IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 September 2023, accepted 12 October 2023, date of publication 16 October 2023, date of current version 19 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325037

RESEARCH ARTICLE

Deep Learning Approaches for Bimodal Speech Emotion Recognition: Advancements, Challenges, and a Multi-Learning Model

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]School of Data Science, Indian Institute of Science Education and Research Thiruvananthapuram (IISER TVM), Vithura, Thiruvananthapuram, Kerala 695551, India
[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 Though acoustic speech emotion recognition has been studied for a while, bimodal speech emotion recognition using both acoustic and text has gained momentum since speech emotion recognition doesn't only involve the acoustic modality. However, there is less review work on the available bimodal speech emotion recognition (SER) research. The review works available mostly concentrate on the use of convolution neural networks (CNNs) and recurrent neural networks (RNNs). However, recent deep learning techniques like attention mechanisms and fusion strategies have shaped the bimodal SER research without explicit analysis of their significance when used singly or in combination with the traditional deep learning techniques. We therefore, review the recently published literature that involves these deep learning techniques in this paper to ascertain the current trends and challenges of bimodal SER research that have hampered it to be fully deployed in the natural environment for off-the-shelf SER applications. In addition, we carried out experiments to ascertain the optimal combination of acoustic features and the significance of the attention mechanisms and their combination with the traditional deep learning techniques. We propose a multi-technique model called the deep learning-based multi-learning model for emotion recognition (DBMER) that operates with multi-learning capabilities of CNNs, RNNs, and multi-head attention mechanisms. We noted that attention mechanisms play a pivotal role in the performance of bimodal dyadic SER systems. However, few publicly available datasets, the difficulty in acquisition of bimodal SER data, cross-corpus and multilingual studies remain open problems in bimodal SER research. Our experiments on the proposed DBMER model showed that though each of the deep learning techniques benefits the task, the results are more accurate and robust when they are used in careful combination with multi-level fusion approaches.

INDEX TERMS Emotion recognition, acoustic and lexical data, deep learning, attention mechanisms.

I. INTRODUCTION

The growing desire to improve the social intelligence of agents that can detect and comprehend human affection has increased the relevancy of affective computing in the research community. The progressive success of affective computing with deep learning techniques has continued to facilitate the improvement of seemingly natural interactions between human beings and intelligent agents without encumbrances. However, besides the success, there are still a number of

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

challenges that are of concern to the research community especially in terms of the transfer of laboratory-generated models to the natural environment [1]. An analysis of the existing models needs to be done in order to improve and make them suitable for the natural environment in which they are applied. Affective computing is deployed in a number of applications; health [2], robotics [3], education [4], customer care [5], psychology [6], etc. Since its proposal in 1997 by Prof. Picard [7], affective computing has been studied in form of emotion recognition and sentiment analysis. Emotion recognition is categorized according to the source of the data. The emotional data is categorized as physical or physiological depending on its source. Physical data is from acoustic, lexical and visual sources and physiological data is from electrocardiogram (ECG), electroencephalogram (EEG) or galvanic skin response (GSR). This paper concentrates on bimodal speech emotion recognition that involves the audio speech signals and their text transcriptions as emotion sources.

The applications of speech emotion recognition (SER) systems need accurate, computationally efficient and robust models to perform the task of SER. Moreover, they need to be more robust when deployed in real-time natural environments [1]. One of the ways to achieve a robust and accurate performance is the choice of sufficient data used for training the SER model. A number of unimodal models have been proposed for SER as reviewed in [1], [8], and [9] and they exhibit a promising performance. However, it decreases when transferred to the natural environment. It is however, prudent to model word-frame interactions between text and acoustic modalities of speech [10] for more robust emotion recognition. In addition, as stated in [11] the clues of how it is said (acoustic) and what is said (lexical) contribute to the emotion portrayed in an utterance. Moreover, Lian et al. [12] asserts that there exists cross-modality relationships between acoustic and lexical features that ought to be carefully fused with intra-modality features for robust and accurate SER. Indeed, bimodal SER has been proven to perform better than unimodal SER either in terms of text or acoustic data. Lian et al. [12] tested their conversational network model that uses a transformer and bidirectional gated unit (BiGRU) with separate acoustic and lexical features before fusing them to allow the model to benefit from both the intra and cross-modality feature relationships and found out that the model performed well on both the interactive emotional dyadic motion capture (IEMOCAP) database and the multi-lines emotion dataset (MELD) datasets. In [13], it is also empirically shown that the unimodal features do not perform as robustly as the fused features. In addition, the results in [14] show that though the models subjected to the text modality perform better than those of the acoustic modality, models subjected to a combination of the two often give better results than unimodal models. The same case is reported in [15], [16], and [17] among others.

The efficiency and robustness of SER models depend on two stages; feature extraction and emotion classification. Robust feature extraction provides the model with elaborate knowledge about acoustic features like pitch, formant frequencies, and vocal tract. Such features are important for the discernment of different emotions in speech. Generally, the two main speech feature extraction techniques are temporal analysis and spectral analysis [18]. Traditional methods involve handcrafted methods of extracting local features from which global features are computed statistically. SER deep learning models dynamically learn local and global features that are important to the task with improvement in generalization, robustness, and accuracy. The overview of the bimodal SER model framework is shown in Fig. 1. For both modalities, the framework includes data processing, feature extraction, model-based learning, and classification. In terms of the acoustic features, the nature of the speech signal dictates the extent of data processing. Data processing is done to ensure feature normalization in order to ensure that the speaker variations and the ambiance of the recording environment [19] do not affect the emotional state recognition process. However, it may involve ensuring equal sequence length of all the speech signals in the dataset which is the input to the deep learning model. The speech signals that are of shorter length than the required are padded and the longer ones are truncated. The data processing stage also involves the removal of the silent regions if they do not carry any emotional clues that are useful to the model. Pre-emphasis and bandpass filters can also be used to allow only the frequencies of the speech signal that are considered to have pertinent cues for emotion recognition. To speed up the fast Fourier transform process and avoid spectral leakage, framing and windowing are normally done. The Hamming and Hanning window functions are used for windowing. The frames are overlapped after windowing to avoid loss of signal information. After the data processing phase, the signal may be fed directly into a deep learning model or low-level features (LLDs) can be first extracted as local features and subsequently, high-level statistical features (HSFs) computed as global features before being fed into the deep learning model.

The same process is followed for the linguistic features however the data processing of the text may involve dropping missing entries in data and standardizing text through tokenization which involves the removal of stop words, stemming, and lemmatization. However, some of these steps are omitted if the bidirectional encoder representations from transformers (BERT) is used since it caters for them. To represent the words in documents (usually transcribed from voice), one hot encoded text vectors are used. However, due to the fact that the indices assigned to the words do not hold any meaning yet similar words occur frequently with different meanings word embeddings proved better representations. Traditional word embeddings like term frequency-inverse document frequency (TF-IDFs), bag of words (BoW), static and dynamic word embeddings like the word2vec, global vectors for word representations (GloVes), fastText, embeddings from language models (ELMO), BERT are used as described in Section II before the deep learning models and fusion with the acoustic feature representations.

To this end, we have described the bimodal SER system overview with an assumption that the decision-level fusion has been used to fuse the two modality feature representations. However, as described in Section II there exist three fusion strategies that can be used to fuse the bimodal features. They include early fusion, intermediate or model-level fusion, and decision-level fusion. Their merits and demerits in addition to alignment strategies of the bimodal features are discussed later in Section II.

The contributions of this paper include;

- A review of the datasets, linguistic and acoustic features used, their alignment and fusion strategies, and deep learning models is presented.
- Since the attention mechanisms have recently enhanced the performance of deep learning models in bimodal SER research, we provide an analysis of bimodal SER deep learning models based on attention mechanisms at early, intermediate and decision fusion levels.
- We also analyze the published results in terms of the accuracy and robustness of the models and point out the strength and challenges that need attention in bimodal SER research.
- We propose a multi-technique model called a deep learning-based multi-learning model for emotion recognition (DBMER) that operates with multi-learning capabilities of CNNs, RNNs and multi-head attention mechanisms and evaluated its performance.

This distinguishes our research from the few published bimodal SER surveys that review only traditional deep learning techniques and do not give adequate attention to the significance of attention mechanisms, alignment, and fusion strategies.

The rest of the paper is organized as follows: we present a review of the methods used in deep learning-based bimodal SER research in Section II. The experiments we used in this paper are described in Section III in which we also propose the DBMER SER model. We describe the results obtained in the different research and our experiments in terms of their robustness and accuracy in Section IV. A detailed discussion of the results of the proposed DBMER model and the recent bimodal SER models and research challenges are presented in the discussion Section V. We finally conclude in Section VI.

II. METHODS

In this section, we describe the datasets, features, fusion and alignment strategies, and the different deep learning techniques used in the bimodal SER research studies. It should be noted that some of these techniques are also used in sequence or concurrently to learn the feature representations of the acoustic and lexical modalities. We therefore, review the combinations of these in the literature in a bid to assess how best the spatial, temporal and grammatical or semantic Intra and inter-modality features can be represented for the best robustness and eventual deployment in the natural environment. It should be noted that, though [20] agrees that spatial and temporal features are important to SER, [21] suggests that these features should be combined with grammatical features to include clues of the semantics of the uttered words.

A. DATASETS

In order to evaluate the performance of proposed bimodal SER models, datasets are used as inputs either in raw signal form or extracted features. The choice of datasets in SER depends on the availability of the datasets and their nature. It is however, a determinant of the performance and robustness of the model in different scenarios. We generally categorize the datasets into three depending on the conditions and environments in which they were recorded or obtained. Natural datasets are those that are recorded during naturalistic tendencies without subjecting the participants to scripted actions or utterances. Most of these are obtained from talk shows that were not primarily collected for speech recognition. The elicited datasets are those that are recorded by stimulating participants' emotions. The actions may include watching movie scenes that consist of different emotions or putting them in situations where the required emotions can be elicited. The last and most common type is the acted datasets which are recorded according to a script by professional actors that bring out the required emotions. Though its expensive to obtain any of these types of datasets, the natural datasets are the scarcest due to the difficulty in recording naturalistic speech emotions. The research community is often left with only the remaining two options which are also not available in many languages in addition to some of them being publicly unavailable. The use of the elicited and acted datasets validates a number of models with good performance but not robust when deployed in the natural environment. This is because the conditions in the natural environment in terms of speakers, age, culture, language, text, and recording conditions do not match with the recording environments. Therefore, there is a need for interventions that will bridge the mismatch between the datasets used in controlled environments which are used to validate SER models and those existent in real-life environments.

