electronics

MDPI

Article

Beyond Speaker Dependency: Adaptive Graph Attention Learning for Low-Resource Speech Emotion Recognition

Jeetamitra Satapathy [1], Tusar Kanti Dash [1], Sambit Kumar Mishra [2,*] and Ganapati Panda [1]

[1] Department of Electronics and Communication Engineering, C. V. Raman Global University, Bhubaneswar 752054, Odisha, India; jeetamitras@gmail.com (J.S.); tusarkantidash@gmail.com (T.K.D.); ganapati.panda@gmail.com (G.P.)
[2] Department of Computer Science and Engineering, SRM University-AP, Amaravati 522240, Andhra Pradesh, India
* Correspondence: sambitkumar.m@srmap.edu.in

Abstract

Speech emotion recognition (SER) remains challenging for low-resource languages because of limited labelled data, speaker variability, and overlapping acoustic characteristics across emotion categories. This study proposes a perceptual graph attention network for five-class speech emotion classification using statistical representations of mel-frequency cepstral coefficients (MFCCs). The MFCC coefficients are represented as graph nodes, with each node described by statistical descriptors extracted from the static MFCC, delta, and delta-delta features. The graph integrates coefficient-adjacency and acoustic-similarity relationships, while multi-head graph attention learns the relative contributions of connected acoustic representations. The proposed framework is evaluated on SantaliESC, BanglaSER, and IITKGP-SEHSC under speaker-dependent and speaker-independent settings. Comparative experiments are conducted against conventional machine-learning classifiers and a convolutional neural network. Further analyses investigate graph connectivity, descriptor sensitivity, class-wise performance, and graph size. The findings demonstrate the potential of graph-based relational learning for modelling statistical acoustic representations and improving speech emotion classification across the selected low-resource speech corpora.

Keywords: speech emotion recognition; Santali language; mel-frequency cepstral coefficients (MFCCs); graph neural network; low-resource languages

1. Introduction

1.1. Background and Motivation

Speech emotion recognition aims to identify emotional states from speech by analysing acoustic characteristics such as pitch, energy, speech rhythm, and spectral patterns. It has applications in affective computing and human–computer interaction. Recent studies have explored graph-based methods for enhancing emotional representation learning. Li et al. [1] presented a graph-LSTM network for speech emotion recognition, and Pentari et al. [2] investigated graph-based representations of speech signals. Wang and Kim [3] proposed a skip graph convolutional network with graph attention network, and Yan et al. [4] proposed the temporal–spatial learnable graph convolutional network for emotion classification. Despite these advances, SER remains challenging for low-resource languages because labelled emotional speech data are limited and models may not generalise reliably to unseen speakers. Santali Emotional Speech Corpus (SantaliESC) contains 80 native speakers' speech samples in five emotion classes: angry, happy, neutral, sad, and surprise. This corpus can be

Check for updates

Academic Editor: Peiying Zhang

Received: 23 October 2026
Revised: 14 November 2026
Accepted: 17 November 2026
Published: 29 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

used for the study of the classification of speech emotion in an under-resourced language, namely Santali. MFCCs are used extensively for speech analysis since they compactly represent the short-term spectral characteristics of speech. Conventional classifiers typically operate on the extracted feature vectors and do not explicitly model the relationships among them. Graph neural networks provide an alternative approach that exploits connectivity information to learn representations. However, their effectiveness depends on the feature representation, the graph construction strategy, and the evaluation protocol.

With these considerations in mind, this study investigates a graph neural network operating on MFCC-derived representations for five-class emotion classification. The proposed framework is compared with conventional MFCC-based classifiers and a neural baseline. Further experiments analyse the graph construction, class-wise performance, descriptor sensitivity, and graph size. Evaluation for SantaliESC, BanglaSER, and IITKGP-SEHSC also gives more insights into the performance of the model for the selected speech emotion corpora.

1.2. Research Gap and Novelty

Graph neural networks have been investigated for speech emotion recognition using different graph representations and learning strategies. Li et al. [1] explored graph-based temporal modelling through a graph-LSTM network, while Pentari et al. [2] investigated graph-based representations of speech signals. Wang and Kim [3] combined skip graph convolutional networks with graph attention networks, and Yan et al. [4] developed a temporal–spatial learnable graph convolutional network. These studies demonstrate the potential of graph-based learning but also indicate that graph representation and topology are important design considerations.

However, the potential of graph-based emotion classification for low-resource languages like Santali has not been explored to a great extent. The recently released SantaliESC [5], a dedicated Santali emotional speech corpus, makes it possible to address this problem. The usefulness of modelling relationships between MFCC-based speech representations, in particular, has to be evaluated against conventional classifiers under both speaker-dependent and speaker-independent evaluation settings. Additionally, the influence of graph structure and speaker variation should be systematically studied to better understand the generalisation capacity of graph-based emotion recognition. In this study, an adaptive graph attention-based framework for low-resource speech emotion recognition is investigated to overcome the above challenges. The framework represents the MFCC coefficients as graph nodes, with their statistical descriptors serving as node features. Graph connectivity describes the relationship between the coefficients, and attention-based feature aggregation learns the relative importance of the coefficients when discriminating different emotional categories. The framework was tested in speaker-dependent and speaker-independent conditions to analyse the classification performance and speaker generalisation. The main contributions are as follows:

- Graph-based acoustic representation: An MFCC-based graph representation is investigated in which cepstral coefficients are modelled as graph nodes and their associated statistical descriptors are used as node features.
- Attention-based relational learning: A graph attention mechanism is employed to learn the relative importance of connected acoustic representations for speech emotion classification.
- Graph construction analysis: Ablation experiments investigate the contribution of graph construction to classification performance.

- Speaker-dependent and speaker-independent evaluation: The proposed framework is evaluated under both evaluation protocols to investigate its classification performance and its ability to generalise to previously unseen speakers.
- Systematic experimental investigation: The effects of graph construction and statistical descriptor groups are examined, alongside comparisons with conventional machine-learning classifiers, to assess the contribution of graph-based learning.

1.3. Paper Organization

The remainder of this paper is organised as follows. Section 2 reviews related work, Section 3 presents the proposed framework, Section 4 presents the experimental results and discussion, and Section 5 concludes this study.

2. Related Work

Research on speech emotion recognition (SER) has focused on acoustic feature extraction, machine-learning classifiers, deep neural networks, and graph-based representation learning. For the present study, the literature on conventional acoustic features, deep-learning approaches, graph-based SER, and low-resource language challenges is reviewed.

