Keep, Delete, or Substitute: Frame Selection Strategy for Noise-Robust Speech Emotion Recognition

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

[1]The University of Texas at Dallas, USA [2]Boston University, USA
[3]Harvard University, USA [4]San Francisco State University, USA

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

Abstract

Speech emotion recognition (SER) system can exploit an *Speech enhancement* (SE) model to increase its noise robustness by suppressing the background noise. However, SE could also suppress emotionally discriminative features, affecting the emotion prediction. We propose an alternative framework, *Keep or Delete* (KoD), to keep the information of the original speech while minimizing the influence of background noise. We train a frame reliability predictor that determines clean frames to keep, discarding the noisy frames. We expand this framework by replacing the dropped frames with those extracted from the enhanced speech to keep the lexical information. We refer to this implementation as *Keep or Substitute* (KoS). Our experiment shows that the KoD model improves the SER results under noisy conditions without fine-tuning the whole model. Also, the KoS framework performs better than enhancing all the frames, indicating the importance of avoiding speech distortion.

Index Terms: speech emotion recognition, noise-robustness, frame selection

1. Introduction

Deploying a *speech-emotion recognition* (SER) system into real-world applications has benefits for multiple domains, such as digital assistants, health care applications [1], and security and defense. The current SER systems built upon pre-trained speech representation models [2–4] have shown to work well in ideal conditions [5]. However, sustaining laboratory performance in a real-world environment is still challenging due to different sources of variability, including the non-stationary background noises in the environment.

For speech-based applications, an attractive approach is the enhancement of the speech signal to provide a more refined input for the system, mitigating the mismatch of a noisy environment. Previous studies have shown that enhancing the acoustic features can improve the SER performance in noisy conditions [6–8]. However, this strategy requires re-training the feature enhancement model whenever we want to change the SER backend. Another appealing approach is to build a cascading system where the pre-trained *speech enhancement* (SE) model enhances the given noisy speech before feeding it into the pre-trained SER model. We can avoid updating the pre-trained SER models by leveraging the off-the-shelf SE model exposed with a large number of speech and noise sound samples [9–11]. However, the SE model is generally designed to improve speech intelligibility, which may affect the emotional discriminative information conveyed on the features [8]. In addition, some acoustic frames might not need enhancement due to the non-stationary nature of real-world background noises.

We investigate why increasing speech intelligibility does not necessarily improve SER performance. Our main assumption is that the speech modifications induced by the pre-trained SE model could impact the SER performance. Exploring this idea, we propose the *Keep or Delete* (KoD) framework, which selectively keeps the original speech's information rather than enhancing all the signals. Our idea is built upon the pre-trained SER model based on the WavLM-large architecture, which has shown strong SER performance in the *speech processing universal performance benchmark* (SUPERB) leaderboard [12]. Without fine-tuning any pre-trained module, we only attach a 1D *convolutional neural network* (CNN) between the convolutional feature encoder and the transformer encoder, which we refer to as the reliable frame selector. The reliable frame selector predicts the reliability score of each acoustic frame extracted from the convolutional feature encoder. To achieve this goal, we mix the clean and noisy acoustic frames and then train the reliable frame selector to determine which frames are from the clean speech. The reliable frame selector can determine which input acoustic frames are noisier than others. The noisy frames are considered unreliable acoustic frames and discarded during the inference. By only maintaining reliable acoustic frames, we can prevent the noisy frames from disrupting the SER prediction while maintaining the original clean frames that do not need enhancement.

We further expand the KoD framework by adding the *Keep or Substitute* (KoS) module. KoS mitigates the influence of acoustic frame gaps induced by the KoD framework by replacing unreliable frames with the ones extracted from the enhanced speech. This framework can avoid destroying the lexical information while keeping the original information of the given speech. Our experiments demonstrate that the proposed frame selection framework can improve SER performance for arousal and dominance without using an additional SE module or adapting the transformer encoder. Furthermore, selectively applying the enhanced frames shows better performance than enhancing all the frames. In a -5dB condition, our best model improves the original model by 17.5%, 16.6%, and 10.6% for arousal, dominance, and valence, respectively.