We opine that since speech data collection is expensive, adversarial data augmentation techniques and transfer learning, especially domain adaptation may be possible solutions to the mismatch since the generated data will be of a similar distribution as the real data. Though these datasets include a number of emotion categories, the annotation process needs to be done carefully. The annotation depends on the ingenuity and perspective of the annotator. For SER databases, it is common for the annotators to be listeners or the speakers themselves and evaluate according to a self-rating system

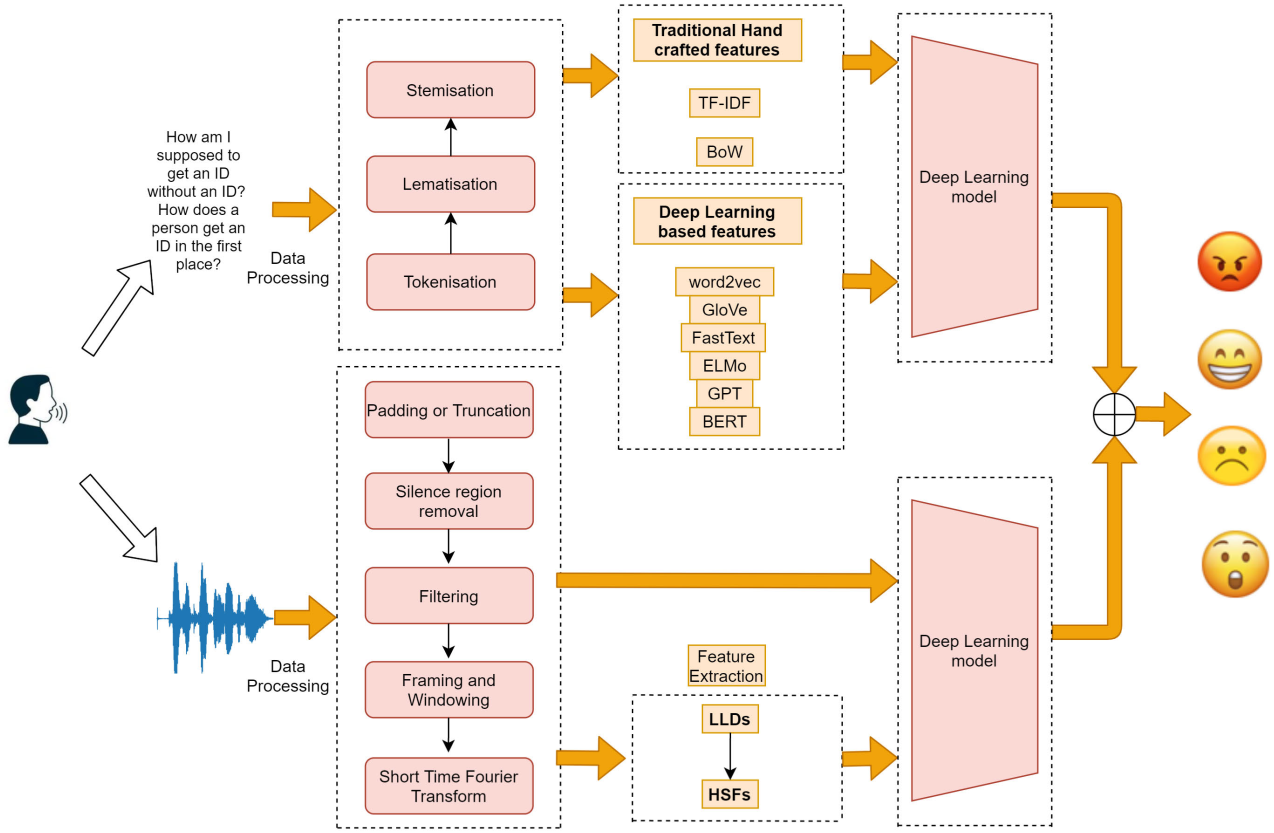


FIGURE 1. The bimodal speech emotion recognition (SER) system overview.

TABLE 1. Summary of the three most common datasets.

Dataset	Kind	Modalities	Number of Utterances	Topic Oriented	Dyadic
IEMOCAP [22]	Acted	Lexical, Acoustic, Visual	10,309	No	Yes
MELD [23]	Natural	Lexical, Acoustic	13,000	No	No
CMU-MOSEI [24]	Natural	Lexical, Acoustic, Visual	23,500	Yes	No

in terms of emotional induction or how they feel. The accuracy of the annotations by both listeners and speakers for the emotion speech databases is still a research issue [25]. Though there are many emotion recognition datasets in existence we only concentrated on the datasets that are used particularly for combined acoustic and lexical SER in this article. Since few databases that have both audio and text modalities exist, we review the three most common speech emotion datasets that are frequently used to validate bimodal SER models. Table 1 shows a summary of the three most common datasets.

1) THE IEMOCAP DATASET

The interactive emotional dyadic motion capture (IEMOCAP) database [22] was collected at the University of Southern California as a multi-modal and multi-speaker emotion recognition database. It contains audio, transcriptions, video, and motion-capture (MoCap) recordings of dialogues between dyadic mixed-gender pairs of ten actors recorded as scripted and improvised utterances in five sessions. This datasets consists of the discrete emotions of anger, happiness, sadness, neutral, frustration, excitement, fear, surprise, and disgust. The actors' utterances were evaluated by at least three different annotators for the discrete emotions and two for the dimensional emotions. The emotions according to dimensional axes are valence (positive or negative), activation (calm or excited), and dominance (passive or aggressive). For each instance of the dialogue to be labeled by the evaluators, the data was partitioned into 3 to 5 seconds length utterances. The database consists of dialog and sentence recordings for about 12 hours. The fact that this dataset consists of participants of mixed gender in an elicited emotional environment makes it partly naturalistic. However, this dataset is heavily imbalanced with one of the emotion categories having only two utterances. The statements used are also few compared to what is found in real life. This explains why it's often challenging for the proposed SER datasets. It should however be noted that most of the articles reviewed in the literature use four of the emotion categories in this dataset that seem balanced or can be worked with when class weights are configured. These are happy or excited, sad, angry, and neutral. Happy and excitement are sometimes fused together since the clues of excitement and happiness are similar in real life and they appear in the same dimension quadrant of the emotional dimension plane. We opine that this dataset can be used as in literature with class weights consideration to alleviate class imbalances. In addition, the dataset can be used with adversarial data augmentation techniques [26] and transfer learning strategies like the one used in [27] to generate feature vectors with similar distribution which can improve SER accuracy and robustness.

2) THE MELD DATASET

Multimodal emotion lines dataset (MELD) [23] has more than 1,300 dialogues and 13,000 utterances from friends TV series. It was formed by the addition of audio and visual modality to the text modality that was contained in the emotion lines dataset (ELD). It includes anger, disgust, sadness, joy, neutral, surprise, and fear as in the discrete emotions. The dataset also consists of positive, negative, and neutral sentiments for each utterance. In addition to having mixed gender, the MELD dataset focuses on understanding emotions in conversations that were collected from TV series and therefore not dyadic like other datasets which provides more natural emotions. This makes this dataset more useful in conversation emotion recognition as compared to IEMOCAP and SEMAINE which are dyadic. In addition, it gives the proposed SER models an opportunity to be evaluated on scenarios that involve more than two interlocutors. The multimodality nature of the dataset coupled with its naturalistic nature enables SER models to compute the context of the utterances which is an important consideration.

3) CMU-MOSEI DATASET

Zadeh et al. [24] proposed the CMU multimodal opinion sentiment and emotion intensity (CMU-MOSEI) dataset that is used in emotion recognition and sentiment analysis. Compared to IEMOCAP and MELD, CMU-MOSEI is the largest with more than 23,500 sentence utterance videos from more than 1,000 online you tube speakers. It consists of data from visual, lexical, and acoustic modalities. The balanced mixed-gender CMU-MOSEI dataset consists of various topics and monologue videos from which sentence utterances are chosen randomly. The researchers who proposed this dataset also published an incite through empirical results on how the three different modalities interact with each other with different magnitudes of weight for a given utterance. Due to the fact that this dataset considers a large variety of speakers and topics in addition to using online videos, we opine that it is naturalistic and a good representation of the natural environment in terms of language, speakers, culture, and recording environments. However, this kind of dataset poses a challenge of diverse conditions which may make bimodal SER models “a jack of all trades and a master of none”. Therefore, the possibility of modality influence is very important to consider when such datasets are used in SER deep learning studies.

B. FEATURES

As mentioned earlier bimodal SER involves the acoustic and lexical modalities data as the input to the deep learning models. In this section, we briefly describe the acoustic and lexical features pointing out their merits and demerits that can influence their choices in bimodal deep learning models.

1) ACOUSTIC FEATURES

Phonemes are the basic building blocks of speech that are used by speech recognition systems to represent features in a sentence. A phoneme is a unit of sound that distinguishes one word from another in a particular language. Allophones represent variations of phonemes. They are caused by accent, age, gender, phoneme position within a word and the emotional states of the speaker. The variability clues about the emotional state of the speaker needs to be represented in such a way that the SER model can understand it. It should however be noted that some researchers urge that raw signal inputs allow the deep learning model to learn the features by itself other than handcraft extraction before subjecting them to the SER model. However, we show in Section III that the results of our simple experiments in which we compared the extracted features with raw signals showed that the models trained on raw signals are robust on some emotion categories but are not as robust on other emotion categories with which they belong in the same dimensional plane. Generally, the acoustic features used in SER models are of four categories; prosodic, spectral, voice quality, and wavelet-based features.

Prosodic features represent the variations in loudness, period of utterance, intonation, and stress. They are expressed in terms of pitch, energy, and speech duration. These features are frequently used in literature because they are less affected by channel mismatch and noise. Psychologically, prosodic features are said to have a convincing correlation with the emotional state of human beings [28]. However, it is suggested in [1] that though these features are robust at distinguishing between low and high arousal emotions, they are not as good at emotions that belong to the same arousal or valence dimension. This same scenario is observed when raw signals are fed into SER models as we mentioned earlier.

The spectral features consist of the signal energy at different frequency bands. They 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). Different from prosodic features that are robust on intra-arousal and intra-valence emotions, spectral features are robust at discriminating emotions that exist in the same valence or arousal plane [1]. Though a number of spectral features have been proposed, mel frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), perceptual linear prediction (PLP), and formant features have been used most in the current SER research studies. Improvement of these spectral features with an emphasis on using local moments of the Gabor spectrograms has also been proposed in [29]. Since they take a similar approach to feature extraction, we describe the most commonly used MFCCs.

MFCCs are acoustic low-level descriptors that describe characteristics of a piece of sound by providing clues about the rate changes in the different spectrum bands. It should be noted that the speech signal is split into multiple intervals (Frames or windows) and short time Fourier transform (STFT) is applied to each interval to generate the input power spectrum. The process involves the implementation of frequencies in terms of the perceived mel scale which mimics the human auditory system. The perceived mel scale frequency (*melfreq*) is described as.

$$melfreq = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

where f denotes the physical frequency in Hz. The STFT is applied to each frame of the speech signal to generate the input power spectrum $K_{(w)}$ computed according to equation 2. $a_{(n)}$ denotes the speech segment input and $w_{(n)}$ is the weight of the window.

$$K_{(w)} = |STFT(w_n \times a_n)|^2 \tag{2}$$

The next steps in the extraction process involve the computation of filter banks and subsequent triangular filters. To compute the mel filter banks, the lowest and highest frequency are converted to mel scale according to equation 2. The mel log frequency of the power spectrum $K_{(w)}$ is given by f_s shown in equation 3. $H_s(w)$ denotes the value of the Sth triangular filter for the Wth frequency.

$$f_s = \log_{10}\left(\sum |K_{(w)}| H_{(s)}(w)\right) \tag{3}$$

The discrete cosine transform (DCT) is applied to the list of mel log frequency sub-bands as described in equation 4

$$c_i = \sum_{j=1}^{fsb} f_j \cos\left(\frac{\pi i}{fsb}(j - 0.5)\right) \tag{4}$$

where $f_{(sb)}$ denotes frequency sub-bands. i is between 0 and the number of mel frequency coefficients ($n_{(mfcc)}$). The amplitudes of the resultant spectrum gives the MFCCs. The DCT is applied since the vocal tract is smooth and therefore the energy levels in adjacent bands tend to be highly correlated and need to be decorrelated. We refer the reader to [30] for an exhaustive description of MFCCs feature extraction process.

The speech features can be extracted in the time domain, frequency domain and time-frequency domain. Some of the time domain features include the amplitude envelope (AE) which gives a rough idea about the audio signal, root mean square energy (RMSE) which is an indicator of the loudness of sound, zero crossing rate (ZCR) which is used to distinguish voiced and unvoiced signals, percussive, and pitched sounds. The frequency domain features include the band energy ratio (BER) which provides comparative information about the energy in the lower and higher frequency bands, the spectral centroid (SC) that shows the frequency band where most of the energy is concentrated (a measure of the brightness of sound) and the spectral spread (SS) which is majorly used in music processing to estimate the range around the centroid. Though speech audio is temporal, the interaction between the time and frequency domain gives better clues to the deep learning models. Therefore, the time-frequency domain features are often used in speech emotion recognition. Examples of time domain features are MFCCs, LPCCs, mel spectrograms, etc. However, there is always a trade-off between time and frequency resolutions using the frame size depending on the application. If the application requires more time or frequency resolution, a small or larger frame/window size is used respectively. These fixed windows are normally used to avoid spectral leakage during the computation of the STFT. The trade-off between time and frequency resolution depends on the application. It should also be noted that the use of the wavelet transforms (WT) was proposed to compute the spectral-temporal information instead of the STFT by using decomposing the signal into low and high-frequency components [1]. The use of the WT to extract features like MFCCs, LPCCs for emotion recognition was proposed in [31] and [32] with improved performance than those extracted using the discrete Fourier transform (DFT).

