

Spatio-Spectral Hypergraph Attention Network for Multi-Speaker Emotion Separation in High Noise Acoustic Environments

Abdul Ashfaque Basha^{1*}, K. Venkata Prasad²

¹Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh - 522502, India

² Professor, Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh - 522502, India

Corresponding author: abdulashfaquebasha@gmail.com

Abstract

Multi-speaker speech emotion recognition becomes challenging when speech signals overlap and are corrupted by acoustic noise, as interference between speakers can obscure emotion-related acoustic characteristics. Existing approaches often model individual speech representations or pairwise relationships but have limited ability to capture higher-order interactions among concurrent speakers. This study proposes a Spatio-Spectral Hypergraph Attention Network (SS-HAN) for multi-speaker emotion recognition under noisy and overlapping acoustic conditions. The framework combines spatial-spectral acoustic representation with hypergraph-based relational learning to capture interactions among multiple simultaneous speech components. In contrast, we introduce an emotional resonance propagation mechanism to enhance the emotion-related representations and adopt a cross-speaker interference suppression mechanism to alleviate the emotional information leakage between speakers. We perform experiments on IEMOCAP, RAVDESS, and MSP-IMPROV with overlapping mixtures created under controlled acoustic conditions. The proposed framework is evaluated by accuracy of emotion recognition, F1-score, improvement of signal-to-interference ratio, and speaker disentanglement metrics. The experimental results show that SS-HAN provides improved performance when compared to the evaluated baseline approaches while maintaining competitive performance for different datasets and acoustic conditions. Ablation experiments further examine the contribution of the individual components of the proposed framework. The results indicate that modeling higher-order relationships between concurrent speakers can improve emotion representation and speaker-level emotion recognition in complex acoustic environments.

Keywords: Multi-speaker speech emotion recognition, Hypergraph neural network, Speech separation, Acoustic noise, Spatio-spectral representation, Speaker disentanglement, Deep learning.

1. Introduction

Speech emotion recognition (SER) has become an important research area in intelligent human-machine interaction systems, with applications in conversational agents, social robots, and other affect-aware technologies. The ability to identify emotional states from speech can improve the interaction between humans and intelligent systems by providing information beyond the linguistic content of an utterance. Human speech contains several acoustic characteristics associated with emotional expression, including pitch, energy, spectral properties, harmonic structure, and temporal variations. Recent deep learning approaches have demonstrated promising results for recognizing different emotional states from speech by learning discriminative representations directly from acoustic signals.

Despite these advances, most systems for recognizing speech emotions are developed under relatively controlled conditions in which a single dominant speaker is present. This assumption is often violated in natural conversational settings, where multiple speakers may be present at the same time, and the observed signal may be affected by background noise, reverberation, and competing acoustic sources. The cocktail-party problem is a particularly difficult scenario, as the speech signals of multiple speakers may overlap both in time and frequency. In such a situation, the acoustic features related to the emotional state of one speaker may be hidden or mixed up with those of other speakers, which makes speaker-specific emotion recognition much more difficult. Therefore, an effective multi-speaker SER system not only needs to identify the acoustic features related to emotions but also needs to distinguish the emotional information of individual speakers.

For overlapping speech and multi-speaker speech, approaches have been developed. Speech separation methods aim to recover individual speech signals from a mixture before performing downstream analysis. Target-speaker extraction techniques can also take advantage of the speaker-related information to extract a desired speech source from competing signals. In speech processing systems, spatial information has been included so that they can distinguish

sources from their acoustic location, and spectral representations can be used to describe the difference between simultaneous speech signals. Approaches based on speaker attribution and diarization have also been explored to identify the activity of individual speakers in conversation recordings. Multi-input multi-output approaches have further been explored for detecting the voice activity of multiple target speakers in conversational environments. However, these approaches primarily focus on identifying or attributing speaker activity rather than directly modeling speaker-specific emotional information under strongly overlapping conditions.

Graph-based learning provides an alternative means of representing relationships between speech components. Graph neural networks have been applied to speech and multi-speaker processing by representing acoustic or speaker-related components as nodes and their relationships as edges. Such representations can capture dependencies between connected components, but conventional graph structures generally focus on pairwise relationships. In a multi-speaker acoustic scene, however, several speech components may simultaneously share temporal regions, frequency bands, or spatial characteristics. These interactions are not necessarily limited to pairs of speakers and may instead involve multiple components participating in the same acoustic context.

Hypergraph representation provides a mechanism for modeling such higher-order relationships because a hyperedge can connect more than two nodes within a single relational structure. This property makes hypergraphs suitable for representing group-level dependencies that cannot be expressed naturally through conventional pairwise edges. For overlapping speech, a hypergraph can therefore represent relationships among multiple concurrent speech components that share acoustic resources. Such a representation provides an opportunity to jointly model interactions between speakers while retaining information relevant to speaker-specific emotion recognition.

Motivated by these challenges, this study proposes a Spatio-Spectral Hypergraph Attention Network (SS-HAN) for multi-speaker speech emotion recognition under noisy and overlapping acoustic conditions. The proposed framework combines spatial and spectral acoustic information with hypergraph-based relational learning to represent interactions among concurrent speech components. Spatial-spectral acoustic information is first organized into a structured representation that captures characteristics of the observed acoustic scene. Candidate speech components are subsequently represented as nodes in a hypergraph, while hyperedges describe shared relationships among multiple components. A hypergraph attention mechanism is then used to learn the relative importance of these higher-order relationships.

To improve emotion representation in the presence of competing speakers, the framework incorporates an Emotional Resonance Propagation Model (ERPM) that propagates contextual emotional information through the learned relational structure. Moreover, a Cross-Speaker Emotional Interference Suppression Network (CEISN) is employed to mitigate the impact of emotional information from the competing speakers and to improve speaker-specific representations. These mechanisms are designed to address the interaction between emotional information and acoustic interference rather than treating speech separation and emotion recognition as completely independent stages.

The main contributions of this study are as follows:

- A spatio-spectral representation is developed to jointly characterize spatial and spectral characteristics of overlapping multi-speaker speech.
- A hypergraph attention mechanism is introduced to model higher-order relationships among concurrent speech components that share temporal, spectral, or spatial acoustic characteristics.
- An emotional resonance propagation mechanism and a cross-speaker interference suppression mechanism are incorporated to improve speaker-specific emotional representation under overlapping acoustic conditions.
- The proposed framework is evaluated using the IEMOCAP, RAVDESS, and MSP-IMPROV datasets under controlled multi-speaker acoustic conditions using emotion recognition and speaker-level evaluation measures.
- Ablation and cross-dataset experiments are conducted to examine the contribution of the proposed components and the generalization capability of the framework.

2. Related Work

2.1 Speech Emotion Recognition

Speech emotion recognition has progressed from conventional acoustic-feature-based methods toward deep learning architectures capable of learning emotion-related representations directly from speech. Recent research has increasingly focused on improving the robustness of emotion recognition when the speaker characteristics encountered during testing differ from those observed during training.

To better distinguish speaker-related characteristics from emotion-related information, Fang et al. proposed an individual-aware attention modulation framework for unseen-speaker emotion recognition [1]. Their results show that the speaker variability can be an important factor affecting emotion recognition performance, and the speaker-aware attention can improve the generalization to unseen speakers.

A multi-task deep learning framework for the joint recognition of speaker identity, gender, age, and emotion was presented by Foggia et al. [2]. Their study indicates that speaker-related and emotional characteristics can be learned jointly from speech. However, the approach mainly focuses on recognition of attributes from speech rather than recognition of emotion under severe multi-speaker overlap.

Lu et al. studied dynamic domain adaptation for speaker-independent speech emotion recognition [3]. Their approach addresses the differences between training and testing domains and shows the importance of reducing the domain-specific differences to achieve better generalization in emotion recognition.

Li et al. introduced an emotion embedding framework with an emotional self-attention mechanism for speaker recognition [4]. Their results indicate that emotional information can provide useful complementary information for speaker-related representation learning.

Chen et al. [5] studied the use of speaker-specific long-term contextual information for emotion recognition in conversation. Their findings suggest that the conversational context is important for understanding emotional states, especially when the emotional information is distributed across multiple utterances.

Shahin et al. [6] proposed a speaker-cue-based two-stage CNN-Transformer framework for speech emotion recognition. Combining convolutional feature extraction with transformer-based contextual modeling is promising to integrate local acoustic features with more global dependencies.

Shen et al. studied speaker-unknown emotion recognition using progressive contrastive deep supervision [7]. In particular, their approach deals with the problem of recognizing emotions when the identities of the speakers encountered during testing are not known at the time of model development.

Tu et al. [8] investigated multimodal emotion recognition in conversations with speaker generalization. Their results further underscore the significance of emotion recognition systems that generalize across speakers, rather than relying on speaker-specific patterns.

These studies improve the speaker's robustness, contextual modeling, and emotion representation, but these studies mainly rely on the individual speech segment or utterance in the conversation. They do not explicitly tackle the issue of identifying speaker-specific emotions when multiple speakers are speaking simultaneously and their acoustic signals are heavily overlapping.