2. Related Work

Researchers have explored how to improve the noise robustness of an SER system that works well in clean conditions. One of the approaches is data augmentation, where the SER model is trained with an augmented training set. For example, Pappagari et al. [13] proposed the CopyPaste method, which augments the training set by concatenating neutral and emotional speech signals or two speech signals from the same emotional category. This augmentation method improves SER performance in noisy conditions even without contaminating the speech with arbitrary noise sounds. Wu et al. [14] proposed the MetricAug strategy, which adaptively samples an augmented noisy speech for

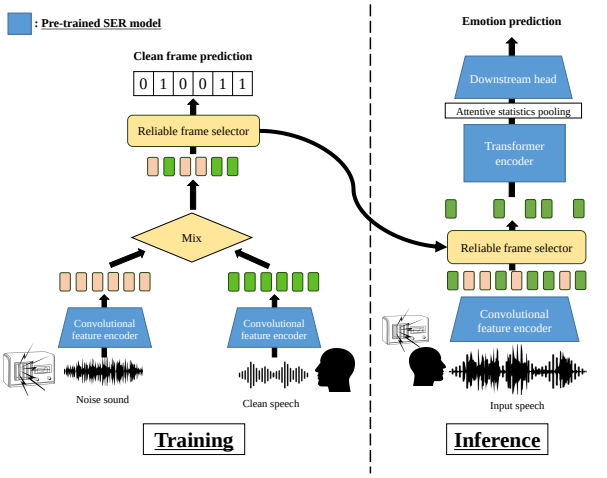

Figure 1: *Training and inference procedure of KoD framework.*

each distortion level based on the validation set's performance improvement. Leem et al. [15] proposed contrastive teacher-student learning for adapting the pre-trained SER model with noisy speech, avoiding catastrophic forgetting of the pre-trained knowledge. Despite these improvements, they require updating the whole pre-trained parameters, which is computationally expensive and requires extra memory for the transformer-based speech representation models [2–4].

Another type of solution is front-end processing by improving the input speech or acoustic feature before feeding it to the pre-trained SER model. The rationale behind this approach is that making the input noisy feature closer to the cleaner one will minimize the environmental difference between the clean training set and the noisy test set. Triantafyllopoulos et al. [7] improved SER performance in noisy conditions using a CNN-based feature enhancement module. Instead of enhancing the acoustic feature, Leem et al. [16] showed that only selecting the noise-robust acoustic feature can yield performance improvement in noisy conditions.

For transformer-based speech representation models, a straightforward approach is to enhance the raw waveform to increase the noise robustness as most of these models accept the raw waveform instead of a handcrafted acoustic feature set. However, in the 4th *deep noise suppression* (DNS) challenge, the authors reported that the submitted SE models could improve the overall audio quality metric but failed to improve the original speech's quality, indicating the suppression of the original speech. This suppression might impact the SER performance. For example, Leem et al. [8] reported that enhancing the speech signal before extracting acoustic features shows worse SER performance than directly enhancing the acoustic features, even though it can improve the speech intelligibility metrics. These observations lead to our proposed frameworks that aim to avoid signal distortion induced by the SE module.

3. Proposed method

We propose the *Keep or Delete* (KoD) framework (Fig. 1) to reduce the influence of noisy frames while keeping reliable frames that do not need enhancement. Figure 1 shows our proposed KoD framework. We mainly focus on improving noise-robustness for a pre-trained transformer-based SER model, which has shown good performance in SER tasks [5, 12, 17]. This type of speech representation model consists of a convolutional feature encoder, which transforms a raw waveform into acoustic frames, and a transformer encoder, which extracts the context information from them. We first fine-tune the pre-trained speech representation model for the SER task under a clean condition. This paper mainly focuses on predicting emotional attributes, including arousal (calm versus active), dominance (weak versus strong), and valence (negative versus positive). Therefore, we fine-tune the model to maximize the *concordance correlation coefficient* (CCC) between the predicted and the ground-truth emotional attribute scores. We freeze the convolutional feature encoder during fine-tuning as it performs better than updating the whole parameters [18].