The relationship between the vocal excitation and the vocal cord in the vocal tract gives clues of the voice quality. The relationship may be described in terms of the period of opening and closing of the vocal tract or the ratio of the two reflex actions. Speech features like shimmer and jitter are used to describe the voice quality of speech audio signals. They give clues about temporal variations of the speech signal. The other voice quality feature that is normally used is the mel spectrogram which uses the mel scale to approximate the human auditory system through in terms of the excitation of the vocal tract.

2) LEXICAL FEATURES

The text transcriptions of the audio speech signal are the other modality that consists of clues that can be leveraged for SER if inter and intra-modality interactions are learned together after a careful fusion strategy. The text in the lexical modality is represented by vectors obtained either by vectorization or word embedding pre-trained models. The techniques of vectorization include the bag of words (BoW) [33] and the term frequency-inverse document frequency (TF-IDF) [34]. The BoW represents words by use of a vector that represents the word count in a document. The TF-IDF represents the information about the important and less important words in a document. However, in addition to their individual challenges in terms of high dimensionality and sparsity, the above two models do not represent semantic and grammatical cues of the utterances which makes it hard to use them in SER systems since the meaning and context of an utterance is important for the inference of emotions. In the bid to solve the context and semantics problems pre-trained word embedding models were proposed to replace vectorization. The model called word2vec [35] that uses the neighboring words to infer semantic and grammatical similarities between them was proposed. The word2vec model maps semantically close words in meaning to approximately similar embedding vectors using cosine similarity. However, this method increases the computational cost. The global vector for word representation (GloVe) [36] was proposed to improve the performance of word2vec as well as reducing on the computational cost using a simpler error function for word embedding representations. GloVe creates the embeddings using the global context of the document as opposed to the local context used in the word2vec model which helps it to produce improved text representations. The other model that uses the composition of a word for its vector representation is fastText [37] which uses the skip-gram method. Thus far, the models discussed compute the meaning and context statically which may not infer the meaning of the utterance according to the way it has been uttered. This degrades their performance hence producing word embeddings that may not be true representations of the interlocutors' intention in SER. The embeddings from language models (ELMo) and BERT are the two most dynamic pre-trained models for word embeddings widely used in in bimodal SER. The ELMo consists of a two-bidirectional language model with the forward pass containing information about prior words and the backward pass containing information about the word after with each producing intermediate vectors that are fed into the next layer. The weighted sum of the vectors and the intermediate vectors from the two layers make the final word representation. Recently, with the advent of the transformers [38] the BERT model has been proven to perform better in terms of text representations than all the other models and is widely used in SER [12], [39] which has made it the state-of-the-art word embedding model. The BERT pre-trained models rely on the multi-head attention mechanism for dynamic contextualization using its parallel operation to produce semantically rich and high-quality word embeddings.

C. GENERAL DEEP LEARNING TECHNIQUES USED IN BIMODAL SER

Since its advent, deep learning has been applied in a number of fields and bimodal SER is no exception. Particularly, recurrent neural networks (RNNs) have been used for ordinal and temporal tasks like bimodal SER. They consist of memory cells that keep track of information from prior sequence inputs to influence the present input and output. To alleviate the vanishing gradient problem and that of short-term memory that the vanilla RNNs would face, the long short-term memory (LSTM) was proposed in [40] to take care of both long short-term dependencies and context in sequence data like speech. The LSTM consists of cell states which enable it to have the ability to remove (forget) or add information regulated by gates. The gate that uses the hidden state and the current state of the input to decide the information to recall or get rid of is called the forget gate. The input gate is used to update the state with the information to be stored and the output gate outputs the filtered version of the cell state. In addition to RNNs, convolutional neural networks (CNNs) which are capable of capturing spatial dependencies and learning high-level representations in speech have been used in bimodal SER research. Since CNNs are good at extracting high-level features, the bimodal speech inputs (acoustic and lexical) are often subjected to them in the local feature learning block (LFLB) before the temporal feature representations are learned in the global feature learning block (GFLB). This is also partly because CNNs are good at dimensionality reduction which is important for speech data that is often of a high dimension.

Among the two models that were proposed by Yoon et al. [15], the multimodal dual recurrent encoder (MDRE) that uses dual RNNs exhibited commendable performance of 71.8% of weighted average precision. The model uses transcripts, MFFCs and prosodic features as inputs to an all-RNN two-branch model. An all-CNN model that combines phones and mel spectrogram representations was proposed in [41]. They claim that emotional cues are lost when the phonemes and mel spectrograms are converted into text or audio respectively. This model achieved an overall accuracy of 73.9% which validates their claim and proposal. Nonetheless, it is worth noting that the model is not robust on anger and happiness that belong to the same dimensional plane of emotional states. Tripathi et al. [42] also reported that among the experiments they carried out, the all-CNN model that uses MFCCs and text transcriptions vectors of the IEMOCAP dataset as inputs exhibited comparable performance of 76.1%. Their experimental results showed that regardless of the feature combinations CNNs are good at modeling high-level emotional cues in speech. However, a similar scenario that was observed in [41] of less robustness for anger and happiness still existed in these results. The robustness of these emotion categories is however different for the results reported in [15] and [43]. Among the two approaches reported in [43], the deep learning approach that involves feed-forward networks and LSTM and uses both speech and text feature inputs exhibits a commendable performance as compared to the other traditional machine learning classifiers. Though it is argued in this reference that textual features singly helped to improve the robustness of the model on the happy and angry emotions, we opine that the use of LSTM is the other reason that made this possible. This is because LSTM as opposed to CNN and DNNs considers the long and short-term dependencies of the current utterance in relation to the history of the speech sequence. The long-term dependencies discriminate the emotions irrespective of their dimensional emotional plane which improves the model's robustness. This scenario suggests that the deep learning models that can employ both RNNs and CNNs can leverage the two deep learning techniques for better bimodal SER performance. Recently, Singh et al. [44] proposed the use of 33 assorted features that depict prosody, spectral, and voice quality of audio features and ELMo extracted embeddings of transcriptions of the IEMOCAP dataset. The model uses a hierarchical deep learning-based neural network to exhibit a comparable performance of 74.5% of accuracy. Besides, bimodal SER CNNs, RNNs, and DNNs have also been applied in unimodal acoustic emotional recognition with promising results [45], [46]. We however opine that since the temporal and spatial features exist in both lexical [47], [48] and acoustic [49], [50] modalities, we find it emotionally rich to use the CNNs and RNNs simultaneously in the LFLB and GFLB to learn the high and low spatial-temporal features from the speech features in both modalities. We experimented with the concurrent feature extraction in the LFLB in [39] and commendable results that affirm our assertion were obtained.

Though these approaches report commendable results, they are weak at learning the context and semantics of the utterances/sentences which affects their robustness when applied in the real natural environment. The attention mechanisms have been applied in bimodal SER research to alleviate this challenge. In addition to dynamically learning the context of the utterances in either the LFLB, GFLB, or both, the attention mechanisms help in modeling the inter and cross-modality interactions between the test and acoustic modality in bimodal SER. A combination of CNNs, RNNs, and attention mechanisms have shown commendable performance in bimodal SER. In some cases, they have been used alone, especially transformer encoder or multi-head attention mechanisms with a commendable performance. We discuss the general attention mechanisms and their impact on bimodal SER in the next sub-section.

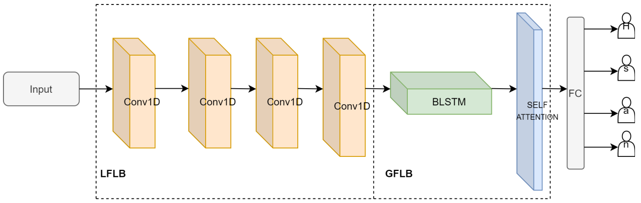


FIGURE 2. The deep learning model used to carry out feature performance evaluation experiments.

D. ATTENTION MECHANISMS USED IN SER

As argued in [25], the idea of focusing on the attention of particular speech features in SER was first suggested in [51] and [52]. In [51], consideration of maximum energy to depict prominent emotional cues was proposed. In [52], the bidirectional long short-term memory (BiLSTM) with extreme learning machine (ELM) was modeled to consider the uncertainty of emotional labels in utterances. With deep learning techniques being applied in various research areas, more robust attention mechanisms [53], [54], and [38] were proposed. In addition to consideration of long-term dependencies, the attention mechanisms compute context vectors of a given input with reference to the surrounding inputs. Additive [53] and multiplicative [54] attention mechanisms are used in combination with CNNs and RNNs separately or in combination in a sequential operation [55], [56], [57], and [58]. The attention mechanism in [38] 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 their performance. It has been deployed in [12], [59], [60], [61], and [62] among others. Because of the merits and drawbacks of the usage of attention mechanisms, we are motivated that careful use of all or a combination of some of them allows the model to take advantage of their merits limiting individual drawbacks. It is however stated in [63] that global attention mechanisms are suitable for speech emotion recognition. It is also worth noting that for the transformer-based multi-head attention mechanism, the decoder part of the transformer is omitted for SER models. Careful considerations ought to be put in mind when selecting an appropriate attention mechanism or their combination [64] for multi-modal SER [65] since it involves the fusion of emotional information from different modalities that are usually implicitly aligned.

Early or feature-level fusion, model-level fusion, and decision/late fusion are the three approaches to multi-modal fusion that exist in SER studies. 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 [66]. 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 splicing of latent representations obtained from different modality channels in order to take advantage of the feature and decision-level fusion simultaneously. It should be noted that early fusion and late fusion prevent models from learning intra and inter-modality interaction characteristics respectively [67]. Multimodal fusion involves the alignment of the features in different modalities explicitly or implicitly [68]. The explicit approach assumes prior alignment of features in order to find relationships between the different modalities. For implicit, the model learns the alignment of the different modality features progressively as it trains [69]. In the next sub sections, we present the aspects in which attention mechanisms have been applied in bimodal SER research studies in terms of alignment and fusion levels.

1) ATTENTION MECHANISMS AT MODALITY EARLY FUSION LEVEL

It is argued in [69] that implicit alignment strategy is more naturalistic than explicit alignment. Therefore, researchers have utilized deep learning techniques like attention mechanisms to implicitly align features with an aim of producing emotionally informative feature representations of multi-modalities. Though automatic speech recognition (ASR) systems are often used to output aligned features, they are more explicit than implicit. Recently implicit alignment has been achieved by applying attention mechanisms at the feature alignment level of the SER deep learning models. Xu et al. [70] used an attention mechanism to enable the model to align audio and text representations. The resultant multi-modal aligned features were combined for emotion recognition. The attention of a constituent of a sequence in terms of the other among different modalities is computed from hidden states of the two different modalities. In [71] the final hidden representation of each modality is used to compute the attention scores of each modality in relation to the other in a multi-hop attention neural network.

At the early fusion, Yoon et al. [15] argue that there is a need for additional knowledge for effective bimodal SER. The cross-attention mechanism [72] has particularly been used in SER to model contextual correlations between the two modalities and provide additional information for the task. They operate in such a way that the query is of one modality and the key and value are of the other modality for better computation of the contextual vectors among the features. In some cases, the last hidden states of the different modalities are used as query and key respectively to compute the global attention between the two modalities. Cross-attention networks apply attention weights of one modality to the other in order to align the emotional cues. In [12], the same cross-attention mechanism arrangement was used in terms of transformer encoders for each modality before concatenation with speaker embeddings for conversation emotion recognition. Furthermore, to model intra and inter-modality interactions Sun et al. [10] proposed a model that uses cross and self-attention mechanisms. Cross attention mechanism was configured to guide one modality to attend to another modality and vice versa while the self-attention mechanism was used to learn the intra-modality characteristics.

