Interspeech 2022
18-22 September 2022, Incheon, Korea

Self-supervised Representation Fusion for Speech and Wearable Based Emotion Recognition

Vipula Dissanayake[1], *Sachith Seneviratne*[2], *Hussel Suriyaarachchi*[1], *Elliott Wen*[1], *Suranga Nanayakkara*[1]

[1]Augmented Human Lab, Auckland Bioengineering Institute, The University of Auckland, New Zealand. [2]Transport, Health and Urban Design Research lab, Melbourne School of Design, The University of Melbourne, Australia.

vipula@ahlab.org, sachith.seneviratne@unimelb.edu.au, hussel@ahlab.org, elliott@ahlab.org, suranga@ahlab.org

Abstract

Even with modern-day advanced machine learning techniques, Speech Emotion Recognition (SER) is a challenging task. Speech signals alone might not provide enough information to build robust emotion recognition models. The widespread usage of wearable devices provides multiple signal streams containing physiological and contextual cues, which could be incredibly beneficial to improving an SER system. However, research around multimodal emotion recognition with wearable and speech signals is limited. Also, the scarcity of annotated data for such scenarios limits the applicability of deep learning techniques. This paper presents a self-supervised fusion method for speech and wearable signals and evaluates its usage in the SER context. We further discuss three different fusion techniques in the context of multimodal emotion recognition. Our evaluations show that pretraining in the fusion stage significantly impacts the downstream emotion recognition task. Our method was able to achieve F1 Scores of 82.59% (arousal), 83.05% (valence) and 72.95% (emotion categories) for K-EmoCon dataset.

Index Terms: speech emotion recognition, multimodal, self-supervised learning, Signal fusion, computational paralinguistics

1. Introduction

Speech emotion recognition (SER) has been drawing considerable attention from researchers recently [1], since it enables many useful applications such as human-robot interaction [2], online learning [3], behaviour assessment [4], and healthcare [5].

Researchers have explored various machine learning technologies and achieved many successes for SER. A recent advancement is multimodal deep learning, which allows a network to exploit supplementary/complementary information from rich data sources to improve performance and robustness. For an SER system, two common modalities combined are text and video [6].

With the proliferation of sensor-rich wearable devices, we now have convenient access to physiological and contextual signals from users. Integrating these signals into an SER system may significantly enhance its performance and reliability, considering they originate from the Autonomous Nervous System (ANS) activity and can hardly be triggered/suppressed by any conscious or intentional control. [7].

However, the research on multimodal emotion recognition with speech and wearable signals is still limited. One major roadblock lies in the scarcity of labelled data. Specifically, many existing works rely on human-annotated datasets, which are small in size.

There is a growing amount of unlabelled data. For example, users upload terabytes of audio and video content daily to social media platforms, and wearable devices capture live streams of signals such as heart rate and accelerometer. However, the scarcity of labelled data is a common problem in the context of SER. The extent of the scarcity increases when it comes to multimodality, such as speech and signals from wearable devices.

In order to expand the limited body of knowledge in the literature and overcome issues with data scarcity, this paper introduces a self-supervised training approach to learning the relationship in multimodal data using pretrained unimodal representations. We reuse existing unimodal signal representations from the literature [8, 9] to represent speech, Blood Volume Pulse (BVP), Electrodermal Activity (EDA), skin temperature (TEMP) and Accelerometer (ACC). We study our self-supervised fusion training approach with three different representation fusion mechanisms (weighted, unweighted and attention-based). We evaluate our proposed method in three downstream tasks related to emotion recognition (arousal, valance and emotion categories) with the K-EmoCon [10] dataset.

Our evaluation shows significant improvement in arousal *(from 66.59% to 82.59%)*, valence *(from 67.60% to 83.05%)* and categorical emotion *(from 59.84% to 72.95%)* recognition performance with multimodal data using our proposed signal representation fusion approach.

The main contribution of this paper is the introduction of a novel self-supervised fusion training approach for pretrained signal representations for multimodal emotion recognition. We extensively evaluate our proposed method with multiple training approaches and multiple training mechanisms.

