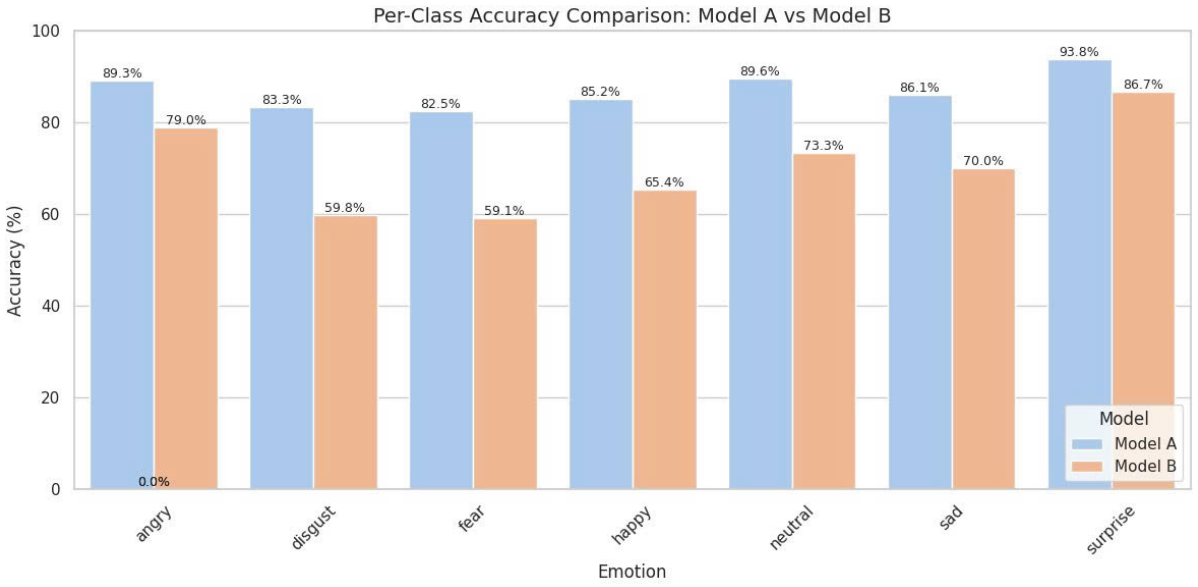


Figure 7: Per-Class Accuracy Bar Chart

4.4 Comparative Benchmarks

To contextualize our results, we reference key prior studies in speech emotion recognition:

- **Mirsamadi et al. (2017)**: Employed an LSTM with attention mechanism on the IEMOCAP dataset, reporting approximately 71% accuracy in a four-class emotion classification task.
- **Zhao et al. (2019)**: Applied a CNN-RNN model on the RAVDESS dataset and achieved around 75% accuracy for emotion-only classification.
- **Akçay et al. (2020)**: Used a CNN-BiLSTM architecture on RAVDESS and SAVEE, reporting an accuracy of 80.6%.

Our proposed **Model A** achieves a validation accuracy of **85.5%**, which exceeds the performance reported in these studies. This improvement can be attributed to several factors:

- Cross-corpus training using four diverse datasets (RAVDESS, TESS, SAVEE, and CREMA-D) compared to one or two datasets in prior work.
- Label fusion strategy combining gender and emotion information to enhance class separability.
- Rich data augmentation and stratified splitting to ensure balanced and robust training.
- Careful architectural tuning to maintain depth without over-parameterization.

However, it is important to note that direct comparisons are not entirely fair due to variations in dataset composition, label granularity, and augmentation strategies across studies.

4.5 Summary

The results confirm that:
• Gender-aware emotion modeling significantly improves performance.
• More CNN layers do not guarantee better results, and may introduce overfitting.
• Neutral and Sad remain difficult to distinguish due to overlapping acoustic profiles.
• Data augmentation and label engineering are vital in maximizing generalization across corpora.

5 Discussion

The goal of this project was to build a well-generalized Speech Emotion Recognition (SER) system using CNN-BiLSTM architectures trained on a diverse, augmented, multi-corpus dataset. While the results are promising—reaching nearly 90% accuracy—the journey to achieving this performance involved several trade-offs and challenges, which we reflect on here.