2) ATTENTION MECHANISMS AT DECISION FUSION LEVEL

Among the experiments carried out in [15], an attention mechanism was applied to compute context weights among the transcriptions in relation to the acoustic features at the decision level. This setup was compared with the no-attention set up which performed better than the earlier one. A similar setup was proposed in [16] where the attention mechanism was applied for the audio modality and not for the text before the concatenation of all modalities with the bimodal representation at the decision level. The results from these experimental setups show that the choice of attention mechanisms and how to configure them matters for commendable performance. Also, for better-aligned feature representations and since the contextual considerations are important in both modalities, we opine that attention mechanisms should be applied in both modalities. In the bid to learn a variety of emotionally rich features in speech signals (audio only) for better accuracy, an ensemble of three deep learning branches some of which use attention mechanism is also proposed in [73] which is a good study of decision level fusion.

3) ATTENTION MECHANISMS AT INTERMEDIATE FUSION LEVEL

Deep learning models that separately learn feature representations before combination at the decision level ignore the interaction between the two modalities. This arrangement does not allow the different modalities to interact explicitly. On the other hand, early fusion also splices features without the model explicitly learning the emotionally rich intra and cross-modality interactions. On the contrary, intermediate or model-level fusion splices resultant representations of the different modalities after learning them which allows the model to understand both the inter and cross-modality features and continue learning other features after concatenation. A model that pays attention to audio and visual features at each time step was proposed in [74]. This model concatenates feature representations obtained from both early and decision-level fusion using learnable attention weights for emotion prediction. Poria et al. [75] proposed multi-level attention for audio, video, and text modalities. In their approach after feature extraction from the utterances, the feature representations are fused using an attention network and the resultant representation is used in the step that follows which learns new utterances representations using LSTM and another attention mechanism before sentiment classification. Zheng et al. [12] proposed a conversation emotion recognition model (CTNet) that fuses the speaker embeddings and features learned from single and cross-modality transformer encoders at the intermediate level that exhibited a commendable performance. This proves that the model benefits from the intermediate fusion of the intra and inter-modality characteristics without ignoring their interactions. This approach motivated us to propose the CoSTGA model [39] which at the intermediate level learns and fuses temporal, spatial, and semantic features at multi-levels for bimodal SER. This model showed commendable performance when validated on the IEMOCAP dataset.

TABLE 2. Ranges of parameters used in the experiments.

Parameter	Range of Values
Optimizer	Adam
Learning rate	0.0004 - 0.001
Batch size	32 & 64
Epochs	50 - 100
Number of heads	2 - 4
Embedding dimension	256
Kernel Regularizer	L2 - 0.00001
Kernel Initializer	Glorot Uniform Initialization
Dropout rate	20% - 50%

However, we argue that with the recent progress in bimodal SER research fueled by the use of transformer encoders and the multi-head attention mechanism which is its basic building block, attention-based models will exhibit better performance without configuration of any other deep learning techniques. This is however possible with a well-thought choice of the modality features, their alignment and appropriate objective functions.

Nonetheless, attention mechanisms result in complex models that are at risk of over fitting during training especially because of the scarcity of speech emotion datasets. In addition, since emotion cues don't appear in the whole sequence of utterances it is possible for attentive models to focus on irrelevant or noisy parts of speech. This is amplified if the data contains biases that may influence the training. Due to the complexity, the models that involve attention mechanisms can also be challenging to interpret especially in terms of understanding why the model attended to specific regions. To avoid the repercussions of complexity that eventually cause training instability, we used optimization, data augmentation, and regularization techniques in our experiments in this paper. We also used ablation studies to ascertain the impact of models with and without attention mechanisms. In addition, some attention mechanisms that operate in sequences may lead to high memory usage which makes it challenging to deploy the models on resource-constrained devices. For this reason, we used self and multi-head attention mechanisms in our experiments since these involve a parallel operation that may not be as memory-intensive as the sequential counterparts. It should also be noted that multi modality model performance is sometimes affected by loss of intra-modality information during the integration of modalities. This challenge is often solved by use of appropriate model fusion strategies discussed earlier and utilized in our experiments. This challenge can also be addressed by use of synchronous and asynchronous representations as used in [76]. To solve most of the challenges of attention mechanisms transfer learning is often utilized. The authors of [77] and [78] solved the possible challenges by use of transfer learning with commendable results. Though transfer learning offers a number of advantages that include improved performance, faster training. domain adaptation and data scarcity solutions, it has some disadvantages. Transfer learning is more applicable when the source and target are related in order to achieve significant performance. In SER studies where there are small datasets, it may not be of benefit to fine tune the pre-trained models since it will cause over fitting. Transfer learning does not also work if there is a domain mismatch. It should also be noted that most of the models are highly parameterized which may distort the training and eventual deployment in low resource devices. This is in addition to not being easy to interpret and dependency on pre-trained models for commendable performance.

III. EXPERIMENTS

In this section, we describe the experiments carried out to further analyze the reviewed concepts in this paper. We carried out experiments to ascertain the effect of three commonly used acoustic features in SER research works. We also present experiments on the use of the different common deep learning techniques and their combination. We eventually propose a multi-technique model called the deep learning-based multi-learning model for emotion recognition (DBMER) that operates with multi-learning capabilities of CNNs, RNNs, self and multi-head attention mechanisms. We evaluated the performance of the proposed DBMER MODEL using the IEMOCAP dyadic datasets described in the previous section. We considered happy, sad, angry and neutral as the emotion categories for these performance evaluation experiments. To avoid challenges that come with the imbalanced nature of the datasets, we configured class weights as functions of the smallest class. Table 2 shows a summary of the parameter ranges used in these experiments.We used Keras 2.6.0 API, TensorFlow 2.6 as the back-end with python programming, and Nvidia GeForce RTX 2080 super graphics processing unit (GPU).

A. EXPERIMENTS ON ACOUSTIC FEATURES IN SER RESEARCH

The choice of the features used for SER models greatly determines their performance. We carried out experiments using the German dataset of Berlin (EMODB) [79] which is a purely unimodal acoustic dataset to show the challenges that exist if the choice of features is not done carefully. As earlier mentioned, each of the feature categories is robust at some dimensional emotions but performs less on the others. We carried out experiments on the performance comparison of the individual mel spectrograms, MFCCs, and a combination of them in addition to chroma grams and found out that a combination of prosodic, spectral, and voice quality features provides a more robust performance as compared to individual features. In these experiments, we considered features that can depict loudness, pitch, and quality of sound. The MFCCs and chroma grams were extracted as spectral features and mel spectrograms as voice quality features. The mean value of these features extracted from each frame was calculated to obtain the high-level statistical features (HSFs) and was separately used as input to the model in the first experiments. In the other two experiments, a combination of either MFCCs and mel spectrograms or MFCCs, chroma grams, and mel spectrograms were used as input to the model after concatenation.

We carried out the experiments using the model shown in Fig. 2 that consists of four convolution layers each with pooling layers where necessary for local feature extraction and self-attention and bidirectional layers of 64 units that were used for global feature learning before the feature representations are fed into a dense layer and a subsequent softmax layer for classification. The self-attention mechanism was configured in order to further consider the global context of the speech representations. We used the exponential linear unit (ELU) as the activation function.

B. EXPERIMENTS ON THE USE OF DIFFERENT DEEP LEARNING TECHNIQUES IN SER RESEARCH

In this subsection, we present the most significant experiments we carried out on the different deep-learning techniques commonly used in SER Research. The experiments were carried out with the goal of ascertaining the significance of; bimodal SER compared to unimodal SER, the use of single deep learning techniques, and the use of a combination of deep learning techniques. In dyadic speech experiments, we used the IEMOCAP dataset to evaluate the models in these experiments. The MFCCs were extracted from the speech signal and used as the acoustic features while the BERT pre-trained model was used to compute the lexical feature vectors for the text modality.

Because of considering long-term dependencies in sequential tasks LSTM [40] and its variants like the BiLSTM have been proposed in most SER studies. It is against this premise that we also chose to use the BiLSTM technique to ascertain the significance of bimodal SER compared to unimodal SER in the first experiment. In this experiment, the model which is composed of two BiLSTM layers of 128 units each, two dense layers with a softmax layer for classification was separately configured for both acoustic and lexical modalities. Layer normalization was configured with a dropout regularization rate of 0.5. The Layer normalization was to ensure a similar scale of the activations from the LSTM layers and therefore stabilize the training process to improve the performance of the model.

We also carried out two different experiments to find out the significance of the use of only one deep learning technique for bimodal SER. Particularly, we configured multi-learning models with either only BiLSTM layers or transformer encoders (TED) which use multi-head attention for both acoustic and lexical branches before splicing them to form inputs for the global feature learning block (GFLB) in an intermediate-level fusion strategy approach. All the BiLSTM layers had 256 units with dropout regularization and layer normalization. For the transformer encoders, we used the positional encoding stated in [38] for the lexical modality and one convolution layer of 128 filters for the acoustic modality. Four heads were used for multi-head attention mechanism in each transformer encoder with layer normalization and dropout regularization of 0.5. Another experiment that resulted into the proposed DBMER model was also carried out. This experiment was aimed at ascertaining the significance of a combination of all the reviewed deep learning techniques with multi-level intermediate-level fusion. The framework of the proposed DBMER model is described in the next subsection. In all these experiments, the BiLSTM layers are configured with the hyperbolic tangent activation unit while the rectified linear unit (ReLU) is used for all the other layers. We also carried out experiments to ascertain the generalization capabilities of the model in multi stream acoustic and non-dyadic SER scenarios.

C. THE PROPOSED DBMER MODEL

As shown in Fig. 3, the proposed DBMER model is a multi-learning model that accomplishes the task using CNNs, RNNs, self, and multi-head attention mechanisms at two subsequent intermediate fusion levels. To learn high-level features of both the acoustic and lexical modalities the extracted features were subjected to two convolution layers of 128 filters with a kernel size of 3, L2 regularization of 0.00001 and the weights were initialized using the Glorot uniform initialization. Each of the learned high-level features is fed into the transformer encoders with a configuration described in the previous subsection. These features learned in each modality branch are spliced at the first level fusion before being fed into a convolution layer of the same configuration that learns the inter-modality high level feature representations. The learned features at this stage are subjected to learning of long-term dependencies among them using the BiLSTM whose configuration is as described in the previous section. However, the BiLSTM in the proposed DBMER model differs from the one used in the previous section since self-attention is configured with it to allow the model to compute the contextual relationships of the inter-modality representations. A similar BiLSTM and self-attention mechanism layer arrangement is subjected to individual modalities to learn the long-term dependencies and their contextual relationships before the second intermediate-level fusion of the individual modality representations with the bimodal representations. After the second fusion level, the resultant feature representations are fed into a transformer encoder to further compute the relationship between them before being subjected to the classification of the emotional states that is done by the softmax layer. It should be noted that throughout this model, dense layers of 128 units are used to ensure the same dimensions for concatenation where necessary.

D. EXPERIMENTS ON THE GENERALIZATION CAPABILITIES OF THE PROPOSED DBMER MODEL

To evaluate the proposed DBMER model's generalization capabilities and performance in the real-world scenarios, we evaluated the proposed DBMER model on datasets that involve presence of noise, variations in speech, and other environmental factors and non dyadic speech.The datasets

TABLE 3. The datasets used in the generalization experiments.

