Advancing Speech Emotion Recognition Via Semantic And Paralinguistic Feature Fusion

Amit Somnath Dombe, Dr. Vaijanath V. Yerigeri
[1, 2]Dept of Department of Post-Graduation (Computer Science and Information Technology)

Abstract- *Speech emotion recognition is an essential component for applications like education and human-computer interaction [1]. While Deep Neural Networks (DNNs) have advanced the field, many studies ignore the semantic information present within the speech signal [2]. This paper proposes a novel framework designed to capture both semantic and paralinguistic information [5]. The model consists of a semantic feature extractor and a paralinguistic feature extractor, which are fused together using a novel attention mechanism into a unified representation. This representation is then processed by a Long Short-Term Memory (LSTM) network to model temporal dynamics [23]. Evaluated on the SEWA dataset from the AVEC challenge [16], the model achieves state-of-the-art results in the valence and liking dimensions.*

Keywords: Speech emotion recognition, Semantic features, Paralinguistic features, Deep learning, Attention mechanism.

I. INTRODUCTION

Affect recognition allows intelligent systems to understand the emotional state of users, enabling natural human-machine interaction [1], [3]. Achieving accurate recognition is challenging because emotions lack clear temporal boundaries and vary heavily between individuals [4]. Recent advancements rely on DNNs to model emotions using multiple modalities, with speech and text being the most prominent [7], [8]. Speech provides low-level prosodic characteristics, while text provides high-level semantic context (e.g., words like "love") [11], [12].

However, since textual information is inherently present in the acoustic speech signal, textual inputs during evaluation might be redundant. To address this, the authors propose an audiotextual training framework:

1) The text modality is utilized strictly during the training phase and discarded during evaluation.
2) Word2Vec [13] and Speech2Vec [14] models are trained, and their embedding spaces are aligned so that Speech2Vec representations closely match Word2Vec features [15].
3) Low-level speech characteristics are captured via a convolutional recurrent neural network [22].
4) To prevent convergence issues—particularly with the "likability" dimension—a novel disentangled attention mechanism fuses these features, ensuring the necessary information per affect dimension is isolated [24].

II. RELATED WORK

Previous studies heavily feature CNNs for spatial information and RNNs for temporal dynamics in speech [23], [25]. Multimodal emotion recognition models combining audio and visual data have also shown significant gains [8], [9]. Current methodologies either use multiple modalities during both training and evaluation or transfer cross-modal knowledge explicitly during the training pipeline [10], [26].

III. PROPOSED METHOD

The cross-modal framework extracts high-level semantic and low-level paralinguistic features, fusing them via an attention strategy before feeding them into a single-layer LSTM.

A. Semantic Feature Extractor

To extract semantic meaning purely from speech, Word2Vec [13] and Speech2Vec [14] models are trained and aligned.

- Let S be the speech embedding matrix and T be the text embedding matrix.
- A linear mapping W is learned so that WS approximates T using domain-adversarial training.
- The generator aims to deceive the discriminator into failing to identify whether an embedding originates from speech or text.
- Because high-frequency words exhibit better embedding quality, a dictionary of frequent words is used to refine W using Singular Value Decomposition (SVD) on a matrix built from k subset vectors.

B. Paralinguistic Feature Extractor

The paralinguistic network utilizes raw waveform inputs processed by three 1-D CNN layers with Rectified Linear Unit (ReLU) activations and max-pooling [22]:

- Layer 1: Convolution (Kernel 8, Stride 1, 50 Channels) followed by Max-pooling (Kernel 10, Stride 10).
- Layer 2: Convolution (Kernel 6, Stride 1, 125 Channels) followed by Max-pooling (Kernel 5, Stride 5).
- Layer 3: Convolution (Kernel 6, Stride 1, 125 Channels) followed by Max-pooling (Kernel 5, Stride 5).

C. Fusion Strategies

Two fusion approaches are compared:

1) Concatenation: A standard feature-level fusion linking semantic and paralinguistic vectors.
2) Disentangled Attention Mechanism: Both feature sets are linearly projected into the same vector space and fused via a softmax attention mechanism. To disentangle the emotion space, the output passes through three separate fully-connected layers (representing arousal, valence, and liking). Hierarchical attention is then applied to combine these specialized vectors for the final prediction.

IV. DATASET

The framework was evaluated using the Sentiment Analysis in the Wild (SEWA) dataset from the AVEC 2017 challenge [16].

- It contains audio, visual, and text data mapping three continuous dimensions: arousal, valence, and liking [20].
- Recordings feature 64 participants (32 pairs) discussing a 90-second commercial for up to 3 minutes.
- The dataset is split into training (17 pairs), development (7 pairs), and test (8 pairs) partitions [16].

V. EXPERIMENTS

A. Experimental Setup

- The Adam optimization method was utilized with a learning rate of 10^-4 [27].
- Mini-batches of 25 samples were used with a sequence length of 300 and a dropout rate of 0.5 [28].
- Raw waveforms were segmented into 10-second sequences at a 22050 Hz sampling rate.

B. Objective Function

The loss function is based on the Concordance Correlation Coefficient (ρ_c), measuring the agreement between predictions and the gold standard [21]. The network trains equally on the arousal, valence, and liking dimensions.

C. Ablation Study

- Vector Spaces: The aligned Speech2Vec model outperformed both standard Speech2Vec and the paralinguistic-only network, proving that paralinguistic information is retained and semantic alignment is beneficial [15].
- Fusion Strategy: The disentangled attention strategy (average ρ_c of .435) significantly outperformed standard concatenation (average ρ_c of .387) across all emotional dimensions.

