DESCRIBE WHERE YOU ARE: IMPROVING NOISE-ROBUSTNESS FOR SPEECH EMOTION RECOGNITION WITH TEXT DESCRIPTION OF THE ENVIRONMENT

Seong-Gyun Leem
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, TX 75080 USA
SeongGyun.Leem@utdallas.edu

Daniel Fulford
Occupational Therapy and
Psychological and Brain Sciences
Boston University
MA 02215 USA
dfulford@bu.edu

Jukka-Pekka Onnela
Department of Biostatistics,
Harvard T.H. Chan School of Public Health
Harvard University
MA 02138 USA
onnela@hsph.harvard.edu

David Gard
Psychology Department
San Francisco State University
CA 94132 USA
dgard@sfsu.edu

Carlos Busso
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, TX 75080 USA
busso@utdallas.edu

July 26, 2024

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 repository. 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. In addition, our proposed approach with an LLM yields better performance than our environment-agnostic baselines, especially in low *signal-to-noise ratio* (SNR) conditions. When testing at -5dB SNR level, our proposed method shows better performance than our best baseline model by 31.8 % (arousal), 23.5% (dominance), and 9.5% (valence).

1 Introduction

Speech *emotion recogntion* (SER) systems have highly improved with the help of pre-trained speech representation models [1, 2, 3] and the creation of larger emotional speech databases [4, 5, 6, 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].

Several solutions have been proposed to improve the robustness of SER systems against acoustic noise. The solutions include data augmentation [11, 12], feature enhancement [13, 14], feature selection [15, 16], and domain adaptation approaches [17, 18]. Since transformer-based speech representation models have been successfully used in speech problems [1, 2, 3], many studies have also worked on increasing the noise robustness of SER systems built with pre-trained speech representation models [19, 20]. 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. [21] 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 focuses on how to effectively use the prior knowledge of a testing condition for an SER model that is adapted to multiple environments. 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 [22], sound event classification [23], and several speech processing downstream tasks, including keyword spotting, and speaker counting [24]. Natural language supervision is also applicable to SER tasks [25, 26]. All these studies indicate that exploiting text information is a promising strategy to 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 description of the testing environment can highly improve the SER performance, especially with *large language model* (LLM). In the -5dB *signal-to-noise ratio* (SNR) condition, our method improves the original SER model built with a *self-supervised learning* (SSL) representation by 163.6% for arousal, 200.0% for dominance, and 91.6% for valence. When we compare the proposed SER model with our best baseline, we observe improvements of 31.8 % for arousal, 23.5% for dominance, and 9.5% for valence (-5dB 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 with freezing the text encoder, the fine-tuning approach improves performance by 72.2% for arousal, 91.6% for dominance, and 21.0% for valence under the -5dB SNR condition. Our solution is highly applicable to SER systems deployed in real-world applications. For example, systems can infer the testing environment from the *global positioning system* (GPS) information by using *geological information service* (GIS) mashups, such as OpenStreetMap [27]. 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.

- 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 2 describes studies relevant to SER in noisy conditions and text-guided training strategies. Section 3 describes the proposed approach, emphasizing the motivations and insights behind the TG-EAT framework. Section 4 provides the experimental setting, including the database, baselines, and implementation details. Section 5 presents the results, discussing the clear benefits of the proposed strategy. Finally, Section 6 concludes the paper, summarizing our study and providing future research directions inspired by the proposed approach.

2 Previous Work

2.1 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. [14] proposed to enhance noisy *low-level descriptors* (LLDs) for an SER model by using a *convolutional neural network* (CNN) with residual blocks. Pandharipande et al. [28]proposed to discard noisy frames to increase the noise robustness of an SER model by using a voice activity detection module. Leem et al. [29] 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 [30, 31, 32, 33, 34, 35], including Wav2Vec2.0 [1], HuBERT [2], and WavLM [3]. Such models have shown higher robustness against the small perturbation on the input speech than the traditional SER model with a Mel-spectrogram [32]. 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. [19] demonstrated that training a HuBERT-based SER model with noisy speech can highly improve the performance in low *signal-to-noise ratio* (SNR) conditions. Leem et al. [20] 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. [21] 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 *global positioning system* (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 *global positioning system* (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.

