Improving Speech Emotion Recognition Using Self-Supervised Learning with Domain-Specific Audiovisual Tasks

Lucas Goncalves and Carlos Busso

Multimodal Signal Processing (MSP) Lab., Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

goncalves@utdallas.edu, busso@utdallas.edu

Abstract

Speech emotion recognition (SER) is a challenging task due to the limited availability of real-world labeled datasets. Since it is easier to find unlabeled data, the use of *self-supervised learning* (SSL) has become an attractive alternative. This study proposes new pre-text tasks for SSL to improve SER. While our target application is SER, the proposed pre-text tasks include audiovisual formulations, leveraging the relationship between acoustic and facial features. Our proposed approach introduces three new unimodal and multimodal pre-text tasks that are carefully designed to learn better representations for predicting emotional cues from speech. Task 1 predicts energy variations (high or low) from a speech sequence. Task 2 uses speech features to predict facial activation (high or low) based on facial landmark movements. Task 3 performs a multi-class emotion recognition task on emotional labels obtained from combinations of *action units* (AUs) detected across a video sequence. We pre-train a network with 60.92 hours of unlabeled data, fine-tuning the model for the downstream SER task. The results on the CREMA-D dataset show that the model pre-trained on the proposed domain-specific pre-text tasks significantly improves the precision (up to 5.1%), recall (up to 4.5%), and F1-scores (up to 4.9%) of our SER system.

Index Terms: self-supervised learning, speech emotion recognition, audiovisual tasks.

1. Introduction

Human communication relies on many signals to effectively transmit a message with the intended semantic content, intended emphasis, and emotional content. We use multimodal cues to convey expressive information during daily interactions, including acoustic, visual and other sensory modalities [1]. Therefore, we expect that combining and leveraging the relationship across different multimodal cues can help us in generating meaningful representations for robust emotion recognition. The conventional approach for building a *speech emotion recognition* (SER) system is to solely focus on the speech modality [2–6]. This study explores the relationship between acoustic and facial cues to enhance acoustic representations for SER. We learn from the complementary information provided by audiovisual modalities, even though only speech is used during inference.

The motivation for this paper is to explore using unlabeled data to generate speech representations which can improve the performance of supervised learning methods for SER, which are often trained with a much lower amount of labeled data. *Self-supervised learning* (SSL) methods can improve performance on downstream tasks. In SSL, models are often pre-trained to make predictions on pre-text tasks for which labels can be automatically obtained. The pre-text tasks are not always related to the downstream tasks. The pre-trained models are subsequently fine-tuned on a labeled dataset to make predictions on a pre-

determined downstream task. Some of the areas where SSL has been successfully used includes question answering [7], action recognition [8], and multimodal emotion recognition [9]. For emotion recognition, SSL methods often make use of speech and/or textual/visual representations to pre-train a model. Previous studies have used *masked language modeling* (MLM) tasks, which consists of replacing some tokens by either the token $< MASK >$ or a random token and asking the model to predict the masked tokens [10, 11]. Then, these models have utilized contrastive objectives to extract speech representations, such as in wav2vec [12], which is optimized to solve a next time-step prediction task using a contrastive loss. We observe that these approaches rely on similar tasks (contrastive objective and/or MLM) when utilizing multimodal data to pre-train SER models. However, some of these tasks are not even related to SER problems. These generic tasks can also be very static, specially with text representations where words always have the same representations. This issue is problematic since emotions are dynamic. Since facial expressions are intrinsically connected with acoustic features [13], we argue that better pre-text tasks can be obtained by relying on audiovisual features that capture the intrinsic and dynamic connection between facial expressions with acoustic features, focusing on tasks that are relevant to SER problem.

This paper proposes using acoustic and facial features to generate pre-text tasks for self-supervised SER. We create three novel unimodal and multimodal pre-text tasks that leverage the complementary information contained in audiovisual features and are designed to produce discriminative SER representations. The first pre-text task is an energy task, where the energy in a speech sequence is classified into binary classes (high or low). This task exploits the fact that several emotions are characterized by high arousal, leading to higher speech energy. The second pre-text task uses speech to predict the activeness of facial expressions, where binary labels (high or low) are defined based on the variance of the movements of facial landmarks. The third pre-text task is a multi-class emotion classifier trained with acoustic features using pseudo-emotional labels obtained from combinations of facial *action units* (AUs) detected across a video sequence. These new pre-text tasks are carefully selected to capture the relationship between facial expressions with acoustic features, and generate representations with emotional content for the downstream SER task. The proposed model is trained in a multi-task environment, and the learned representations are fine-tuned for the SER tasks.

