# Leveraging Content and Acoustic Representations for Speech Emotion Recognition

Soumya Dutta, *Graduate Student Member, IEEE*, and Sriram Ganapathy, *Senior Member, IEEE*

***Abstract*—Speech emotion recognition (SER), the task of identifying the expression of emotion from spoken content, is challenging due to the difficulty in extracting representations that capture emotional attributes. The scarcity of labeled datasets further complicates the challenge where large models are prone to over-fitting. In this paper, we propose CARE (Content and Acoustic Representations of Emotions), where we design a dual encoding scheme which emphasizes semantic and acoustic factors of speech. While the semantic encoder is trained using distillation from utterance-level text representations, the acoustic encoder is trained to predict low-level frame-wise features of the speech signal. The proposed dual encoding scheme is a base-sized model trained only on unsupervised raw speech. With a simple light-weight classification model trained on the downstream task, we show that the CARE embeddings provide effective emotion recognition on a variety of datasets. We compare the proposal with several other self-supervised models as well as recent large-language model based approaches. In these evaluations, the proposed CARE is shown to be the best performing model based on average performance across 8 diverse datasets. We also conduct several ablation studies to analyze the importance of various design choices.**

***Index Terms*—Emotion recognition, representation learning, self-supervised learning, speech-text alignment.**

## I. Introduction

Speech Emotion Recognition (SER) focuses on detecting the speaker's emotional state from the audio signal. Recognizing emotions in speech has significant applications across diverse fields, including human-computer interaction [1], social media analysis [2], customer service call centers [3], and mental health monitoring systems [4]. However, despite considerable progress, SER continues to pose challenges due to the complexity of human emotions and the inherent difficulties in effectively capturing them from limited labeled datasets.

Traditionally, SER systems have relied on various acoustic properties of speech signals. Lieberman et al. [5] emphasize the role of pitch contour in emotion analysis, while additional acoustic features, including energy, intensity, and speaking rate, were recognized as indicators of emotional class [6]. The features identified through the Interspeech para-linguistic challenges were rich in emotional properties while being high-dimensional [7], [8]. Eyben et al. [9] introduced a minimalist feature set to address this dimensionality issue. In recent years, the network architectures in SER commonly include convolutional neural networks (CNN) [10], [11], long short-term memory (LSTM) networks [12], and transformer models [13]. While these models perform well on the specific datasets, they often struggle to generalize across diverse datasets. In such settings, self-supervised learning (SSL) models have emerged as a promising solution. Notable examples of SSL approaches include wav2vec 2.0 [14], HuBERT [15], and WavLM [16]. These models are engineered to capture speech patterns similar to textual models like BERT [17]. Although trained on neutral speech data, these models have demonstrated encouraging results in emotion recognition tasks [18], [19]. The emotion recognition performance may be further enhanced by training these models with emotion-aware self-supervised objectives. Two recent examples are Vesper [20] and emotion2vec [21]. However, emotion in speech is also shaped by its semantic content [22]. For instance, identifying emotions from text transcripts is often more effective than interpreting them from raw audio [19]. The integration of speech content during the pre-training phase of SER models remains an under-explored yet promising area of research.

In this work, we introduce a self-supervised model for speech emotion recognition (SER) called **C**ontent and **A**coustic **R**epresentations of **E**motions (CARE). To the best of our knowledge, our approach is the first effort to pre-train a self-supervised model that integrates both semantic and acoustic components of speech. CARE leverages a dual encoding framework for processing speech signals: a semantic encoder, which aligns speech representations with sentence-level transcripts, and a non-semantic encoder, which aligns speech representations with low-level acoustic features from the PASE+ model [23]. The outputs of both encoders are combined, and a lightweight classification head is then trained to perform emotion recognition. The key contributions are :-

- Proposing a novel self-supervised model for speech emotion recognition (SER) consisting of dual encoders: a semantic encoder and an acoustic encoder.
- Developing an adaptation strategy for aligning pre-trained text models with speech inputs by convolutional adapters.
- Experimenting on 8 benchmark speech datasets with diverse tasks, showcasing the effectiveness of CARE.

Received 19 November 2024; revised 17 June 2025 and 15 July 2025; accepted 11 August 2025. Date of publication 1 September 2025; date of current version 8 September 2025. This work was supported in part by the Prime Ministers Research Fellowship (PMRF), in part by the Ministry of Education, India, in part by the Ministry of Information Technology (MEiTY) NLTM, India, Qualcomm Innovation Fellowship (QIF), and in part by British Telecom. The associate editor coordinating the review of this article and approving it for publication was Dr. Raul Fernandez.*(Corresponding author: Soumya Dutta.)*

The authors are with the Learning and Extraction of Acoustic Patterns (LEAP) laboratory, Electrical Engineering, Indian Institute of Science, Bangalore 560012, India (e-mail: soumyadutta@iisc.ac.in, soumya.besuee@gmail.com).

Digital Object Identifier 10.1109/TASLPRO.2025.3603853

- Identifying the individual and collective impact of semantic and acoustic representations for emotion recognition.

## II. Related Work

### A. Audio Feature Extraction for SER

Recently, deep learning-based representations have gained popularity as low-level acoustic features. Notable examples include the SincNet architecture by Ravanelli et al. [24] and interpretable Gaussian filters by Agrawal et al. [25]. The LEAF front-end [26], was utilized by Dutta et al. [11] for speech emotion classification. Typically, these models require end-to-end training of both feature extractors and classifiers. In contrast, the proposed CARE architecture is a self-supervised model designed to generalize across diverse datasets.

### B. Self-Supervision for SER

One of the earliest self-supervised model for the task of speech emotion recognition was proposed by Pascual et al. [27]. This consisted of processing a speech signal by the SincNet model [24] followed by trainable convolutional blocks to predict a number of speech features such as the waveform, mel-frequency cepstral coefficients (MFCCs), pitch etc. Ravanelli et al. [23] further modified this model by adding more self-supervised tasks such as predicting FBANK and Gammatone features [28] to develop the PASE+ model.

Among the general purpose speech SSL models that were proposed over the years, WavLM [16], was shown to outperform other models such as HuBERT [15] and wav2vec2.0 [14] for emotion recognition. Vesper [20] used a modified masking strategy to emphasize high pitch/energy regions of speech—known indicators of emotion—and derived targets for these masked regions from a WavLM teacher model. A similar strategy was employed by Ma et al. in emotion2vec [21], which utilized a pre-trained data2vec model as the teacher. Emotion2vec [21] also learns a global embedding to enhance SER performance. In contrast, the proposed CARE model integrates semantic content along with acoustic features.

### C. Multimodal Emotion Recognition

The use of speech signals alongside text transcripts for multimodal emotion recognition has been explored in several prior works [11], [29], [30]. These approaches typically involve separate modeling of the two modalities, followed by a fusion stage. In contrast, CARE is designed to model the semantic and acoustic properties of speech with the uni-modal input.

### D. Speech-Text Aligned Representations

The alignment of speech and text modalities has received renewed attention for speech representation learning. The SONAR model [31] aligns a speech encoder with textual representations at the utterance level. With the increasing prominence of large language models (LLMs), recent approaches have integrated speech encoders with LLMs. Notably, the SALMONN model by Tang et al. [32] introduced an audio encoder consisting of Whisper model and a music encoder along with the LLaMA language model [33]. Hu et al. [34] proposed WavLLM, combining Whisper and WavLM encoders with the LLaMA model. These LLM-based approaches harness aligned speech-text representations, enabling prompt-based applications. However, their substantial model sizes (e.g., 7B parameters for SALMONN) present significant computational demands for both training and inference. In contrast, CARE achieves superior performance on various downstream datasets with a much smaller size of 160 M parameters.

