Harnessing Multimodal Unlabeled Data for Enhanced Speech Emotion Recognition

Lucas Goncalves, *Student Member, IEEE,* and Carlos Busso, *Fellow, IEEE*

***Abstract*—*Speech emotion recognition* (SER) often faces challenges due to the lack of large, annotated datasets. The presence of abundance unlabeled data offers a chance to explore methods that could significantly improve SER systems. This study explores the feasibility of enhancing general speech models by incorporating unimodal and multimodal training objectives derived from unlabeled data, specifically tailored to extract emotional content. These multimodal objectives aim to refine *self-supervised learning* (SSL)-based representations that, while effective in SER, were not originally created to extract emotional cues from speech. Our methodology introduces a set of multimodal objectives focused on capturing information from three primary sources: acoustic signals, through a representation objective based on the extended *Geneva Minimalistic Acoustic Parameter Set* (eGEMAPS); facial expressions, via visual representations obtained from a pre-trained facial expression recognition system; and textual content, through pseudo-labels generated by a pre-trained emotion sentiment model. These objectives are automatically generated from 70.7 hours of unlabeled emotional content captured in naturalistic settings.We apply our strategy to four state-of-the-art SSL-based speech models, aiming to enhance their capabilities in SER tasks with multimodal signals while still keeping inference strictly audio-only. Our experimental evaluations across the CREMA-D, MSP-IMPROV, and MSP-Podcast datasets demonstrate that our approach significantly improves SER performance, especially in settings with limited labeled data.**

***Index Terms*—speech emotion recognition (SER), unlabeled multimodal datasets, unsupervised training, pre-trained speech models**

I. Introduction

HUMAN communication utilizes a multitude of signals to accurately convey an intended message, clarifying its meaning and intention, highlighting its emphasis, and conveying its emotions. In daily interactions, we draw from a mix of auditory, visual, and other sensory cues to produce and perceive emotions [1]. Recognizing these signals can be instrumental in developing precise emotion recognition models, even when the goal is to build a *speech emotion recognition system* (SER). While traditional SER systems predominantly focus on acoustic features [2]–[7], this study explores the complementary information present in speech cues, facial expressions, and textual information to refine SER models that are already pre-trained. Spoken language, facial expressions, and written text provide complementary emotional information that can be leveraged in building speech representations. By tapping into the complementary information across these modalities, we aim to further optimize our SER models. The primary objective of this paper is to investigate how multimodal unlabeled data, combined with the strength of pre-trained models, can elevate the performance of SER models. This approach is especially promising for enhancing supervised learning techniques for SER when labeled data is scarce.

The motivation behind this paper is to explore the use of multimodal unlabeled data to craft speech emotional representations, aiming to improve the performance of state-of-the-art pre-trained speech models such as Wav2vec 2.0 [8], WavLM [9], HuBERT [10], and Data2vec [11]. The advances in *self-supervised learning* (SSL) have demonstrated the capability to generalize across numerous downstream tasks. Additionally, studies have shown that fine-tuning [12], [13] or using speech representations [14] from such models has proven effective for achieving good SER performance. Within the SSL framework, models are predominantly pre-trained on pretext tasks, where labels are auto-generated. A strategy often used in this area is the *masked language modeling* (MLM) task. Here, certain tokens are replaced with either the $< \text{MASK} >$ token or a random substitute, tasking the model with predicting these masked tokens [15], [16]. Following this strategy, many models, such as Wav2vec [17], have adopted contrastive objectives to learn speech representations. Specifically, Wav2vec is fine-tuned to predict the subsequent time-step, guided by a contrastive loss. Historically, a few studies have explored the use of modalities outside of speech to enhance SER. Examples include using pre-trained *language models* (LM) to harness the rich information learned from training on large datasets for text sentiment analysis [18], or integrating audio-visual tasks to train models to be used for SER [19]. Given the rich nature of human communication, we hypothesize the benefits of leveraging multimodal data, even when the goal is to build a speech-based emotion recognition system.

This paper proposes to integrate a multi-task objective involving acoustic, visual, and textual features for the pre-training of speech models to improve their SER capabilities. Our fine-tuning strategy consists of predicting feature representations across modalities, where the input is just acoustic features (e.g., the prediction of feature representations for facial expressions using speech). Given that these representations are originally trained to recognize emotional information, we obtain SER models after fine-tuning that are highly discriminative in recognizing emotions from speech. For the speech modality, we incorporate hand-crafted features tailored for emotion recognition [20]. We use these hand-crafted features, designed to capture emotional nuances, to fine-tune our models to be more adept at recognizing emotions

L. Goncalves is with the Erik Jonsson School of Engineering & Commputer Science, The University of Texas at Dallas, Richardson TX 75080.
E-mail: goncalves@utdallas.edu

C. Busso is with the Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213.
E-mail: busso@cmu.edu

Manuscript received June xx, 2025; revised XX XX, 20xx.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2025.3647876

in speech. For the visual and textual modalities, we extract features from two distinct pre-trained models that specialize in visual [21] and textual [22] representations, leveraging the advances and robustness achieved by these models in their respective fields. As these models are trained on diverse datasets, they capture a wide variety of scenarios, nuances, and patterns within their modality representations. It is important to note that our model remains unimodal at inference time, using only speech as input. While we leverage multimodal supervision during pre-training, we do not perform multimodal fusion nor require text or visual input at test time. This feature distinguishes our approach from standard multimodal systems that rely on multi-stream fusion methods during inference. Our strategy demonstrates that multimodal information is useful during training, even when the task is unimodal.

We evaluate our proposed method on three emotion recognition benchmarks: the MSP-Podcast [23], [24], the CREMA-D [25], and the MSP-IMPROV [26] corpora. For pre-training, we extract audio, visual, and textual features from the MSP-Face corpus [27] and utilize these features to pre-train speech models including Wav2vec 2.0 [8], WavLM [9], HuBERT [10], and Data2vec [11]. Our experimental evaluation compares two approaches: one where we pre-train these models using our method and then fine-tune them for the target task, and another where we directly fine-tune the models for the target task without the multimodal pre-training phase. Our findings highlight enhanced performance when leveraging a model pre-trained with multimodal representations during the initial phase, as opposed to methods that only perform the fine-tuning step. Furthermore, we ablate the influence of individual modalities on the performance of our proposed method, specifically focusing on the Wav2vec 2.0 model [8]. Our observations underscore the significance of incorporating all the modality tasks during the pre-training phase to ensure strong results.

The rest of this paper is organized as follows: Section II reviews previous studies that are relevant to this work. Section III describes the proposed approach introduced in this study, the experimental settings for the implementation of the proposed approach, and the description of the corpora used to train and test the models. Section IV presents our experimental evaluations, contrasting the results of the proposed approach with strong baselines and evaluating the proposed approach in low-resource settings. In Section V, we present an ablation study on the proposed approach. Lastly, Section VI summarizes the contributions of this work, highlighting our experiments and discussing possible future research directions.

II. Related Work

Effective modeling of representations from speech signals is crucial for SER tasks. SER problems have been formulated under different settings: supervised learning [28], semi-supervised learning [29], and unsupervised settings [30]. Over the years, SER has evolved through many training setups. Early studies relied on handcrafted features to obtain representations from acoustic inputs [20], [31]–[34]. With the advancements in SSL speech models [8]–[11], studies have underscored the utility of leveraging these pre-trained encoders as foundational elements for SER [12], [13], [35], offering superior performance over conventional approaches [15], [36], [37].

A. Self-Supervised Learning

With the advances of *self-supervised learning* (SSL), several notable methods have emerged [38], including Wav2vec 2.0 [8], WavLM [9], HuBERT [10], and Data2vec [11]. In SSL, models employ techniques that leverage information extracted from the input data itself as labels to learn speech representations, which are beneficial for downstream tasks. These models typically use *masked language modeling* (MLM) tasks as training objectives. This strategy involves replacing some tokens with either the token $< \mathrm{MASK} >$ or a random token, and then prompting the model to predict the masked tokens [15], [16]. Additionally, these models have incorporated contrastive objectives to extract speech representations, exemplified by Wav2vec [17], which is optimized for solving a next-time-step prediction task using a contrastive loss. Wav2vec 2.0 [8] merges contrastive learning with masking, similar to the *contrastive predictive coding* (CPC) model [39], using InfoNCE loss to enhance the similarity between different types of representations.

