INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

ReCLR: Reference-Enhanced Contrastive Learning of Audio Representation for Depression Detection

Pingyue Zhang, Mengyue Wu†, Kai Yu†

Shanghai Jiao Tong University, China

{williamzhangsjtu, mengyuewu, kai.yu}@sjtu.edu.cn

Abstract

Contrastive self-supervised learning has seen great success in computer vision while been less investigated in the audio processing field, in particular depression detection, a socially critical challenge. Detecting depression from one's speech has been examined via various audio representations, including acoustic feature combinations and model-based ones. This paper proposes to obtain depressive audio representations by departing speech via reference features from an emotion recognition model. Furthermore, we propose a reference-enhanced contrastive learning (ReCLR) to select fine-grained positive instances and allocate weight to negative instances. The depression detection results indicate that contrastive learning is effective in such an audio task. Moreover, our modified ReCLR strategy has outperformed contrastive training without references.[1]

Index Terms: self-supervised learning, depression detection, emotion recognition, contrastive learning

1. Introduction

Depression has become one of the most common mental disorders worldwide and a major health challenge to a large population, according to World Health Organizations [2]. Research on automatic depression detection has received increasing amount of attention, mainly including text-based detection from social media posts and audio-based from conversation recordings. Compared with social media posts, conversational audio data with official labels about one's mental state is much more difficult to gather, resulting in a greater challenge for speech-based depression detection. For supervised learning, a large amount of labeled data is necessary. Therefore, despite the recent success of rapidly developing techniques of deep learning, an improvement in speech-based automatic depression detection is not significant.

As a matter of fact, clinical diagnoses for depression are mainly drawn from conversations, where audio cues provide crucial clues for a psychiatrist [1]. Audio-based depression detection from conversational data can be a helpful and trustworthy tool to screen depression. However, due to the complexity of depressive symptoms, robust audio feature extraction remains an arduous task. Previous speech-based detection work has experimented on various acoustic features, like prosodic features, spectral features, and cepstral features (e.g., Mel-Frequency Cepstral Coefficients [2]), and more recently, feature combinations like COVAREP [3], which consists of a high-dimensional feature vector covering common features such as fundamental frequency and peak slope. Deep learning methods have been employed to extract high-level feature representations [4, 5]. Despite the tryout on different features and models, the F1 accuracy generated from speech-based depression detection is average.

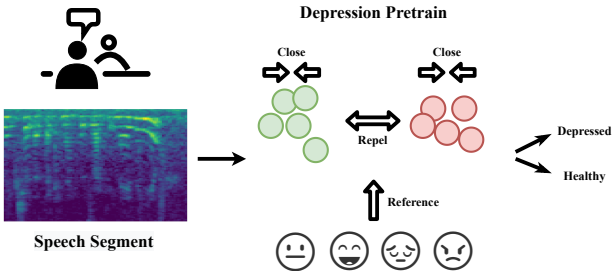


Figure 1: *Overview of the audio-based depression detection framework. We pretrain a depression detection model via contrastive self-supervised learning. Speech emotion features are used to better separate positive and negative instances.*

Self-supervised learning can be an appropriate solution to the aforementioned challenges. It defines a proxy task to pretrain the model and treat the main task as a downstream one. Such a pretraining process can partially alleviate the data sparsity problem, where the model is expected to learn knowledge from the proxy task instead of being trained from scratch. Contrastive learning, on the other hand, can help with the feature extraction as it mainly learns by maximizing agreement between positive instances while minimizing similarity between negative ones.

