IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Robust Speech Emotion Recognition Through Self-Supervised Acoustic-Semantic Fusion and Cross-Attention Under Background Noise

Renuka Bedarkar[1] and Dr. H. B. Jadhav[2]
Student, Department of Computer Engineering[1]
HOD, Department of Computer Engineering[2]
Adsul's Technical Campus, Chas, Ahilyanagar, India

Abstract: *Human speech carries information about both the words being spoken and the emotions behind them. Speech Emotion Recognition (SER) uses artificial intelligence to identify emotions such as happiness, sadness, anger, and neutrality from a person's voice. However, recognizing emotions accurately is difficult when speech contains background noise or when different people express the same emotion in different ways. This paper proposes a speech emotion recognition framework that combines self-supervised speech representations with acoustic features through cross-attention. A pretrained speech encoder is used to extract useful representations from audio, while acoustic features such as pitch, energy, MFCCs, and spectral information provide additional details about vocal expression. The two feature streams are combined through a cross-attention mechanism before the final emotion classification stage. The main aim is to investigate whether this combination can improve emotion recognition under different background-noise conditions. The proposed system will be evaluated on a publicly available emotional speech dataset and compared with acoustic-only, pretrained-representation-only, and conventional feature-fusion methods. Accuracy, unweighted accuracy, macro-F1 score, and class-wise recall will be used for evaluation. The experimental results will determine whether the proposed method offers a meaningful improvement over existing approaches.*

Keywords: Speech Emotion Recognition, Self-Supervised Learning, Cross-Attention, Acoustic Features, Semantic Representations, Noise Robustness

I. INTRODUCTION

Human communication involves more than the words we speak. The way a person speaks can reveal happiness, sadness, anger, excitement, or uncertainty. Changes in pitch, speaking speed, voice intensity, pauses, and word choice all contribute to the emotional meaning of an utterance. Speech Emotion Recognition (SER) is a branch of artificial intelligence that attempts to identify these emotions automatically from speech.

Speech emotion recognition has potential applications in human-computer interaction, virtual assistants, customer-service systems, educational technologies, and other speech-based applications. For example, a conversational system that can recognize frustration may respond differently from one that detects happiness. However, developing a reliable system remains challenging because emotions are expressed differently by different speakers, and background noise can affect important vocal characteristics.

Earlier speech emotion recognition systems often relied on manually extracted features, including Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and spectral characteristics. These features can provide useful information about emotional expression, but they may not capture every relevant pattern in speech.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40501

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

Self-supervised learning has introduced another approach to speech processing. Models such as wav2vec 2.0 learn speech representations from large amounts of audio without requiring emotion labels for their initial pretraining [1]. These representations can then be adapted to tasks such as emotion recognition. However, speech representations and explicit acoustic features may contain complementary information. Two people can say the same sentence while expressing different emotions through their voices. Similarly, the meaning of a sentence can help interpret vocal characteristics that might otherwise be ambiguous.

Cross-attention provides a way for one feature representation to focus on relevant information in another. This paper proposes combining pretrained speech representations and acoustic features through cross-attention, with a particular focus on performance under background noise.

The main objectives of this research are:

- To develop a speech emotion recognition framework that combines pretrained speech representations and acoustic features.
- To investigate the use of cross-attention for integrating the two feature streams.
- To evaluate recognition performance under clean and noisy speech conditions.
- To compare the proposed approach with simpler feature-fusion methods.
- To examine the contribution of each component through controlled experiments.

The main focus is not simply to build another emotion classifier, but to investigate whether combining different types of speech information can make emotion recognition more reliable when recording conditions are less than ideal.

II. EXISTING SYSTEM

A. Overview of Existing Systems

Researchers have explored different ways to identify emotions in speech, ranging from conventional machine-learning methods to advanced deep-learning techniques, pretrained speech models, and attention-based approaches.

Traditional approaches extract acoustic features from speech and use a classifier to predict the emotion. Commonly used features include MFCCs, pitch, energy, and spectral descriptors. Although these methods can be effective in controlled environments, their performance may be affected by speaker differences and background noise.

More recent systems use deep neural networks to learn features directly from audio. Self-supervised speech encoders, including wav2vec 2.0, provide representations learned from large audio collections. Researchers have also explored combining multiple feature streams using attention mechanisms.

For example, Gao et al. investigated the extraction and interaction of acoustic and semantic information for speech emotion recognition in 2024 [2]. Deeb et al. examined cross-attention fusion of acoustic and semantic features using self-supervised learning in 2025 [3]. These studies provide important foundations for the approach proposed in this paper.

