Leveraging Cross-Attention and Speech Separation for Enhanced Stress Detection in Children's Multi-Speaker Environments

Phie Chyan
Department of Informatics, Atma Jaya Makassar University, Makassar, South Sulawesi, Indonesia
phie_chyan@lecturer.uajm.ac.id (corresponding author)

Heni Gerda Pesau
Department of Psychology, Atma Jaya Makassar University, Makassar, South Sulawesi, Indonesia
heni_gerda@lecturer.uajm.ac.id

Norbertus Tri Suswanto Saptadi
Department of Informatics, Atma Jaya Makassar University, Makassar, South Sulawesi, Indonesia
tri_saptadi@lecturer.uajm.ac.id

Received: 26 January 2026 | Revised: 27 February 2026 | Accepted: 15 March 2026

 DOI: https://doi.org/10.48084/etasr.17765

ABSTRACT

Stress detection in children presents unique challenges due to their limited ability to articulate emotional distress, necessitating automated and multimodal assessment approaches. This study presents a framework for stress recognition in noisy, multi-speaker environments by integrating speech separation with cross-attention-based multimodal fusion. The pipeline first employs a speech separation module to disentangle overlapping voices and suppress environmental noise, enabling reliable extraction of discriminative acoustic features. In parallel, transcripts generated via an Automatic Speech Recognition (ASR) system are transformed into linguistic representations using GloVe embeddings enhanced with TF-IDF weighting. The acoustic and linguistic features are projected into a shared latent space and fused through a cross-attention mechanism to model complementary cross-modal interactions. To address domain variability in children's vocal characteristics, the model is pretrained on adult emotional speech data and subsequently fine-tuned on child-specific samples to facilitate domain adaptation. Experimental results demonstrate that the proposed system achieves an accuracy of 89.5%, significantly outperforming unimodal baselines. Ablation studies further validate the critical contributions of speech separation and dynamic multimodal fusion to overall performance. These findings underscore the potential of the proposed framework as a supportive, non-invasive tool for early stress awareness in child-centered environments.

Keywords-multimodal stress detection; cross-attention mechanism; speech separation; acoustic-linguistic fusion

I. INTRODUCTION

Stress is one of the most prevalent mental health issues that affects individuals of all ages, including children. Stress in children has become a significant issue that has garnered considerable attention in the world of education and child health, particularly due to its increasing prevalence worldwide [1]. Global epidemiological data indicate that approximately 12–13% of children and adolescents worldwide experience anxiety-related disorders, including stress and depression [2]. Children are susceptible to stress due to the various pressures they face in their daily lives, ranging from external factors such as academic demands and social interactions to internal issues, including problems within family life. Prevention and management of stress are important to prevent complications that can have long-term negative impacts on children, such as depression and cognitive disorders that can cause decreased motivation and learning achievement [3].

Detecting stress in children requires a distinct approach compared to adults. Younger children, in particular, often lack the communication skills needed to clearly express the psychological challenges they experience. Consequently, traditional diagnostic methods, such as clinical interviews, tend to be less effective for this population [4, 5]. Nevertheless, recent machine learning-based mental stress assessment frameworks, developed primarily using questionnaire-driven data and adult populations, have demonstrated the potential of data-driven approaches to support more objective stress evaluation [6].

A promising alternative for identifying stress in children involves analyzing vocal characteristics as indicative parameters. The human voice is a complex signal shaped by the interaction of the respiratory system, vocal tract resonance, and the coordination of vocal cord muscles. When an individual experiences stress, activation of the sympathetic nervous system increases muscle tension, including the muscles responsible for phonation [7]. These physiological changes influence various acoustic parameters, which can be measured quantitatively. The vocal-based approach offers notable advantages in terms of data collection efficiency, as it is non-invasive and does not cause discomfort to children. This contrasts with physiological-based methods that require sensor attachments to monitor parameters such as heart rate and blood pressure, which may induce physical discomfort in young participants [8].

Stress detection through voice analysis has been extensively investigated in the literature, as summarized in both established and recent survey studies [9, 10]. The accessibility of vocal data allows its direct capture even in multi-speaker environments. However, achieving accurate stress recognition under such conditions remains highly challenging, particularly due to the difficulties in extracting robust and discriminative acoustic features from stressed speech [11]. Speech overlap, background noise, and the inherent variability of children's vocal patterns further exacerbate these challenges. A multi-speaker approach can potentially enable the child's voice to be captured within more natural activity settings, supporting continuous and near real-time monitoring of psychological states. This may assist psychologists in identifying situational stressors and understanding how stress manifests in daily contexts. In comparison, traditional methods rely on recordings conducted in controlled and noise-free environments, where only the child's voice is monitored. Although such setups are easier to manage, they often fail to capture spontaneous stress responses, as acute stress episodes may have subsided by the time recording is performed, thereby reducing the accuracy of subsequent psychological analysis.

Recent developments in multimodal approaches for speech emotion and stress recognition have shown that combining acoustic features with linguistic or semantic information can improve model accuracy [12, 13]. Additionally, advances in speech separation techniques have made it possible to better isolate individual speaker signals in noisy multi-speaker environments, resulting in cleaner audio signals for subsequent processing [14]. However, effectively integrating acoustic features with linguistic information for stress detection remains challenging due to differences in modality representations, contextual dependencies, and the need for complex modeling strategies. Cross-attention mechanisms offer a promising solution by enabling effective alignment and interaction between acoustic and linguistic features. This allows the model to focus on the most informative elements from both modalities, thereby improving the performance of stress detection systems.