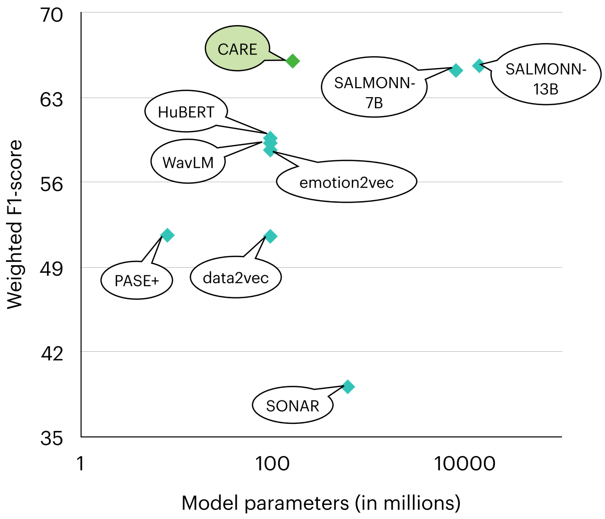

Fig. 1. Scatter plot of inference model size (parameters in millions) versus the SER performance (average weighted F1-score over 8 datasets). CARE is seen to achieve a better trade-off compared to the existing solutions. For more details, refer Tables II and IV and the associated discussions.

*Summary:* The landscape of various SER methods is summarized in Fig. 1. We highlight a clear gap in current modeling frameworks: models either prioritize efficiency with limited performance (those in the lower end of the x-axis), or focus on maximizing performance with increased memory and compute requirements (typically based on LLMs). To address this gap, we propose CARE, that combines the computational efficiency of smaller models with the high performance of large-scale systems, thereby providing a superior trade-off between efficiency and performance.

## III. Proposed Approach

### A. Background

*1) RoBERTa:* One of the significant contributions in creating a text representation model was proposed by Devlin et al. [17]. Liu et al. [35] trained this architecture on a larger corpus of textual data without the next sentence prediction task. This pre-trained model, known as robust optimized BERT approach (RoBERTa), was shown to outperform BERT in a number of downstream tasks.

### B. CARE Model

We propose a dual encoding scheme (semantic and acoustic encoders) to process the speech signal through distinct supervisory signals suited to their respective objectives. The chosen supervision for each encoder is detailed as follows:

*Semantic supervision:* We do not assume the availability of ground-truth text transcripts for the pre-training data. In such a scenario, pre-trained automatic speech recognition (ASR) systems (Whisper-large-v3 [36]) offer an alternative for generating these transcripts. Typically, ASR systems have been shown to exhibit higher word error rates (WER) on emotional speech compared to neutral speech datasets [22]. Podcast recordings, on the other hand, provide sufficiently long context and offer a broad content variety suitable for pre-training the semantic encoder. Specifically, we observe a WER of 12.53%, which may be reasonable for SER tasks.

Since the semantic encoder's purpose is to align the speech signal with its content to facilitate emotion recognition, an ASR-style alignment loss could be applied. However, a sentence-level representation for text is more appropriate for the task of emotion recognition as established by Fan et al. [37]. Therefore, we extract contextual word-level embeddings from the transcripts using a pre-trained RoBERTa model [35] and mean-pool these embeddings to obtain a single feature vector representing the entire transcript. These utterance-level embeddings serve as the supervisory signal, or "teacher", for the semantic encoder in our CARE model. We denote these utterance-level embeddings by $\boldsymbol{y}_{text}$.

*Acoustic Supervision:* In prior works, mean-pooled representations have shown to encode characteristics like speaker identity, accent, and language [38]. However, we speculate that emotion in speech is often contained in fine-grained acoustic attributes such as pitch, rhythm, and their modulations [6]. Thus, a frame-level target is chosen for the acoustic encoder.

A direct approach for the frame level acoustic targets would involve masking parts of the speech signal and reconstructing them. However, prior works show that random masking is less effective for emotion recognition than selectively masking high-energy or high-pitch regions, as demonstrated by Chen et al. [20]. Based on these observations, we choose to predict PASE+ features, which encompass filter-bank energies, pitch, and other low-level descriptors essential for capturing emotion. Specifically, we use frame-level PASE+ features with 256 dimensions as targets for the acoustic encoder in our CARE model. These features are down-sampled by a factor of 2, producing target descriptors at a frequency of 50 Hz. We denote the acoustic targets from the PASE+ model by $\boldsymbol{y}_{pase}$.

*1) Model Architecture:* The speech signal is first processed through a series of convolutional layers designed to produce frame representations every 20 ms. These are followed by a stack of six transformer layers, forming the common encoder that serves both the acoustic and semantic encoder pathways in the proposed model.

The semantic encoder is designed to align the speech representations with its corresponding generated transcript. This encoder consists of six transformer layers which are initialized with the weights from a pre-trained text representation model. Being trained with textual data, the transformer layers in the semantic encoder do not generalize to speech representations. To address this, we propose a novel adaptation strategy by introducing two 1D-convolutional blocks—one placed before and one after each transformer layer.

The first block adjusts the speech representations from the common encoder to align them with the internal representations expected by the text-based model. The second block refines these representations post-transformer processing. Additionally, following established practice for processing speech in text models [32], [39], [40], the time resolution of the speech sequence is reduced before processing by the transformer layers in the semantic encoder. Specifically, the convolutional block preceding each transformer layer down-samples the sequence length by a factor of three, while the block following it up-samples it by the same factor. Each convolutional block consists of a single convolutional layer, with a kernel size of 5, and input and output channels set to 768 in order to match the dimension of the pre-trained transformer layers. While adaptation of speech SSL models with convolution layers has been explored in prior works [41], [42], adapting pre-trained text models for speech tasks, using convolutional adapters, is explored for the first time in this work. Importantly, the transformer layers themselves are not updated during training. Finally, the semantic encoder's output representations are average-pooled to produce an utterance-level representation.

The acoustic encoder also consists of six transformer layers, with its output subsequently mapped to 256 dimensions, using a fully-connected layer, to match the PASE+ feature targets. Fig. 2 provides a block diagram of the CARE model.

*2) Loss:* A semantic loss, $L_{sem}$ and a frame-level acoustic loss, $L_{acoust}$, are employed for training the semantic and acoustic encoders, respectively. Denoting the semantic supervision by $\boldsymbol{y}_{text}$ and the output from the semantic encoder as $\hat{\boldsymbol{y}}_{sem}$, the semantic loss is the mean square error (MSE) loss:

$$L_{sem.} = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{y}^{i}_{text} - \hat{\boldsymbol{y}}^{i}_{sem}||_2^2 \tag{1}$$

where $N$ denotes the batch size.

For the frame level loss, let $\hat{\boldsymbol{y}}_{acoust} \in \mathbb{R}^{N\times T\times D}$ denote the output of the acoustic encoder, where $N$, $T$ and $D$ denote the batch size, number of frames per utterance and the dimension of the representation, respectively. The loss is defined as:

$$L_{acoust.} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T} ||\boldsymbol{y}^{ij}_{pase} - \hat{\boldsymbol{y}}^{ij}_{acoust}||_2^2 \tag{2}$$

where $\boldsymbol{y}_{pase}$ denotes the acoustic target. The total loss during pre-training is given by

$$L_{tot.} = L_{sem.} + \lambda L_{acoust.} \tag{3}$$

where $\lambda$ is decided based on the validation performance.

