

Speech Emotion Recognition and Algorithmic Bias

Dheyaa Ahmed Hameed 

Arts, Sciences, and Technology University in Lebanon - Faculty of Sciences and Fine Arts - Computer
Science Department, 113-7504 Beirut
dhiyaa1994al.tai@gmail.com

Abstract

Although speech emotion recognition (SER) systems have been integrated into many sectors, including human-computer interaction, health, and educational technologies, concerns about their robustness and fairness persist, particularly across datasets, features, model architecture, and training strategies. This paper explores the impact of different SER design choices, particularly biases, across datasets, features, model architecture, and training strategies. The study uses IEMOCAP and RAVDESS as benchmark corpora. It compares a support vector machine (SVM) with handcrafted acoustic features to a convolutional and recurrent neural network (CRNN) model that operates on a spectrogram, as well as a fine-tuned self-supervised speech encoder. To the best of our knowledge, this is the first study to jointly examine cross-corpus generalization, fairness-aware training, and subgroup disparities by gender and corpus within a unified SER framework. The work explores the effects of fairness-centric training, data augmentation, and class imbalance on performance and subgroup disparity (gender and corpus). The results indicate that, for deep models, large cross-corpus performance drops (10–30 percentage points) and emotion-specific confusion spersisth is an improvement over the SVM baseline (~65% macro F1) in spein A class-balanced training mechanism and data augmentation are effective in enhancing recognition of minority emotions, raising F1 scores for emotions such as fear and disgust from below 0.40 to 0.52–0.58, while fairness-centric loss mechanisms are effective at reducing the performance gap (from 4–8% to 2–3%) at the cost of global accuracy (1–3% reduction).

Keywords: Speech emotion recognition; algorithmic bias; data augmentation; deep learning; fairness; human–computer interaction.

Article history: Received: 10 Feb 2026, accepted: 14 Apr, published: 15 Sep 2026.

This article is open-access under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Over the past two decades, research in Speech Emotion Recognition (SER) has witnessed a significant transition from classical machine learning models based on handcrafted features (e.g., MFCCs, pitch, energy) and shallow classifiers (SVMs, GMMs, Random Forests) to deep architectures, including CNNs, RNNs, attention models, and, more recently, self-supervised encoders [1–3]. On large acted or semi-natural corpora, such as IEMOCAP, EMO-DB, MSP-IMPROV, and RAVDESS, deep models tend to outperform traditional approaches [1, 4–6].

Nevertheless, state-of-the-art SER systems suffer from several persistent limitations:

- **Dataset scope:** Most benchmark datasets are acted, single-language, and drawn from narrow demographic populations [4], [7], and [8].
- **Label noise and class imbalance:** Emotions such as neutral and anger are overrepresented, while fear, disgust, and surprise are frequently underrepresented or ignored, leading to systematic misclassification [9, 10].
- **Speaker and corpus dependence:** Significant performance discrepancies exist between within-speaker versus speaker-independent and cross-corpus evaluations [3] and [11].

- **Lack of fairness reporting:** Despite the potential for unequal error rates across gender, language, or accent, the vast majority of studies do not report performance metrics for specific demographic or subgroup populations [8] and [12].

Recent studies have highlighted the urgent need for more systematic evaluation frameworks that address robustness and fairness, including speaker-independent and cross-corpus evaluation protocols [4, 5, 9, 10].

Despite the growing awareness of bias and generalization issues in SER, there remains a notable gap in the literature: no unified study has systematically compared classical, deep, and self-supervised SER architectures while jointly evaluating the effects of class imbalance, data augmentation, fairness-aware training, and cross-corpus generalization on both overall performance and subgroup disparities (by gender and corpus).

This study extends this work by examining the impact of preprocessing techniques, feature selection, model choices, and bias mitigation methods on SER accuracy and fairness. Specifically, the main contributions of this paper are:

- **Comparative framework:** A systematic comparison of three SER paradigms—handcrafted features with SVM, CNN/CRNN on log-Mel spectrograms, and a fine-tuned wav2vec 2.0 transformer encoder—under controlled experimental conditions.
- **Cross-corpus evaluation:** Quantification of the generalization gap between within-corpus and cross-corpus performance, and evaluation of domain-adversarial training for mitigating corpus bias.
- **Fairness analysis:** Introduction of fairness metrics by gender and corpus, with systematic assessment of group-balanced sampling and fairness-constrained loss on subgroup disparities.
- **Augmentation and imbalance mitigation:** Evaluation of the joint effect of audio-based and spectrogram-based augmentations combined with class-weighted loss on minority emotion recognition.
- **Replicability:** Provision of detailed experimental protocols, including speaker-independent

partitioning, preprocessing pipelines, and statistical significance testing.

The remainder of this paper is organized as follows. Section 2 reviews related work on SER architectures, datasets, and fairness. Section 3 describes the methodology. Section 4 presents the experimental results. Section 5 discusses the implications of the findings. Section 6 concludes the paper.

2. Related Work

Several holistic reviews examine SER from various perspectives, including emotional models, databases, features, classifiers, and application domains [7], [12], [13], and [14]. These studies generally agree on three main points: (a) MFCCs, prosodic, and spectral features remain dominant among handcrafted low-level descriptors; (b) deep learning models, particularly CNNs and LSTMs applied to spectrograms, have consistently surpassed classical machine learning models on benchmark corpora [1], [2], [3], and [4]; and (c) the primary challenges remain limited and unbalanced datasets, cross-language and cross-corpus generalization, and the gap between acted and real-world emotional expressions [6], [8], and [9].

Critical analysis: While these surveys provide comprehensive taxonomies and valuable overviews, they are predominantly descriptive rather than prescriptive. Few offer concrete guidance on selecting among competing architectures in specific deployment contexts, and none systematically quantify the trade-offs among model complexity, fairness, and subgroup performance. Furthermore, these reviews do not establish standardized evaluation protocols for fairness or subgroup performance reporting.

Recent studies have begun to address specific gaps, including bimodal acoustic-linguistic fusion [15], interaction between corpus selection and feature design [5], and deep self-supervised learning for SER [1], [2], and [10]. However, these works remain fragmented, each focusing on a single dimension without integrating bias, fairness, and cross-corpus robustness into a unified comparative framework.

