IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 October 2022, accepted 24 November 2022, date of publication 30 November 2022,
date of current version 6 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3225684

RESEARCH ARTICLE

Deep Learning-Based Speech Emotion Recognition Using Multi-Level Fusion of Concurrent Features

SAMUEL KAKUBA[1,2], ALWIN POULOSE[3], AND DONG SEOG HAN[4], (Senior Member, IEEE)

[1]Graduate School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, South Korea
[2]Faculty of Engineering, Technology, Applied Design and Fine Art, Kabale University, Kabale, Uganda
[3]Department of Electrical and Computer Engineering, University of Michigan, Dearborn, MI 48128, USA
[4]School of Electronics Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding author: Dong Seog Han (dshan@knu.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) Funded by the Ministry of Education under Grant 2021R1A6A1A03043144.

ABSTRACT The detection and classification of emotional states in speech involves the analysis of audio signals and text transcriptions. There are complex relationships between the extracted features at different time intervals which ought to be analyzed to infer the emotions in speech. These relationships can be represented as spatial, temporal and semantic tendency features. In addition to emotional features that exist in each modality, the text modality consists of semantic and grammatical tendencies in the uttered sentences. Spatial and temporal features have been extracted sequentially in deep learning-based models using convolutional neural networks (CNN) followed by recurrent neural networks (RNN) which may not only be weak at the detection of the separate spatial-temporal feature representations but also the semantic tendencies in speech. In this paper, we propose a deep learning-based model named concurrent spatial-temporal and grammatical (CoSTGA) model that concurrently learns spatial, temporal and semantic representations in the local feature learning block (LFLB) which are fused as a latent vector to form an input to the global feature learning block (GFLB). We also investigate the performance of multi-level feature fusion compared to single-level fusion using the multi-level transformer encoder model (MLTED) that we also propose in this paper. The proposed CoSTGA model uses multi-level fusion first at the LFLB level where similar features (spatial or temporal) are separately extracted from a modality and secondly at the GFLB level where the spatial-temporal features are fused with the semantic tendency features. The proposed CoSTGA model uses a combination of dilated causal convolutions (DCC), bidirectional long short-term memory (BiLSTM), transformer encoders (TE), multi-head and self-attention mechanisms. Acoustic and lexical features were extracted from the interactive emotional dyadic motion capture (IEMOCAP) dataset. The proposed model achieves 75.50% and 75.82% of weighted and unweighted accuracy, 75.32% and 75.57% of recall and F1 score respectively. These results imply that concurrently learned spatial-temporal features with semantic tendencies learned in a multi-level approach improve the model's effectiveness and robustness.

INDEX TERMS Emotion recognition, spatial features, temporal features, semantic tendency features, multi-head attention.

I. INTRODUCTION

The study of affective computing involves machine learning techniques that can detect, analyze, and predict human emotional states and use them to infer intentions and behaviors. The mood portrayed in one's behavior contributes greatly to the person's intentions. Proper analysis of human emotions allows the intelligent agents used in human-computer interaction (HCI) and human-to-robot interaction (HRI) to mimic human characteristics like empathy, care,

The associate editor coordinating the review of this manuscript and approving it for publication was Kah Phooi (Jasmine) Seng.

and remorse which allow them to react according to human sentiments. Affective computing systems can be applied in a number of areas some of which are; social assistive living, health diagnosis, care and monitoring, fraud detection, home care systems, etc. They can be used to detect skepticism, satisfaction, stress, and frustration which can point to the next course of action in human welfare automated systems. Emotions can be classified according to discrete categories or emotional dimensions. Ekman et al. [1] proposed six discrete categories of emotions that are universal to all human beings. These are happiness, sadness, surprise, anger, disgust, and fear. Because humans may not express any of these emotions at all times, it is a common practice to include neutral as the seventh category. In terms of dimensions, Tsiourti et al. [2] divided the discrete categories of emotions in terms of a two-dimensional plane of valence and arousal. Verma and Tiwary [3] further proposed to include dominance in the two-dimensional emotion space and analyze the emotions in a three-dimensional continuous space.

Generally, emotional analysis is categorized according to the source of data being used. They can be categorized as physiological, lexical, facial, acoustic, or multi-modal emotional analysis if it includes more than one modality. Speech emotion recognition (SER) involves the analysis of features extracted from a varying time speech signal (sometimes in combination with their transcriptions) with the intent of classifying the emotional state of the utterances therein. SER has recently attracted researchers because of the need for robust and reliable systems that can aid interaction between intelligent devices and humans for different activities. The study of emotions in speech does not only consider utterances at a single instance but all the instances in a sequence. This is because psychologically human emotions are perceived from the previous, present and future utterances [4]. Moreover, the emotions in each utterance are triggered by context cues [5] making it so important to consider the context by emotion recognition models. Recurrent neural networks (RNN) like long short-term memory (LSTM) [6] and gated recurrent units (GRU) are often used in combination with attention mechanisms to keep track of long-term dependencies between the features. Memory networks, graph networks, and convolutional neural networks (CNN) are the other deep learning technologies often used sequentially with RNNs in SER literature. However, the sequential consideration of these deep learning techniques leads to weak learning of either the temporal or spatial cues depending on which one comes after the other in the model sequence. Spatial and temporal emotion information occurs concurrently [7] and therefore should be learned concurrently instead of learning the spatial features in the local feature learning block (LFLB) and then temporal features in the global feature learning block (GFLB) or vice versa. In [7], the electroencephalography (EEG) signal's temporal and spatial feature representations are learned and integrated into a unified spatial-temporal dependency model. Grammatical and semantic features in combination with emotional features from all the modalities of emotion recognition improve the recognition performance [8]. We therefore propose a model that concurrently learns spatial, temporal and semantic tendency features for effective and robust SER.

The contribution of this paper is threefold;

- We investigate the impact and significance of multi-level fusion of intra and cross-modality speech feature representations compared to single-level fusion using the multi-level fusion transformer encoder (MLTED) model that we propose in this paper.
- We also propose a deep learning-based concurrent spatial-temporal and grammatical (CoSTGA) model that concurrently learns spatial, temporal and semantic tendency feature representations in the LFLB which are fused as a latent vector to form an input to the GFLB for SER.
- In addition to the evaluation of the proposed model's sub-modules, its performance is also compared with the existing approaches for audio and text bimodal SER.

The rest of the paper is organized as follows: the related work is presented in Section II. The methods used in this paper are discussed in Section III. The experimental evaluation is described in Section IV. Section V reports the results and presents the discussion. A conclusion is drawn in Section VI.

II. RELATED WORK

SER research has been carried out using the acoustic modality alone to enhance improvement in performance. Recently, Yan et al. proposed a model that uses CNN with bidirectional gated recurrent networks (BiGRU) and the attention mechanism to classify emotions from extracted spectrograms and their first and second delta derivatives using the interactive emotional dyadic motion capture (IEMOCAP) dataset [9]. An effective and robust model for acoustic SER was also proposed in [10]. In [11], a late fusion-based model was also proposed for SER. This model uses two branches with the convolutional capsule network (CCN) branch taking in mel spectrograms as input and the BiGRU taking a combination of mel frequency cepstral coefficients (MFCCs), chroma grams, spectral contrast, zero-crossing rate (ZCR), and root mean square energy (RMSE). This model uses double attention mechanism before the late fusion of the extracted features. Xu et al. [12] also proposed a single fusion model based on CNN and multi-head attention to classify emotions according to the extracted MFFCs and tested its performance on the IEMOCAP dataset. In addition, emotion recognition models that use lexical features as input have also been proposed. Hu et al. [5] proposed to classify emotions in text using contextual reasoning. Their model uses LSTM as the perception block and a combination of LSTM and attention mechanism for the cognitive block which extracts emotional cues and integrates them iteratively. Recently, Bekmanova et al. [13] suggested recognition of emotions from word transcriptions for students that participated in distance learning examinations.