This study proposes a novel framework that leverages speech separation and cross-attention to effectively fuse acoustic and textual features for improved stress detection in children's multi-speaker environments. The framework is specifically designed to operate in noisy, multi-speaker conditions involving children. Its first key contribution lies in the integration of speech separation techniques as a dedicated preprocessing stage to disentangle overlapping speech, enabling more reliable extraction of speaker-specific acoustic and linguistic features for stress analysis. The second major contribution is the adoption of a cross-attention-based fusion mechanism that explicitly models interactions between acoustic and textual modalities, allowing dynamic alignment of multimodal representations and leading to improved stress detection performance.

II. METHODOLOGY

This section delineates the methodological framework underpinning the proposed multimodal stress detection system. The proposed architecture integrates speech separation, feature extraction, domain adaptation, and cross-attention mechanisms to facilitate the fusion of acoustic and linguistic cues for stress recognition in children. The subsequent subsections detail the data acquisition and preprocessing protocols, the design of individual modules, the training and fine-tuning strategies, and the evaluation metrics employed.

A. Overview of The Proposed Framework

This study proposes a multimodal stress detection model framework based on the acoustic and linguistic processing of voice data in a multi-speaker environment, combining speech-separation techniques and cross-attention mechanisms. This model is designed to capture direct sound from the environment where children are active, which is typically a multi-speaker environment. Furthermore, through the speech separation process, the recorded sound is isolated individually to produce segregated sounds from each child by reducing interference and overlap between sounds. This step ensures that the extracted acoustic features are clean and represent the original signal of each speaker.

The segregated speech signal for each speaker then undergoes an acoustic feature extraction stage, where multiple acoustic representations such as Mel-spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and other spectral features are derived to characterize the signal. These features capture the dynamic properties of speech that may serve as indicators of stress. In parallel, the proposed framework incorporates linguistic analysis by processing text transcripts generated through an Automatic Speech Recognition (ASR) system using word embedding techniques. The resulting linguistic features capture semantic nuances and contextual cues that can indicate stress through verbal expressions. Both the acoustic and linguistic features are then projected into a shared feature space with a consistent dimensionality. This enables the integration of prosodic information, including intonation, energy, and vocal dynamics, with linguistic representation in a complementary manner.

The dynamic integration of audio modalities with text is performed through a cross-attention mechanism. Using audio features as queries and linguistic features as keys and values (or vice versa), the cross-attention module learns to align and emphasize the most relevant aspects of both modalities. This

fusion approach aims to improve accuracy by focusing attention on important cues that may be weak or ambiguous if viewed separately in only a single modality. This approach enables the model to adapt dynamically by adjusting for variations in acoustic quality or linguistic expression, particularly when working with diverse data. By drawing on both acoustic and linguistic information, this approach allows the model to better recognize signs of stress in a more holistic way. Figure 1 provides a visual overview of the proposed framework.

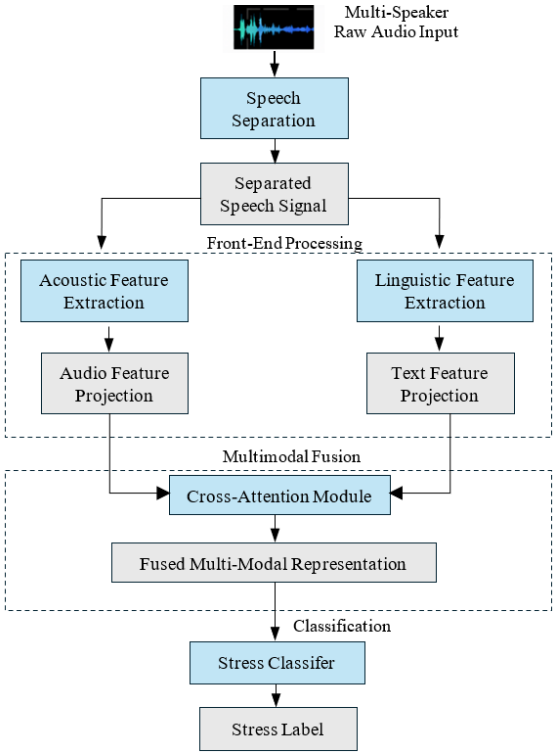

Fig. 1. Proposed framework for multimodal stress detection system.

B. Data Acquisition and Preprocessing

The data collection and preprocessing procedures were designed to ensure the integrity and consistency of both acoustic and linguistic modalities before feature extraction, thereby optimizing input quality for subsequent analysis.

The speech separation component was trained under multi-speaker conditions using the LibriMix corpus [15], a benchmark dataset created by combining clean speech from LibriSpeech with environmental noise from the WHAM! dataset [16]. This corpus is designed to approximate realistic acoustic environments, including conditions similar to classrooms and public areas. It includes mixtures containing two to five simultaneous speakers (LibriMix 2 through LibriMix 5), with each variant offering approximately 20 hours of training data and 10 hours each for validation and testing. The model architecture was configured according to the number of speakers in each mixture, resulting in output channels ranging from two (LibriMix 2) to five (LibriMix 5) to enable robust separation of individual speaker streams.