B. Existing System Architecture

The following diagram illustrates a generalized architecture found in speech emotion recognition systems. It is not intended to represent the exact architecture of any single published paper.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

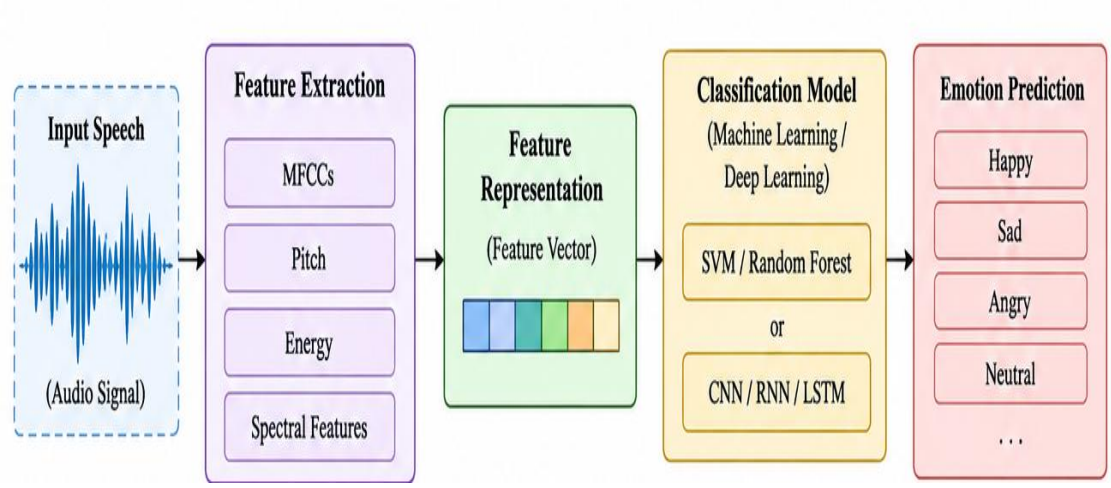


Fig. 1. Existing System Architecture

C. Limitations of Existing Approaches

Although existing approaches have achieved promising results, several issues remain worth investigating.

First, background noise can alter acoustic information and make emotion recognition more difficult. Second, a system that uses only acoustic features may miss useful contextual information, while a pretrained speech representation may not capture every vocal cue required for emotion classification.

Third, simple feature concatenation does not explicitly model which parts of one representation are relevant to another. Attention-based fusion offers one way to investigate these relationships, but it also introduces additional parameters and computational requirements.

Finally, results obtained using different datasets, emotion labels, and evaluation protocols cannot always be compared directly. A consistent evaluation procedure is therefore necessary to determine whether a proposed approach provides a genuine improvement.

III. PROPOSED SYSTEM

A. Overview

This paper proposes a framework called the Noise-Robust Acoustic-Semantic Cross-Attention Network (NR-ASCA). The name is used for the proposed design in this study.

The system combines two types of information. The first comes from a pretrained speech encoder, which produces contextual representations of the input audio. The second comes from explicit acoustic features, which describe characteristics such as pitch, energy, and spectral variation.

Cross-attention is used to allow the two feature streams to exchange information before emotion classification. The model will be evaluated under different background-noise conditions to determine whether this integration improves robustness.

The proposed architecture is a research design rather than a completed, experimentally validated model. Its effectiveness will be determined through the experiments described later in this paper.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