However, in real-life speech, the emotional state is simultaneously inferred from both acoustic and lexical cues. Moreover, it is not enough to infer speech emotions from a single modality since the cross-modality interaction between acoustic and lexical features is as important as the intra-modality feature interactions for emotion classification. Though SER that involves utterances of more than one interlocutor of similar or different accents, gender and culture is complicated, it is more realistic than one that involves one sided utterances. Each utterance ought to be analyzed in terms of the spatial and temporal cues, their context, semantics, and speaker sensitivity [14]. Moreover, it is argued in [15] that the context of a given utterance in relation to others is what differentiates conversational speech between two or more interlocutors from single sentence emotion recognition. This partly explains why researchers have always configured attention mechanisms in SER. In addition to contextualized utterances in conversational speech, Lian et al. [16] proposed to automatically correct errors made by emotion recognition systems by the use of self and inter-speaker influence considerations through the use of gated graph neural networks (GGNN). It is also argued in [17] that syntactic information is as important as semantic information in conversational speech. They propose syntactic analysis alongside the use of graph convolutional neural networks (GCNN) and attention mechanisms.

The complexity of the SER task especially in conversational speech that involves more than one interlocutor necessitates deep learning models which have to be trained for a long period of time. This usually results in the vanishing gradient problem during training. Most models use RNNs like LSTM [6] to keep track of long-term dependencies between the features as well as solve the vanishing gradient problem. Since LSTM only handles forward long-short term dependencies, bidirectional long-short term memory (BiLSTM) is often a better choice for SER. To further solve the vanishing gradient problem in some cases, skip connections are used [18]. A residual BiLSTM combined with multi-head attention was presented in [10] and a commendable performance was registered for acoustic speech emotion recognition with no over fitting and vanishing gradient problems. In [19], a skip connection between the first dense and convolution layers that concatenates the resultant hidden vectors is proposed to improve SER performance. This is enhanced by a mask layer between the convolution and bidirectional LSTM to extract features more relevant to the emotion state before the attention mechanism layer.

In addition, most of the recent work on deep learning-based SER uses attention-based approaches to solve the SER task. The attention mechanisms have been proposed in [20], [21], and [22]. These mechanisms are used to consider long-term dependencies and computation of the context of a given input with reference to the surrounding inputs in the sequence. Additive [20] and multiplicative [21] attention mechanisms are used in combination with CNNs and/or RNNs and they involve a sequential computation of the context vector. The attention mechanism in [22] called transformer-based multi-head attention operates dynamically and involves a parallel computation to obtain context vectors. They also employ residual connections and layer normalization to better the performance. We are motivated that the careful use of attention mechanisms in combination with other deep learning techniques allows the model to take advantage of the merits of each. Moreover, it is stated in [23] that global attention mechanisms are suitable for SER.

Yoon et al. [24] proposed deep learning models; multimodal dual recurrent encoder with attention (MDREA) and multimodal dual recurrent encoder (MDRE) that use text and acoustic information with and without attention mechanisms respectively using dual recurrent neural networks in a multi-modal approach to classifying emotions in audio and text. Chen and Zhao [25] proposed a multi-modal deep learning model called STSER that takes in text and acoustic features as input. It uses CNN for local feature learning and the BiLSTM for long-term dependencies. The multi-head attention mechanism is used to consider the context of speech and focus on the relevant features for the emotion class. Xu et al. [26] used LSTM and attention mechanisms in a model that predicts emotional states from text and audio information. Among the models proposed in [27] is the IEmoNet (BE) that uses both audio and text information for SER. The IEmoNet (BE) model uses pre-trained language models for the text modality that consists of automatically generated transcriptions and a separately trained audio sub-model. Singh et al. [28] proposed an SER model that uses a hierarchy that follows the binary decision tree structure for bimodal SER after early fusion.

The works in [14], [29], and [30] used transformer-based multi-head attention. Multi-head attention is a variant of the self-attention mechanism that computes attention weights of current inputs in relation to all other inputs in the sequence described as [22]

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where query, Q is $W^q x_i$, key, K is $W^k x_i$ and value, V is $W^v x_i$. The term x_i is a word or feature at position i and d_k is the feature dimension of the query and key. If the self-attention mechanism is applied multiple times in parallel then this is called multi-head attention. In [29], multi-head attention was used to detect emotions in conversational speech with multi-modal fusion of video and audio modalities. They consider the two-dimensional space emotions of arousal and valence for emotion classification. The performance of this model is reported in terms of the concordance correlation coefficient (CCC) on the audio-visual emotion challenge (AVEC) 2016 and 2019 datasets. Ho et al. [30] used multi-head attention mechanism in a two-level attention mechanism to classify emotions in dialogues. They used the acoustic and lexical features as inputs to their model. A pre-trained bidirectional encoder representation from transformers (BERT) model was used to extract lexical

features. The acoustic features extracted from the audio files were MFCCs. They report results on improvised and mixed data from the IEMOCAP dataset for audio and text transcriptions. Lian et al. in two papers [31], and [14] proposed fusion of acoustic, lexical, and speaker features using transformer-based multi-head attention to detect and classify emotions in conversations. They used the transformer models to learn the intra-modal and intermodal characteristic interactions before feeding the obtained latent vectors into a fusion layer together with speaker embeddings. In [14], they further use the gated recurrent unit (GRU) and multi-head attention to learn the context of the features and eventually perform the classification task in a model-level fusion approach.

There are three approaches to multi-modal fusion; early or feature level fusion, model level or intermediate fusion, and decision/late fusion. Early fusion involves the concatenation of features at the input stage. However, the results obtained using this approach are affected by the sparsity of data [32]. Decision-level fusion is applied at the classifier level and ensemble techniques are used to obtain the required values according to the performance metrics used. Model-level fusion involves the fusion of latent representations obtained from different modality models at an intermediate level in order to leverage the advantages of both feature and decision-level fusion. Moreover, early fusion and late fusion prevent models from learning intra and inter-modality interaction dynamics [33], [34] respectively. Multimodal fusion involves the alignment of the features in different modalities explicitly or implicitly [35]. The explicit approach assumes prior alignment of features in order to find interaction characteristics between the elements of different modalities. For implicit, the model learns the alignment of the different modality features progressively as it trains. Kimoto et al. [36] assert that implicit alignment strategy is more naturalistic than explicit alignment. We therefore use implicit alignment in this paper.

Spatial features in the text and acoustic modality are as important as the temporal features in behavioral recognition. Sound sources like interlocutors in an interactive speech at different locations have varying intensities in the binaural channels with different frequency spread in the spectrum. This creates complex spatial dependencies which are overlapped between multi-channels of different interlocutors at dissimilar frequencies. Spatial features can be modeled in reverberate and noisy environments to improve speech recognition accuracy. In terms of the text modality, sentences uttered usually express information about the spatial configuration that ought to be modeled. The spatial configuration of the keywords in an utterance or sentence needs to be learned by the model. It is also worth noting that spatial information in languages plays an important role in semantic understanding. The spatial features are combined with temporal and semantic features to improve classification accuracy. In this work, we modeled the spatial features in the spectral-temporal representations and word embeddings of different interlocutors in interactive speech using the dilated causal convolution (DCC) layers.

The recent work discussed thus far considers spatial and temporal features learned one after the other sequentially in the LFLB and GFLB. It should however be noted that spatial and temporal features in emotions occur concurrently [7]. Therefore, learning one after the other in a sequence may not give an accurate representation of these features. In addition, though all modalities consist of emotional cues pertinent to emotion recognition, the uttered sentences in the text modality consist of grammatical and semantic features [8] which can provide supplementary knowledge to the model for effective and robust SER. In [37], a model was suggested that uses multi-channel convolutional neural networks (MCNNs) to learn emotional and grammatical features from the text. However, as asserted in [8], CNNs and RNNs can weakly extract grammatical and semantic information since they are good at learning spatial and temporal features but not the context of the sequences. We therefore propose a model that concurrently learns spatial, temporal and semantic features in the LFLB and their representations are fused and fed into the GFLB. We use dilated causal convolutions (DCC) for spatial features and BiLSTM for temporal features in combination with attention mechanisms. This model particularly uses transformer-based multi-head attention for grammatical and semantic feature representations.

III. METHODS

The main objective of this paper is to propose a deep learning-based concurrent spatial-temporal and grammatical (CoSTGA) model. The CoSTGA model concurrently learns spatial and temporal features from the audio and text modalities of speech that are fused with the semantic tendency features learned from the uttered sentences at the same time in the LFLB. We first present the multilevel fusion transformer encoder (MLTED) model that we propose and use to investigate the performance of multi-level fusion as compared to single-level fusion in SER studies. Secondly, we discuss the proposed CoSTGA model that uses multi-level fusion learning of spatial, temporal, and semantic features. The proposed CoSTGA model learns spatial and temporal representations concurrently which are fused with the semantic tendency latent vector in the LFLB to form an input to the GFLB. This following subsections presents the different approaches proposed in this paper.