When conducting contrastive learning, how to mine positive instances becomes the key question. Traditional contrastive learning methods like SimCLR [6], MoCo [7] treats differently augmented instances of the same sample as positive and those from different samples as negative. CoCLR [8] presents more competitive results by proposing a co-training method to mine hard positive samples by using other complementary views of the data. However, previous contrastive learning methods are independent of the downstream task, nevertheless some instances which could be positive pairs in the downstream task are classified as negative because they belong to different samples in the proxy task. Therefore we believe taking the downstream task into consideration during the process can further maximize data usage and boost downstream performance. Inspired by the relations between depression and emotion established in psy-

[1]† are corresponding authors. This work has been supported by National Natural Science Foundation of China (No.92048205), the Key Research and Development Program of Jiangsu Province (No.BE2022059-2), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and Alibaba Innovative Research.

[2]https://www.who.int/news-room/fact-sheets/detail/depression

10.21437/Interspeech.2023-2474

chology and psychiatry studies [9, 10], we propose to use emotion as a side view to depression detection.

To our knowledge, this is the first time of utilizing contrastive self-supervised learning on pathological audio detection, in particular, with references for training. The main contributions of this work are:

- We propose Reference-enhanced contrastive learning (ReCLR), a novel contrastive learning method which, for the first time, takes downstream task into consideration while pretraining.
- We adopt emotion features as references to depression detection, transferring the knowledge from common speech emotion recognition to depression and investigate the relation between them.
- We compare ReCLR with other contrastive learning methods on the depression detection as well as emotion recognition tasks to demonstrate the superiority of our method.

2. ReCLR Contrastive Self-supervised Pretraining Method

We aim to pretrain a model to extract one single vector from raw acoustic features for each segment. The strategy prevents the too-long-sequence problem caused by concatenating raw features. Furthermore, the strategy will have a better performance if the model can learn to capture depression-related information from raw features during the pretraining process. Therefore, we propose ReCLR, a contrastive learning pretrain method enhanced by emotion-related features as the reference. The detailed framework is introduced as follows, illustrated in Figure 2.

Our pretraining method can be separated into two phases. In the first phase, the model ϕ_e is trained on an emotion recognition task. Then in the second phase, we use the trained model ϕ_e to extract emotion-related features (reference features) from audio as additional information. Since the essence of contrastive learning is to better separate positive instances from negative ones, we believe such a reference-based training scheme is more effective in instance selection and identifying different emotion-related information.

2.1. Phase I: Reference Model Training

In order to utilize additional information as reference for a more effective contrastive training, we first need a model that can extract such reference features. As mentioned previously, we take emotion as a complimentary view to depression detection. For the purpose of extracting emotion-related features in the later pretraining process, we first need to train an audio-based emotion recognition model ϕ_e, illustrated in Figure 2 (a).

Such a feature extraction model can be realized via a standard emotion recognition task, which is usually trained on emotion dataset involving audio clips of different emotion labels. We utilize a commonly-used convolutional neural network Cnn10 proposed in PANNs [11] as ϕ_e and a fully connected to train our emotion recognition model.

Once trained, we select the model ϕ_e which has the best performance on the validation set and use it to extract emotion-related features as reference features, elaborated in Section 2.2.

2.2. Phase II: Contrastive Learning with References

The core of our pretraining method, contrastive learning with references, is illustrated in Figure 2 (b). Given a batch of depression conversational data $\mathcal{D}$ with N audio clips with the same size: $\mathcal{D} = [a_1, a_2, \cdots, a_N]$. For each clips we use ϕ_e (from Section 2.1) to extract emotion-related features as reference features $[v_1, \cdots, v_N]$.

2.2.1. SimCLR: Contrastive Learning without References

First we state one of our baseline method SimCLR, which follows the learning pattern proposed in SimCLR[6], where reference information is not used.

We use random masking method (mask by 0) $\mathcal{T}$, to generate two instances from it, and operate on each sample a_i to construct two instances: $a_k^{(i)}$, $a_q^{(i)}$. Following the MoCo-trick, we use two separate encoders of the same architecture ϕ_k, ϕ_q to extract embeddings of $a_k^{(i)}$, $a_q^{(i)}$: $z_k^{(i)} = \phi_k(a_k^{(i)}), z_q^{(i)} = \phi_q(a_q^{(i)})$, and encoder ϕ_k is momentum updated.

