Describe Where You Are: Improving Noise-Robustness for Speech Emotion Recognition with Text Description of the Environment

Seong-Gyun Leem, *Student Member, IEEE,* Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso, *Fellow, IEEE*

***Abstract*—*Speech emotion recognition* (SER) systems often struggle in real-world environments, where ambient noise severely degrades their performance. This paper explores a novel approach that exploits prior knowledge of testing environments to maximize SER performance under noisy conditions. To address this task, we propose a text-guided, environment-aware training where an SER model is trained with contaminated speech samples and their paired noise description. We use a pre-trained text encoder to extract the text-based environment embedding and then fuse it to a transformer-based SER model during training and inference. We demonstrate the effectiveness of our approach through our experiment with the MSP-Podcast corpus and real-world additive noise samples collected from the Freesound and DEMAND repositories. Our experiment indicates that the text-based environment descriptions processed by a *large language model* (LLM) produce representations that improve the noise-robustness of the SER system. With a *contrastive learning* (CL)-based representation, our proposed method can be improved by jointly fine-tuning the text encoder with the emotion recognition model. Under the -5dB *signal-to-noise ratio* (SNR) level, fine-tuning the text encoder improves our CL-based representation method by 76.4% (arousal), 100.0% (dominance), and 27.7% (valence).**

***Index Terms*—Speech emotion recognition, noise-robustness, text-guided training, multi-modal**

I. Introduction

SPEECH *emotion recogntion* (SER) systems have highly improved with the help of pre-trained speech representation models [1]–[3] and the creation of larger emotional speech databases [4]–[7]. Recently, there has been increased interest in deploying SER systems in real-world applications, opening opportunities across many domains, such as digital assistants [8], health care applications [9], and security and defense. One important barrier in this direction is the degradation of SER performance in real-world environments caused by multiple types of non-stationary background noise [10].

This study is supported by NIH under grant 1R01MH122367-01.

Seong-Gyun Leem is with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: SeongGyun.Leem@utdallas.edu).

Daniel Fulford is with the Occupational Therapy and Psychological and Brain Sciences, Boston University, MA 02215 USA (e-mail: dfulford@bu.edu).

Jukka-Pekka Onnela is with the Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, MA 02138 USA (e-mail: onnela@hsph.harvard.edu).

David Gard is with the Psychology Department, San Francisco State University, CA 94132 USA (e-mail: dgard@sfsu.edu).

Carlos Busso is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: busso@cmu.edu).

Several solutions have been proposed to improve the robustness of SER systems against acoustic noise. The solutions include data augmentation [11]–[13], feature enhancement [14], [15], feature selection [16], [17], and domain adaptation approaches [18], [19]. Since transformer-based speech representation models have been successfully used in speech problems [1]–[3], many studies have also worked on increasing the noise robustness of SER systems built with pre-trained speech representation models [20], [21]. These approaches can increase the performance of transformer-based SER models in target noisy conditions. However, it is challenging to use these models in scenarios with multiple noisy environments since a transformer-based SER model requires important resources to adapt and store its parameters for each target environment. To address multiple noise types in a single SER model, Leem et al. [22] proposed environment-agnostic and -specific adapters. Their work showed that leveraging the prior knowledge of the testing condition is important for an SER model's adaptation to multiple noisy environments.

This paper explores which form of prior knowledge allows an SER model to effectively adapt to multiple unseen environments. Rather than aiming to cover all the environments, our system trained the model to be conditioned by a text embedding describing the environment, which project the unseen condition into the ones that are the closest to the seen environments. With this strategy, the prior knowledge is used as a mechanism for zero-shot learning in new environments with types of noises not considered while training the models. It also provides the mechanism to indirectly identify similar environmental conditions during training (e.g., noise in a bus station and a train station). Exploring this problem, we investigate using text-based environment descriptions as the prior knowledge for a noise-robust SER system. Using natural language prompts during training has shown potential in image classification [23], sound event classification [24], and several speech processing downstream tasks, including keyword spotting, and speaker counting [25]. Natural language supervision is also applicable to SER tasks [26], [27]. All these studies indicate that exploiting text information is a promising strategy for SER systems. We propose a *text-guided environment-aware training* (TG-EAT) strategy to improve the noise robustness of an SER model with text descriptions. We focus on the prediction of arousal (calm to active), valence (negative to positive), and dominance (weak to strong). TG-EAT uses noisy speech and its text-based environmental description to adapt

the SER model. We use a pre-trained text encoder to extract the representation of text-based environment descriptions. This representation is combined with a transformer-based SER model. During adaptation, the SER model learns appropriate denoising functions with respect to the given environment description. During inference, we only need to change the template sentence to guide the SER model with testing environment information. We expect that the pre-trained text encoder can capture similar semantic information from environmental conditions included in the train set, allowing zero-shot environment learning for the SER model. This approach is expected to generalize the SER performance when tested in environmental conditions that are not included in the training process.

Our experiment with the MSP-Podcast corpus shows that using text descriptions of the testing environment can highly improve the SER performance, especially with *large language model* (LLM). In the -5 dB *signal-to-noise ratio* (SNR) condition, our method improves the original SER model built with a *self-supervised learning* (SSL) representation by 7.6% for arousal, 8.3% for dominance, and 45.4% for valence. When we compare the proposed SER model with the DAT baseline, we observe improvements of 16.6% for arousal, 18.1% for dominance, and 23.0% for valence (-5 dB SNR level). With the text encoder from CLAP, pre-trained with paired audio, the SER model can achieve the best performance in the low SNR condition. Compared to freezing the text encoder, the fine-tuning approach improves performance by 76.4% for arousal, 100.0% for dominance, and 27.7% for valence under the -5 dB SNR condition. Our solution is highly applicable to SER systems deployed in real-world applications. For example, systems can infer the testing environment from a *global positioning system* (GPS) by using *geological information service* (GIS) mashups, such as OpenStreetMap [28]. The main contributions of this study are:

- We explore using text embedding for an SER model to increase noise robustness in unseen conditions by explicitly leveraging the environment information. Our method provides a unique advantage by enabling a single model to adapt to multiple noise conditions using text embeddings rather than requiring multiple context-specific expert models. This is particularly beneficial for transformer-based architectures, which demand significant computational resources while delivering SOTA performance for SER. By leveraging text-described target environment information, we maximize performance without the overhead of maintaining multiple models.
- We show the benefits of using LLM to improve SER performance under noisy conditions over using a pre-trained environment classifier, especially in a low SNR condition.
- We show that fine-tuning the text encoder of CLAP can improve SER performance, leading to the possibility of using a paired audio encoder to deal with unknown testing environments.

Our paper is organized as follows. Section II describes studies relevant to SER in noisy conditions and text-guided training strategies. Section III describes the proposed approach, emphasizing the motivations and insights behind the TG-EAT framework. Section IV provides the experimental setting, including the database, baselines, and implementation details. Section V presents the results, discussing the clear benefits of the proposed strategy. Finally, Section VI concludes the paper, summarizing our study and providing future research directions inspired by the proposed approach.

II. Previous Work

A. Speech Emotion Recognition under Noisy Environments

Increasing the noise robustness of an SER system is an essential task when deploying it in real-world applications. Previous studies have mainly focused on improving acoustic features for the SER model. Triantafyllopoulos et al. [15] proposed to enhance noisy waveforms before extracting the SER features. The enhancement models used *convolutional neural network* (CNN) with residual blocks. Pandharipande et al. [29] proposed to discard noisy frames to increase the noise robustness of an SER model by using a voice activity detection module. Leem et al. [30] proposed to select noise-robust LLDs by addressing the performance and robustness of each single LLD.

More recently, SER studies have mainly focused on using transformer-based speech representation models [31]–[36], including Wav2Vec2.0 [1], HuBERT [2], and WavLM [3]. Such models have shown higher robustness against small perturbations on the input speech than the traditional SER model with a Mel-spectrogram [33]. Despite this trend, they still show performance differences from the ones tested in a clean environment. For this reason, studies are currently exploring strategies to improve the noise robustness of the pre-trained speech representation model. A common approach to address this issue is noise-aware training, where the clean training set is augmented with the noise sound during environment adaptation. Mitra et al. [20] demonstrated that training a HuBERT-based SER model with noisy speech can highly improve the performance in low SNR conditions. Leem et al. [21] proposed a contrastive teacher-student learning strategy to address the catastrophic forgetting issue when training a fine-tuned SER model with noisy speech. Wu et al. [12] proposed to dynamically change the distortion level of the augmented speech during adaptation based on the distortion metrics.

The aforementioned methods focused on increasing the SER model's robustness against a single target environment. They might not be the optimal solution for an SER model deployed on a real-world application since it is highly likely that this system will encounter multiple types of environmental noises. We focus on adapting a single transformer-based SER model to multiple noisy environments to efficiently deal with multiple types of environments. To address this issue, Leem et al. [22] proposed to adapt the transformer-based SER model to multiple types of noises with skip connection adapters. They not only trained the SER model with multiple environments but also focused on leveraging the environmental information of the testing conditions to improve SER performance under noisy conditions. The results showed that using the

environment-agnostic and -specific adapters with respect to the testing condition can improve the SER performance under noisy conditions. Such prior knowledge could be achieved using domain knowledge or GPS information. Their result showed that using environmental information during inference is important for a SER model to perform well under noisy conditions. This work indicates that leveraging the prior knowledge of the testing condition is also important for a noise-robust SER model, as well as training it with multiple types of noises. This is beneficial for an SER model deployed on real-world applications where the system can exploit the domain knowledge of the testing environment and the GPS information.