In contrast, HuBERT [10] uses quantized acoustic features, learned via k-means. This strategy discretizes the speech inputs, focusing on learning continuous latent representations for unmasked timesteps and capturing long-range temporal dependencies for the masked prediction. WavLM [9] builds on HuBERT, adding a gated relative position bias to the transformer self-attention mechanism and an utterance mixing strategy for training, which involves combining signals from different speakers to improve the model's ability to distinguish overlapping speech. Data2vec [11] creates targets Y using an *exponential moving average* (EMA) of its parameters, inspired by its application in self-supervised visual learning [40], [41]. It averages hidden representations from the top k layers of the EMA teacher network for unmasked inputs.

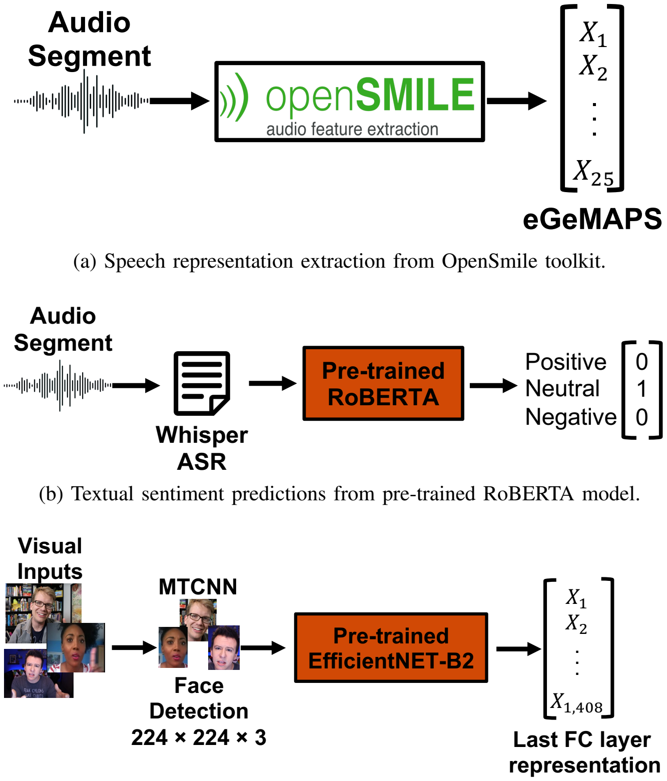
While studies have shown improved performance for SER tasks [12], [13], [42], [43], it is important to note that these SSL models were not designed for SER. Their primary purpose was to create general speech processing frameworks. We hypothesize that adapting these SSL models with emotion-related multimodal objectives can lead to more emotionally discriminative speech representations.

B. Multimodal Learning to Enhance SER

Multimodal learning has been extensively explored in the field of emotion recognition, with various studies investigating methods to combine different modalities with speech. These include audio-text combinations [44]–[47], audio-visual combinations [48]–[54], and the integration of all three modalities [55]–[57]. In addition, studies have explored several methods to combine these modalities, such as hierarchical learning integration [58], cross-modal integration [59]–[61], and decision-level integration [62], [63].

Further extending the scope from multimodal learning, recent studies have explored the use of multimodal cues to enhance speech representations for unimodal systems [18], [19], [64]. Shon et al. [18] investigated the application of a BERT-based system [65] to generate pseudo labels for a dataset (i.e., positive, neutral, negative sentiment). These labels were then utilized to train speech sentiment analysis models in a semi-supervised manner. Srinivasan et al. [44] focused on enhancing contextual speech features by incorporating text modality. They employed joint multimodal embeddings to improve the audio-only representation for emotion recognition, utilizing a teacher-student framework. Shukla et al. [64] investigated the use of visual self-supervision for learning audio features. Their method involves visually guided self-supervised learning of speech representations through facial reconstruction. In our preliminary study [19], we proposed three emotionally related tasks for pre-training a speech model for emotion recognition. The proposed approach introduces two facial-related tasks and one speech task for pretraining the model to leverage the complementary relationship existing between speech and visual cues.

In contrast to previous methods, our approach seeks to leverage pre-trained SSL speech models as a foundation. We aim to enrich these models by introducing emotion-related textual, visual, and acoustic objectives during a pre-training adaptation phase that are specifically designed to create discriminative emotional speech representations. This step is designed to enhance the models' speech representations specifically for SER. The proposed method utilizes unlabeled data to extract multimodal representations. These representations are then employed in a multi-task training objective, which is tailored to improve the accuracy of SER.



(a) Speech representation extraction from OpenSmile toolkit.

(b) Textual sentiment predictions from pre-trained RoBERTA model.

(c) Facial representations extraction using EfficientNet-B2.

Fig. 1. Overview of objective generation for proposed approach.

III. Proposed Approach

As shown in Figure 1, in this study, we employ acoustic, visual, and textual cues from an unlabeled pool of emotionally rich multimodal recordings to improve SER performance in scenarios with limited speech data with emotional labels. Our approach introduces a cross-modal multitask pre-training adaptation step to prepare speech models for downstream fine-tuning on speech emotion recognition. Figure 2 describes the two-stage approach. The first stage takes an SSL off-the-shelf model and pre-trains it with the proposed multimodal objectives that are carefully designed for emotion recognition tasks. The second stage takes the pre-trained model and fine-tunes it using limited data with emotional labels.

We generate training objectives by retrieving emotion-related cues from three sources: speech, text, and face, as depicted in Figure 1. In the following subsections, we describe the process used to retrieve each of these cues. We have conducted extensive experiments with a pre-determined set of multimodal objectives obtained using different models or techniques to enhance the performance of SER systems. This section reports the proposed approach that led to the best performance. Section V discusses alternative implementations and combinations of objectives that we also explored.

A. Speech Representations

For our first objective (Fig. 1a), we use SSL as input to predict acoustic features that have been shown to convey expressive cues in speech. We formulate this objective as a regression problem. We utilize the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [20] as our target, where the goal is to minimize the *mean-squared error* (MSE) (Eq. 1) between the model's outputs and the normalized utterance-wise eGeMAPS features extracted from the speech signal. The eGeMAPS feature set is widely recognized for its ability to effectively represent and analyze emotional states and has been proven to contain relevant information for SER tasks [2]. Its feature set is a comprehensive and standardized set of acoustic features designed for voice research and affective computing. It was developed to provide a minimal set of robust and universally applicable parameters, capturing relevant vocal characteristics across various emotional states and speaking styles. By offering a harmonized and scientifically grounded selection of features, the eGeMAPS set facilitates comparability and reproducibility in voice research, making it a valuable tool for studies in emotion recognition, clinical diagnostics, and human-computer interaction.

The set was extracted using the OpenSmile toolkit [31] with the configuration eGeMAPSv02, which comprises 25 *low-level descriptors* (LLDs), including frequency-related measures, prosody parameters, and temporal aspects of speech. The final parameters are obtained using functionals over each of the 25 LLDs, creating a 25-D vector for each speech sentence regardless of its duration. The features were extracted using a window length of 32ms with a step size of 16ms. We refer to this objective in the paper as O1.

B. Textual Representations

For our second objective (Fig. 1b), we rely on an objective obtained from text. The transcription is assumed to be unknown in the unlabeled set. Therefore, we use the ASR model *Whisper* [66] to obtain automatic transcriptions from each sentence in the unlabeled set. Our task is to use SSL as input to predict pseudo-labels generated by a text-based sentiment prediction tool. We employ the pre-trained multilingual XLM-RoBERTa-base model [22], which has been trained on approximately 198 million tweets and fine-tuned for sentiment analysis. This model serves to extract emotional sentiment predictions from the retrieved textual content derived from audio in our investigated unlabeled dataset. Prior research has validated the efficacy of pre-trained language models in extracting sentiment information from written texts in unlabeled speech datasets and utilizing the extracted sentiment data as pseudo-labels within a semi-supervised training framework for speech sentiment recognition [18]. The selected text-based model outputs three sentiment classes: positive, neutral, and negative. We transform these predictions into a one-hot vector, using speech features to recognize these classes. Our objective is then to train the model to predict the pseudo-labels and minimize the *cross-entropy* (CE) loss (Eq. 2). We refer to this objective in the paper as O2.

