ADAPTING A SELF-SUPERVISED SPEECH REPRESENTATION FOR NOISY SPEECH EMOTION RECOGNITION BY USING CONTRASTIVE TEACHER-STUDENT LEARNING

Seong-Gyun Leem[⋆]*, Daniel Fulford* [†]*, Jukka-Pekka Onnela* [‡]*, David Gard*[⋆]*, and Carlos Busso*[⋆]

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

ABSTRACT

Studies have shown high performance in the *speech emotion recognition* (SER) task by fine-tuning a self-supervised speech representation model. Although this model can provide emotionally discriminative embedding in clean conditions, adapting it to a noisy target environment is still required when deployed on real-world applications. For adaptation, it is essential to balance between acquiring new knowledge from noisy speech and keeping the previous knowledge acquired during the pre-training and fine-tuning of the model. Therefore, we propose a contrastive teacher-student learning framework to retrain a self-supervised speech representation model for noisy SER. To keep the knowledge of the original model, we minimize the root mean square error between the clean embeddings from the original SER model and the noisy embeddings from the retrained model. To acquire the discriminative knowledge in the target noisy condition, we also minimize the InfoNCE loss by selecting the corresponding clean embedding as a positive sample and other noisy embeddings with different emotional labels as negative samples. Our experiment with the clean and noisy version of the MSP-Podcast corpus demonstrates that the contrastive teacher-student learning framework can significantly improve the performance of the model only trained with the clean speech in the target noisy condition for all the emotional attributes.

Index Terms— Speech emotion recognition, noisy speech, transfer learning, contrastive teacher-student learning

1. INTRODUCTION

Self-supervised speech representation models have been successfully adopted for speech processing tasks [1–4]. Self-supervised learning enables the pre-training of the speech representation models with a large amount of speech data without the need of annotating each speech recording with labels for the target task. Studies have used self-supervised learning to improve *speech emotion recognition* (SER) performance. Keesing et al. [5] showed that using the Wav2vec2.0 model for acoustic feature extraction led to better performance than using alternative acoustic features in categorical emotion recognition tasks. Wagner et al. [6] found that fine-tuning the wav2vec2-large-robust model [4] with the downstream head can highly improve the prediction of emotional attributes. Even though self-supervised speech representation models have shown good SER performance in clean conditions, those models are highly likely to generate disrupted embedding in real-world environments due to non-stationary background noises. Using noisy speech with these embeddings can provide inaccurate information to the downstream head, reducing SER performance. There are several approaches to improve SER performance in noisy conditions, including data augmentation [7, 8], domain adaptation [9–11], and feature selection [12, 13]. However, those studies have not investigated their method with self-supervised speech representation models.

Some studies have improved an *automatic speech recognition* (ASR) model by increasing the noise robustness of the self-supervised speech representation model. Zhu et al. [14] trained a feature encoder to generate similar embedding with clean and noisy speech. They applied a contrastive loss to the representation from transformers while pre-training the model. Wang et al. [15] changed the pre-training objective of the Wav2Vec2.0 model to predict the masked noisy embedding with clean speech and the masked clean embedding with noisy speech. Both studies increased the ASR performance in noisy conditions, indicating that improving the speech representation model can lead to performance improvements.

This paper proposes to adapt the self-supervised speech representation model for SER in the presence of noise by using a contrastive teacher-student learning method. Our main goal is to acquire new knowledge from adverse recording conditions without forgetting discriminative information acquired during the pre-training of the self-supervised speech representation model and fine-tuning of the SER model using the clean emotional speech data. We fine-tune the SER model with a clean version of the emotional speech (teacher). Then, we retrain the model with the training set corrupted with the target noise (student). To prevent catastrophic forgetting of the knowledge acquired in the teacher model, we train the student model to minimize the difference between clean embedding from the teacher model and noisy embedding from the student model. However, the model may lose the flexibility to acquire new knowledge of the target recording conditions. We address this issue by relying on the infoNCE loss to provide the student model with emotionally discriminative embedding with noisy speech. For each noisy speech, we make its embedding closer to the one extracted from its corresponding clean speech, while making it distant from the ones extracted from other noisy speech samples with different emotions.

