Bimodal Speech Emotion Recognition using Fused Intra and Cross Modality Features

Samuel Kakuba
Graduate School of Electronics and Electrical Engineering
Kyungpook National University
Daegu, Republic of Korea
2021327392@knu.ac.kr

Dong Seog Han
School of Electronics Engineering
Kyungpook National University
Daegu, Republic of Korea
dshan@knu.ac.kr

***Abstract*—The interactive speech between two or more interlocutors involves the text and acoustic modalities. These modalities consist of intra and cross-modality relationships at different time intervals which if modeled well, can avail emotionally rich cues for robust and accurate prediction of emotion states. This necessitates models that take into consideration long short-term dependency between the current, previous, and future time steps using multimodal approaches. Moreover, it is important to contextualize the interactive speech in order to accurately infer the emotional state. A combination of recurrent and/or convolutional neural networks with attention mechanisms is often used by researchers. In this paper, we propose a deep learning-based bimodal speech emotion recognition (DLBER) model that uses multi-level fusion to learn intra and cross-modality feature representations. The proposed DLBER model uses the transformer encoder to model the intra-modality features that are combined at the first level fusion in the local feature learning block (LFLB). We also use self-attentive bidirectional LSTM layers to further extract intra-modality features before the second level fusion for further progressive learning of the cross-modality features. The resultant feature representation is fed into another self-attentive bidirectional LSTM layer in the global feature learning block (GFLB). The interactive emotional dyadic motion capture (IEMOCAP) dataset was used to evaluate the performance of the proposed DLBER model. The proposed DLBER model achieves 72.93% and 74.05% of F1 score and accuracy respectively.**

***Index Terms*—emotion recognition, intra modality features, inter modality features, Fusion**

I. Introduction

The advent of deep learning techniques has improved how machines handle the tasks that humans are good at. Emotion recognition is one of the tasks that have been improved over the years in the affective domain. Since emotion recognition is a complex task even to human beings, more robust and accurate models are needed. The existing models perform well in laboratory experiments but tend to degrade in the natural environment. The emotion recognition application areas in real life include; assistive living of the elderly [1], [2] and children with health conditions like autism [3], [4], fraud detection, home security among others. Depending on the source of data for the deep learning models, emotional recognition can be physiological, lexical, facial, or acoustic. It should however be noted that speech involves both acoustic and lexical data. A number of studies have been carried out that have led to model propositions in terms of the sources of data singly or in combination. Human emotions can be categorized as discrete or continuous. Tsiourti *et al.* [5] described emotions in a two-dimensional plane of valence and arousal. Verma *et al.* [6] added dominance in the two-dimensional emotion space and analyzed the emotions in a three-dimensional continuous space. However, they all agree with the discrete emotion categories of Ekman *et al.* [7]. Ekman *et al.* described emotions as happiness, sadness, surprise, anger, disgust, and fear. In this paper, we propose a model for speech emotion recognition (SER). We opine that speech involves the acoustic and text modalities and they consist of intra and cross-modality relationships which depict emotionally rich cues. The proposed model leverages these cues for robust performance. Interactive speech between two or more interlocutors involves the analysis of features extracted from a varying time speech signal with the intent of classifying the emotional state of the utterances therein. Interactive SER has recently attracted researchers because of the need for robust and reliable systems that can aid interaction between intelligent devices and humans for different activities. The study of emotions in speech does not only consider utterances at a single instance in one modality but all the instances in a sequence in a multi-model multi-modal approach. As stated in [8], human emotions are perceived from the history, current, and future utterances. Moreover, the emotions in each utterance are triggered by context cues [9] making it so important to consider the contextual representations in the speech signal and text. Recurrent neural networks (RNN) like long short-term memory (LSTM) [10] are often used in combination with attention mechanisms to keep track of long-term dependencies between the features [11]. Memory networks, graph networks, and convolutional neural networks [12] are the other deep learning technologies often used in literature.

Research has been carried out using the acoustic modality to improve the performance of SER models. Recently, Yan *et al.* [13] proposed a model that uses convolutional neural networks with bidirectional gated recurrent networks (BiGRU) and the attention mechanism to classify emotions from ex-

tracted spectrograms and their delta 1 and 2 derivatives using the IEMOCAP dataset. RNNs on the other hand have been replaced with residual dilated convolution blocks [14], [15]. In [16], a late fusion-based model was also proposed for speech emotion recognition. Recently, Bekmanova *et al.* [17] suggested recognition of emotions from word transcriptions for students that participated in distance learning examinations.