2.2 Text-Guided Training

As we discussed in Section 2.1, 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 [22]. 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. [22] used the following prompt template: *"A photo of a {label}"*. They calculate the
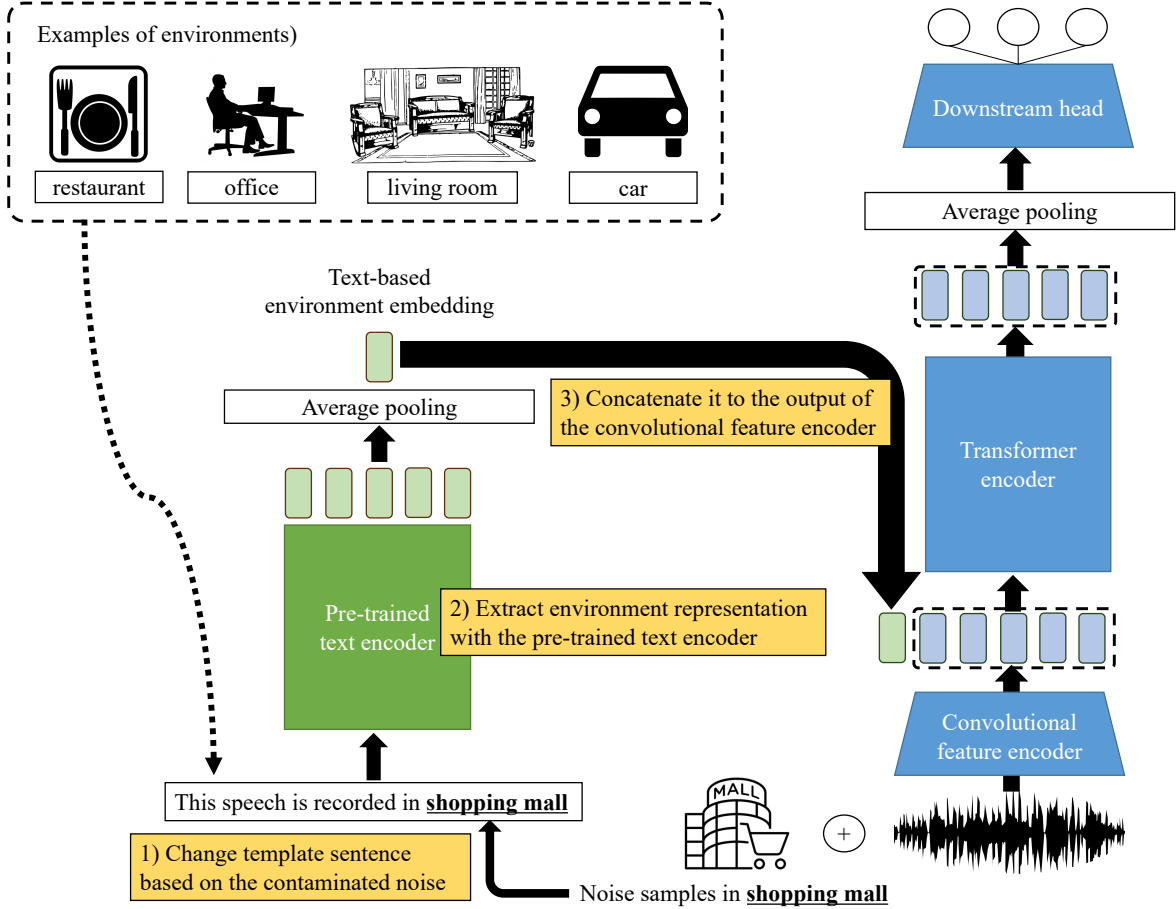
Figure 1: Our proposed text-guided environment-aware training framework. The environment representation is concatenated with the output of the convolutional feature encoder.

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. [23] demonstrated that pre-training audio and text encoder with natural language guidance could improve audio classification performance. The study of Elizalde et al. [24] 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. [25] used word embeddings to encode emotional labels for SER model. Gong et al. [26] used *large language model* (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.

3 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 denoise 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 [23, 24]. 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. [23]. For the LLM-based representation, we use the encoder from the pre-trained RoBERTa model [36]. 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 [37]. 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 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 [38, 32]. 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 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.