In our experiments with the clean and noisy version of the MSP-Podcast corpus [16], our contrastive teacher-student learning improves the model only trained with the clean speech by 42.2% (arousal), 32.7% (dominance), and 48.5% (valence) in the 0db condition. To the best of our knowledge, this is the first study to improve the self-supervised speech representation model for the SER task. Instead of changing the pre-training objective, this is the first study to modify the fine-tuning strategy with noisy speech, which does not require a large amount of additional data for pre-training.

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

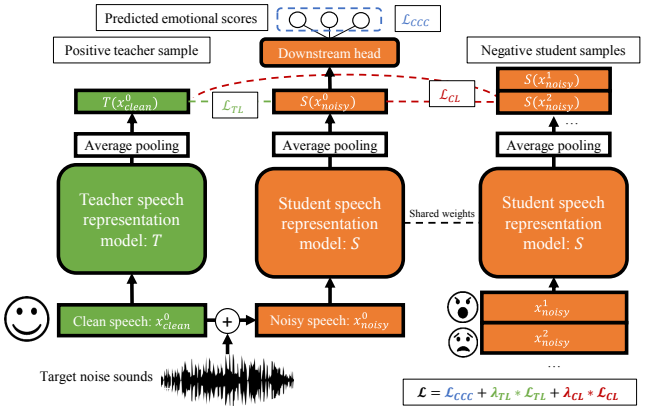


Fig. 1. Diagram of the contrastive teacher-student learning strategy to adapt a self-supervised speech representation model for SER in noisy conditions. The model uses the main task loss $\mathcal{L}_{CCC}$, transfer learning loss $\mathcal{L}_{TL}$, and the contrastive loss $\mathcal{L}_{CL}$.

2. PROPOSED METHOD

We propose a contrastive teacher-student learning framework to prevent catastrophic forgetting of the speech representation model during its adaptation to noisy conditions. Figure 1 illustrates the diagram of our proposed contrastive teacher-student learning framework. We consider the SER model only trained with clean speech as the teacher model (T), and the SER model that we want to retrain with noisy speech as the student model (S). To get noisy training samples for the student model, we contaminate the clean speech samples used to fine-tune the original SER model with the target noise sounds. With these noisy samples, we update the learned self-supervised representation model and the downstream head to predict the emotional label of its corresponding clean speech. We apply average pooling on the frame-level output of the self-supervised speech representation model, $S(x_{noisy})$, to feed it to the downstream head.

We use three loss functions to train the student model: one to predict emotional labels from noisy speech, another to keep the knowledge of the teacher model, and a third to acquire emotionally discriminative information from noisy speech. The first loss is $\mathcal{L}_{CCC}$, which corresponds to the primary loss associated with the SER task. This paper focuses on the prediction of the emotional attributes of arousal, dominance, and valence. Therefore, we train the model to maximize the *concordance correlation coefficient* (CCC) between the prediction from noisy speech and the ground truth labels, $\mathcal{L}_{CCC}$, illustrated in Equation 1,

$$\mathcal{L}_{CCC} = (1 - \mathcal{L}_{Aro}) + (1 - \mathcal{L}_{Dom}) + (1 - \mathcal{L}_{Val}) \quad (1)$$

where $\mathcal{L}_{Aro}$, $\mathcal{L}_{Dom}$, and $\mathcal{L}_{Val}$ correspond to the CCC between the prediction and the ground truth for arousal, dominance, and valence, respectively. There are benefits reported in previous studies on training a model to jointly predict multiple emotional attributes [17–19].

The second loss is $\mathcal{L}_{TL}$, which aims to create embeddings in noisy conditions that are similar to the embeddings from clean conditions. We aim to keep the knowledge acquired from the pre-trained self-supervised model and the fine-tuned model using the clean version of the emotional speech corpus. We decrease the distance between the student model and teacher model embeddings. When we train the student model, we additionally extract the embeddings from the teacher model with the corresponding clean speech ($T(x_{clean})$). We train the student model to minimize the *root mean square error* (RMSE) between the clean teacher and noisy student embeddings. We do not update the teacher model so that the student model can learn with a consistent representation. Equation 2 illustrates our transfer loss $\mathcal{L}_{TL}$,

$$\mathcal{L}_{TL} = \frac{1}{N}\Sigma_{i=0}^{N}\{\sqrt{(T(x^i_{clean}) - S(x^i_{noisy}))^2}\} \quad (2)$$

where N denotes the number of samples in each mini-batch, and x^i_{clean} and x^i_{noisy} denote the i-th clean and noisy speech samples, respectively.