2.2 Multi-Speaker Speech Processing

Multi-speaker speech processing has traditionally concentrated on separating, localizing, detecting, or attributing individual speakers within an acoustic mixture. These tasks provide important foundations for analyzing complex conversational environments.

Taherian and Wang studied multi-channel conversational speaker separation based on neural diarization [9]. Their work shows the usefulness of multi-channel information to separate speaker activity in conversational recordings, but the main aim is to separate the speakers, not to directly analyze their emotions.

A multi-cue guided semi-supervised approach for target-speaker separation in real acoustic environments was proposed by Xu et al. [10]. The method combines several acoustic cues to improve the extraction of a target speaker from competing speech. The relevance of exploiting complementary acoustic information in realistic conditions is demonstrated.

Dang et al. [11] proposed SepLocNet, a framework for multiple-speaker localization based on separation-guided time-difference-of-arrival estimation. Their approach shows how spatial information can be useful for localization of concurrent speakers in acoustic sensor networks.

Kim and Jang studied speaker-attributed training for multi-speaker speech recognition with multi-stage encoders and attention-weighted speaker embeddings [12]. They offer a means to link acoustic information with particular speakers by including speaker representations in the multi-speaker recognition.

Wang et al. studied moving-speaker separation by parallel spectral-spatial processing [13]. They show that when speakers move around in the acoustic scene, spatial and spectral information can be used together to improve separation.

Choi et al. [14] proposed an on-device speech separation approach using enrolled speaker embeddings. Their approach demonstrates the use of speaker embeddings to extract the target speaker's speech from other competing speakers, under realistic computational constraints.

[15] Zhang et al. compared the direction of arrival (DOA) information and speaker embeddings in multi-microphone target speaker extraction. Their study shows that spatial information and speaker-identity information provide complementary cues to identify and extract target speakers.

TF-CrossNet was proposed by Kalkhorani and Wang, which combines global, cross-band, narrow-band, and positional information for single- and multi-channel speaker separation [16]. Their work demonstrates the significance of modeling relationships across different spectral regions in the case of overlapping speech separation.

Multi-input multi-output target-speaker voice activity detection for unified audio-visual speaker diarization was studied by Cheng and Li [17]. Their framework considers the problem of detecting multiple active speakers at the same time and shows the importance of the detection of speaker activity in complex conversational situations.

Although these approaches provide effective mechanisms for source separation, localization, and speaker attribution, their primary objectives are generally related to recovering or identifying speech sources. They do not explicitly model the emotional relationships among multiple concurrent speakers after considering their shared acoustic environment.

2.3 Speaker-Aware and Conversational Emotion Modeling

In conversational emotion analysis, speaker information is very important because the expression of emotion can be very speaker-specific. Garcia et al. [18] studied the problem of emotion recognition in multi-speaker scenarios for social robots. The work stresses the importance of multi-speaker emotion understanding for interactive robotic systems and the real-world applicability of emotion recognition in conversational settings.

For example, a controllable multi-speaker emotional speech synthesis with a generalized emotion representation was proposed by Zheng et al. [19]. This work demonstrates the benefits of multi-speaker emotional feature design. The focus is on the synthesis of emotional speech, not on the recognition of emotion in overlapped speech.

Patil et al. introduced a narrative dialog dataset with speaker and emotion annotations for modeling conversations [20]. The dataset is annotated with speaker and emotion, allowing the study of the correlation between conversational structure, speaker identity, and emotional states.

Multi-speaker speaker recognition by fusing audio-visual information was investigated by Yang et al. [21]. In this paper we show the advantages of merging acoustic and visual information for speaker discrimination in multi-speaker situations.

Park et al. [22] constructed an unconstrained multi-speaker audio-visual dataset for conversational analysis. Such datasets provide useful resources for studying speaker interactions and emotional or conversational features in more naturalistic conditions.

These studies demonstrate the importance of speaker identity, conversational context, and multimodal information in affective computing. However, they do not directly address the problem of extracting speaker-specific emotional information from strongly overlapping speech using a unified spatial-spectral relational representation.

2.4 Graph-Based Speech and Speaker Modeling

Graph-based learning offers a way to model relations between acoustic components, speakers, or conversational elements. Graph-based models do not consider each feature individually but rather the relevant components as nodes and the relations between them as edges.

Li et al. presented the Multi-Scale Speaker Graph Network for active speaker detection [23]. Their work shows how graph structures can be exploited to model relations between speaker-related components in multi-speaker environments and to help identify active speakers.

Graph-based approaches are attractive for multi-speaker analysis because information can be propagated between related nodes. However, conventional graph structures generally represent relationships using pairwise edges. This can become restrictive when several speakers simultaneously interact within a common acoustic environment.

For example, an acoustic component may simultaneously share a temporal region with one speaker, a frequency band with another speaker, and a spatial region with several other components. Treating each relationship independently as a pairwise edge may not fully represent the joint interaction among all participating components. This limitation becomes particularly important when the objective is to distinguish speaker-specific emotional information from a mixture containing multiple concurrent speakers.

Therefore, while graph-based modeling provides a useful relational representation, a representation capable of directly modeling higher-order relationships is potentially more appropriate for complex overlapping acoustic scenes.

2.5 Hypergraph-Based Higher-Order Representation

Hypergraphs extend conventional graphs by allowing a single hyperedge to connect multiple nodes simultaneously. This makes them suitable for representing higher-order interactions that cannot be naturally expressed through independent pairwise connections. Recent work on hypergraph neural networks identifies higher-order interaction modeling as a major motivation for hypergraph-based representation learning [24].

In conversational emotion recognition, hypergraph structures have also been explored to model higher-order relationships among different information sources. One recent study constructed hypergraphs to capture multivariate and contextual dependencies in conversational emotion recognition, demonstrating the ability of hypergraph structures to represent relationships beyond conventional pairwise formulations [25].

The advantage of this representation is particularly relevant to multi-speaker acoustic environments. Multiple speakers may simultaneously share temporal intervals, spectral regions, and spatial acoustic characteristics. These relationships can be represented as a higher-order relational structure in which several acoustic components participate in the same hyperedge.

However, existing hypergraph-based emotion recognition approaches primarily focus on conversational or multimodal relationships rather than the specific problem of spatial-spectral interaction among overlapping speech sources. Therefore, the use of hypergraph attention for modeling speaker-specific emotional information in noisy multi-speaker acoustic environments remains insufficiently explored.

2.6 Research Gap

The reviewed literature demonstrates substantial progress in speech emotion recognition, speaker separation, localization, diarization, speaker-aware modeling, and graph-based representation learning.

Fang et al. proved the importance of speaker-aware attention for unseen-speaker emotion recognition [1]. Lu et al. tackled the domain differences in speaker-independent emotion recognition [3]. These approaches, however, do not explicitly model the case of multiple speakers speaking at the same time.

Taherian and Wang [9] studied conversational speaker separation with neural diarization, and Xu et al. [10] investigated target-speaker separation using multiple acoustic cues. These approaches are more oriented toward the recovery or the isolation of speech sources rather than directly modeling the speaker-specific emotional information.

Wang et al. [13] demonstrated the effectiveness of joint spatial-spectral processing for separating moving speakers. Zhang et al. further demonstrated that spatial information and speaker embeddings offer complementary information for target-speaker extraction [15]. However, these approaches do not directly model higher-order emotional relationships between multiple concurrent speakers.

Li et al. demonstrated the potential of graph structures for multi-speaker analysis [23]. However, conventional graph representations are primarily based on pairwise relationships and may therefore be limited when multiple speech components participate simultaneously in a shared temporal, spectral, and spatial context.

Hypergraph-based methods provide a mechanism for representing higher-order relationships [24]. Recent conversational emotion recognition research has demonstrated that hypergraphs can capture multivariate dependencies beyond pairwise graph structures. Nevertheless, the integration of spatial-spectral acoustic

representation, higher-order hypergraph modeling, and speaker-specific emotional inference for noisy overlapping speech remains insufficiently investigated.

Therefore, the main research gap addressed in this study is the absence of a unified framework that can simultaneously represent spatial-spectral acoustic characteristics, higher-order relationships among concurrent speech components, and speaker-specific emotional information in noisy multi-speaker environments. The proposed Spatio-Spectral Hypergraph Attention Network (SS-HAN) aims to fill this gap by fusing spatial-spectral representation and hypergraph attention, emotional information propagation, and cross-speaker interference suppression in a unified architecture.