2. Related Work

The amount of information covered by multiple modalities is usually higher than unimodality; this increment of information could be identified as supplementary information and complementary information [11]. Adding that information could lead to higher accuracy and higher robustness of prediction models. Out of fusion levels in machine learning (data level, feature level and decision level), feature level fusion has been the most common way of fusing multimodal signals [12]. With the advancement of representation learning techniques, it is common to fuse embeddings of pretrained representations of unimodality

10.21437/Interspeech.2022-11258

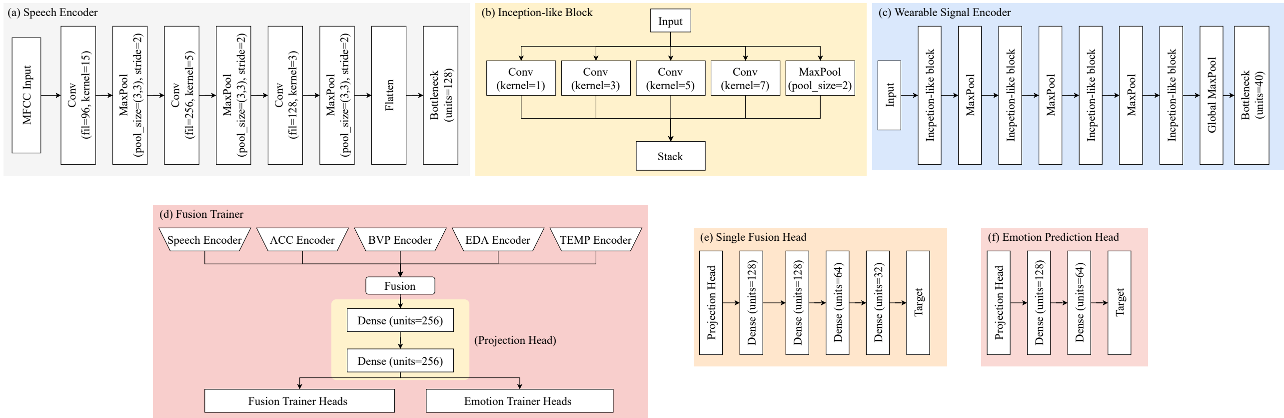

Figure 1: *Building components and Model Architectures. (a) Speech Signal encoder inspired from [8]. (b) Basic building block of the wearable signal encoder models inspired from [9]. (c) Wearable signal encoder architecture, built with the component visualised in (b). (d) Proposed model architecture for fusion. (e) Generic structure of the proposed fusion head. (f) Generic structure of proposed emotion recognition head.*

to construct the multimodal network [13, 14]. In this work, we follow a similar approach of feature-level fusion with pretrained unimodal representations.

In the current body of literature on multimodal emotion recognition, the most common modalities used are speech, text and video [15, 16, 17, 14, 18]. Although wearable devices provide access to multiple streams of physiological and contextual signals, there is a limited amount of research that looks into combining speech and signals from wearables for emotion recognition [19]. Thus, this paper focus on fusing speech and wearable signals in the context of emotion recognition.

A majority of multimodal emotion recognition work has followed a similar approach of concatenating individual modalities at the feature level and proceeded to do an emotion recognition task [15, 14, 18]. However, this approach could lead to over-reliance on a single modality [16]. In order to overcome that issue, Georgiou et al. [16] has proposed a method based on masking to systematically regulate the signal modalities to enhance supplementary/complementary information and show improvements in the downstream emotion recognition task. Further, a majority of multimodal emotion recognition work learns the mutual information of the modality when they train for the final task [17, 14, 16, 19]. Given that the datasets for emotion recognition are limited in size, models have less opportunity to learn complementary/supplementary information.

In this paper, we add to the existing body of literature by investigating the ways to fuse multimodel data for speech emotion recognition. Further, our approach uses a self-supervised method that does not require annotated data.

3. Method

3.1. Model

In developing the framework for our proposed fusion method, we delved into prior work for the initial task of signal encoding. Recent work by Dissanayake et al. presents signal encoder models tailored for both speech and wearable (ACC, BVP, EDA, TEMP) signals [8, 9]. Figure 1(a)) shows the speech signal encoder proposed by the authors, which is pre-trained with the VoxCeleb [20] dataset as an autoencoder. The resulting model offers better generalisability across different corpora [8] and thus inspired us to draw heavily from their work to build our speech encoder.