Dataset	Language	Gender	Kind	Samples Used	Emotions	Emotions Used
RAVDESS [80]	EN	F & M	Acted	1440	8	7
SAVEE [81]	EN	F & M	Acted	480	7	7
TESS [82]	EN	F	Acted	2800	7	7
CREMA [83]	EN	F & M	Acted	7442	6	6
ASVP [84]	CH, EN,FR,RU,OT	F & M	Natural	13829	12	6

EN: English, CH: Chinese, FR:French, RU:Russian, and OT:Others. F: Female, M: Male

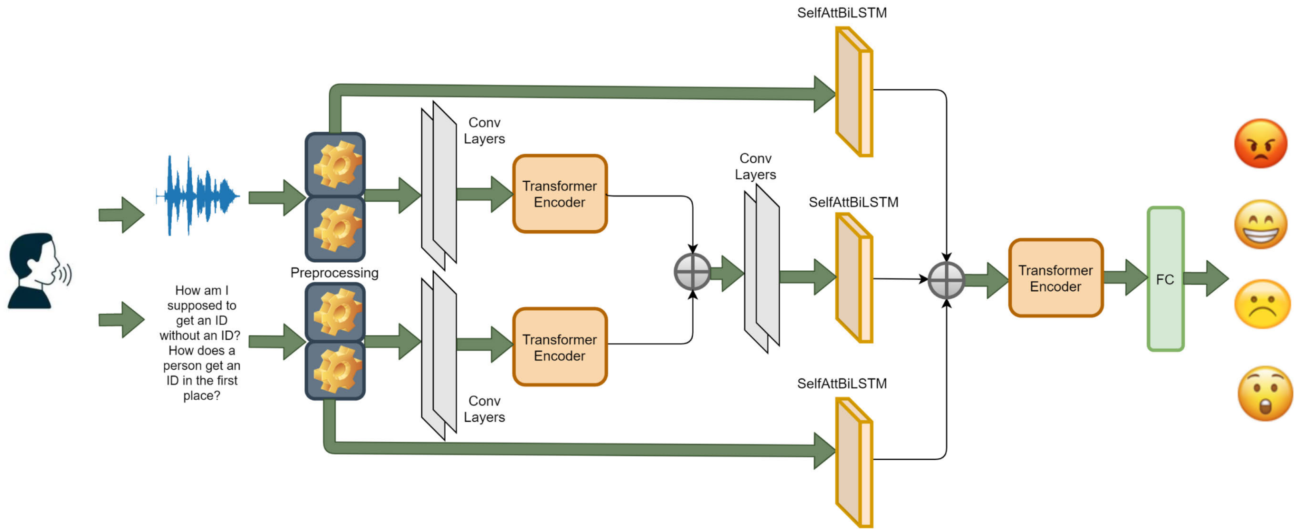


FIGURE 3. The proposed DBMER bimodal SER framework.

that depict real-world scenarios were used. For the multi-stream SER, we used the Ryerson audio-visual database of emotional speech and song dataset (RAVDESS) [80], surrey audio-visual expressed emotion (SAVEE) [81], Toronto emotional speech set (TESS) [82], crowd-sourced emotional multimodal actors dataset (CREMA) [83] and the audio, speech, and vision processing lab emotional sound database (ASVP) [84]. Due to the data scarcity issues alluded to earlier, we combined RAVDESS, TESS, and SAVEE since they were collected in a similar controlled manner and they all contain speech data spoken in the English language with differences in accent and gender distribution. The speech data in the CREMA datasets was collected in the English language from a variety of races and ethnicities that included African American, Asian, Caucasian, Hispanic, and Unspecified. We also chose to evaluate the model on the ASVP dataset released by the South China University of Technology because it contains speech and non-speech emotional data. This data was collected from movies, TV shows, YouTube channels, and other websites in Chinese, English, French, Russian, and other languages which makes the ASVP more realistic and non-scripted with no language restriction. The details of the datasets used in these generalization experiments are shown in Table 3. In the multi stream SER experiments we replaced the text modality with melspectrograms generated from the speech signals using librosa 0.9.0. Due to the data scarcity challenges, we carefully apply data augmentation by adding noise to the acoustic data and then extracting the MFFCs and mel spectrograms from the noisy data. To ensure fixed sizes, we keep track of the maximum number of frames and apply padding. Since we proposed this model for dyadic speech we also attempted to evaluate it on MELD non-dyadic speech datasets.

TABLE 4. Performance analysis of the common acoustic features in bimodal SER models.

Input	A(%)	F1(%)	CH(%)	CS(%)	CA(%)	CN(%)
Raw signal	81.18	80.52	07	92	99	71
Mel	80.00	78.91	73	79	67	60
MFCCs	85.46	85.71	40	93	88	87
Mel & MFCCs	94.55	95.48	40	100	100	93
All	89.09	87.77	53	71	88	87

IV. RESULTS

In this section, we present the results of our experiments. we present the results that show the effect of three different acoustic features and their combination commonly used in SER research. The results that depict the benefits of the combination of CNNs, BiLSTM, self, and multi-head attention that constitute the proposed DBMER model for dyadic bimodal SER are also presented. We also describe the evaluation results of the proposed DBMER model in terms of its generalization capabilities. The results are presented

TABLE 5. Performance of the different SER deep learning techniques on the IEMOCAP dataset.

Model	Modality	Fusion Level	UA(%)	WA(%)	P(%)	R(%)	F1(%)	Loss
BiLSTM	Acoustic (A)	-	58.70	-	60.08	54.63	53.50	1.3956
BiLSTM	Lexical (L)	-	61.23	-	42.13	85.83	56.68	0.9707
BiLSTM	A + L	Multi	70.13	66.70	70.54	69.11	69.90	0.9284
TED	A + L	Single	61.27	50.84	65.15	56.33	60.40	1.0488
TED + BiLSTM	A + L	Single	61.01	55.60	63.69	57.72	60.61	1.0057
Proposed DBMER	A + L	Multi	74.18	74.80	77.02	70.00	73.30	0.7352

TABLE 6. Individual confusion ratio of the different deep learning techniques used for bimodal SER.

Model	Modality	Fusion Level	CA(%)	CH(%)	CN(%)	CS(%)
BiLSTM	A + L	Multi	67	56	74	70
TED	A + L	Single	36	31	44	92
TED + BiLSTM	A + L	Single	40	50	36	95
Proposed DBMER	A + L	Multi	79	71	68	81

CH, CS, CA, and CN are confusion ratios for happy, sad, angry, and neutral respectively.

in terms of unweighted accuracy (UA), weighted accuracy (WA), precision (P) recall (R) and F1 score (F1). For a more informative analysis of the capabilities of these models, we also present the losses they exhibit.

A. EXPERIMENTAL RESULTS ON ACOUSTIC FEATURES

The results of experiments on acoustic features are summarized in Table 4 which shows the performance in terms of accuracy (A) and F1 score (F1) obtained by the simple model. We also present comparative results of the robustness of the model in terms of the confusion ratio of the different classes of emotions for each input. The results show that for all the inputs, the accuracy and F1 score can be commendable however, the robustness especially in terms of the confusion error of high arousal dimension emotions needs to be given more attention. It should however be noted that the performance of MFCCs highly depends on the choice of parameters like the number of filters in the filter bank discussed earlier in the extraction process.

B. EXPERIMENTAL RESULTS ON DEEP LEARNING TECHNIQUES

Results of the Dyadic Bimodal SER Experiments including the Proposed DBMER Model: The results of the experiments described in Section III about the use of different deep learning techniques including the proposed DBMER are presented in Tables 5 and 6. Table 5 also shows the fusion levels and modalities involved in the different experimental models. Table 6 presents the confusion ratios of the different emotional states. These results show how robust the models are at detecting independent emotions. It should be noted that the fusion strategy used for all experiments is the intermediate-level fusion since it was proved to be the best strategy for bimodal SER in the reviewed literature. From the tabulated results, we observe that the bimodal model that uses BiLSTM exhibits comparable performance with the proposed DBMER model that combines all the deep learning techniques. Therefore, we present the confusion matrices of these two models in Fig. 4 for analysis of their robustness.

Results of the Generalization Capability Experiments of the Proposed DBMER Model: The results of the proposed DBMER model evaluation on datasets that depict different real life scenarios are presented in Table 6. These results show an outstanding performance exhibited by the model on a broad spectrum of languages, accents, gender and cultures. The results also show the robustness of the model in different environments since the datasets considered were collected in different real life scenarios. To further facilitate the analysis of the generalization of the proposed DBMER model, we present the confusion matrices obtained when the model is exposed to unseen data in Fig. 5.These confusion matrices continue to show the generalization of the proposed DBMER model. This data is from ASVP datasets which consists of a minimum of five languages and a combination of RAVDESS, SAVEE and TESS which depicts American and British English accent. For the MELD non-dyadic datasets, the proposed DBMER model exhibits an average performance compared to the dyadic IEMOCAP datasets. This is due to the rapid changes of emotion cues as interlocutors and their voices change in a sequence of utterances of non dyadic speech which may need a specialized model.

V. DISCUSSION

A. DISCUSSION OF RESULTS OF THE EXPERIMENTS ON ACOUSTIC FEATURES

The experimental results show that models that use raw signals can achieve a commendable accuracy and F1 score however, they are not robust in terms of discriminating the high arousal emotion states of happy and angry. This is partly because the emotional cues of happy and angry are similar in terms of emotional dimension. Therefore, robust models that aid complex speech signal processing are required if they are to be used in SER systems. The experiments also show that mel spectrograms which depict voice quality cues in a speech signal are quite robust for happy and sad however the model still does not perform well especially for the neutral and angry emotions that tend to be confused with all the other emotions. On the other hand, MFCCs can be used by models if the interest is to achieve robustness for sad, angry and neutral however, the models that use them still confuse happy and other emotions especially anger with which they belong to the same plane. Moreover, a combination of MFCCs and mel spectrograms improves the robustness results further to as high as 100% for sad and angry but the confusion ratio for happy remains the same. A combination of MFCCs, mel spectrograms and chroma grams that takes the pitch of sound into consideration improves the confusion ratio of happy but there is need for its robustness for the other emotions compared to the model that uses a combination of MFCCs and mel spectrograms. These results show that, in terms of the robustness of deep learning-based SER systems, models that use a combination of features perform better than those that either use a single kind of features or those that use raw signals. It should however, be noted that for all the inputs, the accuracy and F1 scores are commendable which further suggests that accuracy and F1 score are not enough for SER studies especially for deployment in real-life situations. This informs why in addition to accuracy and F1 score, it is necessary to analyze the individual confusion ratio of each emotion class.

B. DISCUSSION OF RESULTS OF THE PROPOSED DBMER MODEL

1) PERFORMANCE OF THE PROPOSED DBMER MODEL FOR DYADIC BIMODAL SER

From the results presented in Tables 5 and 6, it is observed that bimodal dyadic SER performs better than unimodal SER. These results show that a combination of both acoustic and lexical features is emotionally richer than either of the modalities. This is especially evident in terms of F1 score and loss values which indicates that the bimodal models learn more emotionally rich features and are able to distinguish between the emotional states accurately as compared to unimodal models. However, it is noted from the results that the lexical features perform better than the acoustic features. This is because the lexical features carry semantic and grammatical cues of the utterances which help the model to easily learn the possible emotional states as compared to the acoustic modality. This is further evidenced by the differences in the loss values with the acoustic modality model having a loss of 1.3956 compared to the lexical modality model's loss of 0.9707. However, in terms of bimodal SER, it is also observed from the results that multi-level fusion is more beneficial than single-level fusion. This confirms the conclusions drawn in [12] and [39] in which multi-level intermediate-level fusion strategy was implemented for bimodal dyadic SER.

In terms of the deep learning techniques, the experimental results show that though models that use BiLSTM and transformer encoders (which use only multi-head attention) are good techniques for SER, they don't perform well when applied singly. However, because of the sequential modeling capability of the BiLSTM, it performs better than the transformer encoder for this task. This is observed in all the metrics used with a difference of 8.86% of unweighted accuracy, 15.86% of weighted accuracy and 9.50% of F1 score. The loss increases from 0.9284 to 1.0488. A model that combines BiLSTM and Transformer encoders at single level fusion does not perform better than a model that uses only BiLSTM with multi-level fusion. This can be explained since understanding the context without analyzing the long-term dependencies is not sufficient enough to infer emotions in addition to the benefits of multi-level fusion which improves the accuracy and robustness of the models.