To formulate, for $z_q^{(i)}$ of sample i,

- Positive set: $\mathcal{P}_i = \{z_k^{(i)}\}$
- Negative set: $\mathcal{N}_i = \{z_k^{(j)}, j \neq i\}$

For sample i, the loss function is:

$$\mathcal{L}_C^{(i)} = -\log \frac{\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i} e^{s_p/\tau}}{\sum_{p\in\mathcal{P}_i} e^{s_p} + \sum_{n\in\mathcal{N}_i} e^{s_n/\tau}}, \quad (1)$$

where $\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i} e^{s_p/\tau} = e^{s_i/\tau}$ in this case and we use cosine similarity as score: $s_j = \frac{z_q^{(i)} \cdot z_k^{(j)}}{||z_q^{(i)}|| \cdot ||z_k^{(j)}||}$.

2.2.2. CoCLR: References used to find potential positives

We follow CoCLR[8] and use emotion-related reference information to enlarge the positive set.

For sample i and corresponding $z_q^{(i)}$, we denote:

- Reference similarity score: $[r_1, r_2, \cdots]$, where r_j is the cosine similarity of reference features v_i and v_j.
- Positive set:

$$\mathcal{P}_i = \{z_k^{(i)}\} \cup \{z_k^{(p)} | p \in \text{topK}(r_j)\},$$

  where topK means the highest K items. The positive set contains not only its own augmented view, but also K instances with the highest reference similarity score. That is, we treat samples which have similar emotions as positives.
- Negative set $\mathcal{N}_i$ contains the remaining z_k.

The loss function is identical to Equation (1)

2.2.3. ReCLR: Reference-enhanced Loss Function

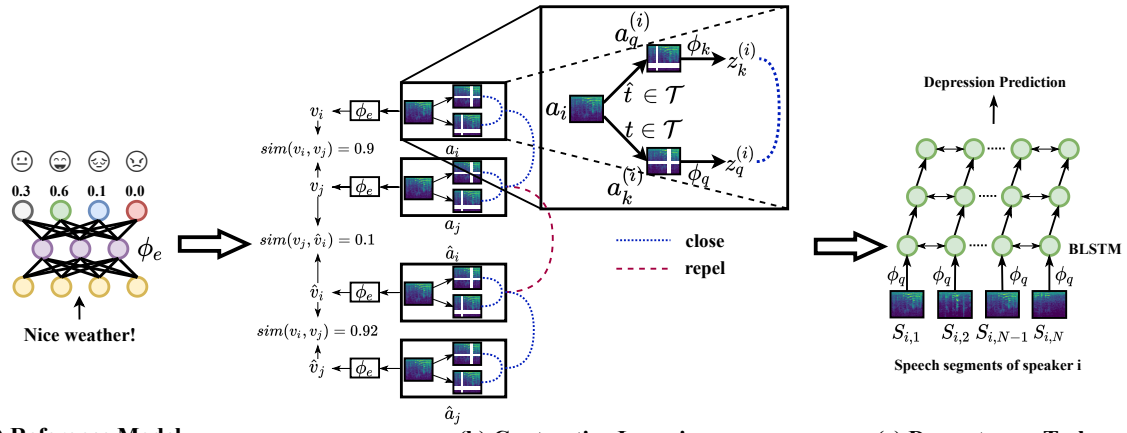
Different from all previous contrastive learning methods which mainly focus on positive instances while treat all negative ones in the same manner, we give each negative instance some weight to distinguish them from each other. We propose a new loss function to take reference similarity score r_j into consideration:

$$\mathcal{L}_R^{(i)} = -\log \frac{\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i} e^{s_p/\tau}}{\sum_{p\in\mathcal{P}_i} e^{s_p/\tau} + \sum_{n\in\mathcal{N}_i} e^{s_n\cdot(-r_n)/\tau}}, \quad (2)$$

where τ is the temperature. The equation is different from the loss function in Equation (1)

And for batch $\mathcal{D}$, the loss function is $\mathcal{L}_R = \frac{\sum_i \mathcal{L}_R^{(i)}}{N}$.