5.1 Accuracy vs. Architectural Complexity

One of the most revealing findings of this study was that a moderate-depth architecture (Model A: 4 CNN layers + 1 BiLSTM) outperformed a deeper variant (Model B with 5 CNN layers). This illustrates a classic trade-off between complexity and performance:
• Deeper networks may offer more representational capacity but are more prone to overfitting, especially when data variability (acoustic, demographic, or linguistic) is insufficient to justify added depth.
• Shallower networks, when combined with carefully crafted augmentations and engineered labels (i.e., gender-emotion), can perform just as well—or better, due to better generalization and faster convergence.

Additionally, increasing CNN layers beyond a point introduced training instability without significant gains in accuracy, suggesting that for SER tasks—particularly those using 2D spectral representations—model simplicity often outperforms raw depth.

5.2 Dataset Limitations

While the combined use of RAVDESS, TESS, SAVEE, and CREMA-D covers a range of speakers, accents, and emotions, several limitations remain:
• Scripted speech: All datasets use pre-defined phrases. This introduces bias toward exaggerated, acted expressions rather than spontaneous, natural emotional speech. Generalizing to in-the-wild audio may reduce accuracy.
• Language homogeneity: All datasets are English-based (North American, British, Canadian). There is limited representation of multilingual or accented speech, which can reduce robustness in real-world deployments.
• Gender binary assumption: The model was trained using male/female categories, excluding non-binary or transgender vocal profiles, which may encode emotion differently. A more inclusive label schema is needed for ethical SER.

Furthermore, while label harmonization across datasets ensured uniformity, there may still be annotation mismatches (e.g., differences in how "happy" or "neutral" is acted across corpora) that introduce hidden noise into training.

5.3 Generalization to Unseen Speakers

The train-test split ensured no speaker overlap, making the evaluation a fair proxy for out-of-distribution generalization. Model A's strong test accuracy (85.5%) and high recall for all classes, especially rare ones like *surprise*, suggest good robustness to unseen voices.

However, true generalization requires evaluation on unseen datasets. For example, a model trained on these four corpora should be tested on external, real-world datasets (e.g., IEMOCAP or MSP-IMPROV) to assess:

- Speaker variation (pitch range, timbre)
- Acoustic environment diversity
- Language and dialect shifts

Such cross-dataset validation is critical for practical deployment, especially in customer-facing applications (e.g., call centers or assistive tech).

5.4 Future Directions

Several opportunities exist to improve both performance and generalization:

1. **Advanced Architectures**
    - Transformer-based models (e.g., wav2vec 2.0 or AudioSpectrogram Transformer) offer potential for capturing long-range temporal dependencies more efficiently than LSTMs.
    - Self-supervised learning on large unlabeled corpora could help reduce reliance on annotated datasets.
2. **Multimodal Integration**
    - Adding visual data (e.g., lip movement, facial expressions) from RAVDESS or CREMA-D could enhance accuracy—particularly for ambiguous emotions (Tzirakis et al., 2017).
    - Incorporating textual transcripts could enable emotion classification from both prosody and semantic content.
3. **Real-Time Evaluation**
    - Current models operate on pre-processed, fixed-length inputs. Adapting the model for real-time, streaming inference is essential for deployment in interactive applications.
4. **Inclusive and Ethical SER**
    - Expanding gender categories and training with diverse demographic groups (age, ethnicity, gender identity) is crucial.
    - Exploring bias auditing—e.g., measuring if the model performs worse on female or minority speakers—can guide responsible SER development.
5. **Cross-Dataset Evaluation**
    - Evaluating trained models on datasets not seen during training will offer a more realistic estimate of performance in the wild.
    - Developing benchmark suites for SER (similar to GLUE for NLP) could standardize progress measurement.

References

Neumann, M., & Vu, N. T. (2018). Attentive convolutional neural networks for speech emotion recognition (*arXiv:1802.05630v2*). *arXiv*. `https://arxiv.org/abs/1802.05630`

Schuller, B., Steidl, S., Batliner, A., et al. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Proceedings of INTERSPEECH 2013*, 148–152.

Trigeorgis, G., Nicolaou, M. A., Zafeiriou, S., & Schuller, B. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5200–5204. https://doi.org/10.1109/ICASSP.2016.7472613

Latif, S., Qayyum, A., Usama, M., & Qadir, J. (2020). Deep architecture enhancement for Speech Emotion Recognition systems. *IEEE Access*, 8, 150530–150542. https://doi.org/10.1109/ACCESS.2020.3016652

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8), 1301–1309. https://doi.org/10.1109/JSTSP.2017.2764438