C. Visual Representations

For our third objective (Fig. 1c), we rely on a visual representation obtained from a facial expression recognition system, where the task is to predict the facial feature representation from speech using the SSL as input. We extract facial feature representations using a pre-trained EfficientNet-B2 model [21]. Facial expressions, stemming from muscle activity, are prominently accentuated during speech articulation, intertwining with emotional expressions [67], [68]. As a result, it is not surprising that facial expressions and speech are intrinsically connected [69]. We employ the *multi-task cascaded convolutional neural network* (MTCNN) face detection algorithm [70] to capture these visual features. This algorithm facilitates the extraction of facial images from each frame of the corpus using bounding boxes. Following the facial image extraction step, the images are resized to a uniform dimension of $224 \times 224 \times 3$. We extract emotional feature representations using the pre-trained EfficientNet-B2 model [21], known for its superior performance on the AffectNet corpus [71]. The target facial feature representation to be predicted from speech is obtained from the last fully connected layer before classification in the EfficientNet-B2 model, which is a 1,408-D vector. Frame-wise representations are extracted and averaged to derive an overall utterance-level feature representation. Similar to the first objective, this task also involves training our model to minimize the *mean-squared error* (MSE) (Eq. 1), aligning the model's outputs with the normalized utterance-wise mean of the EfficientNet-B2 model representations, derived from visual inputs. We refer to this objective in the paper as O3.

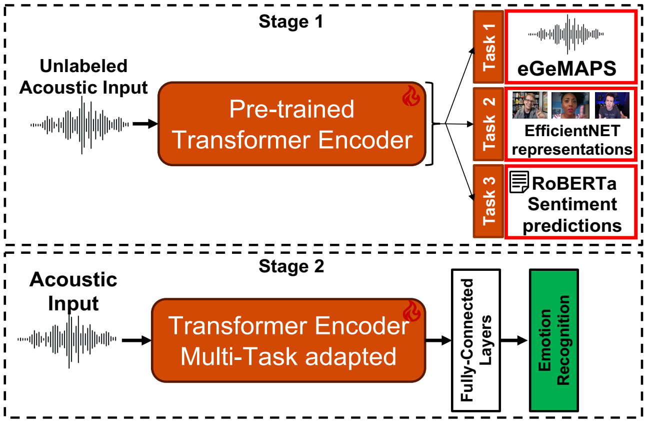

Stage 1
Unlabeled Acoustic Input
Pre-trained Transformer Encoder
Task 1
eGeMAPS
Task 2
EfficientNET representations
Task 3
RoBERTa Sentiment predictions
Stage 2
Acoustic Input
Transformer Encoder Multi-Task adapted
Fully-Connected Layers
Emotion Recognition

Fig. 2. Proposed approach overview. Stage 1 multitask adaptive pre-training consist on training the model to predict on speech, textual, and visual objectives. Stage 2 consisting on downstream finetuning.

D. Training Pipeline

As seen in Figure 2, our training method consists of a two-stage strategy. In stage 1, we perform the proposed adaptive unlabeled pre-training process of the model using a multi-task learning setup using the emotionally-related representations obtained from the three aforementioned modalities. Afterwards, in stage 2, we fine-tune all the parameters of the model on the desired downstream task of SER.

1) Stage 1: In stage 1, we pre-train the transformer layers of the models investigated in this study using a multi-task setup designed to optimize the models based on the textual, visual, and acoustic representations extracted from an emotionally rich audio-visual corpus used as unlabeled data. In our approach, we utilized the MSP-Face corpus [27], which contains close to 70 hours of emotional content from naturalistic recordings obtained from a video-sharing website (Sec. III-E describes the data). The speech representation objective and the facial representation objective are optimized using the MSE loss as shown in Equation 1:

$$\mathcal{L}_{MSE} = \frac{1}{n} \sum_{i=1}^{n} ||y_i - \hat{y}_i||^2 \quad (1)$$

where n is the total number of observations, $\hat{y}_i$ is the predicted value for the i-th data point, and y_i represents the actual value of the i-th data point. For the speech objective, y_i is the extracted 25D vector obtained from the eGEMAPS for the i-th input. For the visual objective, y_i is the 1,408D vector obtained from the EfficientNet-B2 model representations for the i-th input. The term $||y_i - \hat{y}_i||^2$ calculates the squared difference between the actual and predicted values for each data point.

The textual objective consists of a three-class classification task based on the pseudo-labels obtained with the text-based sentiment classifier. The loss function is optimized using the CE loss as shown in Equation 2:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{n} \sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic}) \quad (2)$$

where n is the total number of observations, C represents the number of classes, y_{ic} is a binary indicator (0 or 1) if class label c is the correct classification for observation i, and $\hat{y}_{ic}$ is the predicted probability of observation i belonging to class c. The inner sum $\sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic})$ computes the log loss for each class for a single observation and sums it across all classes. The outer sum then accumulates this loss across all observations.

The overall loss of the multi-task pre-training step consists of a combination of the losses from the three proposed objectives, as shown in Equation 3:

$$\mathcal{L} = \mathcal{L}_{MSE_{O1}} + \lambda * \mathcal{L}_{CE_{O2}} + \mathcal{L}_{MSE_{O3}} \tag{3}$$

where $\mathcal{L}_{MSE_{O1}}$ is the MSE loss from the first objective (speech), $\mathcal{L}_{CE_{O2}}$ is the CE loss from the second objective (text), $\mathcal{L}_{MSE_{O3}}$ is the MSE loss from the third objective (face), and λ is a weighting factor hyper-parameter used for losses combination.

The text objective (O2) uses the *cross-entropy* (CE) loss, which operates on a different scale compared to the MSE losses used for the speech (O1) and face (O3) objectives. To prevent the CE loss from dominating the total loss, we introduce a weighting factor λ to balance its contribution. This factor ensures that all three objectives make meaningful contributions to the optimization process. We adopt a single scalar to prevent the CE term from overwhelming the regression objectives (different natural scales) and to keep training simple and reproducible across backbones/datasets. Dynamic schemes for loss balancing, such as uncertainty-based weighting [72] or GradNorm [73] are also viable directions for future exploration. The effect of different values of λ is explored in the ablation study in Section V. Stage 1 is trained for 100 epochs until it reaches convergence on a held-out set of 15% of the unlabeled data used for the adaptive pre-training.

As described in this section, our proposed approach relies on pseudo-labels or representations derived from pre-trained models, which may have intrinsic bias. To mitigate potential cascading errors or overfitting to biases in the pseudo-labeling sources, we treat these targets as auxiliary objectives in a multi-task setting. The speech model is not trained to exactly reproduce these external predictions, but to extract speech representations that correlate with them. This soft guidance allows the model to generalize without directly inheriting potential artifacts or discrete errors from the original models. We also prevent inheriting the potential bias observed in individual modalities by jointly using text, facial, and acoustic pre-training tasks. We analyze the bias across emotional predictions before and after the proposed pre-training approach in Section IV-E.

2) Stage 2: Stage 2 consists of taking the pre-trained model from stage 1 and further fine-tuning all of its transformer layers for our emotion recognition downstream task using the corpora explored in this study (the audio 1D CNN encoder layers are kept frozen). This study formulates the SER task as a classification problem, where the speech sentence is classified into a set of emotional categories. Therefore, the models are fine-tuned using the CE loss (Eq. 2) as the training objective.

E. Corpora

In this study, we utilize four prominent corpora: the MSP-Podcast [23], [24], the CREMA-D [25], the MSP-IMPROV [26], and the MSP-Face [27] corpora.

1) MSP-Podcast Corpus: The MSP-Podcast corpus [23], [24] contains spontaneous and diverse emotional speech samples collected from various podcast recordings. This paper uses version 1.11 of the corpus containing 151,654 speaking turns, which is 237 hours and 56 minutes. The train set includes 84,030 segments, and the development set includes 19,815 segments. The corpus has three test sets. In our experiments, we utilize "Test1", which contains 30,647 utterances. These sets aim to create speaker-independent partitions. Each sentence was annotated by at least five annotators using a crowdsourcing protocol adapted from the study of Burmania et al. [74]. This study focuses on four emotional categories from this corpus: anger, sadness, happiness, and neutral state.

