ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887569

Noise-Robust Speech Emotion Recognition Using Shared Self-Supervised Representations with Integrated Speech Enhancement

Jing-Tong Tzeng[1], *Seong-Gyun Leem*[2], *Ali N. Salman*[2], *Chi-Chun Lee*[1,3], *Carlos Busso*[2,4]
[1]College of Semiconductor Research, National Tsing Hua University, Taiwan
[2]Department of Electrical and Computer Engineering, The University of Texas at Dallas, USA
[3]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[4]Language Technologies Institute, Carnegie Mellon University, USA

***Abstract*—Recent studies have demonstrated the effectiveness of fine-tuning self-supervised speech representation models for *speech emotion recognition* (SER). However, applying SER in real-world environments remains challenging due to pervasive noise. Relying on low-accuracy predictions due to noisy speech can undermine the user's trust. This paper proposes a unified self-supervised speech representation framework for enhanced speech emotion recognition designed to increase noise robustness in SER while generating enhanced speech. Our framework integrates *speech enhancement* (SE) and SER tasks, leveraging shared *self-supervised learning* (SSL)-derived features to improve emotion classification performance in noisy environments. This strategy encourages the SE module to enhance discriminative information for SER tasks. Additionally, we introduce a cascade unfrozen training strategy, where the SSL model is gradually unfrozen and fine-tuned alongside the SE and SER heads, ensuring training stability and preserving the generalizability of SSL representations. This approach demonstrates improvements in SER performance under unseen noisy conditions without compromising SE quality. When tested at a 0 dB *signal-to-noise ratio* (SNR) level, our proposed method outperforms the original baseline by 3.7% in F1-Macro and 2.7% in F1-Micro scores, where the differences are statistically significant.**

***Index Terms*—Speech emotion recognition, speech enhancement, noisy speech, multitask learning.**

I. Introduction

Speech emotion recognition (SER) automatically identifies human emotional states through speech signals. It is a critical element of *human-computer interaction* (HCI), enabling machines to engage in emotionally-aware communication with humans [1], [2]. In recent years, advancements in SER systems have been facilitated by the availability of large emotional speech datasets [3]–[7] and the integration of pre-trained speech representation models [8]–[10]. These developments have significantly enhanced the performance of SER systems [11]–[13].

As SER systems are increasingly deployed in real-world applications, such as digital assistants [14], the issue of non-stationary background noise has become more critical, significantly degrading system performance. Studies have proposed various approaches to address this challenge, including data augmentation [15], feature selection [16], [17], domain adaptation [18], and pre-processing the speech signal using *speech enhancement* (SE) [19]. While these methods improve SER performance in noisy conditions, most of them, except for SE pre-processing, do not provide cleaned speech, limiting their applicability; for example, in critical applications like emergency response systems [20], where accurate emotion recognition and the availability of intelligible speech are both needed or in the customer call center, having enhanced speech is useful for human intervention beyond machine-generated emotion labels. However, incorporating SE as a front-end module often substantially increases the overall model size, requiring significant computational resources and complicating its implementation. Since SE aims to increase intelligibility rather than SER performance, the aliasing that occurs during the generation of enhanced audio can affect the discriminative acoustic features for SER and limit its performance.

Given the promising results of speech *self-supervised learning* (SSL) pre-trained representations across multiple tasks, recent studies have explored their application to SE tasks and show promising results [21]–[23]. These studies pave the way to investigate the feasibility of combining SER and SE tasks using a shared SSL model, as SER studies have mainly focused on using transformer-based speech representation models [11]–[13]. The ideal implementation is for SER and SE tasks to share the same SSL representation, building a system that enhances discriminative cues relevant to the SER task.

This study proposes a unified self-supervised speech representation framework for enhanced SER to improve the robustness in noisy environments. By integrating SE and SER modules with shared SSL representations, we reduce the model size and leverage the shared information learned from both tasks, leading to improved performance. By adding the SER task, the SE goal is to improve intelligibility and enhance discriminate information to recognize emotions.

Our experiments with the MSP-Podcast [4] corpus show that combining SE and SER helps the model learn more generalized features, making it more robust to unseen noise conditions. In the 0 dB *signal-to-noise ratio* (SNR) condition, our method improves the emotion classification model trained

This study was funded by the NSF under grant CNS-2016719.

with clean speech by 7.3% for F1-Macro and 4.5% for F1-Micro scores. Compared to our best baseline, we observe improvements of 2.1% for F1-Macro and 3.0% for F1-Micro. Additionally, the proposed method shows competitive results with models solely fine-tuned for SE, confirming that combining SE and SER enhances noise robustness without sacrificing SE performance, especially in unseen conditions. This dual-task capability expands potential application scenarios while fostering user trust. The main contributions of this study are:

- We explore using multitask learning, combining SE and SER, to enhance noise robustness, demonstrating improved performance on unseen noise types and levels.
- We adopt a cascade unfrozen training strategy for fine-tuning the SSL model in a multitask setting to stabilize the training process and balance both tasks.