The cross-attention mechanism within the multimodal framework was trained using the IEMOCAP dataset [17], which provides high-quality audio recordings aligned with textual transcripts. Although originally annotated with multiple emotional categories, six representative classes—angry, excited, happy, sad, frustrated, and neutral—were selected to maintain balanced class distributions. These categories were subsequently mapped to a binary stress taxonomy, in which negative affective states were grouped as "Stress" and positive or neutral states were labeled as "Non-Stress," yielding a total of 3000 samples evenly divided between the two classes, which served as the primary pretraining corpus before integration with the child-specific dataset.

This binary categorization is consistent with dimensional affect theory, particularly the circumplex model of affect [18], which characterizes emotional states along valence and arousal dimensions. Emotions such as anger and frustration are positioned within the negative valence and elevated arousal region, conditions commonly associated with stress-related psychophysiological activation. Sadness, while generally lower in arousal, remains within the negative valence domain and is frequently linked to stress exposure and adverse appraisal states. In contrast, positive emotions such as happiness and excitement occupy the positive valence region, and neutral states correspond to baseline affective conditions. Therefore, the adopted mapping represents an operational stress-oriented grouping grounded in established affective theory rather than a direct equivalence between discrete emotions and clinical stress, thereby providing a theoretically grounded abstraction for computational stress modeling in multimodal learning frameworks.

Domain-specific variability in child speech was addressed by augmenting the corpus with a dedicated child speech dataset comprising recordings from 20 elementary school students aged 6 to 8 years. Participation was voluntary, and written informed consent was obtained from the parents or legal guardians of all participating children prior to data collection. All recorded data were anonymized and used exclusively for research purposes. The participants were selected from a single elementary school based on age homogeneity and availability during scheduled classroom activities, ensuring consistency in developmental stage and recording conditions.

Data collection was conducted in two sessions, each involving 10 children in a classroom environment supervised by two licensed child psychologists to ensure ethical compliance and procedural validity. To naturally elicit stress responses, a child-adapted Trier Social Stress Test (TSST) protocol [19] was designed, implemented, and supervised by the psychologists. The activities simulated common academic stressors, including solving time-constrained arithmetic tasks and delivering short storytelling presentations in front of peers.

During the sessions, the psychologists observed each child's facial expressions, gestures, and verbal responses, followed by brief individual interviews with voice recordings. Stress annotation was performed using a predefined stress rating scale based on both behavioral observations and interview responses. To reduce annotation bias, the two psychologists independently evaluated each child's stress condition, and disagreements were

resolved through discussion until consensus was achieved. Each child produced 10 voice samples of 1 to 2 seconds duration, resulting in 200 child-specific speech segments prior to integration with the larger corpus.

After integration with the broader training dataset, the combined corpus consisted of 3200 labeled samples, including 1583 stress instances and 1617 non-stress instances. The integrated dataset was subsequently used to fine-tune the pretrained multimodal model, enabling adaptation to child-specific speech characteristics while preserving the generalized representations learned from the larger corpus. To ensure reliable evaluation and prevent speaker memorization effects, a speaker-independent protocol was adopted, where recordings from each child were assigned exclusively to either the training or testing set.

C. Model Architecture

The proposed multimodal framework integrates speech separation, acoustic and linguistic feature extraction, and a cross-attention fusion mechanism to facilitate dynamic inter-modal interactions for accurate stress detection.

1) Speech Separation

As shown in Figure 2, the recorded multi-speaker audio is first preprocessed to suppress background noise and remove silent segments. The input waveform $x \in \mathbb{R}^T$ is then encoded using a one-dimensional convolutional layer with kernel size L, stride $L/2$, and ReLU activation, producing a latent representation $z \in \mathbb{R}^{N \times T'}$, where $T' = 2\mathrm{T}/L - 1$. The latent features are segmented into overlapping frames and rearranged into a three-dimensional tensor $\mathrm{v} \in \mathbb{R}^{N \times K \times R}$, which is forwarded to a dual-path RNN-based separation network [20].

The separation network consists of multiple MULCAT blocks, each formed by a pair of RNN blocks operating alternately along the spatial and temporal dimensions. Within each block, the input tensor v is processed by two parallel bidirectional LSTM sub-networks, whose outputs are combined via element-wise multiplication and concatenated with the original input. A learnable linear projection maps the concatenated features to the output of the RNN block:

$$B_i(\mathrm{v}) = P_i([M_i^1(\mathrm{v}) \odot M_i^2(\mathrm{v}), \mathrm{v}]) \quad (1)$$

After separation, the resulting representations are passed through parallel decoders implemented as 1×1 convolutional layers with shared PReLU activation. Add-and-overlap operations are then applied to reconstruct continuous time-domain signals for each speaker. Since the number of speakers may vary, the model is initialized with a maximum of five output channels, and silent channels are iteratively removed until all remaining outputs correspond to active speech sources.

The model is trained using the ADAM optimizer with a batch size of 2 and comprises six MULCAT blocks, with each LSTM layer containing 128 hidden units. Separation performance is evaluated using the Scale-Invariant Signal-to-Noise Ratio improvement (SI-SNRi) metric.