*3) Inference:* For evaluating the model across various downstream tasks, we adopt the paradigm proposed in the SUPERB benchmark [43]. The outputs from each transformer layer of the acoustic encoder are concatenated with the outputs from the convolution block following each transformer layer in the semantic

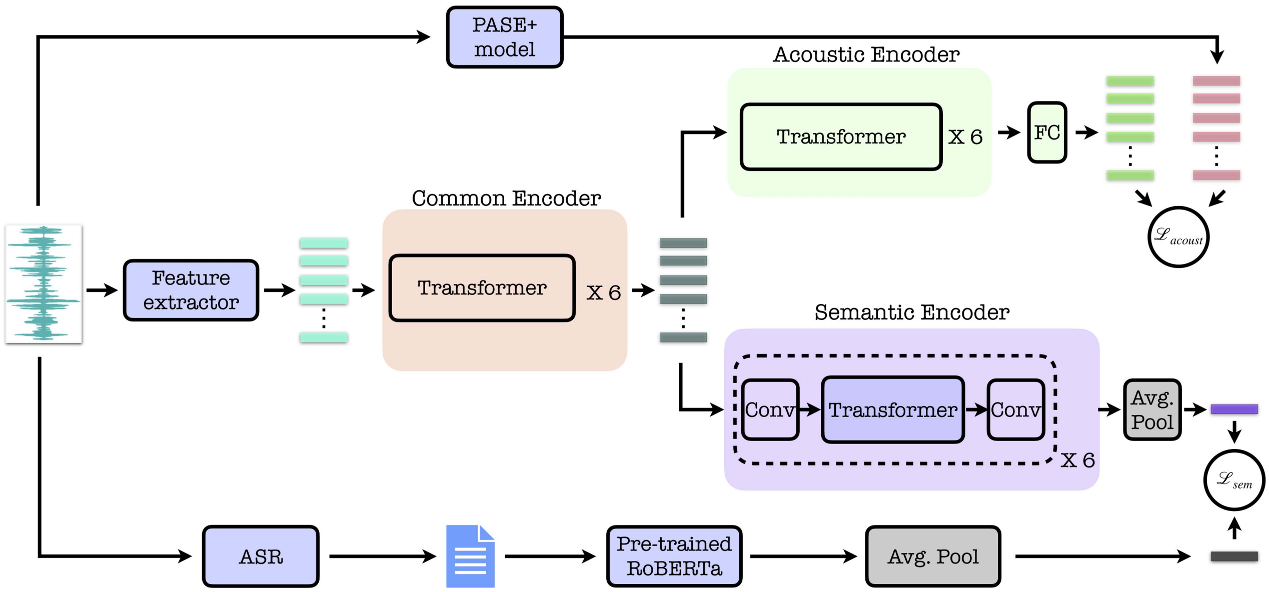

Fig. 2. Block diagram of the proposed CARE model. The acoustic encoder of the model is trained with PASE+ features as targets. Blocks in blue indicate either frozen components or those with no learnable parameters. For the semantic encoder the transformer layers are frozen while the convolutional adapters are trained. As the dimension of the output from the acoustic encoder is 768, a FC layer is attached to match the PASE+ feature dimension of 256. This FC layer and the average pool block after the semantic encoder are not used during inference.

encoder. These are then combined with layer-wise outputs from the common encoder and the convolutional feature extractor. This process yields a total of 13 layer representations—one from the convolutional feature extractor, six from the common encoder, and six from the concatenated semantic and acoustic encoders. A convex combination of these layer representations is then fed into a classification head. It is to be noted that, during inference, the fully-connected layer in the acoustic encoder and the average pooling block in the semantic encoder are not used.

In this setup, the only learnable parameters for the downstream tasks are the weights for the convex combination and those of the lightweight classification head.

## IV. Experiments and Results

### A. Pre-Training

The MSP-PODCAST corpus [44] is used for the task of pre-training. A total of 149,307 samples amounting to 230 hours of emotional speech data are used. Out of these, 80% of the data is randomly chosen as the training set while the remaining 20% serves as the validation set. The Whisper-large-v3 model is used for generating the transcripts (the WER observed is 12.53% ), while the pre-trained RoBERTa model is used for encoding the transcripts. The common encoder is initialized with first 6 layers of the WavLM-base model, while the acoustic encoder is initialized with the last 6 layers of the same. The convolutional feature extractor is also initialized from the WavLM-base model. The 6 transformer layers of the semantic encoder are initialized with the weights of the last 6 layers of a pre-trained RoBERTa base model, while the convolutional adapters are randomly initialized.

TABLE I
Summary of the Evaluation Datasets

| Datasets | # Train Utt. | # Val. Utt. | # Test Utt. | # Classes | Class Bal. | Spkr. Ind. |
|---|---|---|---|---|---|---|
| IEMOCAP-4 [45] | 4425 | 1102 | 1102 | 4 | × | ✓ |
| IEMOCAP-6 [45] | 5947 | 1487 | 1387 | 6 | × | ✓ |
| MELD [46] | 9988 | 1108 | 2610 | 7 | × | × |
| CMU-MOSI [47] | 1188 | 325 | 686 | 2 | ✓ | ✓ |
| DAIC-WOZ [48] | 6003 | 2097 | 2097 | 2 | ✓ | ✓ |
| RAVDESS-Song [49] | 704 | 132 | 176 | 6 | ✓ | ✓ |
| CaFE [50] | 624 | 156 | 156 | 7 | ✓ | ✓ |
| EmoDB [51] | 324 | 105 | 106 | 7 | ✓ | ✓ |

Class balanced denotes if the datasets are balanced across classes. The last column indicates if the training/test data have common speakers. For all the datasets, we perform 5-fold evaluation.

### B. Downstream Tasks

A summary of the different datasets used for evaluation is mentioned in Table I.

*1) IEMOCAP:* The IEMOCAP dataset consists of 151 video recordings split into 5 sessions. Each of these sessions is a conversation between a pair of subjects. Each recording is split into multiple utterances. There are a total of 10,039 utterances, each of which is labeled by human annotators as belonging to one of the 10 emotions - “angry”, “happy”, “sad”, “neutral”, “frustrated”, “excited”, “fearful”, “surprised”, “disgusted” or “other”. Keeping in line with previous works, we do a four-way classification task where we consider “angry”, “happy”, “sad”, “neutral” and “excited” categories (with “excited” and “happy” categories merged). We also have a separate setting of

6 emotional classes [52]. The first 6 of the 10 emotion classes are considered for this setting.

*2) MELD:* The MELD dataset [46] is a dataset created from video clippings of the popular TV show, "Friends". A seven way classification task is performed on this dataset, with each utterance being labeled as one of the 7 emotions - "angry", "sad", "joy", "neutral", "fear", "surprise" or "disgust".

*3) CMU-MOSI:* The CMU-MOSI dataset [47] has a total of 2199 utterances. Each utterance is labeled in the range of $[-3,3]$. Following previous works, we treat this as a binary classification problem with utterances having sentiment values in the range $[-3,0)$ being classified as negative sentiment and those with values in the range $[0,3]$ considered as positive sentiment. The dataset partitioning follows a prior work [53].

*4) DAIC-WOZ:* The DAIC-WOZ dataset [48] is a benchmark dataset for depression detection, consisting of 189 clinical interviews between the patient and the interviewer. Out of these 189 interviews, 107 are part of the training set while 35 interviews are part of the development subset. The dataset suffers from a data imbalance problem, with only 30 interviews labeled as "depressed" in the train set. In order to increase the balance, we follow [54] and extract 100 utterances randomly from each interview, which is labeled as depressed, while only 39 utterances are selected for interviews classified as "normal". The utterances from each interview are chosen randomly for the 5 splits. Following prior work [54], [55], [56], [57], we report results on the development set of this dataset.

*5) RAVDESS-Song:* The RAVDESS-Song dataset [49] has a total of 1012 song recordings by 23 different singers. Each recording in this dataset is sung in one of six different emotions, namely, "neutral", "calm", "happy", "sad", "angry" and "fear". We conduct a speaker independent evaluation for this dataset, and create 5 different splits. For each split, we keep recordings from 16 singers for training, while recordings from 3 separate singers are used for validation. The recordings from the remaining 4 speakers are used for evaluation.

*6) CaFE:* The CaFE dataset [50] is a Canadian French emotional dataset consisting of 936 utterances spoken by 12 speakers. Each utterance in this dataset is categorized as one of the seven emotions - "neutral", "angry", "disgust", "sad", "surprise", "fear" and "happy". Similar to RAVDESS-Song, we create 5 speaker independent splits for this dataset. The utterances belonging to 8 speakers are used for the training, while the remaining 4 speakers are used for validation and testing equally. The speakers used for train, validation and test are chosen randomly for the 5 splits.