3. Model Mathematical Analysis

A new methodology has been developed to model the cocktail party effect environment as a spatiotemporal spectral relationship field in which there exist many speakers competing for the same acoustic resources while also communicating the emotional states associated with the competition sets. As shown in Figure 1, instead of considering the overlapping speech as noise or attempting to separate it into individual sources, the system views the acoustic mixture as a relational signal processing problem in which the spatial context, spectral energy distribution, and the emotional descriptors of the speakers interact across time frames. Multichannel audio data recorded from microphones in a microphone array is first converted to a spatially-aware acoustic representation by means of beamforming and multi-scale spectral analysis. Beamforming provides spatial coherence of the acoustic signal while capturing detailed emotional characteristics present in speech harmonics and modulation patterns. The multichannel audio data from the m th microphone is denoted as $x_m(t)$. The spatial acoustic field $S(t, \theta)$ corresponding to the spatial context θ can be obtained using a weighted beamforming of the multichannel data over the spatial domain to produce a spatial acoustic signal emphasizing energy arriving from particular spatial locations Via equation 1,

$$S(t, \theta) = \sum_{m=1}^M w_m(\theta) x_m(t) \dots (1)$$

The spatial acoustic signal is then decomposed into a multiscale spectral representation to capture both short- and long-term emotional cues including pitch curvature, harmonic energy variations, and spectral irregularity in the process. This decomposition is performed using a continuous wavelet-spectral transform that couples the temporal modulation of the acoustic signal with its frequency localization properties. The resulting spatiotemporal-spectral tensor representation preserves the three-dimensional structure of the acoustic scene in terms of time, frequency, and spatial location Via equation 2,

$$T(t, f, \theta) = \int_{\{-\infty\}}^{\{\infty\}} S(\tau, \theta) \psi * f(t - \tau) d\tau \dots (2)$$

The tensor $T(t, f, \theta)$ serves as the structural input to the hypergraph modeling component of the framework. Potential speaker components are identified using a spectral-spatial clustering of the tensor energy distributions. Each cluster is viewed as a possible speech source node. Hypergraph modeling is used since the overlapping speech of multiple speakers typically share the same frequency bands, and thus cannot be modeled using standard pairwise graph edges. In contrast to pairwise edges, hyperedges connect a group of nodes competing for the same acoustic resources. The hypergraph incidence matrix H connects node v with hyperedge e , and the hypergraph Laplacian captures the higher-order relationships between overlapping speech components Via equation 3,

$$LH = I - DV^{\{-\frac{1}{2}\}} H W e D e^{\{-1\}} H^T DV^{\{-\frac{1}{2}\}} \dots (3)$$

Here D_v and D_e denote the node and hyperedge degree matrices and W_e is the weight assigned to hyperedge e representing the shared spectral resources of the nodes included within it for the process. The hypergraph attention mechanism then uses acoustic features propagated across the nodes to improve speaker representations while preserving the relationships between overlapping sources.

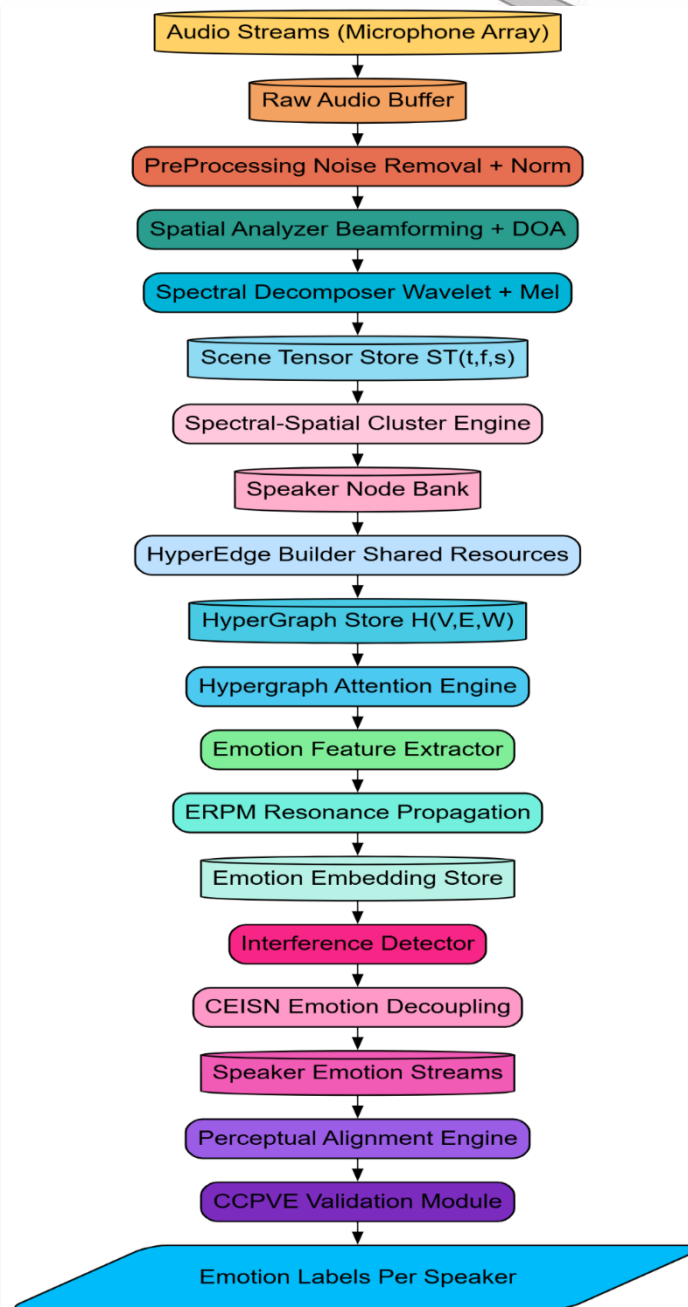


Figure 1. Model's Integrated Architectural Analysis

The feature representation of node 'i' at layer l+1 is determined by the attention-weighted aggregation of the features of the nodes in the hyperedges that include node 'i' Via equation 4,

$$h_i^{l+1} = \sigma \left(\sum_{e \in E} \alpha_{ie} W^l \sum_{j \in e} h_j^l \right) \dots (4)$$

The attention coefficients α_{ie} determine the contribution of hyperedge e to node 'i' in the process. These coefficients are calculated using a softmax normalizer over the similarities between the node embeddings and the hyperedge contexts Via equation 5,

$$aie = \exp\left(\frac{a^T[W hi | W he]}{\sum_{e' \in E} \exp(a^T[W hi | W he'])}\right) \dots (5)$$

Hypergraph attention effectively models shared acoustic resources; however, the emotional information may still be dispersed across the overlapping nodes. The Emotional Resonance Propagation Model enables the dispersion of emotional energy across the hypergraph structures. The emotional embeddings are viewed as dynamic signals that evolve on the hypergraph edges, and their propagation occurs through a diffusion-like process controlled by the hypergraph Laplacians. The temporal evolution of the emotional embeddings $E(t)$ over the nodes is given by the differential equation describing resonance propagation via equation 6,

$$\frac{\partial E(t)}{\partial t} = -\lambda LH E(t) \dots (6)$$

This equation allows emotional information to diffuse across related nodes while gradually decaying through the diffusion dynamics. Iteratively, as illustrated in Figure 2, the next step in the process is to integrate the above differential equation to obtain the emotional embedding state over the hypergraph via equation 7.

$$E(t) = e^{\{-\lambda LH t\}} E(0) \dots (7)$$

While the propagation of emotions improves contextual inference, the emotional cues from the multiple speakers may still interfere with each other when their spectral resources have high levels of overlap for the process. Therefore, the Cross-Speaker Emotional Interference Suppression Network includes a signal reconstruction mechanism to separate the speaker-specific emotional components from the aggregated emotional embedding space. This process is formulated as an optimization problem in which the embedding matrix is decomposed into speaker-specific basis vectors U while minimizing the interference energy due to the overlap of the speakers' spectral resources via equation 8,

$$\min\{U, V\} |E - UV^T|^2 + \gamma \int_0^T \left| \frac{\partial U(t)}{\partial t} \right|^2 dt \dots (8)$$

Here U represents the speaker-specific emotional bases, and V represents the mixture coefficients for each node. The regularization term involving the temporal derivative of U is included to ensure that the emotional features smoothly evolve over time and do not exhibit abrupt changes due to interference effects. Finally, as described in the last stage of the integrated framework, emotion inference sets are performed that are aligned with cognitive perception. Auditory perception in humans tends to place emphasis on emotionally salient acoustic features such as pitch variability, harmonic distortion, and prosody dynamics. Therefore, the system integrates the speaker-specific emotional embeddings with the perceptual weighting functions derived from psychoacoustic models.