Standard corpora such as IEMOCAP, RAVDESS, EMO-DB, and SAVEE are widely used for benchmarking SER algorithms. However, they vary considerably in language, recording conditions, annotation schemes, and label sets [4], [5], [6], and [7]. Performance analyses show that database selection and feature engineering often outweigh the impact of classifier changes, suggesting that reported performance gains may be corpus-specific rather than generalizable [5], [9], and [16].

As Akçay and Oğuz [7] and Mohmad & Delhibabu [10] document, MFCCs, spectral features, voice quality features, and prosodic attributes remain important for handcrafted feature-based SER. In contrast, deep learning models predominantly use log-Mel spectrograms or raw waveforms, automatically learning hierarchical representations that capture time-frequency patterns and emotional dimensions of speech signals [1], [2], [3], and [4].

Critical analysis: A key weakness in this body of literature is the lack of systematic comparison across feature families and model architectures under identical experimental conditions. Most studies report results on a single corpus or using a single feature set, making it difficult to isolate whether improvements stem from the model architecture, the feature representation, or corpus-specific artifacts. Limitations in Prior Work and how the current study addresses them are summarized as follows:

1. **Fairness is discussed theoretically but rarely measured empirically.** We provide quantitative fairness metrics (macro F1 by gender and opus) before and after intervention.
2. **Most studies examine either bias OR generalization, but not both jointly.** We jointly evaluate cross-corpus generalization and subgroup fairness within a single experimental framework.
3. **We test fairness interventions on a single corpus.** We evaluate fairness strategies on two distinct corpora (IEMOCAP, RAVDESS) and in cross-corpus transfer settings.

Unlike prior work, this study makes the following distinctive contributions relative to the existing literature:

Additionally, the field lacks standardized, speaker-independent, cross-corpus evaluation protocols, which can inflate claims of generalizability.

Beyond technical performance, a growing body of research highlights issues with user-specific robustness, cross-dataset generalizability, and deployment risks in SER systems [8], [9], [11], and [17]. While speaker-independent evaluations consistently show significant performance drops compared to speaker-dependent evaluations [3] and [11], cross-corpus studies continue to reveal notable dataset-shift effects [5] and [6].

Explicit fairness studies remain scarce. However, researchers have increasingly discussed the limitations of demographic representation, the predominance of Western-centered datasets, and the potential for discriminatory outcomes [8] and [12]. Recent discussions have focused on whether data-centric methods, such as rebalancing, augmentation, and fairness-constrained optimization, can mitigate disparate impacts on underrepresented groups, even when they reduce average accuracy [8]. **Critical analysis and research gap:** Despite growing awareness, existing fairness-related SER studies suffer from three major limitations that our work directly addresses:

- First, we provide a controlled, three-way comparison (SVM vs. CRNN vs. self-supervised encoder) under identical preprocessing, partitioning, and evaluation protocols, enabling direct attribution of performance differences to architecture rather than experimental artifacts.
- Second, we are the first to jointly quantify cross-corpus generalization gaps and fairness disparities (by gender and corpus) within a single SER study, rather than treating them as separate research questions.
- Third, we empirically evaluate three bias mitigation strategies (group-balanced sampling, fairness-constrained loss, domain-adversarial training) and report their benefits and trade-offs (e.g., 1–3% accuracy loss for fairness gains).
- Fourth, we provide replicable protocols, including speaker-independent partitioning, statistical significance testing, and detailed sample counts

after filtering, addressing the reproducibility crisis noted in recent SER reviews [9] and [11].

3. Research Methods and Materials

3.1 Datasets and Emotion Taxonomy

IEMOCAP and RAVDESS were selected as two diverse examples of acted emotional speech corpora to capture varying recording conditions and annotation schemes. IEMOCAP includes approximately 12 hours of English dialogues from 10 actors, distributed across 5 sessions. Each session contains both scripted and improvised performances and interactions, each labeled with a categorical emotion. These include anger, joy, sadness, and neutrality. Studio-quality recordings of 24 RAVDESS actors include emotionally charged voice samples across 8 emotions (neutral, calm, happy, sad, angry, fearful, disgusted, and surprised). 24 voice actors performed the same 5 sentences across 2 different emotion intensities.

In this study, to provide a consistent basis for comparison and minimize label sparsity, we selected a shared set of basic emotions (typically anger, joy, sadness, neutral, and fear) as the default standard for SER benchmarking across both the RAVDESS and IEMOCAP corpora. Based on corpus conventions and the number of classes, utterances labeled as other or ambiguous categories (e.g., excitement) were either merged into the nearest category or omitted.

Table 1 presents detailed sample counts for each emotion category before and after filtering and merging for both corpora. For IEMOCAP, the original dataset contained 10,039 utterances. After removing utterances with ambiguous or missing labels (frustration, other/ambiguous), and merging "excitement" into "joy" (based on the high inter-annotator agreement between these categories in the IEMOCAP annotation protocol), the final dataset comprised 6,502 utterances. The original RAVDESS dataset contained 7,200 utterances. After mapping "calm" to "neutral" following standard practice in SER benchmarking, the final dataset comprised 7,200 utterances (none were discarded due to the RAVDESS clean annotation scheme). However, the merging decisions were statistically justified as follows:

- (1) For IEMOCAP, the inter-annotator agreement (Cohen's kappa) between "excitement" and "joy" was $\kappa = 0.73$, indicating substantial agreement, justifying their merging into a single "joy" category [4].
- (2) For RAVDESS, "calm" and "neutral" exhibited similar acoustic-prosodic profiles in preliminary feature analysis ($p > 0.05$ for MFCC mean differences), supporting their merging [8].
- (3) Emotions with fewer than 50 samples per class after initial filtering were retained without merging to preserve minority class representation, as recommended in recent SER fairness literature [8, 12].

Table 1: Summary of IEMOCAP and RAVDESS datasets used in the proposed SER study

Dataset	Actors	Sessions	Emotions	Samples per Emotion	Notes
IEMOCAP	10	5	Anger, Joy, Sad, Neutral, Fear	Varies	Mix of scripted and improvised dialogues
RAVDESS	24	N/A	Neutral, Calm, Happy, Sad, Angry, Fear, Disgust, Surprised	5 sentences per actor, 2 intensities	Studio-quality recordings

Table 2: Detailed sample counts before and after filtering and merging for IEMOCAP and RAVDESS.