2.1. Acoustic Features and Conventional Classification

MFCCs, which provide a compact representation of the spectral characteristics of speech, remain widely used in SER. Dar and Delhibabu [6] surveyed speech databases, acoustic features, and classification techniques, including MFCC-based speech classification. Madanian et al. [7] thoroughly synthesized machine-learning techniques for SER and highlighted the challenges of feature extraction, classification, and data. According to their reviews, the performance of models is not just dependent on the classifier but also on the feature representation, the characteristics of the dataset, and the methodology of the evaluation. Research into SER has since been extended using deep-learning methods [8]. Hashem et al. [9] conducted a systematic review of deep-learning architectures, datasets, pre-processing, and feature extraction techniques. Barhoumi and BenAyed [10] studied deep learning approaches for SER with data augmentation to mitigate the lack of training data [11,12]. These studies motivate comparing the proposed graph-based model with conventional and neural baselines under a unified experimental framework.

2.2. Graph-Based Speech Emotion Recognition

Graph-based learning incorporates the relationships among feature representations into the classification process. This concept has recently been expanded upon with multimodal fusion and adaptive graph construction. In this regard, Qi et al. [13] proposed a multimodal fusion graph convolutional network for modeling the intra-modal and inter-modal relationships. Lian et al. [14] designed an adaptive multi-graph convolutional network which can learn the contextual information of the local frequency domain. Attention-pooling has also been explored in the field of graph-based emotion recognition from variable length utterances [15]. These methods have demonstrated the potential of graph structures to learn emotional representations, and they also indicate that graph topology and inter-feature relationships must be carefully designed [16]. Unlike these multimodal approaches, the present study focuses on unimodal graph learning from MFCC-derived representations of Santali speech and evaluates whether the selected graph construction strategy improves classification relative to conventional MFCC-based approaches.

2.3. SER for Low-Resource Languages

Low-resource SER is challenging because of limited labelled data, linguistic variation, and differences in emotional expression. Rathnayake et al. [17] surveyed the SER performance for low-resource and Indigenous languages and identified the need for language-aware approaches and additional studies with limited data learning strategies. The availability of a corpus, however, is no proof that a model is generalising to unseen speakers. It is therefore necessary that the evaluation does not contain the same speakers as the training set.

3. Materials and Methods

3.1. Datasets

Three emotional speech corpora, namely, Santali Emotional Speech Corpus (SantaliESC), BanglaSER, and IITKGP-SEHSC were used in this study. SantaliESC is publicly available via Zenodo (https://doi.org/10.5281/zenodo.20702265) and includes 2000 speech recordings from 80 native Santali speakers ranging in age between 18 and 60 [5]. The recordings are evenly split between the five emotion classes: angry, happy, neutral, sad, and surprise, with 400 recordings in each emotion class. All audio files are provided in WAV format, sampled at 48 kHz with 16-bit resolution, with a total duration of approximately 1 h 53 min. The average emotion recognition accuracy for the human perceptual evaluation was 74.6% [5]. For additional evaluation, emotional Bangla speech from BanglaSER [18] and emotional Hindi speech from IITKGP-SEHSC [19] were used. These data enabled the proposed method to be explored in three languages. All of the audio recordings underwent a standardized MFCC extraction process before evaluation. In the case of SantaliESC, MFCC features were given as input to the proposed GNN framework. For speaker-independent evaluation, the data were split at the speaker level so that recordings from any given speaker appeared in only one of the training, validation, and test sets. To make fair comparisons, all baseline methods used the same partitions. Speaker-independent partitioning was also used for BanglaSER and IITKGP-SEHSC for which speaker identifiers and adequate speaker-level data were available. Fivefold data augmentation was performed using time stretching, pitch shifting, volume gain, and time shifting, applied exclusively to the training recordings [20]. To ensure a consistent five-class classification setting across the selected speech emotion corpora, a common label space comprising Angry, Happy, Neutral, Sad, and Surprise was considered. The original emotion annotations of each dataset were examined before harmonisation. Labels were retained only when their definitions corresponded to one of the five target emotion categories. Dataset-specific categories outside this common label space were excluded rather than merged into another category. In particular, Disgust, Fear, and Sarcastic were excluded from IITKGP-SEHSC.

3.2. Cepstral Feature Extraction

MFCCs are used to represent the short-term spectral characteristics of speech and are widely employed in speech analysis and classification [21]. Each speech recording is divided into overlapping frames of 25 ms with a frame shift of 10 ms. A Hamming window is applied to each frame to reduce spectral leakage, followed by Fast Fourier transform (FFT) to obtain the frequency-domain representation. A mel filter bank is then applied to the power spectrum, and the logarithm of the filter-bank energies is transformed using the discrete cosine transform (DCT) to obtain 13 MFCC coefficients per frame. The mel scale is defined as follows:

$$f_{mel} = 2595\, log_{10}\,(1 + \frac{f}{700}) \quad (1)$$

where f and f_{mel} denote the frequency on the linear and mel scales, respectively [22]. To capture the temporal dynamics of speech, first-order (delta (Δ)) and second-order (delta-delta (Δ^2)) temporal derivatives are calculated from the static MFCCs. For the graph representation, statistical descriptors are computed across frames for each coefficient in each of the three feature streams. Ten descriptors are extracted: mean, median, root-mean-square (RMS) value, maximum, minimum, first quartile, third quartile, standard deviation, skewness, and kurtosis. These descriptors characterise the distribution of each coefficient over the recording. Each of the 13 MFCC coefficients is represented as a graph node. Accordingly, each recording is represented by 13 MFCC coefficient indices, with each coefficient associated with a 30-dimensional feature vector comprising ten statistical descriptors from each of the three feature streams (MFCC, delta, and delta-delta coefficients). The resulting representation contains 390 feature values per recording.

3.3. Perceptual Graph Attention Network (PGAT)

The proposed PGAT framework integrates statistical MFCC descriptors with graph attention learning to classify speech emotions. The relationships among the 13 MFCC coefficient nodes are represented through graph edges based on acoustic similarity and coefficient adjacency. The graph attention network (GAT) uses node features and their connections to generate a graph-level representation for emotion classification. The framework consists of four stages: graph construction, node feature transformation, graph attention learning, and graph-level classification (as illustrated in Figure 1).