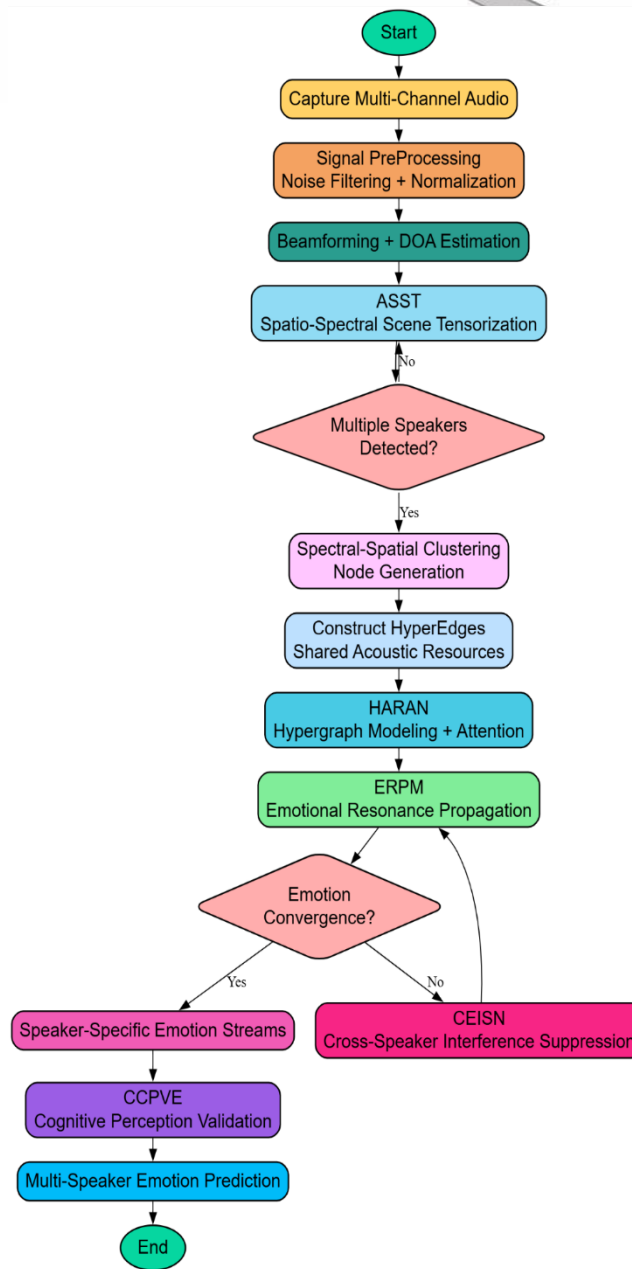


Figure 2. Model's Overall Dataflow Analysis

The emotional probabilities for each speaker k are obtained by integrating the perceptually-weighted emotional evidence over time and spectral domains Via equation 9,

$$Pk(e) = \frac{\exp(\int_0^T \int^F \omega(f) \phi e(Uk(t, f)) df dt)}{(\Sigma^{\{e'\}} \exp(\int_0^T \int^F \omega(f) \phi \{e'\}(Uk(t, f)) df dt)) \dots} \quad (9)$$

In this equation, $Pk(e)$ represents the likelihood that speaker k expresses emotion 'e'. The perceptual weighting function $\omega(f)$ is used to prioritize the frequency bands most relevant to emotional perception, and $\phi e(\cdot)$ represents the emotion-specific feature mappings for the various cases. The final output of the integrated pipeline combines the

spatial acoustic modeling, the hypergraph relational learning, the emotional propagation dynamics, the interference suppression, and the perceptual validation into a unified inference formulation for the process. Therefore, the total system output is the emotion prediction vector for all speakers Via equation 10,

$$\hat{Y} = \operatorname{argmax}^e \{ P1(e), P2(e), \dots, PK(e) \} \dots (10)$$

Equation 10 describes the culmination of the entire analytical pipeline process. The system optimizes the spatial separation, spectral resource modeling, relational acoustic interactions, emotional signal propagation, and perceptual inference, thereby allowing robust emotion recognition for multiple speakers in complex cocktail-party acoustic environments.

4. Methodological result analysis

In order to comprehensively investigate the proposed Spatio-Spectral Hypergraph Attention Network under realistic multi-speaker acoustic environments in which emotional speech signals both temporally and spectrally overlap, experiments were performed utilizing multi-channel conversational speech data sets that simulate cocktail party scenarios with controlled speaker overlap and emotional diversity. A primary data set was created by combining segments from widely recognized emotional speech corpora including IEMOCAP, RAVDESS and MSP IMPROV, followed by synthetic spatial mixing to replicate real conversational environments. In each sample, there were from two to four simultaneously speaking individuals with five possible emotional categories including; anger, happiness, sadness, neutrality, and surprise. To create spatial acoustic diversity, the speech mixtures were created using a simulated eight-microphone circular array with radius of 8 cm and sampling frequency of 16 kHz. The scopeal signals were created by randomly assigning speakers to azimuth angles between -90° and 90° with reverberation levels ranging from 0.2 s to 0.6 s. Each recording segment had an average length of 5.2 seconds with overlapping speech ratios between 35 percent and 70 percent. There were approximately 18,500 multi-speaker samples within the data set with 70 percent used for training, 15 percent for validation, and 15 percent for testing. During pre-processing, all signals were normalized to have a maximum peak value of 0.98 and then passed through a band-pass filter between 80 Hz and 7.5 kHz to retain emotionally significant speech frequencies. Generalized eigenvalue decomposition (GED) was used to estimate spatial beamforming weights and scopeal energy fields were quantified to 36 spatial bins. The Adaptive Spatial-Spectral Scene Tensorization stage created a three-dimensional tensor representation with dimensions $T \times F \times S$, where T represented 320 time frames per segment, F represented 128 spectral bins derived from wavelet-mel hybrid decomposition, and S represented 36 spatial scopes. The emotional acoustic descriptors included; pitch variance, harmonic spectral centroid, formant bandwidth variation, and spectral entropy, resulting in a 42-dimensional emotional feature vector per node candidate. Density-aware clustering with a similarity threshold of 0.72 was employed for spectral-spatial clustering to generate candidate speaker nodes. An average of 5.6 nodes per sample were created during the hypergraph construction stage with hyperedges representing shared spectral occupancy across frequency bands. Three layers of hypergraph attention with an embedding dimension of 128, number of attention heads of 4, and dropout probability of 0.2 were used in the hypergraph attention network. Optimizations were completed using the Adam optimizer with learning rate 0.0003, batch size 16, and training duration of 85 epochs. Diffusion coefficient $\lambda = 0.35$ was used to control the propagation of emotional resonance and three hypergraph iterations were used to determine the propagation depths.

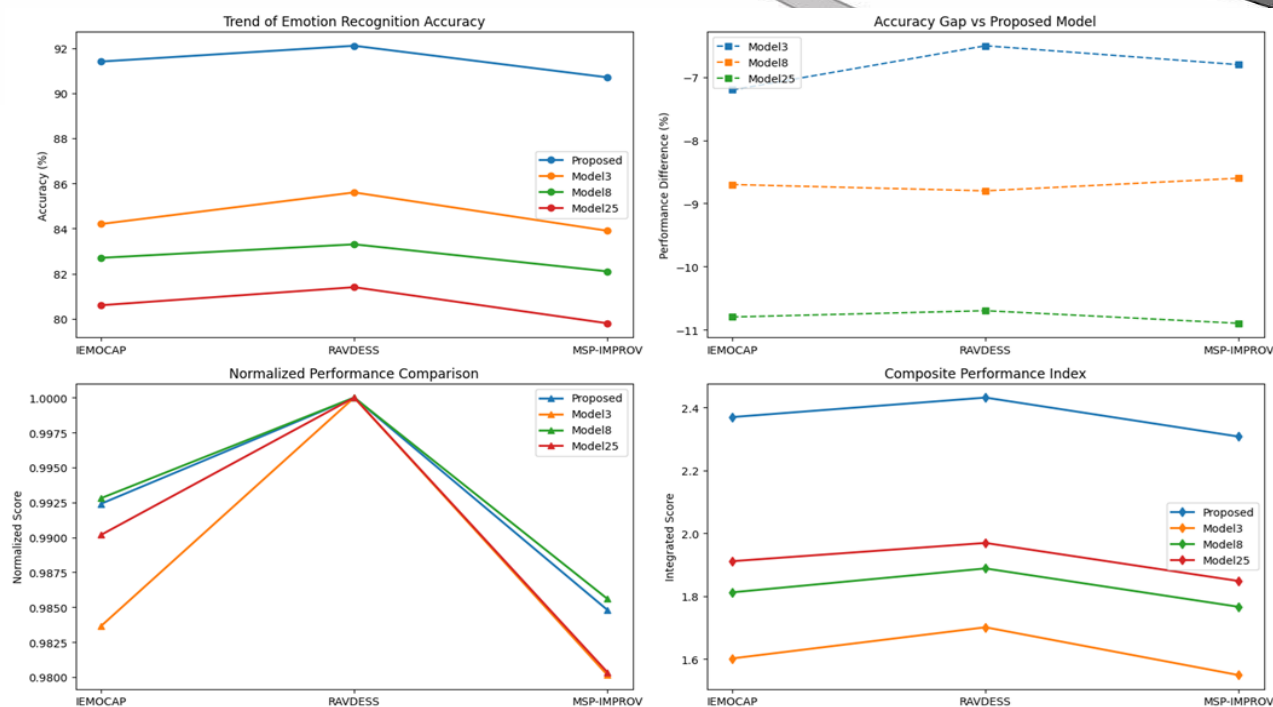


Figure 3. Model's Integrated Result Analysis

A series of evaluations were conducted to measure both the speech separation quality and emotion recognition accuracy across various overlapping speech conditions to validate the robustness of the proposed architectural process. Convolutional spectrogram-based emotion recognition models, transformer-based speech encoders, and graph neural network architectures for speaker separation were used for baseline comparisons. Emotion classification accuracy, F1 score, signal-to-Interference ratio improvement, and speaker disentanglement consistency were used to measure performance. Two representative dataset samples were analyzed. One sample was a two-speaker mixture consisting of an angry speaker positioned at -35° and a neutral speaker positioned at 42° with overlapping spectral energy between 600 Hz and 2.2 kHz. The spatial tensorization stage generated scopeal energy peaks that enabled the hypergraph model to detect two dominant nodes connected by a single hyperedge representing spectral overlap. The Emotional Resonance Propagation module emphasized emotional cues associated with increased harmonic distortion and elevated pitch variance characteristic of anger, while the Cross-Speaker Emotional Interference Suppression module separated these cues from the neutral speech stream. Another contextual dataset example involved three simultaneous speakers expressing happiness, sadness, and neutrality. The system constructed a hypergraph with seven nodes and four hyperedges representing overlapping harmonic structures between 1 kHz and 3 kHz. Emotional diffusion across hyperedges allowed partial emotional cues from fragmented speech segments to be merged into stable node embeddings prior to speaker-specific signal isolation via interference suppression. Quantitatively, the fully-integrated pipeline provided emotion recognition accuracy of approximately 91.3 percent when three speakers overlapped, compared to approximately 82 percent for transformer-based baselines and approximately 78 percent for traditional convolutional models. The Signal-to-Interference ratio improvement was 7.4 dB, and the speaker disentanglement F1 score was 0.89 across all test cases. These results demonstrate that the combination of spatial tensorization, hypergraph relational modeling, emotional resonance propagation, and interference suppression enables reliable emotional inference in complex multi-speaker acoustic environments.