We evaluate the proposed approach with the CREMA-D corpus [14], pre-training the models with the proposed pre-text tasks on the MSP-Face corpus [15]. The experimental evaluation compares our method with a supervised method and a baseline which uses the wav2vec 2.0 method [16] to pre-train the architecture. The results show improved performance by using a pre-trained model using our pre-text tasks over alter-

native methods. We also analyze the effect that each pre-text task has on our proposed method's performance. We observe that having all the proposed pre-text tasks during pre-training is crucial to achieve strong results.

2. Related Work

The use of SSL has led to performance improvements in several downstream tasks, such as question answering [7], action recognition [8], and multimodal emotion recognition [9]. One SSL-based approach that has shown impressive results in speech related tasks is the wav2vec 2.0 framework [16]. Wav2vec 2.0 has obtained strong representations from raw audio for *automatic speech recognition* (ASR) [16]. Studies have shown that this framework is also a powerful alternative for SER tasks [17]. Keesing et al. [18] compared different feature representations for SER tasks, showing that wav2vec 2.0 was one of the best alternatives. Previous studies using SSL in SER tasks have used text-based self-supervised methods to improve performance in SER. For example, Tseng et al. [19] used a word prediction pre-text task which included acoustic features in addition to word tokens to generate joint representations of speech and text through modeling a bidirectional language model with word-aligned acoustic features. The pre-trained model was later fine-tuned to the downstream SER task. Although SSL approaches have shown good performance with text-based tasks, some models do not work well when the SSL is implemented without text masking tasks. Khare et al. [9] used a SSL method with cross-modal transformers for emotion recognition. The method achieved good performance with text. However, the ablation study showed that the text modality was essential in the architecture to achieve good performance.

We take inspiration from previous SSL studies and from studies on multimodal emotion recognition [20,21] to derive the proposed pre-text tasks. Although studies have also explored text-based or speech-based pre-text tasks, we argue that leveraging acoustic and facial cues is crucial for multimodal pre-text tasks. Facial expressions are intrinsically connected with acoustic features [13]. Therefore, we can create better models by considering audiovisual features, even if the target application is a unimodal system. The novelty on our formulation is to use acoustic information to predict on audiovisual tasks to generate emotionally meaningful acoustic representations for SER by leveraging the complementary information provided by acoustic and facial features.

3. PROPOSED APPROACH

The goal of this study is to use SSL to build emotional representations from audiovisual pre-text tasks to improve SER performance. We designed three pre-text tasks to train with a large number of unlabeled videos available on the MSP-Face dataset [15]. All datapoints used to train the pre-text tasks are obtained from two-second segments with an overlap of 0.5 seconds. The unimodal and multimodal pre-text tasks are carefully designed to (1) create representations that are discriminative for the SER task, and (2) leverage the complementary relationship between acoustic and facial features in the externalization of emotions. After presenting the three proposed pre-text tasks, this section presents the SER framework used to evaluate our SSL solution.

3.1. Task of Speech Energy

The first pre-text task is unimodal, relying only on speech. Several emotions such as happiness and anger are characterized
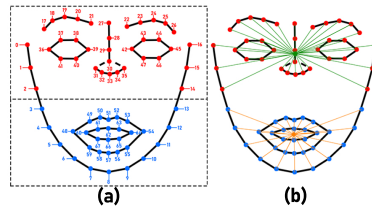

Figure 1: *Task of variance of facial landmarks. (a) Facial landmarks for the top (red) and bottom (blue) sections. (b) Centroids created for each frame for the top and bottom facial regions.*

by high level of speech energy. Motivated by this observation, our pre-text task consists of predicting audio with low and high energy. Audios are extracted from the two-second datapoints and converted into mono channel using the ffmpeg library [22]. Then, the librosa package [23] is used to compute the *short-time Fourier transform* (STFT) of audio samples and compute the *root-mean-square* (RMS) value for each segment. We dichotomized the segments into two classes using the mean RMS energy computed across all datapoints as a threshold. If the RMS energy value is above the threshold, the segment is labeled as "1." Otherwise, the segment is labeled as "0."