Our proposed strategy does not update any parameters to adapt the SER model to noisy conditions after fine-tuning it with clean speech. Instead, it discards the noisy acoustic frames that degrade the SER prediction by disrupting the transformer encoder. We achieve this goal by training a reliable frame selector that determines whether each acoustic frame is extracted from the clean or noisy speech. We use 1D-CNN hidden layers and a sigmoid activation function for the output layer to build a reliable frame selector. This module predicts the reliability score of each frame. During training, we mix clean and noisy acoustic frames extracted from the convolutional feature encoder to train the reliable frame selector. The portion of mixing noisy frames to the clean frames is randomly selected between 20% and 80%. This range gives us the best result in our development set. We also tested by fixing the mixture ratio to {20%, 40%, 60%, 80%}, but none of these options significantly improved our initial setting. With these mixed frames, we train the reliable frame selector to predict if the frame is from clean or noisy speech. We minimize the binary cross-entropy loss, $\mathcal{L}_{bce}$, as illustrated in Equation 1.

$$\mathcal{L}_{bce} = \frac{1}{N}\sum_{i=0}^{N}\left[\frac{1}{T_i}\sum_{t=0}^{T_i}\{y_t \log p_t + (1-y_t)\log(1-p_t)\}\right] \quad (1)$$

where N denotes the mini-batch size, T_i denotes the number of frames for the i-th sample, y_t denotes the ground truth (0 if the frame is from noisy speech, 1 otherwise), and p_t denotes the prediction of the t-th frame.

After training, we attach the reliable frame selector between the convolutional feature encoder and the transformer encoder. The reliable frame selector predicts the reliability score of each frame extracted from the convolutional feature encoder. We regard the frame as reliable if the reliability score is above the threshold. We discard the unreliable frames and only keep the reliable ones, feeding them to the transformer encoder. We set the threshold as 0.5, as it gives the best performance for the development set among $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Without adapting the whole SER model to the noisy speech, this approach can compensate for the environmental difference between a clean training set and a noisy test set by discarding noisy frames. Furthermore, this approach avoids the distortion of original speech information that does not need an enhancement, which avoids affecting the SER discrimination of the features. To avoid removing too many frames, which could destroy the input speech, we ensure that at least 50 frames remain after the selection. If the number of reliable frames is less than 50, we select 50 from the top reliability score even if they are not above the threshold.

The possible risk of removing the acoustic frames is the disruption of lexical information of the given speech. Therefore, we test an alternative method that we refer to as *Keep or Substitute* (KoS). KoS keeps all the frames while maintaining the strength of our proposed framework. As in the KoD framework, we define the reliable and unreliable frames by using the

Table 1: *Overall audio, speech, and noise suppression quality for each enhancement model and SNR condition*

Metric	SE Model	**10db**	**5db**	**0db**	**-5db**	**Mean**
OVRL	Noisy	2.31	2.00	1.66	1.40	1.84
	MetricGAN+	2.43	2.26	2.04	1.77	2.12
	FRCRN	3.10	3.07	2.99	2.83	3.00
SIG	Noisy	3.13	2.77	2.24	1.76	2.48
	MetricGAN+	2.86	2.70	2.46	2.14	2.54
	FRCRN	3.30	3.11	2.82	2.42	2.91
BAK	Noisy	2.53	2.14	1.76	1.48	1.98
	MetricGAN+	3.44	3.40	3.33	3.21	3.34
	FRCRN	3.95	3.95	3.94	3.89	3.93

trained reliable frame selector. However, we replace the unreliable frames with the corresponding enhanced frames instead of dropping the unreliable ones. We extract the enhanced frame by directly enhancing the noisy speech signal and then extracting the acoustic frames from it with the convolutional feature encoder. This formulation differs from the cascading system where the SER model uses all the acoustic frames extracted from the enhanced speech. Our KoS framework selectively applies the enhanced features, which can keep the reliable information of the original speech and the enhanced information for the unreliable segments. Compared to the KoD framework, KoS can fill in missing information on acoustic features, keeping the lexical information of the original speech.