II. Related Work

A. SER in noisy speech

Noise robustness in SER systems has become increasingly important for real-world applications. Lakomkin *et al.* [24] and Wu *et al.* [15] used data augmentation strategies to close the gap between training and real-world conditions. Leem *et al.* [16], [25] focused on enhancing noise robustness by selecting robust *low-level descriptors* (LLDs) and strengthening weaker ones. However, these methods do not provide clean speech, limiting the practical use of these systems. Triantafyllopoulos *et al.* [19] incorporated an SE front-end to improve SER in low SNR, while Chen *et al.* [26] introduced SNR-level detection to mitigate SE's impact on clean speech. However, these two-stage approaches are resource-intensive, so they are not optimal for deployment.

B. Speech self-supervised representations for SE

SSL representations have shown promise for various tasks, including SE. Huang *et al.* [21] demonstrated that SSL models outperform traditional features such as STFT in speech enhancement. Song *et al.* [23] used WavLM representations to compensate for limited data, while Hung *et al.* [22] integrated weighted SSL representations with spectrogram features to deal better with the noise.

SSL representations have significantly improved SER tasks [11]–[13], [27]. Given the progress in applying SSL representations to SE tasks, our method explores the potential of integrating both modules using shared SSL representations. This approach successfully reduces the overall model size and avoids the potential aliasing issues that can arise when generating enhanced speech.

III. Proposed Method

A. Unified SSL Speech Representation for Enhanced Speech Emotion Recognition

This paper proposed a unified self-supervised speech representation for enhanced SER. Figure 1 illustrates our proposed framework. We integrate SE and SER heads with shared speech SSL representations. Let S_θ^l is the representation of layer $l \in \mathbb{N}$ with the SSL model parameterized by θ. Let

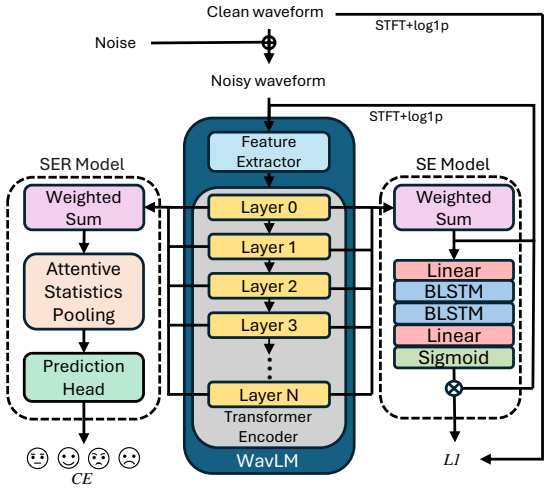

Fig. 1. Framework of our proposed unified self-supervised speech representations for enhanced speech emotion recognition.

$x_{clean} \in \mathbb{X}_{clean}$ and $x_{noisy} \in \mathbb{X}_{noisy}$ denote the clean and noisy speech, paired with the SER task labels $y \in \mathbb{Y}$. We apply a weighted sum of features from each layer to fully exploit the information captured by the SSL model. Specifically, for SER, we define the weighted sum as $F_{\theta_{ser}} = \sum_{l=0}^{L-1} w_{l_{ser}} S_\theta^l(x_{noisy})$, where $w_{l_{ser}}$ denotes the weight assigned to each layer for the SER task. Similarly, for SE, we define $F_{\theta_{se}} = \sum_{l=0}^{L-1} w_{l_{se}} S_\theta^l(x_{noisy})$, with $w_{l_{se}}$ representing the corresponding weights for SE. The SER head, parameterized by θ_{SER}, is denoted as $H_{\theta_{SER}}(Pooling(F_{\theta_{ser}})) : \mathbb{X}_{noisy} \rightarrow \mathbb{Y}$. For the SE task, we concatenate the spectral representation X, derived from the magnitude component of the STFT and compressed using the $log1p$ function $(log1p(z) = log(1+z))$ [28], with the weighted sum of features from all layers to enrich the information. The SE head, parameterized by θ_{SE}, is denoted as $H_{\theta_{SE}}(F_{\theta_{se}}, X_{noisy}) : \mathbb{X}_{noisy} \rightarrow \mathbb{X}_{clean}$.