We also adopted a model and method introduced by Dissanayake et al. named ‘SigRep’ for wearable signal encoding [9]. Figure 1(c) illustrates the architecture of the signal encoder proposed in SigRep for a single modality. The encoder is built with a basic building block called ‘inception like block’ (see Figure 1(b)). This mechanism demonstrated robustness against signal losses expected in wearable devices [9]. Following SigRep, we pre-trained encoder networks using contrastive learning for each wearable signal modality. For this training, we used Affective Road [21], CASE [22], CLAS [23], PPG Field Study [24], WESAD [25] datasets.

To construct a neural network architecture for fusing these pre-trained signal encoders, we developed the fusion training network outlined in Figure 1(d). Output embeddings from each encoder network are fused in the fusion layer with three different fusion strategies (see Section 3.3). We then projected the fused embeddings with two fully connected layers to feed the fusion training heads. The structure of a single fusion head can be visualised in Figure 1(e). We used multiple fusion trainer heads, which we will discuss in the fusion training section.

Upon training of the fusion network, we discarded the fusion trainer heads and replaced them with emotion recognition heads (see Figure 1(f)) for the downstream task of emotion recognition. Each emotion recognition head has two fully connected layers and a softmax layer for the emotion target.

3.2. Self-supervised Fusion Training

Different signal modalities may contain information that complement each other. We investigated this using the VerBio dataset [26], which contains both speech and wearable signals and tried to extract any complementary information among them. In order to do that, we used a multi-task self-supervised approach. First, we randomly masked a part of each signal modality with zeros for each training example. Then, we calculated the masked area’s statistical features (mean, standard deviation, min, max, skewness, kurtosis, number of peaks), resulting in 40 values for five signal modalities. We assigned one fusion head (see Figure 1(e)) per statistical value and built our fusion trainer network (see Figure 1(d)). During the training process of our fusion network, we fed the network with five masked signals and let the network predict the 40 statistical values simultaneously. We treated our prediction tasks as regression problems; thus, optimisation was performed using the Adam algorithm [27] with a learning rate of 0.001, while the loss for each task was calculated using mean absolute error.

3.3. Fusion Mechanisms

We studied three different fusion techniques in our fusion training strategy.

1. **Unweighted Fusion.** With unweighted fusion, we concatenated embeddings from each signal modality. Figure 2(a) illustrates the unweighted fusion strategy. The concatenated embedding is then used in further fusion training and emotion training.
2. **Weighted Fusion.** In weighted fusion strategy, we introduced a trainable scalar to the embeddings of individual signal modalities. As illustrated in Figure 2(b), each embedding is multiplied by the corresponding scalar value before concatenation for the fused embedding. The weights $(\alpha 1, \alpha 2, \alpha 3, \alpha 4, \alpha 5)$ of each embedding is learned through the fusion training stage.
3. **Attention Based Fusion.** We used multi-head self-attention for each modality in the attention-based fusion strategy as shown in Figure 2(c). The output of each attention layer is then concatenated to construct the fusion embedding.