4 Experimental Settings

4.1 Data preparation

Our experiment uses the MSP-Podcast corpus, 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 [39], 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. We use 20 noisy environments to contaminate the training and development set, consisting of {mall, restaurant, office, airport, station, city, park, street, traffic, home, kitchen, living room, bathroom, bedroom, metro, bus, car, construction site, pedestrian, beach}. For the evaluation, we use six environments, including {plaza, garden, school, tram, sea, boat}. 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 if our proposed method can capture this semantic similarity during the 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.

4.2 Fine-Tuning Transformer-Based Architecture

We implement our proposed approach with two different pre-trained speech representation models: wav2vec2-large-robust [40] and the wavlm-base-plus models [3]. The wav2vec2-large-robust model has shown good performance in the emotional attribute prediction task [32]. The wavlm-base-plus model has shown good performance for emotion recognition in the *speech processing universal performance benchmark* (SUPERB) [41]. 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 [32]. We import the pre-trained models from the HuggingFace library [42]. 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 32 utterances per mini-batch and update the model for ten epochs. We use the Adam optimizer [43] with a learning rate warmup scheduling, which shows good performance when fine-tuning a pre-trained transformer architecture [44]. 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.

4.3 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 3. 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.

4.4 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 4.1, it uses 20 environmental conditions for adaptation. The evaluation uses six other environmental conditions.

Domain adversarial training (DAT): Inspired by Huang et al. 's work [45], 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.

Table 1: Average CCC of the ten experiments for the proposed text-guided environment-aware methods and the baselines. We report the performance with models implemented using either the wav2vec2-large-robust or wavlm-base-plus feature vectors. We denote with $*$, $\dagger$, and $\star$ when a model shows significantly better performance than the Original, RT, and DAT models, respectively. We highlight in bold the best performance per condition.

SNR	Model	Arousal	Dominance	Valence
		wav2vec2-large-robust		
5dB	Original	0.59	0.51	0.40
	RT	0.60	0.52	0.45*
	DAT	0.61	0.5	0.45*
	TG-EAT-CL	0.62*	0.52	0.46*
	TG-EAT-LLM	**0.63***	**0.53***	**0.47***
0dB	Original	0.53	0.47	0.33
	RT	0.56*	0.46	0.4*
	DAT	0.54	0.44	0.4*
	TG-EAT-CL	0.52	0.42	0.4*
	TG-EAT-LLM	**0.57****	**0.48**†*	**0.41***
-5dB	Original	0.27	0.25	0.14
	RT	0.24	0.22	0.19*
	DAT	0.24	0.22	0.16*
	TG-EAT-CL	0.21	0.2	0.18 *
	TG-EAT-LLM	**0.29***†*	**0.27***†*	**0.21***†*
		wavlm-base-plus		
5dB	Original	0.53	0.46	0.45
	RT	0.57*	0.48*	0.41
	DAT	**0.59***	**0.49***	**0.48***†
	TG-EAT-CL	0.57*	0.47	0.47*†
	TG-EAT-LLM	0.58*	0.48*	0.45†
0dB	Original	0.40	0.32	0.35
	RT	0.53*	0.43*	0.34
	DAT	0.53*	**0.45***	**0.42***†
	TG-EAT-CL	0.52*	0.43*	**0.42***†
	TG-EAT-LLM	**0.55***†*	**0.45***†	0.40*†
-5dB	Original	0.11	0.07	0.12
	RT	0.19*	0.12*	0.13
	DAT	0.22*†	0.17*†	0.21*†
	TG-EAT-CL	0.18*	0.12*	0.19*†
	TG-EAT-LLM	**0.29***†*	**0.21***†*	**0.23***†*

5 Results

5.1 Emotion recognition performance

We report the SER performance of our text-guided environment-aware training with our baselines. As described in Section 4.1, 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.