However, it is not enough to infer speech emotions from only one modality since the cross-modality interaction between acoustic and lexical features is as important as the intra-modality feature interactions for emotion classification and enriches the emotional cues. The works in [18], [19], [20], [21] and [22], utilise transformer-based multi-head attention [23] to model multimodal SER. In this paper, we propose a deep learning-based bimodal speech emotion recognition (DLBER) model that uses multi-level fusion to learn intra and cross-modality feature representations in speech. The proposed DLBER model uses the transformer encoder to model the intra-modality features that are combined at the first level fusion in the local feature learning block (LFLB). The self-attentive bidirectional LSTM layers are used to further extract intra-modality features before the second level fusion to learn the cross-modality features. The resultant feature representation is fed into another self-attentive BiLSTM block in the global feature learning block (GFLB). This work's contributions are as follows:

- We propose a deep learning based bimodal speech emotion recognition (DLBER) model that progressively learns intra and cross-modality feature interaction representation in speech with multi-level fusion.
- We carry out a performance evaluation to show that progressive multi-level fusion of intra and cross-modality features learned by the proposed DLBER model improves the performance of SER systems.

The rest of the paper is organized as follows: the proposed model is presented in Section II. Section III presents the experimental evaluation. Section IV gives the results and discussion. The paper is concluded in Section V.

II. The Proposed Model

In this section, we describe the proposed deep learning-based bimodal speech emotion recognition (DLBER) model that learns intra and cross-modality feature representations fused at two separate levels. The intra and cross-modality feature interaction characteristics of audio and transcription files are learned using transformer encoders and self-attentive BiLSTM layers at the two levels. As shown in Fig. 1, the local feature learning block (LFLB) of the proposed DLBER model uses the transformer encoder block (TEB) to learn intra-modality features in the acoustic and text modalities separately before being fused at the first level. The resultant feature representation is further fed into a self-attentive BiLSTM block. The features from the individual modalities are also concurrently fed into self-attentive BiLSTM blocks. The resultant features from these channels are combined at the second level fusion. The resultant feature representation is fed as input to another self-attentive BiLSTM block in the GFLB. It should be noted that the fully connected layers (FC) are used in this model to ensure similar feature dimensions so as to be able to fuse them. The deep learning techniques used to configure the proposed DLBER model allows it to progressively learn the long-term dependencies and pay attention to the most relevant intra and cross-modality features. This enables the model to preserve the context in which they were uttered. The transformer encoder and attention mechanisms ensure that the model learns contextualized features of the lexical and acoustic modalities and the relationship characteristics that may exist between them. The emotional states of excited, sad, neutral, and anger are considered in this paper. They are chosen because they have a representative number of samples in the dataset compared to the rest of the emotions. Drop-out regularization, the addition of Gaussian noise, and layer normalization are configured to overcome overfitting. We also configured class weights to cater for the class imbalances in the dataset.

III. Experimental Evaluation

We used Librosa 0.9.2 to read audio files and extract acoustic features from them. The Keras 2.8.0 and TensorFlow 2.6 frameworks were used. We used the Nvidia GeForce RTX 2080 super graphics processing unit (GPU). The batch size and initial learning rate were 32 and 0.0001 respectively with the Adam optimizer and cross-entropy loss objective function. The evaluation metrics in the Sci-kit-learn toolbox were used. The 4936 samples were apportioned in a ratio of 80% for training, 10% for validation, and 10% for testing. The experiments carried out, datasets, and features used are presented in this section.

A. Experiments

We carried out experiments in form of ablation studies to investigate the significance of the components of the proposed DLBER model. The experiments included the single modality SER models for acoustic and lexical features separately. Experiments about single-level fusion cross-modality transformer encoder network with self-attentive BiLSTM (TESAB) were also carried out to investigate the performance of the first-level modality fusion. In this experiment, the transformer encoder was used to extract intra-modality features from the individual modalities and then they were fused before being fed into the self-attentive BiLSTM block and softmax layer for emotion classification. In another experiment, the output of the sub-model described in the previous experiment which now represents the cross-modality feature representations is fused with intra-modality features learned by other self-attentive BiLSTM blocks before the GFLB that consists of other self-attentive BiLSTM model. The aim of this experiment was to fuse intra-modality features obtained from individual modalities using the self-attentive BiLSTM block and those progressively