*7) EmoDB:* The EmoDB dataset [51] has a total of 535 utterances spoken by 10 different speakers for the task of emotion recognition in German. Each utterance in this dataset is categorized as one of the seven emotions - "neutral", "angry", "disgust", "sad", "boredom", "fear" and "joy". Similar to RAVDESS-Song and CaFE, we create 5 different speaker independent splits for this dataset. The utterances belonging to 6 different speakers are chosen for training while the remaining 4 speakers are used for validation and testing equally. The speakers used for train, validation and testing are chosen randomly for each of the 5 splits.

### *C. Loss and Evaluation Metrics for Downstream Tasks*

The cross-entropy loss is used for training the downstream model weights (the convex combination weights and the lightweight classification head parameters). For testing, we use the weighted F1-score as the evaluation metric as many of the datasets are class-imbalanced (Table I). Denoting the F1 score of class $c$ with $N_c$ samples, by $F1_c$, the weighted F1-score is

$$WF1 = \frac{1}{\sum_{c=1}^{C} N_c} \sum_{c=1}^{C} N_c \times F1_c \tag{4}$$

We also report the unweighted average recall (UAR) for all cases, which is the mean of the class-wise recall scores.

### *D. Implementation Details*

*1) Pre-Training:* During pre-training, all the speech utterances from MSP-PODCAST are padded or randomly cropped to a duration of 5 seconds. The model is trained with a learning rate of $1e$-5 and a batch size of 128 with AdamW [58] as the optimizer. The model is trained for a total of 200,000 steps and the best model parameters based on validation set performance are chosen for evaluation of downstream datasets. We experiment with different values of $\lambda$ (Eq. 3) to balance the two losses during pre-training. Setting $\lambda = 0.1$ results in degraded performance, while increasing it to $\lambda = 10$ does not yield any significant improvement over $\lambda = 1$. Therefore, we fix $\lambda = 1$ for all the subsequent experiments.

*2) Fine-Tuning and Evaluation:* For the downstream task training, the speech signals are cropped to a maximum duration of 30 seconds or padded to a minimum duration of 1 s. For the depression detection dataset, DAIC-WOZ, each speech segment has a duration of 10 seconds [54].

Each layer output in the common, semantic, and acoustic encoders has a dimensionality of $T \times 768$, where $T$ denotes the number of frames in the speech signal, sampled at 50Hz. For the CARE model, as outputs from the 6 semantic and acoustic encoder layers are concatenated, the combined output dimension is $6 \times T \times 1536$. To align with this dimensionality, the output from the convolutional feature extractor and the common encoder's 6 layers are duplicated to yield features of dimension $7 \times T \times 1536$. Representations from these 13 layers are combined through a convex combination approach with learnable weights producing features of dimension $T \times 1536$. Following this, features are mean-pooled along the temporal dimension, producing a single 1536-dimensional vector per audio file. This is input into a classification head consisting of a two-layer feed-forward neural network that employs ReLU activation [59]. Only the weights for the convex combination of layer representations and those in the two-layer fully connected classification head are trained on each downstream dataset, consistent with the SUPERB framework [43].

We use a batch size of 32 with a learning rate of $1e$-4 and train the model for 50 epochs. The hidden dimension of the two-layer classification head is set to be 256. The AdamW optimizer is used here as well. All the models, including the

TABLE II
COMPARISON WITH OTHER WORKS FOR DOWNSTREAM DATASETS

| Datasets | WavLM [16] | HuBERT [15] | data2vec [60] | emotion2vec [21] | SONAR [31] | SALMONN [32] | | CARE |
|---|---|---|---|---|---|---|---|---|
| | **Params**:94M | **Params**:94M | **Params**:94M | **Params**:94M | **Params**:600M | **Params**:7B | **Params**:13B | **Params**:160M |
| IEMOCAP-4 | $65.9^{\pm0.5}(67.2)$ | $65.0^{\pm0.2}(68.0)$ | $62.7^{\pm0.7}(64.0)$ | $67.5^{\pm0.6\#}(69.0)$ | $59.4^{\pm0.4}(61.0)$ | $\mathbf{75.8}^{\pm0.6\#}(76.9)$ | $\underline{72.9}^{\pm2.3\#}(74.9)$ | $69.4^{\pm0.5}(70.1)$ |
| IEMOCAP-6 | $51.7^{\pm0.5}(48.4)$ | $50.7^{\pm0.9}(46.5)$ | $46.0^{\pm0.4}(42.3)$ | $54.1^{\pm0.6\#}(51.9)$ | $43.5^{\pm0.2}(41.0)$ | $\mathbf{59.3}^{\pm1.4\#}(55.7)$ | $\underline{58.1}^{\pm1.6\#}(55.2)$ | $55.0^{\pm0.4}(52.1)$ |
| MELD | $45.6^{\pm0.4}(24.3)$ | $45.3^{\pm0.6}(24.0)$ | $41.9^{\pm0.5}(23.1)$ | $47.6^{\pm0.3\#}(27.4)$ | $43.2^{\pm0.2}(23.3)$ | $\mathbf{53.3}^{\pm0.7}(33.4)$ | $\underline{52.6}^{\pm0.4}(32.8)$ | $48.1^{\pm0.8}(28.8)$ |
| CMU-MOSI | $64.1^{\pm0.8}(64.2)$ | $62.5^{\pm0.6}(62.5)$ | $59.7^{\pm0.4}(58.9)$ | $66.5^{\pm0.6}(65.9)$ | $\underline{74.6}^{\pm0.3}(73.9)$ | $\mathbf{78.0}^{\pm0.7}(77.0)$ | $72.8^{\pm1.0}(72.0)$ | $66.7^{\pm1.0}(66.2)$ |
| DAIC-WOZ | $63.2^{\pm1.5}(61.5)$ | $65.9^{\pm2.0}(61.9)$ | $\underline{67.8}^{\pm1.4}(65.7)$ | $61.6^{\pm0.7}(61.0)$ | $64.3^{\pm0.4}(63.7)$ | $62.6^{\pm3.4}(60.4)$ | $64.7^{\pm3.0}(61.1)$ | $\mathbf{68.5}^{\pm2.1}(67.1)$ |
| RAVDESS | $50.5^{\pm3.6}(49.1)$ | $\underline{53.5}^{\pm1.1}(55.7)$ | $38.5^{\pm5.2}(40.8)$ | $48.5^{\pm1.0}(51.0)$ | $11.8^{\pm2.0}(10.8)$ | $50.2^{\pm1.3}(54.2)$ | $51.9^{\pm3.6}(53.4)$ | $\mathbf{60.1}^{\pm1.6}(62.0)$ |
| CaFE | $66.6^{\pm2.6}(69.0)$ | $66.5^{\pm4.5}(69.1)$ | $48.8^{\pm4.3}(51.0)$ | $59.3^{\pm3.8}(62.6)$ | $5.7^{\pm1.4}(7.1)$ | $59.9^{\pm2.0}(62.8)$ | $\underline{69.8}^{\pm3.3}(71.4)$ | $\mathbf{77.0}^{\pm1.5}(78.1)$ |
| EmoDB | $66.5^{\pm4.8}(68.5)$ | $66.9^{\pm3.9}(68.2)$ | $48.9^{\pm3.1}(49.9)$ | $64.4^{\pm2.6}(66.7)$ | $10.2^{\pm2.0}(12.6)$ | $\underline{82.8}^{\pm2.9}(85.3)$ | $82.2^{\pm4.0}(84.1)$ | $\mathbf{83.4}^{\pm2.0}(83.9)$ |
| Avg. | 59.3(56.5) | 59.5(57.0) | 51.8(49.5) | 58.7(56.9) | 39.1(36.7) | 65.2(63.2) | $\underline{65.6}(63.1)$ | $\mathbf{66.0}(63.5)$ |

# models which include downstream dataset in pre-training. Results in bold, underlined indicate the best and the second-best model, respectively. All numbers are weighted f1-scores computed over 5 random initializations (mean and standard deviation shown). The unweighted average recall is also shown in brackets.