The third loss is $\mathcal{L}_{CL}$, which is a contrastive loss to make the student model more discriminative in noisy conditions. We use the InfoNCE loss [20] so that the student model acquires contrastive emotional knowledge with clean and noisy speech. The idea is to decrease the distance in the embeddings of the speech representation for samples that have similar values for all the attributes, and increase it for samples that have different emotional labels, regardless of the environmental conditions. We set the corresponding clean speech as a positive sample and the other noisy speech with different ground truth within a mini-batch as negative samples for each noisy speech. As explained in Section 3.1, the emotional attributes are annotated with a seven-point Likert scale. We define *negative* samples if the difference in the emotional attributes between the positive sample and the other samples in the mini-batch is larger than 0.5. With the positive and negative samples for each noisy speech, we extract the teacher embedding from the clean positive sample and the student embeddings from the noisy negative samples. By using these embeddings, we train the student model to minimize the contrastive loss, $\mathcal{L}_{CL}$, which is illustrated in Equation 3,

$$\mathcal{L}_{CL} = -\frac{1}{N}\Sigma_{i=0}^{N} log \frac{exp(cos(T(x^i_{clean}), S(x^i_{noisy}))/\tau)}{\Sigma_{j=0}^{k} exp(cos(S(x^j_{noisy}), S(x^i_{noisy}))/\tau)} \quad (3)$$

where k denotes the number of negative samples for the i-th noisy speech and τ denotes the temperature coefficient to smooth the cosine similarity, which is set to 0.1 in this study.

Equation 4 illustrates the final objective of the proposed student model in our contrastive teacher-student learning, $\mathcal{L}$.

$$\mathcal{L} = \mathcal{L}_{CCC} + \lambda_{TL}\mathcal{L}_{TL} + \lambda_{CL}\mathcal{L}_{CL} \quad (4)$$

where λ_{TL} and λ_{CL} are the coefficients to weigh the transfer loss and the contrastive loss. These hyper-parameters balance the objective function between keeping the original knowledge and acquiring new discriminative knowledge.

3. EXPERIMENT SETTINGS

3.1. The MSP-Podcast Corpus

We use the MSP-Podcast corpus to train and evaluate the proposed model [16]. The corpus contains natural and diverse emotional speech samples collected from various audio recordings. All the audios are selected to have duration between 2.75s and 11s, such that they do not include background music or overlapped speech. The data collection protocol also considers samples with predicted *signal-to-noise ratio* (SNR) above 20dB. This study focuses on predicting the emotional attributes of arousal, dominance, and valence. The corpus is annotated with emotional labels using a crowdsourcing protocol inspired by the study of Burmania et al. [21]. The perceptual score for each emotional attribute was annotated by at least five raters using a seven-point Likert scale with values between 1 and 7. The ground truth values are estimated by averaging the scores provided by the raters to 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. 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.

The study also uses the noisy version of the MSP-Podcast corpus, which was introduced in Leem et al. [9]. We simultaneously played speech and noise sounds with two portable speakers, recording those mixed sounds on a smartphone in a single-walled sound booth. We use noise sounds collected from traditional radio shows without copyright containing human voices, background music, and various types of sound effects. We collect three noisy conditions at different levels of SNR by calibrating the distance between two speakers and the smartphone with the estimated SNR of each condition. According to their estimated SNR, we name these three conditions as 10dB, 5dB, and 0dB, respectively. For the emotional label, 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.

Since it is impractical to directly get a noisy parallel speech set in the target condition, we do not use the train set in the noisy version of the MSP-Podcast corpus to retrain the model for the target condition. Instead, we contaminate the clean train set with 30 minutes of noise sound samples. We use only 30 minutes, since collecting noisy recordings with a longer duration from the target environment is less practical for real applications. Those noise samples are collected using the same setting used in the recording of the noisy version of the MSP-Podcast corpus by recording just noise sounds without playing speech samples from the corpus.