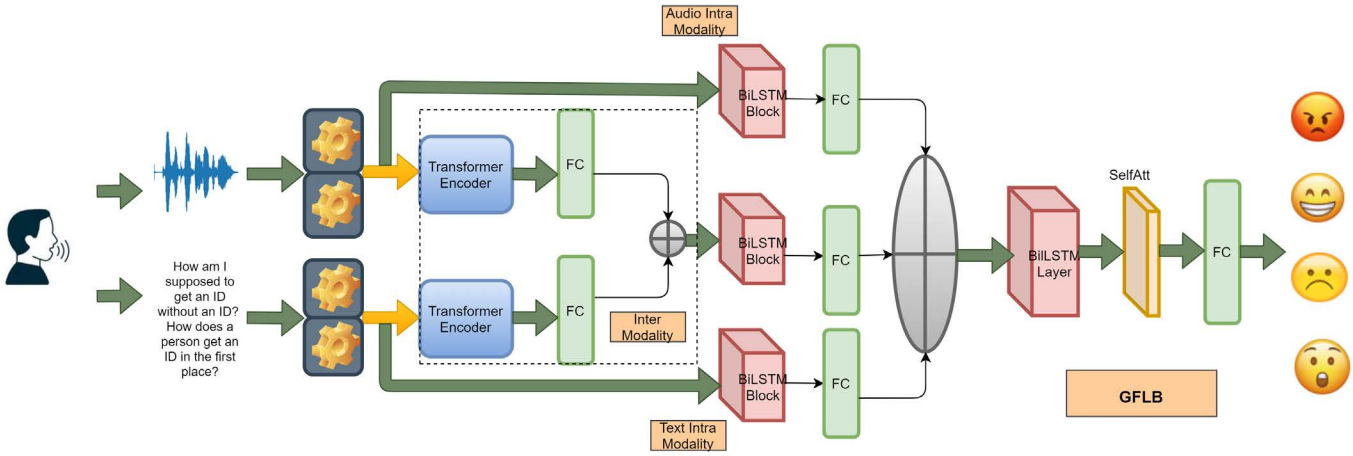


Fig. 1: The proposed deep learning-based bimodal speech emotion recognition (DLBER) mode.

learned from the individual transformer encoders and another subsequent BiLSTM block before second-level fusion. In this experiment, all the feature representations from the described setups discussed thus far were fused in an end-to-end proposed DLBER model.

Though we carried out more experiments, there was no significant change in results yet the complexity of the model was increasing. Nonetheless, we report the results of the experiments with confidence intervals to cater for the results obtained in the other experiments not reported in this paper.

B. Datasets and Features

We evaluated the experiments in this paper using the interactive emotional dyadic motion capture (IEMOCAP) [24] which is a multi-modal and multi-speaker database. To extract lexical features, we used the pre-trained bidirectional encoder representation for transformers (BERT) [25] to extract word embedding vectors from transcriptions. The mel frequency cepstral coefficients (MFCCs) that depict the vocal tract frequency response in sound were used as acoustic features.

IV. Results and Discussion

In this section, we present the results obtained from our experiments and discuss the significance of the proposed DLBER and its constituent models.

A. Results

As shown in Table I, we report results in terms of accuracy (A), precision (P), recall (R), and F1 score (F1) as our performance metrics. We report on experiments for the individual modalities which are modeled using the BiLSTM block only, single-level fusion of intra-modality feature vectors to make cross-modality features using the TESAB model, Multi-level fusion of Intra and cross-modality feature vectors using the proposed DLBER model. We also report the performance of the proposed DLBER model on the individual classes of the testing dataset in terms of the F1 score in Table II. The confusion matrices for the experiments carried out are shown in Fig. 2. The performance of the proposed DLBER model compared to the other experimental approaches used in this paper shows that each of the constituent components is significant in the performance of the proposed model however a combination of all in a multi-level approach is more beneficial to SER.