CARE and the baseline systems, utilize the same classification backend. Thus, the design allows fair comparison of the different representations. [1]

### E. Performance of CARE

The results on the 8 downstream datasets using representations from the proposed CARE model are shown in Table II. These baseline models are categorized into two groups based on the number of parameters used during inference: base models (parameter size $< 200$M), and large models ( $> 500$M), which also include LLM based models. The following observations are made for each category:

*1) Base Models:* We compare HuBERT [15], WavLM [34], data2vec [60] and emotion2vec [21] representations as the baseline models in this category. Among these baseline systems, the emotion2vec is also pre-trained on IEMOCAP and MELD datasets, partially explaining the improved results seen on the downstream tasks on these datasets. While CARE performs similar to emotion2vec on CMU-MOSI, it improves over all the base-sized models on other datasets. On the average, the proposed CARE achieves a relative improvement of 15.6% over the best baseline model (HuBERT).

*2) Large Models:* SONAR [31] is selected as the speech encoder in this category. For the six English-based datasets, the pre-trained English speech encoder[2] is used, while the French and German speech encoders are utilized for the CaFE and EmoDB datasets, respectively. Similar to the CARE backend, the layer representations from the SONAR encoder are linearly combined and the classification head is trained on the downstream task. Although SONAR has nearly four times the parameter size of CARE, our proposed model outperforms SONAR across all datasets except CMU-MOSI.

*3) LLM Based Models:* Two versions of SALMONN [32] (7B and 13B)[3] are considered as examples of LLM-based models. These are typically applied in a zero-shot setting; however, due to variability in emotion classes across datasets, their zero-shot performance is inconsistent. E.g. while SALMONN-13B model achieves 68.75% weighted F1-score on the IEMOCAP-4 dataset (on which it is trained), it achieves only 24.06% for MELD. Thus, for fair comparison, the same framework used in CARE and other baseline models is followed for the LLM based evaluations as well. The internal representations from all layers (41 layers for SALMONN 13B and 33 layers for SALMONN 7B) are aggregated using a convex combination, and the classification head (similar to CARE) is trained for each downstream dataset. Similar to emotion2vec, SALMONN includes IEMOCAP in its pre-training, leading to superior performance on IEMOCAP-4 and IEMOCAP-6 compared to CARE. The larger model size and extensive pre-training data allows SALMONN to outperform CARE by 10% and 34% (relative improvements) on the MELD and CMU-MOSI datasets, respectively. However, on the remaining four tasks, CARE surpasses the SALMONN models, achieving relative improvements of 17% and 24% on the RAVDESS-song and CaFE datasets, respectively. Notably, though music datasets are used to pre-train SALMONN, CARE emerges as the best model on the RAVDESS-Song dataset.

*Key takeaways:* 1) On average, CARE emerges as the top-performing model across the eight datasets, surpassing even the SALMONN 13B model, which has nearly 80 times more parameters. Although LLM-based models show strengths in in-domain emotion recognition datasets, their performance declines on out-of-domain tasks, indicating limited generalizability across diverse tasks and multilingual emotional speech. 2) CARE's advantage over speech SSL models like WavLM, HuBERT, and data2vec is expected, given that these models are trained on non-emotional data (see Sec. V-G for a related experiment). 3) Notably, CARE outperforms the multilingual SONAR model on CaFE and EmoDB datasets although it is trained on English speech only. This showcases the generalizability of our pre-training technique to out-of-domain tasks in SER.

### F. Emotional Attribute Prediction

The emotion recognition can be posed as a regression problem, where valence, arousal and dominance of a particular utterance are predicted [61]. We use the MSP-IMPROV [62] for this purpose. This is an audio-visual dataset that consists of 12 actors eliciting a set of sentences in different emotions. The

[1] Code available at https://github.com/iiscleap/CARE.
[2] https://dl.fbaipublicfiles.com/SONAR/spenc.eng.pt
[3] https://huggingface.co/tsinghua-ee/SALMONN

TABLE III
RESULTS FOR THE MSP-IMPROV DATASET

| Method | CCC-V | CCC-A | CCC-D |
|---|---|---|---|
| WavLM-base [16] | 0.51 | 0.64 | 0.47 |
| emotion2vec [21] | 0.5 | 0.61 | 0.49 |
| SALMONN-7B [32] | 0.53 | **0.67** | 0.52 |
| SALMONN-13B [32] | 0.56 | 0.65 | 0.51 |
| CARE | **0.57** | 0.66 | **0.53** |

CCC stands for concordance correlation coefficient while V, A, D stand for valence, arousal and dominance respectively.

dataset consists of 8438 utterances with valence, arousal and dominance values (ranging from 1 to 5). We split the dataset in 12 parts, where each part contains utterances corresponding to 10 training speakers, while speech from the two other speakers are used for validating and testing the model. The performance is measured as the average over these 12 parts.

We use the concordance correlation coefficient (CCC) as the metric. Denoting the mean, variance of ground truth by $\mu_g$, $\sigma_g^2$ and predicted scores by $\mu_p$, $\sigma_p^2$, the CCC is defined as

$$CCC = \frac{2\rho\sigma_g\sigma_p}{\sigma_p^2 + \sigma_g^2 + (\mu_g - \mu_p)^2} \tag{5}$$

In Eq. 5, $\rho$ is the Pearson's correlation coefficient between the ground truth and the predicted scores. For training the downstream model, the representations from CARE and other models are aggregated similar to the categorical datasets. This is followed by a two-layer regression head with 256 as the hidden dimension and 3 as the output dimension (1 for each of the three attributes). The objective is to increase the CCC between the ground truth and the predicted values for each of the dimensions of valence, arousal and dominance. The results for this dataset along with other baseline models are shown in Table III. We note that for this task, the CARE embeddings achieve the best results in terms of the valence and dominance attributes, while the performance on arousal is marginally better for the SALMONN-7B model.

## V. DISCUSSION

### A. Comparison With Baselines

Four baseline systems (Table IV) are considered:-

*PASE+:* For each downstream dataset, PASE+ features are extracted and a classification network is trained to predict the emotion class of each utterance similar to CARE. The total number of parameters used during inference is 8 M.

*Whisper:* For each downstream dataset, the representations from the 33 encoder layers of the Whisper-large-v3 model [36] are linearly combined with learnable weights. A two-layer classification head is trained on top of these representations for the task of emotion recognition. The total number of parameters used during inference is 800 M.

*Whisper+RoBERTa:* The transcripts are generated using the Whisper-large-v3 model and subsequently processed by a pre-trained RoBERTa model. The internal representations from RoBERTa are linearly combined by learnable weights, followed by training a two-layer classification head. This has a total of 1.6B parameters in use during inference.

*Teacher-fusion:* The PASE+ and Whisper+RoBERTa representations are concatenated and a two-layer classification head is trained for each downstream dataset. This baseline also has a total of 1.6B parameters during inference.

*Key takeaways:* 1) The performance of CARE surpasses that of the acoustic supervisory signal by 29.76% (relative) on average across the 8 datasets. This improvement can be attributed to the larger parameter size of CARE compared to the PASE+ model. 2) CARE is seen to outperform Whisper and Whisper+RoBERTa systems by 41.79% and 41.18% in relative terms. This indicates that, although the Whisper-based baselines are much larger in size, the combination of the acoustic and semantic information in CARE results in effective emotion recognition. 3) On MELD and CMU-MOSI, CARE is outperformed by the Whisper+RoBERTa baseline. For these datasets, text-based models are known to significantly outperform speech-only systems [19], [64]. In the Whisper+RoBERTa setup, the RoBERTa model is fine-tuned on transcripts generated by Whisper-large-v3 (1.6B sized model). In contrast, CARE is a smaller model (160 M), and does not use directly use the ASR transcripts during inference. To further elucidate the fairness in model-size, we replace Whisper-large-v3 with a Whisper-base model for the ASR, followed by the RoBERTa modeling. Then, the performance drops from 49.29% to 46.02% on MELD and from 75.14% to 71.91% on CMU-MOSI. This underscores the importance of accurate transcriptions and large model capacity in settings where the textual information is emotion rich. 4) While the teacher-fusion baseline is competitive for a number of datasets involving English speech, CARE outperforms this baseline on average by 5.24% absolute. This also motivates why CARE was pre-trained using knowledge distillation as it outperforms the fusion baseline with only 10% of the parameters.