4. Experimental Settings

4.1. Datasets

Our experiment uses the MSP-Podcast corpus, which consists of natural and diverse emotional speech samples from various podcast recordings [19]. The audios do not include background music or overlapped speech, and their predicted SNR is above 20 dB. At least five raters annotate each sample for arousal, valence and dominance using a seven-point Likert scale. We average the scores provided by raters for each sample to establish its ground truth values. This paper uses version 1.11 of the corpus (151,654 speaking turns; 237 hours and 56 mins). We use the train set to fine-tune the pre-trained speech representation model, using it as the original SER model. We use samples from the development set to select the best model during the fine-tuning process.

During training the reliable frame selector, we use publicly available noise sounds from the DNS challenge-2 dataset [20]. This dataset contains the additive noise sounds collected from the AudioSet corpus [21], the Freesound repository [22], and the DEMAND database [23]. We remove the noise sounds collected from the Freesound repository for training since we collected the noise sounds for contaminating the test set from the same repository. Each sample is contaminated with different SNR level randomly chosen from the range from 10dB to 0dB. Therefore, we train the model with multiple SNR conditions.

We simulate real-world noisy environments for the testing conditions by collecting diverse ambient noise sounds from the Freesound repository [22]. We use the queries related to the indoor, outdoor, and in-vehicle conditions. We filter the collected samples by using a voice activity detector to ensure that the speech activity does not exist in the noise sounds. We use PyAnnote [24] to detect the speech segment, dropping the samples if the speech activity continues for more than one second. After the collection, we randomly picked the noise sounds to contaminate the Test1 set of the clean MSP-Podcast corpus. We repeat this process five times, creating five different sets for four different SNR levels: 10dB, 5dB, 0dB, and -5dB.

4.2. Speech Emotion Recognition model

We implement our proposed approach with the Wavlm-large model [4], which showed the best emotion recognition performance in the SUPERB benchmark [12] (observed on 02/28/2024). We fine-tune the transformer encoder of the pre-trained speech representation model and the downstream head with the MSP-Podcast corpus. We import the pre-trained models from the HuggingFace library [25]. We use two fully connected layers for the downstream head, where each layer has 512 nodes, layer normalization, and the *rectified linear unit* (ReLU) as the activation function. We use dropout in all the hidden layers to increase regularization, with a rate set to $p = 0.5$. We use a linear output layer with three nodes to predict emotional attribute scores, where each node predicts the scores for arousal, dominance, and valence. We apply attentive statistics pooling [26] on top of the last transformer layer's representation to feed it to the downstream head. This pooling method gives us a better SER performance than the global average pooling under a clean condition.

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 into the range from 0 to 1. We use 32 samples per mini-batch and update the model for ten epochs. We use the AdamW optimizer [27] with a learning rate of 0.00001.

4.3. Implementation of Proposed KoD and KoS Methods

The key component of the KoD framework is the reliable frame selector. The reliable frame selector includes three 1D-CNN blocks and the linear output layer with a sigmoid activation function. Each 1D-CNN block consists of a 1D convolutional layer with a kernel size of three, an instance normalization layer [28], and the ReLU activation function. We put dropout between the CNN blocks to increase regularization, with a rate set to $p = 0.5$.

For training, we contaminate the clean speech from the training set of the MSP-Podcast corpus with the noise sound collected from the DNS challenge-2 dataset. We bring the convolutional feature encoder from the fine-tuned SER model and then extract acoustic feature frames from the clean and the corresponding noisy speech. The resulting feature has 512 dimensions for each frame. We mix the clean and noisy acoustic feature frames with a randomly selected ratio, as described in Section 3. We run 20 epochs to train the reliable frame selector. The other hyperparameters are the same as the ones used for fine-tuning the Wavlm-large model with clean speech.