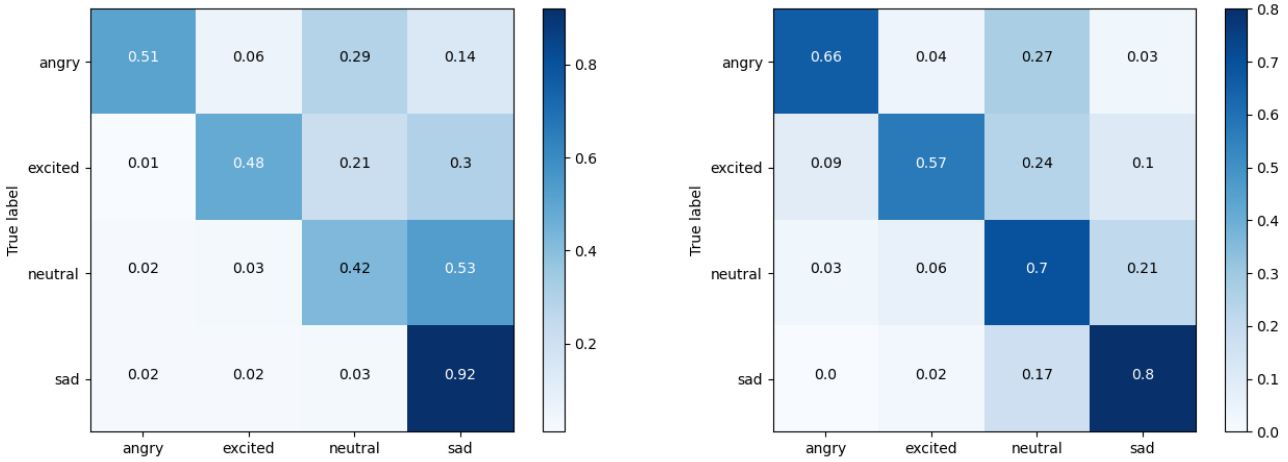
B. Discussion

In this section, we discuss the significance of the results obtained from our ablation study through the experiments we carried out.

1) Ablation Study: According to the results shown in Tables I and II, the proposed DLBER exhibits a commendable performance. From the accuracy and F1 score metrics, it is observed that progressive multi-level fusion of the cross and intra-modality feature relationships improves performance. We also further analyzed the proposed DLBER model's performance on the individual classes using the F1 score in Table II. From these results, it is clear that the individual class performance benefits from the different parts of the proposed model since a progressive F1 score improvement is observed. The improvement is because the proposed DLBER model leverages the feature level and decision level fusion at the intermediate level to progressively learn the intra and cross-modality feature representations at the first and second fusion levels. The performance improvement is also because of the dynamic and parallel operation of the transformer encoder that uses the multi-head attention mechanism used to learn the intra modality features. These results show the superiority of modeling intra and cross-modality features for SER systems which are learned in a multi-level approach. This performance also shows that a careful combination of deep learning technologies allows the model to benefit from their capabilities in the training process for better results. The global attention approach used by self-attention mechanisms in computing the context vector complemented with the dynamic and parallel

TABLE I: Performance of the Individual Modality Models, the TESAB Model and the Proposed DLBER Model.

Model	Fusion level	A(%)	P(%)	R(%)	F1(%)
BiLSTM (Acoustic)	No fusion	58.70±0.21	60.08±0.10	54.63±0.50	53.50±0.10
BiLSTM (Lexical)	No fusion	61.23±1.00	42.13±5.00	85.83±0.20	56.68±2.00
TESAB	Single	63.42±2.00	65.11±1.90	60.00±3.00	62.39±2.50
DLBER	Multi	74.05±1.10	75.17±2.00	70.89±1.90	72.93±2.10



(a) (b)

Fig. 2: The confusion matrix results. (a) TESAB. (b) DLBER.

TABLE II: F1 score values for Individual Classes that were obtained by the TESAB and the Proposed DLBER Model.

	F1 Score(%)	
Emotion	TESAB	DLBER
Angry	65.0±1.00	73.0±2.00
Excited	59.0±2.00	66.0 ±3.00
Neutral	48.0±4.00	66.0±4.00
Sad	58.0±3.00	71.0±2.00

operation done by the multi-head attention mechanism does not only improve the learning of the grammatical features but is also applied in the GFLB to attentively learn the cross-modality feature representations at the low level. Multi-head attention consists of more than one self-attention head that can be used to compute different relationship representations of the acoustic and lexical modality features for the benefit of the proposed DLBER model. Therefore, progressive learning of intra and cross-modality feature relationships in a multi-level approach allows the model to generalize better. The self-attentive BiLSTM considers the long-term dependencies as well as solving the vanishing gradient problem which contributes to better generalization of the proposed DLBER model.