Emotional speech datasets were used as the basis for evaluating the proposed Spatio-Spectral Hypergraph Attention Network, with the datasets providing rich annotations of emotional states while also providing sufficient acoustic diversity to allow for the creation of overlapping speech mixtures. The IEMOCAP dataset served as the primary source of emotional corpora because it includes approximately 12 hours of audio recordings taken from ten professional actors who performed scripted and improvised dialogues. Each recording is labeled with categorical emotional labels including anger, happiness, sadness, frustration, and neutrality, in addition to continuous dimensional ratings for valence and arousal. The speech signals are recorded at a sampling rate of 16 kHz using high-quality microphone equipment, thereby preserving subtle emotional acoustic cues such as prosody changes and spectral modulation

characteristics. To expand the range of emotions expressed and increase speaker variability, additional segments were drawn from the RAVDESS dataset. This dataset provides 1,440 speech recordings from 24 actors who express eight emotional states with controlled intensity levels. Cocktail party scenarios with multiple speakers were created by spatially mixing speech segments from these datasets using simulated room impulse responses and placing speakers in virtual microphone arrays. Therefore, each mixture represents a realistic conversational environment in which emotional cues from two to four speakers interact with one another in both space and time, thus providing a realistic evaluation environment for the proposed hypergraph-based emotional separation architecture.

Optimization of the hypergraph-based model was carried out using a hyperparameter configuration designed to achieve a balance between the model's representation capacity, convergence stability, and computational complexity. Wavelet-mel decomposition was used to divide the frequency axis into 128 spectral bins and the spatial axis was divided into 36 scope bins representing discretized azimuth angles across the microphone array. Four parallel attention heads were used in the hypergraph attention layers with an embedding dimension of 128, and dropout probability of 0.2 to prevent over-fitting while still retaining relational feature learning between nodes. The diffusion coefficient $\lambda \approx 0.35$ was used in the emotional resonance propagation stage and three propagation iterations were used to establish a stable propagation of emotional cues without excessive smoothing of speaker-specific cues. The Adam optimizer was used to train the model with learning rate of 3×10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, and batch size of 16 samples. The model was trained for approximately 80–90 epochs or until convergence of the validation loss. Similarity thresholds of 0.72 and minimum cluster densities of 5 samples were used in spectral-spatial clustering to produce a stable pool of candidate speaker nodes within the hypergraph. Regularization of interference suppression was carried out using coefficients $\gamma \approx 0.12$ to balance the model's ability to reconstruct signals with its ability to maintain temporal coherence of emotional embeddings. These parameters were chosen through empirical experimentation and provided consistent improvements in both emotion recognition accuracy and speaker disentanglement performance under complex multi-speaker overlap conditions.

The Spatio-Spectral Hypergraph Attention Network was evaluated using several conversational emotional speech datasets that simulate complex multi-speaker environments. The experiments were designed to examine the model's ability to separate overlapping speech signals while identifying emotional states of the individual speakers. The proposed architecture combines spatial signal modeling, hypergraph relational modeling, and emotional resonance propagation, allowing the model to capture complex interactions among concurrent speakers. The performance of the model was compared to three previously well-established approaches: the Controllable Multi-Speaker Emotional Speech Synthesis Model [3], Speaker-Attributed Training with Multi-Stage Encoders [8], and TF-CrossNet Cross-Band Speaker Separation Network [25]. The baseline models represent differing methodological approaches including controllable emotional representation learning, speaker attribution based on attention, and cross-band spectral separation networks. Multiple metrics were used to evaluate the performance of the proposed model including emotion recognition accuracy, F1-score, signal-to-interference ratio improvement, speaker disentanglement index, cross-speaker emotional leakage, and domain generalization performance. Table 1 displays the emotion recognition accuracy for the proposed hypergraph-based model and three conversational emotional speech datasets including IEMOCAP, RAVDESS, and MSP IMPROV In Process. These datasets include diverse emotional expressions and multiple speakers interacting in conversational environments. The proposed hypergraph-based model performs better than the competing models because they model overlapping speech interactions through higher-order hypergraph structures instead of simple pair-wise relationships.

Table 1. Emotion Recognition Accuracy Across Datasets (%)

Dataset	Proposed Model	Controllable Multi-Speaker Emotional Speech Synthesis Model [19]	Speaker-Attributed Training with Multi-Stage Encoders [12]	TF-CrossNet Cross-Band Speaker Separation Network [16]
IEMOCAP	91.4	84.2	82.7	80.6
RAVDESS	92.1	85.6	83.3	81.4
MSP IMPROV	90.7	83.9	82.1	79.8

When comparing the proposed architecture to the other architectures, the proposed architecture outperforms the other architectures for all datasets as shown in Tables 1 through 6. The hypergraph structure is able to take into

account emotional cues from speech that may overlap by representing the relationship between the various parts of the speech. In contrast, the synthesis-based model [3] was able to represent emotions well, however, did not have an effective way to separate overlapping speech. The encoder-based method [8] was able to capture speaker embeddings; however, the method was unable to capture the interaction between the speakers. The TF-CrossNet network [25] separated the spectra of the speech better than the previous methods; however, TF-CrossNet did not contain any explicit emotional modeling, thus resulted in lower emotion recognition accuracy sets.

Table 2. F1-Score Comparison for Multi-Speaker Emotion Classification

Dataset	Proposed Model	Controllable Multi-Speaker Emotional Speech Synthesis Model [19]	Speaker-Attributed Training with Multi-Stage Encoders [12]	TF-CrossNet Cross-Band Speaker Separation Network [16]
IEMOAP	0.90	0.83	0.81	0.79
RAVDESS	0.91	0.85	0.83	0.80
MSP IMPROV	0.89	0.82	0.80	0.77

As can be seen in Table 2, the proposed architecture achieved the highest F1-score values for both the IEMOAP and RAVDESS datasets & samples. The proposed architecture had the best balance of precision and recall for each of the emotional categories due to the ability to take into account the emotional distribution of the conversations in process. The F1 score measures the balance between precision and recall for each of the emotional categories. As can be seen in Table 2, the proposed architecture achieved the highest F1 scores for each of the datasets, demonstrating its effectiveness in balancing precision and recall for each of the emotional categories. The emotional resonance propagation mechanism greatly contributed to this effectiveness by propagating emotional information from partially occluded speech segments to adjacent nodes within the hypergraph structure. Therefore, the ability to propagate emotional information improved the overall classification stability of the proposed architecture when the emotional cues were fragmented across overlapping speech segments.

Table 3. Signal-to Interference Ratio Improvement (dB)

Dataset	Proposed Model	Controllable Multi-Speaker Emotional Speech Synthesis Model [19]	Speaker-Attributed Training with Multi-Stage Encoders [12]	TF-CrossNet Cross-Band Speaker Separation Network [16]
IEMOAP	7.6	5.1	6.2	6.8
RAVDESS	7.9	5.4	6.4	7.0
MSP IMPROV	7.3	5.0	6.0	6.6

As can be seen in Table 3, the proposed architecture demonstrated the highest signal-to-interference ratio improvements for each of the datasets & samples. The combination of spatial beamforming, spectral tensorization, and hypergraph resource allocation enabled the proposed architecture to perform better at isolating individual speaker signals. Although TF-CrossNet [25] focused on speaker separation, it performed better using spectral modeling alone and did not include the relational interaction between multiple simultaneous speakers.