For the KoS approach, we use the same reliable frame selector as used in the KoD framework. We tested two different publicly available SE models: MetricGAN+ [11] and FRCRN [9]. The MetricGAN+ framework is trained with the VoiceBank-DEMAND dataset, which is used in the study of Valentini et al. [29]. The FRCRN model is trained with the 4th DNS challenge dataset, achieving one of the top performances in this challenge [30]. We first enhance the noisy speech with each SE model and then extract the acoustic frames with the SER model's convolutional feature encoder. We replace the frames dropped by the reliable frame selector with the corresponding frames extracted from the enhanced speech. We denote the KoS framework with MetricGAN+ as *KoS-MetricGAN+*, and the one with FRCRN as *KoS-FRCRN*.

We check the enhanced speech quality to ensure that the

Table 2: *Average CCC of arousal (Aro.), dominance (Dom.), and valence (Val.) for different SNR and models. We highlight the best model in bold for each attribute and SNR level. We mark $*$ if it significantly improves the performance of Original.*

	10db			**5db**			**0db**			**-5db**		
	Aro.	**Dom.**	**Val.**	**Aro.**	**Dom.**	**Val.**	**Aro.**	**Dom.**	**Val.**	**Aro.**	**Dom.**	**Val.**
Original	0.59	0.51	0.62	0.56	0.46	0.60	0.49	0.40	0.56	0.40	0.30	0.47
CS-MetricGAN+	0.49	0.39	0.52	0.42	0.32	0.47	0.33	0.23	0.40	0.24	0.17	0.30
CS-FRCRN	0.58	0.50	0.59	0.56	0.47	0.58	0.51	0.41	0.56	0.42*	0.31	0.51*
KoD	**0.63***	**0.55***	0.62	**0.60***	**0.51***	0.60	**0.55***	**0.43***	0.55	0.45*	0.33*	0.45
KoS-MetricGAN+	0.60	0.51	0.62	0.57	0.46	0.59	0.50	0.39	0.54	0.39	0.28	0.43
KoS-FRCRN	0.61	0.52	**0.63**	0.58*	0.48*	**0.61**	0.54*	0.43*	**0.58***	**0.47***	**0.35***	**0.52***

KoS framework uses the properly enhanced frames. We use DNSMOS P.835 [31], including the scores for the *overall audio quality* (OVRL), *speech quality* (SIG), and the *background noise quality* (BAK). All the metrics range from 0 to 5, where 5 indicates the maximum quality. Table 1 illustrates the P.835 scores for noisy and enhanced speech. Both SE models show an improvement in OVRL and BAK scores compared to the original noisy speech condition, validating the use of SE models to successfully improve the overall audio quality and suppress the background noise sounds. However, we can see that the MetricGAN+ model shows a decrease in speech quality metrics in the 10dB and 5dB conditions. This observation matches the results reported in the 4th DNS challenge [30], where all the SE models exhibited speech distortion. We will discuss the impacts of this distortion on the SER performance in Section 5.

4.4. Baselines

We compare our approach with the SER model without using the SE module, which we refer to as *Original*. We also use a *cascading system* (CS) that first enhances the speech and then predicts the emotional attributes. We use the same SE models as those used for the KoS framework. Notice that this baseline uses all the enhanced frames. We do not update any parameters of the pre-trained SER model when using the SE module. We denote the cascading systems as *CS-MetricGAN+* and *CS-FRCRN*, following the name of the SE method.

5. Results

We compare the SER performance of our proposed KoD and KoS frameworks with the baselines in noisy conditions. We report the average CCC of five different test sets for each SNR level, as we described in Section 4.1. We conduct an one-tailed Welch's t-test to assess whether the system performs significantly better than the original SER model in noisy conditions. We assert significance at p-value < 0.05.