3.2. Task of Variance of Facial Landmarks

The second pre-text task is a multimodal problem, relying on acoustic and facial features. Facial expressions involve muscle activity that increases the movements in the face. The articulatory movements to produce speech are also visible in the face, interplaying with the externalization of emotions [24]. The multimodal pre-text task is to identify from speech whether the level of facial activity is low or high. The input of the model is speech features. All the video segments are processed with the OpenFace 2.0 toolkit [25] to extract frame-wise facial landmarks. In this task, 68 facial landmarks are obtained for each video frame. Speech articulation mostly affects the lower facial region. Therefore, we split the faces into upper and lower regions as represented in Figure 1(a). Then, we estimate the landmark centroid in each frame for the upper and lower facial regions, as illustrated in Figure 1(b). The positions of the centroids across frames are compiled and used to estimate the variance of the centroids as a proxy of facial movements. We estimate the mean variance of the traces for the entire set of datapoints to estimate a threshold to dichotomize the segments into binary classes for the pre-text task. If the variance of the trace of a sequence is above the threshold, the segment is labeled as "1." Otherwise, the segment is labeled as "0." This approach is separately done for the lower and upper facial regions creating two binary problems.

3.3. Task of Action Units

The third pre-task task is a multimodal multi-class problem involving acoustic and facial features. This pre-text task makes an explicit attempt to create an emotional discriminative feature representation by predicting emotional classes, where the labels are automatically obtained by detecting the co-occurrence of facial *action units* (AUs). AUs describe specific facial muscle movements. They are the basic coding units of the *facial action coding system* (FACS). We rely on the *Emotional Facial Action Coding System* (EMFACS) [26], which was created to infer emotional classes from the AUs present in the face. Table 1 shows the combinations of AUs provided by the EMFACS that are often observed for four major emotional states: happiness, sadness, surprise, and anger. We obtain the detected AUs (true or false) for each frame with the OpenFace 2.0 toolkit [25].

Table 1: *Set of AUs needed to define four emotional classes obtained from EMFACS.*

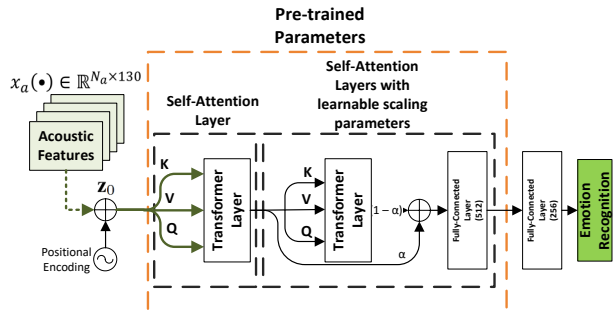| Emotion	Top AUs	Bottom Aus
Happiness	AU6	AU12
Sadness	AU1+AU4	AU15
Surprise	AU1+AU2+AU5	AU26
Anger	AU4+AU5+AU7	AU23

Figure 2: *Overview of the architecture used to perform speech emotion recognition.*

Then, we estimate the proportion of frames that the AUs are detected in each video segment. We create two separate AU rankings, sorting the top (AUs 1-9) and bottom (AUs 10-26) AUs according to their detection proportion. We compare the two groups of AUs obtained from each video sequence with the AU combinations listed in Table 1. An emotional label is given to any sequence that has a match of at least one AU present in the top AU list and one AU present in the bottom AU list, while also matching the AUs emotion combinations seen in Table 1. Sequences that might match more than one emotion are decided based on which combination of AUs were detected more times across the entire sequence. If there is not a combination that closely relates to happiness, sadness, surprise, or anger, we label the segment as "other," creating a five-class problem.

3.4. Speech Emotion Recognition Framework

The proposed pre-text tasks are general and can be implemented with different *deep neural networks* (DNNs). This study relies on an attention-based framework to build our SER architecture. Recent studies using attention models [27–29] have shown great performance in emotion recognition. Recent studies have showed that attention models work well with multimodal emotion recognition as well [20, 21]. Although attention-based architectures work well with acoustic data in supervised settings, this study explores potential improvements using SSL.