2) CREMA-D Corpus: The CREMA-D corpus is an audiovisual dataset with high-quality recordings from a diverse group of 91 actors (48 male and 43 female), representing various racial and ethnic backgrounds. These actors were instructed to articulate a set of sentences, each time aiming for a specific emotional class. The recordings utilized a green screen as a backdrop. The data acquisition was facilitated by two directors: 51 actors collaborated with the first director, while the remaining 40 paired with the second director. The emotional categorizations of these clips were assessed by at least seven annotators across three distinct settings: solely audio, solely video, and combined audiovisual. The corpus consists of 7,442 clips, evaluated by 2,443 individual raters. Our study uses the labels provided after evaluating the audiovisual stimuli. The distribution of the consensus labels is: 1,067 clips for anger (10,054 ratings), 1,222 for disgust (11,429 ratings), 1,180 for fear (11,153 ratings), 1,230 for happiness (11,730 ratings), 672 for sadness (6,347 ratings), and 2,071 for a neutral state (19,450 ratings). We formulate the SER problem for the CREMA-D corpus as a six-class task: anger, disgust, fear, happiness, sadness, and neutral state.

3) MSP-IMPROV Corpus: The MSP-IMPROV corpus [26] serves as the second audiovisual resource that we consider. This corpus was devised to probe the nuances of emotional perception [75]. A unique aspect of this corpus was the need to maintain consistent lexical content for target sentences, spoken while the speaker conveyed different emotions. Instead of merely instructing actors to recite sentences with varied emotions, a protocol was employed to create authentic emotional renditions by designing hypothetical two-person scenarios that would prompt a participant to utter the target sentence with the desired emotion. With 20 target sentences spanning four emotional states (happiness, sadness, anger, and neutrality), 80 unique scenarios were produced. This segment of the corpus has 652 speech instances. Additionally, the corpus includes all other interactions leading up to the target utterance (4,381 spontaneous speech turns) and natural interactions between the actors during the breaks of the improvisations (2,785 natural speech turns). It further consists of read renditions of the target sentences in the four emotions (620 rehearsed speech instances). In total, the MSP-IMPROV corpus comprises 7,818 spontaneous speech instances and 620 read phrases. The annotation was conducted through a crowdsourcing protocol, which meticulously monitored worker quality in real-time, stopping evaluations if their performance fell below a threshold [74]. A minimum of five workers annotated each sentence, and the consensus labels were established based on the plurality rule. In this research, we utilized four emotional categories from this corpus: anger, sadness, happiness, and neutral state.

4) MSP-Face Corpus: The MSP-Face dataset [27] is an audiovisual emotional dataset, sourced in-the-wild from recordings obtained from video-sharing websites. This corpus encompasses recordings of 491 individuals who express their opinions on various topics or share personal experiences. The participants exhibit a broad spectrum of rich emotional behaviors in their videos, contributing to the dataset's diversity. In total, the MSP-Face dataset comprises 70.7 hours of audiovisual data, split into 24.7 hours of labeled content (with emotional labels) and 46 hours of unlabeled content. For our framework's pre-training, we utilize the MSP-Face corpus, treating the entire dataset as an unlabeled set, without employing the available labels.

F. Implementation Details

The stage 1 adaptive pre-training step involves tuning all transformer layers of the original off-the-shelf SSL speech model using our proposed multitask setup that incorporates all three objectives. In this process, the audio 1D CNN encoder layers of the speech models are kept frozen. Additionally, we introduce a mean pooling layer at the output of the last transformer layer. This layer is employed to obtain an averaged representation from the transformer encoder layer, which is then fed into three separate prediction heads, each corresponding to one of the objectives. The model undergoes training for 100 epochs, with a learning rate set at 1E-5, a batch size of 32, and a λ value of 0.5 for the loss. The model is trained using the Adam optimizer, and the most effective model checkpoint is preserved based on performance evaluated using a held-out test set that comprises 15% of the unlabeled data utilized in the adaptive pre-training step. In stage 2, for downstream fine-tuning, we modify the model architecture by replacing the multitask prediction heads with a two-layer fully connected prediction head. This alteration is specifically tailored for SER, building upon the model obtained from stage 1. As in the previous stage, the CNN encoder layers of the speech models remain frozen. The rest of the model is fine-tuned for 20 epochs, with a learning rate set at 1E-5 and a batch size of 32. During training, the model is optimized using the Adam optimizer and cross-entropy loss as the objective function. All implementations were carried out in PyTorch, and training was performed on an NVIDIA Tesla V100.

IV. Results

A. Baseline Models Description

The baseline models used for comparison throughout this study, referred to explicitly as "Original" in the results tables, are the off-the-shelf versions of SSL pre-trained speech models, including Wav2vec 2.0, WavLM, HuBERT, and Data2vec. These models maintain their original architectures, comprising a convolutional feature extractor (audio 1D *convolutional neural network* (CNN) encoder layers) followed by multiple transformer encoder layers. During fine-tuning for the SER downstream task, following common SSL fine-tuning practice [9], [10], [76], [77], we fix the convolutional feature encoder and update only the Transformer and task heads. This strategy preserves low-level acoustic invariances and improves stability and computational efficiency on small labeled targets. Additionally, a mean pooling layer is introduced at the output of the last transformer layer to generate utterance-level representations. Finally, we append a two-layer fully connected head specifically for emotion classification. We also include a purely speech-based baseline using the eGeMAPS handcrafted acoustic features [20], which serves as a reference point for unimodal SER performance without any pre-trained speech model. This model is trained using the same downstream classifier setup for fair comparisons.

B. Evaluation of Proposed Multimodal Adaptation Strategy

This section compares our proposed approach with the performance of the off-the-shelf SSL base versions of Wav2vec 2.0 [8], WavLM [9], HuBERT [10], and Data2vec [11]. Table I reports the average and standard deviation of the F1-score performances across five trials on each dataset. To ensure fair comparisons, all models are fine-tuned using the same experimental setup, including training batches, hyperparameters, and implementation details. This consistency isolates the impact of the proposed adaptation strategy, allowing for a direct assessment of its effectiveness. Additionally, we include a handcrafted feature-based baseline (eGeMAPS) to provide a conventional unimodal reference for SER.

In the CREMA-D dataset, all models showed improved performance after adaptation and downstream fine-tuning, with WavLM showing the most substantial improvement, from 0.675 ± 0.03 in the original to 0.724 ± 0.02 in the adapted model. In the MSP-IMPROV dataset, adaptations led to notable improvements for all models, especially for Wav2vec 2.0 and WavLM. For Wav2vec 2.0, our approach results in a notable increase from 0.501 ± 0.11 to 0.614 ± 0.04. For WavLM, the performance gain is also strong, increasing the F1-score from 0.555 ± 0.12 to 0.636 ± 0.05. In both CREMA-D and MSP-IMPROV corpora, these results are also substantially better than the eGeMAPS-based baseline, which reaches an F1-score of 0.387 (CREMA-D) and 0.354 (MSP-IMPROV), underscoring the benefit of using large pretrained speech models for unimodal SER.

The results on the MSP-Podcast corpus did not exhibit the same trends as in the other two databases. Table I shows that there is even a performance decrease observed when using WavLM with the proposed multimodal adaptation strategy. We hypothesize that this outcome may be attributed to the extensive size of the MSP-Podcast corpus. In comparison to the MSP-IMPROV and CREMA-D databases, the size of the MSP-Podcast corpus is about 1900% higher than each of the other datasets. In such cases, pre-training might not significantly expand the models' capabilities. The inherent

TABLE I
COMPARISON BETWEEN THE PROPOSED APPROACH (ADAPTED) AND THE ORIGINAL OFF-THE-SHELF VERSIONS OF THE SSL FINETUNED SPEECH BASELINES (ORIGINAL), ALONG WITH A HANDCRAFTED FEATURE-BASED BASELINE (EGEMAPS). RESULTS ARE AVERAGED OVER FIVE TRIALS (± STANDARD DEVIATION).