This paper also explores the multi-condition training approach where the fine-tuned SER model is adapted to multiple types of noise. Different from other methods, our strategy relies on a text embedding that describes the testing environment to deal with multiple unseen environments.

B. Text-Guided Training

As we discussed in Section II-A, exploiting environmental information can improve SER performance in a noisy environment. This paper mainly focuses on using text prompts to infuse environmental information into an SER model. Using natural language prompts does not require the recognition model to use a fixed set of predetermined labels during training. *Contrastive language-image pre-training* (CLIP) is a good example of this approach [23]. It consists of an image encoder and a text encoder, trained with pairs of images and their corresponding text descriptions. These encoders are trained in a contrastive learning manner, which maximizes the similarity of both representations if the image and the description are paired and minimizes the similarity if they are unpaired. After training, these encoders can perform zero-shot classification by checking the similarity between the given image and the candidate prompts. The study of Radford et al. [23] used the following prompt template: *"A photo of a {label}"*. They calculate the similarity between the representation from the given image and the representations from the prompts with different {*label*}, selecting the {*label*} that shows the maximum similarity.

The contrastive pre-training strategy with natural language supervision is also successful in universal audio and speech processing. Wu et al. [24] demonstrated that pre-training audio and text encoder with natural language guidance could improve audio classification performance. The study of Elizalde et al. [25] showed that such natural language guidance can improve speech processing tasks, including keyword spotting, speaker counting, and SER tasks.

Previous studies have found that natural language supervision can apply to SER tasks. Stanley et al. [26] used word embeddings to encode emotional labels for SER model. Gong et al. [27] used LLM to infer weak emotion labels for unlabeled data for weakly-supervised learning of an SER model. All these findings have shown that exploiting text information is highly applicable to SER systems. To the best knowledge of the authors, the use of natural language supervision to address SER robustness against unknown noisy environments is a novel research direction.

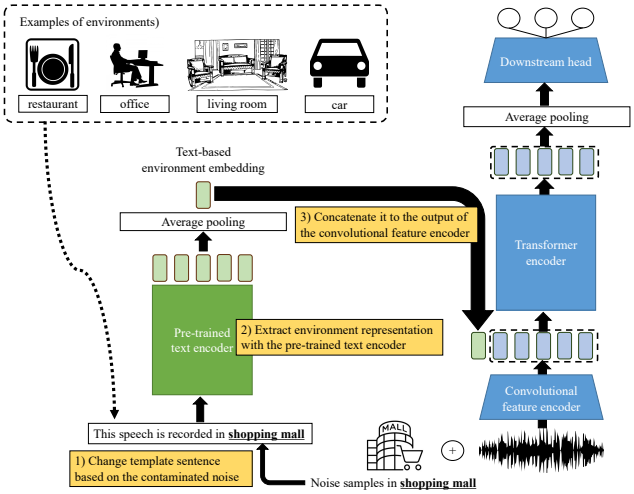


Fig. 1: Our proposed text-guided environment-aware training framework. The environment representation is concatenated with the output of the convolutional feature encoder.

III. Proposed Method

This paper proposes *text-guided environment-aware training* (TG-EAT), which leverages environmental information to improve an SER model in noisy conditions. Figure 1 illustrates our proposed TG-EAT framework, which uses a pair of noisy speech and its corresponding environmental description. The text embedding extracted from the environmental description is combined with the acoustic representation in the SER model, allowing it to improve the representation for the given environmental description.

The key contribution of this study is how we use the text description from the target environment. We used prompts to generate the text description where the target environment is changed. As a preliminary experiment, we tested different prompts to describe the target environment such as *"The type of background noise is {environment},"* or *"The input is recorded with a sound of {environment}."* We change {*environment*} in the prompts according to the target environment during training and testing. We found that all the prompts showed similar emotion recognition performance for all the attributes. Therefore, we consistently use the following prompt in this study: *"This speech is recorded in {environment}."* We extract the text-based environment embedding from this text description using a pre-trained text encoder. We test two different text representations: *contrastive learning* (CL)-based representation and LLM-based representation. For the CL-based representation, we use the text encoder pre-trained with the *contrastive language audio pre-training* (CLAP) strategy [24], [25]. CLAP consists of an audio encoder and a text encoder. It uses a pair of acoustic events and their text description during pre-training (e.g., *bird chirping sound* with the description, "Bird is chirping in the given audio"). With these audio-text pairs, the training objective is to maximize the similarity between the audio and text representation if

they are from the same pair and minimize it if they are from a different pair. Since CLAP uses an audio-text pair during pre-training, we assume that its text encoder can generate an appropriate representation from the given environment description coherent with the target acoustic condition. This paper uses the pre-trained text encoder from the unfused CLAP model proposed in the study of Wu et al. [24]. We take a 768-dimensional latent embedding from the text encoder, using it as our text-based environmental embedding. For the LLM-based representation, we use the encoder from the pre-trained RoBERTa model [37]. RoBERTa is pre-trained with *masked language modeling* (MLM) and *next sentence prediction* (NSP) tasks. RoBERTa has shown good performance in various benchmarks for evaluating natural language understanding systems, such as GLUE [38]. Although it is not pre-trained with audio data, we assume that its encoder can extract enriched semantic information from the given prompt. We use RoBERTa-large, which has 24 transformer layers. For each text encoder, we use the same tokenizer used in its pre-training to tokenize the text description of the environment. We extract token-level text embeddings from the tokenized prompt and then apply average pooling, resulting in a 1024-dimensional single representation vector for each prompt.

After the environmental representation is obtained, the next step is to introduce this information into the model. We mainly focus on a transformer-based SER model, which has shown good performance in SER tasks [33], [39]. An important task is to fine-tune the model with clean and emotional speech data. We first fine-tune the SER model with clean speech to maximize the *concordance correlation coefficient* (CCC) between the predicted and the ground-truth emotional attribute scores of arousal, dominance, and valence. After fine-tuning with clean speech, the SER model is continuously updated with the training set contaminated with multiple types of noise and their corresponding text description. We insert the text representation from the given environment description into the fine-tuned transformer-based SER model. We achieve this goal by combining the text embedding with the acoustic representation, which is the output of the convolutional encoder. We apply trainable linear projection to the text embedding to match its dimension to the acoustic representation embeddings. We concatenate the projected text embedding to the acoustic representation embeddings along the time axis, then feed them into the transformer encoder. We choose this approach to allow the self-attention module in the transformer encoder to attend to the text embedding to all acoustic representation embeddings. Previous studies have proposed alternative approaches to add text embeddings into a prediction model [40], [41], but we leave this research direction as our future investigation to further improve its performance. We update the transformer encoder and the downstream head with the concatenated embeddings. We use the same training objective as the one used when training with clean speech. From this framework, we want to evaluate if the SER model can learn the denoising function given a noisy acoustic representation with its text embedding.

TABLE I: Keywords that are used for contaminating *training*, *development*, and *testing* sets. *Freesound* illustrates the keywords that are used for crawling the ambient recordings from the Freesound repository. *DEMAND* illustrates the keywords paired with the recorded sounds in the DEMAND corpus.

Data Split	Corpus	Keywords
Training, Development	*Freesound*	mall, restaurant, office, airport, station, city, park, street, traffic, home, kitchen, living room, bathroom, bedroom, metro, bus, car, construction site, pedestrian, beach
Testing	*Freesound*	plaza, garden, school, tram, sea, boat, amusement park, aquarium, arcade, art gallery, backyard, balcony, bank, bar, barn, beach, bridge, cafe, campground, canyon, carnival, cave, cemetery, church, circus, classroom, creek, crowd, dessert, dock, elevator, exhibition hall, factory, fairground, farmyard, festival, field, forest, fountain, gallery, gas station, grocery store, gym, harbor, highway, hospital, hotel, ice rink, industrial site, jungle, lake, laundromat, library, lobby, machine shop, market, meadow, mountain, museum, nightclub, parade, parking lot, patio, pet store, playground, pub, river, rooftop, shopping center, stadium, subway, swimming pool, theater, valley, waiting room, warehouse, waterfall, wetland, workshop, yard
	DEMAND	washroom, kitchen, living room, sports field, river, park, office, hallway, meeting, subway station, cafeteria, restaurant, traffic intersection, town square, cafe terrace, subway, bus, car

IV. Experimental Settings

A. Data preparation

Our experiment uses the MSP-Podcast corpus [42], which consists of natural and diverse emotional speech samples from various podcast recordings [6]. The audios do not include background music or overlapped speech, and their predicted SNR is above 20 dB. We consider this corpus a clean emotion speech database for these reasons. This study focuses on predicting the emotional attributes of arousal (calm to active), dominance (weak to strong), and valence (negative to positive). Labels for these attributes were annotated by at least five raters using a seven-point Likert scale. We average the scores provided by raters for each sample to establish its ground truth values. This paper uses version 1.10 of the corpus, which consists of 104,267 annotated utterances. We use the train set to fine-tune the pre-trained speech representation model, using it as the original SER model. We use samples from the development set to select the best model during the fine-tuning process.