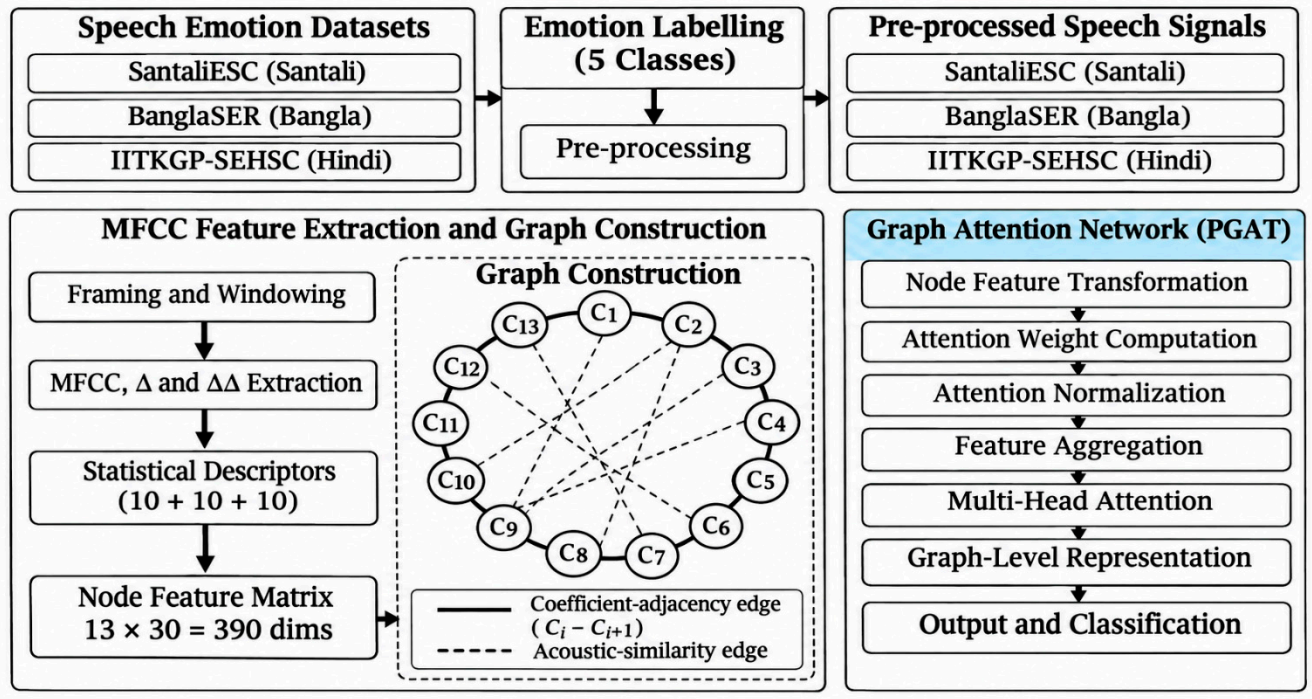


Figure 1. Block diagram of the proposed PGAT implementation.

3.3.1. Graph Construction

In the present model, the graph is built with the statistical representation of the MFCCs obtained from each recording. Each MFCC coefficient forms one node, and its 30-dimensional feature vector comprises 10 statistical descriptors from each of the three feature streams. Cosine similarity is used to measure the acoustic similarity between nodes. Graph connectivity is determined by top-K neighbour selection (k = 4), supplemented by edges whose cosine similarity exceeds the similarity threshold (Table 1). Cepstral adjacency

edges are also added to connect nodes of adjacent MFCC coefficient indices which represent local relationships between neighbouring cepstral coefficients, as shown in Figure 1. The resulting graph captures both cepstral adjacency and acoustic-similarity relationships among the MFCC coefficient nodes within the same recording. The resulting graph is defined as follows:

$$G = (V, E) \tag{2}$$

where V and E denote the node and edge sets, respectively. The resulting graph preserves cepstral-adjacency and acoustic-similarity relationships among the MFCC coefficient nodes within the same recording.

Table 1. Hyperparameter configuration of the proposed PGAT model.

Hyperparameter	Value/Description
MFCC coefficients per recording	13
Statistical descriptors per node	10 per feature stream
Number of graph nodes	13
Feature dimension per node	30
Graph connectivity	Cosine-similarity-based, top-K and cepstral-adjacency connections
Cosine-similarity threshold	0.64
Number of GAT layers	2
Attention heads per layer	8
Hidden units per head	32
Global pooling	Global mean pooling
Dropout rate	0.20
Optimizer	Adam
Learning rate	0.001
Batch size	16
Maximum training epochs	200, with early stopping
Loss function	Categorical cross-entropy

3.3.2. Node Feature Transformation

Each node v_i is associated with a feature vector $x_i \in \mathbb{R}^{30}$ obtained from the statistical MFCC descriptors. The input node features are first projected into a higher-dimensional latent space using a learnable linear transformation:

$$h_i = W x_i \tag{3}$$

where W denotes the trainable weight matrix. This transformation maps the handcrafted statistical features into a representation suitable for graph attention learning.

3.3.3. Graph Attention Learning

For every connected node pair (i, j), an attention coefficient is computed, following the graph attention formulation of Veličković et al. [23], to estimate the importance of the neighbouring node:

$$e_{ij} = LeakyReLU\left(a^T \left[h_i \parallel h_j \right]\right) \tag{4}$$

where a denotes the learnable attention vector and $\|$ represents feature concatenation. The coefficients are normalised using the softmax function:

$$\alpha_{ij} = \frac{exp\left(e_{ij}\right)}{\Sigma_{k\in N(i)}\, exp\,(e_{ik})} \tag{5}$$

where $N(i)$ denotes the neighbourhood of node i. The updated node representation is obtained by aggregating the neighbouring features:

$$h_i{}' = \sigma\left(\Sigma_{j\in N(i)}\, \alpha_{ij}\, h_j\right) \tag{6}$$

where $\sigma(\cdot)$ denotes the activation function. Multi-head attention is employed to improve feature robustness and to stabilise training. Residual connections and learnable feature-stream attention are incorporated to improve feature integration and information propagation across the graph. The graph structure provides an interpretable representation of relationships among MFCC coefficient nodes. Cepstral adjacency connections preserve local relationships between neighbouring coefficient indices, whereas acoustic-similarity connections identify nodes with similar statistical feature patterns. The graph attention mechanism further assigns learnable attention weights to connected nodes, indicating their relative contributions during feature aggregation. Thus, the proposed PGAT captures complementary relationships among acoustic representations for speech emotion classification.

3.3.4. Graph-Level Classification

After attention-based feature aggregation, the node embeddings are pooled to form a graph-level representation using a global mean-pooling operation:

$$h_G = \frac{1}{|V|}\, \Sigma_{i\in V} h_i{}' \tag{7}$$

The graph-level embedding is fed to a fully connected layer followed by a softmax classifier that estimates the posterior probability of each emotion category. The network parameters are optimized using the categorical cross-entropy loss, which quantifies the discrepancy between the predicted class probabilities and the ground-truth labels. Optimization is performed using the Adam optimizer to minimise the categorical cross-entropy loss. The complete procedure is summarised in Algorithm 1.

Algorithm 1. Proposed PGAT Model

Input: Speech recordings and their emotion labels.
Output: Predicted emotion category for each recording.

1. Extract 13 MFCC coefficients from each speech recording.
2. Calculate delta and delta-delta coefficients from the static MFCCs.
3. Calculate ten statistical descriptors for each coefficient across the three feature streams.
4. Represent each MFCC coefficient as a graph node with a 30-dimensional feature vector.
5. Normalise node features using statistics estimated from the training data.
6. Construct a graph using cosine similarity, top-K neighbour selection (K = 4), and cepstral-adjacency connections between adjacent coefficient indices.
7. Learn a linear transformation to project node features into latent space.
8. Compute multi-head graph attention coefficients and aggregate neighbouring node representations.
9. Global mean pooling to get a graph-level representation.
10. Predict the emotion category using a fully connected softmax classifier.
11. Use the Adam optimizer and categorical cross-entropy loss to train the network and use the given classification metrics to evaluate it.