### B. Importance of the Two Encoders

We present the performance of CARE when we use only one of acoustic and semantic encoders along with the common encoder for the downstream datasets in Table IV. For evaluating the combination of the semantic and common encoders, we use the 768-dimensional representations from the convolutional feature extractor, the common encoder, and the semantic encoder, excluding outputs from the acoustic encoder. Similarly, the semantic encoder representations are disregarded during the evaluation of the acoustic-common encoder combination. Note that, while CARE has more number of parameters (160 M) as compared to models like WavLM or emotion2vec, both these combinations have similar number of parameters during inference. While the semantic-common encoder combination has an inference time parameter size of 110 M, the acoustic-common encoder has a total of 94 M parameters during evaluation on each downstream dataset.

*Key takeaways:* 1) The combination of the acoustic and common encoder representations outperforms the best performing

TABLE IV
BASELINE RESULTS ON THE DIFFERENT DOWNSTREAM DATASETS IN TERMS OF WEIGHTED F1-SCORE

| Datasets | PASE+ [23]<br>**Params**:8M | Whisper [36]<br>**Params**:800M | Whisper [36]+<br>RoBERTa [35]<br>**Params**:1.6B | Teacher-fusion<br>**Params**:1.6B | Semantic +<br>Common Enc.<br>**Params**:110M | Acoustic +<br>Common Enc.<br>**Params**:94M | CARE<br>**Params**:160M |
|---|---|---|---|---|---|---|---|
| IEMOCAP-4 [45] | 56.68 | 56.40 | 61.97 | 69.49 | 66.44 | 65.91 | **69.39** |
| IEMOCAP-6 [45] | 41.38 | 40.62 | 49.28 | 56.61 | 53.05 | 52.09 | **55.02** |
| MELD [46] | 35.86 | 40.11 | **49.29** | 49.72 | 47.37 | 46.98 | 48.05 |
| CMU-MOSI [47] | 50.69 | 55.60 | **75.14** | 74.12 | 64.23 | 64.17 | 66.74 |
| DAIC-WOZ [48], [63] | 66.84 | 62.08 | 64.04 | 67.63 | 66.32 | 66.89 | **68.49** |
| RAVDESS-Song [49] | 46.05 | 34.20 | 9.58 | 48.48 | 55.23 | 56.17 | **60.11** |
| CaFE [50] | 52.86 | 19.22 | 13.59 | 53.42 | 69.23 | 71.62 | **76.98** |
| EmoDB [51] | 62.59 | 24.77 | 14.98 | 66.75 | 75.42 | 78.63 | **83.41** |
| Avg. | 51.62 | 41.63 | 42.23 | 60.78 | 62.16 | 62.81 | **66.02** |

All numbers are averaged over 5 random initializations of the downstream network. We also show the results of the different components of CARE in this table.

SSL model (HuBERT) by 8.08% (relative) on average for the 8 datasets (Table II). Given the similar parameter count, this performance suggests an advantage of our pre-training approach. 2) For the three out-of-domain datasets, the acoustic-common combination fares better than its semantic counterpart. 3) Across all datasets, the combination of both encoders in CARE yields the highest performance, suggesting that while the individual encoder performances are comparable, they capture distinct characteristics of the speech signal.

### C. Modifications in the Semantic Encoder

To evaluate the suitability of our design choices for the semantic encoder, we made three architectural modifications:

1) *CARE-No init.:* Removing the convolutional adapters, the transformer layers in the semantic encoder are initialized randomly (instead of pre-trained RoBERTa weights).
2) *CARE-Trans.:* Removing the convolutional adapters while the RoBERTa transformer layers are updated.
3) *CARE-FT:* Keeping the convolutional adapters, we update all the parameters (conv. adapters and transformer weights) in the semantic encoder.

The results for these modifications are shown in Table V .

*Key takeaways:* 1) Initializing the transformer weights with RoBERTa is essential for CARE's performance. Random initialization of the semantic encoder leads to performance drops across all five datasets, suggesting that a randomly initialized semantic encoder struggles to regress to the semantic supervisory signal. 2) Removing convolutional adapters (in CARE-Trans) negatively impacts performance, highlighting the necessity of our convolution-based adaptation technique for aligning speech representations with RoBERTa's transformer layers. 3) Updating the transformer layers in the semantic encoder decreases performance. Since RoBERTa is pre-trained on text, fine-tuning with speech data degrades its effectiveness.

TABLE V
RESULTS ON THE DOWNSTREAM DATASETS (WEIGHTED F1-SCORE) WITH MODIFICATIONS IN THE SEMANTIC ENCODER

| Dataset | Method | WF1 |
|---|---|---|
| IEMOCAP(4-class) | CARE-No init. | 65.71 |
| | CARE-Trans. | 65.89 |
| | CARE-FT | 68.16 |
| | CARE | **69.39** |
| IEMOCAP(6-class) | CARE-No init. | 50.96 |
| | CARE-Trans. | 51.19 |
| | CARE-FT | 52.37 |
| | CARE | **55.02** |
| MELD | CARE-No init. | 45.16 |
| | CARE-Trans. | 46.57 |
| | CARE-FT | 47.93 |
| | CARE | **48.05** |
| CMU-MOSI | CARE-No init. | 62.16 |
| | CARE-Trans. | 65.97 |
| | CARE-FT | 64.27 |
| | CARE | **66.74** |
| DAIC-WOZ | CARE-No init. | 64.57 |
| | CARE-Trans. | 65.93 |
| | CARE-FT | 67.43 |
| | CARE | **68.49** |

TABLE VI
RESULTS ON THE DOWNSTREAM DATASETS (WEIGHTED F1-SCORE) WITH DIFFERENT INITIALIZATIONS OF THE ACOUSTIC ENCODERS

| Datasets | Random init. | HuBERT init. | Data2vec init. | WavLM init. |
|---|---|---|---|---|
| IEMOCAP-4 | 66.72 | 67.65 | 66.76 | **69.39** |
| IEMOCAP-6 | 51.47 | 51.40 | 52.98 | **55.02** |
| MELD | 46.41 | 46.97 | 47.19 | **48.05** |
| CMU-MOSI | 65.07 | 66.26 | **68.14** | 66.74 |
| DAIC-WOZ | 67.56 | 65.19 | 66.59 | **68.49** |
| RAVDESS-Song | 57.81 | **60.30** | 55.09 | 60.11 |
| CaFE | 73.83 | 72.23 | 62.46 | **76.98** |
| EmoDB | 78.61 | **86.51** | 77.19 | 83.41 |
| Avg. | 63.44 | 64.56 | 62.05 | 66.02 |

### D. Initialization of Acoustic and Common Encoders

As indicated in Section III, the common and the acoustic encoders of CARE are initialized with the WavLM-base model weights. We present the results of our method when this initialization is modified to i) random, ii) HuBERT-base [15] or iii) data2vec-base [60] (Table VI).

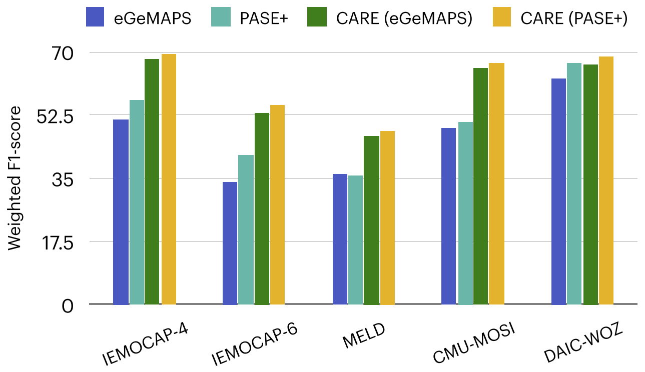

Fig. 3. Performance of CARE when different acoustic targets are used. The model with eGeMAPS as features is trained similarly to that of the PASE+ baseline. All numbers are shown as the average of 5 random initializations.