Figure 2 shows an overview of the architecture used in this study. The framework consists of a transformer-based architecture with self-attention layers to learn representations within the input audio sequences. The framework receives the audio feature matrix ($x_a \in \mathbb{R}^{N_a \times 130}$, see Sec. 4.2). Then, we use the method from Vaswani et al. [30] and add a 1-D positional embedding to the input feature vectors (Eq. 1). The positional embedding is important for the model to retain temporal information from the input sequences, since the model accepts input sequences in parallel. The inputs are then entered into self-attention layers to compute within sequence representations from the acoustic features. After the first set of self-attention layers, we add a set of self-attention layers that are equipped with residual connections, which have layer-independent learnable scaling weights α and β. This self attention layer has, for the most part, the same structure of the previous self-attention layers. The Q, V, K vectors are contained within their own self-attention layers. The self-attention layers compute representations within the same sequence. The use of self-attention with learnable scaling parameter module after the first attention layers helps our model scales areas of the representations extracted from the previous layers to give more emphasis to areas that are more useful for the final task. The self-attention representations are then extracted and passed through two fully-connected layers to generate the predictions.

$$\mathbf{z}_0 = [x^1\mathbf{E}; x^2\mathbf{E}; \cdots ; x^{N_a}\mathbf{E}] + [\mathbf{E}_{pos} \in \mathbb{R}^{N_a \times 130}] \quad (1)$$

4. EXPERIMENTAL SETTINGS

4.1. Resources

This study uses the CREMA-D [14] and MSP-Face [15] corpora. CREMA-D is a crowd-sourced audiovisual dataset, which contains videos of subjects saying sentences while expressing pre-defined emotions (happiness, fear, disgust, anger, sadness and neutral state). This corpus was collected from an ethnically and racially diverse group consisting of 91 actors (48 male and 43 female). The videos were annotated with emotional labels by seven raters after watching the videos. In total, the CREMA-D corpus contains 7,442 videos with the following distribution: 1,230 happy clips, 1,180 fear clips, 1,222 disgust clips, 1,067 angry clips, 672 sad clips, and 2,071 neutral clips. We only use the speech signal from this corpus, focusing on a SER task consisting of predicting the six emotional states included in the corpus. We generate random splits in a speaker-independent manner using 70% of the data for the train set, 15% of the data for the development set, and 15% of the data for the test set.

The MSP-Face [15] is a natural audiovisual emotional dataset collected in-the-wild from recording obtained from video-sharing websites. The corpus includes 491 individuals expressing their opinions about certain topics or sharing their experiences. The individuals in this dataset express a wide range of rich emotional behaviors in their videos. In total, the dataset contains 70.7 hours of audiovisual data, where 24.7 hours are labeled and 46 hours are unlabeled. The MSP-Face corpus is used for pre-training our framework. We do not make use of the labels available for this corpus, considering the entire dataset as our unlabeled set. As we described in Section 3, we use two second windows with 0.5 seconds overalp. We use data from all the 491 subjects present in the MSP-Face corpus [15]. In total, we use 109,656 datapoints (60.92 hrs).

4.2. Acoustic Features

For training our SER model, we use the *OpenSmile* toolkit [31] to extract the *low level descriptors* (LLDs) that were proposed for the paralinguistic challenge in Interspeech 2013 [32]. This set consists of 65 acoustic frame-based features, which include the fundamental frequency, energy, and *Mel-frequency cepstral coefficients* (MFCCs). The feature set also includes the respective first order derivatives of these LLDs, generating a 130 dimensional feature vector. The LLDs are directly extracted from the raw audio for each video clip, which are pre-processed with a window length of 32 ms and a step size of 16 ms. The extracted LLDs are subsequently z-normalized before sequentially concatenating row-wise into a $N_a \times 130$ matrix, where N_a is the number of 32ms-segments. This matrix is used as an input to our network.