Table 2 lists the results. Simply enhancing the speech with the pre-trained SE model fails to significantly improve the performance of the original SER model, except for the CS-FRCRN approach in the -5dB condition. For example, the FRCRN model slightly improves the performance in the 0dB and the -5dB conditions but fails to achieve good results in 10dB and 5dB conditions. While both SE models successfully suppress the background noise and improve the audio quality, as shown in Table 1, they still degrade or cannot significantly improve the performance when the SNR level is high. These results imply that improving the audio quality does not guarantee improving the SER performance. The SE speech manipulations can adversarially affect the discriminative information relevant to SER.

In contrast to the baselines, the KoD framework significantly improves arousal and dominance prediction performance under all SNR levels. In the 10dB condition, it improves the original SER model's performance by 6.7% for arousal and 7.8% for dominance without adapting any parameters to the noisy conditions. In the -5dB condition, it provides 7.1% (arousal) and 6.4% (dominance) improvements over the CS-FRCRN baseline. The KoD strategy performs better than using all the enhanced frames by simply dropping the noisy frames. This result demonstrates that avoiding the distortion of the original speech is important for arousal and dominance predictions. Despite these improvements, the KoD framework fails to improve the valence performance of the original SER model. The results are worse than enhancing the speech with the FRCRN model in the 0dB and -5dB conditions. The KoD framework drops unreliable frames, which could destroy the linguistic information of the given speech. According to the study of Wagner et al. [5], valence performance correlates to the linguistic information implicitly included in the transformer-based *self-supervised learning* (SSL) speech representation models such as WavLM, which were primarily built for *automatic speech recognition* (ASR). It is expected that dropping frames affects the linguistic information in WavLM, decreasing the valence performance.

Compared to the KoD framework, the KoS-FRCRN shows better performance for valence prediction. It yields 5.4% and 15.5% improvement for valence in the 0dB and -5dB conditions, respectively. In addition, it successfully improves the original SER model's performance for all the attributes, except for the 10dB condition. Compared to the *Original* model, it improves the performance by 17.5% (arousal), 16.6% (dominance), and 10.6% (valence) under -5dB conditions. KoS-MetricGAN+ does not perform better than the KoD framework but performs much better than enhancing all the frames with the MetricGAN+ method. All these results indicate that avoiding the distortion of original speech information is important for emotion prediction. Also, it is important not to drop frames to preserve the lexical information for valence prediction.

6. Conclusions

We proposed a KoD framework that selects the reliable frames and discards the unreliable ones to improve the noise robustness of the SER model. This framework can avoid the distortion introduced by the SE module on the speech signal by learning what to keep while minimizing the influence of background noise. Our experiment demonstrated that the KoD framework can improve the performance for arousal and dominance by keeping the original speech information. Our experiment also suggests that the selective application of enhanced frames performs better than enhancing all the frames, implying the importance of keeping lexical information conveyed on SSL speech representations for valence prediction. For real-world applications, we plan to investigate how we can decrease the computational overhead of KoD and KoS frameworks while keeping their strengths under noisy conditions.

7. Acknowledgements

Study supported by NIH under grant 1R01MH122367-01.

8. References

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

[2] 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.

[3] 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.

[4] 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.

[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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[6] C. Huang, , G. Chen, H. Yu, Y. Bao, and L. Zhao, "Speech emotion recognition under white noise," *Archives of Acoustics*, vol. 38, no. 4, pp. 457–463, 2013.

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

[8] 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.

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

[10] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9122–9126.

[11] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 201–205.

[12] S.-W. 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 *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[13] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6324–6328.

[14] 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*, 2023, pp. 3587–3591.

[15] 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.

[16] ——, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 6447–6451.

[17] A. 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)*, vol. To appear, Seoul, Republic of Korea, April 2024.

[18] 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.

[19] 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.

[20] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[21] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[22] 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.

[23] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. AIP Publishing, 2013.

[24] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.

[25] 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.

[26] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.

[28] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[29] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

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

[31] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.