Computation and Memory Efficient Noise Adaptation of Wav2Vec2.0 for Noisy Speech Emotion Recognition with Skip Connection Adapters

Seong-Gyun Leem[1], *Daniel Fulford* [2], *Jukka-Pekka Onnela* [3], *David Gard*[4], *and Carlos Busso*[1]

[1]Department of Electrical and Computer Engineering, The University of Texas at Dallas
[2] Occupational Therapy and Psychological & Brain Sciences, Boston University
[3] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University
[4]Psychology Department, San Francisco State University

SeongGyun.Leem@utdallas.edu,dfulford@bu.edu,onnela@hsph.harvard.edu,dgard@sfsu.edu,busso@utdallas.edu

Abstract

An appealing approach for *speech emotion recognition* (SER) is to pre-train a large speech representation model, such as Wav2Vec2.0 or HuBERT. However, this large model should be adapted to different environments when deployed in real-world applications. This approach demands additional training time and stored parameters for each target environment. This paper proposes a computation and memory-efficient adaptation method. The approach trains skip connection adapters that generate environmental representations from the convolutional encoder, and denoise the self-supervised speech representations. Our experiments with the clean and contaminated versions of the MSP-Podcast corpus show that our adapter-based approach not only improves the performance of the original fine-tuned SER model, but also reduces the computation and memory requirements. For each environment, the approach requires 59.16% less adaptation time and only 0.98% of the parameters of the transformer encoder.

Index Terms: Speech emotion recognition in noisy environments, environment adaptation.

1. Introduction

The use of *speech emotion recognition* (SER) can serve as an important tool for real-world applications in different domains, including healthcare, marketing, education, and entertainment. Recent studies have shown that fine-tuning a large pre-trained transformer model, such as Wav2Vec2.0 [1] and HuBERT [2], performs well in SER tasks [3–5]. However, when deployed in a real-world scenario, a large transformer model must be adapted to noisy conditions to compensate for the environmental difference between the source and target domain.

Several studies have presented solutions to increase the robustness of SER models against noisy environments [6–8]. Most of these studies have not leveraged popular pre-trained self-supervised feature representations. Recently, some studies have started to leverage a large transformer encoder to be robust against noisy environments in speech tasks [9–11]. However, these methods still require a large amount of pre-training data. The adaptation methods also take time to update the transformer encoder, which has to be separately implemented for each target environment. In addition, a large number of parameters must be stored to deal with each environment. These limitations are crucial when using a large transformer model for SER deployed in a real-world application, where adapting to multiple noisy environments is essential. Some studies have explored minimizing the parameter requirements for adapting a large transformer architecture to multiple domains or tasks [12–14]. Such approaches can reduce the space requirements for domain-specific parameters in each adaptation condition. However, they still require large computation resources to adapt the models, since they need back-propagation through the transformer encoder, which includes *multi-head self-attention* (MHSA) layers.

This paper proposes environment-agnostic and environment-specific skip connection adapters for adjusting a large self-supervised speech representation model to multiple noisy environments. The environment-agnostic adapter learns the general characteristics of all non-speech background noises, while the environment-specific adapter learns the granular characteristics of the background noise in the target environment. Our approach avoids the back-propagation of gradients through the transformer encoder, since it updates both adapters without modifying the transformer encoder and the downstream head. Our method decreases the space requirements for an SER model to deal with each noisy environment by only storing the skip connection adapters.

We demonstrate the computation and memory efficiency of our proposed method, while improving the system's performance. We evaluate the skip connection adapters on the MSP-Podcast corpus and contaminated versions of its recordings with various noise sounds. Our experiments show that using skip connection adapters leads to 16.2% (arousal), 12.2% (dominance), and 9.46% (valence) performance improvements from the original model in the 0dB condition. Compared to adapting a transformer encoder per environment, our solution decreases the adaptation time by 59.16%, while only using the equivalent of 0.98% of the transformer encoder parameters.

2. Background

Various self-supervised speech representations have been proposed [15–17]. This work focuses on the Wav2Vec2.0 architecture, which performs well in SER and other speech-processing tasks [1]. While many studies have considered solutions for SER tasks in noisy environments [6–8, 11, 18–20], the focus of this section is on applications of Wav2Vec2.0 in SER tasks.