A. THE MULTI-LEVEL FUSION TRANSFORMER ENCODER (MLTED) MODEL

The aim of this model is to investigate the significance of multilevel fusion as compared to single-level fusion in SER studies. As shown in Fig. 1, the multi-level fusion transformer encoder (MLTED) model uses transformer encoders (TE) to learn the lexical and acoustic features which are fused at the first level and the resultant feature representation is fed into multi-head attention layer of two heads. The resultant feature representation from this level is again fused

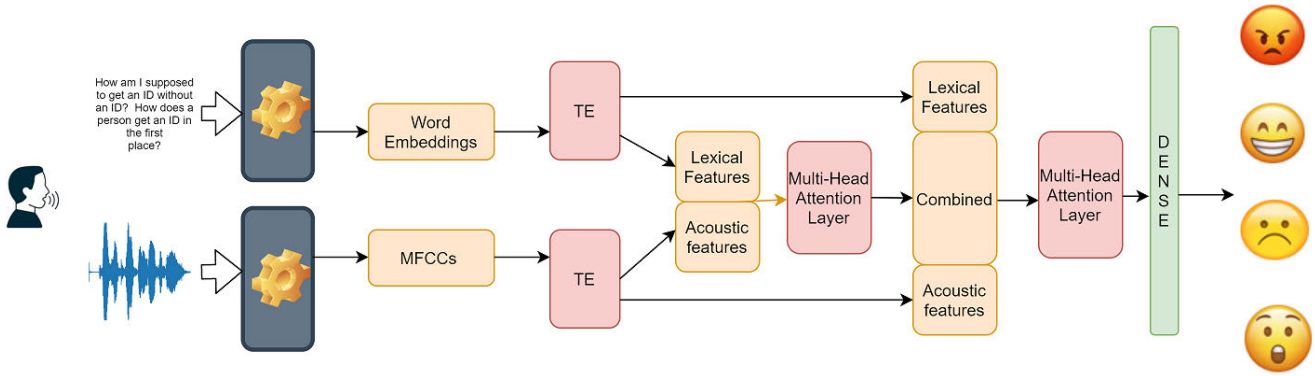


FIGURE 1. Multi-level fusion transformer encoder (MLTED) model.

with the separate lexical and acoustic feature representation from the transformer encoder before being fed into another multi-head attention layer of two heads and later through the fully connected (FC) layer. The softmax activation function is used for emotion classification at all levels without separate consideration of spatial, temporal or explicit semantic tendencies. We used the transformer encoder and multi-head attention layer in this architecture because of the global nature of operation to extract intra and cross-modal interaction characteristics progressively to improve the reliability and robustness of the model. The results from this architecture are compared with results from a model that consists of only the first-level fusion.

The transformer encoder (TE) block consists of four heads which take in inputs from projections done by linear models as shown in Fig. 2. The position encoding is handled by the use of a one-dimensional convolutional layer that extracts the relationship between acoustic features. We used similar position encoding as was used in [22] for lexical features. The encoding result is passed through linear models to obtain the query Q, key K, and value V that would later be used in the scaled dot product computation to obtain similarities between features. We used feed-forward neural networks (FFN) as linear models. It should be noted that in self attention mechanism, every token in an utterance computes the similarity or attention weight in regard to all the other tokens. This implies that each vector of the query is compared in terms of similarity to all the keys of the utterance or sentence. Therefore, the attention of a target word or acoustic feature with respect to the input word is calculated by using the query of the current word Q_2 and the keys K_1, K_2, K_n of all the other words, a matching score normalized with a softmax function is obtained and multiplied as weights with the value V. Each head h_i (self-attention layer) uses different weights W_i^q, W_i^k, W_i^v for each word or acoustic feature U_n and the results are concatenated. It also involves residual connections and layer normalization. The described attention score computations are repeated as many times as the number of attention heads in parallel. Through the use of parallel self-attention layers that obtain similarity by use of a scaled

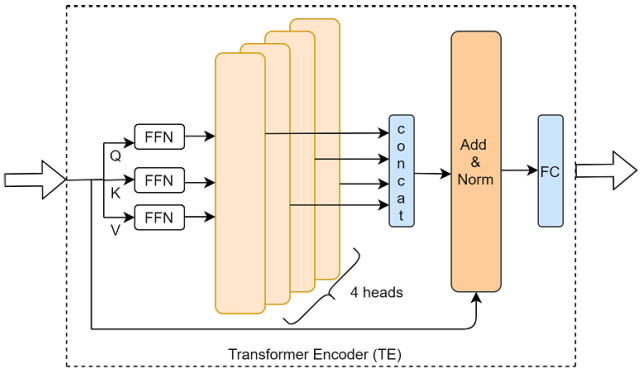


FIGURE 2. Transformer encoder(TE) block of four heads.

dot product, contextualized attention of features in relation to others is performed in a dynamic way. The transformer-based multi-head attention considers the dynamic context of the past, present, and future representations for the analysis of the semantics depicted by the lexical features along with the paralinguistic cues in SER. The transformer encoder operates in parallel which improves its performance, reduces complexity and training time. The output of each transformer encoder *te* is represented as

$$\begin{aligned} \text{Given a conversation file, } U &= U_1, U_2, \ldots.., U_n \\ U_n &= Q_n, K_n, V_n \\ h_i &= Attention(Q_n, K_n, V_n) \\ \text{output } te &= dense(concat(h_i, \ldots., h_z)) \end{aligned} \tag{2}$$

B. THE PROPOSED CoSTGA MODEL

The architecture of the proposed deep learning-based CoSTGA model that uses concurrent spatial, temporal and semantic tendency features is shown in Fig. 3. The proposed end-to-end CoSTGA model learns low-level representations of these features concurrently as opposed to the sequential approach used by the existing models. This model consists of a feature extraction block, the LFLB and the GLFB. The

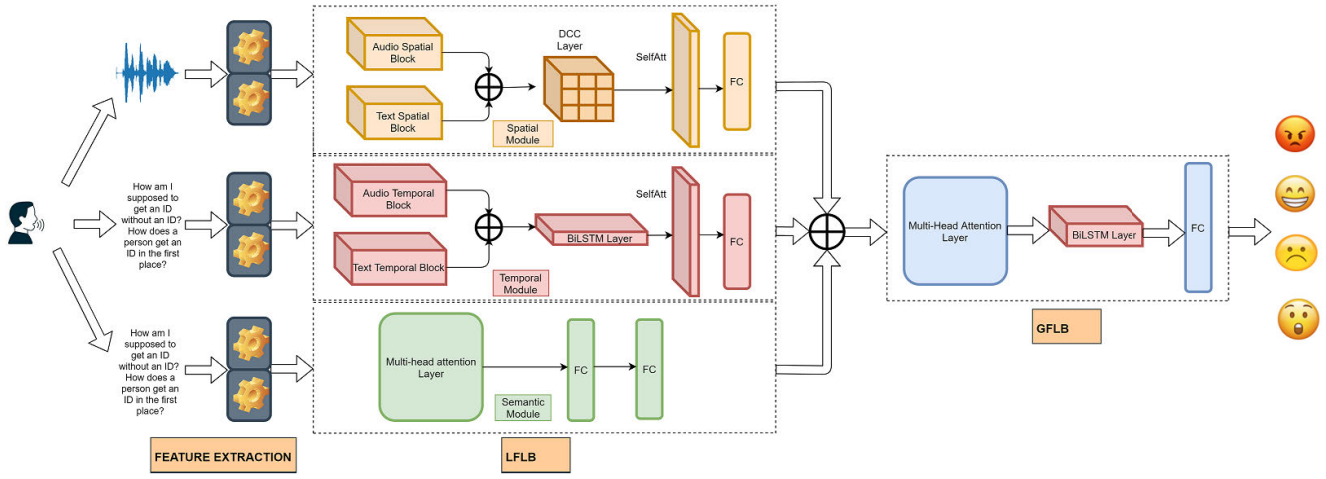


FIGURE 3. Proposed CoSTGA Model that uses concurrent spatial, temporal and semantic features.

feature extraction block is used to pre-process and extract acoustic and lexical features from each modality as explained later in Section IV. The LFLB consists of three modules that are used to concurrently learn the low-level spatial, temporal and semantic tendency features. The spatial and temporal modules are used by the model to concurrently learn low-level spatial and temporal features of each modality respectively. The semantics module is used to learn the grammatical and semantic tendencies of the uttered sentences as in [8]. The resultant vector is fused at the second level to form inputs to the GFLB that is responsible for learning the high-level feature representations. The GFLB consists of a multi-head attention layer of two heads and a BiLSTM layer of 256 units whose output is fed into a dense layer and softmax layer for emotional classification. We chose to use multi-head attention in combination with BiLSTM layers in the GFLB to allow the model to learn globally contextualized high-level features of the combined feature representations before classification. In addition to Fig. 3, we describe the proposed CoSTGA model in Algorithm 1. The algorithm shows that given an utterance U that contains an audio signal a and its transcription text t, the model's goal is to extract features X_1 and X_2, analyze and learn from them the spatial features S_f, temporal features T_f and the semantic tendency features G_f of the uttered sentences at the LFLB. The resultant vectors are fused and fed into the GFLB to allow the model to learn the global feature representations ST that are later used by the softmax function to recognize one of the emotions in the set $y = (happy, angry, sad, neutral)$. In the subsequent sections, we describe the spatial, temporal and semantic modules in detail.