We optimize the model using two loss functions: a weighted multi-class cross-entropy loss $\mathcal{L}_{WCE}$ to handle class imbalance in predicting emotional labels and a $\mathcal{L}_1$ loss for reconstructing clean speech. The objective of our approach is to simultaneously enhance speech and perform robust SER on x_{noisy}, formally expressed as:

$$\min_{\theta,\theta_{SER},\theta_{SE}} \mathcal{L}_{WCE}(H_{\theta_{SER}}(Pooling(F_{\theta_{ser}})), y) + \mathcal{L}_1(H_{\theta_{SE}}(F_{\theta_{se}}, X_{noisy}), X_{clean}) \quad (1)$$

The first term evaluates the performance of speech emotion classification, and the second term assesses the quality of the enhanced spectral representation. When we combine these losses, we expect the SE model to improve not only the signal's intelligibility but also the SER discriminative information in the features for the SER task.

B. Cascade Unfrozen Training

SSL leverages large amounts of unlabeled data to extract meaningful representations. To prevent these features from being misled by the back-propagation from randomly initialized task heads and to avoid over-fitting to a specific task, we

adopt the *cascade unfrozen training* (CUT) strategy. In this approach, we first train the SE and SER models with the SSL model's parameters frozen. Once these models are stable, we unfreeze the SSL model and fine-tune all components. This method helps stabilize the training and ensures that the SSL representations remain general and effective for both tasks.

IV. Experimental settings

A. Data preparation

We train and evaluate the proposed model using the MSP-Podcast corpus [4], which contains natural and diverse emotional speech samples collected from various podcast recordings. The selected audio segments range from 2.75 to 11 seconds in duration, ensuring they are free of background music or overlapping speech. The recordings have a predicted SNR above 20 dB, making this corpus a reliable, clean emotional speech database. In this study, we focus on predicting four categorical emotions: anger, sadness, happiness, and neutral state. We use version 1.11 of the corpus, which comprises 100,896 annotated utterances across these emotion classes (Anger: 10,342; Sadness: 8,347; Happiness: 29,454; Neutral: 52,753). We use the training set to fine-tune the pre-trained speech representation model. We employ the development set to select the best model during fine-tuning.

To simulate real-world noisy environments, we overlay speech from the CRSS-4ENGLISH-14 corpus [29] and generate babble noise, contaminating the training and development sets with an SNR level of 5 dB. We contaminate the Test 1 set of the MSP-Podcast corpus for evaluation with three different SNR levels: 0 dB, 5 dB, and 10 dB. Additionally, we collect ambient noise sounds from the Freesound repository [30] to assess the robustness of our method to unseen noise. Similarly, we contaminate the Test 1 set with the same three SNR levels.

B. Fine-tuning Self-Supervised Model

In this study, we implement our proposed method using the WavLM Large model [10], which has demonstrated top performance in SER tasks within the *speech processing universal performance benchmark* (SUPERB) [13], [31]. This model is pre-trained with noisy speech, making it likely to produce more robust representations compared to other pre-trained speech SSL models. For the SE task, we use the BSSE-SE model [22]. For the SER task, we employ the baseline model from the Odyssey 2024 Speech Emotion Recognition Challenge [13].

We apply Z-normalization to the raw waveform during fine-tuning, estimating the mean and standard deviation across the entire training set. Initially, we freeze the SSL model and train the SE and SER heads separately. The SE head is trained with a batch size of 16 for 130 epochs using the AdamW optimizer [32] with a learning rate of 5×10^{-5}. The SER head is trained with a batch size of 32 for 20 epochs using the same optimizer and learning rate. After this, we fine-tune the entire pipeline using the pre-trained head weights, with a batch size of 32 for 20 epochs. We continue using the AdamW optimizer, with a learning rate of 5×10^{-5} for the task heads and 2.5×10^{-5} for the SSL model for another 20 epochs. Throughout all training steps, the WavLM CNN feature extractor remains frozen.