B. Proposed System Architecture

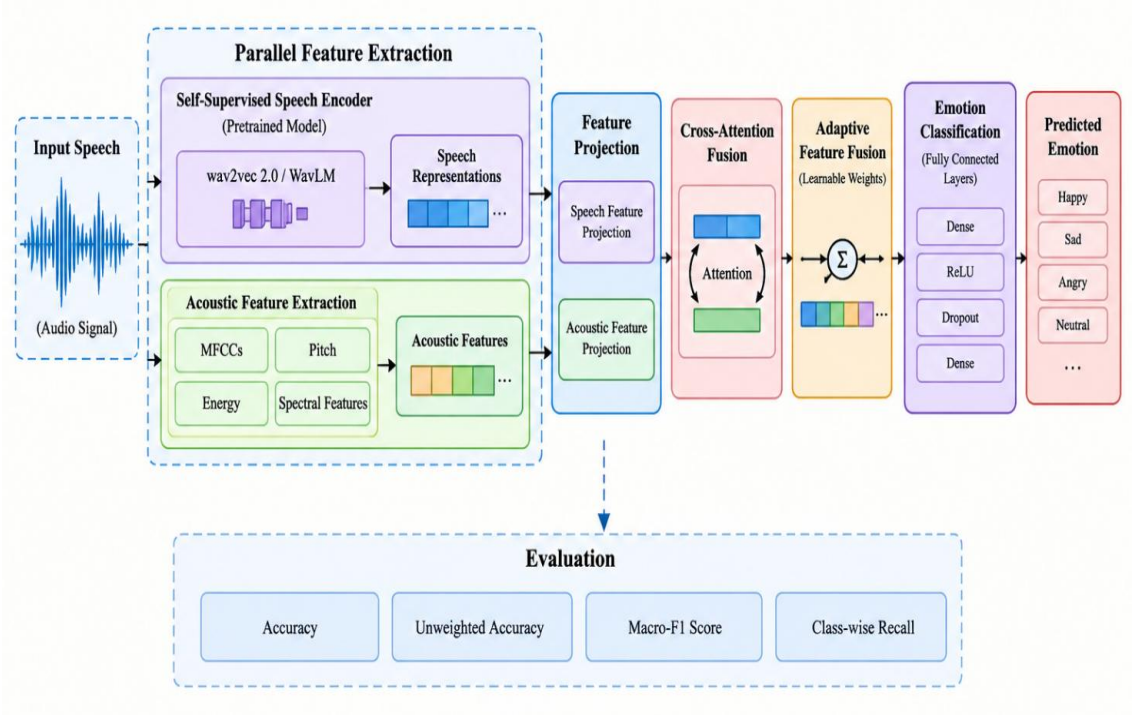


Fig. 2. Proposed System Architecture

C. Audio Preprocessing

The first stage prepares the audio for feature extraction. All recordings will be converted to a consistent sampling rate compatible with the selected pretrained encoder. A sampling rate of 16 kHz is a suitable starting point for a wav2vec 2.0-based implementation.

The preprocessing stage will also ensure consistent audio formatting and handle differences in recording length. Any normalization procedure will be applied consistently across the experimental conditions.

To evaluate robustness, copies of the speech recordings will be corrupted with controlled background noise at predefined signal-to-noise ratios. The clean recordings and noisy versions will be assigned to the same data partition so that copies of the same utterance cannot appear in both training and testing.

D. Self-Supervised Speech Representation

A pretrained speech encoder, such as wav2vec 2.0 or WavLM, will be used to obtain high-level speech representations. Let the input speech waveform be represented by

$$x = (x_1, x_2, \ldots, x_n)$$

The pretrained encoder transforms the waveform into a sequence of feature vectors:

$H\square = E\square(x)$

where $E\square$ represents the pretrained encoder and $H\square$ represents the resulting speech features.

These features can contain information about acoustic patterns, phonetic structure, and linguistic content. They provide a useful starting point for the emotion classification task, although they should not be considered purely semantic representations.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

E. Acoustic Feature Extraction

The acoustic branch extracts features that describe the vocal properties of the input speech. The proposed implementation will investigate MFCCs, log-Mel spectrograms, pitch, and short-time energy.

MFCCs describe aspects of the short-term spectral envelope, while pitch and energy provide information about vocal frequency and intensity. These features may help distinguish emotional expressions that are difficult to identify from linguistic content alone.

The acoustic representation is expressed as:

$H_a = E_a(F(x))$

where $F(x)$ represents the extracted acoustic features and E_a is the acoustic encoder.

The two branches may produce feature sequences with different dimensions and time resolutions. Linear projection layers and appropriate temporal alignment will therefore be used before applying cross-attention.

F. Cross-Attention-Based Feature Fusion

Cross-attention allows one feature sequence to focus on relevant information in another feature sequence. In the proposed system, the speech representation and acoustic representation will interact in both directions.

The scaled dot-product attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V$$

where Q represents queries, K represents keys, V represents values, and d_k is the key dimension.

For speech-to-acoustic attention:

$$C_{s\to a} = \text{Attention}(H_s W_Q, H_a W_K, H_a W_V)$$

For acoustic-to-speech attention:

$$C_{a\to s} = \text{Attention}(H_a U_Q, H_s U_K, H_s U_V)$$

The matrices WQ, WK, WV, UQ, UK, and UV are trainable projection matrices.

These operations allow the system to learn which acoustic features are relevant to the learned speech representation and which speech features are relevant to the acoustic representation. The resulting features are then passed to the fusion stage.

G. Adaptive Feature Fusion

After cross-attention, the two feature sequences will be converted into compatible representations and summarized over time. Let the resulting vectors be z_s and z_a.