2) Feature Extraction

Systematic data augmentation was applied to the combined audio corpus comprising the IEMOCAP dataset and a curated child speech dataset to improve generalization under diverse real-world recording conditions. Each original utterance was augmented into two synthetic variants: one with controlled background noise and another with pitch modulation to simulate natural acoustic variability. This process expanded the dataset to 9200 samples and improved robustness against environmental noise and speaker-dependent variations.

Raw speech waveforms are inherently irregular and high-dimensional, which limits their suitability for direct analysis by machine learning models. Discriminative acoustic features were therefore extracted to reduce dimensionality while preserving stress-relevant information. Time-domain features, including Zero Crossing Rate (ZCR) and Root Mean Square (RMS) energy, were computed to capture amplitude dynamics and energy fluctuations associated with emotional states. Frequency-domain representations, namely Mel-spectrograms, were also employed to characterize spectral energy distributions on the perceptually motivated Mel-scale and to facilitate the identification of stress-induced spectral shifts. In addition, MFCCs were extracted to encode combined spectral and temporal characteristics, enabling robust modeling of vocal traits such as intonation patterns and vocal tension.

Linguistic features were derived from textual transcriptions provided by the IEMOCAP corpus, which were reclassified into the Stress and Non-Stress categories. The Natural Language Processing (NLP) pipeline begins with text normalization, followed by tokenization and the removal of non-informative elements such as stop words and punctuation. Lemmatization is subsequently applied to consolidate morphological variants into their canonical forms, yielding a standardized textual representation suitable for feature modeling.

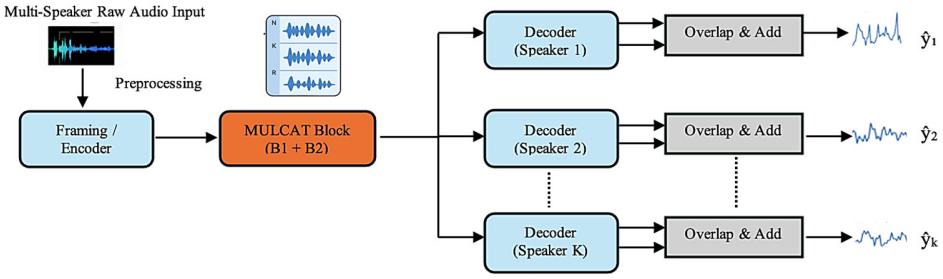

Fig. 2. Speech separation model.

Two complementary linguistic feature extraction strategies were employed. Pretrained GloVe embeddings were used to generate dense semantic representations that capture contextual relationships learned from large-scale corpora [21]. Term Frequency-Inverse Document Frequency (TF-IDF) weighting was additionally applied to emphasize lexically salient terms based on their distribution across documents. The combination of GloVe embeddings and TF-IDF scores produces a hybrid linguistic representation that integrates semantic richness with statistical relevance. These linguistic features are subsequently fused with acoustic representations through a cross-attention mechanism within the proposed multimodal architecture.

3) Cross-Attention Module

Once the acoustic and linguistic features have been extracted, both need to be aligned into a common feature space to allow effective integration between the two modalities. To achieve this, projection layers are used to transform the feature vectors from each modality into a uniform dimensional form before they are passed to the cross-attention module. For clarity, let $f_a' \in \mathbb{R}^{d_a}$ represent the acoustic feature vector and $f_l' \in \mathbb{R}^{d_l}$ denote the linguistic feature vector. To make these vectors compatible, they are projected into a common feature space of dimension d using learned linear transformations. The projection process is defined as

$$f_a' = w_a f_a + b_a, \quad f_l' = w_l f_l + b_l \tag{2}$$

where $w_a \in \mathbb{R}^{d \times d_a}$ and $w_l \in \mathbb{R}^{d \times d_l}$ are the weight matrices, while b_a and b_l are the corresponding bias terms. This transformation ensures that both modalities share a consistent representation in the same latent space, thereby reducing discrepancies in feature scale and distribution. By mapping acoustic and linguistic features into the shared feature space, the model can more naturally integrate the two modalities together through cross-attention. This shared space makes it easier for the system to pick up on how the two types of information relate to each other, which in turn supports more reliable and accurate stress detection.

After the acoustic and linguistic features have been aligned into a common feature space via the projection layers, the next step is to effectively integrate these modalities using a cross-attention mechanism. The cross-attention mechanism brings together the acoustic and linguistic features in a way that allows the model to recognize how they influence one another. This interaction helps the system better identify signs of stress by making use of both acoustic cues and the linguistic context. The cross-attention mechanism treats one modality's projected features as queries and the other's as keys and values. In particular, let $Q \in \mathbb{R}^{n \times d}$ be the query matrix, which comes from the projected acoustic features. The key and value matrices K and V, each in $\in \mathbb{R}^{m \times d}$ are taken from the projected linguistic features, where d_k is the dimensionality of the key vectors, acting as a scaling factor to stabilize gradients during training. This mechanism allows the model to adjust the weighting of linguistic features dynamically, depending on how relevant they are to the surrounding acoustic context. After processing through the cross-attention module, the model produces a fused multimodal representation that reflects information from both the acoustic and linguistic inputs.

The combined output is subsequently passed to the final classifier to determine whether the input corresponds to a stress condition. The cross-attention process is carried out as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

III. RESULTS AND DISCUSSION

This section presents the empirical results and discussion of the proposed multimodal stress detection framework. Performance is evaluated across the speech separation, feature extraction, and classification stages, with additional ablation and component-level analyses conducted to assess the contribution of individual modules.