Algorithm 1 The proposed Deep Learning-Based CoSTGA SER Model

Given an utterance input: audio a, text t and output: $y = hap,ang,sad, neu$
for epoch in epochs:
  Word embeddings $X_1 = \text{BERT}(t)$
  MFCCs $X_2 = \text{librosa}(a)$
  $X_2 = \text{mean}\ (X_2)$
  for LFLB in CoSTGA:
    a) $S_t = \text{spatial features}(X_1)$
    b) $S_a = \text{spatial features}(X_2)$
    c) $T_t = \text{temporal features}(X_1)$
    d) $T_a = \text{temporal features}(X_2)$
    e) $G_f = \text{semantic features}(X_1)$
    f) $S_f = \text{fuse}\ S_t \text{ and } S_a$
    g) $T_f = \text{fuse}\ T_t \text{ and } T_a$
    h) $S_f = \text{DCC}(f)$
    i) $S_f = \text{selfAtt}(h)$
    j) $S_f = \text{Dense}(i)$
    k) $T_f = \text{BiLSTM}(g)$
    l) $T_f = \text{selfAtt}(k)$
    m) $T_f = \text{Dense}(l)$
    n) $ST = \text{fuse}(\ G_f, S_f, T_f)$
  end for
  for GFLB in CoSTGA:
    o) $ST = \text{multi-head attention layer}(ST)$, number of heads $= 2$
    p) $ST = \text{BiLSTM layer}(ST)$, units $= 256$
    q) $ST = \text{Dense}(ST)$, units $= 128$
    Output: $y = \text{Softmax}(ST)$
  end for
end for

1) THE SPATIAL MODULE

The spatial module involves learning spatial features that exist in the audio and text modalities separately and later concatenated at the first level fusion. We argue that this preliminary learning process allows the model to learn the intra-modality feature interaction characteristics of audio and text transcription files before being concatenated at the second level to learn the cross-modality feature representations. The resultant feature representation is fed into a DCC layer and self-attention mechanism before a fully connected (FC) layer that resizes the feature vector to be able to be fused with other modality feature representations. In this module, we use the DCC in combination with the transformer encoder (TE) and sequential attention as shown in Fig. 4 to form the

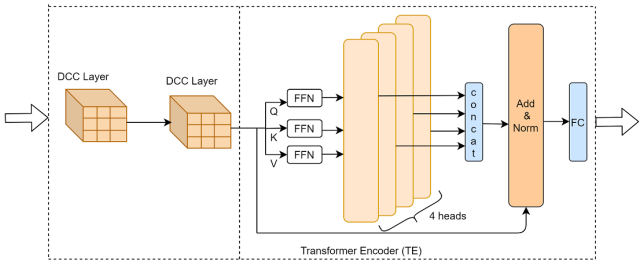


FIGURE 4. Audio or text spatial block used in the spatial module of the proposed CoSTGA model.

first part of the spatial module that we call either audio or text spatial block depending on the modality. We particularly use DCC layers in the proposed model to learn the spatial feature representations while covering a large receptive field. The dilated convolutions spread causal filters by skipping values in the input sequence in specified predetermined steps. It should be noted that dilated convolution layers provide a large receptive field with a few layers which allow faster convergence with minimal use of resources. We gradually increase the receptive fields by increasing the dilation rates from 1 to 4 for the layers considered in the architecture. The DCC allows filters to be applied over a large receptive area without an increase in the kernel inputs and the number of parameters which improves spatial feature learning. After the DCC layers, the resultant representation is fed into a transformer encoder similar to the one discussed earlier. The resultant representations from each modality then undergoes first-level fusion with the other to form cross-modality spatial feature representations that are fed into another DCC layer and a self-attention layer before the dense layer that ensures a resultant vector similar in size to the output of the other two modules. We chose to use the self-attention mechanism at this stage to benefit the model with its advantages including computing a global context of the inputs as discussed earlier but with fewer parameters as compared to multi-head attention.

2) THE TEMPORAL MODULE

The temporal module consists of BiLSTM and self-attention layers used to learn the intra-modality temporal feature relationships before first-level fusion. After the first level fusion, we configured BiLSTM and self-attention layers again to pay attention to the global context of the cross-modality features before feeding them into a dense layer that ensures the same size as the outputs of the other modules to be fused with before the GFLB. The BiLSTM layers used in this module consist of 256 units each. The operation of the BiLSTM layer involves sequential updates of the cell and hidden states in the forward and backward directions to ensure vivid long-term dependencies. The BiLSTM's final hidden state is fed into the self-attention mechanism and later the dense layer. BiLSTM helps in solving the vanishing gradient problem that would otherwise occur. Fig. 5 shows the

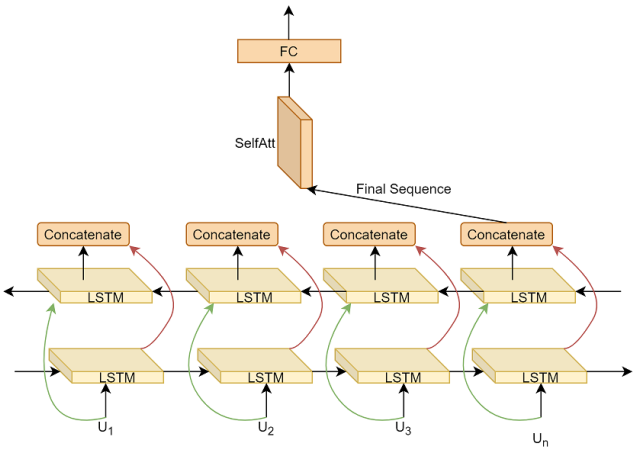


FIGURE 5. Audio or text temporal block used in the temporal module.

temporal audio or text temporal blocks used in the proposed CoSTGA model to learn the intra and cross-modality features before and after first-level fusion respectively. The output of the temporal blocks for audio or text modalities before or after first-level concatenation is represented as

$$\begin{aligned} \text{Given a conversation file, } U &= U_1, U_2, \ldots.., U_n \\ s^i &= U_n \\ f_n^i &= (\overrightarrow{h_n}, \overleftarrow{h_n}) \\ \text{block output, } h_2(s) &= selfAtt(f_n^i) \end{aligned} \tag{3}$$

The speech file U consists of utterances U_n for a given length. $\overrightarrow{h_n}$ is the forward hidden state sequence and $\overleftarrow{h_n}$ is the backward hidden state sequence of the bidirectional LSTM f_n^i. The term i is the ith layer, n is the time in a sequence. The utterances in the speech consists of acoustic and lexical clues which are separately at a given instance.

The self-attention mechanism is particularly used in this block to learn the global context vector of the features in the individual lexical and acoustic modalities separately before being fused to learn the cross-modality features. The input sequence is encoded into vectors and a subset of these that represent the most relevant feature representations is chosen. The cross-modality temporal feature representations are learned using a similar architecture that consists of the BiLSTM, self-attention and dense layers. The dense layer ensures the same size that can be fused with other outputs of the other modules at the next fusion level.