C. SER baselines

We compare our proposed method with four SER baselines:

- **Original**: Fine-tunes the SER model using clean emotional speech without adapting to noisy conditions.
- **SE Pre-process (SE-P)**: Utilizes speech enhancement as a pre-processing step and fine-tunes the SER model with the enhanced speech. The SE model has been fine-tuned on the MSP-Podcast corpus from the pre-trained weight trained on the VCTK-DEMAND dataset [33].
- **Fine-tuning Head (FT-H)**: Updates only the downstream classification head with noisy speech, keeping the fine-tuned speech representation model parameters frozen.
- **Fine-tuning Entire Model (FT-M)**: Updates both the speech representation model and the downstream SER head using noisy speech.

We denote our proposed method as *fine-tuning with multi-task learning* (FT-MTL): Our proposed method trains the SE and SER models simultaneously using CUT.

D. SE baselines

In this paper, we compare our proposed method against two baselines:

- **Noisy**: Speech contaminated by overlaying speech from the CRSS-4ENGLISH-14 corpus or adding ambient noise from the Freesound repository.
- **Fine-tuned**: Speech enhanced by a model fine-tuned exclusively for the SE task using the MSP-Podcast corpus training set. The starting model is the pre-trained weight obtained using the VCTK-DEMAND dataset.

V. Results

A. Emotion Recognition

We evaluate the SER performance of each training strategy under three different SNR conditions, considering both seen and unseen noise.

We select models trained with four different random seeds that showed the best performance on the development set. The test set is divided into five groups for each condition, resulting in 20 values used for statistical analysis (4 runs $\times$ 5 test sets). We perform a one-tailed Welch's t-test between the original and other models to determine if the training strategy significantly improves the performance of the original SER model in noisy conditions, asserting significance at p-value $\leq$ 0.05.

Table I presents the SER performance across noisy conditions. The model fails to improve when only the task head is fine-tuned to adapt to noise. However, fine-tuning the entire model, including the SSL speech representation, leads to significant gains. Adding the SE module before the SER model further enhances performance. Nevertheless, our proposed FT-MTL method performs better in unseen noise conditions, particularly in high-noise scenarios, indicating that SE concatenation may cause over-fitting to seen noise and

TABLE I

AVERAGE SCORES OF OUR PROPOSED METHOD (FT-MTL) AND BASELINES IN SER. WE DENOTE WITH $\star$, $*$, †, AND ‡ WHEN A MODEL PERFORMS SIGNIFICANTLY BETTER THAN THE ORIGINAL, FT-H, FT-M, AND SE-P MODELS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

SNR	Model	Precision	Recall	F1-Macro	F1-Micro
CRSS-4ENGLISH-14 corpus (Seen Noise)					
0dB	Original	0.474	0.414	0.416	0.553
	FT-H	0.391	0.431*	0.352	0.398
	FT-M	0.519**	0.531**	0.520**	0.597**
	SE-P	**0.552****†	**0.581****†	**0.557****†	**0.624****†
	FT-MTL	0.529**†	0.530**	0.525**†	0.607**†
5dB (Matched)	Original	0.536	0.502	0.508	0.612
	FT-H	0.430	0.495	0.413	0.457
	FT-M	0.550**	0.577**	0.557**	0.623**
	SE-P	**0.566****†	**0.598****†	**0.572****†	**0.636****
	FT-MTL	0.560**†	0.576**	0.562**†	0.635**†
10dB	Original	0.565	0.553	0.553	0.637
	FT-H	0.446	0.520	0.438	0.480
	FT-M	0.562*	0.593**	0.569**	0.633*
	SE-P	**0.569***†	**0.602****†	**0.576****	0.638*
	FT-MTL	**0.569***†	0.588**	0.573**	**0.644***
Freesound repository (Unseen Noise)					
0dB	Original	0.536	0.512	0.508	0.598
	FT-H	0.427	0.500	0.410	0.444
	FT-M	0.539*	**0.565****	0.543**	0.610*
	SE-P	0.545*	0.547**	0.534**	0.607*
	FT-MTL	**0.552****†	0.551**	**0.545****‡	**0.625****†‡
5dB	Original	0.563	0.565	0.554	0.629
	FT-H	0.445	0.521	0.432	0.469
	FT-M	0.557*	**0.591****	0.564**	0.627*
	SE-P	0.565*†	0.584**	0.564**	0.629*
	FT-MTL	**0.566***†	0.578**	**0.565****	**0.640****†
10dB	Original	**0.571**	0.586	0.570	0.639
	FT-H	0.452	0.529	0.443	0.480
	FT-M	0.565*	**0.600****	0.572*	0.634*
	SE-P	**0.571***	0.597*	**0.574***	0.636*
	FT-MTL	**0.571***	0.588*	0.572*	**0.644***†

reduce robustness to unseen conditions. At 0 dB, FT-MTL outperforms the original model by 3.7% in F1-Macro and 2.7% in F1-Micro. FT-MTL achieves improvements of 1.1% and 1.8% in F1-Macro and F1-Micro over the SE-P model. These results are statistically significant, particularly in the unseen noise condition (Freesound repository dataset).