Dataset	**eGeMAPS**	**Wav2vec 2.0**		**WavLM**		**HuBERT**		**Data2vec**	
		Original	**Adapted**	**Original**	**Adapted**	**Original**	**Adapted**	**Original**	**Adapted**
CREMA-D	0.387 ± 0.02	0.650 ± 0.04	**0.681 ± 0.03**	0.675 ± 0.03	**0.724 ± 0.02**	0.639 ± 0.01	**0.704 ± 0.04**	0.653 ± 0.03	**0.672 ± 0.02**
MSP-IMPROV	0.354 ± 0.04	0.501 ± 0.11	**0.614 ± 0.04**	0.555 ± 0.12	**0.612 ± 0.05**	0.538 ± 0.07	**0.636 ± 0.04**	0.533 ± 0.02	**0.564 ± 0.04**
MSP-Podcast	0.328 ± 0.02	0.614 ± 0.02	**0.619 ± 0.02**	**0.632 ± 0.01**	0.618 ± 0.02	0.618 ± 0.02	**0.621 ± 0.01**	0.642 ± 0.01	**0.644 ± 0.03**

diversity and complexity of the large corpus can well tune the model to various emotional contexts on its own, which may limit the additional benefits of pre-training. This result suggests that pre-training with multimodal emotion-related tasks might be more effective in low-resource scenarios with a narrower emotional range and fewer data points available for training. We explore this scenario in Section IV-C.

C. Results on Low Resource Subsets

As indicated in Table I, the results on the MSP-Podcast corpus did not exhibit the performance gain observed with the MSP-IMPROV and CREMA-D corpora. We hypothesize that the pre-training of the models with multimodal unlabeled objectives could be more beneficial in low-resource scenarios, such as those observed in the CREMA-D and MSP-IMPROV corpora. To investigate this hypothesis, we conduct additional experiments on the MSP-Podcast corpus by restricting the amount of training data to determine if the observed improvements in performance persist under these conditions.

For our experiments, we create training subsets of the MSP-Podcast corpus of different sizes, preserving the distribution of the emotional classes in the subsets. This setting mimics the original dataset content in a low-resource environment, preserving the original full dataset's emotional class distribution. We create three subsets containing 1,000, 5,000, and 10,000 files, respectively. Moreover, we randomly re-sampled each training subset five times and conducted experiments five times for each model and setup.

1) Classification Results: Table II presents an evaluation of the models trained with the three subsets created from the MSP-Podcast corpus. The results clearly indicate that using our approach leads to a performance gain over the original model when the training set is limited. This trend suggests that the proposed adaptation strategy is beneficial for this corpus as well. Notably, the relative improvements in the adapted models are higher in smaller subsets (1K and 5K). For instance, the Wav2vec 2.0-based model improves the F1-score from 0.569 to 0.577 in the 1K subset. The F1-score improvement is from 0.589 to 0.610 in the 5K subset. These results support the hypothesis that pre-training with multimodal unlabeled objectives is particularly beneficial in low-resource scenarios. We also observe that the improvements vary among the different models, hinting at the varying sensitivities of these models to the adaptation process. HuBERT and Data2vec, for example, show significant performance boosts when adapted, especially with the smallest dataset, which might imply their greater efficacy in capitalizing on the adaptation process under resource constraints.

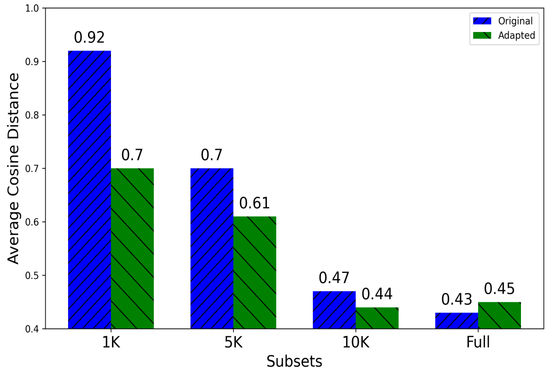


Fig. 3. Plot showing the average cosine distance obtained from comparing the last transformer layer's embeddings of each resulting model obtained from using each subset derived from MSP-Podcast 1.11 against their respective downstream fine-tuning starting points (original or adapted)

The low standard deviation values reported in Table II indicate that the performance enhancements are robust and repeatable across different randomized re-samplings of the training data. In essence, the experiments reinforce the utility of adapting pre-trained models for SER tasks, demonstrating that such adaptations can yield performance benefits in low-resource environments and pointing towards a promising direction for future research in the field of emotion recognition with limited labeled data.

2) Embeddings Distance Analysis: In this section, we analyze the average cosine distances of models obtained using subsets ranging from 1K to 10K training samples derived from the MSP-Podcast dataset. We compare each subset setup in both scenarios: fine-tuning directly from the original model and fine-tuning after the adaptation step. Through this experiment, we aim to measure the divergence of the transformer models' last-layer embeddings with respect to those of the original and adapted models under the limited resource setup. We seek to understand the extent to which the fine-tuned models have moved away from their initial configuration following the fine-tuning process. The analysis of embedding distance reveals insightful patterns about the fine-tuning process of SER models using the MSP-Podcast dataset.

Figure 3 shows the results. As the training subset size increases from 1K to 10K, the embedding distance to the original model correspondingly decreases. We note that the primary goal of this analysis is not to evaluate the absolute magnitude of the cosine distances across subset sizes, but

TABLE II
COMPARISON BETWEEN THE PROPOSED APPROACH (ADAPTED) AND THE ORIGINAL OFF-THE-SHELF VERSIONS OF THE SSL FINETUNED SPEECH BASELINES (ORIGINAL). THE TABLE REPORTS THE AVERAGE PERFORMANCE METRICS ACROSS FIVE TRIALS AND ITS STANDARD DEVIATION OF MSP-PODCAST 1.11 SUBSETS.

	Wav2vec 2.0		**WavLM**		**HuBERT**		**Data2vec**	
	Original	**Adapted**	**Original**	**Adapted**	**Original**	**Adapted**	**Original**	**Adapted**
1K Subset	.569 ± .02	**.577 ± .03**	.572 ± .01	**.591 ± .01**	.544 ± .03	**.563 ± .02**	.480 ± .05	**.569 ± .01**
5K Subset	.589 ± .01	**.610 ± .01**	.604 ± .01	**.616 ± .01**	.586 ± .02	**.602 ± .01**	.563 ± .01	**.598 ± .01**
10K Subset	.588 ± .02	**.612 ± .01**	.611 ± .01	.611 ± .01	.614 ± .01	**.616 ± .01**	.592 ± .01	**.607 ± .01**

TABLE III
MACRO-F1 (± STD.) FOR THE AUDIO-ONLY AV-HUBERT BASELINE VERSUS THE **BEST** ADAPTED SSL MODEL ON EACH CORPUS.

Dataset	***AV* -HuBERT**	**Best Adapted SSL**
CREMA-D	0.645 ± 0.027	0.724 (WavLM)
MSP-IMPROV	0.452 ± 0.068	0.636 (HuBERT)
MSP-Podcast	0.507 ± 0.032	0.644 (Data2vec)

TABLE IV
F1 BIAS BEFORE AND AFTER ADAPTATION FOR THE HUBERT MODEL AND COMPARISON WITH AV-HUBERT. LOWER VALUES INDICATE LESS CLASS-LEVEL BIAS.

Dataset	**Original**	**Adapted**	**AV-HuBERT**
CREMA-D	0.1864	**0.1319**	0.1892
MSP-IMPROV	0.1426	**0.0647**	0.1033

rather to compare the relative representational shifts between the original and adapted models within each subset condition (1K, 5K, 10K, Full). This trend indicates that with the addition of more training data, the fine-tuned models start to rely less on the adaptation step. The distance to the model adapted for emotion recognition is lower across all data subsets compared to that to the original model, which aligns with expectations since the adapted model was specifically tailored for SER tasks. This consistent pattern supports the hypothesis that adaptation provides a strong initialization for emotion-relevant representations, especially when labeled data is scarce. Notably, there is an initial decrease in distance as we move from the 1K to the 5K subset, followed by a slight increase from the 10K subset to the full training set in the database. The increase in distance from the 10K subset to the full training set might suggest a point of diminishing returns from using the adaptation step, where enough data is available to fully fine-tune the transformer layers of the SSL representations.

D. Comparison with AV-HuBERT

We evaluate AV-HuBERT [78] to examine whether a repurposed multimodal model can rival our adaptation strategy when restricted to speech-only inference. AV-HuBERT is an audio-visual model originally designed for lip reading. For the experiment, we retain only the audio branch; the visual front-end and fusion layers were disabled so that training and inference remain strictly unimodal, matching our problem formulation which is audio-only at inference. We fine-tune the model following exactly the same procedure used for the SSL baselines in Section IV-B.