We simulate real-world noisy environments by collecting noise sounds from the Freesound repository [43], which contains publicly available ambient noise sounds. We use diverse queries related to each environment to collect noise sounds, including indoor, outdoor, and in-vehicle conditions. Additionally, we included the DEMAND dataset for additional testing conditions. DEMAND contains 15 different recording

conditions that simulate indoor, outdoor, and in-vehicle environments. We directly use the metadata of each recording sample to define the keyword for the testing conditions. Table I illustrates the keywords that are used to contaminate train, development, and test sets. We use 20 noisy environments for the train and development sets and 89 environments to contaminate the test set. Although these noise sounds are not used during adaptation, they have common characteristics with the noise sounds used during adaptation (e.g., indoor, outdoor, or in-vehicle conditions). We want to evaluate whether our proposed method can capture this semantic similarity during inference. We randomly pick the noise sounds to contaminate the Test1 set of the clean MSP-Podcast corpus. We repeat this process 10 times, creating 10 different sets for three different SNR levels, 5dB, 0dB, and -5dB. We also create a *random* set, where the SNR levels are randomly selected from 5dB to -5dB. This set simulates the testing condition in a real-world application where the SNR level varies.

B. Fine-Tuning Transformer-Based Architecture

We implement our proposed approach with two different pre-trained speech representation models: wav2vec2-large-robust [44] and the wavlm-base-plus models [3]. The wav2vec2-large-robust model has shown good performance in the emotional attribute prediction task [33]. The wavlm-base-plus model has shown good performance for emotion recognition in the *speech processing universal performance benchmark* (SUPERB) [45]. This model is pre-trained with noise, creating representations that are expected to be more robust to noise than other SSL representations. We fine-tune the transformer encoder of the pre-trained speech representation model and the downstream head with the clean version of the MSP-Podcast corpus. For wav2vec2-large-robust, we remove the top 12 transformer layers from the model to preserve the recognition performance with fewer parameters [33]. We import the pre-trained models from the HuggingFace library [46]. We use two fully connected layers for the downstream head, where each layer has 512 nodes, layer normalization, and the *rectified linear unit* (ReLU) as the activation function. We use dropout in all the hidden layers to increase regularization, with a rate set to $p = 0.5$. We use a linear output layer with three nodes to predict emotional attribute scores, where each node predicts the scores for arousal, dominance, and valence. We apply average pooling on top of the last transformer layer's representation to feed it to the downstream head.

During fine-tuning, we apply Z-normalization to the raw waveform by using the mean and standard deviation estimated over the training set and min-max normalization to the emotional labels, mapping them to the range of 0 to 1. We use the same mean and standard deviation estimated over the training set to normalize the test set's raw waveform. We use 32 utterances per mini-batch and update the model for ten epochs. We use the Adam optimizer [47] with a learning rate warmup scheduling, which shows good performance when fine-tuning a pre-trained transformer architecture [48]. For the first 1,000 mini-batches, we linearly increase the learning rate from $1e^{-8}$ to $1e^{-5}$. After the 1,000 mini-batches, we fix the learning rate to $1e^{-5}$. All of our experiments are conducted on a single NVIDIA GeForce RTX 3090.

C. Text-Guided Environment-Aware Training

After fine-tuning with the clean speech, we adapt the SER model to the noisy environmental conditions. We randomly select one of the 20 noise conditions for each mini-batch during adaptation. We then use 32 different noise samples in the selected condition to contaminate 32 clean speech samples from the training set of the MSP-Podcast corpus. We build text prompts with respect to the picked environment for each mini-batch, as described in Section III. In real-world applications, it is difficult to assume the exact SNR level of the testing condition. Therefore, we introduce an SER mismatch between our experiment's adaptation and testing stages. We randomly select the SNR level for the adaptation of the models among these options: $\{2.5, 7.5, 12.5\}$dB. We use the same hyperparameters as the ones used for fine-tuning the SER model with clean speech during adaptation. We tested two variations of our proposed text-guided environment-aware training: the CL-based representation TG-EAT-CL, and the LLM-based representation TG-EAT-LLM.

D. Baselines

Original: This model fine-tunes the model with clean emotional speech, with no adaptation to the noisy conditions.

Retrain the original model with noisy speech (RT): This baseline updates the transformer encoder and the downstream head of the Original model with noisy speech. It does not use environmental information during adaptation and inference. As described in Section IV-A, it uses 20 environmental conditions for adaptation. The evaluation uses 89 other environmental conditions.

Domain adversarial training (DAT): Inspired by Huang et al. [49], we test a domain adversarial training strategy to adapt an SER model to multiple noisy conditions. Along with the downstream head for the SER task, we attach an environment classifier on top of the average-pooled transformer representations. The environment classifier has the same architecture as the downstream head for the SER task. The environment classifier is trained to minimize the cross-entropy loss between the predicted and the ground-truth noise types. We applied a *gradient reversal layer* (GRL) between the environment classifier and the transformer encoder to train the transformer encoder to normalize the environment information in the resulting representations. Like the RT baseline, this baseline does not use environmental information during inference.

Enhance the noisy speech (SE): This baseline denoises the input noisy speech before feeding it into the original SER model. We use the *frequency recurrent convolutional recurrent network* (FRCRN) framework [50] to enhance the input speech. The FRCRN model is trained with the 4th DNS challenge dataset, achieving one of the top performances in this challenge [51].

TABLE II: Average CCC for models using wav2vec2-large-robust feature vectors. We denote with *, †, *, and ‡ when a model shows significantly better performance than the Original, RT, DAT, and SE models, respectively. We also mark $^{\diamondsuit}$ and $^{\clubsuit}$ when a baseline significantly perform better than the TG-EAT-CL and TG-EAT-LLM, respectively. We highlight in bold the best performance per condition.

SNR	Model	Arousal	Dominance	Valence
Clean	Original ($*$)	0.63	0.53	0.41
	RT ($\dagger$)	0.63	0.53	0.46*
	DAT ($\star$)	0.63	0.51	0.45*
	SE ($\ddagger$)	0.53	0.48	0.37
	TG-EAT-CL ($\diamondsuit$)	0.63	0.53	0.45*
	TG-EAT-LLM ($\clubsuit$)	0.63	0.53	0.46*
5dB	Original ($*$)	0.60‡	0.51‡	0.40‡
	RT ($\dagger$)	**0.63***‡	**0.52**‡	0.44*‡
	DAT ($\star$)	0.62‡	0.50‡	0.44*‡
	SE ($\ddagger$)	0.50	0.44	0.35
	TG-EAT-CL ($\diamondsuit$)	0.62‡	0.51‡	0.45*‡
	TG-EAT-LLM ($\clubsuit$)	0.62‡	**0.52**‡	**0.46***‡
0dB	Original ($*$)	0.54‡	0.46$^{\ddagger\diamondsuit}$	0.31
	RT ($\dagger$)	0.55$^{\ddagger\diamondsuit}$	0.46$^{\ddagger\diamondsuit}$	0.38*‡
	DAT ($\star$)	0.54$^{\ddagger\diamondsuit}$	0.44‡	**0.39***‡
	SE ($\ddagger$)	0.47	0.41	0.35*
	TG-EAT-CL ($\diamondsuit$)	0.52‡	0.42	0.38*‡
	TG-EAT-LLM ($\clubsuit$)	**0.56**$^{**\ddagger\diamondsuit}$	**0.47**$^{\star\ddagger\diamondsuit}$	**0.39***‡
-5dB	Original ($*$)	0.26$^{\diamondsuit}$	0.24$^{\ddagger\diamondsuit}$	0.11
	RT ($\dagger$)	0.22	0.21	0.15**
	DAT ($\star$)	0.24$^{\diamondsuit}$	0.22‡	0.13
	SE ($\ddagger$)	0.23	0.19	**0.19**$^{*\dagger\star\diamondsuit\clubsuit}$
	TG-EAT-CL ($\diamondsuit$)	0.21	0.20	0.15**
	TG-EAT-LLM ($\clubsuit$)	**0.28**$^{*\dagger\star\diamondsuit}$	**0.26**$^{*\dagger\star\ddagger\diamondsuit}$	0.16**
Random	Original ($*$)	0.47$^{\ddagger\diamondsuit}$	0.41$^{\ddagger\diamondsuit}$	0.29
	RT ($\dagger$)	0.47$^{\ddagger\diamondsuit}$	0.39$^{\ddagger\diamondsuit}$	0.35*‡
	DAT ($\star$)	0.47$^{\ddagger\diamondsuit}$	0.39‡	0.34*‡
	SE ($\ddagger$)	0.37	0.32	0.30
	TG-EAT-CL ($\diamondsuit$)	0.44‡	0.37‡	0.34*‡
	TG-EAT-LLM ($\clubsuit$)	**0.50**$^{*\dagger\star\ddagger\diamondsuit}$	**0.42**$^{\dagger\star\ddagger\diamondsuit}$	**0.36****‡

TABLE III: Average CCC for models using wavlm-base-plus feature vectors. We use the same notations as in Table II. We highlight in bold the best performance per condition.