*Key takeaways:* 1) The model's performance decreases with data2vec initialization, likely due to data2vec's lower baseline performance compared to HuBERT and WavLM (see Table II). An exception is the CMU-MOSI dataset, where this initialization improves over the WavLM initialized model by 4.21% (relative). 2) The HuBERT-initialized model performs best on the RAVDESS-Song and Emo-DB datasets. Notably, HuBERT outperforms WavLM for these two out-of-domain datasets (Table II). 3) Initialization impacts the acoustic and common encoders less than the semantic encoder, as the latter requires alignment with text representations.

### E. Choice of Acoustic Targets

We run an experiment where the acoustic encoder is trained with targets based on eGeMAPS [9] features extracted from the openSMILE toolkit [65]. The PASE+ targets of the acoustic encoder of the CARE model is replaced by the eGeMAPS features. The performance of this model, called CARE (eGeMAPS), is shown in Fig. 3.

*Key takeaway:* The baseline model using eGeMAPS input features performs worse than the baseline with PASE+ features, as expected, since eGeMAPS are handcrafted. Consequently, the average performance of CARE with eGeMAPS is also lower than that of CARE with PASE+ targets.

### F. Choice of Semantic Targets

We conduct an experiment where the Whisper encoder representations serve as supervisory signals for the semantic encoder. We explore two variants of this: 1) We pool the Whisper representations to serve as semantic targets while pre-training. This model is called CARE (Whisper-pool). 2) We pre-train a model with the frame-level representations of Whisper as the targets. We call this model CARE (Whisper-frame). 3) We also use the frame level alignments between speech and the RoBERTa tokens and use the frame-level RoBERTa representations as the semantic targets. We call this model CARE (RoBERTa-frame). The comparative performances of the different variants along with the proposed model-CARE (RoBERTa-pool) are shown in Fig. 4.

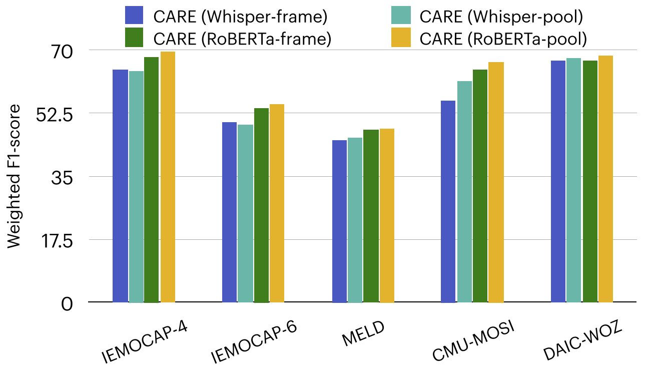

Fig. 4. Performance of CARE when different semantic targets are used. All numbers are shown as the average of 5 random initializations.

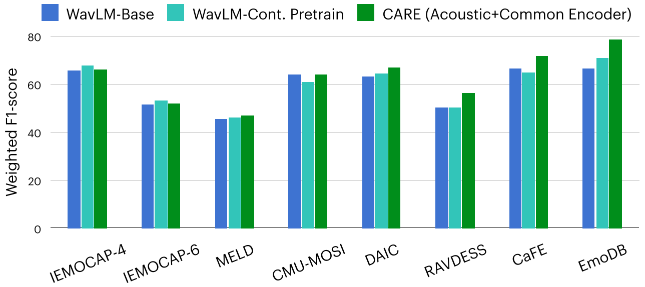

Fig. 5. Comparison of the performance when WavLM is continually pre-trained on MSP-PODCAST. The performance of the combination of acoustic and common encoders of CARE is shown for reference. Here, RAVDESS refers to the RAVDESS-Song dataset.

*Key takeaways:* 1) The performance of CARE (RoBERTa-pool) is seen to be superior to both variants trained with Whisper encoded representations. 2) The performance of the systems when the semantic encoder is trained with the pooled targets is observed to be better than those trained with frame-level representations.

### G. Continued Pre-Training of WavLM

Since all self-supervised learning (SSL) models are trained on neutral data, their ability to accurately discern emotions from speech signals is typically limited. The emotion recognition performance of these SSL models when pre-trained on emotion datasets thus becomes crucial. To explore the impact of pre-training setup in the proposed CARE, we continued the pre-training of the publicly available WavLM-base model, using the MSP-PODCAST dataset. This was done following the WavLM pre-training procedure, with masked language modeling loss, for an additional 200,000 steps (similar to the CARE). The results of this experiment are shown in Fig. 5, wherein the performance of the continually pre-trained WavLM model is denoted by WavLM-Cont. Pretrain. The performance of the combination of the common and acoustic encoders is also shown for comparison.

*Key takeaway:* Continued pre-training improves WavLM-base performance on certain downstream tasks, like IEMOCAP.

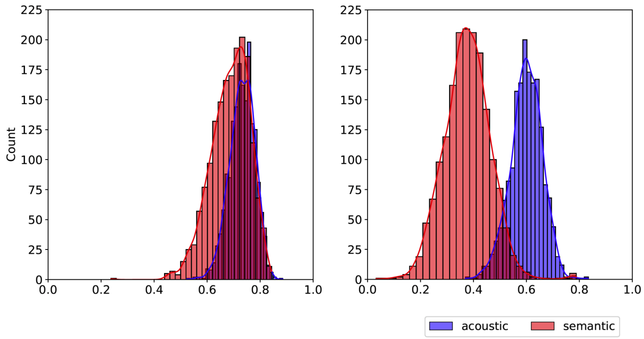

Fig. 6. Distribution of cosine similarities for layer 7 representations for the acoustic and the semantic encoders. The left plot is when two speech signals belonging to different speakers, same emotion and same content are processed by CARE. The plot on the right refers to the setting with different speakers, same emotion and different content.

However, except for IEMOCAP, WavLM performs worse than CARE's acoustic-common encoder combination. As all the models in Fig. 5 have the same size (94 M) during inference, this experiment highlights the benefits of our proposed distillation-based pre-training.

### H. Multimodal Emotion Recognition

To assess CARE's utility in the multimodal speech-text setting, we design a model using speech (WavLM-base or CARE), and text (RoBERTa) fusion. After combining the layer representations in SUPERB style, we concatenate the representations and train a classification head. We experiment on IEMOCAP-4 and IEMOCAP-6 and observe that the weighted F1-score for CARE+RoBERTa improves from 73.02% to 75.19% for IEMOCAP-4 and from 60.41% to 62.21% for IEMOCAP-6 over WavLM-base+RoBERTa system. In addition to the uni-modal improvements reported in Table IV, these results highlight that multi-modal speech-text emotion recognition systems can also benefit from the enhanced representations provided by CARE.

### I. Visualization of Layers of CARE

In order to interpret the pre-trained model representations learnt by the acoustic and semantic encoder of CARE, we probe the representations from each encoder. We use the English part of the Emotional Speech Dataset (ESD) [66] for this analysis. We form pairs of utterances, where both the utterances of a pair have the same emotional label in all cases and they are derived from two different speakers. A total of 1750 pairs are considered and the cosine similarities of the pooled representations (for transformer layer 7) are shown in Fig. 6. The figure on the left indicates the setting where the speech content in the two utterances is the same whereas the plot on the right indicates different speech content. We note that when the spoken content is different, the acoustic encoder has higher similarity than the semantic encoder, indicating that the acoustic encoder is beneficial when the emotion information cannot be reliably predicted from the textual content of the audio.

## VI. Summary

*Key Highlights:* In this paper, a pre-training technique for content and acoustic encoding of emotional speech is provided. The proposed architecture, termed CARE, learns an enriched representation of acoustic and semantic information. The acoustic encoder uses supervision from low-level descriptors of speech, while the semantic encoder is distilled using text representations of the speech transcripts. We also propose an adaptation strategy for text-based models in speech representation learning using convolutional neural network layers. The CARE model, with experiments on 8 downstream tasks, is seen to outperform models of comparable sizes on most of the datasets. Further, the CARE is also observed to generalize better than LLM based models with large parameter sizes. The importance of the different components of the proposed model, along with the different design choices, are established through ablation studies.