2.1. Wav2Vec2.0 Architecture and Pre-training Procedure

The Wav2Vec2.0 model leverages the raw waveforms as its input to generate frame-level contextual representations through a convolutional feature encoder and a transformer encoder. The convolutional feature encoder uses a *convolutional neural network* (CNN) to transform the raw waveform into latent speech representations. Such representations are fed into the transformer encoder, including MHSA layers [21] to generate frame-level contextual representations.

When pre-training the Wav2Vec2.0 model, some of the frames in the latent speech representations are masked and fed into the transformer encoder. The model is trained to minimize the contrastive loss, where the query is the contextual representation of the masked frame, its positive sample is the quantized

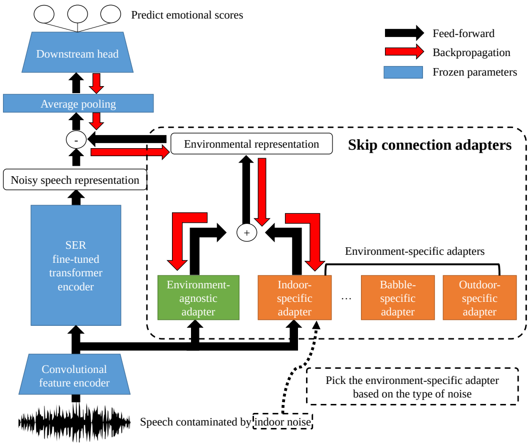

Predict emotional scores
Downstream head
Average pooling
Feed-forward
Backpropagation
Frozen parameters
-
Environmental representation
Skip connection adapters
Noisy speech representation
+
SER fine-tuned transformer encoder
Environment-specific adapters
Environment-agnostic adapter
Indoor-specific adapter
…
Babble-specific adapter
Outdoor-specific adapter
Convolutional feature encoder
Pick the environment-specific adapter based on the type of noise
Speech contaminated by indoor noise

Figure 1: *Our proposed environment adaptation method using skip connection adapters.*

latent representation of the masked frame, and its negatives are the quantized latent representations of the other frames. This pre-training objective does not require any downstream labels, enabling the utilization of a large amount of unlabeled speech data regardless of the downstream tasks.

2.2. Fine-tuning Wav2Vec2.0 to SER tasks

Although Wav2Vec2.0 was originally designed for *automatic speech recognition* (ASR) tasks, it also performs well in utterance-level SER tasks [3–5]. Unlike ASR, which requires a frame-level prediction, a common formulation for an SER task uses an average time pooling of the frame-level contextual representations to generate an utterance-level prediction. Such averaged representation is fed to the downstream head to predict categorical or dimensional emotional labels. Wang et al. [3] showed that only fine-tuning a transformer encoder with the categorical SER task yields better performance than freezing the Wav2Vec2.0 model or fine-tuning the whole model, including the convolutional feature encoder. This trend was also observed in dimensional SER tasks [5]. Wagner et al. [5] also verified that fine-tuning from the pre-trained parameters performs better than fine-tuning from randomly initialized parameters, even with the same architecture. All these studies indicate that both the pre-training and fine-tuning stages are essential for performance improvement when using a large pre-trained transformer model for the SER task.

3. Proposed Approach

The Wav2Vec2.0 model requires much time to fine-tune the large transformer encoder and memory spaces to store the parameters for each fine-tuned condition, which is problematic if the model needs to be adapted for multiple environments. To solve this issue, we aim to develop a computation- and memory-efficient adaptation scheme using Wav2Vec2.0 for SER tasks under multiple noisy conditions. Our main objective is to avoid the gradient back-propagation of the transformer encoder, T, which requires much time during adaptation and large memory space to store its parameters. Our solution should also improve the SER performance in multiple noisy conditions.

Figure 1 shows our proposed approach. We start from a Wav2Vec2.0 model fine-tuned with the SER task under a clean speech condition. After fine-tuning, we freeze all the fine-tuned parameters and attach the skip connection adapters, using the output of the convolutional feature encoder (E) as the input for these modules. The skip connection adapters with environment-agnostic and environment-specific modules transform these latent speech representations into the target environmental representations, as illustrated in Figure 1. The objective of the skip connection adapters is to denoise the speech representation of the transformer encoder to maximize the SER performance. This goal is achieved by subtracting the environmental representation generated by the skip connection adapters from the speech representation generated by the transformer encoder.