A. Quantitative Evaluation

Before initiating feature extraction and classification, the performance of the speech separation module was rigorously assessed using the LibriMix benchmark. Evaluation employed the SI-SNRi metric, which is a widely recognized measure of separation fidelity. Four distinct models were trained, each optimized for a specific speaker cardinality (two, three, four, and five speakers), and subsequently tested on mixtures containing between two and five simultaneous speakers. For each evaluation instance, a channel-activity detection algorithm dynamically selected the model whose separation capacity most closely aligned with the estimated number of speakers in the input mixture. Table I summarizes the SI-SNRi scores obtained from these experiments.

TABLE I. SI-SNRI (DB) FOR MODEL TRAINED AND TESTED ON DIFFERENT SPEAKER COUNTS

Used model	**# of Speakers in sample**			
	2	**3**	**4**	**5**
2-speakers	17.5	-	-	-
3-speakers	12.3	13.6	-	-
4-speakers	9.6	10.8	9.7	-
5-speakers	6.5	8.7	8.4	7.5

As shown in Table I, optimal separation performance is obtained when the model's speaker count matches the number of speakers in the mixture, with a two-speaker model achieving an SI-SNRi of 17.5 dB on two-speaker mixtures. When the model capacity exceeds the actual number of speakers, redundant output channels generate silent signals, resulting in degraded SI-SNRi. The proposed dynamic selection algorithm mitigates this issue by selecting the most appropriate model for each unknown mixture, thereby maximizing separation quality.

Following separation, acoustic features (MFCCs and log-Mel spectrograms) and linguistic features (GloVe embeddings and TF-IDF) are extracted and fused via a cross-attention mechanism prior to classification. Evaluation was conducted on a dataset combining adult speech from IEMOCAP and the collected child speech dataset, with fine-tuning performed using 200 labeled child utterances to improve domain generalization. The proposed multimodal model was compared against acoustic-only and text-only baselines, and classification performance was assessed using Accuracy, F1-score, and ROC-AUC on the child speech test set, as shown in Table II.

TABLE II. PERFORMANCE COMPARISON OF UNIMODAL AND MULTIMODAL STRESS DETECTION MODELS

Model	Accuracy (%)	F1-score	ROC-AUC
Acoustic-only (CNN)	78.4	0.78	0.85
Text-only (GloVe +TF-IDF)	80.1	0.80	0.86
Proposed (Cross-attention)	89.5	0.86	0.91

TABLE III. STRESS DETECTION PERFORMANCE OF THE PROPOSED MODEL UNDER DIFFERENT NUMBERS OF CONCURRENT SPEAKERS

Used model	Accuracy (%)	F1-score	ROC-AUC
2-speakers	89.5	0.86	0.91
3-speakers	85.3	0.83	0.88
4-speakers	78.5	0.77	0.80
5-speakers	69.7	0.67	0.71

The proposed cross-attention–based multimodal model outperformed both unimodal baselines, achieving an accuracy of 89.5%, an F1-score of 0.86, and a ROC-AUC of 0.91, indicating strong discriminative capability. This performance gain is primarily attributed to two factors: high-quality speech separation, which provides cleaner acoustic inputs and facilitates reliable feature extraction, and the cross-attention mechanism, which dynamically aligns and weights complementary acoustic and linguistic information, enabling more effective multimodal integration than static fusion approaches. The integration of dynamic speaker-count model selection with robust multimodal fusion demonstrates strong potential for stress detection in realistic, noisy, multi-speaker environments involving children.

Robustness under increasing conversational complexity was further examined by analyzing stress classification performance as the number of overlapping speakers increased. As reported in Table III, performance gradually declines with additional speakers, with accuracy decreasing from 89.5% in two-speaker mixtures to 85.3%, 78.5%, and 69.7% in three-, four-, and five-speaker conditions, respectively. This trend is consistent with the reduction in separation quality observed in Table I, where SI-SNRi decreases as the number of speakers increases. Despite this degradation, the proposed model maintains practically useful accuracy even under highly overlapped conditions. The gradual performance decline indicates graceful degradation, suggesting that the combination of speech separation and cross-attention-based fusion enhances robustness. When acoustic cues are compromised by residual inter-speaker interference, complementary linguistic information can still be effectively leveraged to support classification.

A joint analysis of SI-SNRi values (Table I) and classification results (Table III) reveals a strong correlation between separation quality and downstream stress detection performance, confirming that the separation module plays a critical role in preserving stress-relevant cues rather than functioning solely as a preprocessing stage. In addition, training convergence and evaluation across different speaker configurations exhibit stable behavior with minimal performance variability, indicating that the reported results are reliable and reproducible under the evaluated experimental conditions.

B. Ablation Study and Component Analysis

A series of ablation experiments was conducted to quantify the contribution of individual components within the proposed multimodal stress detection framework. Each experiment systematically removed or modified a key module, and classification performance was evaluated on the child speech test set. Table IV reports the comparative results for configurations involving the removal of the speech separation module, the exclusion of the projection layer, the replacement of the cross-attention mechanism with simple concatenation, the omission of data augmentation, and alternative fine-tuning strategies based on selective parameter freezing.