Table III reports macro-F1 results averaged over five runs. We observe across all corpora that the audio-only AV-HuBERT baseline lags behind the best adapted SSL model by 6% to 18% (absolute). We believe this performance gap is due in part to the fact that the AV-HuBERT's architecture and pre-training are optimized for phonetic-level representations used in visual speech recognition. Our models, in contrast, focus on pre-trained tasks that are relevant for paralinguistic information, specifically emotion, leading to clear improvements.

E. Bias Analysis from Tasks

To examine whether the use of pre-text tasks from models (RoBERTa for text and EfficientNet-B2 for visual features) introduces performance bias into our SER model, we analyze the distribution of F1-scores across emotional classes. Specifically, we use the F1 Bias metric to quantify how uniformly the model performs across emotional categories. F1 Bias is computed as the average absolute difference between all pairs of per-class F1-scores. Formally, let A_g denote the F1-score for emotion class g, and G is the number of classes. Then:

$$\text{F1 Bias} = \frac{1}{G(G-1)} \sum_{g=1}^{G} \sum_{\substack{i=1 \\ i \neq g}}^{G} |A_g - A_i|$$

A lower F1 Bias indicates more balanced performance across classes and, hence, less bias. We perform this analysis using the HuBERT model results. Table IV reports the F1 Bias of the HuBERT model before and after applying the proposed adaptation strategy on the CREMA-D and MSP-IMPROV datasets. We observe a consistent reduction in F1 Bias following adaptation, indicating that the additional supervision does not inject bias into the model. Instead, it improves class-level balance. These findings suggest that incorporating multimodal emotion-relevant signals during pre-training improves both overall performance and fairness across emotional classes. Rather than introducing cascading errors or bias from the pseudo-labeling models, the auxiliary supervision contributes to more equitable model behavior across emotions.

To contextualize these findings further, we compared our adapted HuBERT model against AV-HuBERT [78], a multimodal pre-trained model, which integrates visual and audio information during pre-training. We selected HuBERT and AV-HuBERT for this comparison because they share a similar architecture and training origin, making them the most appropriate point of reference. The results in Table IV show that our adapted HuBERT model not only achieves lower bias than the original SSL baseline but also exhibits less per-class bias than AV-HuBERT, especially on the MSP-IMPROV

TABLE V
CROSS-CORPUS MACRO-F1 (± STD.) WHEN MODELS ARE TRAINED ON THE MSP-PODCAST CORPUS AND EVALUATED ON THE MSP-IMPROV CORPUS.

Model (trained on MSP-Podcast)	MSP-IMPROV F1
HuBERT Original	0.538 ± 0.048
HuBERT Adapted	**0.558 ± 0.046**

corpus. Importantly, our model maintains a purely speech-based inference path, while AV-HuBERT is fundamentally designed for audio-visual fusion tasks such as lip reading. The observed bias reduction improvements suggest that our adaptation strategy helps reduce bias while preserving generalization. These findings support our design choice and indicate that our approach does not overfit to biases in the pre-training data. Rather, it leads to improved performance balance across emotion classes compared to both standard SSL and AV-based models.

F. Cross-Corpus Generalization Ability

We conduct a cross-dataset experiment with the HuBERT backbone to understand how well the emotion-adapted representations transfer to an unseen domain. We use the MSP-Podcast corpus as the source domain and the MSP-IMPROV corpus as the target domain. Both the Original (SSL-only) model and the Adapted model were fine-tuned on the full MSP-Podcast corpus and then evaluated, without any additional training, on the MSP-IMPROV test set. The two corpora share four categorical emotions (anger, happiness, sadness, and neutral state), allowing a direct label match, but differ markedly in recording style (naturalistic podcasts vs. scripted studio dialogues), yielding a realistic domain-shift scenario.

Table V shows the results. The adapted model delivers an absolute gain of roughly 2% macro-F1 gain (absolute) over the SSL-only baseline, confirming that the emotion-centric pre-training yields more transferable speech representations. Although overall performance remains lower than the in-corpus result obtained when training directly on the MSP-IMPROV corpus (0.636 ± 0.04), the improvement under zero-shot transfer indicates enhanced robustness to domain shift. This result suggests that the proposed adaptation can complement future domain-adaptation techniques [13], [79]–[81], supporting scalability to real-world applications featuring diverse recording conditions and label distributions.

V. ABLATION STUDIES

In this section, we present ablation experiments to explore the multimodal objectives used in this study for emotion adaptation. We detail the preliminary experiments that informed the selection of these objectives and ablations related to varying the weights of objective losses in our proposed approach. The experiments in this section are all conducted using the base model of Wav2vec 2.0 [8] as the baseline and starting point for adaptation. Additionally, we conduct all experiments using the CREMA-D [25] and the MSP-IMPROV [26] corpora. We use only one model and the two smaller corpora for these experiments because it would not be feasible for us to run all models and corpora for every experiment.

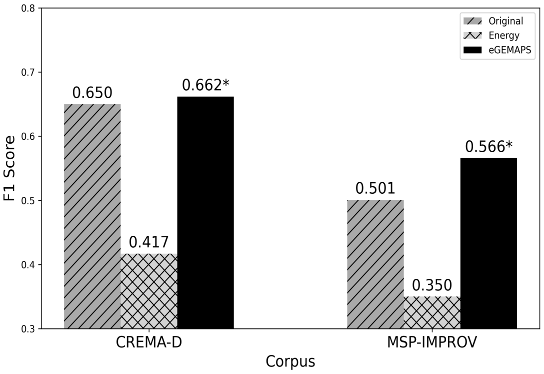


Fig. 4. Comparative Analysis of Model Performance Using Energy or eGEMAPS as objectives for Emotion Adaptation in Wav2vec 2.0-base. This figure illustrates the performance impact when using energy alone (resulting in a significant decline) versus employing eGEMAPS (leading to a marked improvement) for emotion adaptation, starting from the baseline of an original off-the-shelf Wav2vec 2.0 model. The asterisk symbol above the bars indicates that the model's performance using eGEMAPS as the objective is statistically significant compared to the other two explored formulations.

A. Pre-training the Models with Speech-Based Objective

We begin the analysis considering our preliminary study as a starting point [19], which showed that we can improve SER by integrating a speech energy-related task with facial activity-related tasks as pre-training objectives. First, we consider the speech energy-related task, which consists of the binary task of predicting if the speech energy is higher or lower than the median energy across sentences. We notice that using only this objective results in a marked decrease in SER performance. This decline occurs because classifying speech energy as high or low is overly simplistic for a speech model, leading to a shift in the model's representation towards a space that fails to adequately capture the nuances of emotional speech. To address this problem, we revised our approach. Instead of using a binary task to determine high/low energy, we now employ regression objectives to estimate the average of the eGEMAPS [20] features as the pre-training objectives (O1, Sec. III-A). The eGEMAPS features have been extensively and successfully utilized in various speech emotion recognition applications [82]. Therefore, we posit that using the eGEMAPS feature set extracted from speech as the training objective, instead of focusing solely on energy, will more effectively capture emotion-related cues for our model.

Figure 4 presents the results of employing either the binary energy-based task or the eGEMAPS regression task as the objective for emotion adaptation, starting with an off-the-shelf Wav2vec 2.0-base model (original). These results indicate that using energy as the sole objective significantly deteriorates the model's performance. In contrast, when eGEMAPS is utilized, there is a significant increase in performance compared to fine-tuning directly from the original off-the-shelf Wav2vec 2.0 model.

B. Combining Speech and Text Based Objectives

Leveraging pre-trained language models for generating pseudo labels has proven effective in enhancing speech sen-

TABLE VI
SUMMARY OF MODEL PERFORMANCE WITH DIFFERENT WEIGHTINGS OF OBJECTIVES O1 AND O2. THIS TABLE DISPLAYS AVERAGE F1-SCORES FROM FIVE RUNS WITH RANDOM INITIALIZATIONS, HIGHLIGHTING THE IMPACT OF VARYING THE WEIGHTINGS OF O1 AND O2 LOSSES IN THE ADAPTIVE PRE-TRAINING PIPELINE. THE TOP SECTION SHOWS RESULTS WITH DIFFERENT BALANCES BETWEEN O1 AND O2, NOTING THAT A λ VALUE OF .5 FOR O2 YIELDS OPTIMAL RESULTS. THE LOWER SECTION PRESENTS OUTCOMES WHEN O3 IS INCORPORATED ALONGSIDE O1 AND O2, DEMONSTRATING THE NECESSITY OF A BALANCED APPROACH DUE TO THE DIFFERING LOSS FUNCTIONS OF MSE FOR O1 AND O3, AND CROSS-ENTROPY FOR O2. THE ASTERISK SYMBOL (*) INDICATES THAT THE RESULTS ARE STATISTICALLY SIGNIFICANT. IN THE TOP SECTION, THE EVALUATION IS COMPUTED AGAINST THE PERFORMANCE OF THE O1 MODEL, WHILE IN THE LOWER SECTION, IT IS COMPUTED AGAINST THE PERFORMANCE OF THE BEST COMBINATION OBTAINED FROM INCORPORATING THE O1 AND O2 OBJECTIVES WITHOUT ANY ADDITIONAL OBJECTIVES.

