Speech Emotion Recognition Using Multi-Layer Heterogeneity Acoustic Feature Fusion

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

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

***Abstract*—Though there have been endeavors to realize commendable performance for speech emotion recognition in recent years, more robust and accurate models are still needed amidst data scarcity and performance disparity between lab experiments and the real world. In this paper, we propose a deep learning-based multi-layer heterogeneity acoustic feature fusion (MHAFF) model for Speech Emotion Recognition. We also carried out a performance evaluation of the model to show the significance of fusion with weight score configuration (WSC) of the shallow, intermediate and higher-level learned features in SER systems. We evaluated the model on the emotional German speech dataset (EMODB) and Toronto Emotional Speech Set (TESS) datasets with the emotional states of happy, sad, neutral, and angry. The proposed MHAFF model achieves 81.82% and 99.65% of average accuracy on EMODB and TESS respectively.**

***Index Terms*—emotion recognition, heterogeneity features, Fusion**

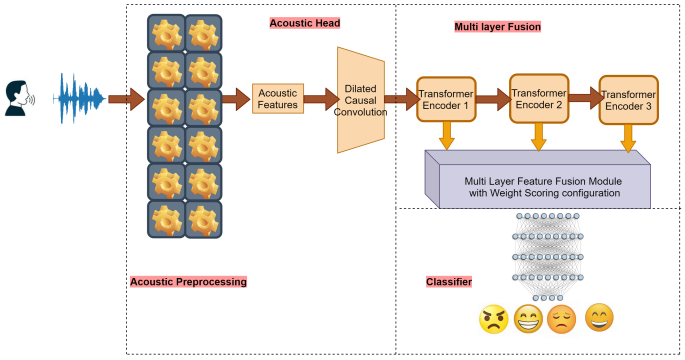

Fig. 1: The proposed multi-layer heterogeneity acoustic feature fusion (MHAFF) model for Speech Emotion Recognition.

I. Introduction

Emotion recognition is a complex task that involves intelligent devices empowered by artificial intelligence (AI) techniques to recognize and process human emotional states and take appropriate actions. Human emotions can be categorized as discrete or continuous. A two-dimensional plane of valence and arousal was proposed by Tsiourti *et al.* in [1]. Verma *et al.* [2] added dominance in the two-dimensional emotion space and analyzed the emotions in a three-dimensional continuous space. There are also discrete emotion categories that were described by Ekman *et al.* [3]. Ekman *et al.* described emotions as happiness, sadness, surprise, anger, disgust, and fear.

Though there have been endeavors like in [4], [5], [6] to realize commendable performance in speech emotion recognition (SER), more robust and accurate models are still needed amidst data scarcity and performance disparity issues between lab experiments and the real world. The existing models learn heterogeneity features sequentially however, upon learning the higher-level features in a deep complex network, the shallow-level features which might consist of rich cues for the prior emotion context are discarded. It is therefore important to propose models that learn heterogeneity features but employ a mechanism to allow the shallow as well as the intermediate and higher level features an opportunity to participate in the emotion classification. In this paper, we propose a fusion mechanism that uses weighted multi-layer fusion to highlight the importance of each level-learned feature.

Learning features from sequential speech utterances involves understanding patterns for long-term dependencies, context, and spatial representations. The existing approaches utilize long short-term memory (LSTM) [7] and its variants, convolution neural networks (CNN) like in [8] and [9], and attention mechanisms especially transformer-based self and multi-head attention mechanisms first proposed in [10]. However, the robust approaches like [11], [6], [12], [13], [14], and [5] use a combination of all these techniques.

Nonetheless, it may not be enough to learn spatial and contextualized long-term dependencies in emotional speech using deep networks. It is important to devise means of utilizing heterogeneity acoustic features learned at shallow, intermediate, and higher levels of the deep learning models. In line with this fact, we propose a deep learning-based model called multi-layer heterogeneity acoustic feature fusion (MHAFF) that uses a weight scoring configuration (WSC) module to gain knowledge of shallow, intermediate, and higher-level heterogeneity features for SER. The proposed MHAFF model uses three transformer encoders to model the acoustic heterogeneity features which are fused using the WSC module that consists

TABLE I: Performance of the proposed MHAFF model on EMODB and TESS speech datasets.