TABLE 7. Performance evaluation of generalization capabilities of the proposed DBMER model.

Datasets	UA(%)	WA(%)	P(%)	R(%)	F1(%)	Loss
RAVDESS	93.65	94.03	94.40	93.65	94.04	0.3390
CREMA	84.00	84.00	84.50	83.60	84.24	0.7934
ASVP	84.40	83.40	84.97	83.84	84.41	0.8696
RAVDESS + SAVEE + TESS	93.40	93.42	94.86	92.16	93.44	0.3698

Well knowing that CNN is better than RNNs and attention mechanisms in learning high-level features we chose to use the BiLSTM, self and multi-head attention mechanisms in combination with CNNs. This model is named the deep learning-based multi-learning model for emotion recognition (DBMER) which we propose in this paper. The proposed DBMER model's performance showed that a careful combination of all the common deep learning techniques coupled with multi-level fusion benefits bimodal dyadic SER as compared to the other approaches experimented on earlier. The performance in terms of the considered metrics is improved compared to the all-BiLSTM model tested in similar conditions. The unweighted and weighted accuracy improved from 70.13% to 74.18% and 66.70% to 74.80% respectively. The loss exhibited by the proposed DBMER model is reduced from 0.9284 to 0.7352. This confirms that the proposed model learns emotionally rich cues to be able to infer emotional states. To confirm these observations, an analysis of the robustness of the model compared to the other approaches is presented in Table 6. It is observed that, compared to the other approaches the proposed DBMER model is uniformly robust for the four individual emotions depicted from dyadic speech. There is a tremendous improvement in the confusion ratios of the proposed DBMER model compared to the single-level fusion models that use the transformer encoder or in combination with the BiLSTM. The prediction analysis of the BiLSTM model and the proposed DBMER model is shown in the confusion matrices presented in Fig. 4. It is further affirmed that the proposed DBMER model is more robust than the BiLSTM model due to the multi-technique learning approach that benefits the model with emotionally rich cues in terms of the high-level

TABLE 8. Performance analysis of the the proposed DBMER on the datasets used in the generalization experiments.

Dataset	CA(%)	CD(%)	CF(%)	CH(%)	CS(%)	CP(%)	CN(%)
RAVDESS	100	91	95	94	96	100	82
CREMA	87	78	71	93	82	-	93
ASVP	86	88	86	80	82	-	78
RAVDESS + SAVEE + TESS	92	91	92	89	92	97	88

features learned by the CNNs, the long-term dependencies learned by the BiLSTM and the context computed using the self and multi-head attention mechanisms at all levels.

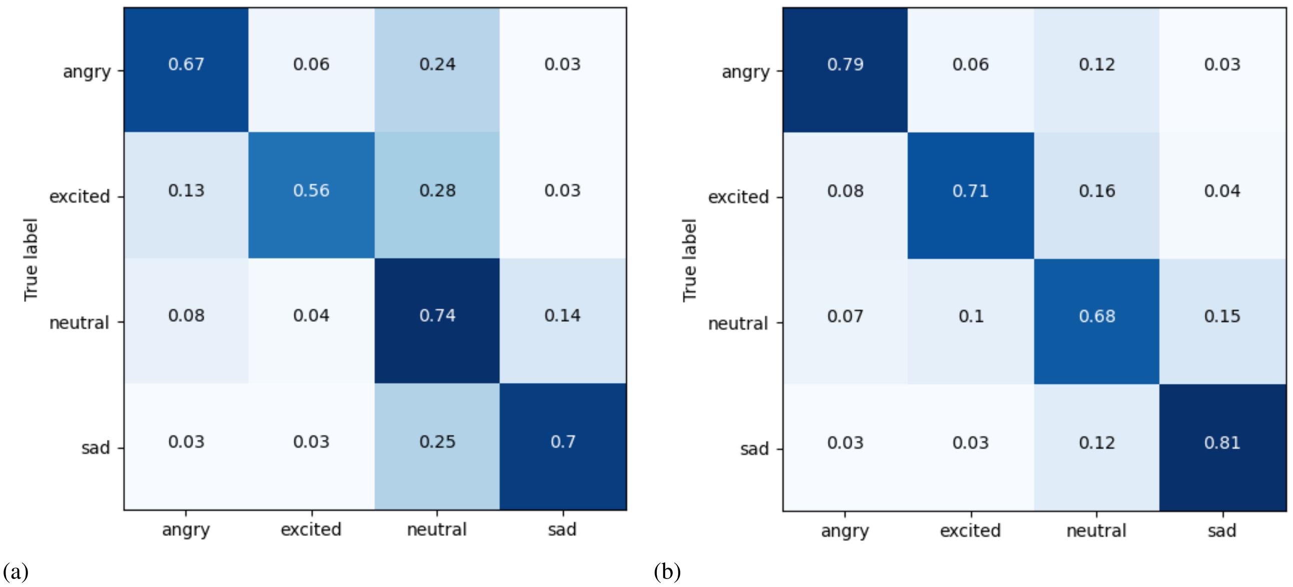
2) PERFORMANCE OF THE PROPOSED DBMER MODEL IN TERMS OF ITS GENERALIZATION CAPABILITIES

As alluded to in the results section the proposed model is robust on all the datasets it was evaluated on in a multi stream approach. The purpose of the experiments was to evaluate how the model performs in real-world scenarios, where noise, variations in speech, and other environmental factors can affect its performance. Because of the careful combination of the deep learning techniques and the ability of the proposed DBMER model to learn both intra and inter modality representations at two fusion levels, the model learns most of the cues it needs to be robust in any condition it is faced. The local features learned by the CNN, long term dependencies learned by the BiLSTM and the context learned at all the levels of the model makes it robust in presence of noise, accent, gender and cultural issues evident in the multi-language ASVP datasets, the multi-cultural CREMA datasets and the combination of RAVDESS, SAVEE and TESS datasets. An analysis of the confusion ratios of the different emotions considered in the generalization experiments is presented in Table 8. In this table CA, CD, CF, CH, CS, CP and CN are confusion ratios for angry, disgust, fearful, happy, sad, surprised, and neutral respectively. These results show that the model uniformly recognizes all the emotions and predicts them with minimal confusion. We however opine that the model is more robust on dyadic and purely acoustic data but less robust on non-dyadic speech data because of the rapid changes in emotion cues as more than two interlocutors participate in a speech or conversation which complicates the training. We never carried out cross corpus experiments because as [85] suggests it is obvious that there will be a challenge of feature distribution discrepancy that will affect the SER performance besides the need for diverse datasets that are non-existent. A solution to this problem could be domain adaptation and other transfer learning strategies that are out of scope of this paper.

C. COMPARISON OF RESULTS OF THE SOTA DEEP LEARNING MODELS

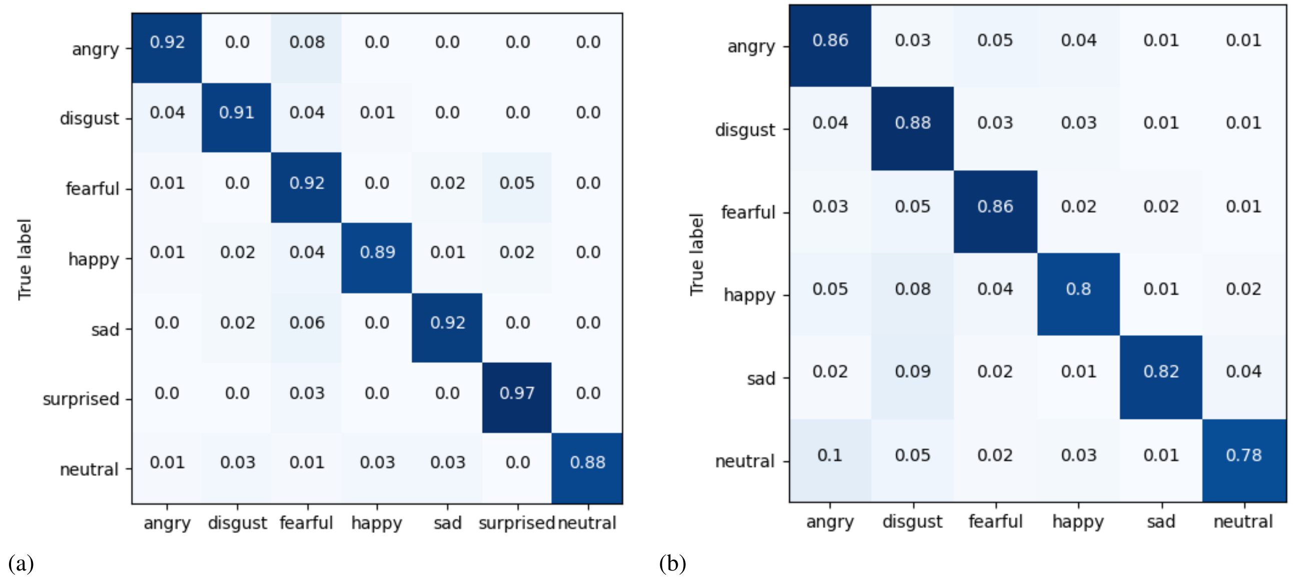
In this section, we present a comparison of the results of recently proposed bimodal SER models that use different deep-learning techniques. Table 9 presents these results in terms of weighted accuracy (WA), unweighted accuracy (UA), and F1 Score (F1) with details of datasets, the deep learning techniques, and fusion strategies used.

From the results we sampled in recent bimodal SER literature shown in Table 9, it is evident that researchers have deployed deep learning techniques to enhance performance in emotion prediction. However, it is also evident that the reported performance depends on the dataset and the number of emotion categories chosen. The common datasets for dyadic bimodal SER is IEMOCAP dataset. We however present a few studies that attempt to incoporate non dyadic bimodal SER using CMU-MOSEI, and MELD datasets. Each of the datasets has merits and demerits as discussed earlier and should be carefully considered to avoid models that exhibit good performance in laboratory experiments and poor performance on deployment. There also exist many emotions in daily life however, only four (sad, happy, neutral, and angry) are commonly used in literature. The sampled results also show that a careful combination of the different deep learning techniques benefits automatic speech emotion prediction. As an example, attention mechanisms have been shown to benefit every model in which they are deployed except in [15] where the model without attention performed better than the attention-based models. This performance could be because of the way the attention mechanism was deployed in this paper where the attention is computed as a similarity score between the text and audio representations yet the intra-modality features also need attention since some utterances are more emotionally rich as compared to others. Nonetheless, we note that the advent of attention mechanism techniques especially self and multi-head attention mechanisms improve robustness in SER systems. This is because in real life emotions are inferred from contextual speech. Therefore, an attempt to compute the context in audio or text or both modalities should exhibit better performance. Among all the attention mechanisms, Multi-head attention that operates in a parallel and dynamic manner for all the utterances improves the execution speed as well as computing the pre and prior contexts in a speech signal sequence. However, recently [92] showed that transfer learning could be utilized without the use of attention mechanisms with comparable performance. This is because attention mechanisms especially multi-head attention takes a lot of time to train and execute yet there is insufficient labeled SER data that the long training requires. This may result into complexity problems explained in Section II. To alleviate these challenges, the authors of [77] and [78] applied transfer



(a) (b)

FIGURE 4. The confusion matrix results for dyadic bimodal SER. (a) BiLSTM. (b) DBMER.



(a) (b)

FIGURE 5. The confusion matrix results for the generalization experiments of the proposed DBMER model. (a) A combination of RAVDESS, SAVEEE and TESS datasets. (b) ASVP datasets.

learning either in one branch or both to register commendable performance. The results in [65] inform researchers that the choice of the cost and/or loss function determines the performance of the proposed model. They chose to use the additive angular margin loss (ArcLoss) primarily used in face recognition to achieve state-of-the-art comparable results. ArcLoss focuses on the angles between features and weights to achieve comparable performance. The approach proposed in [94] configures cross attention between the speech and text features. This informs why their models perform slightly better than our proposed DBMER model. It should also be noted that we compared our proposed DBMER model with models that include techniques other than those that we reviewed and subsequently used to constitute it. The addition of these techniques that we don't use in our experiments is to show that these deep learning techniques can exhibit an even better performance when enhanced with other techniques. Besides this, our emphasis was on the possibility of using the reviewed deep learning techniques in line with suitable fusion strategies.