We may know the testing environments of each testing speech sample for real-world applications by exploiting domain knowledge or *global positioning system* (GPS) information. To utilize such prior knowledge, we use two skip connection adapters: the environment-agnostic adapter, $A_{agn.}$, which is updated using all types of background noises, and the environment-specific adapter, $A_{spe.}$, which is updated only using a specific type of background noise (e.g., vehicle noise). Using these two different adapters allows the environmental representations to learn the general characteristics that all noisy speech samples share and the granular characteristics conditioned by the given target environments. The environment-agnostic and environment-specific modules are added to create the environmental representation. Equation 1 illustrates the proposed operations:

$$z(\hat{x}^i) = T(E(\hat{x}^i)) - \{A_{agn.}(E(\hat{x}^i)) + A^i_{spe.}(E(\hat{x}^i))\} \quad (1)$$

where $\hat{x}^i$ denotes the noisy speech contaminated with the i-th type of background noise, and $z(\hat{x}^i)$ denotes the denoised speech representation. The resulting representation is fed into an average pooling layer to be used as an input to the fine-tuned downstream head. The environment-agnostic and environment-specific modules use the same architecture, consisting of a down-sample *fully-connected* (FC) layer, the attention module, and an up-sample *fully-connected* (FC) layer. The down-sample FC layer projects the frame-level latent speech representations from a 512-dimensional vector into a 256-dimensional vector. In our preliminary experiment on the development set, we observed that using the 256-dimensional feature representation preserves the emotion recognition performance while reducing the number of parameters. The attention module uses the same architecture of the single transformer module in the transformer encoder of the wav2vec2-large-robust network [1, 22], which consists of an MHSA layer, two FC layers implemented with the *Gaussian error linear unit* (GELU) as the activation function [23], and layer normalization. The only difference from the transformer encoder is that the embedding dimension of its MHSA layer is 256 instead of 1,024. The up-sample FC layer applies a linear projection to the output of the attention module. The features are projected from the 256-dimension vector to the 1024-dimension vector to match the dimension of the output of the transformer encoder.

In order to avoid the gradient back-propagation through the transformer encoder, we only update the skip connection adapters $A_{agnostic}$ and $A_{specific}$. We freeze the parameters of the convolutional feature encoder, transformer encoder, and downstream head during the adaptation stage. We update the skip connection adapters with noisy speech training samples, synthesized by manually adding the noise sounds to the clean emotional speech samples used in the fine-tuning stage.

This paper focuses on dimensional SER tasks to predict the emotional attributes for arousal (calm versus active), dominance (weak versus strong), and valence (negative versus positive). We train the adapters to maximize the *concordance correlation coefficient* (CCC) between the prediction from the noisy speech and the ground truth labels. With this objective function, the SER model learns to denoise the noisy contextual speech representation by only updating the adapters to maximize the CCC.

4. Experimental settings

4.1. The MSP-Podcast Corpus

We use the MSP-Podcast corpus [24] for the clean emotional speech corpus, which consists of natural and diverse emotional speech samples gathered from various podcast recordings. All the audios do not include background music or overlapped speech, and their predicted *signal-to-noise ratio* (SNR) is above 20dB. This study focuses on predicting the emotional attributes of arousal, dominance, and valence, annotated by at least five raters. These attributes are annotated with a seven Likert-scale. The ground truth values are estimated by averaging the scores provided by the raters for each speaking turn. We use release 1.8 of the corpus, which has 15,326 samples in the test set, 7,800 samples in the development set, and 44,879 samples in the train set. The partitions aim to create sets that are speaker independent. We use the train set to fine-tune the SER model. We use samples from the development set to select the best model during the fine-tuning process.

4.2. Noise Preparation

To simulate the various noisy environments, we collect noise sounds from six different conditions: radio, babble, indoor, outdoor, house, and vehicle. For the radio condition, we use the noise sounds collected from traditional radio shows without copyright containing human voices, background music, and various sound effects. For the test set, we use recordings from the noisy version of the MSP-Podcast corpus introduced by Leem et al. [7]. It was collected by playing speech and noise sounds with two portable speakers, and recording those mixed sounds on a smartphone in a single-walled sound booth.