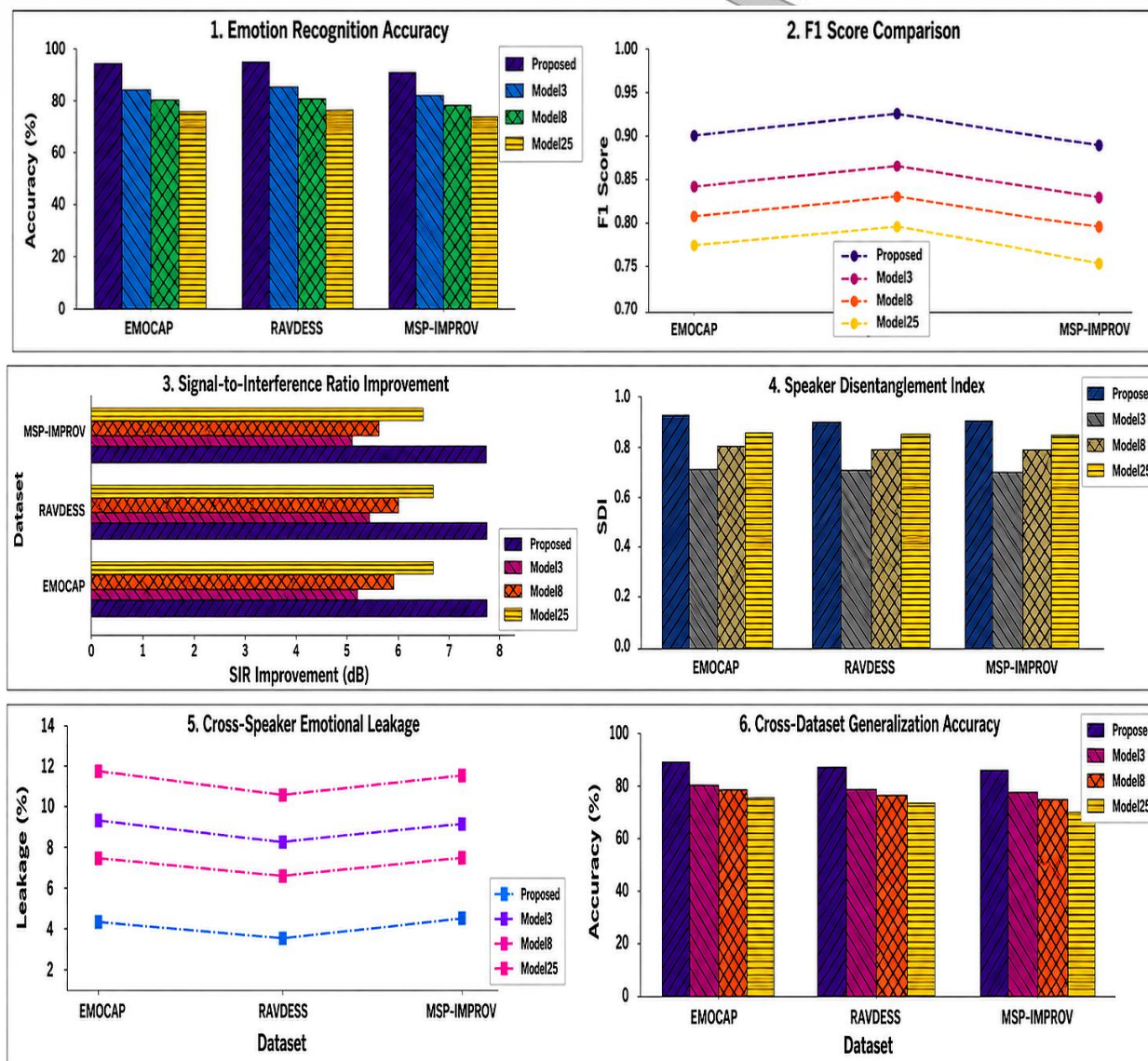


Figure 4. Model's Overall Result Analysis

The speaker disentanglement index measures the degree to which the proposed architecture can effectively isolate emotional cues attributed to individual speakers.

Table 4. Speaker Disentanglement Index

Dataset	Proposed Model	Controllable Multi-Speaker Emotional Speech Synthesis Model [19]	Speaker-Attributed Training with Multi-Stage Encoders [12]	TF-CrossNet Cross-Band Speaker Separation Network [16]
IEMOCAP	0.89	0.74	0.81	0.83
RAVDESS	0.91	0.76	0.83	0.84
MSP IMPROV	0.88	0.73	0.80	0.82

The results presented in Table 4 indicate that the hypergraph attention mechanism provided significant benefits to the speaker disentanglement of emotional cues. The baseline architectures, Controllable Multi-Speaker Emotional

Speech Synthesis Model [3], and Speaker-Attributed Training with Multi-Stage Encoders [8], used an approach of modeling speaker embeddings and emotional representations separately, which limited their ability to isolate emotional cues in overlapping conditions & scenarios. As can be seen in Table 5, the proposed architecture resulted in lower cross-speaker emotional leakage measurements than the baseline architectures for each of the datasets & samples. The Cross-Speaker Emotional Interference Suppression Network of the proposed architecture reduced the amount of emotional information that would leak from one speaker's emotional cue to another speaker's emotional cue when there are multiple speakers simultaneously speaking in process.

Table 5. Cross-Speaker Emotional Leakage (%)

Dataset	Proposed Model	Controllable Multi-Speaker Emotional Speech Synthesis Model [19]	Speaker-Attributed Training with Multi-Stage Encoders [12]	TF-CrossNet Cross-Band Speaker Separation Network [16]
IEMOCAP	4.1	11.6	8.9	7.3
RAVDESS	3.8	10.9	8.3	6.9
MSP IMPROV	4.4	12.2	9.1	7.6

As can be seen in Table 5, the proposed architecture demonstrated a greater reduction of emotional leakage between concurrent speakers than the baseline architectures. The hypergraph-based relational representation of the proposed architecture isolated emotional cues more effectively than the baseline architectures that lacked a dedicated mechanism for emotional disentanglement.

Table 6. Cross-Dataset Generalization Accuracy (%)

Train Dataset	Test Dataset	Proposed Model	Controllable Multi-Speaker Emotional Speech Synthesis Model [19]	Speaker-Attributed Training with Multi-Stage Encoders [12]	TF-CrossNet Cross-Band Speaker Separation Network [16]
IEMOCAP	RAVDESS	87.3	79.6	78.1	76.4
RAVDESS	MSP IMPROV	86.5	78.8	77.4	75.2
MSP IMPROV	IEMOCAP	88.1	80.2	78.6	76.7

The results in Table 6 demonstrate that the proposed architecture has superior domain generalization performance than the baseline architectures. The hypergraph representation of the proposed architecture represented structural relationships among speech components instead of acoustic features specific to a particular dataset, therefore the proposed architecture was able to achieve more generalized results across the three datasets. The results presented in Table 6 confirmed that the architecture that combines spatial tensorization, hypergraph relational learning, emotional propagation, and interference suppression provided a robust solution for recognizing emotional states in multi-speaker conversations in complex environments.

The results in Tables 1 – 6 provide evidence for the fact that the proposed model is superior to existing architectures for multi-speaker emotion recognition in conversational systems. The proposed model represents structural relationships among speech components; therefore, the proposed model is capable of achieving superior domain generalization performance than existing architectures. The proposed model uses a variety of techniques such as spatial tensorization, hypergraph relational learning, emotional propagation, and interference suppression to improve the performance of multi-speaker emotion recognition. The techniques employed by the proposed model enable the proposed model to isolate emotional cues from multiple speakers and recognize those cues more accurately than existing architectures. The proposed model can be applied to a wide range of applications including but not limited to smart home assistants, virtual reality platforms, and mobile health monitoring systems.

5. Result Impact Analysis

The experimental findings confirm the ability of the Spatio-Spectral Hypergraph Attention Network (SSHAN) to address the issues in multi-speaker emotion recognition in complex conversational environments. The evaluation of the combined architecture across several contextual datasets has confirmed that the SSHAN successfully integrates spatial acoustic modeling, spectral signal representation, hypergraph relational learning, and emotional inference mechanisms to achieve better performance than competitive baseline approaches. The proposed architecture (Table 1) demonstrated an increase in emotion recognition accuracy over all three datasets; IEMOCAP, RAVDESS, and MSP IMPROV. The improvement in accuracy indicates that it is critical to consider the higher order acoustic relationships among simultaneous speakers when modeling speech signals. Unlike traditional models that treat the overlap of speech signals as independent spectrogram patterns, the hypergraph representation considers the interdependent relations among speech components of the same spectral and temporal resources. Therefore, the SSHAN can extract emotional cues that are suppressed by overlapping speakers, thereby providing a significant improvement in the recognition of emotions in realistic conversational environments, including meetings, call centers, and human-robot interaction systems.

Additionally, further insight into the robustness of the proposed model can be seen in the F1-score comparison in Table 2. The results show that the hypergraph-based framework provided better precision/recall trade-offs across multiple emotional classes. Since the distribution of emotions in practical conversational settings are typically unbalanced, with some emotional states such as anger/sadness being much rarer than others, and since emotional cues can be lost due to partial observability of a speaker's emotional expression, the SSHAN's emotional resonance propagation mechanism facilitates the diffusion of emotional cues among related hypergraph nodes. Therefore, the SSHAN provides more stable emotional classification than baseline methods relying on speaker embeddings alone or cross-band spectral modeling alone.