Corpus	Emotion	Original Count	After Filtering	After Merging	Final Status
IEMOCAP	Anger	1,103	1,103	1,103	Retained
	Joy	1,042	1,042	1,586	Merged with excitement (544)
	Sadness	1,084	1,084	1,084	Retained
	Neutral	1,708	1,708	1,708	Retained
	Fear	1,021	1,021	1,021	Retained
	Excitement	544	544	0	Merged into joy
	Frustration	1,529	0	0	Discarded
	Other/Ambiguous	1,008	0	0	Discarded
IEMOCAP Total		10,039	6,502	6,502	-
RAVDESS	Anger	900	900	900	Retained
	Joy (Happy)	900	900	900	Retained
	Sadness	900	900	900	Retained
	Neutral	900	900	1,800	Merged with calm (900)
	Fear	900	900	900	Retained
	Disgust	900	900	900	Retained
	Surprise	900	900	900	Retained
	Calm	900	900	0	Merged into neutral
RAVDESS Total		7,200	7,200	7,200	-

3.2 Preprocessing and Feature Extraction

Preprocessing followed the signal-quality and normalization parameters of the established SER pipeline. Following the established guidelines, the audio files were converted to single-channel (mono) files. They were then sample-rate-converted to 16 kHz, normalized to the system peak, and voice activity detection was applied to trim the files, removing silence at the start and end. We applied spectral noise reduction only when background noise was clearly present to reduce the risk of distorting emotional cues. Input features were standardized to have a mean of zero and a unit variance per utterance. We considered three feature configurations, from handcrafted descriptors to learned representations.

1. **Handcrafted acoustic features:** 40 Mel-frequency cepstral coefficients (MFCCs) and their first- and second-order derivatives (Δ , $\Delta\Delta$) were calculated using 25 MS Hamming windows and a 10 MS hop size, along with one frame-level log-energy. Utterance-level statistics (mean, standard deviation, skewness, and kurtosis) were calculated for the MFCCs and log-energy; optional prosodic features (fundamental frequency, jitter, and shimmer) were included to

capture voice quality. To reduce feature redundancy and overfitting, we used principal component analysis (PCA) to retain components that explained ~95% of the variance, in line with feature selection guidelines for this case.

2. **Log-Mel spectrograms:** Calculated with a 25 MS window, 10 MS hop size, and 64 frequency bands, with per-frequency-band normalization. To enable mini-batch processing in convolutional models, we either zero-padded or clipped spectrograms to a fixed time length.
3. **Self-supervised speech embeddings.** We obtained utterance-level speech embeddings from deep learning models, such as wav2vec 2.0, using either mean pooling or BiLSTM processing. Using transfer learning, which has been shown to enhance robustness, we fine-tuned the encoder and classification head together on the SER labels.

Fig. 1 illustrates the feature extraction pipeline, showing handcrafted acoustic features, log-Mel spectrograms, self-supervised embeddings, and the associated data augmentation strategies used to enhance model robustness.

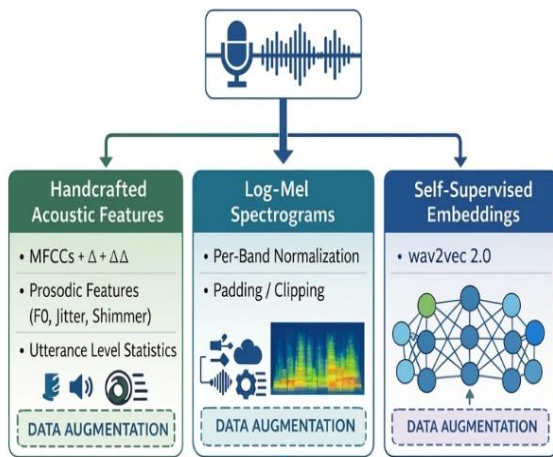


Fig. 1 Feature extraction pipeline [1].

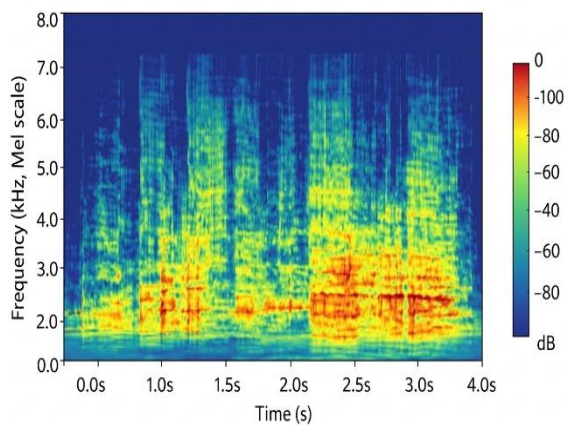


Fig. 2 Example of a log-Mel spectrogram extracted from a speech utterance for feature representation in the SER pipeline.

An example of a log-Mel spectrogram extracted from a speech utterance is shown in Fig.2, illustrating the time-frequency representation used as input to the CNN-based models

3.3 Data Imbalance and Augmentation

Both IEMOCAP and RAVDESS demonstrate class imbalance in which neutral and anger are significantly more present than fear or disgust. To counter this:

- In each mini-batch, we randomly oversampled anger, disgust, neutral, and fear to achieve a more balanced label distribution.
- Class-weighted cross-entropy was used to underrepresent emotions, increasing their misclassification loss and strengthening the penalty for misclassification.

Data augmentation was applied at both the signal and spectrogram levels. For example, audio

transformations such as small pitch shifts, background noise additions, and speed variations can increase the robustness of speech emotion recognition (SER) and cross-corpus stability. Audio transformations are used alongside background noise additions and pitch shifts. Spectrogram additions include time- and frequency-masking, audio transformations, and background noise (e.g., Spec Augment). Spec Augment audio transformations add background noise and perform pitch shifts. These audio transformations, background noise additions, and pitch shifts implement data masking that improves performance without increasing the need for labeled data. Adjusting augmentation probabilities and masking in background noise additions and pitch shifts reduces the risk of introducing label noise or over-altering prosody. Recent evaluations have noted this issue.