2) Prediction Error Analysis: As observed in confusion matrices shown in Fig. 2 (a and 2 (b, the individual class accuracy predicted by the models progressively improves in the proposed DLBER model compared to the TESAB model. It is observed from both of the confusion matrices in Fig. 2 that there is an attempt to generalize prediction of the emotions by the models as more levels of fusion are added to learn intra and cross-modality representations. This is also evident in the observed changes in the confusion ratio of the sad emotion class which had 92% of the support samples predicted correctly as seen in Fig. 2 (a) at the expense of all the other emotions. However, with a progressive increase in levels of fusion, the confusion ratio of the sad emotion is comparable to the other emotion classes. These confusion matrices also show that there is better generalization during inference of the proposed DLBER model as compared to the control experiments. It should however be noted that the emotion classes are mostly confused with the neutral class. This is because the neutral emotion class is situated at the center of the two-dimensional arousal-valence space of emotions which complicates the discriminatory capability of the model [26]. The other reason could be because it consists of the largest number of samples in this dataset.

V. Conclusion

We proposed the DLBER model which leverages the benefits of fusing intra and cross-modality feature interactions of acoustic and lexical modalities for SER. The feature relationship representations were learned using self-attentive BiLSTM and transformer-encoded blocks at two intermediate fusion levels. We compared the performance of the proposed DLBER model with the TESAB and single modality models in ablation study experiments. Compared to the TESAB model, the proposed DLBER model achieves performance improvement in terms of accuracy and F1 score. These results show that SER systems benefit from multi-level fusion learning of intra and cross-modality features for improved and robust performance. In terms of future work, since human emotional states encompass visual and physiological cues in addition to speech, it is worth exploring other modalities for emotion recognition. We also plan to evaluate the performance of the proposed DLBER model on other datasets.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2021R1A6A1A03043144).

References

[1] M. Visser, "Emotion recognition and aging. comparing a labeling task with a categorization task using facial representations," *Frontiers in psychology*, vol. 11, p. 139, 2020.

[2] D. S. Cortes, C. Tornberg, T. Bänziger, H. A. Elfenbein, H. Fischer, and P. Laukka, "Effects of aging on emotion recognition from dynamic multimodal expressions and vocalizations," *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[3] S. Fridenson-Hayo, S. Berggren, A. Lassalle, S. Tal, D. Pigat, S. Bölte, S. Baron-Cohen, and O. Golan, "Basic and complex emotion recognition in children with autism: cross-cultural findings," *Molecular autism*, vol. 7, no. 1, pp. 1–11, 2016.

[4] E. Nagy, L. Prentice, and T. Wakeling, "Atypical facial emotion recognition in children with autism spectrum disorders: Exploratory analysis on the role of task demands," *Perception*, vol. 50, no. 9, pp. 819–833, 2021.

[5] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots," *International Journal of Social Robotics*, vol. 11, no. 4, pp. 555–573, 2019.

[6] G. K. Verma and U. S. Tiwary, "Affect representation and recognition in 3d continuous valence–arousal–dominance space," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2159–2183, 2017.

[7] P. Ekman, "i friesen, wv (1971). constants across cultures in the face and emotion," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1972.

[8] J. J. Gross and L. Feldman Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion review*, vol. 3, no. 1, pp. 8–16, 2011.

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

[10] S. Hochreiter, "Ja1 4 rgen schmidhuber (1997).'long short-term memory'," *Neural Computation*, vol. 9, no. 8, 1997.

[11] S. Kakuba and H. Dong, Seog, "Residual bidirectional lstm with multi-head attention for speech emotion recognition," in *Korea Communications Association Summer General Academic Conference*. KICS, 2022, pp. 1419–1421.

[12] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 1058–1064.

[13] Y. Yan and X. Shen, "Research on speech emotion recognition based on aa-cbgru network," *Electronics*, vol. 11, no. 9, p. 1409, 2022.

[14] S. K. Pandey, H. S. Shekhawat, and S. Prasanna, "Emotion recognition from raw speech using wavenet," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 1292–1297.

[15] 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. 122 302–122 313, 2022.

[16] B. Maji and M. Swain, "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, 2022.

[17] G. Bekmanova, B. Yergesh, A. Sharipbay, and A. Mukanova, "Emotional speech recognition method based on word transcription," *Sensors*, vol. 22, no. 5, p. 1937, 2022.

[18] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 4171–4183, 2021.

[19] 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. 61 672–61 686, 2020.

[20] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[21] S. Kakuba and D. S. Han, "Speech emotion recognition using context-aware dilated convolution network," in *2022 27th Asia Pacific Conference on Communications (APCC)*. IEEE, 2022, pp. 601–604.

[22] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning-based speech emotion recognition using multi-level fusion of concurrent features," *IEEE Access*, 2022.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] 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," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.