B. Speech Enhancement

This section evaluates the SE performance across two noise datasets at three different SNR levels. As shown in Table II, fine-tuning with the SER task slightly drops SE performance. However, this performance gap narrows under unseen noise conditions, demonstrating that our proposed method maintains SE effectiveness without sacrificing robustness in unseen environments. While our primary goal is to improve SER performance, it is encouraging that the SE metrics of the resulting speech are similar to the original SE method.

C. Ablation Study

In the ablation study, we analyze the effectiveness of the adapted CUT strategy. As shown in Table III and Table IV, CUT introduces a trade-off between SER and SE performance. Paired t-tests reveal no significant difference in SER performance with or without CUT, confirming its negligible

TABLE II

SE PERFORMANCE OF OUR PROPOSED METHOD (FT-MTL) AND BASELINES. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

SNR	Model	PESQ	CSIG	CBAK	COVL	SSNR	STOI
CRSS-4ENGLISH-14 corpus (Seen Noise)							
0dB	Noisy	1.10	2.16	1.68	1.57	-2.35	0.62
	Fine-tuned	**1.76**	**3.51**	**2.72**	**2.66**	**5.24**	**0.83**
	FT-MTL	1.59	3.34	2.55	2.48	4.35	0.81
5dB (Matched)	Noisy	1.20	2.61	2.06	1.88	1.20	0.74
	Fine-tuned	**2.31**	**4.04**	**3.19**	**3.21**	**8.03**	**0.89**
	FT-MTL	2.16	3.92	3.10	3.07	7.74	0.88
10dB	Noisy	1.42	3.09	2.50	2.26	5.15	0.83
	Fine-tuned	**2.81**	**4.46**	**3.63**	**3.68**	**10.82**	**0.93**
	FT-MTL	2.68	4.36	3.55	3.56	10.67	0.92
Freesound repository (Unseen Noise)							
0dB	Noisy	1.15	2.50	1.79	1.79	-2.28	0.73
	Fine-tuned	**1.26**	**2.69**	**1.99**	**1.95**	**-0.51**	0.75
	FT-MTL	1.23	2.68	1.96	1.94	-0.80	**0.76**
5dB	Noisy	1.29	2.91	2.16	2.09	1.25	0.82
	Fine-tuned	**1.66**	**3.38**	**2.59**	**2.53**	**4.39**	**0.86**
	FT-MTL	1.59	3.32	2.51	2.46	3.80	**0.86**
10dB	Noisy	1.57	3.36	2.60	2.48	5.17	0.88
	Fine-tuned	**2.28**	**4.03**	**3.23**	**3.19**	**8.99**	**0.92**
	FT-MTL	2.20	3.97	3.16	3.12	8.48	**0.92**

TABLE III

COMPARISON OF SER PERFORMANCE WITH AND WITHOUT THE CUT METHOD ACROSS DIFFERENT NOISE CONDITIONS.

SNR	Method	Precision	Recall	F1-Macro	F1-Micro
0dB	FT-MTL w/o CUT	**0.530**	**0.530**	**0.528**	**0.610**
	FT-MTL	0.529	**0.530**	0.525	0.607
5dB	FT-MTL w/o CUT	**0.560**	**0.579**	**0.567**	**0.636**
	FT-MTL	**0.560**	0.576	0.562	0.635
10dB	FT-MTL w/o CUT	**0.571**	**0.595**	**0.580**	**0.647**
	FT-MTL	0.569	0.588	0.573	0.644

TABLE IV

COMPARISON OF SE PERFORMANCE WITH AND WITHOUT THE CUT ACROSS DIFFERENT NOISE CONDITIONS.

SNR	Method	PESQ	CSIG	CBAK	COVL	SSNR	STOI
0dB	FT-MTL w/o CUT	1.39	3.05	2.31	2.21	2.79	0.77
	FT-MTL	**1.59**	**3.34**	**2.55**	**2.48**	**4.35**	**0.81**
5dB	FT-MTL w/o CUT	1.91	3.67	2.91	2.82	6.95	0.87
	FT-MTL	**2.16**	**3.92**	**3.10**	**3.07**	**7.74**	**0.88**
10dB	FT-MTL w/o CUT	2.40	4.12	3.37	3.30	9.99	0.91
	FT-MTL	**2.68**	**4.36**	**3.55**	**3.56**	**10.67**	**0.92**

impact on this task. Conversely, CUT yields substantial gains in SE, improving SSNR by 55.9% under 0 dB conditions and enhancing speech quality metrics across all SNR levels. These results demonstrate that CUT effectively balances the demands of both tasks without severely compromising either.