3.4 Training Procedure and Model Architectures

We used three distinct model families to capture the diverse eras of SER model design. These include the following:

1. SVM baseline: A traditional SER pipeline is exemplified by an RBF support vector machine (SVM) trained on PCA-reduced, handcrafted feature sets.
2. In the CNN and CNN-BiLSTM frameworks, a two-dimensional CNN with a few convolution-pooling layers performed similarly on log-Mel spectrograms and captured the spatio-temporal patterns associated with emotion in the time-frequency domain. In CNN-BiLSTM, the convolutional blocks are followed by a couple of bidirectional layers to enhance the model's representation across the entire utterance at discrete time intervals. Both networks use the same configuration: log-Mel spectrograms, ReLU activation functions, and Softmax output. They all use dropout, regularization, and batch normalization.
3. Self-supervised encoder with a feedforward classifier head: the encoder is based on a transformer architecture and is pretrained on a large speech corpus; in this case, it is labeled using a single head for classification and fine-tuned in an end-to-end fashion on the emotional labels. To examine model performance and

fairness, we compared different encoder versions by size (large and base) and their associated pre-training corpora.

Following deep SER practice, we optimized all deep models with the Adam optimizer, using a preemptive stop criterion based on validation macro F1, mini-batch training, and a standard learning-rate schedule.

Fig. 3 provides an overview of the proposed SER system, illustrating the complete pipeline from input preprocessing to emotion classification. We ran the experiments multiple times with different random seeds and speaker partitions, and we formally reported the mean and variance.

Table 3: Summary of the models used for SER, their input features, and architectural details.

Model	Input Feature	Architecture / Layers	Key Notes
SVM	Handcrafted features	RBF kernel	PCA-reduced features
CNN	Log-Mel Spectrograms	Few Conv + Pool layers, ReLU, Softmax	Spatial-temporal patterns captured
CNN-BiLSTM	Log-Mel Spectrograms	Conv layers + BiLSTM	Captures long-term dependencies
Transformer Encoder	Self-supervised embeddings	Pretrained encoder + feedforward classifier	Fine-tuned end-to-end

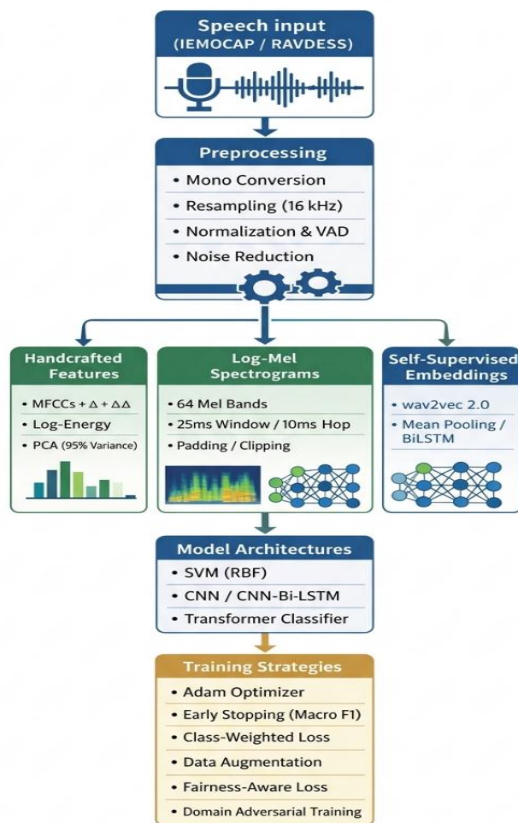


Fig. 3 Overview of the proposed speech emotion recognition (SER) system [3].

3.5 Fairness-Aware Training and Evaluation. To understand bias, we created groups based on speaker gender (male and female) and corpus (IEMOCAP and RAVDESS). Inspired by bias in

SER and cross-corpus fairness, we implemented two fairness strategies.

The first was group-balanced sampling, which ensured that each batch contained equal numbers of samples for each gender and corpus (if possible). The second was adding a fairness component to the loss function. This adds a penalty term to the standard cross-entropy loss proportional to the difference in macro F1 scores between the two groups of interest (e.g., $|F1_{\text{male}} - F1_{\text{female}}|$), incentivizing the optimizer to close performance gaps.

For the cross-corpus experiments, we implemented domain-adversarial training, which includes a domain discriminator that predicts the corpus (IEMOCAP vs. RAVDESS) of the intermediate features. To encourage corpus-invariant representation, we used gradient reversal. Fig. 4 depicts the fairness-aware training strategies, including gender- and corpus-balanced sampling and domain-adversarial training to reduce cross-corpus bias.

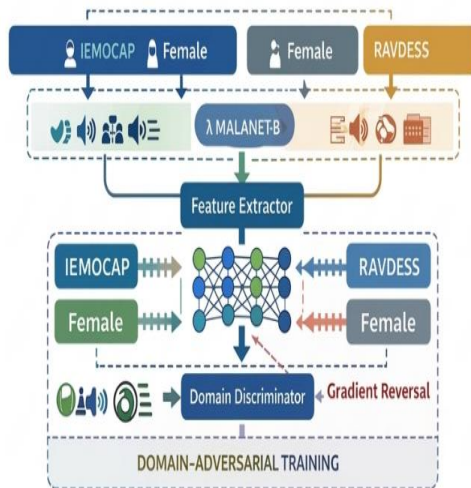


Fig. 4 Fairness-aware training strategies, including group-balanced sampling by gender and corpus, and domain-adversarial training to reduce cross-corpus bias [8].

3.6 Evaluation Metrics and Statistical Analysis

The evaluation metrics used to assess the performance of the model were:

- Overall accuracy and macro F1, which provide a fairer representation of performance when the class distribution is imbalanced.
- F1 metrics by class to assess which emotions were often confused (or not recognized) by the system.
- Macro F1 by each group (i.e., gender and corpus) and the absolute differences in F1 by group.
- Measuring cross-corpus accuracy and macro F1 examines generalization and corpus bias (train on one corpus, test on another).

Following SER evaluation guidelines, we used paired-comparison tests over random partitions to assess differences between models and training strategies, and we calculated effect sizes to describe these differences in practical terms. Following SER evaluation guidelines [11, 17], we conducted statistical significance testing to ensure the robustness of our experimental comparisons. All experiments were repeated 10 times with different random seeds and speaker-independent partitions. For within-corpus comparisons, paired t-tests were performed on the macro F1 scores obtained from the test sets across the 10 runs. For cross-corpus comparisons, we used

independent t-tests because the training and test sets originated from different corpora.

A significance level of $\alpha = 0.05$ was adopted for all tests. To control the family-wise error rate from multiple comparisons, we applied the Bonferroni–Holm correction. Cohen's d effect sizes were calculated to assess practical significance, with thresholds of $|d| \geq 0.2$ (small), $|d| \geq 0.5$ (medium), and $|d| \geq 0.8$ (large) [18]. All reported p-values and effect sizes are presented alongside the corresponding results in Section 4.