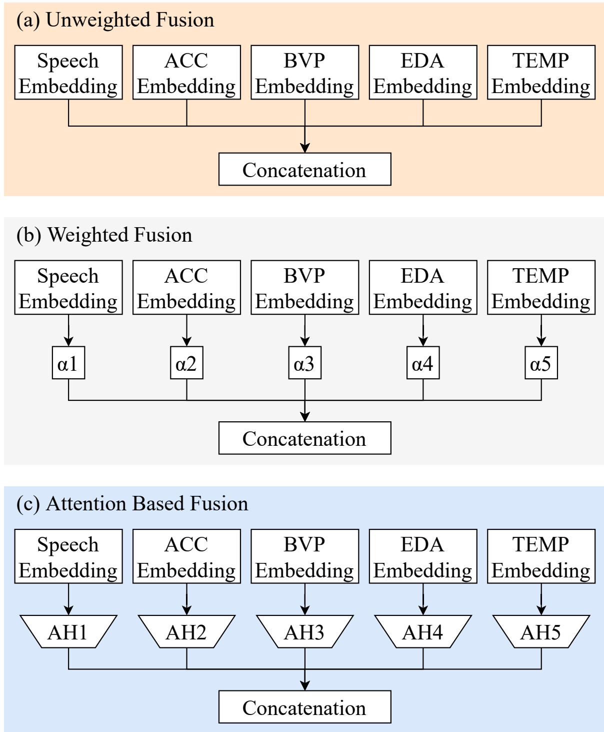

Figure 2: *Fusion Strategies. (a) embedding from each signal representation is concatenated to create the fusion embedding. (b) Each representation is multiplied by a trainable scalar α before concatenation. (c) Embedding of each signal goes through a multi-headed self attention layer before concatenation for the fusion embedding.*

4. Evaluation

4.1. Datasets

We evaluated our method using the K-EmoCon [10] dataset, which contains multimodal signals captured from 28 participants during a dyadic debate scenario. The K-EmoCon dataset consists of speech, video, ACC, BVP, EDA, TEMP, electrocardiogram (ECG) and electroencephalogram (EEG) signals. Furthermore, the dataset is annotated for arousal, valence and categorical emotions (cheerful, happy, angry, nervous, sad) in three different perspectives (self, debate opponent and external annotator). We used speech, ACC, BVP, EDA and TEMP signals together with annotations of self-reported arousal, valence and emotion categories from the dataset for our study. The dataset contained signal data recording roughly about 170 minutes.

4.1.1. Data Preprocessing

Preprocessing of the K-EmoCon [10] dataset followed a similar protocol to that conducted by prior work [8, 9]. We split all signals into four-second windows with a one-second overlap. Mel frequency cepstral coefficients (MFCCs) were extracted from speech signals as features with 40 coefficients. Although the K-EmoCon [10] dataset provides arousal and valence annotations in five stages, the benchmarking work for emotion recognition put forward by Dissanayake et al. [9] binned these values into three levels. Furthermore, they used the annotations for emotion categories, usually a score between 1 and 4, to identify the most prominent emotion and map it to one of six classes (cheerful, happy, angry, nervous, sad and neutral). Consequently, we applied the same procedure concerning both the binning of arousal and valence values and the mapping of emotion categories in our work.

4.2. Experiment Setup

All experiments investigated our models' success at predicting arousal, valence and emotion categories. Accuracy and F1-Scores served as the measures for determining model performance in our evaluation. These results were then validated using the leave-one-out cross-validation technique.

Our experiments involved using different model configurations to assess the effect of fusion training and its mechanisms on emotion recognition. To explore the effect of fusion training, we created two primary configurations where one contained pre-trained weights from the fusion training stage while the other did not.

- **With fusion training:** We loaded the model (*see Figure 1(d) with emotion trainer heads*) with pretrained weights from the fusion trainer stage. We then froze the weights of the model upto the projection head.
- **Without fusion training:** We loaded the model with pre-trained weights of encoders. We then froze the weights in encoders.

Three further sub-configurations allowed us to evaluate the effect of the three different fusion mechanisms (*see Section 3.3*) in both the presence and absence of fusion training. In all experiments we used Adam optimiser [27] with a learning rate of 0.001 and categorical cross-entropy as the loss function for this training phase. Further, we fixed the number of epochs at 32 with early stopping.