VI. CONCLUSIONS

This paper introduced a unified self-supervised speech representation framework to enhance SER robustness in noisy environments. By integrating SE and SER with shared SSL representations, the model preserves competitive performance across SNR levels. The results show strong SER performance without sacrificing SE quality. Notably, the framework excels in unseen noise, outperforming the best baseline. This result demonstrates the potential of jointly training SE and SER to improve generalization and noise robustness, which is crucial for real-world applications where diverse noise scenarios are inevitable. Furthermore, this study highlights that SSL speech representations can effectively handle both SE and SER tasks simultaneously.

REFERENCES

[1] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, (Calgary, AB, Canada), pp. 5084–5088, IEEE, April 2018.

[2] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. 14, pp. 1215–1227, April-June 2023.

[3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[4] 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, pp. 471–483, October-December 2019.

[5] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[6] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-Conversation Corpus," in *Interspeech 2020*, (Shanghai, China), pp. 1823–1827, October 2020.

[7] S. G. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. N. Salman, C. Busso, and C.-C. Lee, "An Intelligent Infrastructure Toward Large Scale Naturalistic Affective Speech Corpora Collection," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, 2023.

[8] 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* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 12449–12460, Curran Associates, Inc., 2020.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[10] 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, 2022.

[11] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, pp. 3400–3404, 2021.

[12] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech Emotion Recognition Using Self-Supervised Features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6922–6926, 2022.

[13] L. Goncalves, A. N. Salman, A. R. Naini, L. M. Velazquez, T. Thebaud, L. P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024-Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results," *Development*, vol. 10, no. 9,290, pp. 4–54, 2024.

[14] 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, 2021.

[15] Y.-T. Wu and C.-C. Lee, "MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion Recognizer," in *Proc. INTERSPEECH 2023*, pp. 3587–3591, 2023.

[16] 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.

[17] S.-G. Leem, D. Fulford, J. Onnela, D. Gard, and C. Busso, "Keep, delete, or substitute: Frame selection strategy for noise-robust speech emotion recognition," in *Interspeech 2024*, (Kos Island, Greece), September 2024.

[18] 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 *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[19] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement," in *Proc. Interspeech 2019*, pp. 1691–1695, 2019.

[20] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in *Interspeech 2005*, pp. 1841–1844, 2005.

[21] Z. Huang, S. Watanabe, S.-w. Yang, P. García, and S. Khudanpur, "Investigating Self-Supervised Learning for Speech Enhancement and Separation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6837–6841, 2022.

[22] K.-H. Hung, S. wei Fu, H.-H. Tseng, H.-T. Chiang, Y. Tsao, and C.-W. Lin, "Boosting Self-Supervised Embeddings for Speech Enhancement," in *Proc. Interspeech 2022*, pp. 186–190, 2022.

[23] H. Song, S. Chen, Z. Chen, Y. Wu, T. Yoshioka, M. Tang, J. W. Shin, and S. Liu, "Exploring WavLM on Speech Enhancement," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 451–457, 2023.

[24] 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 *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 854–860, 2018.

[25] 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 *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6447–6451, 2022.

[26] Y.-W. Chen, J. Hirschberg, and Y. Tsao, "Noise robust speech emotion recognition with signal-to-noise ratio adapting speech enhancement," *arXiv preprint arXiv:2309.01164*, 2023.

[27] A. Reddy Naini, M. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2024)*, (Seoul, Republic of Korea), pp. 12031–12035, April 2024.

[28] S.-W. Fu, C.-F. Liao, T.-A. Hsieh, K.-H. Hung, S.-S. Wang, C. Yu, H.-C. Kuo, R. E. Zezario, Y.-J. Li, S.-Y. Chuang, Y.-J. Lu, Y.-C. Lin, and Y. Tsao, "Boosting Objective Scores of a Speech Enhancement Model by MetricGAN Post-processing," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 455–459, 2020.

[29] F. Tao and C. Busso, "Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1290–1302, 2018.

[30] 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 *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.*, International Society for Music Information Retrieval (ISMIR), 2017.

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

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[33] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pp. 146–152, 2016.