TABLE 9. Comparison of results of the SOTA deep learning models.

Ref.	Dataset	Techniques	Fusion Level	No. of Emotions	WA(%)	UA(%)	F1(%)
[15]	IEMOCAP	Dual RNNs	Decision	4	-	71.8	-
	IEMOCAP	Dual RNNs + Attention	Decision	4	-	69.0	-
[70]	IEMOCAP	BiLSTM + Attention for Alignment	Intermediate	4	72.5	70.9	-
[85]	IEMOCAP	TDNN + BiLSTM +Pretrained models for text	Decision	4	73.5	71.0	-
[16]	IEMOCAP	CNN + BiLSTM + Attention	Feature and Decision	4	71.06	72.05	-
[86]	MELD	Transformer based Attention + Graph Neural Networks	Feature	7	61.8	-	-
[14]	IEMOCAP	RNN + Multi-Head Attention	Intermediate	4	74.33	73.23	73.77
	CMU-MOSEI	RNN + Multi-Head Attention	Intermediate	7	99.19	99.19	-
	MELD	RNN + Multi-Head Attention	Intermediate	7	59.94	63.26	59.66
[44]	IEMOCAP	Hierarchical Deep Neural Network	Feature	4	-	74.5	-
[87]	IEMOCAP	CNN based multi stage fusion	Intermediate	4	-	72.6	-
	MSP-Podcast	CNN based multi stage fusion	Intermediate	4	-	56.0	-
[10]	IEMOCAP	Self and Cross Attention	Intermediate	7	61.2	56.0	-
	MELD	Self and Cross Attention	Intermediate	7	-	-	59.2
[12]	IEMOCAP	Multi-Head Attention + BiGRU	Intermediate	4	83.6	-	83.8
	IEMOCAP	Multi-Head Attention + BiGRU	Intermediate	6	68.0	-	67.5
	MELD	Multi-Head Attention + BiGRU	Intermediate	7	62.0	-	60.5
[65]	IEMOCAP	BiLSTM + Memory Compressed Attention + GRU + ArcLoss	Intermediate	7	72.8	62.5	-
	IEMOCAP	BiLSTM + Memory Compressed Attention + GRU + ArcLoss	Intermediate	4	82.4	80.6	-
[88]	IEMOCAP	Self Attentional BiLSTM and Multi channel CNN (MCNN)	Feature and Decision	4	74.98	75.05	-
[89]	MELD	CNN + LSTM + Meaningful Neural Network (MNN)	Feature and Decision	7	-	86.69	-
[90]	IEMOCAP	CNN + BiLSTM + Attention + Autoencoders	Intermediate	4	74.8	-	-
	CMU-MOSI	CNN + BiLSTM + Attention + Autoencoders	Intermediate	2	79.85	-	-
	MELD	CNN + BiLSTM + Attention + Autoencoders	Intermediate	7	63.85	-	-
[39]	IEMOCAP	Multi-level fusion with DCC + BiLSTM + Multi-head and Self Attention	Intermediate	4	75.50	75.82	75.57
[91]	IEMOCAP	Transfer Learning with RoBERTa and Inception ResNet-V2	Intermediate	4	72.8	-	-
	CMU-MOSEI	Transfer Learning with RoBERTa and Inception ResNet-V2	Intermediate	6	99.2	-	-
	MELD	Transfer Learning with RoBERTa and Inception ResNet-V2	Intermediate	7	63.8	-	-
[92]	IEMOCAP	CNN + BiLSTM + Cross Attention	Decision	4	80.51	79.22	-
[93]	IEMOCAP	CNN + Transformer + Score fusion	Decision	4	73.5	73.0	-
[94]	IEMOCAP	BiLSTM + Self Attention with weight correction and confidence measures	Intermediate	4	76.6	76.8	-
[77]	IEMOCAP	CNN + Attention with only transfer learning in the text branch	Decision	4	85.5	80.7	-

The sampled results showed in Table 9 also conform with the assertions in [12] and [39] that intermediate fusion of text and audio representations benefits the bimodal SER systems compared to using either feature or decision level fusion singly. It should however be noted that the models that use both feature-level and decision-level fusion also perform well. This is because intermediate-level fusion is a hybrid of both feature and decision-level fusion that only combines representations in their intermediate form. This is usually coupled with intra-modality representations that enable repetitive learning for better performance.

D. COMPLEXITY AND PERFORMANCE IN SER MODELS

Though combining multiple deep learning techniques is beneficial to the SER models, there is an increase in the complexity of such models. The complexity brought about by the increase in trainable parameters can potentially improve the model performance however there is a scarcity of datasets for SER studies to sustain the training of such highly parametrized models without overfitting or causing gradient descent challenges. Due to this scarcity, it's advisable to use low-parameterized models to avoid overfitting. However, we found out that the use of pre-trained models can improve the performance of the multiple deep-learning technique models. This was evident when we used the BERT embeddings in the text branch which increased the total number of parameters to 116,600,969 with only 7,116,424 trainable parameters registering a good performance for four emotions from dyadic speech. The models that did not include the BERT embeddings had only 5,585,159 total number of parameters and 5,582,855 training parameters which did not negatively affect the performance significantly. It should however be noted that in order to achieve good performance of complex multi-technique models with a small number of datasets, regularization techniques like dropout, L2 regularization, gaussian noise addition, and other data augmentation strategies should be utilized. In addition, the use of transfer learning to improve the performance of complex multi-technique SER models is a good solution to this challenge as observed from the results in [77] and [78]. On the whole, though the models appear complex, the strategies configured to optimize the training determine their performance.

E. CHALLENGES OF BIMODAL SER AND FUTURE WORK

Bimodal dyadic SER has improved to a great extent with the advent of deep learning. Nonetheless, a number of issues need to be addressed by researchers to be able to deploy models in natural environments in real-time. Most of the literature available models bimodal dyadic SER as a non-multilingual task yet the real world is not only multilingual but multi-cultural with different accents and emotional behavior. This derails the development of off-the-shelf models for SER applications and needs to be given attention by researchers. We have attempted to assess the generalization of our proposed DBMER model for similar aspects in a multi-stream approach and a commendable performance was registered. Similarly, cross-corpus bimodal SER remains a challenge. It is suggested in [85] that there is a feature distribution discrepancy that affects the SER performance besides the need for diverse datasets that are non-existent. A solution to this problem could be domain adaptation and other transfer learning strategies that are out of the scope of this paper. Also, little work has been done in this aspect due to the lack of an all-round corpus that can enable multilingual and multi-cultural SER studies. In [96], the authors only related data collected from YouTube as the source dataset, and the IEMOCAP dataset was used as the target. It however remains a challenge to train and test models of this nature or even consider the source and target datasets of different languages with commendable results. In addition, limited datasets coupled with the complexity of speech pose challenges to bimodal SER studies. This leads to models with commendable performance in laboratory experiments but poor results in real environments. The ambiance of the environments in which the models are to be deployed also needs to be given attention by researchers from the time of data acquisition to modeling and design with noise and voice perturbations as major aspects found in the real environment. In addition, the purpose of the models being proposed should be clear since there are many dyadic and non-dyadic bimodal SER use cases in the real world.

Ethical issues are an important aspect during the development and deployment of artificial intelligence (AI) systems like SER and ought to be considered. There should be consent from the participants during the emotional speech data collection. The collected data should be safeguarded from privacy infringement. Developers and users should be aware of the environmental conditions in which the data was collected and its characteristics before it is used to train any models. This is because biases related to gender, race, accent, culture, and the environment affect performance when deployed in the real world. In this paper, we carried out an evaluation putting all these discriminatory distributions into consideration especially during the choice of the datasets to use for experiments. Another emerging ethical issue that will require specialized models to support SER systems is an assessment of the emotional impact on a human being after an emotion is recognized and revealed to him/her by a machine. Music recommendation systems have been proposed in [97] and [98] to improve or maintain the users' mood upon facial emotion detection. This should be done for all AI systems to handle the impact that HCI systems can have on humans. Recently, the authors of [99] proposed to recognize emotions and influence YouTube to play music that can stabilize one's mood. In [100], contextual affective hashtag information in tweets was used to rank music recommendations in an unsupervised approach. Therefore, ethical considerations are an important factor to consider during SER deployment in the real world.

VI. CONCLUSION

In this paper, we reviewed the different aspects involved in deep learning-based bimodal SER research. We presented recent literature on the datasets, features, deep learning techniques, and some of the recent results published in the literature. We opine that all the aspects in the bimodal SER framework are important for robust performance. We found out that there are few publicly available datasets for bimodal SER research which hampers the full deployment of proposed models. We also noted that attention mechanisms when used with other deep learning techniques play a pivotal role in the performance of bimodal SER systems by computing the context score of the features. We also carried out experiments on the significance of CNNs, RNNs, and attention mechanisms which are the common techniques used in literature. Subsequently, we proposed a deep learning-based multi-learning model for emotion recognition (DBMER) that operates with multi-learning capabilities of CNNs, RNNs, and multi-head attention mechanisms and evaluated its performance on acoustic and dyadic bimodal speech. It was found that a careful combination of these techniques improves bimodal SER performance. On the other hand, to avoid the demerits of attention mechanisms that include long training periods and the need for sufficient data that is not available in addition to complexity, transfer learning can be used. However, cross-corpus and multilingual research and the acquisition of all-round datasets remain open problems in bimodal SER research.

REFERENCES

[1] M. S. Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digit. Signal Process.*, vol. 110, Mar. 2021, Art. no. 102951.

[2] G. N. Yannakakis, "Enhancing health care via affective computing," *Malta J. Health Sci.*, vol. 5, no. 1, pp. 38–42, 2018.

[3] L. Devillers, "Human–robot interactions and affective computing: The ethical implications," in *Robotics, AI, and Humanity*. Cham, Switzerland: Springer, 2021, pp. 205–211.

[4] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Comput. Educ.*, vol. 142, Dec. 2019, Art. no. 103649.

[5] F. Ren and C. Quan, "Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: An application of affective computing," *Inf. Technol. Manage.*, vol. 13, no. 4, pp. 321–332, Dec. 2012.

[6] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas, *The Oxford Handbook of Affective Computing* (Oxford Library of Psychology). Oxford, U.K.: Oxford Univ. Press, 2015.

[7] R. W. Picard, *Affective Computing*. Google Scholar Digital Library Digital Library, 1997.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[9] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[10] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4275–4279.

[11] B. T. Atmaja, A. Sasou, and M. Akagi, "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion," *Speech Commun.*, vol. 140, pp. 11–28, May 2022.

[12] 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.

[13] J. Zhang, L. Xing, Z. Tan, H. Wang, and K. Wang, "Multi-head attention fusion networks for multi-modal speech emotion recognition," *Comput. Ind. Eng.*, vol. 168, Jun. 2022, Art. no. 108078.

[14] 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.

[15] 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.

[16] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," in *Proc. Interspeech*, Oct. 2020, pp. 374–378.

[17] 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.

[18] M. P. Kesarkar and P. Rao, "Feature extraction for speech recognition," Electron. Syst. Group, EE Dept., IIT Bombay, India, Tech. Rep., 2003.

[19] B. M. Nema and A. A. Abdul-Kareem, "Preprocessing signal for speech emotion recognition," *Al-Mustansiriyah J. Sci.*, vol. 28, no. 3, pp. 157–165, Jul. 2018.

[20] 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, Mar. 2019.

[21] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," 2022, *arXiv:2203.13504*.

[22] 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, Dec. 2008.

[23] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," 2018, *arXiv:1810.02508*.