Table 1 illustrates the SER performance of each model in noisy testing environments. When comparing our baselines (RT, DAT) with the original model, they do not consistently yield performance improvement for all the attributes. RT does not improve the performance neither for arousal and dominance with the wav2vec2-large-robust feature vector, nor for valence with the wavlm-base-plus feature vector. Although the DAT shows significant performance improvement with wavlm-base-plus feature vector, it fails to improve arousal and dominance prediction performance with wav2vec2-large-robust feature vector. Since these baselines do not use environmental information, we can observe the importance of using environmental information when adapting the SER model to multiple noisy environments.
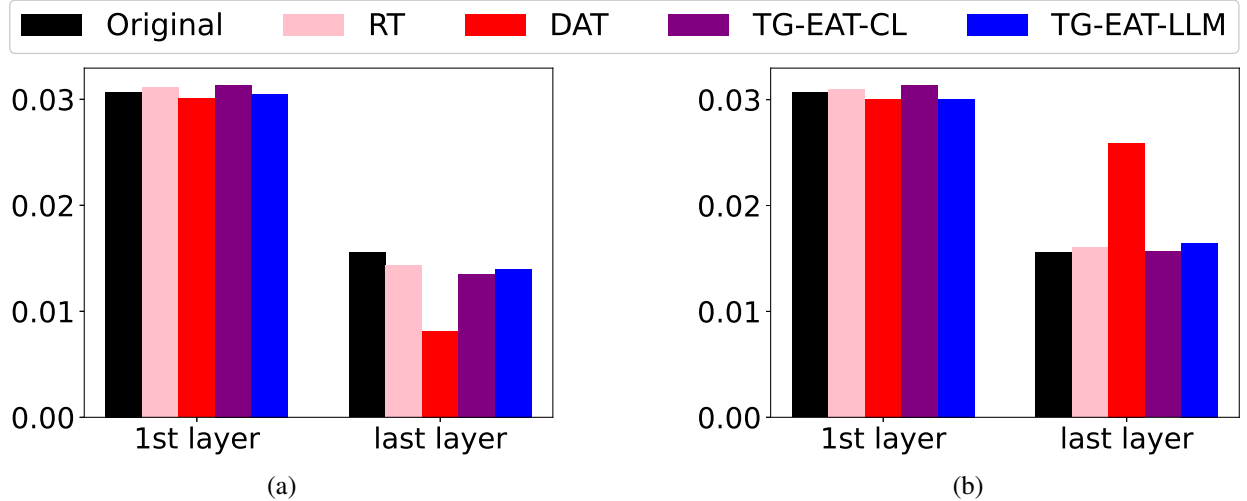
Figure 2: Embedding differences in the first and the last transformer encoder layers using clean and noisy speech in the -5dB condition. We use the wavlm-base-plus feature vector in this analysis. (a) it extracts and compares clean and noisy representations from each of the models. (b) it extracts and compares the clean representation from the Original model and the noisy representation from each of the models.

Compared with the baselines, our proposed TG-EAT-LLM performs best when using the wav2vec2-large-robust feature vector. In the 5dB condition, TG-EAT-LLM improves the original model's performance by 6.7 % (arousal), 3.9% (dominance), and 17.5% (valence). It does not yield the best performance with the wavlm-base-plus feature vector in the 5dB conditions. However, as the SNR level decreases, TG-EAT-LLM shows higher performance than the baselines. In -5dB condition, TG-EAT-LLM shows performance gains of 31.8% (arousal), 23.5% (dominance), and 9.5% (valence) compared to the best baseline, DAT. In spite of having a mismatch in SNR and environment conditions, TG-EAT-LLM shows robust results under all the conditions. 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. Although the DAT approach is effective when using the wavlm-base-plus model for noise conditions above 0dB SNR, using LLM-based representation is more helpful when dealing with low SNR conditions.

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 a audio-text pair.

5.2 Embedding analysis

Section 5.1 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 has robustness between 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.

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
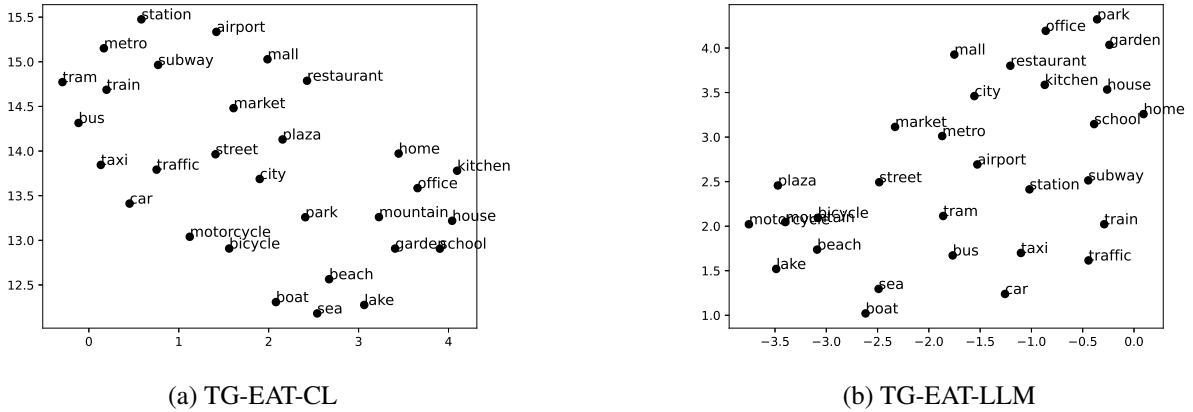
(a) TG-EAT-CL · (b) TG-EAT-LLM