Experiment	Arousal		Valence		Categorical	
	Accuracy	F1Scores	Accuracy	F1Scores	Accuracy	F1Scores
Speech Only	0.5770 (± 0.0180)	0.5673 (± 0.0138)	0.6078 (± 0.0197)	0.6066 (± 0.192)	0.5682 (± 0.0215)	0.5552 (± 0.0199)
Wearable Only (SigRep)	0.6493	0.6291	0.5906	0.5615	0.5121	0.5028
Speech + Wearable (Unweighted) w/o fusion training	0.7382 (± 0.0355)	0.7064 (± 0.0328)	0.7312 (± 0.0353)	0.7024 (± 0.0346)	0.6589 (± 0.0277)	0.6565 (± 0.0251)
Speech + Wearable (Weighted) w/o fusion training	0.7049 (± 0.0419)	0.6821 (± 0.0416)	0.7242 (± 0.0248)	0.7021 (± 0.0243)	0.6343 (± 0.0376)	0.6352 (± 0.0353)
Speech + Wearable (Attention Based) w/o fusion training	0.7048 (± 0.0376)	0.6659 (± 0.0385)	0. 7006 (± 0.0332)	0.6760 (± 0.0323)	0.6038 (± 0.0250)	0.5984 (± 0.0253)
Speech + Wearable (Unweighted) w/ fusion training	0.8152 (± 0.0236)	0.7820 (± 0.0226)	0.7923 (± 0.0281)	0.7701 (± 0.0276)	0.7205 (± 0.0360)	0.7084 (± 0.0342)
Speech + Wearable (Weighted) w/ fusion training	0.8236 (± 0.0175)	0.7978 (± 0.0174)	0.7918 (± 0.0261)	0.7734 (± 0.0243)	0.7186 (± 0.0369)	0.6906 (± 0.0324)
Speech + Wearable (Attention Based) w/ fusion training	**0.8469 (± 0.0141)**	**0.8259 (± 0.0140)**	**0.8488 (± 0.0158)**	**0.8305 (± 0.0153)**	**0.7512 (± 0.0308)**	**0.7295 (± 0.0297)**

Table 1: *Emotion Prediction Performance for K-EmoCon [10] dataset. We report prediction accuracy, and macro F1Score for speech signal only, wearable signals only (results reported in [9]), without fusion training and with fusion training for each fusion strategy. All the results tabulated here are based on leave one subject out cross validation. Standard error value for each performance measurement is reported within parenthesis.*

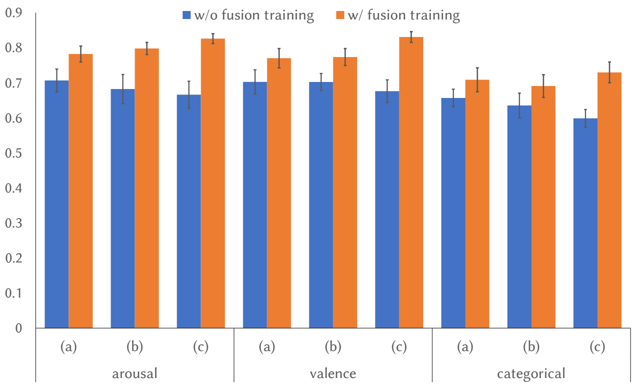

Figure 3: *Emotion recognition performance with vs without fusion training. F1Score for each setting is plotted in the graph. (a), (b) and (c) represents unweighted, weighted and attention-based fusion mechanism. Error bars show the standard error*

5. Results & Discussion

Table 1 tabulates the overall results of the series of experiments. In addition to these results, we also report the emotion recognition performance recorded in SigRep [9], which uses signals from speech and wearable sources independently. As anticipated, the prediction performance for all emotion recognition tasks improved when speech and wearable signals were combined; Student's t-test has confirmed the statistical significance ($p < 0.05$). This observation indicates that supplementary information of different modalities positively affects the downstream task of a multimodal prediction system.

To assess the effect of fusion training, we compared emotion recognition performance with and without fusion training. We evaluated three emotion targets (arousal, valence and categorical) for each fusion mechanism (unweighted, weighted and attention). Thus, we obtained nine accuracy and F1-Score pairs for comparison between the two primary model configurations. When comparing emotion recognition performance of fusion training against without fusion training, we observed improved mean accuracy and mean F1-Scores for all three prediction tasks independent of their fusion mechanism. We further analysed the difference in performance among the two configurations using T-tests, where results revealed a statistically significant improvement ($p < 0.05$) in seven out of nine pairs. Valence prediction with unweighted fusion and categorical emotion prediction with weighted fusion did not indicate a significant improvement. Figure 3 demonstrates the improvement of F1Sores with and without fusion training approach. This observation indicates that complementary information learned in the fusion training process has significantly positively affected the emotion recognition tasks.