For the train and the development set, we manually add the noise sound samples to the clean samples, which is more realistic than having a noisy parallel speech set in the target condition. To simulate the babble noise, we mix speech samples of seven speakers collected from the TIMIT dataset [25] and the CRSS-4English-14 corpus [26]. Krishnamurthy and Hansen showed that this simulation approach makes individual words indistinguishable, resulting in babble-like noise [27].

For the rest of the noisy conditions, we collect noise sounds from the Freesound repository [28], which contains publicly available ambient noise sounds. We use queries related to each environment to collect noise sounds. For example, we use {mall, restaurant, office, airport, school, station} queries for indoor, {city, park, street, traffic, construction, plaza} for outdoor, {home, kitchen, living room, bathroom, bedroom} for house, and {metro, bus, car, tram, boat, driving} for vehicle environments. We manually add the noise signals to the clean speech recordings of the MSP-Podcast corpus. We collect more than 120 hours of samples for each environment so that each speech sample in the clean MSP-Podcast corpus can be contaminated with a different noise sound. We directly transfer the emotional labels of the clean version of the MSP-Podcast corpus to the corresponding noisy speech samples. This study uses the 15,326 samples from the test set for each noisy condition to test the model in noisy conditions.

4.3. Fine-tuning Wav2Vec2.0 with clean speech

In this work, we build our base SER model by fine-tuning the transformer encoder of the wav2vec2-large-robust model [22] and the downstream head, which showed the best performance in dimensional SER tasks [5]. The wav2vec2-large-robust model consists of 24 transformer layers in the transformer encoder, pre-trained with diverse speech sets. We import the pre-trained wav2vec2-large-robust model from the HuggingFace library [29]. For the efficiency and reproducibility of this study, we prune the top 12 transformer layers from the model during the fine-tuning stage, which is shown to preserve the recognition performance with fewer parameters [5]. We use two fully connected layers for the downstream head, where each layer has 1,024 nodes, layer normalization, and the *rectified linear unit* (ReLU) as the activation function. We use dropout, with a rate set to $p = 0.5$, in all the hidden layers to increase regularization. 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.

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 Adam optimizer [30] with a learning rate of 0.00001. We use 32 utterances per mini-batch and update the model for 20 epochs. All of our experiments are conducted on a single NVIDIA GeForce RTX 3090.

4.4. Adaptation to Multiple Noisy Environments

After fine-tuning with the clean speech, we adapt the SER model to six different environmental conditions, as described in Section 4.2. For each mini-batch, we randomly select one of the six noise conditions, then use 32 samples of its noise sounds to contaminate 32 clean speech samples from the training set of the MSP-Podcast corpus. Since it is difficult to define the exact SNR level of the testing condition, we assume that the SNR level can be mismatched between the adaptation and test stages. For this reason, we randomly select the SNR level for the adaptation of the models among these options: {2.5, 7.5, 12.5}dB. For evaluation, the test set is created by randomly selecting the SNR level among these options {0, 5, 10}dB. We run five epochs to adapt the environment-specific and environment-agnostic adapters to all the noisy environments.

We refer to the proposed model as *skip connector adapters* (SCA). We compare this model with the fine-tuned model without adaptation, which we refer to as *Original*. We also use two extra baselines. The first baseline is the *retrained transformer* (RT), which retrains the downstream head and the transformer encoder for each environmental condition. Second baseline is the *retrained head* (RH), which only retrains the downstream head without updating the transformer encoder. This baseline does not use the proposed skip connection adapters. We train multiple downstream heads, each trained with a single type of environmental condition. As an ablation study, we also compare the implementation of the proposed model with only the environment-agnostic adapter (SCA-a), and with only the environment-specific adapters (SCA-s).