4.3. Implementation Details

The multi-head attention layers and self-attention layers are implemented with five layers, where each layer has 10 attention

Table 2: *Comparison of SSL pre-trained and fine-tuned models with the supervised learning method. The table reports the average F1-Score values across 20 experiments using different random seeds (✶ indicates that our model is significantly better than the other two methods)*

CREMA-D						
	Macro			**Micro**		
Architecture	Prec.	Rec.	F1	Prec.	Rec.	F1
Supervised	54.8	53.4	51.9	53.1	53.1	53.1
wav2vec 2.0 (retrain)	55.3	54.5	53.3	54.3	54.3	54.3
Our Method✶	**59.7**	**57.9**	**57.0**	**57.7**	**57.7**	**57.7**

Table 3: *Performance comparison obtained from ablations performed with five different combinations of pre-text tasks used during pre-training.*

CREMA-D									
Architecture				**Macro**			**Micro**		
	T1	T2	T3	Prec.	Rec.	F1	Prec.	Rec.	F1
SSL 1	✓	✓	✓	**59.7**	**57.9**	**57.0**	**57.7**	**57.7**	**57.7**
SSL 2	✓	~✓	✓	59.0	57.2	56.4	57.0	57.0	57.0
SSL 3	✓	✓		53.0	53.4	51.5	53.3	53.3	53.3
SSL 4		✓	✓	56.7	55.0	53.6	54.8	54.8	54.8
SSL 5	✓			33.6	43.2	35.8	42.8	42.8	42.8

heads. For the supervised learning baseline, we set a dropout rate of 0.25 at the output embeddings obtained from the attention layers, along with a 0.1 dropout rate for the residual connections. The model uses *adaptive moment estimation* (ADAM) as the optimizer with an initial learning rate set to 0.000725. We set a five epoch-patience learning decay with a 0.1 factor. We use a batch size of 32, implementing the activrfunction with *rectified linear units* (ReLUs). The model is trained for 20 epochs with an early stopping criterion based on the development loss.

The SSL pre-training method has mostly the same setting as the supervised learning training, except the model is pre-trained for 300 epochs with the learning rate set to 0.0001. The loss used for the binary tasks (speech energy and facial landmarks) is the binary cross-entropy loss and the loss used for the multi-class problem (AU task) is the cross-entropy loss. The losses obtained from each task are summed during training. We use ReLU as the activation function and ADAM as the optimizer. After pre-training on the pre-text tasks, the model is fine-tuned on the CREMA-D acoustic data using the same setting as the supervised learning. The model at this step is only trained for five epochs. Everything is implemented in PyTorch and trained using a Nvidia QUADRO RTX 8000.

We compare our SSL approach with two baselines: Baseline 1 consists of training the framework shown in Figure 2 with a supervised setting, without the pre-training the model. Baseline 2 consists of using the wav2vec 2.0 [16] method for pre-training the transformer architecture used in this study. For fair comparison, our implementation of wav2vec 2.0 [16] is pre-trained from scratch using the raw audio from the recordings of the MSP-Face [15] data used to pre-train our proposed model. Another adaptation is that the transformer layers present in the original wav2vec 2.0 model are modified to match the same transformer architecture we used in this study. Our implementation of wav2vec 2.0 is pre-trained for 400k iterations and fine-tuned for 1,200 iterations.

5. EXPERIMENTAL RESULTS

This section compares the performances of our proposed model with results of baselines. Table 2 lists the results. Each model is trained 20 times with different speaker-independent train, development, and test partitions every time. The models' performances are evaluated using both the macro-averaged and micro-averaged precision, recall, and F1 scores. Table 2 reports the averaged results across the 20 experiments. We compare the results using a one-tailed matched pair t-test over the 20 results with p-value <0.05 to assert statistical significance.

Table 2 shows that using the pre-text tasks proposed in this study generates meaningful representations which significantly improves the performance over the purely supervised SER method. Our method achieves increased SER performance by up to 5.1% on precision, up to 4.5% on recall, and up to

4.9% on F1 scores. The results also show improved performance over the model using the wav2vec 2.0 [16] framework pre-trained with the same dataset. This result shows that the proposed domain-specific pre-text tasks for SER system lead to better performance than using the general purpose pre-text tasks in wav2vec 2.0 when we pre-train with similar data.

5.1. Pre-text Tasks Ablation

In this section, we study the effect each pre-text task has on our proposed method's performance. For this ablation study we generate five different combinations of pre-text tasks to pre-train the architecture used in this study. The five combinations are: SSL1 is our standard method which uses all tasks described in Section 3; SSL2 consists of using the energy task, action units task, and a modified version of the facial landmarks task where instead of retrieving two centroids for each frame, we only generate a single centroid for each face frame; SSL3 uses combination of the energy task and the facial landmarks task; SSL4 uses the facial landmarks task and the action units task; and, SSL5 uses only the energy task for pre-training the model.