3.2. Fine-Tuning Self-Supervised Model with Clean Speech

We use the Wav2vec2.0 architecture [1] that has shown good performances for SER tasks [6,22]. Among the variants of the Wav2vec2.0 model, we use wav2vec2-large-robust, pre-trained with databases from multiple speech domains [4]. Wagner et al. [6] reported the best SER performance with this model. We prune the top 12 transformer layers from the model, since this approach preserves the recognition performance, while reducing the number of parameters [6]. We aggregate the outputs of the Wav2vec2.0 model across frames by using average pooling per utterance. This vector is processed by the downstream head, implemented with two hidden layers, where each layer has 1,024 nodes. We use the *rectified linear unit* (ReLU) activation function, and a linear output layer with three nodes for the downstream head. We apply dropout, with a rate set to $p = 0.5$, and layer normalization [23] for all the hidden layers of the downstream head to regularize the model. During fine-tuning, we use the Adam optimizer [24] with a learning rate set to 0.00001. We use 32 utterances per mini-batch and update the model for 20 epochs, after which we select the model with the best development set performance.

3.3. Contrastive Teacher-Student Implementation

The contrastive teacher-student model is implemented by adapting the fine-tuned SER model to the target noisy condition. We copy and retrain the fine-tuned SER model to use it as a student model. We set the 0dB condition in the noisy version of the MSP-Podcast corpus as our target condition. We use the parallel version of the train set contaminated with the target noise sound at 0dB SNR level. By randomly extracting segments from the 30 minutes of the target noise, we change the noise sounds to contaminate each clean speech for each epoch, preventing it from being overfitted to a fixed noisy speech set. We also contaminate the clean speech in the development set using the same approach.

With these noisy speech samples, we update the speech representation model and the downstream head with the objective of our contrastive teacher-student learning, $\mathcal{L}$. We set λ_{TL} and λ_{CL} to 100 and 10. These conditions showed the best performance in the noisy development set. We use the same optimizer and hyper-parameters as the ones used for fine-tuning the model with clean speech. Once the adaptation is completed, we use the student model to evaluate the SER performance in the matched SNR condition (0dB), as well as in the mismatched SNR condition (10dB, 5dB).

3.4. Baseline models

Original: This model fine-tunes the model with clean emotional speech, with no adaptation to the noisy condition.
Retrain head (RH): This baseline retrains the model without the proposed transfer learning and contrastive losses (e.g., $\mathcal{L}_{TL}$, and $\mathcal{L}_{CL}$). We only update the downstream head with the noisy speech, freezing the parameters of the fine-tuned speech representation model. Although this model can keep the original representation learned from clean data while pre-training and fine-tuning the model, it restricts its capability to acquire new knowledge from the noisy condition.
Retrain entire model (RM): This baseline updates the speech representation model and the downstream head with noisy speech. This method acquires new knowledge from noisy speech. However, there is no constraint for preventing the catastrophic forgetting of the knowledge previously learned by the SER model.
Retrain with transfer learning loss (RM +TL): This baseline trains the model without using the InfoNCE loss, $\mathcal{L}_{CL}$, to assess the effectiveness of including the proposed contrastive learning strategy.
Retrain with contrastive loss (RM +CL): This baseline train the model without using the transfer learning loss, $\mathcal{L}_{TL}$.

The complete contrastive teacher-student learning strategy is referred to as RM+TL+CL.

4. RESULTS

4.1. Emotion Recognition Performance

We compare the SER performance of each training strategy in clean and noisy conditions. For each training method, we select three models that showed the best performances in the development set, without selecting models in consecutive epochs during training. We also split the test set into 30 groups for each condition. This process results in 90 values (3 models $\times$ 30 test sets), over which we conduct statistical analysis of the results. We conduct a one-tailed Welch's t-test between the original and the other models to check if the training strategy helps to improve the performance of the original SER model in noisy conditions. We assert significance at p-value ≤ 0.05.