SNR	Model	Arousal	Dominance	Valence
Clean	Original ($*$)	0.60	0.49	0.46
	RT ($\dagger$)	0.59	0.49	0.43
	DAT ($\star$)	0.58	0.48	0.48
	SE ($\ddagger$)	0.58	0.46	0.43
	TG-EAT-CL ($\diamondsuit$)	0.57	0.47	0.47
	TG-EAT-LLM ($\clubsuit$)	0.59	0.48	0.46
5dB	Original ($*$)	0.54	0.45	0.44
	RT ($\dagger$)	**0.58***‡	**0.48***‡	0.41
	DAT ($\star$)	**0.58***‡	**0.48***‡	**0.47***†‡
	SE ($\ddagger$)	0.55	0.45	0.40
	TG-EAT-CL ($\diamondsuit$)	0.57*‡	0.47*‡	0.46*†‡
	TG-EAT-LLM ($\clubsuit$)	**0.58***‡	0.47*‡	0.44†‡
0dB	Original ($*$)	0.40	0.31	0.33
	RT ($\dagger$)	0.53*	0.43*	0.33
	DAT ($\star$)	0.53*	**0.45**$^{*\dagger\diamondsuit}$	**0.41***†
	SE ($\ddagger$)	0.53*	0.44*	**0.41***†
	TG-EAT-CL ($\diamondsuit$)	0.51*	0.42*	0.40*†
	TG-EAT-LLM ($\clubsuit$)	**0.55**$^{*\dagger\star\ddagger\diamondsuit}$	**0.45**$^{*\dagger\diamondsuit}$	0.38*†
-5dB	Original ($*$)	0.11	0.07	0.10
	RT ($\dagger$)	0.18*	0.11*	0.12
	DAT ($\star$)	0.22$^{*\dagger\diamondsuit}$	0.16$^{*\dagger\diamondsuit}$	0.17*†
	SE ($\ddagger$)	0.28$^{*\dagger\star\diamondsuit}$	**0.22**$^{*\dagger\star\diamondsuit}$	**0.23**$^{*\dagger\star\diamondsuit\clubsuit}$
	TG-EAT-CL ($\diamondsuit$)	0.17*	0.11*	0.18*†
	TG-EAT-LLM ($\clubsuit$)	**0.29**$^{*\dagger\star\diamondsuit}$	0.20$^{*\dagger\star\diamondsuit}$	0.20*†*
Random	Original ($*$)	0.34	0.25	0.31
	RT ($\dagger$)	0.45*	0.33*	0.30
	DAT ($\star$)	0.46$^{*\dagger\ddagger\diamondsuit}$	0.37$^{*\dagger\diamondsuit}$	**0.38***†‡
	SE ($\ddagger$)	0.44*	0.35*	0.35*
	TG-EAT-CL ($\diamondsuit$)	0.43*	0.33*	0.37*†
	TG-EAT-LLM ($\clubsuit$)	**0.50**$^{*\dagger\star\ddagger\diamondsuit}$	**0.39**$^{*\dagger\star\ddagger\diamondsuit}$	0.36*†

V. Results

A. Emotion recognition performance

We report the SER performance of our text-guided environment-aware training with our baselines. As described in Section IV-A, we use ten different evaluation sets for three SNR levels. We report the average CCC of ten experiments for each SNR level. We conduct a one-tailed Welch's t-test between the baselines and our proposed models to assess if the training strategy shows significantly better SER performance in noisy conditions. We assert significance at p-value < 0.05.

Tables II and III illustrate the SER performance of each model in noisy testing environments. When comparing our baselines (RT, DAT, SE) with the original model, they do not consistently yield performance improvement for all the attributes. RT does not improve performance for either arousal or dominance with the wav2vec2-large-robust feature vector, or for valence with the wavlm-base-plus feature vector. Although the DAT and SE show significant performance improvements with the wavlm-base-plus feature vector, both baselines fail to improve arousal and dominance prediction performance with the wav2vec2-large-robust feature vector. Since these baselines do not use environmental information, we can observe the importance of incorporating it when adapting the SER model to multiple noisy environments.

Compared with the baselines, our proposed TG-EAT-LLM performs the best when using the wav2vec2-large-robust feature vector. In the random condition, TG-EAT-LLM improves the original model's performance by 6.3% (arousal), 2.4% (dominance), and 24.1% (valence). It yields the best performance with the wavlm-base-plus feature vector for arousal and dominance prediction tasks. In the random condition, TG-EAT-LLM shows performance gains of 8.6% (arousal) and 5.4% (dominance) compared to the best baseline, DAT. Unlike with the wav2vec2-large-robust representation, DAT significantly improves the original model's performance for all the attributes with the wavlm-base-plus representation. The wavlm-base-plus is pre-trained with noisy data, while the wav2vec2-large-robust is trained with a diverse speech corpus under clean conditions. This difference makes the wavlm-base-plus inherently more robust to noise, which leads to the successful improvement with the baselines that do not use environmental information. We note that TG-EAT-LLM consistently outperforms the original model across all SSL representations. These results indicate that guiding the SER model with LLM-based representation can improve the noise-robustness for the SER task. It shows good generalization to unknown environments.

For the valence prediction, the SE baseline shows the best performance under the -5dB condition. Previous studies have shown that valence performance correlates with the linguistic

information [33]. This phenomenon could explain how the SE baseline can improve valence performance by explicitly enhancing speech intelligibility. However, it does not always yield the best performance for arousal and dominance. Both arousal and dominance are related to acoustic characteristics rather than linguistic information. Thus, this observation implies that the enhancement module can manipulate the acoustic characteristics of the original speech. We can also see that the SE baseline does not yield the best performance for valence under 5dB conditions, where the impact of acoustic distortion could be higher than the impact of intelligibility improvement.

When we compare the TG-EAT-CL and TG-EAT-LLM models, we conclude that the CL-based representation does not show a performance improvement over the original SER model, especially with the wav2vec2-large-robust feature vector. We can clearly see that the TG-EAT-CL model does not improve the performance for arousal and dominance in the 0dB and -5dB conditions. This result indicates that pre-training the text encoder to have enriched semantic information is more helpful for the noise-robust SER model than pre-training the text encoder with an audio-text pair.

An interesting and counter-intuitive finding here is that the models trained with noisy speech (e.g., RT, DAT, TG-EAT-CL, TG-EAT-LLM) outperform the Original model under clean conditions when experimenting with the wav2vec2-large-robust architecture. We assume this improvement is caused by exposing the noisy speech to the wav2vec2-large-robust representation, which is not trained with noisy speech in its pre-training stage. Previous studies have shown that augmenting the training set with multiple conditions not only improves *automatic speech recognition* (ASR) performance under noisy conditions but also improves under a clean condition [52], [53]. As discussed in the previous section, we hypothesize that improvements in speech intelligibility leads to improvement of valence prediction. Based on these observations, we conclude that this phenomenon, while unintuitive, demonstrates the benefits of data augmentation under clean conditions.

B. Embedding analysis

Section V-A demonstrated that the TG-EAT-LLM approach shows better performance than the environment-agnostic baselines and the TG-EAT-CL approach. Our initial assumption is that the proposed TG-EAT-LLM can learn appropriate denoising functions for the transformer encoder. To verify this assumption, we analyze the difference between the clean and noisy representations (Fig. 2(a)). We use the wavlm-base-plus feature vector and the noisy speech from the -5dB condition for this analysis. The first analysis compares the clean and noisy representation extracted from each model. We want to assess with this analysis if the model is robust by comparing the representation obtained with clean and noisy speech. The second analysis compares the clean representation from the Original framework and the noisy representation from each of the models (Fig. 2(b)). In this analysis, we want to assess if the model can keep the knowledge of the original SER model. We extract the representations from the first and the last transformer encoder layers and then calculate the mean square difference between clean and noisy representations for each layer.

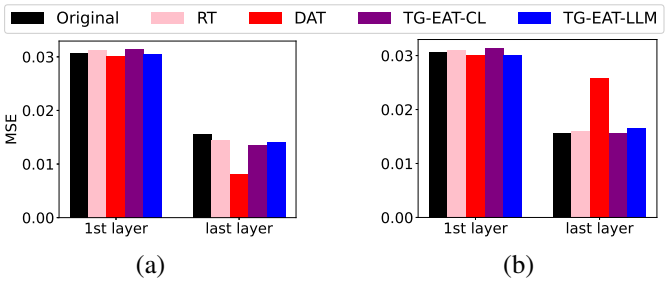


Fig. 2: Embedding differences in the first and the last transformer encoder layers using clean and noisy speech in the -5dB condition. We use the wavlm-base-plus feature vector in this analysis. (a) illustrates the *mean square error* (MSE) between the clean and noisy representations, where both representations are extracted from each of the final models. (b) illustrates the MSE between the clean representation extracted from the Original model and the noisy representation extracted from each final model.