[24] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics Long Papers*, 2018, pp. 2236–2246.

[25] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, May 2021.

[26] L. Yi and M.-W. Mak, "Improving speech emotion recognition with adversarial data augmentation network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 172–184, Jan. 2022.

[27] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 511–516.

[28] E. Liebenthal, D. A. Silbersweig, and E. Stern, "The language, tone and prosody of emotions: Neural substrates and dynamics of spoken-word emotion perception," *Frontiers Neurosci.*, vol. 10, p. 506, Nov. 2016.

[29] H. Tao, R. Liang, C. Zha, X. Zhang, and L. Zhao, "Spectral features based on local Hu moments of Gabor spectrograms for speech emotion recognition," *IEICE Trans. Inf. Syst.*, vol. 99, no. 8, pp. 2186–2189, 2016.

[30] M. Ezz-Eldin, A. A. M. Khalaf, H. F. A. Hamed, and A. I. Hussein, "Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition," *IEEE Access*, vol. 9, pp. 19999–20011, 2021.

[31] Y. Huang, A. Wu, G. Zhang, and Y. Li, "Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition," *IET Signal Process.*, vol. 9, no. 4, pp. 341–348, Jun. 2015.

[32] H. K. Palo and M. N. Mohanty, "Wavelet based feature combination for recognition of emotions," *Ain Shams Eng. J.*, vol. 9, no. 4, pp. 1799–1806, Dec. 2018.

[33] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, nos. 1–4, pp. 43–52, Dec. 2010.

[34] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *Int. J. Gen. Syst.*, vol. 46, no. 1, pp. 27–36, Jan. 2017.

[35] J. Bhatta, D. Shrestha, S. Nepal, S. Pandey, and S. Koirala, "Efficient estimation of nepali word representations in vector space," *J. Innov. Eng. Educ.*, vol. 3, no. 1, pp. 71–77, Mar. 2020.

[36] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[37] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[39] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning-based speech emotion recognition using multi-level fusion of concurrent features," *IEEE Access*, vol. 10, pp. 125538–125551, 2022.

[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[41] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proc. Interspeech*, Sep. 2018, pp. 3688–3692.

[42] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions," 2019, *arXiv:1906.05681*.

[43] G. Sahu, "Multimodal speech emotion recognition and ambiguity resolution," 2019, *arXiv:1904.06022*.

[44] 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.

[45] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence LSTM architecture," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6474–6478.

[46] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101894.

[47] P. Kordjamshidi, M. Van Otterlo, and M.-F. Moens, "Spatial role labeling: Towards extraction of spatial relations from natural language," *ACM Trans. Speech Lang. Process.*, vol. 8, no. 3, pp. 1–36, Dec. 2011.

[48] T. Li, H. Xu, Z. Liu, Z. Dong, Q. Liu, J. Li, S. Fan, and X. Sun, "A spatiotemporal multi-feature extraction framework for opinion mining," *Neurocomputing*, vol. 490, pp. 337–346, Jun. 2022.

[49] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 771–775.

[50] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, ''Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4380–4384.

[51] K. Han, D. Yu, and I. Tashev, ''Speech emotion recognition using deep neural network and extreme learning machine,'' in *Proc. Interspeech*, Sep. 2014, pp. 223–227.

[52] J. Lee and I. Tashev, ''High-level feature representation using recurrent neural network for speech emotion recognition,'' in *Proc. Interspeech*, Sep. 2015, pp. 1–4.

[53] D. Bahdanau, K. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' 2014, *arXiv:1409.0473*.

[54] M.-T. Luong, H. Pham, and C. D. Manning, ''Effective approaches to attention-based neural machine translation,'' 2015, *arXiv:1508.04025*.

[55] Y. Yan and X. Shen, ''Research on speech emotion recognition based on AA-CBGRU network,'' *Electronics*, vol. 11, no. 9, p. 1409, Apr. 2022.

[56] 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.

[57] 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.

[58] D. Hu, L. Wei, and X. Huai, ''Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations,'' 2021, *arXiv:2106.01978*.

[59] 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.

[60] 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.

[61] S. Kakuba and D. S. Han, ''Speech emotion recognition using context-aware dilated convolution network,'' in *Proc. 27th Asia Pacific Conf. Commun. (APCC)*, Oct. 2022, pp. 601–604.

[62] S. Kakuba, A. Poulose, and D. S. Han, ''Attention-based multi-learning approach for speech emotion recognition with dilated convolution,'' *IEEE Access*, vol. 10, pp. 122302–122313, 2022.

[63] 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.

[64] S. Kakuba and D. S. Han, ''Bimodal speech emotion recognition using fused intra and cross modality features,'' in *Proc. 14th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2023, pp. 109–113.

[65] Y. Tang, Y. Hu, L. He, and H. Huang, ''A bimodal network based on Audio–Text-Interactional-Attention with ArcFace loss for speech emotion recognition,'' *Speech Commun.*, vol. 143, pp. 21–32, Sep. 2022.

[66] 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, pp. 1–18, 2014.

[67] P. Koromilas and T. Giannakopoulos, ''Deep multimodal emotion recognition on human speech: A review,'' *Appl. Sci.*, vol. 11, no. 17, p. 7962, Aug. 2021.

[68] T. Baltrusaitis, 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.

[69] 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. Human Agent Interact.*, Oct. 2016, pp. 193–200.

[70] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, ''Learning alignment for multimodal emotion recognition from speech,'' *arXiv:1909.05645*, 2019.

[71] S. Yoon, S. Byun, S. Dey, and K. Jung, ''Speech emotion recognition using multi-hop attention mechanism,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2822–2826.

[72] C. R. Chen, Q. Fan, and R. Panda, ''CrossViT: Cross-attention multi-scale vision transformer for image classification,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.

[73] C. Zheng, C. Wang, and N. Jia, ''An ensemble model for multi-level speech emotion recognition,'' *Appl. Sci.*, vol. 10, no. 1, p. 205, Dec. 2019.

[74] S. Chen and Q. Jin, ''Multi-modal conditional attention fusion for dimensional emotion prediction,'' in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 571–575.

[75] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, ''Multi-level multiple attentions for contextual multi-modal sentiment analysis,'' in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1033–1038.

[76] W. Wu, C. Zhang, and P. C. Woodland, ''Emotion recognition by fusing time synchronous and time asynchronous representations,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6269–6273.

[77] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, ''Jointly fine-tuning 'BERT-like' self supervised models to improve multimodal speech emotion recognition,'' 2020, *arXiv:2008.06682*.

[78] N. Braunschweiler, R. Doddipatla, S. Keizer, and S. Stoyanchev, ''Factors in emotion recognition with deep learning models using speech and text on multiple corpora,'' *IEEE Signal Process. Lett.*, vol. 29, pp. 722–726, 2022.

[79] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, ''A database of German emotional speech,'' in *Proc. Interspeech*, vol. 5, Sep. 2005, pp. 1517–1520.

[80] S. R. Livingstone and F. A. Russo, ''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*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[81] P. Jackson and S. Haq, ''Surrey audio-visual expressed emotion (SAVEE) database,'' Univ. Surrey, Guildford, U.K., 2014.

[82] M. K. Pichora-Fuller and K. Dupuis, ''Toronto emotional speech set (TESS),'' Scholars Portal Dataverse, Univ. Toronto, Toronto, ON, Canada, 2020, vol. 1, p. 2020.

[83] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, ''CREMA-D: Crowd-sourced emotional multimodal actors dataset,'' *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

[84] D. Landry, Q. He, H. Yan, and Y. Li, ''ASVP-ESD: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances,'' *Global Sci. J.*, vol. 8, no. 5, pp. 1793–1798, 2020.

[85] H. Fu, Z. Zhuang, Y. Wang, C. Huang, and W. Duan, ''Cross-corpus speech emotion recognition based on multi-task learning and subdomain adaptation,'' *Entropy*, vol. 25, no. 1, p. 124, Jan. 2023.

[86] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, ''Bimodal speech emotion recognition using pre-trained language models,'' 2019, *arXiv:1912.02610*.

[87] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, ''Conversational emotion recognition using self-attention mechanisms and graph neural networks,'' in *Proc. Interspeech*, Oct. 2020, pp. 2347–2351.

[88] A. Triantafyllopoulos, U. Reichel, S. Liu, S. Huber, F. Eyben, and B. W. Schuller, ''Multistage linguistic conditioning of convolutional layers for speech emotion recognition,'' 2021, *arXiv:2110.06650*.

[89] 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.

[90] H. Filali, J. Riffi, C. Boulealam, M. A. Mahraz, and H. Tairi, ''Multimodal emotional classification based on meaningful learning,'' *Big Data Cogn. Comput.*, vol. 6, no. 3, p. 95, Sep. 2022.

[91] P. Shixin, C. Kai, T. Tian, and C. Jingying, ''An autoencoder-based feature level fusion for speech emotion recognition,'' *Digit. Commun. Netw.*, Oct. 2022.

[92] Y. Khurana, S. Gupta, R. Sathyaraj, and S. P. Raja, ''RobinNet: A multimodal speech emotion recognition system with speaker recognition for social interactions,'' *IEEE Trans. Computat. Social Syst.*, early access, Dec. 26, 2022, doi: 10.1109/TCSS.2022.3228649.

[93] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, ''Attention driven fusion for multi-modal emotion recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3227–3231.

[94] M. R. Makiuchi, K. Uto, and K. Shinoda, ''Multimodal emotion recognition with high-level speech and text features,'' in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 350–357.

[95] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, ''Speech emotion recognition based on self-attention weight correction for acoustic and text features,'' *IEEE Access*, vol. 10, pp. 115732–115743, 2022.

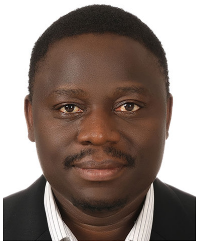
[96] L. Yunxiang and Z. Kexin, ''Cross-corpus bimodal speech emotion recognition,'' *Authorea Preprints*, 2022.

[97] J. S. Joel, B. E. Thompson, S. R. Thomas, T. R. Kumar, S. Prince, and D. Bini, "Emotion based music recommendation system using deep learning model," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Apr. 2023, pp. 227–232.

[98] M. Pant, S. Trivedi, S. Aggarwal, R. Rani, A. Dev, and P. Bansal, "Driver's companion-drowsiness detection and emotion based music recommendation system," in *Proc. Int. Conf. Comput., Commun., Intell. Syst. (ICCCIS)*, Nov. 2022, pp. 1–6.

[99] V. Mounika and Y. Charitha, "Mood-enhancing music recommendation system based on audio signals and emotions," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Apr. 2023, pp. 1766–1772.

[100] E. Zangerle, C.-M. Chen, M.-F. Tsai, and Y.-H. Yang, "Leveraging affective hashtags for ranking music recommendations," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 78–91, Jan. 2021.



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 an Assistant Lecturer with the Department of Electrical Engineering, 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, computer vision, affective computing, human behavior prediction, 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, Bengaluru, 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 localization and mapping system for autonomous vehicles. He was a Research Fellow at the Department of Electrical and Computer Engineering, University of Michigan–Dearborn, USA, from November 2022 to December 2022. He has been an Assistant Professor at the School of Data Science, Indian Institute of Science Education and Research Thiruvananthapuram (IISER TVM), Vithura, Thiruvananthapuram, Kerala, India, since January 2023. 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.

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 with Samsung Electronics Company Ltd., where he developed the transmission system for the ATSC HDTV receiver. Since 1996, he has been a Professor at the School of Electronic and Electrical Engineering, KNU. He was a Courtesy Associate Professor at the Department of Electrical and Computer Engineering, University of Florida, in 2004. He was the Director at the Center of Digital TV and Broadcasting, Institute for Information Technology Advancement (IITA), from 2006 to 2008. He is currently the Director of the Center for ICT and Automotive Convergence, KNU, where he is also the Dean of the College of IT Engineering. His main research interests include intelligent signal processing and autonomous vehicles.