As seen in Table 3, the model's performance gradually drops as less tasks or different combinations are used during pre-training. There's a major decline in performance specially when we only use the energy task for pre-training. This result shows the importance of having the face-related tasks when training our model. It is also important to mention that the energy task helps. Although the energy task alone does not result in performance improvement, the performance obtained from SSL4, which only uses face-related tasks, is lower than the performance of SSL1, showing the need for this pre-text task.

6. Conclusions

This study presented new unimodal (speech task) and multimodal (facial landmark and action unit tasks) pre-text tasks for SSL carefully created to generate meaningful representations for SER. The pre-text task leverages the complementary relationship between speech and facial cues. With only 60.92 hours of unlabeled data and using acoustic features, the proposed pre-text tasks significantly enhance the representations obtained by a model that would otherwise be trained under a supervised regime. This study showed the benefits of creating multimodal pre-text tasks even when the focus of the study is solely on a speech-based system. A future research direction for this study is to expand the architecture to include more modalities and more domain specific pre-text tasks that are relevant for SER. We will also explore implementing this approach with other emotional audiovisual databases with a wider range of speakers such as the CMU-MOSEI [33] and SEWA [34] corpora.

7. Acknowledgements

8. References

[1] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[2] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

[3] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[4] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. Early Access, 2022.

[5] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. Calgary, AB, Canada: IEEE, April 2018, pp. 5084–5088.

[6] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186.

[8] L. Lin, S. Song, W. Yang, and J. Liu, "MS2L: Multi-task self-supervised learning for skeleton based action recognition," in *ACM International Conference on Multimedia (MM 2020)*, Seattle, WA, USA, October 2020, pp. 2490–2498.

[9] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *IEEE Spoken Language Technology Workshop (SLT 2021)*, Shenzhen, China, January 2021, pp. 381–388.

[10] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 3755–3759.

[11] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *IEEE Spoken Language Technology Workshop (SLT 2021)*, Shenzhen, China, January 2021, pp. 373–380.

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 3465–3469.

[13] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[14] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, October-December 2014.

[15] A. Vidal, A. Salman, W.-C. Lin, and C. Busso, "MSP-face corpus: A natural audiovisual emotional database," in *ACM International Conference on Multimodal Interaction (ICMI 2020)*, Utrecht, The Netherlands, October 2020, pp. 397–405.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, Virtual, Dec. 2020, pp. 12 449–12 460.

[17] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using Wav2vec 2.0 embeddings," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 3400–3404.

[18] A. Keesing, Y. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 3415–3419.

[19] S. Y. Tseng, S. Narayanan, and P. Georgiou, "Multimodal embeddings from language models for emotion recognition in the wild," *IEEE Signal Processing Letters*, vol. 28, pp. 608–612, Mar. 2021.

[20] Y.-H. Tsai, S. Bai, P. Liang, J. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Association for Computational Linguistics (ACL 2019)*, vol. 1, Florence, Italy, July 2019, pp. 6558–6569.

[21] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *International Conference on Multimodal Interaction (ICMI 2020)*, Utrecht, The Netherlands, October 2020, pp. 400–404.

[22] S. Tomar, "Converting video formats with FFmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, June 2006.

[23] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Python in Science Conference (SciPy 2015)*, Austin, TX, USA, July 2015, pp. 18–25.

[24] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 549–556.

[25] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *IEEE Conference on Automatic Face and Gesture Recognition (FG 2018)*, Xi'an, China, May 2018, pp. 59–66.

[26] W. V. Friesen, P. Ekman *et al.*, "Emfacs-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, no. 36, p. 1, 1983.

[27] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2227–2231.

[28] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Annual Workshop on Audio/Visual Emotion Challenge (AVEC 2017)*, Mountain View, California, USA, Oct. 2017, pp. 19–26.

[29] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7357–7361.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[32] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[33] A. Zadeh, P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *ACM Association for Computational Linguistics (ACL 2004)*, vol. 1, Melbourne, Australia, July 2018, pp. 2236–2246.

[34] J. Kossaifi *et al.*, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, March 2021.