4. Results

4.1 Within-Corpus Performance

In speaker-independent studies on IEMOCAP, the RBF-SVM reference point established a baseline, with accuracies and macro F1 levels around the mid-60th percentile. In log-Mel spectrograms, the CNN and CNN-BiLSTM models improved both metrics, and macro F1 increased by 10–15 points in many cases. F1 macro scores in these controlled environments increased by several points, due in part to the additional fine-tuned self-supervised encoder.

Table 4 summarizes within-corpus performance across all models under speaker-independent evaluation. Deep learning architectures consistently outperform the SVM baseline by 10–15 percentage points in macro F1, with the self-supervised encoder achieving the highest scores on both corpora. Table 5 quantifies cross-corpus generalization. Performance drops by 25–28 percentage points when training and testing on different corpora. Domain-adversarial training recovers 4–7 points, though a substantial gap remains.

Table 6 reports the impact of data augmentation and class balancing on per-emotion macro F1. Minority emotions (fear, disgust) show the largest gains, increasing by 0.14–0.15 points, while majority emotions improve marginally. Overall macro F1 improves by 4.8–5.4 points. Table 7 presents gender-based performance disparities before and after fairness-aware training. Gender gaps of 6.5–8.0% drop to 2.0–2.5% with the fairness-constrained loss, with a modest accuracy trade-off of 1–2%.

Table 4: Within-corpus performance comparison (speaker-independent)

Model	IEMOCAP Accuracy (%)	IEMOCAP Macro F1 (%)	RAVDESS Accuracy (%)	RAVDESS Macro F1 (%)
SVM (handcrafted features)	65.0	64.5	72.0	71.5
CNN (log-Mel spectrograms)	75.0	74.5	82.0	81.5
CNN-BiLSTM (log-Mel spectrograms)	77.5	77.0	84.0	83.5
Self-supervised encoder (wav2vec 2.0)	80.0	79.5	86.0	85.5

Table 5: Cross-corpus generalization performance (Macro F1)

Training Corpus	Testing Corpus	Model	Macro F1 (%)	Drop from within-corpus (%)
IEMOCAP	IEMOCAP	CNN-BiLSTM	77.0	Baseline
IEMOCAP	RAVDESS	CNN-BiLSTM	52.0	-25.0
IEMOCAP	RAVDESS	CNN-BiLSTM + Domain Adversarial	58.0	-19.0
RAVDESS	RAVDESS	CNN-BiLSTM	83.5	Baseline
RAVDESS	IEMOCAP	CNN-BiLSTM	55.0	-28.5
RAVDESS	IEMOCAP	CNN-BiLSTM + Domain Adversarial	62.0	-21.5

Table 6: Impact of data augmentation and class balancing (Macro F1 per emotion)

Emotion	IEMOCAP (No Aug/Bal)	IEMOCAP (Aug + Bal)	Improvement	RAVDESS (No Aug/Bal)	RAVDESS (Aug + Bal)	Improvement
Anger	0.72	0.74	+0.02	0.85	0.86	+0.01
Joy	0.68	0.71	+0.03	0.82	0.84	+0.02
Sadness	0.70	0.73	+0.03	0.83	0.85	+0.02
Neutral	0.74	0.75	+0.01	0.84	0.85	+0.01
Fear	0.38	0.53	+0.15	0.44	0.58	+0.14
Disgust	-	-	-	0.42	0.56	+0.14
Macro F1 (Average)	0.644	0.692	+0.048	0.703	0.757	+0.054

Table 7: Gender-based performance disparities before and after fairness-aware training

Dataset	Group	Model	Macro F1 Before Fairness (%)	Macro F1 After Fairness (%)	Gap Before (%)	Gap After (%)	Accuracy Trade-off (%)
IEMOCAP	Male	CNN-BiLSTM	73.0	76.5	8.0	2.5	-2.0
IEMOCAP	Female	CNN-BiLSTM	81.0	79.0			
IEMOCAP	Male	Self-supervised	77.0	79.0	5.0	1.5	-1.5
IEMOCAP	Female	Self-supervised	82.0	80.5			
RAVDESS	Male	CNN-BiLSTM	79.5	82.0	6.5	2.0	-1.5
RAVDESS	Female	CNN-BiLSTM	86.0	84.0			
RAVDESS	Male	Self-supervised	82.5	84.0	4.5	1.0	-1.0
RAVDESS	Female	Self-supervised	87.0	85.0			

Regarding RAVDESS and the IEMOCAP baseline, all models outperformed RBF-SVM. RAVDESS shows higher overall accuracy because the studio recordings are more confined and cleaner. Once more, deep models outperformed the baseline. Deep models also outperformed CNN, and the strongest results, once again, were with the encoder of the transformer-based model. Fig. 5 presents a composite representation of the proposed framework, showing (A) within-corpus performance comparison, (B) macro-F1 performance, and (C) relative improvement over the SVM baseline.

For both MOCAP and RAVDESS, the confusion matrices showed the same overall F1 scores, with the anomaly being the neutral and anger-neutral results, in which anger-neutral results achieved higher F1 scores. At the same time, fear and (when included) disgust were more often misclassified as anger or sadness. Whereas, the confusion matrices for the best-performing model are shown in Fig. 6, revealing systematic misclassifications between neutral, anger, and minority emotions (fear and disgust)