We demonstrate the effectiveness of our loss function here:



(a) Reference Model Training **(b) Contrastive Learning with References** **(c) Downstream Task: Depression Detection**

Figure 2: *Illustration of the two-phase ReCLR and its application on depression detection; (a) trains a model on emotion recognition to extract emotional audio features. (b) uses emotion features as references to pretrain a depression model by enclosing similar instances while repelling different ones; (c) use the pretrained model to extract features for the downstream depression detection task.*

$$\nabla_{z_q^{(i)}} \mathcal{L}_{R/C} = \sum_{n \in \mathcal{N}_i} \nabla_{s_n} \mathcal{L}_{R/C}^{(i)} \cdot \nabla_{z_q^{(i)}} s_n + \sum_{p \in \mathcal{P}_i} \nabla_{s_p} \mathcal{L}_{R/C}^{(i)} \cdot \nabla_{z_q^{(i)}} s_p, \tag{3}$$

where,

$$\nabla_{z_q^{(i)}} s_n = z_k^{(n)}, \nabla_{z_q^{(i)}} s_p = z_k^{(p)},$$

and $\nabla_{s_p} \mathcal{L}_{R/C}^{(i)}$ is similar to both $\mathcal{L}_C^{(i)}$ and $\mathcal{L}_R^{(i)}$. As a result, we mainly consider the difference between $\nabla_{s_n} \mathcal{L}_C^{(i)}$ and $\nabla_{s_n} \mathcal{L}_R^{(i)}$.

Respectively, the gradients passed to score s_n of negative samples in $\mathcal{N}_i$ will be:

$$\nabla_{s_n} \mathcal{L}_C^{(i)} = \frac{e^{s_n/\tau}}{D_C} \cdot \frac{1}{\tau}, \nabla_{s_n} \mathcal{L}_R^{(i)} = \frac{e^{s_n \cdot (-r_n)/\tau}}{D_R} \cdot \frac{-r_n}{\tau}, \tag{4}$$

where D_C, D_R are the denominators of $\mathcal{L}_C^{(i)}, \mathcal{L}_R^{(i)}$ in Equation (1) and Equation (2).

It can be inferred:

- when $r_n < 0$ ($-r_n > 0$), which means sample n has negative reference similarity score. As a result, the gradients of two loss function have the same sign, that is, the optimization goal is to make s_n smaller.
- when $r_n > 0$ ($-r_n < 0$), which means sample n has positive similarity score on reference information. As a result, the gradients of two loss function have different signs, that is, our modified function will make s_n larger.

Furthermore, we consider when s_n stays the same, how $|r_n|$ will affect the norm of gradient, for a larger $|r_n|$ means more significant relevance on reference. For convenience, we consider function $f(x) = |x \cdot e^{s \cdot x}|$, where we let $x = -r_n \in [-1, 1], s = s_n \in [-1, 1]$.

$$\nabla_x f(x) = \begin{cases} (1 + x \cdot s)e^{s \cdot x} \geq 0, & x \in [0, 1] \\ -(1 + x \cdot s)e^{s \cdot x} \leq 0, & x \in [-1, 0], \end{cases} \tag{5}$$

and under almost no circumstance can $\nabla_x f(x)$ be 0, which means the norm of gradient increases as $|r_n|$ increases. As a result, a negative instance which has closer relation with instance i on reference will receive more attention. Further, among negative instances which have close s_n, those have larger $|r_n|$ will get more updates than ones with smaller $|r_n|$, that is, our modified loss function gives weight to the negative instances according to the reference information, which is different from traditional contrastive loss functions. Hence, negative instances which have close similarity score s_n but differ in reference score r_n will be distinguished.