3.4. Experimental Setup

The proposed Perceptual Graph Attention Network (PGAT) was completely implemented in MATLAB R2024b for signal preprocessing, feature extraction, graph construction, graph attention learning, and visualisation. All experiments were conducted at C. V. Raman Global University, Bhubaneswar, Odisha, India, using a Lenovo laptop equipped with an 11th Generation Intel® Core™ i5-11320H processor, and 24 GB RAM. Table 1 summarises the model configuration. The network consists of two graph attention layers, each having eight attention heads and 32 hidden units per head. Global mean pooling was performed to produce a graph-level representation for emotion classification. The model uses a dropout rate of 0.20, the Adam optimizer, a learning rate of 0.001, and a batch size of 16. Categorical cross-entropy was used as the loss function.

The accuracy, precision, recall, F1 score, and specificity were used to evaluate the performance of the model. In the speaker-independent protocol, no speaker appears in both the training and test sets. The feature-normalisation parameters were computed on the training partition and then used on the test partition. This procedure helps to reduce the leakage of information and offers a more accurate estimation of performance on unseen speakers. PGAT is compared with SVM, k-NN, and MLP classifiers using the MFCC statistical features, and with a CNN using log-mel spectrogram features. Additional analyses include alternative graph configurations, class-wise performance, descriptor sensitivity, and graph size. To ensure a fair comparison, the same data partitions and evaluation procedures were used for all the model configurations.

4. Results

This section presents the proposed perceptual graph attention network (PGAT) for speech emotion classification based on the statistical MFCC coefficient representation and graph attention learning. The evaluation compares the classification performance of PGAT with traditional machine-learning and neural-network baselines and examines the impact of graph construction, descriptor groups, and graph size. SantaliESC is the primary evaluation dataset, while BanglaSER and IITKGP-SEHSC are used to assess performance on emotional speech in other languages. Comparisons are made using the same data partitions and evaluation procedures for each dataset.

4.1. Comparison with Baseline Models

To ensure a dimension-matched comparison, the conventional machine-learning baselines (SVM, k-NN, and MLP) were trained using the same 390-dimensional statistical feature representation supplied to the proposed PGAT model. Specifically, the 390-dimensional representation was obtained by concatenating ten statistical descriptors calculated for each of the 13 MFCC coefficients across three feature streams: static MFCC, delta, and delta-delta coefficients. For the conventional classifiers, this representation was flattened into a single feature vector without graph construction or graph-based message passing. In contrast, PGAT retained the structured representation of 13 nodes, each containing 30 features, and employed graph connectivity and attention-based feature aggregation for classification. The comparative evaluation was conducted under two protocols: speaker-dependent (SD) and speaker-independent (SI). In the SD evaluation, recordings from the same speakers could appear in both the training and testing partitions, whereas in the SI evaluation, the speakers in the training and testing partitions were mutually exclusive. All models were evaluated using the corresponding identical data partitions and consistent feature-normalisation procedures, with normalisation parameters estimated exclusively from the training data. Thus, the comparison maintains the same underlying acoustic information and overall feature dimensionality while examining model performance under both evaluation protocols.

PGAT achieved the highest observed accuracy among the compared models on all three datasets, followed by the CNN with log-mel features. The values obtained from the conventional classifiers based upon MFCC were comparatively lower. This pattern suggests a potential advantage of using graph attention to model relationships among MFCC coefficient nodes over conventional feature-based classification or local spectro-temporal learning. To further evaluate the reliability of the proposed PGAT model, Wilson confidence intervals and Cohen's kappa coefficients were calculated using the test-set confusion matrices [24]. Under speaker-dependent evaluation, the proposed PGAT model achieved an accuracy of 91.3% and a Cohen's kappa coefficient of 0.891. Similarly, under speaker-independent evaluation, the model achieved an accuracy of 90.1% and a Cohen's kappa coefficient of 0.876. These results indicate strong agreement between the predicted and actual emotion labels, further supporting the reliability of the proposed model.

4.2. Ablation Study of Graph Construction

Figure 2 shows the result of combined graph connectivity in PGAT. Acoustic-similarity edges outperform coefficient-adjacency edges on all the datasets (81.5% vs. 80.0% for SantaliESC, 81.0% vs. 79.0% for BanglaSER, and 80.0% vs. 78.0% for IITKGP-SEHSC). The combined graph further improves the accuracy to 91.3%, 90.7%, and 90.0%, respectively. This improvement may be due to the fact that the two types of edges, coefficient -adjacency and acoustic-similarity, are complementary, in the sense that coefficient-adjacency edges represent local relations among neighbouring MFCC coefficients, while acoustic-similarity edges represent relations between features that are not necessarily neighbours to each other. They are integrated to incorporate both structural and acoustic dependencies of speech, which helps the attention mechanism to generate more discriminative representations for speech emotion classification.

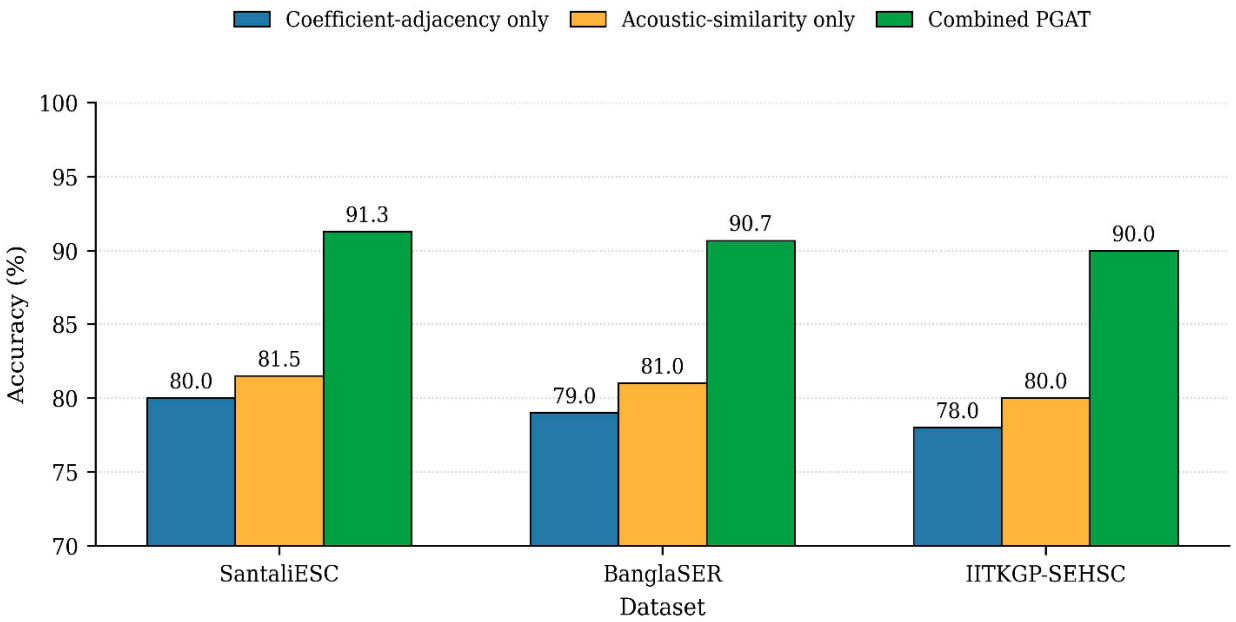


Figure 2. Effect of graph connectivity on PGAT accuracy: coefficient-adjacency edges only, acoustic-similarity edges only, and the combined graph.