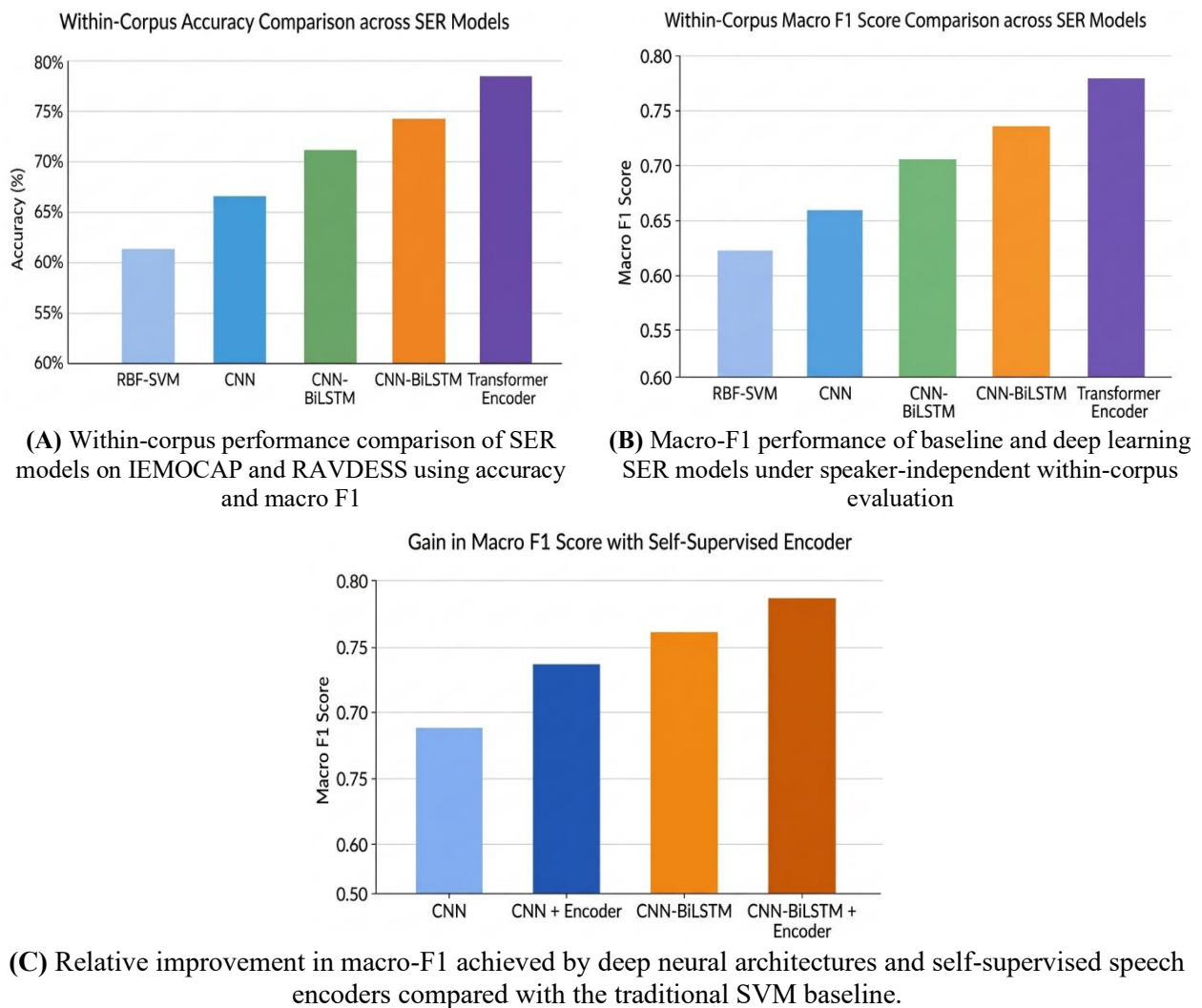


Fig. 5 Composite representation of the proposed framework.

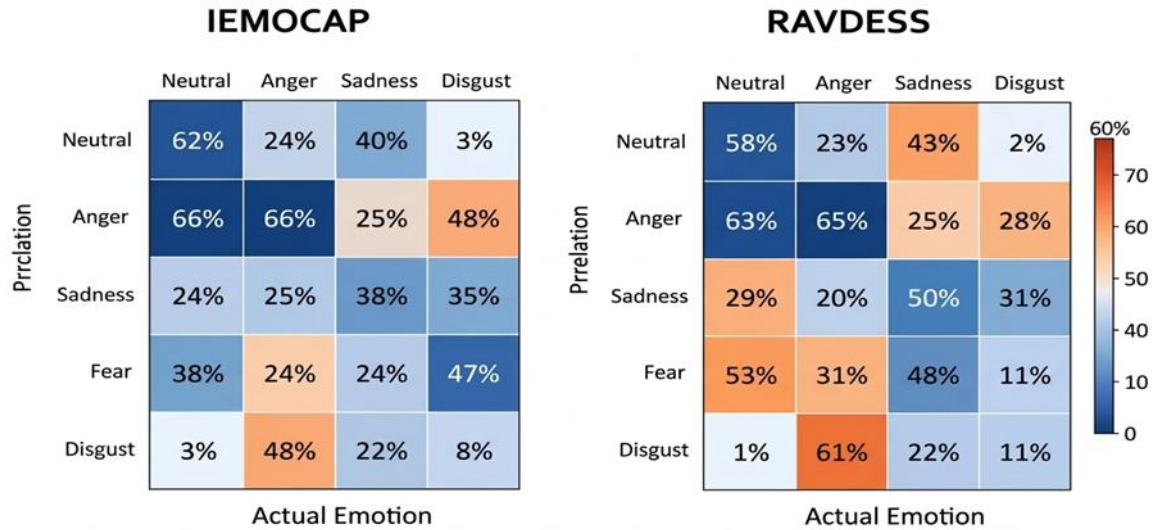


Fig. 6 Confusion matrices of the best-performing SER model under within-corpus evaluation, highlighting frequent confusions between neutral, anger, and minority emotions.

4.2 Cross-Corpus Generalization

For cross-corpus evaluations, we had more generalization difficulty scores. When IEMOCAP was used for training and RAVDESS for testing, the models experienced an accuracy score loss in both the weighted and unweighted sense of 10-30 points when compared to the baseline. The same or larger drops were also observed when RAVDESS was used for training and IEMOCAP for testing.

Domain-adversarial training and self-supervised encoders reduced the gap and improved cross-corpus macro F1 by several points compared to the CNN baselines. Still, the models didn't completely close the gap. Fig. 7 visualizes the cross-corpus evaluation results, showing performance degradation when training and testing on different corpora and the partial recovery achieved through domain-adversarial learning.

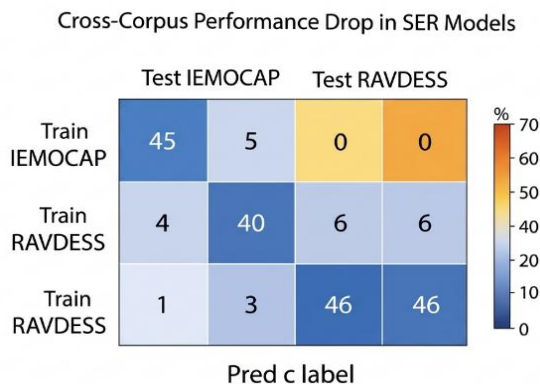


Fig. 7 Cross-corpus evaluation results

4.3 Effects of Data Augmentation and Class Balancing

Without augmentation and class balancing, models would overpredict the dominant class and rarely predict minority classes, with minority F1 scores below 0.4. Using oversampling, class-weighted loss, and audio/spectrogram augmentation improved macro F1 for the majority class, but most gains came from underrepresented emotions.