Because the reference features extracted by ϕ_e may contain noise, our final loss function is a weight sum of $\mathcal{L}_R$ and $\mathcal{L}_C$ to be more robust:

$$\mathcal{L} = \lambda \mathcal{L}_R + (1 - \lambda) \mathcal{L}_C. \tag{6}$$

When $\lambda = 0$, the loss function reduces to a CoCLR one. Once the training process is done, we use ϕ_q to extract features for the downstream task.

3. Experiments

In this section, we introduce our datasets and experimental setup for different stages of our method. Downstream tasks with different methods are reported on MDD. In ablation study, we also finetune our pretrained model on emotion recognition datasets.

3.1. Dataset

MDD corpus is a large conversational dataset for major depression disorder detection (MDD) [1]. It consists of 1000 hours of speech conversation between interviewers and subjects. Among these data, we pick 588 healthy subjects and 545 depressed subjects. We conduct speaker diarization to select only the subject's contents. Then we concatenate all these contents and cut into several sequential 5-second segments as their utterances. Each segment is treated as an individual sample during pretraining. We also use the full dataset to perform depression detection, detailed in Section 3.2.

CMUMOSEI corpus is a multimodal dataset [12] for sentiment analysis and emotion recognition. Following previous

studies, we neglect sentences with a score of 0 and label them in a binary manner: label a sentence as 0 if its score is less than 0, otherwise as 1. We use this dataset to train ϕ_e and choose the model with the best performance on the validation set.

IEMOCAP corpus is a dyadic conversational dataset [13]. We select utterances labeled as "angry", "happy", "excited", "sad", "frustrated", and "neutral", where utterances with "excited" are merged into "happy" class.

3.2. Settings

When pretraining using MDD dataset, we use Cnn10 as ϕ_q and train the model using a SGD optimizer with an initial learning rate of 0.1 and reducing learning rate using a cosine annealing strategy. We train the model for 200 epochs where each epoch has 2000 iterations with a batch size of 64. We set $K = 5$ and $\tau = 0.1$. We perform emotion recognition and depression detection downstream tasks on IEMOCAP and MDD dataset. We randomly split them into a training set (70%) and a test set (30%). For emotion recognition, we use a Cnn10 followed by a fully connected layer and initialize Cnn10 with the pretrained model. For MDD, we use the pretrained Cnn10 to extract segment-level features and input to LSTM.

3.3. Results

We present results of different methods on MDD dataset in Table 1, with 10 times running under different seeds. Clearly, by using references to select potential positives, CoCLR outperforms SimCLR. By further utilizing reference information, ReCLR enhances the performance. A larger λ means assigning heavier weight to emotion-related reference information in the pretraining stage, which proves to be helpful for depression detection. The previous methods can be seen as using a CRNN which consists of a Cnn10 and LSTM with pretrained and fixed Cnn10. We further train the whole CRNN on MDD from scratch and show result in Table 1. We also load the Cnn10 with CoCLR-pretrained parameters before training to highlight the necessity of pretraining process. Note that for all methods we treat all utterances from the same speaker as positives, hence even for SimCLR method the positive set $\mathcal{P}$ may contain more than one instance. The marginal increase of CoCLR suggests that SimCLR already considers multiple positive samples, and they may weaken the top K samples for averaging $\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i} e^{s_p/\tau}$ in Equation (2).

Therefore, we also test performance of not treating different segments from the same speaker as positives, with results presented in Table 2. Results indicate that CoCLR outperforms SimCLR more significantly under this setting, further demonstrating that using emotion-related reference information benefits depression detection.