The 30 input features were grouped into four functional categories based on their statistical properties: central tendency (mean and median), signal magnitude (RMS, maximum and minimum), dispersion (first quartile, third quartile and standard deviation), and distribution shape (skewness and kurtosis). This grouping follows established statistical interpretations of location, variability, and distribution shape, with signal magnitude treated

separately to distinguish amplitude-related descriptors. Similar statistical functionals have been employed in acoustic analysis for speech research [25]. The same grouping was applied to the MFCC, delta-MFCC, and delta-delta-MFCC streams, resulting in 6, 9, 9, and 6 features, respectively. A fixed-weight descriptor sensitivity analysis was then performed by individually setting each group's normalized feature values to zero while keeping the trained PGAT parameters unchanged. The reference model achieved 91.30% accuracy, which decreased to 62.25%, 49.25%, 49.25%, and 42.50% after occluding central tendency, signal magnitude, dispersion, and distribution-shape descriptors, respectively. The results are shown in Figure 3. The largest performance reduction occurred for distribution-shape descriptors, indicating their substantial contribution to the classification performance. This analysis evaluates descriptor importance through feature occlusion without retraining separate models.

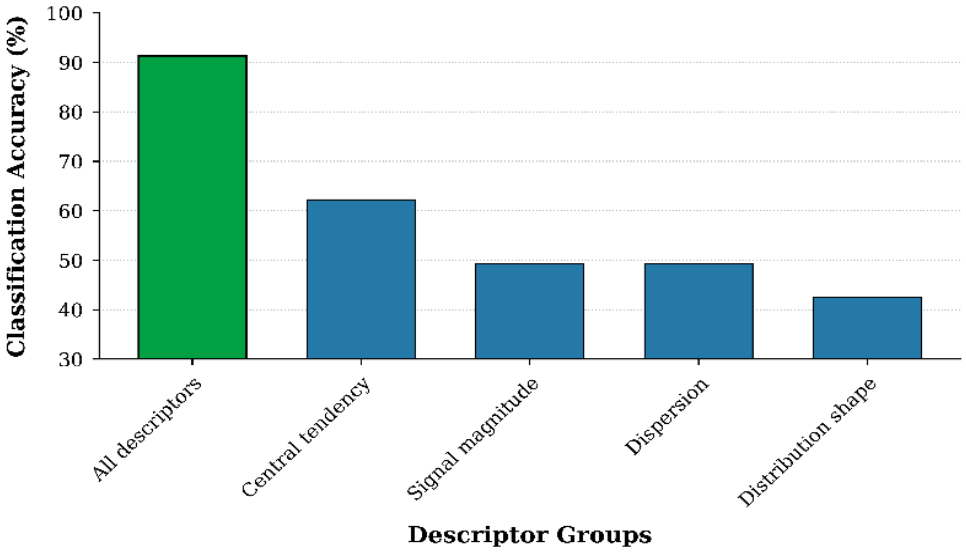


Figure 3. Descriptor-group sensitivity analysis (fixed-weight feature occlusion) on SantaliESC.

4.3. Class-Wise Performance Analysis

The class-wise results of the proposed PGAT model on SantaliESC are shown in Figure 4, which includes the confusion matrix and class-wise precision, recall, and F1 score. The diagonal values, as shown in Figure 4a, range from 0.910 to 0.920, indicating that the classification is consistent over the five emotion classes. Neutral achieves the highest recognition rate (0.920), followed by Happy (0.915) and Angry, Sad, and Surprise (0.910 each). The largest off-diagonal value (0.040) corresponds to Surprise being misclassified as Sad, suggesting some similarity in the acoustic properties of these two emotions. As seen in Figure 4b, Surprise has the highest precision (0.948), while the highest recall (0.920) is obtained by Neutral. The F1 scores are between 0.899 and 0.929, where Surprise has the highest score, and Sad has the lowest score. Results show that the classes performed consistently, but there is some confusion between the classes with similar emotions.

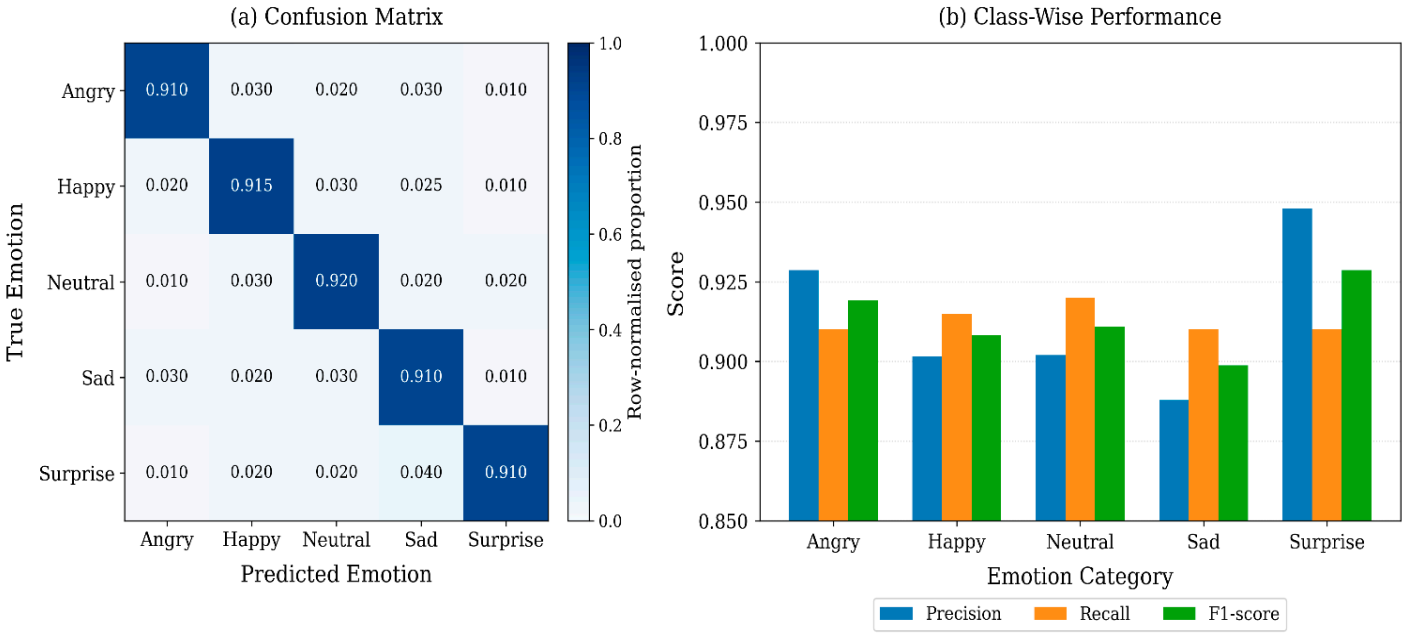


Figure 4. Class-wise performance of the proposed PGAT model on SantaliESC: (**a**) confusion matrix; (**b**) precision, recall, and F1 score by emotion category.