Figure 2 illustrates our analysis results. When extracting clean and noisy representations from the same model, we can first see that DAT shows the lowest difference in the last transformer layer. On the contrary, it shows the highest difference when extracting the clean representation from the original model. This result demonstrates the risk of catastrophic forgetting when using the DAT method. Although it can normalize the environmental difference in the adapted model, its representation can deviate from the original SER model's representation. However, our TG-EAT method does not highly increase the difference compared to the original model's clean representation. This result indicates that TG-EAT can minimize the risk of catastrophic forgetting during adaptation by introducing environmental information about the speech.

Compared with the TG-EAT-LLM method, TG-EAT-CL shows a higher representation difference in the first layer. When comparing the clean and noisy representations from the same model, TG-EAT-LLM shows 7.7% less representation difference than the TG-EAT-CL method in the first transformer layer. However, TG-EAT-CL shows less representation difference than the TG-EAT-LLM in the last layer. Even though the downstream head uses the representation from the last transformer layer, TG-EAT-CL shows worse performance than the TG-EAT-LLM approach. LLM-based representation can better denoise the acoustic representation than the CL-based representation. In addition, we speculate that the embedding difference in the lower transformer layer might be the crucial factor for increasing the robustness to noise of the SER system.

We also investigate if the proposed text-based environment embedding clusters similar environments together, which is the key premise of the proposed approach to deal with unseen environments. First, we randomly select 21 different keywords, each representing an indoor, outdoor, and in-vehicle environ-

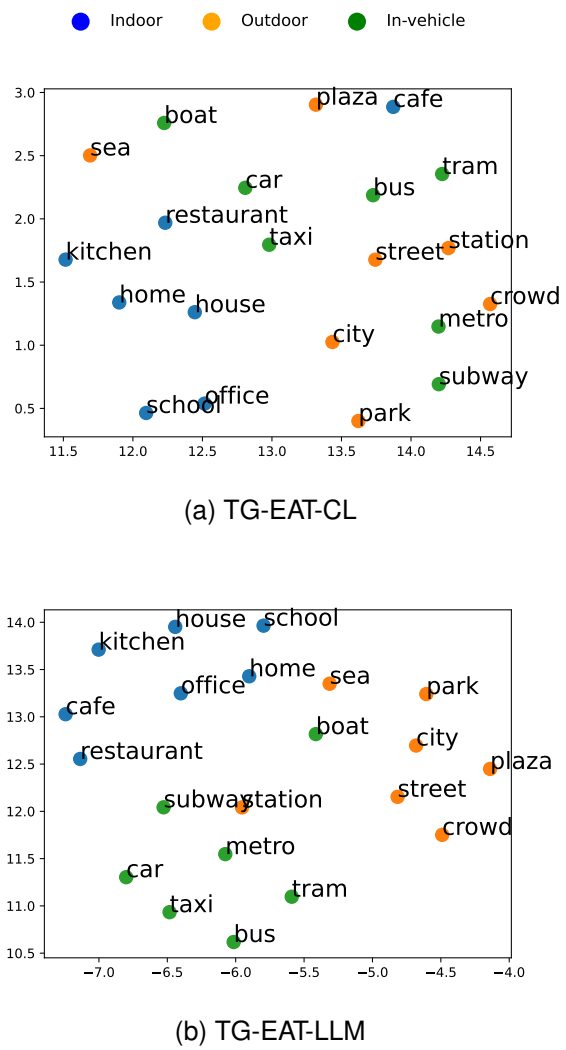


Fig. 3: Visualization of text-based environment embeddings. We use UMAP to project text embeddings into 2D space.

ment. Each environment includes seven keywords extracted from the train and test sets, aiming to illustrate the model's capability to cluster similar environments in both seen and unseen environments. We extract the text embedding from these 21 keywords by using the same template that we used for our TG-EAT frameworks (i.e., *"This speech is recorded in* {*environment*}*."*) We project these embeddings into the 2D space to visualize the embedding space using the *uniform manifold approximation and projection* (UMAP) method [54]. Figure 3 illustrates the text embedding space of TG-EAT-CL and TG-EAT-LLM. The figure shows that both frameworks cluster semantically similar environmental conditions together. For example, we observe the embeddings for "boat" and "sea," together. We also observe the ones for "subway" and "station" clustered together. Both encoders cluster the house environments ("house", "home", "kitchen") and the vehicle environments ("bus", "taxi", "car"), which indicates that the text encoder can cluster acoustically similar environments. This analysis implies that our proposed frameworks can handle unseen environments by clustering acoustically and semantically similar environments.

TABLE IV: Silhouette score of text embedding space of TG-EAT-CL and TG-EAT-LLM. We apply K-means clustering on the projected environmental embedding with K clusters ($K = 3, 5, 7$).

	K = 3	K = 5	K = 7
TG-EAT-CL	0.11	0.10	0.10
TG-EAT-LLM	0.57	0.43	0.40

TABLE V: Average CCC of the ten experiments for the seen environment. The environmental conditions for the train set and the test set are the same. We compare the proposed method with the baselines by using the wavlm-base-plus model.

SNR	Model	Arousal	Dominance	Valence
5dB	Original	0.54	0.46	0.45
	One-hot	0.59	0.48	0.47
	TG-EAT-LLM	0.59	0.48	0.47
0dB	Original	0.40	0.32	0.35
	One-hot	0.56	0.45	0.42
	TG-EAT-LLM	0.56	0.46	0.40
-5dB	Original	0.09	0.06	0.10
	One-hot	0.29	0.20	0.21
	TG-EAT-LLM	0.27	0.18	0.21

To provide a quantitative analysis of the environmental embeddings and their impact on the model's representations, we evaluated the clustering quality of the environmental text embeddings. We extracted embeddings for all environments listed in Table I from both the TG-EAT-CL and TG-EAT-LLM text encoders. We calculated the silhouette score for each set of embeddings using the K-means clustering [55]. Table IV illustrates the silhouette score of each embedding projection with a different number of clusters. With three clusters, the TG-EAT-LLM embeddings achieved a score of 0.57, substantially higher than the 0.10 score from the TG-EAT-CL embeddings. This result indicates that the LLM-based encoder generates more separable and well-defined clusters for different environments. This higher-quality embedding structure correlates with the superior performance of the TG-EAT-LLM model in noisy conditions, suggesting that more discriminable environmental representations are key to achieving robust performance.

C. Evaluation of Different Types of Environmental Embedding

Our proposed method uses the embedding extracted from the text encoder to represent the testing environmental condition. To verify the benefits of using a text-based environmental embedding, we compare it with three different types of environmental embedding: *one-hot encoding* (One-hot), *global vectors for word representation* (GloVe) [56], and *audio spectrogram transformer representation* (AST) [57]. One-hot uses 20-dimensional binary vectors, where 1 represents the target environment condition, and 0 represents the others. Each dimension corresponds to the environmental condition of the training set. This embedding fully represents a seen environment with a simple vector; however, it cannot represent unseen environments, which is inappropriate for real-world services. GloVe is a word-level vector representation extracted from the regression model that considers the co-occurrences

TABLE VI: Average CCC of the ten experiments for the unseen environment. We compare the proposed method with the baselines by using the wavlm-base-plus model. We denote with * when a model shows significantly better performance than the Original model.

SNR	Model	Arousal	Dominance	Valence
5dB	Original	0.54	0.45	**0.44**
	RT	0.58*	0.48*	0.41
	GloVe	0.58*	0.47*	0.42
	AST	**0.59***	**0.49***	0.41
	TG-EAT-LLM	0.58*	0.48*	**0.44**
0dB	Original	0.40	0.31	0.33
	RT	0.53*	0.43*	0.33
	GloVe	0.53*	0.42*	0.37*
	AST	**0.55***	0.44*	0.34
	TG-EAT-LLM	**0.55***	**0.45***	**0.38***
-5dB	Original	0.11	0.07	0.10
	RT	0.18*	0.11*	0.12
	GloVe	0.24*	0.16*	0.18*
	AST	0.28*	**0.20***	0.14*
	TG-EAT-LLM	**0.29***	**0.20***	**0.20***

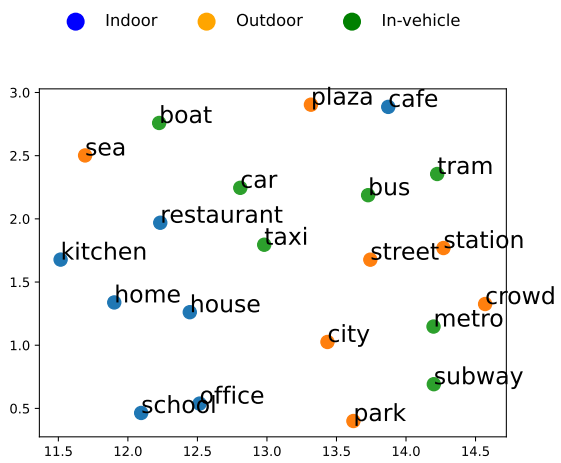
of words. We import the pre-trained GloVe vector collections, which consist of a 2.2 million-word vocabulary. We select the word vector representation that corresponds to the target noisy environment. The resulting representation is a 300-dimensional vector. This representation can handle unseen environments through text description, but it is semantically limited compared to our proposed text encoders. AST uses a transformer architecture to map the spectrogram patches into an audio-level representation. The model is fine-tuned with sound event classification tasks using AudioSet, which serves as the noise sound corpus for our training set. We directly import the pre-trained checkpoint from HuggingFace and extract the patch-wise embedding sequence from the given input. We apply average pooling to the extracted sequence to yield a single environment embedding, which is then fused with the pre-trained SER model. We do not fine-tune the pre-trained checkpoint jointly with the SER model, following the same strategy we use to train the TG-EAT-LLM framework. This model can automatically capture the acoustic characteristics from the audio-only input. However, it cannot explicitly use the semantic information of the testing environment.