The most substantial performance degradation occurred when the speech separation module was removed, with accuracy decreasing to 76.4% and the F1-score to 0.724, highlighting its critical role in isolating speaker-specific information under multi-speaker conditions. Replacing cross-attention with static concatenation also resulted in a notable performance drop, yielding an F1-score of 0.781, although the impact was less severe than that of removing speech separation. More moderate declines were observed when eliminating the projection layer or data augmentation, with F1-scores of 0.751 and 0.798, respectively, indicating their supportive contributions to feature alignment and model generalization. These results demonstrate the hierarchical importance of system components, with speech separation and dynamic multimodal fusion emerging as key factors for robust stress detection in complex acoustic environments.

TABLE IV. ABLATION STUDY RESULTS

Model variant	Accuracy (%)	F1-score
Full Model	89.5	0.86
Without Speech Separation	76.4	0.72
Without Cross-Attention	81.7	0.78
Without Projection Layer	79.1	0.75
Without Data Augmentation	83.0	0.79

Figure 3 presents a normalized heatmap depicting relative declines performance across ablation configurations, enabling direct comparison of component-level contributions. Darker regions indicate greater performance degradation and highlight the criticality of the removed modules. The ablation results reveal a clear hierarchy among architectural components within the proposed framework. Removing the speech separation module produces the most pronounced reduction in predictive performance, confirming its essential role in disentangling overlapping speech streams for reliable stress detection in multi-speaker child-focused scenarios. Excluding the cross-attention mechanism results in a more moderate performance decline, suggesting that clean, speaker-specific acoustic representations exert a stronger influence on classification accuracy than fusion complexity under high-overlap conditions. The findings indicate that front-end speech separation functions as a foundational determinant of system effectiveness, while advanced multimodal fusion primarily serves as a refinement stage once adequate signal isolation has been achieved. This hierarchical dependency underscores the importance of robust preprocessing strategies for stress recognition in realistic, noisy, multi-speaker environments.

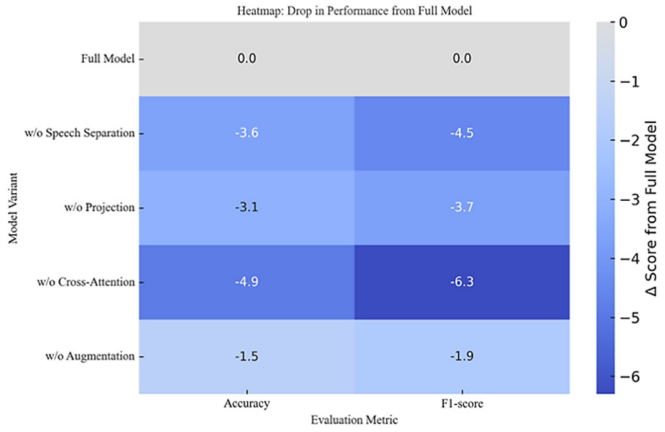

Fig. 3. Heatmap illustrating the performance degradation relative to the full model (Δ Accuracy and Δ F1-score), represented on a normalized scale. Darker shades indicate greater decreases in performance.

C. Extended Discussion and Practical Implications

To situate the proposed framework within the broader landscape of speech-based stress and emotion recognition, this section provides a comparative analysis of methodological approaches reported in prior studies. Historically, much of the research in this domain has focused on adult emotional speech collected under controlled laboratory conditions, typically involving mono-speaker and noise-free recordings. While such settings facilitate feature extraction and model training, they may not fully capture the complexity of real-world environments where stress frequently co-occurs with background noise, overlapping speech, and spontaneous child vocalizations. Under these conditions, conventional acoustic models and unimodal pipelines may experience reduced discriminative robustness, particularly when prosodic cues are affected by multi-speaker interference or the intrinsic variability of children's pitch and articulation patterns.

Over the past decade, acoustic-based stress and emotion detection has evolved from traditional statistical classifiers and support vector machines to advanced deep learning architectures, including convolutional and recurrent neural networks. For example, in [22], speech-based prediction of physiological stress markers was investigated under controlled experimental protocols using adult participants. Similarly, in [23], a 2D-CNN architecture was employed on MFCC representations to enhance spectral feature learning, but the evaluation was conducted on conventional segmented utterances without explicit modeling of overlapping speech. More recent works, such as [24, 25], introduced multimodal or deep acoustic frameworks that improved recognition performance by integrating multiple feature streams or unscripted speech data. Nevertheless, these studies primarily focused on adult populations and did not explicitly address multi-speaker scenarios, which may limit their applicability to complex child-centered acoustic environments.

Table V summarizes the methodological distinctions between these representative studies and the proposed framework. Due to differences in datasets, labeling protocols, and evaluation settings across prior research, direct numerical comparison would not provide a fair assessment. Consequently, the comparison emphasizes dataset characteristics, modality integration, and robustness to environmental complexity rather than absolute performance metrics. In contrast to prior approaches, the proposed model integrates speech separation with cross-attention-based multimodal fusion and is explicitly evaluated under multi-speaker conditions involving child speech, thereby extending stress detection toward more ecologically representative settings.