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

Table 2: 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

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 noise-robustness of the SER system.

We also investigate if the proposed text-based environment embedding clusters similar environments together, which it the key premise of the proposed approach to deal with unseen environments. First, we extract the 26 different text embeddings using the different templates to describe each environmental condition. We project these embeddings into the 2D space to visualize the embedding space using the *uniform manifold approximation and projection* (UMAP) method [46]. Figure 3 illustrates the text embedding space of TG-EAT-CL and TG-EAT-LLM. The figure shows that both frameworks cluster environmental conditions that are semantically similar 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 deal with unseen environments by clustering acoustically and semantically similar environments.

Table 3: 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.

SNR	Model	Arousal	Dominance	Valence
5dB	Original	0.53	0.46	0.45
	RT	0.57	0.48	0.41
	GloVe	0.58	0.47	0.42
	AST	**0.60**	**0.49**	**0.48**
	TG-EAT-LLM	0.58	0.48	0.45
0dB	Original	0.40	0.32	0.35
	RT	0.53	0.43	0.34
	GloVe	0.53	0.43	0.37
	AST	**0.56**	**0.46**	**0.43**
	TG-EAT-LLM	0.55	0.45	0.40
-5dB	Original	0.11	0.07	0.12
	RT	0.19	0.12	0.13
	GloVe	0.25	0.17	0.20
	AST	0.25	0.18	0.21
	TG-EAT-LLM	**0.29**	**0.21**	**0.23**

5.3 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) [47], and *audio spectrogram transformer representation* (AST) [48]. One-hot uses 20-dimension 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 of words. We import the pre-trained GloVe vector collections consisting of 2.2 million vocabularies. We select the word vector representation that corresponds to the target noisy environment. The resulting representation is a 300-dimension vector. This representation can deal with unseen environments using 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 by using AudioSet, which is the same noise sound corpus for our training set. 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 One-hot in the seen environment scenario (Table 2) and with the other baselines in the unseen environment scenario (Table 3). For the seen environment scenario, we used the same environmental condition 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. Table 2 and 3 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 0dB and -5dB conditions. It also shows a better performance for valence in the 10dB 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 shows better performance for valence than our proposed model under the 5dB and 0dB conditions. In comparison, our proposed model performs better for all the emotional attributes under the -5dB 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. Considering that the -5dB SNR level is not presented while training the model, the result demonstrates that the AST works well for the seen SNR level but not for the unseen SNR level. In contrast, our proposed method works well for the unseen SNR level since the text description is independent of the SNR level. This result demonstrates that our proposed method is robust against the unseen SNR level, which is practical for real-world scenarios.

10

Table 4: 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	0.47
	TG-EAT-CL-FT	0.58	0.48	**0.49**
	TG-EAT-LLM	0.58	0.48	0.45
	TG-EAT-LLM-FT	0.58	0.48	0.46
0dB	TG-EAT-CL	0.52	0.43	0.42
	TG-EAT-CL-FT	0.56	0.46	**0.45**
	TG-EAT-LLM	0.55	0.45	0.40
	TG-EAT-LLM-FT	0.55	0.45	0.41
-5dB	TG-EAT-CL	0.18	0.12	0.19
	TG-EAT-CL-FT	**0.31**	**0.23**	**0.23**
	TG-EAT-LLM	0.29	0.21	**0.23**
	TG-EAT-LLM-FT	0.27	0.19	0.21

5.4 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 4 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. This observation illustrates the importance of compensating the embedding space gap 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.

6 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 denoise the speech representation 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. We assume that the CL-based representation can address the scenario when the 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.

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] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

12

[21] ——, "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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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