We compare our proposed method with the one-hot vector in the seen environment scenario (Table V) and with the other baselines in the unseen environment scenario (Table VI). For the seen environment scenario, we used the same environmental conditions as the train set to contaminate the clean test set, but with different audio samples. We use ten different test sets and report the average CCC for both cases. Tables V and VI report the results for the seen and unseen environments, respectively. In the seen environment, our proposed method and the one-hot environment encoding model improve the original SER performance for all the conditions and attributes. Both models show similar performances in the seen environments. However, the one-hot encoding cannot cover unseen environments. This result demonstrates that the proposed text embedding can deal with both seen and unseen environments. Compared to the model that uses GloVe embeddings, our proposed method shows better SER performances in the 0dB and -5dB conditions. It also shows a better performance for valence in the 5dB condition. The GloVe model only considers word co-occurrence to get a word embedding, while our proposed text encoder model is pre-trained to understand the semantic information of a sentence. This result implies the importance of pre-training the text encoder with language modeling to get a robust environment embedding for performance improvement. The AST strategy significantly improves the performance for arousal and dominance. However, it fails to improve the performance for valence when the SNR level is high (e.g., 5dB and 0dB conditions). AST does not use semantic information from the testing environment to get environmental embedding; instead, it extracts the environmental information from the given audio. We hypothesize that AST confuses the environmental condition when the background noise amplitude is comparably lower than the speech sound. Unlike this approach, our proposed model relies on the text description, which is independent of the SNR level. Therefore, it performs better than AST for valence. In the 0 dB and -5 dB conditions, our method significantly improves the original models' performances for all the attributes. Considering that those low SNR levels are not presented while training the model, the result demonstrates that our proposed method is robust against unseen SNR levels, which is practical for real-world scenarios.

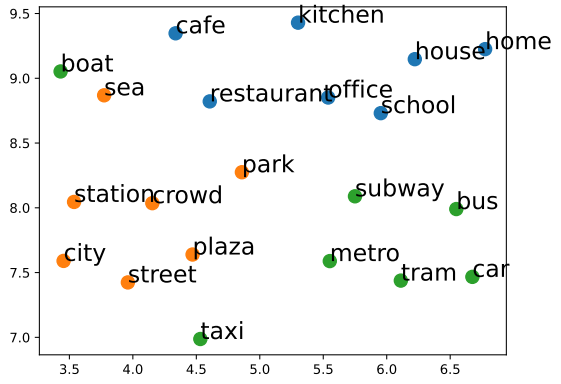
D. Benefit of Fine-Tuning the Text Encoder

Our results demonstrate that using the text encoder pre-trained with the CLAP strategy shows worse SER performance than using the pre-trained LLM. Despite this observation, we assume that this type of text encoder should have the potential to improve since the text encoder is pre-trained with the audio modality. Our assumption is that jointly fine-tuning the text encoder with the SER model could further improve the performance. Therefore, we compare the performance of an SER model by either freezing the text encoder or updating the encoder while adapting the SER model with the text-based environment embedding. We refer to the models that fine-tune the text encoder of the *TG-EAT-CL* and *TG-EAT-LLM* approaches during adaptation as *TG-EAT-CL-FT* and *TG-EAT-LLM-FT*, respectively.

Table VII reports the average CCC of ten different test sets for each model. When comparing the *TG-EAT-LLM* and *TG-EAT-LLM-FT* implementations, they do not show significantly different performance. However, the *TG-EAT-CL-FT* approach shows meaningful performance improvement over the *TG-EAT-CL* implementation. For the -5dB conditions, it even reaches the best performance among all the models. When compared with TG-EAT-CL, fine-tuning the text encoder improves the recognition performance by 76.4% (arousal), 100.0% (dominance), and 27.7% (valence). To analyze how fine-tuning benefits the model, we visualize in Figure 4 the text embedding projection from the text encoder used for the TG-EAT-CL and TG-EAT-CL-FT models. We use the same experiment setting as we used for illustrating Figure 3. We can see that some of the embeddings that were not clustered well in TG-EAT-CL are corrected in TG-EAT-CL-FT. For



(a) TG-EAT-CL



(b) TG-EAT-CL-FT

Fig. 4: Comparison of the text embedding projection obtained before fine-tuning (TG-EAT-CL) and after fine-tuning (TG-EAT-CL-FT). Similar to the plots in Figure 3, we use UMAP to project text embeddings into a 2D space.

example, "cafe" is distant from other indoor environments in TG-EAT-CL. However, when fine-tuning the text encoder, its embedding gets closer to those environments. We also observe that such cluster alignments could lead to performance improvement. Figure 5 illustrates the relative performance improvement of TG-EAT-CL-FT for each environmental condition. We can see that the TG-EAT-CL-FT model improves performance in the "cafe" environments. We can also see that the fine-tuning strategy can improve the performance for all the attributes, except for five environments in valence ("creek", "playground", "stadium", "arcade", and "fountain"). This observation illustrates the importance of compensating for the gap in the embedding space between the pre-trained text encoder space and the acoustic embedding. Although jointly fine-tuning the text encoder and the SER model can cost more memory space and computation time for the adaptation, this strategy can fully utilize the potential of the text encoder pre-trained with the audio modality.

TABLE VII: Comparison of freezing the text encoder and updating it while adapting the SER model for the TG-EAT-CL and the TG-EAT-LLM models. We report the average CCC of the ten experiments for all the methods. We implement all the approaches with wavlm-base-plus feature vectors. We highlight in bold the best performance per condition.

SNR	Model	Arousal	Dominance	Valence
5dB	TG-EAT-CL	0.57	0.47	0.46
	TG-EAT-CL-FT	**0.58**	**0.48**	**0.48**
	TG-EAT-LLM	**0.58**	0.47	0.44
	TG-EAT-LLM-FT	0.57	0.46	0.46
0dB	TG-EAT-CL	0.51	0.42	0.40
	TG-EAT-CL-FT	**0.55**	**0.45**	**0.44**
	TG-EAT-LLM	**0.55**	**0.45**	0.38
	TG-EAT-LLM-FT	0.54	0.44	0.41
-5dB	TG-EAT-CL	0.17	0.11	0.18
	TG-EAT-CL-FT	**0.30**	**0.22**	**0.23**
	TG-EAT-LLM	0.29	0.20	0.20
	TG-EAT-LLM-FT	0.27	0.19	0.21

TABLE VIII: Average CCC of the six sessions with the clean and noisy version of the MSP-IMPROV corpus. We compared the proposed method with the baselines by using the wavlm-base-plus model.

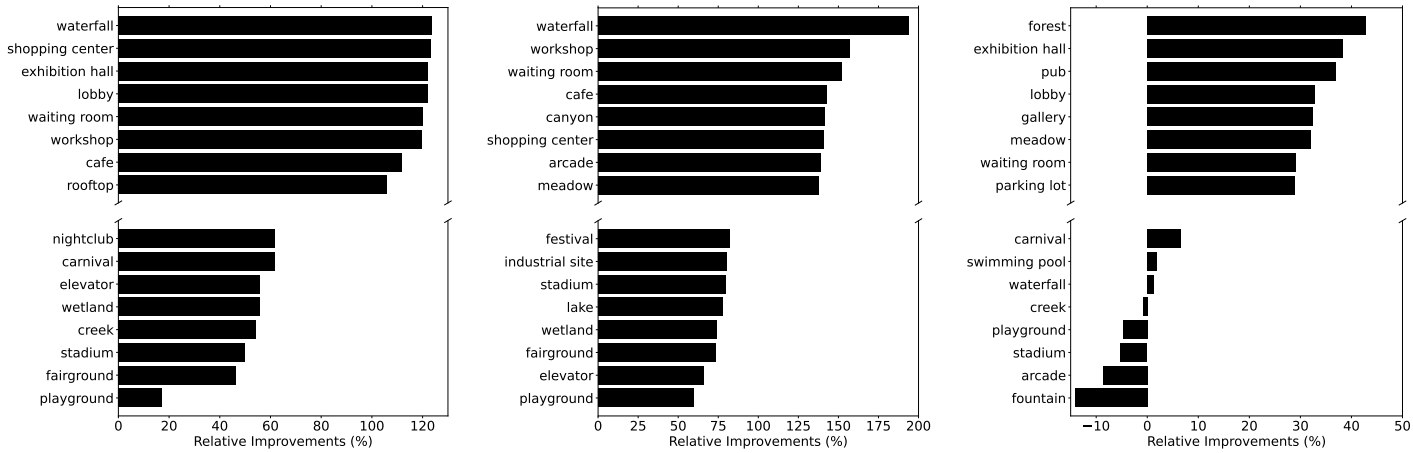
SNR	**Model**	**Arousal**	**Dominance**	**Valence**
Clean	Original	0.38	0.44	0.41
	RT	0.39	0.44	0.40
	TG-EAT-CL	0.38	0.45	**0.44**
	TG-EAT-LLM	**0.40**	0.45	0.42
Random	Original	0.32	0.36	0.25
	RT	**0.40**	0.42	0.30
	TG-EAT-CL	0.36	0.42	**0.34**
	TG-EAT-LLM	**0.40**	0.42	0.32