Another interesting observation in our result is that unweighted fusion demonstrated the highest performance without fusion training, while attention-based fusion had the lowest performance. However, attention-based fusion outperformed weighted and unweighted fusion mechanisms in the fusion training configuration. Although we did not observe a statistical significance across fusion mechanisms in either configuration, this observation could indicate that simpler models may have an advantage in low data scenarios; however, a complex system can outperform simpler methods when there is enough data. We suggest this observation requires further investigation with a substantial amount of data.

6. Conclusion

This paper proposed a self-supervised fusion training approach for speech and wearable signal representations. Our proposed fusion training approach attempts to enhance multiple modalities' complementary information. We studied three different fusion mechanisms (weighted, unweighted and attention-based) with our fusion training approach. Our evaluations show a significant improvement in emotion recognition performance in K-EmoCon dataset with the proposed fusion approach. The evaluation results indicate that the complementary information in a multimodal system plays a critical role when fusing speech and wearable signals for emotion recognition. We also observed an improved emotion recognition performance with the attention-based fusion mechanism compared to weighted and unweighted fusion. We suggest further investigations are required to evaluate the effect of different fusion mechanisms with a larger dataset to train representation fusion.

7. Acknowledgement

This work has been supported by the doctoral scholarship of the University of Auckland and the Assistive Augmentation Research Grant through the Entrepreneurial Universities (EU) initiative of New Zealand.

8. References

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[3] L. Cen, F. Wu, Z. L. Yu, and F. Hu, "A real-time speech emotion recognition system and its application in online learning," in *Emotions, technology, design, and learning*. Elsevier, 2016, pp. 27–46.

[4] N. Kamaruddin and A. Wahab, "Driver behavior analysis through speech emotion understanding," in *2010 IEEE Intelligent vehicles symposium*. IEEE, 2010, pp. 238–243.

[5] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[6] M. Ren, W. Nie, A. Liu, and Y. Su, "Multi-modal correlated network for emotion recognition in speech," *Visual Informatics*, vol. 3, no. 3, pp. 150–155, 2019.

[7] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. IEEE, 2011, pp. 410–415.

[8] V. Dissanayake, H. Zhang, M. Billinghurst, and S. Nanayakkara, "Speech Emotion Recognition 'in the Wild' Using an Autoencoder," in *Proc. Interspeech 2020*, 2020, pp. 526–530.

[9] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara, "Sigrep: Toward robust wearable emotion recognition with contrastive representation learning," *IEEE Access*, vol. 10, pp. 18 105–18 120, 2022.

[10] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, pp. 1–16, 2020.

[11] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[12] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[13] S. Seneviratne, "Contrastive representation learning for natural world imagery: Habitat prediction for 30,000 species," *CLEF working notes*, 2021.

[14] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," *Proc. Interspeech 2020*, pp. 3755–3759, 2020.

[15] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition," *Proc. Interspeech 2021*, pp. 4518–4522, 2021.

[16] E. Georgiou, G. Paraskevopoulos, and A. Potamianos, "M3: MultiModal Masking Applied to Sentiment Analysis," in *Proc. Interspeech 2021*, 2021, pp. 2876–2880.

[17] P. Kumar, V. Kaushik, and B. Raman, "Towards the Explainability of Multimodal Speech Emotion Recognition," in *Proc. Interspeech 2021*, 2021, pp. 1748–1752.

[18] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[19] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[21] N. E. Haouij, J.-M. Poggi, S. Sevestre-Ghalila, R. Ghozi, and M. Jaïdane, "Affectiveroad system and database to assess driver's attention," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018, pp. 800–803.

[22] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific data*, vol. 6, no. 1, pp. 1–13, 2019.

[23] V. Markova, T. Ganchev, and K. Kalinkov, "Clas: A database for cognitive load, affect and stress recognition," in *2019 International Conference on Biomedical Innovations and Applications (BIA)*. IEEE, 2019, pp. 1–4.

[24] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg: large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, p. 3079, 2019.

[25] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[26] M. Yadav, M. N. Sakib, E. H. Nirjhar, K. Feng, A. Behzadan, and T. Chaspari, "Exploring individual differences of public speaking anxiety in real-life and virtual presentations," *IEEE Transactions on Affective Computing*, 2020.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.