5. Experiment Results

5.1. Emotion Recognition Performance

We first compare the SER performance of our proposed adaptation strategy with the baselines in six noisy conditions (radio, babble, indoor, outdoor, house, and vehicle). Table 1 presents the results with three different SNR levels (10dB, 5dB, and 0dB). We run three experiments for each training method by changing the seed value, which changes the minibatch order

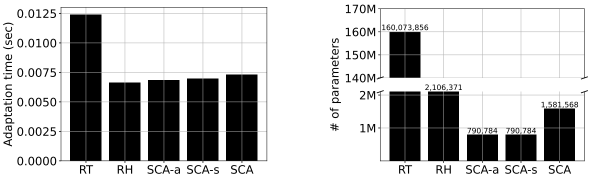
Table 1: *Average CCC of 18 experiments (3 trials × 6 noises) for each adaptation method under different SNR levels (10, 5, and 0dB). Clean denotes the performance of the original models without noise (∗ indicates results that are significantly better than the ones achieved by original model without adaptation).*

SNR	Model	Arousal	Dominance	Valence
Clean	Original	0.666	0.599	0.529
10dB	Original	0.596	0.562	0.473
	RT	0.634∗	**0.587**∗	**0.507**∗
	RH	**0.637**∗	0.553	0.484
	SCA-a	0.613	0.571	0.499∗
	SCA-s	0.629∗	0.580∗	0.464
	SCA	0.633∗	0.573∗	0.506∗
5dB	Original	0.526	0.506	0.424
	RT	0.581∗	**0.541**∗	0.453∗
	RH	**0.590**∗	0.535∗	0.433
	SCA-a	0.561	0.534	0.450∗
	SCA-s	0.581∗	0.536∗	0.410
	SCA	0.583∗	0.540∗	**0.461**∗
0dB	Original	0.432	0.418	0.338
	RT	0.492∗	0.465∗	0.369∗
	RH	0.497∗	0.458∗	0.349
	SCA-a	0.446	0.445	**0.371**∗
	SCA-s	0.473∗	0.460∗	0.307
	SCA	**0.502**∗	**0.469**∗	0.370∗

and the initial weights of the skip connection adapters. This process results in 18 values (6 environments × 3 trials) for each SNR level. We conduct a one-tailed matched-pair t-test of the original and the other models to evaluate if the adaptation strategy helps improve the performance of the original SER model in noisy conditions. We assert significance at p-value ≤ 0.05.

Table 1 reports the average CCC of 18 experiments for each adaptation strategy and the original model in 10dB, 5dB, and 0dB conditions. We also report the CCC performance of the *original* model tested with the clean version of the MSP-Podcast corpus (row labeled "Clean"). Compared with the *original* model, RT yields significant performance improvement in all the prediction tasks under 10dB, 5dB, and 0dB conditions. However, it does not always yield the best performance among all the tested model. For example, it does not show significant performance differences compared with the SCA model. This result shows that retraining the transformer encoder is not always the best method to adapt the model to noisy conditions. Although RH shows significant performance improvements for arousal in all the noisy conditions, these trends are not observed for valence. However, SCA-a and SCA always yield significant improvements for valence. The RH model only uses the output of the fine-tuned transformer encoder, indicating that relying only on the fine-tuned transformer encoder is insufficient to adapt the model to noisy environments.

In all the conditions, SCA-a shows significant improvement for valence, but not for arousal and dominance. For valence, SCA-a improves the *original*'s performance by 5.4% (10dB), 6.1% (5dB), and 9.7% (0dB). In contrast, SCA-s significantly improves the prediction for arousal and dominance, but not for valence. The SCA-s model shows 5.2% (10dB), 10.4% (5dB), and 9.4% (0dB) improvements for arousal, and 3.2% (10dB), 5.9% (5dB), and 12.2% (0dB) improvements for dominance. Such contrastive results are compensated for when using both adapters in our proposed approach. The SCA model achieves significant improvements for all the prediction tasks. In the 0dB condition, SCA yields 16.2% (arousal), 12.2% (dominance), and 9.46% (valence) performance improvements from the *original* model. This result demonstrates that using both environment-agnostic and environment-specific adapters is crucial for SER in noisy speech.