3) THE SEMANTICS MODULE

Though both the audio and text modalities consist of emotional factors and features that provide clues about the human emotional state, the text modality includes the grammatical and semantics cues [8] of the spoken sentence that can avail further knowledge about the speech utterances. Due to this fact, we model the grammar and semantics of the spoken sentences using the multi-head attention layer that consists of four self-attention heads. The resultant

vector is fed into two dense layers of 768 and 128 units. We call the feature representations obtained from this module semantic tendency feature representations since they give the model a clue of the emotional intention from the semantics perspective. The multi-head attention layer that we used to learn these feature representations takes the whole sentence and a global context of each word in terms of another within the sentence is computed using equation 1. The output of this module is fused with the spatial and temporal feature representations at the second level fusion and input into the GFLB to further learn high-level feature representations and subsequently predict the emotion state.

IV. EXPERIMENTAL EVALUATION

In this section, we present the dataset, features and experiments carried out to investigate multi-level fusion and evaluate the performance of the proposed CoSTGA model. To carry out the experiments we used Librosa 0.9.2, with python programming, Keras 2.8.0 Application Programming Interface (API), and Tensorflow 2.6 backend. The Nvidia GeForce RTX 2080 super graphics processing unit (GPU) was used. An initial learning rate of 0.0001, a batch size of 32, and the Adam optimizer were used for all the experiments. The cross-entropy loss was used as the objective function. The performance metrics used in this paper are computed using the Sci-kit-learn toolbox with modifications where need be.

A. DATASET

The IEMOCAP [38], which was collected at the University of Southern California as a multi-modal and multi-speaker emotion recognition database was used in this paper. It is a dyadic database that contains audio and video data with transcriptions in addition to motion capture recordings. It consists of five sessions of dialogues that are improvised and/or scripted to depict discrete emotions annotated by more than three experienced evaluators. It consists of happy, angry, neutral, sad, frustrated, fearful, excited, disgusted, and surprised with another category named other. For each instance of the dialogue to be labeled by the evaluators, the data was partitioned into 3 to 5 seconds length utterances. They also classified the data obtained in terms of emotional dimensional spaces of valence, activation, and dominance. The database consists of the dialog and sentence recordings for about 12 hours.

In this paper, we chose to use mixed (improvised and scripted files) audio dialog (conversation) files, and transcriptions. Consistent with prior research we chose to evaluate the performance of the model on four discrete emotions of anger, sadness, excitement, and neutral. These were also chosen because of their seemingly balanced nature except for neutral and substantial class size as compared to the other emotions. As shown in Fig. 6, the data distribution included; 1103 samples for angry, 1084 for sad, 1041 for excited, and 1708 for neutral which is a total of 4936 samples. The 4936 samples were apportioned in a ratio of 80% for training,

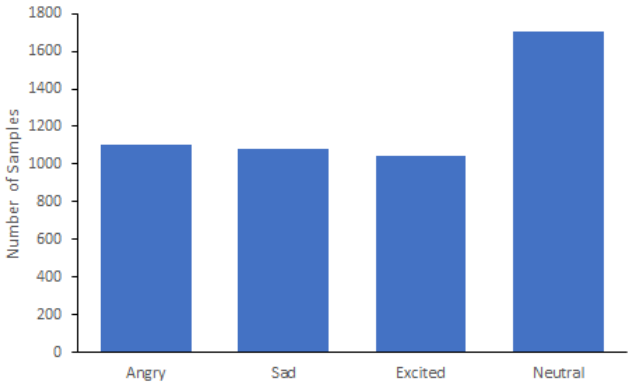


FIGURE 6. Data distribution of the considered discrete emotions.

10% for validation, and 10% for testing. To alleviate the class imbalance problem, we configured class weights depending on the number of samples per emotion category in the Keras model.

B. FEATURE EXTRACTION

We followed the procedure in [39] to pre-process the data from the IEMOCAP dataset. The details of the acoustic and lexical features used are described below.

1) ACOUSTIC FEATURES

We used a sampling rate of 16 kHz since the dataset was collected at the same. A maximum length of 100 frames which is about 10 seconds of the acoustic signal was chosen. To ensure equal sequence length, the speech signal was zero-padded if the audio file was shorter and the longer ones were truncated to have fixed-length inputs. We also removed the silent regions since they do not carry any emotional cues. A Pre-emphasis filter was used to allow only the frequencies of the speech signal that are considered to have pertinent clues for emotion recognition. To speed up the fast Fourier transform (FFT) process and avoid spectral leakage we performed framing and windowing of the speech signal. We particularly used the hamming window function for the windowing process. We ensured that the frames are appropriately overlapped after windowing to avoid loss of signal information. Instead of the 34 speech features extracted in the previous research, 40 MFCCs were extracted using Librosa. Mel spectrograms, a combination of spectral and prosodic features were also experimented on, however, similar or worse results were obtained. The frame size and hop length were 512 and 128 respectively. MFFCs are low-level descriptors of sound that describe changes per time interval of different sound spectrum bands. They depict the vocal tract frequency response in sound. They are obtained by creating triangular filters on already constructed log mel spectra and decorrelating the obtained filter banks using the discrete cosine transform (DCT). Since they use the mel scale which mimics the human auditory system, MFCCs provide

the model with perceptual frequency representations that are relevant for emotion recognition.

2) LEXICAL FEATURES

Because static pre-trained models for word embeddings do not dynamically understand the logical meaning of some words in sentences, we chose to use dynamic word embedding models. The dynamic word embedding model that has recently been used in recent research is the BERT model [40]. We therefore used the pre-trained BERT model to extract word embedding vectors from transcriptions. We particularly used the distilBERT because of its lightweight yet effective compared to others that do the same task. However, a smaller BERT may yield poorer results. The BERT model operates in such a way that the first token depicts the class of the sentence and the last token acts as a sentence separator. BERT produces dynamic word embeddings that are aware of their surrounding in the input sequence which helps the model to achieve contextualized learning that eventually improves its robustness and reliability. The BERT model also represents the grammatical and semantic features more effectively compared to the other existing dynamic embedding models.

C. EXPERIMENTS

The first category of experiments we carried out were to ascertain the significance of multilevel fusion of learned intra and cross-modality features using the MLTED model. We then carried out other experiments on the proposed CoSTGA model that uses multi-level fusion to concatenate the learned features at two progressive levels as explained in Section III.

1) EXPERIMENTS ON MULTI-LEVEL FUSION

The significance of multilevel feature fusion was investigated in form of ablation studies of the MLTED model. The experiments on multi-level fusion involved the single-level fusion individual modality transformer encoder (SLTED) to find out the performance of the first-level modal fusion and the MLTED for multi-level fusion. The first experiment used a model named SLTED where the transformer encoder (TED) was used to extract intra-modality features from the individual modalities and then they were fused. In the second experiment the whole MLTED model was used. The first fusion extracted intra-modality features which were fused as in the first experiment as cross-modality features. The intra-modality features extracted using single modality TED were then fused with the cross-modality features at the second level as shown in Fig. 1.

2) EXPERIMENTS ON THE PROPOSED CoSTGA MODEL

We carried out experiments on the proposed CoSTGA model in form of ablation studies to ascertain the impact and significance of its constituent modules and the entire model for the SER task. We particularly carried out experiments to ascertain the performance and robustness of the model for each individual emotion. Experiments were also carried out to ascertain the significance of the spatial module (SA), temporal module (TA), the temporal and spatial module (TASA) and finally the proposed CoSTGA model that consists of the spatial, temporal and semantic modules.

V. RESULTS AND DISCUSSION

In this section, we present the results of the experiments on the multi-level fusion MLTED model and the proposed CoSTGA model. We discuss the significance and impact of multi-level fusion compared to single-level fusion in SER. We also present results and discuss the impact and significance of the proposed CoSTGA model in SER in relation to our experimental results and those obtained by existing approaches.

A. RESULTS

We report results in terms of unweighted accuracy (UA), weighted accuracy (WA), precision (P), recall (R), F1 score (F1), loss (L), area under the curve (AUC) and individual class confusion ratios (CR) as our performance metrics. We also present the confusion matrices obtained from all the experiments. Tables 1 and 2 present the results of the performance of the MLTED model and its constituent single-level sub-model, the SLTED model. Table 1 presents the general performance of the models while Table 2 presents the models' performance on individual emotional class. These results confirm the superiority of multi-level fusion as compared to single-level fusion. This prompted us to propose a model that uses multi-level fusion.

The results of the proposed CoSTGA model which employs multi-level fusion and those of its sub-models are presented in Tables 3, 4 and 5. Table 3 presents the proposed CoSTGA model's performance results together with those obtained from ablation experiments of spatial (SA), temporal (TA) modules and the TASA model that combines the spatial and temporal representations without consideration of the grammatical and semantic information. Table 4 presents the performance of the TA and SA models on individual classes. Table 5 consists of the performance of the TASA model and our proposed CoSTGA model on individual emotional classes.