Cnn10	Method	F1 macro	F1 micro
Fixed	SimCLR	76.56 ± 2.11	76.59 ± 2.11
	CoCLR	76.63 ± 1.95	76.65 ± 1.95
	ReCLR ($\lambda = 0.2$)	77.64 ± 2.58	77.68 ± 2.59
	ReCLR ($\lambda = 0.4$)	$\mathbf{77.90 \pm 2.56}$	$\mathbf{77.94 \pm 2.56}$
Tuning	Random	69.83	70.59
	CoCLR-pretrained	74.7	74.7

Table 1: *Results of different methods on MDD*

Method	F1 macro	F1 micro
SimCLR	74.98 ± 2.62	75.00 ± 2.61
CoCLR	75.66 ± 2.63	75.68 ± 2.63
ReCLR ($\lambda = 0.4$)	$\mathbf{76.60 \pm 2.49}$	$\mathbf{76.62 \pm 2.50}$

Table 2: *Results of different methods on MDD, where difference segments from the same speaker are not positives*

3.4. Ablation study

Besides depression detection, we also finetune our model to conduct emotion recognition on IEMOCAP dataset and present results in Table 3. Generally, CoCLR and ReCLR exhibit similar improvement, compared with random initialization and SimCLR. However, a higher λ leads to slight performance drop. This might be due to the fact that binary emotion classification on MOSEI is not consistent with multiple emotions on IEMOCAP.

Method	F1 macro	F1 micro
Random	52.20	51.40
SimCLR	55.67 ± 0.91	54.85 ± 0.92
CoCLR	$\mathbf{57.14 \pm 0.88}$	$\mathbf{56.44 \pm 0.83}$
ReCLR ($\lambda = 0.2$)	57.10 ± 0.89	56.42 ± 0.85
ReCLR ($\lambda = 0.4$)	56.59 ± 0.89	55.92 ± 1.02

Table 3: *Results of different methods finetuning on IEMOCAP.*

To further explore the effect of λ value, we finetune our model on MOSEI and present results in Table 4. Here, the ablation results indicate that larger λ should contain more emotion-related information learned from MOSEI dataset.

Method	F1 macro	F1 micro
CoCLR	66.65 ± 1.20	69.25 ± 1.23
ReCLR ($\lambda = 0.2$)	67.14 ± 1.22	69.58 ± 1.25
ReCLR ($\lambda = 0.4$)	$\mathbf{67.21 \pm 1.11}$	$\mathbf{70.12 \pm 0.82}$

Table 4: *Results of different methods finetuning on MOSEI.*

4. Conclusion

Due to the limited data problem, depression detection is quite a challenging task. Hence, we use the self-supervised method to pretrain the model and treat depression detection as its downstream task. This work proposes ReCLR, a reference-enhanced contrastive learning method of audio representation for depression detection. We propose a novel reference-enhanced loss function to conduct contrastive learning in a more fine-grained manner, with the help of reference information that relates to the downstream task. We choose emotion-related information as reference, transferring the knowledge from the emotion recognition dataset to depression and examine the relation between them. By comparing our method with some baseline contrastive learning methods, we demonstrate the superiority of our approach. Ablation studies show that our pretraining also benefits emotion recognition tasks.

5. References

[1] Y. Di, J. Wang, W. Li, and T. Zhu, "Using i-vectors from voice features to identify major depressive disorder," *Journal of Affective Disorders*, vol. 288, pp. 161–166, 2021.

[2] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *arXiv preprint arXiv:1909.07208*, 2019.

[3] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[4] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3d facial expressions," *arXiv preprint arXiv:1811.08592*, 2018.

[5] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech 2018*, 2018, pp. 1716–1720. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2522

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in . *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in . *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[8] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *arXiv preprint arXiv:2010.09709*, 2020.

[9] J. Joormann and M. Quinn, "Cognitive processes and emotion regulation in depression," *Depression and anxiety*, vol. 31, 04 2014.

[10] W. M. Vanderlind, Y. Millgram, A. R. Baskin-Sommers, M. S. Clark, and J. Joormann, "Understanding positive emotion deficits in depression: From emotion preferences to emotion regulation," *Clinical Psychology Review*, vol. 76, p. 101826, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0272735820300143

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[12] A. B. 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[13] 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.