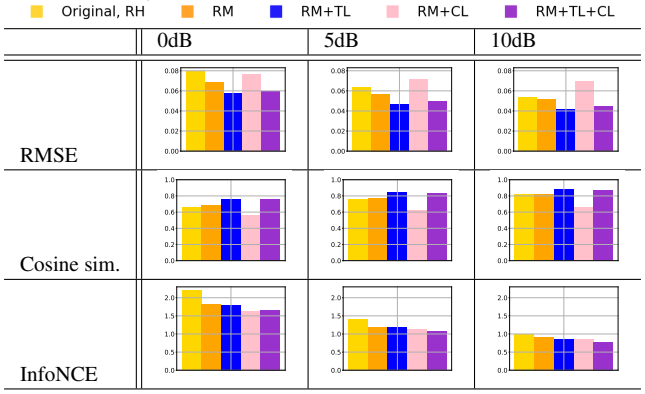
Table 1 shows the average CCC across the 90 values per model and SNR condition. As the SNR level is reduced in the test set, the performance of the original SER model drops in all the attributes. Therefore, adapting the SER model to the noisy condition is essential, even when using a self-supervised speech representation model. When comparing RH and RM with the original model, we find that neither model consistently improves the performance of each attribute. While RH increases the arousal and dominance performance, it lowers the performance of valence. RM improves valence performance, but it does not significantly improve the predictions for arousal and dominance. These results illustrate the limitations of freezing the original speech representation for noisy SER tasks and show that retraining the model with noisy speech without applying any constraints is not an appropriate method for noisy SER tasks.

Compared with the original model, our contrastive teacher-student learning strategy shows significant improvements for all the attributes. When the model is trained with either the transfer learning loss or the contrastive loss, we observe significant improvements for arousal and dominance over the model that is just retrained with

Table 1. Average CCC for the proposed contrastive teacher-student learning model (RM+TL+CL) and the baselines, including the model implemented without the contrastive loss (RM+TL), or without the transfer learning loss (RM+CL). The results are reported for arousal (Aro.), dominance (Dom.), and valence (Val.) across noise conditions. We denote with *, †, and * when a model shows significantly better performance than the original, RH, and RM models, respectively. We highlight in bold the best performance per condition.

	Matched condition			Mismatched condition						Clean condition		
	0dB			5dB			10dB			Clean		
	Aro.	Dom.	Val.	Aro.	Dom.	Val.	Aro.	Dom.	Val.	Aro.	Dom.	Val.
original	0.244	0.226	0.227	0.379	0.335	0.308	0.438	0.374	0.351	**0.586**	**0.506**	0.473
RH	0.323*	0.278*	0.164	0.443*	0.390*	0.236	0.494	0.424*	0.278	0.549	0.433	0.418
RM	0.272	0.215	0.330*	0.412	0.328	**0.418***	0.459	0.350	0.454*	0.546	0.429	0.498
RM+TL	0.345**	0.289**	**0.337***†	0.474**	0.403**	0.416*†	0.513*	0.430**	**0.459***†	0.565	0.472	**0.505**†
RM+CL	0.339**	0.296**	0.314*†	0.451**	0.391**	0.399*†	0.487*	0.402**	0.433*†	0.574	0.485	0.495†
RM+TL+CL	**0.347****	**0.300****	0.335*†	**0.477****	**0.410****	0.417*†	**0.523****	**0.435****	0.452*†	0.566	0.472	0.492†

Table 2. Analysis of clean and noisy embeddings created by the proposed and baselines models. The table shows the RMSE and cosine similarity and InfoNCE loss across different noise conditions.



noisy speech. The models RM+TL and RM+CL also show significant improvements for valence over the RH model. Table 1 shows that combining these two losses yields the best performance for arousal and dominance, even in mismatched noise conditions (i.e., using an SNR level that is different from 0dB SNR, which is the noise level used to adapt the model). Compared with the original model, our contrastive teacher-student learning strategy shows significant improvement for all the attributes. In the matched condition, our contrastive teacher-student learning approach yields a 42.2% gain for arousal, 32.7% gain for dominance, and 48.5% gain for valence. These results clearly show that our transfer learning loss and contrastive loss can prevent catastrophic forgetting of learned information and acquire new knowledge from noisy speech.