Although speaker-independent evaluation was enforced to ensure that child speakers in the test set were not present in the training partition, certain factors may influence the broader generalizability of the reported findings. The child speech dataset was collected from a single elementary school and limited to the 6–8 age range, which may not fully capture wider demographic, linguistic, or socio-cultural variability. Although the integration of publicly available corpora and domain-specific child recordings enhances ecological relevance, further validation across multiple institutions, age groups, and diverse acoustic settings would strengthen the robustness and external validity of the proposed framework.

The integration of speech separation and cross-attention-based multimodal fusion introduces additional computational overhead compared to conventional unimodal stress detection systems. In particular, the speech separation module operates on time-domain mixtures and requires multiple output streams proportional to the number of concurrent speakers, while the cross-attention mechanism performs joint feature alignment across acoustic and linguistic embeddings. Despite this increased architectural complexity, both components are implemented using modern deep learning frameworks that are optimized for parallel processing on contemporary GPU hardware.

TABLE V. COMPARISON WITH REPRESENTATIVE PREVIOUS STUDIES AND THE PROPOSED MODEL

Study	Dataset	Modality	Multi-speaker handling	Key remarks
[22]	Adult stress corpora (Controlled)	Audio / multimodal physiologic	Not explicitly addressed	Regression of physiological stress markers from speech under controlled conditions.
[23]	Adult emotion corpus (Controlled)	Acoustic (2D-CNN)	Not explicitly addressed	CNN approach on conventional datasets without multi-speaker focus.
[24]	IEMOCAP (adult)	Multimodal (Acoustic + Text)	Not explicitly addressed	Multimodal emotion recognition; not focused on overlapping/noisy speech.
[25]	Adult (Unscripted) speech dataset	Acoustic (MFCC+CNN)	Not explicitly addressed	MFCC + CNN for multiclass stress labels on unscripted recordings.
Proposed	IEMOCAP + Child Corpus	Multimodal (Acoustic + Text)	Yes (2-5 Speakers)	Integrates speech separation with cross-attention–based multimodal fusion and evaluated under multi-speaker child-centered acoustic conditions.

From a deployment perspective, the proposed framework is well-suited for controlled environments such as classroom monitoring systems or institutional screening tools, where inference can be performed on dedicated edge devices or centralized servers. Although strict real-time processing under high speaker density may require hardware acceleration, near-real-time performance is feasible under typical classroom conditions with a limited number of concurrent speakers. Therefore, the computational requirements represent a practical trade-off for achieving robustness in multi-speaker and child-centered acoustic environments.

Beyond methodological and computational considerations, the proposed framework has meaningful practical implications for early stress identification in child-centered environments. Early detection of elevated stress responses may help teachers, school counselors, and caregivers recognize children who experience sustained emotional strain during academic or social activities. Rather than serving as a diagnostic instrument, the system is intended as a supportive screening tool that can complement existing psychological observation practices. By leveraging non-invasive vocal analysis, the framework enables continuous and context-aware monitoring without disrupting natural classroom interactions. This capability may contribute to timely interventions, improved emotional support strategies, and the promotion of healthier learning environments.

IV. CONCLUSION

This study presents a multimodal stress detection framework tailored to children's speech in multi-speaker environments by integrating speech separation with cross-attention-based fusion of acoustic and linguistic representations. By explicitly addressing speaker overlap and heterogeneous feature alignment, the proposed architecture advances stress recognition toward more ecologically realistic classroom scenarios. The experimental results demonstrate that the model achieves superior performance over unimodal baselines, reaching an accuracy of 89.5% while maintaining robustness as the number of concurrent speakers increases.

The ablation analysis further confirms the complementary contribution of speech separation and dynamic multimodal fusion in enhancing discriminative reliability. Despite this study being limited by dataset scale and demographic scope, the findings provide a strong foundation for extending stress detection to broader child-centered contexts. Future research will focus on expanding multi-site data collection, improving computational efficiency for scalable deployment, exploring more advanced contextual language models for richer linguistic representation, and incorporating additional modalities to further strengthen generalizability and practical applicability in real-world pediatric environments. Ultimately, this line of research may contribute to the development of supportive, non-invasive technologies for early stress awareness in child-centered environments.

DECLARATION OF COMPETING INTERESTS

The authors declare that there are no competing interests.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the Fundamental Research Grant (PFR), under Contract No. 656/LL9/PL/2025 and 038/LPPM/UAJM/PFR/VI/2025.

DATA AVAILABILITY

The dataset generated during this study consists of recorded child speech data and is not publicly available due to privacy and ethical considerations. However, it is available from the corresponding author upon reasonable request. Publicly available datasets, including LibriMix [15] and IEMOCAP [17], were also used in this study.

AI USE AND DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used Microsoft Copilot to assist in translation, paraphrasing, and language editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of this publication.

REFERENCES

[1] A. Sood, D. Sharma, M. Sharma, and R. Dey, "Prevalence and repercussions of stress and mental health issues on primary and middle school students: a bibliometric analysis," *Frontiers in Psychiatry*, vol. 15, Sept. 2024, Art. no. 1369605, https://doi.org/10.3389/fpsyt.2024.1369605.

[2] M. Solmi *et al.*, "Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies," *Molecular Psychiatry*, vol. 27, no. 1, pp. 281–295, Jan. 2022, https://doi.org/10.1038/s41380-021-01161-7.

[3] P. Morgado and J. J. Cerqueira, "Editorial: The Impact of Stress on Cognition and Motivation," *Frontiers in Behavioral Neuroscience*, vol. 12, Dec. 2018, Art. no. 326, https://doi.org/10.3389/fnbeh.2018.00326.