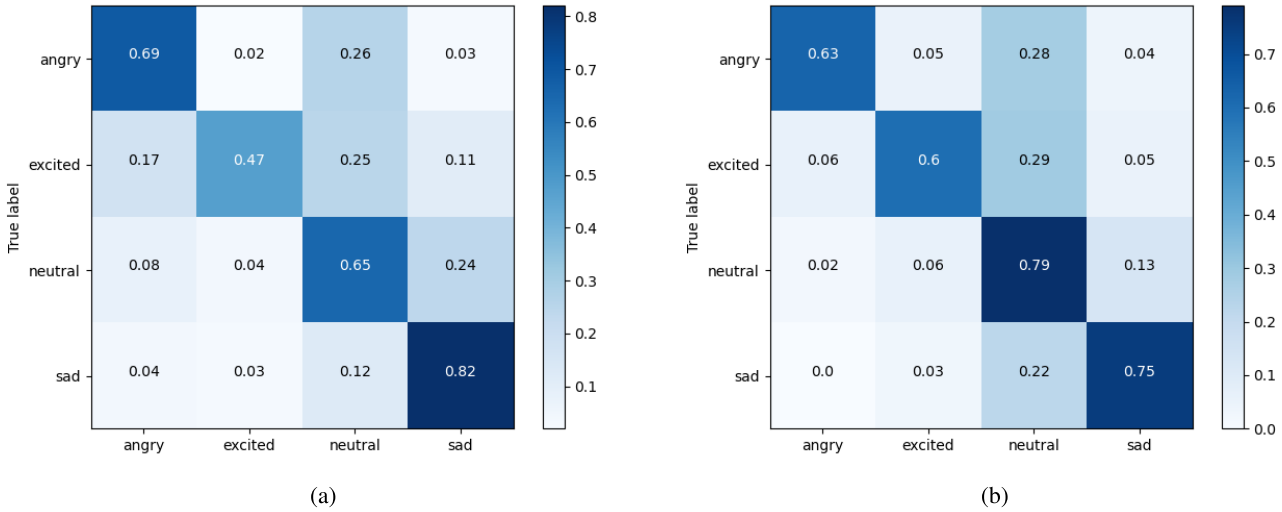
The confusion matrices for the multi-level fusion experiments are shown in Figs. 7 (a) and 7 (b). These figures further show the significance of multi-level fusion compared to single-level fusion in terms of how the models predict the individual emotion classes. The confusion matrices for the TA, SA, TASA and the proposed CoSTGA model are shown in Figs. 8 (a), 8 (b), 8 (c) and 8 (d). These confusion matrices show the progressive improvement in even robustness from the models considered for ablation experiments to the proposed model in terms of individual emotion class prediction compared to being robust on one emotional class and poor on another.

TABLE 1. Performance of the SLTED model and MLTED model.

Model	Fusion level	UA(%)	WA(%)	P(%)	R(%)	F1(%)	AUC(%)	Loss
SLTED	Single	69.75	65.67	72.42	63.16	67.50	76.92	0.8198
MLTED	Multi	73.54	70.00	75.69	68.99	72.18	79.36	0.7306

TABLE 2. Performance of the SLTED model and MLTED model on individual emotional classes.

Model	SLTED					MLTED				
Emotion	P(%)	R(%)	F1(%)	AUC(%)	CR(%)	P(%)	R(%)	F1(%)	AUC(%)	CR(%)
Excited	80	47	59	72	47	76	60	67	78	60
Sad	62	82	70	84	82	72	75	73	83	75
Angry	68	69	69	80	69	87	63	73	80	63
Neutral	63	65	64	72	65	62	79	69	76	79



(a) (b)

FIGURE 7. Confusion matrix results; (a) SLTED model. (b) MLTED model.

TABLE 3. Performance of the proposed CoSTGA Model and its sub models.

Model	UA(%)	WA(%)	P(%)	R(%)	F1(%)	AUC(%)	Loss
TA	73.42	72.42	75.86	69.62	72.61	81.38	0.7280
SA	73.92	70.00	75..03	72.28	73.65	79.41	0.8050
TASA	74.1	73.67	74.45	73.04	73.57	82.2	0.8900
CoSTGA	75.82	75.50	75.89	75.32	75.57	83.50	0.7793

B. DISCUSSION

In this subsection, we discuss the implications of the results obtained from the experiments we carried out. We also compare our proposed CoSTGA model's performance with the existing approaches.

1) SINGLE AND MULTI-LEVEL FUSION

According to the results shown in Tables 1 and 2, the multi-level fusion approach improves the performance of speech emotion recognition systems. The results show that the progressive multi-level fusion of the cross and intra-modality feature relationships improves performance. From the experimental results in Table 1, we realize that the performance of the SER model improved by 4.33% and 3.79% of weighted and unweighted accuracy respectively. The loss exhibited by the model on the testing dataset reduces from 0.8198 to 0.7306. The multilevel fusion further obtains superior performance over single-level fusion in terms of precision, recall, F1 score, and the area under the curve which proves its effectiveness and robustness in the SER task. The robustness and effectiveness of multi-level fusion are further shown in Table 2 in terms of the individual emotion class performance. Different from existing approaches the results show that progressive multilevel fusion improves the model's performance on individual class prediction in an even manner compared to performing so well on one emotion with poor performance on the other. This scenario happens in previous research especially for emotions that exist in the same dimensional plane like anger and happiness. However, a balanced area under the curve, confusion ratio, F1 score and recall are registered by the MLTED model that uses multi-level fusion compared to single-level fusion. This evenly effective and robust performance is further visualized in Figs. 7 (a) and (b). Fig. 7 (a) shows that the model is more effective at detecting the sad emotion but very poor at detecting the excited emotion, however with multi-level fusion the MLTED model whose confusion matrix is shown in Fig. 7 (b) is evenly effective for all the

TABLE 4. Performance of the temporal (TA) and spatial (SA) sub models on individual emotional classes.

Model	TA					SA				
Emotion	P(%)	R(%)	F1(%)	AUC(%)	CR(%)	P(%)	R(%)	F1(%)	AUC(%)	CR(%)
Excited	73	66	69	80	66	76	53	63	75	53
Sad	78	70	74	82	70	71	80	75	85	80
Angry	73	84	78	87	84	81	67	73	81	67
Neutral	68	70	69	76	70	65	78	71	76	78

TABLE 5. Performance of the TASA and proposed CoSTGA models on individual emotional classes.

Model	TASA					CoSTGA				
Emotion	P(%)	R(%)	F1(%)	AUC(%)	CR(%)	P(%)	R(%)	F1(%)	AUC(%)	CR(%)
Excited	69	74	71	83	74	72	60	73	83	74
Sad	75	75	75	84	75	77	74	75	84	74
Angry	80	76	78	85	76	79	82	80	88	82
Neutral	71	70	70	77	70	73	72	73	79	72

emotion classes involved. This is because the proposed model learns the intra and cross-modality feature representations progressively at the first and second fusion levels. This is done by leveraging the benefits of both feature-level and decision-level fusion by the model-level fusion approach. The results show the superiority of multi-level fusion models as compared to single-level fusion models. It should also be mentioned that the multi-head attention mechanism used in the MLTED model to compute context vectors of the inputs that are most relevant for the emotion recognition task in the sequence improves the SER performance. A careful combination of attention mechanisms and the traditional deep learning approaches for learning spatial and temporal features in a multi-learning approach allows the model to benefit from the parallel computation of global attention for better results. We were therefore motivated by these results to apply a combination of multi-head attention mechanisms and traditional deep learning techniques to allow the model to learn the spatial, temporal and semantic features.