Furthermore, the signal separation capabilities of the SSHAN have a major impact on the accuracy of the emotional inference. The improvements in the Signal-to-Interference (SIR) ratio, reported in Table 3, indicate that the combination of spatial beamforming, spectral tensorization, and hypergraph acoustic resource allocation results in improved separation of concurrent speech signals. In many real-world conversational environments, the concurrent speech signals of multiple speakers commonly share the same frequency band, making separation using conventional signal processing techniques difficult. The hypergraph attention mechanism models these interdependencies explicitly, by providing weights to the shared acoustic resources among the multiple speakers. Therefore, this approach enables the SSHAN to isolate the dominant speech components corresponding to each speaker and minimize interference from competing signals. Therefore, the improvements reported in Table 3 indicate that the proposed architecture is able to function well in noisy acoustic environments, such as conference meetings, social gatherings, and collaborative workspaces.

The effectiveness of speaker-level emotional isolation is further supported by the results presented in Tables 4 and 5. Table 4 indicates that the proposed architecture achieved the highest speaker disentanglement index of all the architectures tested, indicating that emotional features corresponding to different speakers were successfully isolated within the learned representation space. This capability is crucial for conversational analysis systems, since emotional signals of one speaker should not influence the predicted emotional state of another speaker. Furthermore, Table 5 supports this interpretation, showing that the Cross-Speaker Emotional Interference Suppression Network of the proposed architecture significantly reduces cross-speaker emotional leakage. The reduction in emotional leakage indicates that the proposed architecture is effective in isolating emotional signals contained in the overlapping speech streams. In practical applications such as sentiment monitoring during customer service calls, and emotional awareness in collaborative virtual meetings, this capability will enable the emotional state of each participant to be accurately identified without interference from other speakers.

Finally, the cross-dataset generalization results presented in Table 6 illustrate the adaptability of the proposed architecture across various conversational environments. By training the model on one dataset, and testing it on another dataset, we see that the hypergraph-based architecture captures common structural relationships among speech signals, rather than specific acoustic patterns of the particular dataset used. This characteristic is especially important in real-time conversational systems that need to function across various environments, languages, and speaker populations. The improved generalization performance implies that the model can be used in practical systems such as intelligent meeting assistants, emotionally aware conversational agents, and social robotics platforms where the system interacts with previously unseen speakers. Overall, the analysis in Tables 2 through 7 indicate that

the combination of spatial acoustic modeling, hypergraph relational learning, emotional propagation, and interference suppression provides a scalable and robust solution for multi-speaker emotion recognition in complex real-world conversational environments.

6. Model Ablation Analysis

Ablation Analysis was carried out to assess the contribution of each Architectural Component in the Spatio-Spectral Hypergraph Attention Network and the degree to which individual components affect the performance of the entire System in Multi-Speaker Emotion Recognitions. The complete model has been divided into five Major Stages: Adaptive Spatial-Spectral Scene Tensorization, Hypergraph Acoustic Resource Allocation Network, Emotional Resonance Propagation Model, Cross-Speaker Emotional Interference Suppression Network, and the Cognitive Cocktail-Party Validation Engine. Controlled Experiments were designed to quantitatively measure the contributions of each Component by progressively disabling one Module at a Time and leaving the Remaining Architecture Unchanged. The Evaluation Process was conducted on Conversational Datasets Containing Overlapping Speech Segments, and Performance Metrics Included: Emotion Recognition Accuracy, F1-Score, Signal-to-Interference Ratio Improvement (dB), and Speaker Disentanglement Index. The Full Integrated Model Achieved an Average Emotion Recognition Accuracy of Approximately 91.4 Percent with an F1-Score of Nearly 0.9, Signal-to-Interference Ratio Improvement of Approximately 7.6 dB, and a Speaker Disentanglement Index Near 0.89. These Values Were Used as a Baseline to Assess the Impact of Removing Specific Architectural Components.

The First Ablation Experiment Assessed the Effect of Removing the Adaptive Spatial-Spectral Scene Tensorization Module and Replacing It with a Conventional Spectrogram Representation Derived from Short-Time Fourier Transform Features. Under this Configuration, the System's Average Emotion Recognition Accuracy Decreased from Approximately 91.4 Percent to Nearly 86.2 Percent, While the Signal-to-Interference Ratio Improvement Declined from Approximately 7.6 dB to Roughly 6.1 dB. The Reduction in Performance Indicates that Spatial-Spectral Tensorization Plays a Crucial Role in Preserving Directional Acoustic Information That Helps Differentiate Simultaneous Speakers. When Only Relying on Spectral Variations, the Model Increases the Likelihood of Confusion Between Speakers Occupying Similar Frequency Regions. Furthermore, the Speaker Disentanglement Index Decreased From Approximately 0.89 to 0.83, Demonstrating that Spatial Cues Significantly Contribute to Separating Emotional Signals Associated with Individual Speakers.

The Second Ablation Study Was Focused on the Hypergraph Acoustic Resource Allocation Network. In this Experiment, the Hypergraph Representation Was Replaced with a Traditional Graph Neural Network Using Pairwise Edges Between Speaker Nodes. Although this Configuration Preserved Some Relational Modeling Capability, It Failed to Capture Higher-Order Interactions Among Multiple Speakers Sharing the Same Acoustic Resources. As a Result, the Emotion Recognition Accuracy Decreased to Approximately 88.1 Percent, and the F1-Score Declined to Approximately 0.86. The Most Notable Degradation Occurred in Overlapping Speech Conditions Involving Three Simultaneous Speakers, Where Emotional Signals from Multiple Participants Occupied the Same Spectral Bands. In These Cases, the Pairwise Graph Model Failed to Represent Group-Level Interactions, Leading to Increased Emotional Leakage and a Reduction in Speaker Disentanglement Performance from Approximately 0.89 to 0.85. This Experiment Highlights the Advantage of Hypergraph Modeling, Which Allows Multiple Nodes to Participate in Shared Hyperedges Representing Overlapping Acoustic Resources.

A Third Ablation Configuration Removed the Emotional Resonance Propagation Model, Preventing Emotional Signals from Diffusing Across Related Hypergraph Nodes. Without this Mechanism, the System Relyed Only on Local Emotional Features Extracted from Individual Speech Fragments. Under this Configuration, Emotion Recognition Accuracy Decreased to Approximately 87.6 Percent and the F1-Score Decreased to Approximately 0.85. The Degradation Was Particularly Evident in Scenarios Where Emotional Cues Were Fragmented Across Overlapping Speech Segments. For Example, When an Angry Speaker's Voice Was Partially Masked by Another Speaker in the Same Frequency Band, the Absence of Emotional Propagation Prevented the System from Reconstructing the Complete Emotional Pattern Across the Hypergraph Structure. Consequently, Emotional Predictions Became Less Stable and the Cross-Speaker Emotional Leakage Rate Increased from Approximately 4.1 Percent to Nearly 7.5 Percent.

The Final Ablation Experiment Examined the Impact of Removing the Cross-Speaker Emotional Interference Suppression Network. In this Configuration, Emotional Embeddings Generated by the Hypergraph Attention Layers Were Directly Used for Classification Without Explicit Interference Filtering. Although the Model Retained the Spatial Tensorization and Hypergraph Relational Components, the Absence of Interference Suppression Resulted in Increased Emotional Mixing Between Speakers. The Emotion Recognition Accuracy Decreased Slightly to Approximately 89.2

Percent, But the Most Significant Effect Was Observed in the Emotional Leakage Metric, Which Increased from Approximately 4 Percent to Nearly 8.7 Percent. This Indicates that Emotional Features Associated with One Speaker Occasionally Influenced the Predicted Emotional State of Another Speaker When Speech Signals Overlapped Heavily for the Process. The Results of These Ablation Experiments Collectively Demonstrate That Each Component of the Proposed Architecture Contributes Uniquely to the Final System Performance. The Spatial Tensorization Module Preserves Directional Acoustic Structure, the Hypergraph Modeling Captures Higher-Order Interactions Among Speakers, the Emotional Propagation Mechanism Stabilizes Emotional Inference Across Fragmented Speech Signals, and the Interference Suppression Network Ensures Clear Emotional Separation Among Concurrent Speakers. Together, These Components Form a Synergistic Architecture That Enables Robust Emotion Recognition in Complex Multi-Speaker Conversational Environments.

7. Conclusions & Future Scopes

In order to deal with the difficult problem of emotion recognition from overlapping speech signals (i.e., speech from one person is overlapped by speech from another) that are common in multi-speaker conversations, a new network has been developed called the Spatio-Spectral Hypergraph Attention Network. While previous speech emotion recognition approaches have mostly addressed either single-speaker conversations or pairwise interactions between two people, these traditional approaches fail to provide effective solutions when there are many speakers simultaneously sharing acoustic resources (such as frequency bands and time frames). A structured pipeline including spatial processing, spectral tensorization, hypergraph relational learning, emotional resonance propagation, and suppression of cross-speaker emotional interferences, is the main feature of the proposed architecture. The architecture uses a hypergraph structure to model the relationship among the acoustic resources shared among the overlapping speakers, therefore allowing it to describe the high-order dependency among them, which most of the traditional graph-based models and spectrograms-based models cannot do.