Adding Textual Objective		
Weighting	CREMA-D	MSP-IMPROV
O1	0.662	0.566
O1+O2	0.684*	**0.608***
O1+.5*O2	**0.686***	0.607*
O1+10*O2	0.668	0.582
10*O1+O2	0.673*	0.556
.5*O1+O2	0.668	0.565
Adding Facial Objective		
Weighting	CREMA-D	MSP-IMPROV
O1+O2+O3	0.679	0.613*
O1+.5*O2+O3	**0.681**	**0.614***
O1+1.5*O2+O3	0.674	0.580

timent analysis, particularly in semi-supervised training scenarios [18]. This study investigates the integration of pseudo labels, derived from a pre-trained language model [22], as an additional objective in adaptive pre-training (O2, Sec. III-B), in conjunction with eGEMAPS features.

Table VI reports results obtained from including O2 (text) in the adaptive pre-training pipeline of the model, using different weights for the total loss. The results are reported as average F1-scores of 5 randomly initialized runs. We can see through different weightings of the O1 and O2 losses that a more balanced weighting of the losses or weighting O2 with a λ = .5 yields the best results for the proposed approach. The table also shows that adding O2 leads to substantial performance improvements over a model only pre-trained with O1.

C. Combining Speech, Text and Face based Objectives

Studies have also explored using visual inputs to improve speech representations [19], [64]. While these studies demonstrate the effectiveness of visual inputs in enhancing speech representations for emotion recognition, the integration of speech with both text and visual cues remains unexplored. This study explores adding a visual-based objective (O3, Sec. III-C), focusing on utilizing representations from a pre-trained EfficientNet-B2 model [21], which has been trained to recognize facial emotion expressions in the AffectNet corpus [71]. With O3 (visual), we leverage a system that has already mastered identifying and interpreting a wide range of facial expressions. Our goal is to encourage our model to concentrate on aspects of speech that closely align with visible emotional expressions. This approach could potentially lead to a model that is more finely attuned to the nuances of emotional speech. The objective tasks O1 (speech) and O3 (visual) employ the *mean-squared error* (MSE) for optimization, sharing a common loss function. In contrast, O2 uses the cross-entropy loss. This discrepancy requires a balanced approach among O1, O2, and O3, given their differing error scales and loss functions. For O2, we adjust the λ weight in relation to the other objectives. This adjustment is crucial since cross-entropy's error characteristics and loss function scale differ significantly from MSE. To address these variations, we strategically modify the λ weight for O2 relative to O1 and O3. This ensures the training process appropriately emphasizes the classification task of O2 while maintaining equilibrium with the regression tasks of O1 and O3.

The lower section of Table VI presents the results obtained by incorporating O3 alongside O1 and O2. We see from Table VI that keeping λ = .5 leads to the best balance in performance for the CREMA-D and MSP-IMPROV corpora with statistically significant improvements over the other setups. Additionally, looking at Figure 5, we see the overall progression in performance from the original model to utilizing our proposed adaptive step as we added more objectives into the training process. Note that although there is a small drop in performance for the CREMA-D corpus when we added O3 to the adaptive training, we see a larger gain in performance for the MSP-IMPROV corpus using O3.

VI. CONCLUSIONS

This study explored the use of mixed objectives generated from unlabeled data from speech, text, and facial expressions to improve speech emotion recognition systems. We found that combining these different types of objectives is useful in helping the speech representation to be more emotionally discriminative, especially when the training data is limited. Our approach uses pre-trained speech models that are further pretrained with the proposed strategy using information from all three modalities. This approach helps the speech representation

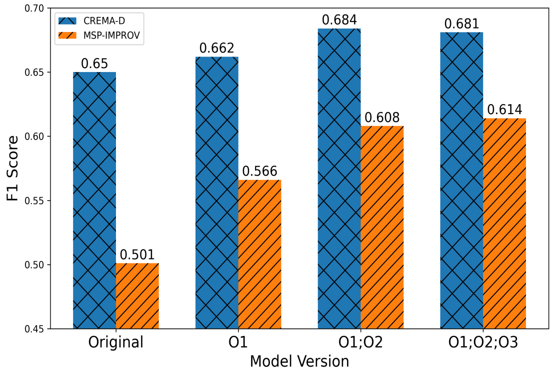


Fig. 5. Evolution of Model Performance through the Addition of Objectives in the Adaptive Training Process. This figure tracks the progression in performance of the model, starting from the original version and moving through various stages of incorporating additional objectives (O1, O2, and O3) into the proposed approach.

to learn emotional cues by leveraging the multimodal relationships observed in the externalization of emotions conveyed across acoustic, facial, and lexical features. We evaluated this strategy on three datasets, observing clear improvements in SER performance, particularly when there is not much labeled data. The evaluation also explored ablation studies considering subsets of the proposed multimodal pre-training objectives. Our results show that including all explored types of information makes our models better at identifying emotions.

This paper provides an effective strategy to improve an SER model by using a variety of unlabeled data. By tapping into the rich, complementary information provided by speech, text, and visual data, we can build more accurate and reliable emotion recognition models, especially in situations where labeled data is scarce. This work opens up new possibilities for improving SER systems and highlights the importance of using multimodal data in machine learning, even when the final objective is to design a unimodal system (e.g., a speech emotion recognition system). Future work includes extending this multimodal pre-training strategy to other emotion tasks (e.g., video emotion recognition) and/or exploring additional modalities (e.g., physiological signals). We will also explore perturbations of upstream text/ASR signals and usage of different sentiment models in our ablations, evaluate multi-task weighting strategies (uncertainty weighting, GradNorm) during training, and evaluate broader multi-domain and cross-lingual transfer methods.

Acknowledgment

This material is based upon work done, in whole or in part, in coordination with the Department of Defense (DoD). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DoD and/or any agency or entity of the United States Government.

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. 14, no. 2, pp. 1215–1227, April-June 2023.

[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] A. Reddy Naini, L. Goncalves, A. Salman, P. Mote, I. Ülgen, T. Thebaud, L. Moro-Velazquez, L. Garcia, N. Dehak, B. Sisman, and C. Busso, "The Interspeech 2025 challenge on speech emotion recognition in naturalistic conditions," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025, pp. 4668–4672.

[7] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*, Cambridge, UK, September 2019, pp. 441–447.

[8] 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, December 2020, pp. 12 449–12 460.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, October 2022.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[11] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning (ICML 2022)*. Honolulu, HI, USA: Proceedings of Machine Learning Research (PMLR), July 2022, vol. 162, pp. 1298–1312.

[12] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, September 2023.

[13] A. Reddy Naini, M. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*, Seoul, Republic of Korea, April 2024, pp. 12 031–12 035.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] S. Shon, P. Brusco, J. Pan, K. J. Han, and S. Watanabe, "Leveraging pre-trained language model for speech sentiment analysis," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235422693

[19] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Interspeech 2022*, Incheon, South Korea, September 2022, pp. 1168–1172.

[20] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.

[21] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132–2143, October-December 2022.

[22] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados, "XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 258–266. [Online]. Available: https://aclanthology.org/2022.lrec-1.27

[23] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[24] C. Busso, R. Lotfian, K. Sridhar, A. Salman, W.-C. Lin, L. Goncalves, S. Parthasarathy, A. Reddy Naini, S.-G. Leem, L. Martinez-Lucas, H.-C. Chou, and P. Mote, "The MSP-Podcast corpus," *ArXiv e-prints (arXiv:2509.09791)*, pp. 1–20, September 2025.

[25] 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.

[26] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, “MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception,” *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[27] 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.