2) PERFORMANCE ANALYSIS OF THE PROPOSED CoSTGA MODEL

The results presented in Tables 3 and 5 suggest that the proposed CoSTGA model is effective and robust for the SER task. Table 3 shows that the concurrent learning of spatial, temporal and semantic features in the LFLB and subsequent fusion at the second level in the GFLB yields a performance of 75.50% and 75.82% of weighted and unweighted accuracy respectively. The proposed model's effectiveness and robustness are further shown by the precision, recall, F1 score and AUC values of 75.89%,75.32%, 75.57% and 83.50% respectively. These effective results are due to the fact that the model learns the three main factors that constitute an emotion concurrently without first learning one and then the other in a sequential manner as is done by the existing approaches. Table 3 further presents details of the significance of each module that constitutes the proposed CoSTGA model. It is observed that a combination of the concurrent temporal and spatial feature representations obtains 74.1%, 73.67%, 73.04%, 73.57% and 82.2% in terms of unweighted accuracy, weighted accuracy, recall and AUC respectively. These results are however further improved by the addition of the semantic information to make the proposed CoSTGA model as mentioned earlier. The CoSTGA model also exhibits a loss of 0.7793 compared to 0.8900 of the TASA model that only combines the concurrent temporal and spatial information without the semantics information. These results uphold the significance of semantic and grammatical information in SER system's effectiveness and robustness. The temporal and spatial modules' contribution in terms of the individual emotion classes are presented in Table 4, From these tabular results, it is clear that the sub-models are effective on some emotion classes but poor on others. An example is the confusion ratio of excited and angry compared with sad and neutral, excited and the others for spatial and temporal sub-models respectively. This explains why the existing approaches that use either of the representations obtain uneven effectiveness and robustness of emotion classes. It is however, clear from all the results presented in this section that each of the sub-models contributes positively to the general performance of the proposed CoSTGA model. The evenly distributed and effective individual emotion class recognition exhibited by the proposed CoSTGA model compared to the TASA model is shown in Table 5. This is especially evidenced in terms of the AUC, confusion ratio and F1 score in which the proposed CoSTGA model obtains values that are in the same range for all emotion classes.

It is further observed from the confusion matrices shown in Figs. 8 (a), (b), (c) and (d) that there is balanced generalization by the proposed CoSTGA model compared to the constituent sub-models. The prediction of the happy emotion class which is poorly done by the TA and SA sub-models improves with a combination of the spatial and temporal feature representations. It is also observed that this combination lowers the angry emotion recognition accuracy but it is improved by the addition of grammatical and semantic

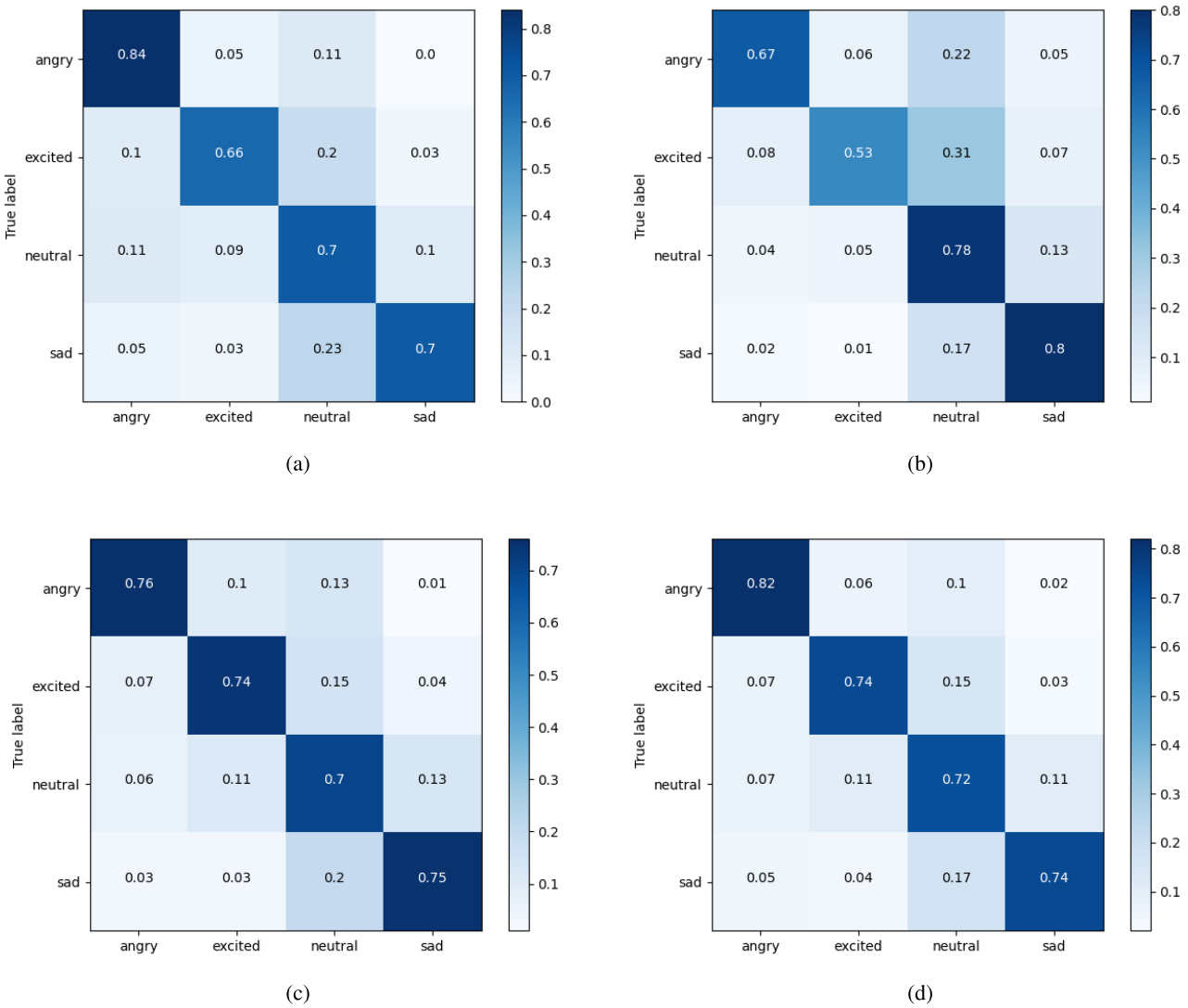


FIGURE 8. Confusion matrix results; (a) TA model. (b) SA model. (c) TASA model. (d) Proposed CoSTGA model.

knowledge in the proposed CoSTGA model. These results also confirm that the progressive increase of levels of fusion, concurrent spatial, temporal and semantic feature learning improves the performance of SER models. The confusion matrices also show that there is better generalization during inference of the proposed CoSTGA model as compared to the control experiments. It should however be noted that the emotion classes are mostly confused with the neutral class. This is because the neutral emotion class is situated at the center of the two-dimensional arousal-valence spaces of emotions which complicates the discriminatory capability of the model [41]. Since the proposed CoSTGA model's performance has been evaluated on a dyadic database that includes interactive speech between a pair of interlocutors, we anticipate that the results reported may fluctuate if tested on datasets that were collected in more natural environmental conditions with a variety of interlocutors in terms of gender, culture, age and accent, ambiance and noise attributes.

TABLE 6. Performance comparison of the proposed CoSTGA model with the existing models.

Fusion	Model	UA(%)	WA(%)	R(%)	F1(%)
Single-Level	MDRE [24]	-	-	71.8	-
	STSER [25]	72.05	71.06	-	-
	LAMER [26]	70.9	72.5	-	-
	IEmoNet (BE) [27]	72.05	74.98	-	-
	Ho et al [30]	-	-	73.23	73.32
	Singh et al [28]	74.5	-	73.2	-
Muli-level	Proposed CoSTGA	**75.82**	**75.50**	**75.32**	**75.57**

On the whole, there is an effective and generally robust SER for multi-level fusion compared to single-level fusion with a combination of concurrently learned temporal, spatial and semantic features in the proposed CoSTGA model.

3) COMPARISON WITH EXISTING APPROACHES

We compared the performance of the proposed model with the existing approaches. It should be noted that most of the existing approaches do not use semantic tendencies

except [8] which does not consider spatial feature learning. Three modalities of video, text and audio are considered in [8] for SER. The other approaches use either spatial, temporal feature learning or both sequentially but not in a concurrent manner. In terms of multi-level fusion, most of the approaches apply early or late fusion but not model or intermediate fusion used in this paper. For a fair comparison, we chose to compare the performance of our proposed model with existing models that use lexical and acoustic modalities only leaving out those that use visual and speaker sensitivity features in combination with text and audio modalities. In addition, we compared with only models that use the IEMOCAP dataset. The models chosen for comparison used either additive, multiplicative or multi-head attention mechanisms in either one modality branch or a combination of them in a multi-modality approach. These models also use single-level fusion. Table 6 shows the performance comparison between the proposed CoSTGA model and the existing approaches. The existing approaches we compared the proposed CoSTGA model with are; MDRE [24], STSER [25], LAMER [26], IEmoNet (BE) proposed in [27], the model proposed in [30] that uses mixed files of the IEMOCAP dataset and the hierarchical approach for bimodal SER model proposed in [28]. The model in [28] handles the task by use of early fusion and subsequent progressive hierarchical deep learning-based SER. The IEmoNet (BE) uses pre-trained language models for the transcriptions in combination with audio signals trained separately and combined in a late fusion approach for SER.