4.4. Evaluation Across Datasets

The performance of the proposed PGAT model was tested on the SantaliESC, BanglaSER, and IITKGP-SEHSC datasets with five common emotion categories (Angry, Happy, Neutral, Sad, and Surprise) to investigate the consistency of the proposed PGAT model across different speech emotion datasets. To ensure a consistent definition of classes, the Disgust, Fear, and Sarcastic categories were dropped from the IITKGP-SEHSC. As shown in Table 2, the speaker-dependent (SD) accuracy of PGAT is 91.3%, 90.7%, and 90.0% on SantaliESC, BanglaSER, and IITKGP-SEHSC, respectively. The corresponding accuracy for speaker-independent (SI) evaluation is 90.1%, 89.2%, and 88.1%. The performance of the three datasets was fairly uniform, which suggests that the proposed feature-level graph representation is applicable across the three evaluated language corpora. PGAT was tested using an integrated dataset which was created by merging the five common emotion categories of all three corpora. Speaker-dependent accuracy was lower at 87.2%, which may be due to the different linguistic characteristics, speaker attributes, and recording conditions among the datasets.

Table 2. Performance comparison with baseline models.

Dataset	SVM SD	SVM SI	k-NN SD	k-NN SI	MLP SD	MLP SI	CNN SD	CNN SI	PGAT SD	PGAT SI
SantaliESC	90.3	89.0	86.8	85.2	89.5	87.8	90.0	88.2	91.3	90.1
BanglaSER	88.7	87.1	85.9	84.3	88.0	86.4	89.0	87.3	90.7	89.2
IITKGP-SEHSC	87.2	85.6	84.5	82.9	86.5	84.8	87.8	86.0	90.0	88.1

4.5. Effect of Graph Size on Recognition Accuracy and Computational Complexity

The effect of graph size was evaluated using 5, 7, 9, 11, and 13 nodes. The classification accuracies were approximately 50.5%, 51.5%, 45.5%, 90.5%, and 91.3%, respectively. The mean inference times were about 6.69, 5.53, 5.96, 5.54, and 5.55 ms per recording, respectively. The 13-node configuration was the most accurate with an inference time of 5.55 ms per recording. Inference time did not decrease proportionally with the number of nodes. Thus, the 13-node configuration performs best in terms of recognition while having a low inference

time. Compared with the conventional SVM classifier (3910 learned parameters), PGAT (25,413 trainable parameters) is a more expressive model that captures relationships among acoustic feature representations, at the cost of increased model complexity.

4.6. Discussion

Table 3 presents the performance of some of the most representative SER methods based on the most widely used benchmark datasets, including CNN, attention-based, transformer, and graph-based methods [26–30]. The reported results show that there are variations in SER performance between the different datasets and architectures. However, direct comparison requires caution because of differences in datasets, emotion classes, and evaluation protocols. The proposed PGAT further investigates speaker-dependent and speaker-independent recognition in low-resource multilingual speech datasets.

Table 3. Reported SER accuracy on established speech emotion datasets.

Dataset	Method	Year	Emotion Classes	Overall Accuracy (%)	Reference
RAVDESS [31]	Temporal Bucketing SER	2025	8	91.11	[28]
RAVDESS [31]	CTENet (CNN + Multi-head Convolutional Transformer)	2023	8	82.31	[26]
IEMOCAP [32]	Temporal Bucketing SER	2025	5	89.36	[28]
IEMOCAP [32]	CTENet (CNN + Multi-head Convolutional Transformer)	2023	5	79.42	[26]
Emo-DB [33]	CNN-LSTM with Attention	2023	7	96.72	[27]
Emo-DB [33]	Temporal Bucketing SER	2025	7	95.14	[28]
SAVEE [34]	CNN-LSTM with Attention	2023	7	97.13	[27]
SAVEE [34]	Temporal Bucketing SER	2025	7	92.91	[28]
RAVDESS [31]	Enhanced Graph Attention + Bi-LSTM	2026	8	85.7	[29]
RAVDESS [31]	Attention-Based Multi-Learning	2022	8	85.89	[30]
SAVEE [34]	Attention-Based Multi-Learning	2022	7	93.75	[30]
Emo-DB [33]	Attention-Based Multi-Learning	2022	7	95.93	[30]

In addition to the representative SER methods presented in Table 3, several studies have investigated graph-based learning for speech emotion recognition. Li et al. [1] proposed a graph-LSTM method and reported weighted accuracy (WA) of 71.83% and unweighted accuracy (UA) of 65.39% on IEMOCAP. Wang and Kim [3] introduced SkipGCNGAT, which achieved WA and UA values of 67.37% and 65.61%, respectively. Similarly, Yan et al. [4] proposed TLGCNN, reporting WA of 66.82% and UA of 64.21% on the same dataset. These studies demonstrate the application of graph-based architectures to model relationships among speech representations for emotion recognition. The proposed PGAT adopts a different graph construction strategy by representing MFCC coefficient indices as graph nodes and integrating acoustic-statistical descriptors with graph attention learning. It achieves 91.3% speaker-dependent (SD) accuracy and 90.1% speaker-independent (SI) accuracy on SantaliESC. However, these results should not be interpreted as a direct performance

comparison with the IEMOCAP studies because the datasets, evaluation protocols, and reported metrics differ.