[28] H.-C. Chou, L. Goncalves, S.-G. Leem, C.-C. Lee, and C. Busso, “The importance of calibration: Rethinking confidence and performance of speech multi-label emotion classifiers,” in *Interspeech 2023*, vol. To appear, Dublin, Ireland, August 2023.

[29] W.-C. Lin, K. Sridhar, and C. Busso, “DeepEmoCluster: A semi-supervised framework for latent cluster representation of speech emotions,” in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2021)*, Toronto, ON, Canada, June 2021, pp. 7263–7267.

[30] Z. Zhang, F. Weninger, M. Wollmer, and B. Schuller, “Unsupervised learning in cross-corpus acoustic emotion recognition,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, Waikoloa, HI, USA, December 2011, pp. 523–528.

[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] R. Lotfian and C. Busso, “Practical considerations on the use of preference learning for ranking emotional speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.

[33] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, “Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 6447–6451.

[34] T. Rahman and C. Busso, “A personalized emotion recognition system using an unsupervised feature adaptation scheme,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

[35] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, “Speech sentiment analysis via pre-trained features from end-to-end asr models,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7149–7153.

[36] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, “Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts,” in *Proc. Interspeech 2018*, 2018, pp. 247–251.

[37] 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.

[38] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, “Self-supervised speech representation learning: A review,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[39] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *ArXiv e-prints (arXiv:1807.03748)*, pp. 1–12, July 2018.

[40] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, “Bootstrap your own latent a new approach to self-supervised learning,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[41] M. Caron, H. Touvron, I. Misra, H. J’egou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233444273

[42] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S. R. Jang, C.-C. Lee, and H.-Y. Lee, “Emo-superb: An in-depth look at speech emotion recognition,” 2024.

[43] L. Goncalves, A. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, P. Garcia, N. Dehak, B. Sisman, and C. Busso, “Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results,” in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, Quebec, Canada, June 2024, pp. 247–254.

[44] S. Srinivasan, Z. Huang, and K. Kirchhoff, “Representation learning through cross-modal conditional teacher-student training for speech emotion recognition,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6442–6446.

[45] W.-C. Lin, L. Goncalves, and C. Busso, “Enhancing resilience to missing data in audio-text emotion recognition with multi-scale chunk regularization,” in *ACM International Conference on Multimodal Interaction (ICMI 2023)*, vol. To appear, Paris, France, October 2023.

[46] B. T. Atmaja, K. Shirai, and M. Akagi, “Speech emotion recognition using speech feature and word embedding,” *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 519–523, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:212648833

[47] B. Zhang, S. Khorram, and E. Mower Provost, “Exploiting acoustic and lexical properties of phonemes to recognize valence from speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, United Kingdom, May 2019, pp. 5871–5875.

[48] L. Goncalves and C. Busso, “Learning cross-modal audiovisual representations with ladder networks for emotion recognition,” in *IEEE International Conference onicassp Acoustics, Speech and Signal Processing (ICASSP 2023),*, Rhodes island, Greece, 2023.

[49] L. Schoneveld, A. Othmani, and H. Abdelkawy, “Leveraging recent advances in deep learning for audio-visual emotion recognition,” *Pattern Recognition Letters*, vol. 146, pp. 1–7, June 2021.

[50] F. Ma, S. L. Huang, and L. Zhang, “An efficient approach for audio-visual emotion recognition with missing labels and missing modalities,” in *IEEE International Conference on Multimedia and Expo (ICME 2021)*, Shenzhen, China, July 2021, pp. 1–6.

[51] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, “Versatile audio-visual learning for emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 16, no. 1, pp. 306–318, January-March 2025.

[52] 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.

[53] ——, “Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features,” *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2156–2170, October-December 2022.

[54] L. Goncalves, H.-C. Chou, A. Salman, C.-C. Lee, and C. Busso, “Jointly learning from unimodal and multimodal-rated labels in audio-visual emotion recognition,” *IEEE Open Journal of Signal Processing*, vol. 6, pp. 165–174, January 2025.

[55] 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.

[56] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues,” in *AAAI Conference on Artificial Intelligence (AAAI 2020)*, vol. 34, New York, NY, USA, February 2020, pp. 1359–1367.

[57] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, “A survey of deep learning-based multimodal emotion recognition: Speech, text, and face,” *Entropy*, vol. 25, no. 10, 2023. [Online]. Available: https://www.mdpi.com/1099-4300/25/10/1440

[58] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, “Multimodal sentiment analysis using hierarchical fusion with context modeling,” *Knowledge-Based Systems*, vol. 161, no. 1, pp. 124–133, December 2018.

[59] 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.

[60] W. Rahman, M. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, “Integrating multimodal information in large pretrained transformers,” in *Association for Computational Linguistics (ACL 2020)*, Online, July 2020, pp. 2359–2369.

[61] Y. Li, T. Zhao, and T. Kawahara, “Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning,” in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2803–2807.

[62] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, June 2011.

[63] B. Sun, L. Li, X. Wu, T. Zuo, Y. Chen, G. Zhou, J. He, and X. Zhu, "Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 125–137, June 2016.

[64] A. Shukla, S. Petridis, and M. Pantic, "Does visual self-supervision improve learning of speech representations for emotion recognition?" *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 406–420, 2023.

[65] 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.

[66] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[67] 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.

[68] A. N. Salman and C. Busso, "Style extractor for facial expression recognition in the presence of speech," in *IEEE International Conference on Image Processing (ICIP 2020)*, Abu Dhabi, United Arab Emirates (UAE), October 2020, pp. 1806–1810.

[69] 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.

[70] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, October 2016.

[71] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, January-March 2019.

[72] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.

[73] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 794–803. [Online]. Available: http://proceedings.mlr.press/v80/chen18a.html

[74] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[75] E. Mower Provost, Y. Shangguan, and C. Busso, "UMEME: University of Michigan emotional McGurk effect data set," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 395–409, October-December 2015.

[76] X. Zhang, X. Zhang, W. Chen, C. Li, and C. Yu, "Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments," *Scientific Reports*, vol. 14, no. 1, p. 9543, Apr 2024.

[77] S. Amiriparian, F. Packań, M. Gerczuk, and B. W. Schuller, "Exhubert: Enhancing hubert through block extension and fine-tuning on 37 emotion datasets," in *Proceedings of Interspeech 2024*, 2024, pp. 2635–2639.

[78] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[79] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1959–1972, October-December 2022.

[80] P. Mote, D. Robinson, E. Richerson, and C. Busso, "Vector quantized cross-lingual unsupervised domain adaptation for speech emotion recognition," in *Interspeech 2025*, Rotterdam, The Netherlands, August 2025.

[81] P. Mote, B. Sisman, and C. Busso, "Unsupervised domain adaptation for speech emotion recognition using K-Nearest neighbors voice conversion," in *Interspeech 2024*, Kos Island, Greece, September 2024, pp. 1045–1049.

[82] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, "Speech emotion recognition using machine learning — a systematic review," *Intelligent Systems with Applications*, vol. 20, p. 200266, 2023.

Lucas Goncalves (S'22) received his BS in Electrical Engineering from University of Wisconsin - Platteville, in 2018. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas. At UTD, he is a Research Assistant at the multimodal Signal Processing (MSP) laboratory. In 2022, he was awarded the Excellence in Education Doctoral Fellowship from the Erik Jonsson School of Engineering and Computer Science. His research interests include areas related to affective computing, deep learning, and multimodal signal processing. He is also a student member of the IEEE Signal Processing Society.

Carlos Busso (S'02-M'09-SM'13-F'23) is a Professor at Language Technologies Institute, Carnegie Mellon University, where he is also the director of the *Multimodal Speech Processing* (MSP) Laboratory. He received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. His research interest is in human-centered multimodal machine intelligence and applications, focusing on the broad areas of speech processing, affective computing, multimodal behavior generative models, and foundational models for multimodal processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. His students received the third prize IEEE ITSS Best Dissertation Award (N. Li) in 2015, and the AAAC Student Dissertation Award (W.-C. Lin) in 2024. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). In 2023, he received the Distinguished Alumni Award in the Mid-Career/Academia category by the *Signal and Image Processing Institute* (SIPI) at the University of Southern California. He received the 2023 ACM ICMI Community Service Award. He is currently a Senior Area Editor of IEEE/ACM Speech and Language Processing. He is a member of AAAC and a senior member of ACM. He is an IEEE Fellow and an ISCA Fellow.