D. Results

When compared against the top AVEC 2017 challenge models [17], [18], [19] (using only audio modalities for fairness), the proposed architecture achieved:

- State-of-the-art performance in the valence dimension (.503) and the liking dimension (.312).
- Second-best performance in the arousal dimension (.429).
- Superior generalization capabilities, particularly for the likability metric, avoiding the significant performance drops seen in competing models when moving from development to test datasets [17].

VI. CONCLUSION

This paper successfully demonstrates an audiotextual training framework that relies purely on the acoustic signal during evaluation. By aligning Word2Vec and Speech2Vec embedding spaces, the model accurately extracts semantic features from speech. Combined with a novel disentangled attention mechanism, the network produced state-of-the-art results on the SEWA dataset for valence and liking [16]. Future work aims to consolidate the architecture into a single network to reduce parameter count and test the framework on categorical emotion datasets.

VII. APPENDIX

Appendix A: System Specifications
The deep learning models and feature extraction pipelines were implemented using Python 3.8. The neural network training process required specialized hardware acceleration. Key libraries utilized include TensorFlow/Keras for deep learning architectures, LibROSA for paralinguistic audio

feature extraction, and Gensim for generating Word2Vec and Speech2Vec embeddings.

VIII. ACKNOWLEDGMENT

I would like to express my profound gratitude to my project guide and co-author, Dr. Vaijanath V. Yerigeri, for his invaluable guidance, continuous support, and motivation throughout this research work. I also extend my sincere thanks to the Department of Post-Graduation (Computer Science and Information Technology) at M.B.E.S. College of Engineering, Ambajogai, for providing the necessary infrastructure, academic environment, and resources to successfully complete this project.

REFERENCES

[1] R. Picard, Affective Computing, MIT Press, 1997.

[2] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011." Artificial Intelligence Review, pp. 155-177, 2015.

[3] P. Tzirakis, S. Zafeiriou, and B. Schuller, "Chapter 18 - Real-world automatic continuous affect recognition from audiovisual signals," in Multimodal Behavior Analysis in the Wild, pp. 387-406. Elsevier, 2019.

[4] D. Kollias et al., "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," International Journal of Computer Vision, pp. 907-929, 2019.

[5] B. Schuller. "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," Communications of the ACM, pp. 90-99, 2018.

[6] L. Stappen et al., "Muse 2020 challenge and workshop... Emotional car reviews in-the-wild," in Proc. ACM International on Multimodal Sentiment Analysis, 2020, pp. 35-44.

[7] J. Zhang et al., "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," Information Fusion, pp. 103-126, 2020.

[8] P. Tzirakis et al., "End-to-end multimodal emotion recognition using deep neural networks," IEEE Journal of Selected Topics in Signal Processing, pp. 1301-1309, 2017.

[9] S. Albanie et al., "Emotion recognition in speech using cross-modal transfer in the wild," in Proc. ACM Multimedia, 2018, pp. 292-301.

[10] P. Tzirakis, S. Zafeiriou, and B. Schuller, "End2You-The Imperial Toolkit for Multimodal Profiling by End-to-End Learning.," arXiv preprint arXiv: 1802.01115, 2018.

[11] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in Proc. IEEE Spoken Language Technology, 2018, pp. 112-118.

[12] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis-A review of research topics, venues, and top cited papers," Computer Science Review, pp. 16-32, 2018.

[13] T. Mikolov et al., "Distributed representations of words and phrases and their compositionality," in Proc. Advances in neural information processing systems (NeurIPS), 2013, pp. 3111-3119.

[14] Y.-A. Chung and J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," Proc. Interspeech, pp. 811-815, 2018.

[15] Y.-A. Chung et al., "Unsupervised cross-modal alignment of speech and text embedding spaces," in Proc. Advances in neural information processing systems (NeurIPS), 2018, pp. 7354-7364.

[16] F. Ringeval et al., "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in Proc. ACM Multimedia Workshop, 2017, pp. 3-9.

[17] S. Chen et al., "Multimodal multi-task learning for dimensional and continuous emotion recognition," in Proc. ACM Multemedia Workshops, 2017, pp. 19-26.

[18] T. Dang et al., "Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017," in Proc. ACM Multemedia Workshops, 2017, pp. 27-35.

[19] J. Huang et al., "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in Proc. ACM Multemedia Workshops, 2017, pp. 11-18.

[20] B. Schuller et al., "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," 2018, pp. 122-126.

[21] P. Tzirakis et al., "End-to-end multimodal affect recognition in real-world environments," Information Fusion, vol. 68, pp. 46-53, 2021.

[22] G. Trigeorgis et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. IEEE ICASSP, 2016, pp. 5200-5204.

[23] P. Tzirakis, J. Zhang, and B. Schuller, "End-to-end speech emotion recognition using deep neural networks," in Proc. IEEE ICASSP, 2018, pp. 5089-5093.

[24] L. Tarantino, P. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," Proc. Interspeech 2019, pp. 2578-2582, 2019.

[25] 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," Proc. Interspeech 2017, pp. 1263-1267, 2017.

[26] J. Han et al., "Implicit fusion by joint audiovisual training for emotion recognition in mono modality," in Proc. IEEE ICASSP, 2019, pp. 5861-5865.
[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[28] N. Srivastava et al., "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014.