(a) (b)

Figure 2: *Analysis of the adaptation time and memory efficiency of each adaptation method. (a) illustrates each method's adaptation time normalized by the duration of the adaptation speech samples, and (b) illustrates the number of parameters needed for adaptation per environment.*

5.2. Computation and Memory Efficiency

We first evaluate the adaptation time for each adaptation method by measuring the time required for feed-forward, back-propagation, and parameter update during the adaptation in one epoch. We average the times over five epochs of three runs (i.e., 15 numbers), dividing the result by the total duration of the adaptation samples. Figure 2-(a) reports the results. Our SCA method requires 59.16% less adaptation time than the RT method, while achieving similar performance. The RH method requires 9.4% less adaptation time than SCA, but the performance is lower than our proposed method. These results demonstrate that SCA is a time-efficient adaptation method that increases the SER performance of the fine-tuned model.

We also check the number of parameters for each adaptation method to deal with each noisy environment. We first report the number of parameters for each method in a single environment in Figure 2-(b). Compared with RT, all the models require less than 2% of the number of parameters to deal with each noisy environment. Our SCA requires the equivalent of 0.98% of the number of parameters in RT for a single environment. When dealing with N environments, RT requires $160,073,856 \times N$ parameters, while SCA only requires $790,784 \times (N+1)$ parameters in addition to the pre-trained transformer model. This analysis shows that our proposed method is memory efficient for multiple noisy environments.

6. Conclusion

This paper proposed the skip connection adapters for computation- and memory-efficient noise adaptation of the Wav2Vec2.0-based SER model. We combined environment-agnostic and environment-specific adapters to avoid back-propagation through the transformer encoder during the adaptation. Our experiment verified that using skip connection adapters yields significant performance improvements over the original fine-tuned SER model for predicting arousal, dominance, and valence. The approach decreases the time and memory requirements to adapt the model to the new domain, compared to fine-tuning the entire transformer to the target noisy condition. In our future work, we plan to analyze why the environment-agnostic adapter helps the valence prediction, and the environment-specific adapter helps the arousal and dominance predictions. Studies have shown particular patterns in the expression of valence that make it different from arousal and dominance [31]. We also plan to work on applying this approach to HuBERT [2] and WavLM [32], which have also shown good performance in SER tasks. This model can also be useful for other speech processing tasks.

7. Acknowledgement

Study supported by NIH under grant 1R01MH122367-01.

8. 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, Dec. 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] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned Wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding," *ArXiv e-prints (arXiv:2111.02735)*, pp. 1–7, November 2021.

[4] A. Reddy Naini, M. Kohler, and C. Busso, "Unsupervised domain adaptation for preference learning based speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[5] 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," *ArXiv e-prints (arXiv:2203.07378)*, pp. 1–25, March 2022.

[6] 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.

[7] 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.

[8] 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, Sept.-Oct. 2021, pp. 1–8.

[9] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, "Wav2vec-Switch: Contrastive learning from original-noisy speech pairs for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7097–7101.

[10] Q.-S. Zhu, J. Zhang, Z.-Q. Zhang, M.-H. Wu, X. Fang, and L.-R. Dai, "A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 3174–3178.

[11] 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, p. 1.5.

[12] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proceedings of Machine Learning Research (PMLR 2019)*, K. Chaudhuri and R. Salakhutdinov, Eds. Vancouver, Canada: PMLR, June 2017, vol. 54, pp. 2790–2799.

[13] S. Kessler, B. Thomas, and S. Karout, "An adapter based pre-training for efficient and scalable self-supervised speech representation learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, Singapore, May 2022, pp. 3179–3183.

[14] B. Thomas, S. Kessler, and S. Karout, "Efficient adapter transfer of self-supervised speech models for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, Singapore, May 2022, pp. 7102–7106.

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

[16] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 146–150.

[17] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 3465–3469.

[18] Ł. 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *ArXiv e-prints (arXiv:1606.08415)*, pp. 1–8, June 2016.

[24] 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.

[25] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Timit acoustic-phonetic continuous speech corpus," 1993.

[26] F. Tao and C. Busso, "Bimodal recurrent neural network for audiovisual voice activity detection," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1938–1942.

[27] N. Krishnamurthy and J. H. L. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1394–1407, September 2009.

[28] 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.

[29] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[30] 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.

[31] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *ArXiv e-prints (arXiv:2201.07876)*, pp. 1–16, January 2022.

[32] S. Chen *et al.*, "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.