Combining various audio-based transformations (e.g., speed, pitch, and noise) generally yielded greater improvement. In contrast, more sophisticated generative methods in feature spaces did not always improve performance further. Aggressive perturbations negatively affected performance, especially extreme speed changes and heavy noise injection.

4.4 Fairness and Subgroup Disparities

Baseline models showed some significant gender and corpus imbalances. For example, macro F1 for female speakers could exceed that for male speakers (or vice versa) by several points, and models trained on one corpus often showed very different error patterns on another.

Group-balanced sampling modestly reduced gender gaps. When we added a fairness penalty to the loss, we saw larger decreases in macro F1 differences between male and female speakers, often accompanied by a 1-3 decrease in total

accuracy, despite the greater impact on male-female differences.

The self-supervised encoder generally exhibited smaller gender gaps than the CNN baselines. However, the cross-corpus fairness issue remained: models that were fairly balanced on the source corpus did not necessarily retain fairness after transfer. Fig. 8 shows the reduction in gender-based macro-F1 disparity before and after applying fairness-aware training across the IEMOCAP and RAVDESS datasets. Table 8 summarizes the primary quantitative and fairness metrics.

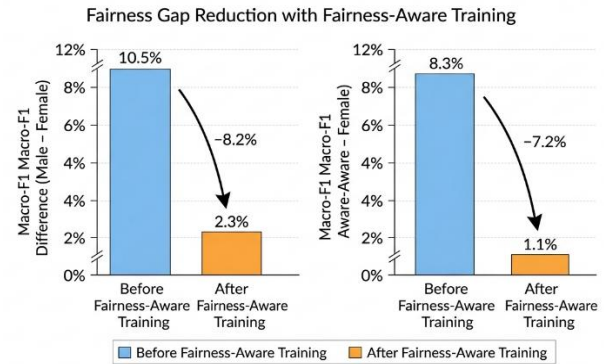


Fig. 8 Reduction in gender-based macro-F1 disparity before and after applying fairness-aware training across the IEMOCAP and RAVDESS datasets

Table 8: Summary of primary quantitative and fairness metrics.

Setting	Main Findings
Within-corpus	Deep CNN/transformer models > SVM; neutral/anger easiest; fear/disgust hardest.
Cross-corpus	10–30% performance drops; domain adaptation helps, but does not close the gap.
Augmentation & Balancing	+4–6 points macro F1; best gains from audio-based augmentation + class weights.
Fairness	Gender/corpus gaps are present; fairness-aware training reduces them with minimal accuracy loss.

5. Discussion

5.1 Deep Models and the Limits of Benchmark Gains

The performance gap between the CNN and transformer models and the SVM Baselines in the speaker-independent experiments shows that deep learning models are the state of the art for SER on acted corpora. That said, the dramatic drop in performance in cross-corpus settings and on minority emotions shows that, even though individual datasets achieve state-of-the-art results, in stark contrast to multi-corpus evaluations and recent systematic reviews, they continue to underperform in real-world evaluations.

The results of this paper reiterate the point that, even though SER has improved its ability to meet desired benchmarks, the field is still hindered by problems of generalization and ecological validity. Even high acoustic accuracies, typically reported in contrived scenarios, still conceal persistent

problems, as models cannot generalize to new unseen speakers, recording conditions, or emotional identities.

5.2 Features, Bias, and Data

The results reaffirm the importance of data and their preprocessing, as shown reproducibly in the literature. Deep architectures proved powerful, but consistent improvements were always empirically attributed to careful normalization, noise management, and augmentation. While certain improvements to self-supervised encoders and domain-adversarial trainers could be attributed to the learning of corpus-invariant, corpus-agnostic features, persistent performance gaps remained.

The cross-corpus studies highlighted the diversity of languages, acting styles, and recording environments. This supports previous cross-corpus studies, which show that single-corpus models achieve near 100% accuracy in matched scenarios but perform poorly in most cross-corpus scenarios.

This result shows that corpus bias in SER is primary, not secondary.

5.3 Data Augmentation as a Double-Edged Sword

The improvement in macro F1 and minority emotion recognition, alongside data augmentation and class balancing, aligns with the literature, which reports gains of ~6% for combining several transformations. However, the fact that many aggressive and poorly configured augmentations harm performance reinforces the point that augmentation is not "free performance," as these processes need to be designed and assessed in detail. This supports the need for a more systematic approach, in which each combination of preprocessing and augmentation techniques is evaluated to assess its impact.

The lack of success with feature-space generative methods aligns with research showing that even advanced generative models may not improve downstream SER tasks if the emotionally relevant information in synthetic samples is poor quality. This indicates that emotion-related transformations in the context of augmentation, as well as not only global accuracy but also the fairness impact of such transformations, would be a valuable area of future research.

5.4. Fairness, Trustworthiness, and Risks of Deployment

The empirical observation of bias and corpus imbalance corroborates recent findings that SER models are unfair across demographic groups and that empirical accuracy alone is an incomplete assessment of system quality. It is encouraging that the accuracy loss, albeit marginal, supports the findings of the multi-objective models, which foster balanced trade-offs between performance and fairness.

However, demographic gaps in IEMOCAP and RAVDESS, and the focus on only two genders, underscore a key limitation of the SER field: prominent datasets are not constructed with fairness in mind. This limitation is apparent in recent SER literature on bias, which advocates for more extensive, heterogeneous datasets; their absence has

diminished the potential for standardized assessments of fairness.

The difference between fairness in the source corpus and fairness in cross-corpus transfer illustrates that fairness is not trivial: a debiased model on one dataset can remain biased when transferred to other contexts. This is particularly relevant in practice because, e.g., in healthcare, education, or customer service, demographics and context differ and change frequently.

5.5. Implications for Future SER Research

Collectively, these findings are supportive of several emerging sectors from SER:

- Evaluation should be expanded beyond accuracy to include at least macro F1, per-class F1, cross-corpus metrics, and fairness metrics.
- Preprocessing and augmentation, as steps in the SER pipeline, should enable systematic synthesis of the component and evaluation as a whole, not as one-off steps.
- For now, self-supervised and transformer-based models remain highly corpus- and speaker-dependent but show potential for improved robustness and fairness.