Dataset	Loss	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
EMODB	0.690±0.130	81.82±0.01	82.82±0.12	80.30±0.30	87.30±0.02
TESS	0.18±0.321	99.65±2.00	99.65±1.50	99.65±0.40	99.65±5.00

(a) (b)

Fig. 2: The confusion matrix results. (a) EMODB. (b) TESS.

of a score generation mechanism for each layer before fusion and subsequent classification.

II. Method

A. The proposed model

The proposed MHAFF SER model learns shallow, intermediate, and higher-level heterogeneity intra-modality feature representations which are fused before emotion classification. As shown in Fig. 1, the proposed MHAFF SER model consists of the acoustic head, multi-layer fusion, and classifier modules. The acoustic head consists of acoustic preprocessing to obtain Mel frequency cepstral coefficients (MFCCs) that depict the vocal tract frequency response in speech which are fed into two dilated causal convolution layers of 64 channels each and a dilation rate of 2 and 4. The shallow, intermediate, and higher-level acoustic representations are learned using transformer encoders which consist of multi head attention with an embedding dimension of 64, four heads, and layer normalization. The transformer encoder outputs y_1, y_2, and y_3 are then passed through the WSC to obtain the magnitude of the influence each has on the final feature representation. The subsequent feature representation is given by equation 1.

$$WSC_{out} = (WSC_{y_1}) + (WSC_{y_2}) + (WSC_{y_3}) \quad (1)$$

This subsequent feature representation is fed into one fully connected layer of 128 units and a softmax layer of four nodes for classification.

B. Datasets

For model performance evaluation we used the emotional German speech dataset (EMODB) [15] and Toronto emotional speech set (TESS) datasets [16]. The emotional states considered in this paper are happy, sad, neutral, and angry. They were chosen because they are common to all the datasets used.

C. Experiments

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 16 and 0.00025 respectively with the Adam optimizer and cross-entropy loss objective function. The evaluation metrics in the Sci-kit-learn toolbox were used.

The addition of Gaussian noise, dropout of 0.2 to 0.5 and L2 regularization of 0.0001 are configured to overcome overfitting. We also configured class weights to cater for the class imbalances.

III. Results

As shown in Table I, the model exhibits good performance on the datasets that were used in this paper. It achieves an average accuracy of 81.82% on the EMODB German dataset and 99.65% on the TESS English dataset with commendable loss precision, recall, and F1 score values. This gives the insight that a combination of shallow, intermediate, and higher-level feature representation with appropriate weight configuration scores benefits the SER models in learning the heterogeneity acoustic features for better performance. The confusion matrices for the considered datasets are shown in Fig. 2. They show commendable performance in terms of individual emotional class confusion ratio.

IV. Conclusion

We proposed the MHAFF SER model which through fusion and weight scoring configuration, leverages the shallow, intermediate, and higher-level heterogeneity acoustic features learned using a transformer encoder for SER. The feature relationship learned helps the model to utilize all the representations at all levels without the loss of any information. This also reduces the possibility of gradient descent since the shallow representations that would have been lost are also utilized. In the future, we shall explore this approach for performance enhancement with data scarcity which is a persistent problem in SER systems research.

Acknowledgment

This research was financially supported by the Ministry of Trade, Industry and Energy, Korea, under the "Regional Innovation Cluster Development Program (R&D, P0025274)" supervised by the Korea Institute for Advancement of Technology (KIAT).

References

[1] 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.

[2] 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.

[3] 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.

[4] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," *arXiv preprint arXiv:1910.11263*, 2019.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model," *IEEE Access*, 2023.

[12] 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.

[13] 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.

[14] 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.

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[16] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (tess)," *Scholars Portal Dataverse*, vol. 1, p. 2020, 2020.