The difference between human perceptual accuracy (74.6%) and the classification performance of the proposed PGAT model (91.3% under speaker-dependent evaluation and 90.1% under speaker-independent evaluation) deserves further consideration. Human perception of vocal emotion is inherently subjective and can vary across listeners, particularly when emotional categories have overlapping acoustic characteristics or when expressions are relatively subtle. The speaker-dependent and independent cases are explored along with data augmentation techniques used in machine learning models, which are quite different from the human studies. Previous research has also found discrepancies in automatic emotion classification and human perception. For instance, the accuracy in recognizing five emotions reported by Petrushin [35] was approximately 49.5% to 72.2%. Jeon et al. [36] showed that the recognition patterns of human perception and automatic emotion classification can be different in different languages and speech corpora. Likewise, earlier studies of non-acted speech have shown significant interlistener agreement differences across emotion categories, and machine classification results varied as a function of the interlistener agreement and evaluation protocol [37]. The results show that the accuracy of human perception is not necessarily the upper limit for automated emotion recognition. The speaker-independent result of 90.1% (91.3% speaker-dependent) provides further evidence of the model's ability to generalise to unseen speakers. However, this result alone does not rule out annotation-related or dataset-specific effects. The results of human perceptual classification accuracy and the automated classification accuracy should thus be interpreted with caution. To achieve reliability and generalisability, the approach could be evaluated further with other speaker-independent datasets, consensus-based labels and independent annotators. Recent research has also indicated problems in the perception of more nuanced and ambiguous emotions or behaviours in fully automated systems. Rasool et al. [38] showed that the use of human annotations helped the AI-based behavioural classification, notably in the case of ambiguous cases and subtle emotions. This observation further supports the notion of perceptual uncertainty in the recognition of emotions. The human perceptual accuracy of 74.6% reported for SantaliESC [5] indicates that emotion interpretation remains challenging even for human listeners. The proposed PGAT partially addresses this challenge through relational learning, which models complementary relationships among MFCC-derived acoustic features rather than treating each feature in isolation.

5. Conclusions

This study presents a perceptual graph attention network (PGAT) for speech emotion recognition using statistical representations derived from static MFCC, delta-MFCC, and delta-delta-MFCC features. The proposed framework represents the 13 MFCC coefficient indices as graph nodes, each associated with a 30-dimensional statistical feature vector. By integrating coefficient-adjacency and acoustic-similarity relationships, the graph attention mechanism learns complementary dependencies among acoustic representations for five-class emotion classification. The experimental findings demonstrate the effectiveness of PGAT across SantaliESC, BanglaSER, and IITKGP-SEHSC. The model achieved speaker-dependent and speaker-independent accuracies of 91.3% and 90.1%, respectively, on SantaliESC. The corresponding accuracies were 90.7% and 89.2% on BanglaSER and 90.0% and 88.1% on IITKGP-SEHSC. PGAT outperformed the evaluated SVM, k-NN, MLP, and CNN baselines across all three datasets. The graph construction analysis further established the benefit of combining coefficient-adjacency and acoustic-similarity connections. The combined graph achieved accuracies of 91.3%, 90.7%, and 90.0% across the

three datasets, respectively, compared with lower performance from either connection type used independently.

The descriptor sensitivity analysis identified distribution-shape descriptors as particularly influential. Occluding skewness and kurtosis produced the largest reduction among the descriptor groups, lowering the reference accuracy from 91.30% to 42.50%. Class-wise evaluation showed relatively consistent recognition across the five emotion classes, although some confusion remained between Surprise and Sad. The graph-size analysis showed that the maximum accuracy of 91.3% was achieved with the 13-node configuration at an average inference time of around 5.55 ms per recording. Additionally, evaluation on the combined multilingual dataset yielded a lower speaker-dependent accuracy of 87.2%, reflecting the additional variability across languages and recording conditions. Notwithstanding these positive results, there are some limitations in the study. The evaluation only covers three speech emotion corpora and five common emotion categories, and this study therefore does not establish generalisation to unseen languages or unrestricted natural conversational speech. Cross-dataset performance may also vary due to differences in the corpora, speakers, recording conditions, and emotion annotation protocols. In addition, the descriptor sensitivity experiment occludes descriptor groups while keeping the trained model weights fixed, rather than retraining the model without each group; its findings should therefore be interpreted as a sensitivity analysis rather than as evidence of feature necessity. The reported inference time is dependent on the experimental workstation and may also differ depending on the hardware platform.

The proposed framework will be explored with more extensive and varied multilingual speech corpora, with natural conversational speech recordings and more rigorous cross-corpus and unseen-speaker evaluation procedures. Independent annotators and consensus-based emotion labels will also be considered to better account for perceptual uncertainty. Finally, model compression and computational optimization will be explored to assess the feasibility of deploying graph-based speech emotion recognition in real-time and resource-constrained applications.

Author Contributions: Conceptualization, T.K.D. and S.K.M.; methodology, T.K.D. and S.K.M.; software, J.S. and S.K.M.; validation, T.K.D. and S.K.M.; formal analysis, J.S. and S.K.M.; investigation, J.S. and G.P.; writing—original draft preparation, J.S. and S.K.M.; writing—review and editing, T.K.D. and S.K.M.; supervision, T.K.D. and G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the Telecom Centre of Excellence India under the Telecom Technology Development Fund (TTDF), proposal number TTDF/6G/352.

Institutional Review Board Statement: This study used previously collected speech emotion datasets, namely, SantaliESC, BanglaSER, and IITKGP-SEHSC, under their respective access and licensing terms. SantaliESC was developed by the authors' research group, while BanglaSER and IITKGP-SEHSC were obtained from their respective sources. No new data were collected from human participants for the present study.

Informed Consent Statement: This study did not involve new human participants; informed consent for the original recordings was obtained by the creators of the respective datasets.