An adaptive gate can be used to control their contribution:

$$g = \sigma(W_g[z_s; z_a] + b_g)$$

The fused representation is:

$$z = g \odot z_s + (1 - g) \odot z_a$$

Here, σ denotes the sigmoid function, $[z_s; z_a]$ denotes concatenation, and $\odot$ denotes element-wise multiplication.

The purpose of the gate is to let the model learn how much information to use from each feature stream. The experiments will also compare this design with simpler concatenation to determine whether adaptive fusion is useful.

H. Emotion Classification

The fused representation is passed to a fully connected classification layer. A softmax function converts the resulting scores into class probabilities.

$$p(y \mid x) = \text{softmax}(W_c z + b_c)$$

The predicted emotion is the class with the highest probability:

$$\hat{y} = \text{argmax}_c\, p(y = c \mid x)$$

The emotion categories will depend on the selected dataset and its label mapping. A four-class experiment could use angry, happy, sad, and neutral speech, provided that the chosen dataset supports this configuration.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

The model will be trained using categorical cross-entropy. Class weighting may be considered if the training data contains substantially different numbers of examples for different emotions.

IV. EXPERIMENTAL METHODOLOGY

A. Dataset Selection

The IEMOCAP dataset is a suitable candidate for evaluating the proposed framework because it is widely used in speech emotion recognition research. Access and use must follow the dataset's applicable terms.

A four-class classification protocol can be used, with a clearly documented mapping of the original emotion labels. The same label mapping must be applied to every baseline and the proposed model.

B. Data Partitioning

A speaker-independent or session-independent evaluation protocol will be used to investigate generalization to unseen speakers.

For IEMOCAP, leave-one-session-out evaluation is a suitable option. In each fold, one session will be held out for testing, while the remaining sessions will be used for training and validation.

Hyperparameter selection will be performed without accessing the held-out test data. Clean recordings and their artificially corrupted versions will remain within the same partition to avoid data leakage.

C. Baseline Systems

The proposed system will be compared with the following baselines:

Acoustic-only model using handcrafted acoustic descriptors.

Self-supervised-only model using a pretrained speech encoder.

Early-fusion model using concatenation of the two feature streams.

Single-direction cross-attention model.

Proposed NR-ASCA model using bidirectional cross-attention and adaptive fusion.

These comparisons will help determine whether any performance improvement comes from the pretrained representation, acoustic features, cross-attention, or adaptive fusion.

D. Background-Noise Evaluation

The model will be evaluated on clean speech and speech corrupted with controlled background noise.

The proposed test conditions include clean audio and noisy audio at signal-to-noise ratios of 20 dB, 10 dB, 5 dB, and 0 dB. The specific noise types, noise files, mixing procedure, and random seeds will be documented.

Noise augmentation may be used during training. However, the evaluation recordings and their noise configurations will remain separate from model selection.

E. Evaluation Metrics

The following metrics will be used to evaluate the models.

Weighted accuracy (WA):

$WA = (\text{Number of correct predictions} / \text{Total number of predictions}) \times 100$

This metric measures the proportion of correctly classified test examples.

Unweighted accuracy (UA):

$UA = (1/C) \Sigma [TPc / (TPc + FNc)] \times 100$

where C is the number of emotion classes, TPc is the number of true positives for class c, and FNc is the number of false negatives.

Unweighted accuracy gives equal importance to the recall of each emotion class.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Macro-F1 score: This is calculated by obtaining the F1-score for each class and averaging the scores equally. It is useful when the dataset contains an unequal number of examples for different emotions.
Class-wise recall, confusion matrices, model size, and inference latency may also be reported.

V. RESULTS AND DISCUSSION

The proposed model must be trained and evaluated before numerical results can be reported. Therefore, the tables below define the experiments to be completed rather than presenting fabricated measurements.

Comparison of Model Performance

TABLE I. COMPARISON OF BASELINE AND PROPOSED MODELS

Model	Clean UA	Noisy UA	Macro-F1
Acoustic-only model	To be measured	To be measured	To be measured
Self-supervised-only model	To be measured	To be measured	To be measured
Early-fusion model	To be measured	To be measured	To be measured
Single-direction cross-attention	To be measured	To be measured	To be measured
Proposed NR-ASCA	To be measured	To be measured	To be measured

The results will be used to examine whether combining pretrained speech representations with explicit acoustic features improves recognition performance. Particular attention will be given to the difference between clean and noisy conditions.