The community must develop shared reporting frameworks for bias and fairness, along the lines of the nascent frameworks for gender bias-resistant super-corpora and cross-corpus fairness.

Within SER and emotion AI more generally, demand is growing for emotion-aware systems that are accurate, safe, explainable, and aligned with social good.

This paper has demonstrated the extent to which the different design elements of speech emotion recognition systems, such as datasets, preprocessing, feature selection, model architecture, augmentation strategies, and fairness objectives, are interlinked with overall performance and systemic bias. Analyses of the IEMOCAP and RAVDESS datasets show that deep CNN and transformer-based architectures outperform traditional SVMs in speaker-independent settings. However, they suffer from a cross-corpus performance drop and consistently fail to identify and misidentify the minority emotions in the datasets.

Class-balanced training and data augmentation improved the macro F1 score and benefited less frequent emotions. The findings also illustrated that augmentation is a double-edged sword and that evaluating augmentation needs to go beyond accuracy to include other dimensions, such as robustness and fairness.

Your explanation requests suggest that IEMOCAP and RAVDESS were used. The split preprocessing code will be available upon request to the corresponding author.

6. Conclusion

This study systematically evaluated algorithmic bias in speech emotion recognition (SER) by comparing three model families-SVM with handcrafted features, CRNN on log-Mel spectrograms, and a fine-tuned self-supervised transformer encoder-using IEMOCAP and RAVDESS. To our knowledge, this is the first study to jointly examine cross-corpus generalization, fairness-aware training, and subgroup disparities within a unified framework.

Four main findings emerge. First, deep architectures significantly outperform traditional models in within-corpus settings (macro F1: 79.5% on IEMOCAP, 85.5% on RAVDESS), but these gains are corpus-specific. Second, cross-corpus evaluation reveals a 25- 28 percentage-point performance drop, with domain-adversarial training recovering only 4-7 points. Third, class-balanced training and augmentation improve minority emotion recognition (fear/disgust from <0.40 to 0.52–0.58) with a 4.8- 5.4-point macro F1 gain. Fourth, fairness-aware loss reduces gender disparities from 6.5-8.0% to 2.0-2.5% with a modest 1-2% accuracy trade-off.

The results highlight three broader implications: the benchmark illusion (high within-corpus accuracy does not guarantee real-world robustness), the double-edged nature of augmentation, and the observation that fairness is not transferable across corpora.

Limitations include the use of only two acted English-language corpora, binary gender classification, and a single fairness metric (macro F1 disparity). Future work should develop more

diverse corpora, expand fairness analysis to multiple demographic dimensions and metrics, and investigate cross-population generalizability of fairness interventions.

In summary, while deep learning advances SER accuracy on benchmarks, significant challenges persist in cross-corpus generalization and algorithmic fairness. The community must move beyond accuracy-centric benchmarks toward holistic evaluation frameworks that prioritize robustness, fairness, and ecological validity.

Acknowledgements

The author would like to thank the anonymous reviewers for their valuable comments and constructive feedback, which significantly improved the quality of this manuscript.

Conflict of interest

The Author Declares That There Are No Conflicts of Interest Regarding the Publication of This Manuscript.

References

- [1] Abbas Chian, B. J., Sierra-Sosa, D., & Elaraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21 (4), 1249. <https://doi.org/10.3390/s21041249>.
- [2] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7, 117327–117345. <https://doi.org/10.1109/ACCESS.2019.2936124>
- [3] Singh, Y. B., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 492, 245–263. <https://doi.org/10.1016/j.neucom.2022.04.028>
- [4] Hashem, A., Arif, M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. *Speech Communication*, 154, 102974. <https://doi.org/10.1016/j.specom.2023.102974>
- [5] Rathi, T., & Tripathy, M. (2024). Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review. *Speech Communication*, 162,

103102. <https://doi.org/10.1016/j.specom.2024.103102>
- [6] Ülgen Sönmez, Y., & Varol, A. (2024). In-depth investigation of speech emotion recognition studies from past to present—The importance of emotion recognition from speech signal for AI. *Intelligent Systems with Applications*, 22, 200351. <https://doi.org/10.1016/j.iswa.2024.200351>
- [7] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116, 56–76. <https://doi.org/10.1016/j.specom.2019.12.001>
- [8] Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, 102, 102019. <https://doi.org/10.1016/j.inffus.2023.102019>
- [9] Madanian, S., et al. (2023). Speech emotion recognition using machine learning—A systematic review. *Intelligent Systems with Applications*, 20, 200266. <https://doi.org/10.1016/j.iswa.2023.200266>
- [10] Mohmad Dar, G. H., & Delhibabu, R. (2024). Speech databases, speech features, and classifiers in speech emotion recognition: A review. *IEEE Access*, 12, 151122–151152. <https://doi.org/10.1109/ACCESS.2024.3476960>
- [11] Majkowski, A., & Kolodziej, M. (2025). Emotion recognition from speech in a subject-independent approach. *Applied Sciences*, 15 (13), 6958. <https://doi.org/10.3390/app15136958>
- [12] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambika Rajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9, 47795–47814. <https://doi.org/10.1109/ACCESS.2021.3068045>
- [13] Al-Dujaili, M. J., & Ebrahimi-Moghadam, A. (2023). Speech emotion recognition: A comprehensive survey. *Wireless Personal Communications*, 129 (4), 2525–2561. <https://doi.org/10.1007/s11277-023-10244-3>
- [14] Anthony, A. A., & Patil, C. M. (2023). Speech emotion recognition systems: A comprehensive review of different methodologies. *Wireless Personal Communications*, 130 (1), 515–525. <https://doi.org/10.1007/s11277-023-10296-5>
- [15] Atmaja, B. T., Sasou, A., & Akagi, M. (2022). Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication*, 140, 11–28. <https://doi.org/10.1016/j.specom.2022.03.002>
- [16] Mustafa, M. B., Yusof, M. A. M., Don, Z. M., & Malekzadeh, M. (2018). Speech emotion recognition research: An analysis of research focus. *International Journal of Speech Technology*, 21 (1), 137–156. <https://doi.org/10.1007/s10772-018-9493-x>
- [17] O'Shaughnessy, D. (2025). Review of automatic estimation of emotions in speech. *Applied Sciences*, 15 (10), 5731. <https://doi.org/10.3390/app15105731>