*Limitations and future scope:* The MSP-PODCAST dataset is used for pre-training CARE, which has only 230 hours of emotional speech data. Another limitation of this work, is the relatively lower performance on some in-domain speech datasets compared to LLM-based models, like the CMU-MOSI. In future, we plan to extend the CARE approach to multi-modal speech-text emotion recognition tasks.

## References

[1] M. Pantic et al., "Affective multimodal human-computer interaction," in *Proc. ACM Int. Conf. Multimedia*, 2005, pp. 669–676.
[2] B. Gaind, V. Syal, and S. Padgalwar, "Emotion detection and analysis on social media," 2019, *arXiv:1901.08458*.
[3] B. Li, D. Dimitriadis, and A. Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 5876–5880.
[4] S. Ghosh, S. Sahu, N. Ganguly, B. Mitra, and P. De, "EmoKey: An emotion-aware smartphone keyboard for mental health monitoring," in *Proc. 11th Int. Conf. Commun. Syst. Netw.*, 2019, pp. 496–499.
[5] P. Lieberman and S. B Michaels, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech," *J. Acoustical Soc. Amer.*, vol. 34, no. 7, pp. 922–927, 1962.
[6] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proc. Artif. Neural Netw. Eng.*, 1999, vol. 710, p. 22.
[7] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.
[8] B. Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, 2013, pp. 148–152.
[9] F. Eyben et al., "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.
[10] P. Yenigalla et al., "Speech emotion recognition using spectrogram & phoneme embedding," in *Proc. Interspeech*, 2018, pp. 3688–3692.
[11] S. Dutta and S. Ganapathy, "Multimodal transformer with learnable frontend and self attention for emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6917–6921.
[12] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2526–2530.
[13] C. S. A. Kumar et al., "Speech emotion recognition using CNN-LSTM and vision transformer," in *Proc. Int. Conf. Innov. Bio-Inspired Comput. Appl.*, 2022, pp. 86–97.
[14] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *in Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12449–12460.
[15] 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 Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[16] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[17] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[18] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[19] S. Dutta and S. Ganapathy, "HCAM–Hierarchical cross attention model for multi-modal emotion recognition," 2023, *arXiv:2304.06910*.

[20] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1711–1724, Jul.–Sep. 2024.

[21] Z. Ma et al., "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Find. ACL*, 2024, pp. 15747–15760.

[22] Y. Li et al., "ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition," in *Proc. Interspeech*, 2023, pp. 1449–1453.

[23] M. Ravanelli et al., "Multi-task self-supervised learning for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6989–6993.

[24] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 1021–1028.

[25] P. Agrawal and S. Ganapathy, "Interpretable representation learning for speech and audio signals based on relevance weighting," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2823–2836, 2020.

[26] N. Zeghidour et al., "LEAF: A learnable frontend for audio classification," in *Proc. Int. Conf. Learn. Representations*, 2021.

[27] Santiago Pascual et al., "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proc. Interspeech*, 2019, pp. 161–165.

[28] R. Schluter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. IV-649–IV-652.

[29] Q. Fan et al., "Leveraging contrastive learning and self-training for multimodal emotion recognition with limited labeled samples," in *Proc. 2nd Int. Workshop Multimodal Responsible Affect. Comput.*, 2024, pp. 72–77.

[30] G. Hu et al., "Recent trends of multimodal affective computing: A survey from NLP perspective," 2024, *arXiv:2409.07388*.

[31] P.-A. Duquenne, H. Schwenk, and B. Sagot, "SONAR: Sentence-level multimodal and language-agnostic representations," 2023, *arXiv:2308.11466*.

[32] C. Tang et al., "SALMONN: Towards generic hearing abilities for large language models," in *Proc. 12th Int. Conf. Learn. Representations*, 2024.

[33] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[34] S. Hu et al., "WavLLM: Towards robust and adaptive speech large language model," in *Proc. Find. Empirical Methods Natural Lang. Process.*, 2024, pp. 4552–4572.

[35] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

[36] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.

[37] S. Fan et al., "Sentiment-aware word and sentence level pre-training for sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 4984–4994.

[38] V. Krishna and S. Ganapathy, "Towards the next frontier in speech representation learning using disentanglement," 2024, *arXiv:2407.02543*.

[39] W. Yu et al., "Connecting speech encoder and large language model for ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 12637–12641.

[40] Y. Fathullah et al., "Prompting large language models with speech recognition abilities," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 13351–13355.

[41] Y. Li et al., "Evaluating parameter-efficient transfer learning approaches on SURE benchmark for speech understanding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[42] K. Kim et al., "Convolution-augmented parameter-efficient fine-tuning for speech recognition," in *Proc. Interspeech*, 2024, pp. 2830–2834.

[43] S. wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[44] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct.–Dec. 2019.

[45] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[46] S. Poria et al., "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[47] A. Zadeh et al., "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.

[48] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews," in *Proc. LREC*, 2014, pp. 3123–3128.

[49] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS One*, vol. 13, no. 5, 2018, Art. no. e0196391.

[50] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 399–402.

[51] F. Burkhardt et al., "A database of German emotional speech," in *Proc. Interspeech*, 2005, vol. 5, pp. 1517–1520.

[52] N. Majumder et al., "Dialoguernn: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 6818–6825.

[53] S. Poria et al., "Context-dependent sentiment analysis in user-generated videos," in *Proc. Assoc. Comput. Linguistics (Volume 1: Long papers)*, 2017, pp. 873–883.

[54] W. Wu, C. Zhang, and P. C Woodland, "Self-supervised representations in speech-based depression detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[55] Y. Shen, H. Yang, and L. Lin, "Automatic depression detection: An emotional audio-textual corpus and a Gru/Bilstm-based model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6247–6251.

[56] V. Ravi et al., "A step towards preserving speakers' identity while detecting depression via speaker disentanglement," in *Proc. Interspeech*, 2022, vol. 2022, Art. no. 3338.

[57] W. Wu, M. Wu, and K. Yu, "Climate and weather: Inspecting depression detection via emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6262–6266.

[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.

[59] V. Nair and G. E Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[60] A. Baevski et al., "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1298–1312.

[61] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Interspeech*, 2017, vol. 2017, pp. 1103–1107.

[62] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan.–Mar. 2017.

[63] D. DeVault et al., "SimSensei kiosk: A virtual human interviewer for healthcare decision support," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2014, pp. 1061–1068.

[64] Z. Lian, B. Liu, and J. Tao, "SMIN: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2415–2429, Jul.–Sep. 2023.

[65] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[66] K. Zhou et al., "Emotional voice conversion: Theory, databases and ESD," *Speech Commun.*, vol. 137, pp. 1–18, 2022.

**Soumya Dutta** (Graduate Student Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Engineering Science and Technology (IIEST), Shibpur, India, in 2015, and the M.Tech. degree in control and computing from the Indian Institute of Technology (IIT) Bombay, India, in 2018. He is currently working toward the Ph.D. degree with the Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science (IISc), Bangalore, India. From 2018 to 2021, he was a Data Scientist with IBM India Pvt. Ltd. His research interests include multimodal emotion recognition, speech synthesis, and multimodal large language models. He is the recipient of the Prime Minister's Research Fellowship (PMRF) in 2022 and the Qualcomm Innovation Fellowship (QIF) in 2025.

**Sriram Ganapathy** (Senior Member, IEEE) received the B.Tech. degree from the College of Engineering, Trivandrum, India, the M.E. degree from IISc, Bangalore, and the Ph.D. degree from the Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA, He is currently an Associate Professor with the Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India, where he heads the Learning and Extraction of Acoustic Patterns (LEAP) Lab. From 2022 to 2024, he was a Visiting Research Scientist with Google DeepMind, Bangalore. From 2011 to 2015, he was a Research Staff Member with the IBM Watson Research Center, Yorktown Heights, NY, USA. He has also held research positions with the Idiap Research Institute, Switzerland. He is a member of the Editorial Board of Elsevier *Speech Communication* and Nature Scientific Data. His research interests include signal processing, speech processing, machine learning, deep learning, and auditory neuroscience.