4.2. Embedding Comparison

In section 4.1, we demonstrated that our contrastive teacher-student learning strategy improves SER performance for all the attributes. This section analyzes the embedding differences and contrastive losses between clean and noisy speech to validate that the robust representation causes such improvements. We use clean and noisy speech samples in the development set of the clean and noisy version of the MSP-Podcast corpus. We compare the RMSE and cosine similarity between the embedding extracted from the clean speech with the original model and the embedding from the noisy speech with the retrained model to assess if the learned representation can keep the knowledge acquired by the original model. We also calculate the infoNCE loss to assess if the learned representation can provide emotionally discriminative information regardless of the recording conditions. Similar to the approach used to estimate the contrastive loss described in Section 2, we select 32 clean and noisy speech pairs to define positive and negative samples. For each noisy speech, we select the corresponding clean speech as a positive sample. The negative samples are the other noisy speech samples with different emotional labels in the set. We repeat this process multiple times until we use all the samples in the development set.

Table 2 illustrates the result of our analysis. When comparing the difference between the clean and noisy embeddings, applying the transfer loss reduces the distance in the embeddings (i.e., RM+TL, RM+TL+CL). Those two models show less differences between clean and noisy representations than the original model and the retrained model without using the transfer loss (i.e., RM, RM+CL). It shows the same trends even in the mismatched SNR condition (10dB, 5dB), indicating that making the noisy embedding closer to the clean embedding leads to SER performance improvement for all the attributes. The analysis of the InfoNCE loss shows that using the contrastive loss (i.e., RM+CL, RM+TL+CL) leads to lower losses than the other models. In the 0dB condition, the RM+CL model shows 26.0% lower infoNCE loss than the original model, and 10.1% lower infoNCE loss than the RM model. By combining the transfer loss with the contrastive loss (i.e., RM+TL+CL), it further decreases the loss in the mismatched SNR conditions (10dB, 5dB). Only applying the transfer loss (i.e., RM+TL) does not lead to a clear difference in the infoNCE loss, when compared with the RM model. Generating emotionally contrastive and noise-robust representation further improves the predictions for arousal and dominance.

5. CONCLUSIONS

This paper proposed a contrastive teacher-student learning strategy to adapt the SER model with a self-supervised speech representation model to the noisy condition. Our method retrains the SER model with noisy speech using a transfer learning loss and a contrastive loss that aim to keep the knowledge learned with the pre-trained model and the fine-tuning process, while still acquiring new contrastive knowledge using the noisy speech in the target condition. Our experiments with the clean and noisy version of the MSP-Podcast corpus showed that contrastive teacher-student learning improves the performance of a fine-tuned SER model for arousal, dominance, and valence in noisy conditions. The improvements were observed even in mismatched SNR conditions (5dB, 10dB) that are different from the target noise condition used to adapt the model (0dB). The contrastive teacher-student learning strategy can generate a robust embedding for noisy speech, leading to performance improvement.

Our method still relies on a parallel corpus with noisy speech, which requires a large number of recordings with the target noise to adapt the model. We plan to investigate alternative implementations that use non-parallel unlabeled noisy speech samples to further improve our contrastive teacher-student learning framework.

6. 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)*, Virtual, December 2020, vol. 33, pp. 12449–12460.

[2] W.-N. Hsu, Y.-H. H. Tsai 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] 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.

[5] A. Keesing, Y.S. 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.

[6] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B.W. 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.

[7] 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.

[8] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 7194–7198.

[9] 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.

[10] A. Wilf and E. Mower Provost, "Dynamic layer customization for noise robust speech emotion recognition in heterogeneous condition training," *ArXiv e-prints (arXiv:2010.11226)*, pp. 1–5, October 2020.

[11] 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.

[12] 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, May 2006, ISCA.

[13] 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.

[14] 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 2022)*, Singapore, May 2022, pp. 3174–3178.

[15] 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.

[16] 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.

[17] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

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

[19] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[20] 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.

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

[22] 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.

[23] L.J. Ba, J.R. Kiros, and G.E. Hinton, "Layer normalization," *ArXiv e-prints (arXiv:1607.06450)*, pp. 1–12, July 2016.

[24] D.P. 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.