It is shown that MDRE achieved a maximum average recall of 71.8%, LAMER obtained 70.9% and 72.5%, STSER 72.05% and 71.06%, IEmoNet (BE) 72.05% and 74.98% of unweighted and weighted accuracy respectively. The model in [30] achieved 73.23% of recall and 73.32% of F1 score on mixed files of the IEMOCAP dataset. 74.5% of unweighted accuracy and 73.2% of recall were reported in [28]. These results place the accuracy of our proposed CoSTGA model higher than the existing approaches by a range of 1.32% to 4.92%, 0.52% to 4.44%, 2.12% to 3.52%, 2.25% in terms of unweighted accuracy, weighted accuracy, recall and F1 score respectively. This is an indicator that multi-level fusion models that concurrently and progressively learn spatial, temporal and semantic intra and cross-modality feature representations improve performance in SER.

VI. CONCLUSION

In this paper, we investigated the significance of multi-level fusion for SER and proposed the CoSTGA model which leverages the benefits of fusing concurrently learned spatial, temporal and semantic feature representations in a multi-level approach. The proposed model learns intra and cross-modality feature of acoustic and lexical modalities at two fusion levels using model-level fusion. A combination of traditional deep learning techniques and multi-head attention mechanisms were used. The impact and significance of each of the concurrent feature representations were evaluated. The proposed CoSTGA model's performance was compared with the existing approaches in terms of weighted and unweighted accuracy, recall, AUC and F1 score. The results show that a multi-level fusion of concurrently learned spatial, temporal and semantic feature representations improves the effectiveness and robustness of SER models. We plan to explore concurrent feature learning of spatial, temporal and semantic tendencies in all modalities for emotion recognition since human emotional states encompass other modalities like visual and physiological cues.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, p. 124, 1972.
[2] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *Int. J. Social Robot.*, vol. 11, no. 4, pp. 555–573, 2019.
[3] G. K. Verma and U. S. Tiwary, "Affect representation and recognition in 3D continuous valence–arousal–dominance space," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2159–2183, Jan. 2017.
[4] J. J. Gross and L. Feldman Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion Rev.*, vol. 3, no. 1, pp. 8–16, Jan. 2011.
[5] D. Hu, L. Wei, and X. Huai, "DialogueCRN: Contextual reasoning networks for emotion recognition in conversations," 2021, *arXiv:2106.01978*.
[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[7] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Jan. 2018.
[8] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," 2022, *arXiv:2203.13504*.
[9] Y. Yan and X. Shen, "Research on speech emotion recognition based on AA-CBGRU network," *Electronics*, vol. 11, no. 9, p. 1409, Apr. 2022.
[10] S. Kakuba and D. S. Han, "Residual bidirectional LSTM with multi-head attention for speech emotion recognition," in *Proc. Korea Commun. Assoc. Summer General Academic Conf.*, 2022, pp. 1419–1421.
[11] B. Maji, M. Swain, and M. Mustaqeem, "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-GRU features," *Electronics*, vol. 11, no. 9, p. 1328, Apr. 2022.
[12] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *Proc. 10th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2020, pp. 1058–1064.
[13] G. Bekmanova, B. Yergesh, A. Sharipbay, and A. Mukanova, "Emotional speech recognition method based on word transcription," *Sensors*, vol. 22, no. 5, p. 1937, Mar. 2022.
[14] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 985–1000, 2021.
[15] W. Li, W. Shao, S. Ji, and E. Cambria, "BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73–82, Jan. 2022.
[16] Z. Lian, B. Liu, and J. Tao, "DECN: Dialogical emotion correction network for conversational emotion recognition," *Neurocomputing*, vol. 454, pp. 483–495, Sep. 2021.
[17] Y. Shou, T. Meng, W. Ai, S. Yang, and K. Li, "Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis," *Neurocomputing*, vol. 501, pp. 629–639, Aug. 2022.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.
[19] H. Zhang, H. Huang, and H. Han, "Attention-based convolution skip bidirectional long short-term memory network for speech emotion recognition," *IEEE Access*, vol. 9, pp. 5332–5342, 2021.
[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[21] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[23] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, p. 7530, Nov. 2021.
[24] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 112–118.
[25] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," in *Proc. Interspeech*, Oct. 2020, pp. 374–378.
[26] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," 2019, *arXiv:1909.05645*.
[27] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," 2019, *arXiv:1912.02610*.
[28] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107316.
[29] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multimodal multi-head attention for continuous affect recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 4171–4183, 2021.
[30] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.
[31] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," 2019, *arXiv:1910.11263*.
[32] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. 1, p. e12, 2014.
[33] P. Koromilas and T. Giannakopoulos, "Deep multimodal emotion recognition on human speech: A review," *Appl. Sci.*, vol. 11, no. 17, p. 7962, Aug. 2021.
[34] S. Lee, D. K. Han, and H. Ko, "Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification," *IEEE Access*, vol. 9, pp. 94557–94572, 2021.
[35] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
[36] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Alignment approach comparison between implicit and explicit suggestions in object reference conversations," in *Proc. 4th Int. Conf. Hum. Agent Interact.*, Oct. 2016, pp. 193–200.
[37] Y. Liu, H. Sun, W. Guan, Y. Xia, and Z. Zhao, "Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework," *Speech Commun.*, vol. 139, pp. 1–9, Apr. 2022.
[38] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[39] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," 2018, *arXiv:1804.05788*.
[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[41] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," 2017, *arXiv:1706.00612*.

SAMUEL KAKUBA received the B.Sc. degree in computer engineering from Busitema University Tororo, Uganda, in 2011, and the M.Sc. degree in data communication and software engineering from Makerere University Kampala, Uganda, in 2018. He is currently pursuing the Ph.D. degree with the Graduate School of Electronic and Electrical Engineering, College of IT Engineering, Kyungpook National University (KNU), Republic of Korea. He is currently an Assistant Lecturer with Kabale University, Uganda. He has worked as a Researcher for projects in the fields of data communication systems, embedded systems engineering, the Internet of Things, emotion recognition, computer vision, affective computing, and other machine and deep learning systems.

ALWIN POULOSE received the B.Sc. degree in computer maintenance and electronics from the Union Christian College (affiliated to Mahatma Gandhi University), Aluva, India, in 2012, the M.Sc. degree in electronics from the MES College (affiliated to Mahatma Gandhi University), Marampally, India, in 2014, the M.Tech. degree in communication systems from Christ University, Bangalore, India, in 2017, and the Ph.D. degree in electronics and electrical engineering from Kyungpook National University, Daegu, South Korea, in 2021. From 2021 to 2022, he was a Researcher at the Center for ICT and Automobile Convergence (CITAC), Kyungpook National University, where he developed a Multi-intelligence-based Human-Centric Autonomous Driving Core Technology. His research interests include localization, human activity recognition, facial emotion recognition, and human behavior prediction. He is a Reviewer of prominent engineering and science international journals and has served as a technical program committee member/session chairing at several international conferences. He is currently a Research Fellow with the Department of Electrical and Computer Engineering, University of Michigan, Dearborn, USA.

DONG SEOG HAN (Senior Member, IEEE) received the B.S. degree in electronic engineering from Kyungpook National University (KNU), Daegu, South Korea, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1989 and 1993, respectively. From 1987 to 1996, he was at Samsung Electronics Company Ltd., where he developed the transmission systems for QAM HDTV and Grand Alliance HDTV receivers. Since 1996, he has been a Professor with the School of Electronics Engineering, KNU. He was a Courtesy Associate Professor at the Department of Electrical and Computer Engineering, University of Florida, in 2004. He was a Director at the Center of Digital TV and Broadcasting, Institute for Information Technology Advancement (IITA), from 2006 to 2008. He is currently the Director with the Center for ICT and Automotive Convergence, KNU. He is also the Dean of the IT College, KNU. His main research interests include intelligent signal processing and autonomous vehicles.