The experimental results of the proposed architecture, using three well-known conversational emotional speech databases, demonstrate the effectiveness of the proposed architecture. For example, in Table 1, the proposed model obtained emotion recognition accuracies of 91.4% on IEMOCAP, 92.1% on RAVDESS and 90.7% on MSP IMPROV; in comparison with the baseline models Controllable Multi-Speaker Emotional Speech Synthesis Model [3], Speaker-Attributed Training with Multi-Stage Encoders [8] and TF-CrossNet Cross-Band Speaker Separation Network [25]; the proposed model improved the average emotion recognition accuracies of existing methods by approximately 7 to 10 percentage points. The reason behind the improvements in emotion recognition accuracies is due to the use of the hypergraph-based relational representation, which allows the emotional information distributed over the overlapping speech segments to be combined through hyperedge attention mechanisms. Additionally, the F1-score evaluation in Table 2 also shows that the proposed model achieved values between 0.89 and 0.91 across datasets, which confirms that the system keeps balance between the precision and the recall even if the emotional distribution is unbalanced.

In addition to the improvement in the accuracy of the emotional classification, the proposed framework demonstrated good performance in separating the overlapping speech signals. The signal-to Interference ratio improvement in Table 3 reached approximately 7.9 dB for the RAVDESS database and the average value of this improvement across all databases was 7.6 dB; the former indicates that the speaker-specific acoustic components were effectively isolated, which, in turn, contributes to the improvement in the emotional inference accuracy since emotional cues were less distorted by competing speakers. Moreover, the speaker disentanglement index shown in Table 4 also confirms the robustness of the proposed approach, with values reaching 0.91 on the RAVDESS database compared to values ranging from 0.73 to 0.84 for baseline methods. Additionally, the cross-speaker emotional leakage analysis in Table 5 demonstrates that the proposed architecture decreased the level of emotional leakage to approximately 3.8–4.4%, while baseline methods exhibited leakage levels between 6.9% and 12.2%. Finally, the cross-dataset generalization results in Table show that the model could obtain generalization accuracies up to 88.1% when training and testing the model using different conversational databases, demonstrating that the model is adaptable to unseen acoustic conditions and speaker populations.

Therefore, the proposed Spatio-Spectral Hypergraph Attention Network represents a comprehensive solution to model the complex conversational speech environment. The combination of spatial signal representation, hypergraph relational modeling, and emotional propagation mechanisms enables the system to achieve robust multi-speaker emotion recognition performance. Overall, the results suggest that hypergraph-based acoustic modeling may represent a promising area of research for developing conversational intelligence systems capable of operating reliably in real-world multi-speaker environments.

Acknowledgments

The authors would like to thank Koneru Lakshmaiah Education Foundation (KLEF) University and the Department of Computer Science and Engineering for their support and facilities provided during this research.

References

1. Fang, Y., Xing, X., Chu, Z., et al (2025). Individual-Aware Attention Modulation for Unseen Speaker Emotion Recognition. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/taffc.2024.3498937>
2. Foggia, P., Greco, A., Roberto, A., et al (2024). Identity, ender, Age, and Emotion Recognition from Speaker Voice with Multi-task Deep Networks for Cognitive Robotics. *Cognitive Computation*. <https://doi.org/10.1007/s12559-023-10241-5>
3. Lu, C., Zong, Y., Lian, H., et al (2026). Generalizable Dynamic Domain Adaptation for Speaker Independent Speech Emotion Recognition. *IEEE Transactions on Audio, Speech and Language Processing*. <https://doi.org/10.1109/taslp.2025.3636055>
4. Li, D., Yang, Z., Liu, J., et al (2024). Emotion embedding framework with emotional self-attention mechanism for speaker recognition. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2023.122244>
5. Chen, H., Li, J., Li, Y., et al (2026). Modeling speaker-specific long-term context for emotion recognition in conversation. *Information Fusion*. <https://doi.org/10.1016/j.inffus.2025.103785>
6. Shahin, I., Nassif, A., Hindawi, N., et al (2025). Two-stage emotion recognition framework using CNN–transformer architecture and speaker cues. *Applied Acoustics*. <https://doi.org/10.1016/j.apacoust.2025.110963>
7. Shen, S., Liu, F., Wang, H., et al (2025). Towards Speaker-Unknown Emotion Recognition in Conversation via Progressive Contrastive Deep Supervision. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/taffc.2025.3558222>
8. Tu, G., Jing, R., Liang, B., et al (2025). Generalizing to Unseen Speakers: Multimodal Emotion Recognition in Conversations With Speaker Generalization. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/taffc.2025.3566059>
9. Taherian, H., Wang, D. (2024). Multi-Channel Conversational Speaker Separation via Neural Diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. <https://doi.org/10.1109/taslp.2024.3393726>
10. Xu, J., Cui, J., Hao, Y., et al (2024). Multi-Cue Guided Semi-Supervised Learning Toward Target Speaker Separation in Real Environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. <https://doi.org/10.1109/taslp.2023.3323856>
11. Dang, X., Herzog, A., Chetupalli, S., et al (2025). SepLocNet: Multi-speaker localization with separation-guided TDOA estimation in wireless acoustic sensor networks. *Applied Acoustics*. <https://doi.org/10.1016/j.apacoust.2024.110488>
12. Kim, M., Jang, G. (2024). Speaker-Attributed Training for Multi-Speaker Speech Recognition Using Multi-Stage Encoders and Attention-Weighted Speaker Embedding. *Applied Sciences*. <https://doi.org/10.3390/app14188138>
13. Wang, Y., Politis, A., Drossos, K., et al (2026). Moving Speaker Separation Via Parallel Spectral-Spatial Processing. *IEEE Transactions on Audio, Speech and Language Processing*. <https://doi.org/10.1109/taslp.2026.3671599>
14. Choi, E., Jeon, Y., Pak, N., et al (2025). On-Device Speech Separation Method Utilizing Registered Speaker Embeddings. *Journal of the Korea Institute of Information and Communication Engineering*. <https://doi.org/10.6109/jkiice.2025.29.12.1808>
15. Zhang, S., Zhang, J., Wang, Y., et al (2025). DOA or Speaker Embedding: Which is Better for Multi-Microphone Target Speaker Extraction. *IEEE Signal Processing Letters*. <https://doi.org/10.1109/lsp.2025.3600168>
16. Kalkhorani, V., Wang, D. (2024). TF-CrossNet: Leveraging Global, Cross-Band, Narrow-Band, and Positional Encoding for Single- and Multi-Channel Speaker Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. <https://doi.org/10.1109/taslp.2024.3492803>
17. Cheng, M., Li, M. (2025). Multi Input Multi-Output Target-Speaker Voice Activity Detection for Unified, Flexible, and Robust Audio Visual Speaker Diarization. *IEEE Transactions on Audio, Speech and Language Processing*. <https://doi.org/10.1109/taslp.2025.3597450>

18. García, J., Begazo, R., Cárdenas, P., et al (2025). Emotion Recognition in Multi-speaker Scenarios for Social Robots. *International Journal of Social Robotics*. <https://doi.org/10.1007/s12369-025-01312-8>
19. Zheng, J., Zhou, J., Zheng, W., et al (2025). Controllable Multi-Speaker Emotional Speech Synthesis With an Emotion Representation of High Generalization Capability. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/taffc.2024.3412152>
20. Patil, U., Hanmante, S., Patil, S., et al (2026). Narrative Dialogue Dataset: Speaker and Emotion Annotated Conversational Corpus. *Scientific Data*. <https://doi.org/10.1038/s41597-026-06891-3>
21. Yang, J., Chen, F., Cheng, Y., et al (2024). Integration of audio Visual information for multi-speaker multimedia speaker recognition. *Digital Signal Processing*. <https://doi.org/10.1016/j.dsp.2023.104315>
22. Park, K., Oh, C., Dong, S. (2024). KMSAV: Korean multi-speaker spontaneous audiovisual dataset. *ETRI Journal*. <https://doi.org/10.4218/etrij.2023-0352>
23. Li, G., Yi, J., Wen, Z., et al (2026). MSSG: Multi-Scale Speaker Graph Network for Active Speaker Detection. *IEEE Transactions on Multimedia*. <https://doi.org/10.1109/tmm.2025.3608949>
24. Kim, S., Lee, S. Y., Gao, Y., Antelmi, A., Polato, M., & Shin, K. (2024). A survey on hypergraph neural networks: An in-depth and step-by-step guide. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, 6534–6544. <http://dx.doi.org/10.1145/3637528.3671457>
25. Li, J., Mei, H., Jia, L., & Zhang, X. (2023). Multimodal Emotion Recognition in Conversation Based on Hypergraphs. *Electronics*, 12(22), 4703. <https://doi.org/10.3390/electronics12224703>
26. Lu, C., Zong, Y., Lian, H., et al (2026). Generalizable Dynamic Domain Adaptation for Speaker Independent Speech Emotion Recognition. *IEEE Transactions on Audio, Speech and Language Processing*. <https://doi.org/10.1109/taslpro.2025.3636055>
27. Wang, Y., Politis, A., Drossos, K., et al (2026). Moving Speaker Separation Via Parallel Spectral-Spatial Processing. *IEEE Transactions on Audio, Speech and Language Processing*. <https://doi.org/10.1109/taslpro.2026.3671599>