Data Availability Statement: This study used three publicly available speech emotion datasets: SantaliESC (https://doi.org/10.5281/zenodo.20702265, accessed on 20 June 2026), BanglaSER (https://data.mendeley.com/datasets/t9h6p943xy/5, accessed on 20 June 2026), and IITKGP-SEHSC (https://doi.org/10.1109/ICDECOM.2011.5738540, available from IIT Kharagpur on request under its research database sharing guidelines). The datasets are described in Section 3.1 and were used under their respective access and licensing terms. No new data were collected from human or animal subjects.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Y.; Wang, Y.; Yang, X.; Im, S.-K. Speech Emotion Recognition Based on Graph-LSTM Neural Network. *EURASIP J. Audio Speech Music Process.* **2023**, *2023*, 40. [CrossRef]
2. Pentari, A.; Kafentzis, G.; Tsiknakis, M. Speech Emotion Recognition via Graph-Based Representations. *Sci. Rep.* **2024**, *14*, 4484. [CrossRef] [PubMed]
3. Wang, H.; Kim, D.-H. Graph Neural Network-Based Speech Emotion Recognition: A Fusion of Skip Graph Convolutional Networks and Graph Attention Networks. *Electronics* **2024**, *13*, 4208. [CrossRef]
4. Yan, J.; Li, H.; Xu, F.; Zhou, X.; Liu, Y.; Yang, Y. Speech Emotion Recognition Based on Temporal-Spatial Learnable Graph Convolutional Neural Network. *Electronics* **2024**, *13*, 2010. [CrossRef]
5. Dash, T.K.; Satapathy, J.; Sarangi, S.; Mahapatra, S. *Santali Emotional Speech Corpus (SantaliESC)*, Version 1; Zenodo: Geneva, Switzerland, 2026. [CrossRef]
6. Dar, G.H.M.; Delhibabu, R. Speech Databases, Speech Features, and Classifiers in Speech Emotion Recognition: A Review. *IEEE Access* **2024**, *12*, 151122–151152. [CrossRef]
7. Madanian, S.; Chen, T.; Adeleye, O.; Templeton, J.M.; Poellabauer, C.; Parry, D.; Schneider, S.L. Speech Emotion Recognition Using Machine Learning—A Systematic Review. *Intell. Syst. Appl.* **2023**, *20*, 200266. [CrossRef]
8. Nagarajan, S.; Nettimi, S.S.S.; Kumar, L.S.; Nath, M.K.; Kanhe, A. Speech Emotion Recognition Using Cepstral Features Extracted with Novel Triangular Filter Banks Based on Bark and ERB Frequency Scales. *Digit. Signal Process.* **2020**, *104*, 102763. [CrossRef]
9. Hashem, A.; Arif, M.; AlGhamdi, M. Speech Emotion Recognition Approaches: A Systematic Review. *Speech Commun.* **2023**, *154*, 102974. [CrossRef]
10. Barhoumi, C.; BenAyed, Y. Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation. *Artif. Intell. Rev.* **2025**, *58*, 49. [CrossRef]
11. Atmaja, B.T.; Sasou, A. Effects of Data Augmentations on Speech Emotion Recognition. *Sensors* **2022**, *22*, 5941. [CrossRef] [PubMed]
12. Roger, V.; Farinas, J.; Pinquier, J. Deep Neural Networks for Automatic Speech Processing: A Survey from Large Corpora to Limited Data. *EURASIP J. Audio Speech Music Process.* **2022**, *2022*, 19. [CrossRef]
13. Qi, X.; Wen, Y.; Zhang, P.; Huang, H. MFGCN: Multimodal Fusion Graph Convolutional Network for Speech Emotion Recognition. *Neurocomputing* **2025**, *611*, 128646. [CrossRef]
14. Lian, H.; Lu, C.; Chang, H.; Zhao, Y.; Li, S.; Li, Y.; Zong, Y. AMGCN: An Adaptive Multi-Graph Convolutional Network for Speech Emotion Recognition. *Speech Commun.* **2025**, *168*, 103184. [CrossRef]
15. Liu, J.; Wang, H.; Sun, M.; Wei, Y. Graph Based Emotion Recognition with Attention Pooling for Variable-Length Utterances. *Neurocomputing* **2022**, *496*, 46–55. [CrossRef]
16. Khaertdinov, B.; Jeruis, P.; Sousa, A.; Hortal, E. Exploring Self-Supervised Multi-view Contrastive Learning for Speech Emotion Recognition with Limited Annotations. In *Proceedings of the Annual Conference of the International Speech Communication Association*; International Speech Communication Association: Grenoble, France, 2024; pp. 4708–4712. [CrossRef]
17. Rathnayake, H.; James, J.; Leoni, G.; Nicholas, A.; Watson, C.I.; Keegan, P. A Review on Speech Emotion Recognition for Low-Resource and Indigenous Languages. *Speech Commun.* **2026**, *176*, 103342. [CrossRef]
18. Das, R.K.; Islam, N.; Ahmed, M.R.; Islam, S.; Shatabda, S.; Islam, A.K.M.M. BanglaSER: A Speech Emotion Recognition Dataset for the Bangla Language. *Data Brief.* **2022**, *42*, 108091. [CrossRef] [PubMed]
19. Koolagudi, S.G.; Reddy, R.; Yadav, J.; Rao, K.S. IITKGP-SEHSC: Hindi Speech Corpus for Emotion Analysis. In Proceedings of the International Conference on Devices and Communications (ICDeCom), Mesra, India, 24–25 February 2011. [CrossRef]
20. Salamon, J.; Bello, J.P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [CrossRef]
21. Davis, S.B.; Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [CrossRef]
22. Abdul, Z.K.; Al-Talabani, A.K. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access* **2022**, *10*, 122136–122158. [CrossRef]
23. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
24. Yang, F.; Zamzmi, G.; Angara, S.; Rajaraman, S.; Aquilina, A.; Xue, Z.; Jaeger, S.; Papagiannakis, E.; Antani, S.K. Assessing Inter-Annotator Agreement for Medical Image Segmentation. *IEEE Access* **2023**, *11*, 21300–21312. [CrossRef] [PubMed]
25. Weninger, F.; Eyben, F.; Schuller, B.W.; Mortillaro, M.; Scherer, K.R. On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. *Front. Psychol.* **2013**, *4*, 292. [CrossRef] [PubMed]
26. Ullah, R.; Asif, M.; Shah, W.A.; Anjam, F.; Ullah, I.; Khurshaid, T.; Wuttisittikulkij, L.; Shah, S.; Ali, S.M.; Alibakhshikenari, M. Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. *Sensors* **2023**, *23*, 6212. [CrossRef] [PubMed]

27. Alsabhan, W. Human-Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention. *Sensors* **2023**, *23*, 1386. [CrossRef] [PubMed]
28. Gurowiec, I.; Nissim, N. Improving speech emotion recognition capabilities in the short and long term using temporal bucketing and active learning. *Comput. Biol. Med.* **2025**, *196*, 110912. [CrossRef] [PubMed]
29. Dong, W.; Dong, M.; Zhao, Y.; Pan, S. Enhanced graph attention network by integrating Long Short-Term Memory for artificial emotion representation in multi-modality datasets. *PLoS ONE* **2026**, *21*, e0339946. [CrossRef] [PubMed]
30. Kakuba, S.; Poulose, A.; Han, D.S. Attention-Based Multi-Learning Approach for Speech Emotion Recognition with Dilated Convolution. *IEEE Access* **2022**, *10*, 122302–122313. [CrossRef]
31. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
32. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
33. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In *Proceedings of the Annual Conference of the International Speech Communication Association*; International Speech Communication Association: Grenoble, France, 2005; pp. 1517–1520. [CrossRef]
34. Jackson, P.; Haq, S. Surrey Audio-Visual Expressed Emotion (SAVEE) Database. Available online: https://kahlan.eps.surrey.ac.uk/savee (accessed on 20 June 2026).
35. Petrushin, V.A. Emotion recognition in speech signal: Experimental study, development, and application. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 16–20 October 2000*; ISCA International Speech Communication Association: Grenoble, France, 2001; Volume 2, pp. 222–225. [CrossRef]
36. Jeon, J.H.; Le, D.; Xia, R.; Liu, Y. A preliminary study of cross-lingual emotion recognition from speech: Automatic classification versus human perception. In *Proceedings of the Annual Conference of the International Speech Communication Association*; International Speech Communication Association: Grenoble, France, 2013; pp. 2837–2840. [CrossRef]
37. Deshpande, G.; Viraraghavan, V.S.; Duggirala, M.; Vempada, R.R.; Patel, S. Comparing manual and machine annotations of emotions in non-acted speech. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 4241–4244. [CrossRef] [PubMed]
38. Rasool, A.; Dunlap, K.; Husic, A.; Kline, A.; Azizian, P.; Wall, D.P.; Washington, P. Balancing Human and AI Features for Digital Autism Diagnostics from Home Videos: Human-In-The-Loop Machine Learning Study. *JMIR AI* **2026**, *5*, e84699. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.