[4] C. A. Kearney, A. Freeman, and V. Bacon, "Structured and semistructured interviews for children," in *Handbook of Psychological Assessment*, Elsevier, 2019, pp. 337–353.

[5] E. Macleod, J. Woolford, L. Hobbs, J. Gross, H. Hayne, and T. Patterson, "Interviews with children about their mental health problems: The congruence and validity of information that children report," *Clinical Child Psychology and Psychiatry*, vol. 22, no. 2, pp. 229–244, Apr. 2017, https://doi.org/10.1177/1359104516653642.

[6] S. S. Shinde and A. S. Ghotkar, "From Questionnaires to Actionable Insights: Machine Learning for Mental Stress Detection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 29240–29250, Dec. 2025, https://doi.org/10.48084/etasr.13513.

[7] M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn, "Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance," *Frontiers in Psychology*, vol. 9, Nov. 2018, Art. no. 1994, https://doi.org/10.3389/fpsyg.2018.01994.

[8] Y. Choi, Y. M. Jeon, L. Wang, and K. Kim, "A Biological Signal-Based Stress Monitoring Framework for Children Using Wearable Devices," *Sensors*, vol. 17, no. 9, Aug. 2017, Art. no. 1936, https://doi.org/10.3390/s17091936.

[9] G. M. Slavich, S. Taylor, and R. W. Picard, "Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations," *Stress*, vol. 22, no. 4, pp. 408–413, July 2019, https://doi.org/10.1080/10253890.2019.1584180.

[10] L. Lavanya and N. Vasavya, "Stress Recognition in Speech – A Survey of The State of The Art," *Journal of Neonatal Surgery*, vol. 14, no. 5S, pp. 793–798, Mar. 2025, https://doi.org/10.52783/jns.v14.2153.

[11] P. Tiwari and A. D. Darji, "Pertinent feature selection techniques for automatic emotion recognition in stressed speech," *International Journal*

of Speech Technology, vol. 25, no. 2, pp. 511–526, June 2022, https://doi.org/10.1007/s10772-022-09978-5.

[12] P. Lu, L. Tsao, and L. Ma, "Daily stress detection from real-life speeches using acoustic and semantic information," *Ergonomics*, vol. 68, no. 10, pp. 1694–1717, Oct. 2025, https://doi.org/10.1080/00140139.2024.2430370.

[13] P. Chyan, A. Achmad, I. Nurtanio, and I. S. Areni, "Multi-Stage Approach for Stress Detection Using Speech Lexical Analysis," in *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Purwokerto, Indonesia, Aug. 2023, pp. 157–162, https://doi.org/10.1109/ICITISEE58992.2023.10404529.

[14] M. Liu and Y. Zhang, "A Review of Speech Separation Focusing on TasNet, Conv-TasNet, and DPRNN," in *2025 5th International Conference on Sensors and Information Technology*, Mar. 2025, pp. 880–885, https://doi.org/10.1109/ICSI64877.2025.11009960.

[15] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation." arXiv, 2020, https://doi.org/10.48550/ARXIV.2005.11262.

[16] G. Wichern *et al.*, "WHAM!: Extending Speech Separation to Noisy Environments," in *Interspeech 2019*, Sept. 2019, pp. 1368–1372, https://doi.org/10.21437/Interspeech.2019-2821.

[17] C. Busso *et al.*, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008, https://doi.org/10.1007/s10579-008-9076-6.

[18] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980, https://doi.org/10.1037/h0077714.

[19] N. F. Narvaez Linares, V. Charron, A. J. Ouimet, P. R. Labelle, and H. Plamondon, "A systematic review of the Trier Social Stress Test methodology: Issues in promoting study comparison and replicable research," *Neurobiology of Stress*, vol. 13, Nov. 2020, Art. no. 100235, https://doi.org/10.1016/j.ynstr.2020.100235.

[20] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 46–50, https://doi.org/10.1109/ICASSP40776.2020.9054266.

[21] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, https://doi.org/10.3115/v1/D14-1162.

[22] A. Baird *et al.*, "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress," *Frontiers in Computer Science*, vol. 3, Dec. 2021, Art. no. 750284, https://doi.org/10.3389/fcomp.2021.750284.

[23] Y. Eom and J. Bang, "Speech Emotion Recognition Using 2D-CNN with Mel-Frequency Cepstrum Coefficients," *Journal of Information and Communication Convergence Engineering*, vol. 19, no. 3, pp. 148–154, Sept. 2021, https://doi.org/10.6109/jicce.2021.19.3.148.

[24] S. W. Byun, J. H. Kim, and S. P. Lee, "Multi-Modal Emotion Recognition Using Speech Features and Text-Embedding," *Applied Sciences*, vol. 11, no. 17, Aug. 2021, Art. no. 7967, https://doi.org/10.3390/app11177967.

[25] N. A. Zainal, A. L. Asnawi, A. Z. Jusoh, S. N. Ibrahim, and H. A. Mohd. Ramli, "Integration of MFCCs and CNN for Multi-Class Stress Speech Classification on Unscripted Dataset," *IIUM Engineering Journal*, vol. 25, no. 2, pp. 381–395, July 2024, https://doi.org/10.31436/iiumej.v25i2.3207.