E. Cross-corpus Generalization

To evaluate the generalization ability of our proposed TG-EAT models under unseen, out-of-domain dataset, we conduct a cross-corpus analysis. We test our baselines and the proposed TG-EAT models on the MSP-IMPROV dataset [58], where their data acquisition process is different from our training set, the MSP-Podcast Corpus. For this evaluation, we evaluate the performance for each of the six sessions in the MSP-IMPROV corpus. We create a noisy version of this test set by contaminating the clean audio with noise from the DEMAND database at a randomly selected SNR level between -5dB and 5dB.

Table VIII illustrates our experiment results. Under the clean condition, we can see that our proposed TG-EAT models do not degrade the performance of baselines. Indeed, they outperform the RT baseline for valence. As we discussed in Section V-A, training with multiple conditions could help performance improvement for valence even under the clean condition. We can see that guiding the model with environmental conditions can keep this benefit under a cross-corpus scenario, while preserving the Original SER model's generalization ability.

As expected, under noisy conditions, all models retrained with noisy speech (RT, TG-EAT-CL, and TG-EAT-LLM) significantly outperform the Original model, confirming that noise-aware training is crucial for robust performance on out-of-set corpora. Our models show a clear improvement in valence. As discussed in Section V-A, arousal and dominance are



(a) Arousal (b) Dominance (c) Valence

Fig. 5: Relative improvement of fine-tuning the text encoder (TG-EAT-CL-FT) in the TG-EAT-CL framework under the -5dB condition. We illustrate 16 environments, including the eight highest and the eight lowest improvements for each attribute.

more closely related to acoustic characteristics, while valence is correlated with linguistic content. By providing explicit environmental information, our TG-EAT framework may allow the SER model to better normalize for acoustic variability, thereby improving the extraction of linguistic content critical for valence prediction.

Interestingly, the TG-EAT-CL model yields the best valence performance under noisy, out-of-set conditions, despite not being the top performer in the in-set evaluation. This result indicates that its learned representation possesses a strong capability for generalization, particularly under challenging mismatched conditions. While our framework shows clear benefits, we note that the significant improvements observed for arousal and dominance in the in-set condition did not fully transfer to this cross-corpus task. This suggests that future work could explore methods to further enhance the generalization of acoustically-related emotion attributes in out-of-domain scenarios.

F. Impact of Mislabeled Audio Tags

In a real-world scenario, location tags could be mislabeled due to an inaccurate GPS signal or ambiguous locations. To evaluate the impact of mislabeling audio tags, we present an ablation study to report the performance of the proposed TG-EAT-LLM framework, manipulating the environmental description. To simulate mislabeling, we intentionally manipulated the environmental tags associated with the input noisy speech at varying levels of distortion. Specifically, the environmental tags were randomly replaced with ratios of 25%, 50%, and 75%. The original model trained without any mislabeled tags (0% manipulation) was used as the baseline for comparison.

To ensure a robust evaluation, we tested the model on 10 different testing sets, each contaminated with noise at a randomly selected SNR between 5dB, 0dB, and -5 dB. The environmental conditions in the testing data were also chosen randomly to simulate diverse real-world scenarios.

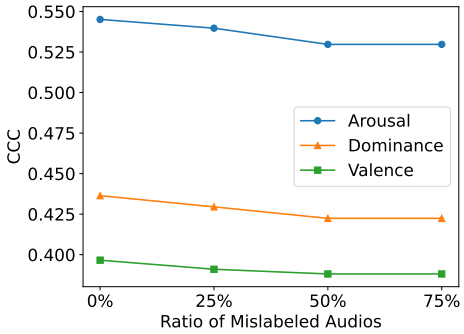


Fig. 6: CCC of the TG-EAT-LLM model trained with noisy speech paired with mislabeled environment conditions.

TABLE IX: Silhouette score of the last hidden layer's embedding space of the RT and TG-EAT-LLM models.

	RT	TG-EAT-LLM
Using the correct environmental tag	0.363	0.133
Using a semantically similar but incorrect tag	0.363	0.133
Using a semantically different tag	0.360	0.140
Using without an environmental tag	0.361	0.138

The average CCC across these 10 sets was computed for each manipulation ratio. Figure 6 illustrates the result of the ablation study. The performance of the TG-EAT-LLM model gradually decreases as the levels of mislabeled audio tags increase. The trend indicates that as the model is exposed to higher degrees of mislabeling, it struggles to cluster recordings accurately from similar environmental conditions. This performance degradation highlights the sensitivity of the TG-EAT-LLM framework to the quality of audio tags and the need for accurate labeling during training. We can see that environmental conditions play a significant role in our proposed framework. The model relies heavily on this information to achieve robust SER performance, highlighting the importance of minimizing labeling errors when describing environmental conditions.

We investigate the degree of impact on our proposed model of using an audio tag that is mislabeled but either semantically similar or completely dissimilar to the correct environmental information. We measured the separability between the final layer embeddings of speech under clean versus noisy conditions. A robust model should normalize the environmental difference, which should lead to a low clustering quality when the embedding space is clustered by environmental differences. We compared our TG-EAT-LLM model against the RT baseline under four prompt conditions: (1) the correct environmental tag, (2) a semantically similar but incorrect tag, (3) a semantically different tag, and (4) no tag at all. We select five similar conditions ("kitchen", "house", "living room", "school", "office"), contaminating the test set with these noise types. SNR levels are randomly chosen from -5dB to 5dB. Condition (1) uses the same tag as the noise label in the input audio. Condition (2) randomly selects the tags from the four other similar conditions. Condition (3) randomly selects tags from five different tags that are semantically different from this group ("playground", "subway station", "town square", "construction site", "sports field"). Condition (4) does not use any environmental tags (i.e., the model only accepts the input audio).

Table IX shows the result of our experiment. The TG-EAT-LLM model consistently achieves much lower silhouette scores than the RT baseline. When using a correct tag, the RT baseline yields a 0.363 score, while our TG-EAT-LLM's score achieves a 0.133 score. This result illustrates that our framework effectively normalizes environmental differences from the speech representation. When compared to using the correct tag, using a semantically similar tag does not change the clustering score in TG-EAT-LLM. The score slightly increases when a semantically different tag or no tag is provided. The RT baseline's score remains unchanged regardless of the text input. These observations imply that guiding the model with environmental information during training can introduce sensitivity to semantically incorrect environmental information. The model does not ignore the prompt but instead uses it as intended to disentangle environmental noise from emotional content.

G. Limitations

Our proposed TG-EAT framework heavily relies on the assumption that the recorded speech is paired with accurate GPS location data, which is crucial for acquiring accurate environmental tags. However, the recorded speech could be associated with inaccurate or missing GPS points in real-world scenarios, leading to irrelevant or unavailable environmental tags. As discussed in Section V-F, having irrelevant tags could degrade our system's performance, and missing tags would not provide the information for our model to work properly. Additionally, GIS mashups may fail to retrieve meaningful tags in areas with sparse or incomplete annotations, further limiting the system's ability to leverage the information of the recording conditions. Furthermore, even with accurate data, a single static tag may be an oversimplification for complex acoustic scenes with overlapping speech or rapidly changing soundscapes, potentially degrading performance. Those limitations demonstrate the need for future work to address potential inaccuracies in GPS data and missing GPS modality. In addition to the availability of accurate GPS information, our architectural approach to fuse the text embedding to the SER model was limited to concatenating a text embedding to audio tokens. The exploration of more dynamic fusion strategies [40], [41] remains a key area for future work to potentially build upon our findings and further enhance performance.

VI. Conclusions

We proposed the TG-EAT method, which uses a text description of the testing environment for noise-robust SER. This approach inserts a text-based environment representation into an SER model, leading it to improve the prediction with respect to the given environmental information. Our experiment demonstrated that the LLM-based representation can improve SER performance under noisy conditions, especially when dealing with low SNR conditions. Our analysis indicates that the pre-trained text encoder can cluster acoustically and semantically similar environments into the same embedding, which is crucial for generalizing the models for unseen environments. Our result also shows that the CLAP-based text encoder can be highly improved by updating the text encoder. This result demonstrates the importance of minimizing the embedding space gap between the text encoder and the acoustic embedding.