B. Evaluation Under Different Noise Levels

TABLE II. NOISE ROBUSTNESS EVALUATION

Test condition	Acoustic-only	Self-supervised-only	Proposed NR-ASCA
Clean speech	To be measured	To be measured	To be measured
20 dB SNR	To be measured	To be measured	To be measured
10 dB SNR	To be measured	To be measured	To be measured
5 dB SNR	To be measured	To be measured	To be measured
0 dB SNR	To be measured	To be measured	To be measured

The results will show how the performance of each model changes as the amount of background noise increases. If the proposed method maintains better performance under noisy conditions, this may indicate that cross-attention helps combine complementary information. However, such a conclusion will depend on the measured results.

C. Ablation Study

An ablation study will be performed to understand the contribution of each component of the proposed system.

TABLE III. ABLATION STUDY CONFIGURATIONS

Reference	Configuration
Without cross-attention	Evaluate the contribution of cross-attention
Without acoustic features	Evaluate the contribution of explicit acoustic descriptors
Without adaptive gating	Evaluate the contribution of adaptive fusion
Without noise augmentation	Evaluate the effect of noise augmentation

Each configuration will be evaluated using the same data partitions and metrics. This will help identify which components contribute to the overall performance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

D. Discussion

The evaluation will focus on three questions. First, does the combination of acoustic features and pretrained speech representations outperform the individual feature streams? Second, does cross-attention provide an advantage over simple concatenation? Third, does the proposed approach retain its performance more effectively when background noise increases?

The study will also consider whether any improvement is consistent across emotion classes. For example, a model may perform well for neutral speech but struggle with emotions that have similar acoustic patterns. A confusion matrix and class-wise recall will help identify these differences.

Computational cost will also be considered because a more complex model may provide only a small improvement over a simpler baseline. The final assessment will therefore consider recognition performance, robustness, and model efficiency together.

VI. CONCLUSION

This paper proposes a noise-robust speech emotion recognition framework that combines self-supervised speech representations and explicit acoustic features through cross-attention. The framework is designed to investigate whether the two feature streams can complement one another and improve emotion classification when speech contains background noise.

The proposed system includes audio preprocessing, pretrained speech representation extraction, acoustic feature extraction, bidirectional cross-attention, adaptive fusion, and emotion classification. A speaker-independent evaluation protocol, controlled noise conditions, baseline comparisons, and ablation experiments have been outlined to assess the approach systematically.

The effectiveness of the proposed framework has not yet been established and will depend on the results of the planned experiments. Future work may investigate cross-dataset generalization, multilingual speech, alternative pretrained encoders, and methods for making attention-based predictions easier to interpret.

ACKNOWLEDGMENT

The authors would like to acknowledge the institution, research supervisor, and computing resources that genuinely support the work. The final acknowledgment should be completed with accurate details before submission.

REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[2] Y. Gao, H. Shi, C. Chu, and T. Kawahara, “Speech Emotion Recognition with Multi-level Acoustic and Semantic Information Extraction and Interaction,” in *Proceedings of Interspeech 2024*, 2024.

[3] B. M. Deeb, A. V. Savchenko, and I. Makarov, “Enhancing Emotion Recognition in Speech Based on Self-Supervised Learning: Cross-Attention Fusion of Acoustic and Semantic Features,” *IEEE Access*, vol. 13, pp. 56283–56295, 2025, doi: 10.1109/ACCESS.2025.3554454.

[4] S. G. Leem, D. Fulford, J. P. Onnela, D. Gard, and C. Busso, “Keep, Delete, or Substitute: Frame Selection Strategy for Noise-Robust Speech Emotion Recognition,” in *Proceedings of Interspeech 2024*, pp. 3734–3738, 2024, doi: 10.21437/Interspeech.2024-1218.

[5] “A Cross-Modal Attention Model with Contextual Enhancements for Speech Emotion Recognition,” *Speech Communication*, vol. 173, article 103285, 2025, doi: 10.1016/j.specom.2025.103285.

[6] “Cross-Feature Fusion Speech Emotion Recognition Based on Attention Mask Residual Network and Wav2vec 2.0,” *Digital Communications and Networks*, vol. 11, no. 5, pp. 1567–1577, 2025, doi: 10.1016/j.dcan.2024.10.007.

[7] “Joint Acoustic–Semantic Alignment Network for Advanced Speech Emotion Recognition,” *Speech Communication*, vol. 181, article 103409, 2026, doi: 10.1016/j.specom.2026.103409.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

[B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in Proceedings of Interspeech 2009, pp. 312–315, 2009.
[8] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011, doi: 10.1016/j.patcog.2010.09.020.
[9] C. Busso et al., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," Language Resources and Evaluation, vol. 42, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.