We plan to expand this approach to cases where we cannot obtain information on the testing environment. While AST embeddings demonstrate competitive performance for arousal and dominance, they do not show improvements for valence compared to models that explicitly use text embeddings. The CL-based representation can address scenarios where noise information is not provided by introducing its audio encoder. CLAP trains the audio encoder to have a similar representation to the ones from the text encoder, which could be useful for extracting environmental information from the audio. For this reason, we plan to investigate how we can improve the noise-robustness of the SER model with a CLAP encoder. We also plan to investigate the alternative approach of leveraging the inferred caption from an *automated audio captioning* (AAC) model [59], using it as an environmental descriptor.

References

[1] 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.

[2] W.-N. Hsu, Y.-H. H. T. B. Bolte, 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.

[3] 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.

[4] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 1823–1827.

[5] l. Kondratenko, N. Karpov, A. Sokolov, N. Savushkin, O. Kutuzov, and F. Minkin, "Hybrid dataset for speech emotion recognition in Russian language," in *ISCA Interspeech 2023*, Dublin, Ireland, August 2023, pp. 2958–1796.

[6] 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.

[7] S. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. Salman, C. Busso, and C.-C. Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023, pp. 1–8.

[8] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 68–76, February 2021.

[9] D. Fulford, J. Mote, R. Gonzalez, S. Abplanalp, Y. Zhang, J. Luckenbaugh, J.-P. Onnela, C. Busso, and D. Gard, "Smartphone sensing of social interactions in people with and without schizophrenia," *Journal of Psychiatric Research*, vol. 137, pp. 613–620, May 2021.

[10] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C.Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, November 2021.

[11] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018)*, Madrid, Spain, October 2018, pp. 854–860.

[12] Y.-T. Wu and C.-C. Lee, "MetricAug: A distortion metric-lead augmentation strategy for training noise-robust speech emotion recognizer," in *ISCA Interspeech 2023*, Dublin, Ireland, August 2023, pp. 3587–3591.

[13] S. Ranjan, R. Chakraborty, and S. K. Kopparapu, "Reinforcement Learning based Data Augmentation for Noise Robust Speech Emotion Recognition," in *Interspeech 2024*, 2024, pp. 1040–1044.

[14] L. Juszkiewicz, "Improving noise robustness of speech emotion recognition system," in *Intelligent Distributed Computing VII*, ser. International Symposium on Intelligent Distributed Computing (IDC 2013), F. Zavoral, J. Jung, and C. Badica, Eds. Prague, Czech Republic: Springer International Publishing, 2014, vol. 511, pp. 223–232.

[15] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 1691–1695.

[16] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *ISCA Speech Prosody*. Dresden, Germany: ISCA, May 2006.

[17] 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.

[18] A. Wilf and E. Mower Provost, "Towards noise robust speech emotion recognition using dynamic layer customization," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2021)*, Nara, Japan, September-October 2021, pp. 1–8.

[19] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 2871–2875.

[20] V. Mitra, V. Kowtha, H.-Y. S. Chien, E. Azemi, and C. Avendano, "Pre-trained model representations and their robustness against noise for speech emotion analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, June 2023, pp. 1–5.

[21] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Adapting a self-supervised speech representation for noisy speech emotion recognition by using contrastive teacher-student learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[22] ——, "Computation and memory efficient noise adaptation of Wav2Vec2.0 for noisy speech emotion recognition with skip connection adapters," in *Interspeech 2023*, Dublin, Ireland, August 2023, pp. 1888–1892.

[23] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML 2021)*, M. Meila, , and T. Zhang, Eds. Virtual: Proceedings of Machine Learning Research (PMLR), July 2021, vol. 139, pp. 8748–8763.

[24] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, June 2023, pp. 1–5.

[25] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes Island, Greece, June 2023, pp. 1–5.

[26] E. Stanley, E. DeMattos, A. Klementiev, P. Ozimek, G. Clarke, M. Berger, and D. Palaz, "Emotion label encoding using word embeddings for speech emotion recognition," in *ISCA Interspeech 2023*, Dublin, Ireland, August 2023, pp. 2418–2422.

[27] T. Gong, J. Belanich, K. Somandepalli, A. Nagrani, B. Eoff, and B. Jou, "LanSER: Language-model supported speech emotion recognition," in *ISCA Interspeech 2023*, Dublin, Ireland, August 2023, pp. 2408–2412.

[28] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. F. ao, "OpenStreetMap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 184–199, March 2021.

[29] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "An unsupervised frame selection technique for robust emotion recognition in noisy speech," in *European Signal Processing Conference (EUSIPCO 2018)*, Rome, Italy, September 2018, pp. 2055–2059.

[30] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Selective acoustic feature enhancement for speech emotion recognition with noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 32, pp. 917–929, 2024.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] L. Goncalves, D. Robinson, E. Richerson, and C. Busso, "Bridging emotions across languages: Low rank adaptation for multilingual speech emotion recognition," in *Interspeech 2024*, Kos Island, Greece, September 2024.

[36] S. Upadhyay, C. Busso, and C.-C. Lee, "A layer-anchoring strategy for enhancing cross-lingual speech emotion recognition," in *Interspeech 2024*, Kos Island, Greece, September 2024.

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *ArXiv e-prints (arXiv:1907.11692)*, pp. 1–12, July 2019.

[38] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA, May 2019, pp. 1–20.

[39] 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.

[40] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[41] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[42] 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.

[43] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *International Society for Music Information Retrieval (ISMIR 2017)*, Suzhou, China, October 2017, pp. 486–493.

[44] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *ArXiv e-prints (arXiv:2104.01027)*, pp. 1–9, April 2021.

[45] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Lin, A. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-T. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-Y. Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 1194–1198.

[46] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, and Q. L. amd A.M. Rush, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[47] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.

[48] M. M. Popel and O. Bojar, "Training tips for the transformer model," *The Prague Bulletin of Mathematical Linguistics*, vol. 110, pp. 43–70, April 2018.

[49] K. Huang, Y.-K. Fu, Y. Zhang, and H.-Y. Lee, "Improving distortion robustness of self-supervised speech processing tasks with domain adaptation," in *ISCA Interspeech 2022*, Incheon, Korea, September 2022, pp. 2193–2197.

[50] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9281–9285.

[51] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper *et al.*, "Icassp 2022 deep noise suppression challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9271–9275.

[52] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5220–5224.

[53] Pablo Peso Parada and Agnieszka Dobrowolska and Karthikeyan Saravanan and Mete Ozay, "pMCT: Patched Multi-Condition Training for Robust Speech Recognition," in *Interspeech 2022*, 2022, pp. 3779–3783.

[54] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *Journal of Open Source Software,*, vol. 3, no. 29, p. 861, September 2018.

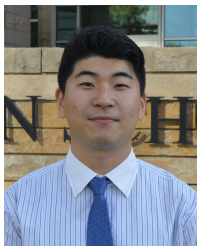
[55] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 747–748.

[56] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, October 2014, pp. 1532–1543.

[57] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *ISCA Interspeech 2021*, Brno, Czechia, August-September 2021, pp. 571–575.

[58] 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.

[59] X. Mei, X. Liu, M. D. Plumbley, and W. Wang, "Automated audio captioning: an overview of recent progress and new challenges," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, 2022.



Seong-Gyun Leem (S'21) received his B.S. and M.S. degree in Computer Science and Engineering at Korea University, Seoul, South Korea in 2018 and 2020, respectively, and his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas, Richardson, Texas, USA in 2024. He is currently a research scientist with Meta Platforms, Inc, Redmond, Washington, USA. His current research interests include speech emotion recognition, noisy speech processing, and machine learning.

Daniel Fulford is Associate Professor of Occupational Therapy, Rehabilitation Sciences, and Psychological & Brain Sciences at Boston University. He holds a PhD in Clinical Psychology from the University of Miami and a BA in psychology from UCLA. His research includes laboratory-based and ambulatory studies to better understand motivation and emotion in psychopathology, using smartphones as tools for experience sampling and behavioral sensing. Much of this research informs the development and testing of digital interventions to support psychosocial functioning in individuals with schizophrenia and other serious mental illness.

Jukka-Pekka "JP" Onnela is Associate Professor of Biostatistics in the Department of Biostatistics at the Harvard T.H. Chan School of Public Health of Harvard University. After completing his doctorate in Finland, he completed a junior research fellowship at the University of Oxford, was a Fulbright Scholar at Harvard University, and a postdoctoral fellow at Harvard Medical School. His main interest is in developing quantitative methods in two areas: statistical network science and digital phenotyping.

David Gard is a Professor of Psychology and the Director of the Motivation and Emotion Research Lab at San Franscico State University (SFSU). He received his Ph.D. in clinical psychology from the University of California at Berkeley in 2005 and started as an Assistant Professor at SFSU that year. His research interests are broadly in the area of emotion and motivation dysfunction in various mental health disorders including schizophrenia, bipolar disorder, and depression. He also researches novel treatments in severe mental illness.

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.