Selective Acoustic Feature Enhancement for Speech Emotion Recognition with Noisy Speech

Seong-Gyun Leem, *Student Member, IEEE,* Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso, *Fellow, IEEE*

***Abstract*—A *speech emotion recognition* (SER) system deployed on a real-world application is highly likely to encounter speech contaminated with unconstrained background noise. To deal with this issue, a *speech enhancement* (SE) module can be attached to the SER system to compensate for the environmental difference of an input. Although the SE module can improve the quality and intelligibility of a given speech, there is a risk of affecting discriminative acoustic features for SER that are resilient to environmental differences. Exploring this idea, we propose to enhance only weak features that degrade the emotion recognition performance, while keeping strong features that are resilient to environmental differences. Our model first identifies weak feature sets by using multiple models trained with one acoustic feature at a time using clean speech. After training the single-feature models, we rank each speech feature by measuring three criteria: performance, robustness, and a joint rank ranking that combines performance and robustness. We group the weak features by cumulatively incrementing the features from the bottom to the top of each rank. Once the weak feature set is defined, we only enhance those weak features, keeping the resilient features unchanged. We implement these ideas with the *low-level descriptors* (LLDs). We show that extracting LLDs from an enhanced speech signal does not improve the performance of weak features. Instead, directly enhancing the LLDs lead to better performance. Our experiment with clean and noisy versions of the MSP-Podcast corpus shows that the selective feature enhancement approach proposed in this study yields a 17.7% (arousal), 21.2% (dominance), and 3.3% (valence) performance gains over a system that enhances all the LLDs for the 10dB *signal-to-noise ratio* (SNR) condition.**

***Index Terms*—Speech emotion recognition, noisy speech, speech enhancement, feature selection.**

I. Introduction

INFERRING human behavior using speech is appealing given the ubiquitousness of speech-based devices in daily life. Important information for *human-computer interaction* (HCI) is the emotion of a person, which plays a key role in her/his decision-making process [1]. Therefore, recognizing emotion from speech has been an active research area [2], with applications in diverse areas such as health informatics, education, entertainment, and surveillance. One challenge for *speech emotion recognition* (SER) systems is the background acoustic noise observed in recordings collected on real-world applications. A corrupted speech signal can greatly disrupt the acoustic features, reducing the prediction performance of the SER systems given the mismatch between train and test conditions.

One way of attenuating the effect of noisy recordings in SER tasks is to denoise the signal or acoustic feature so that the SER system can receive a cleaner input. For example, Huang et al. [3] improved the predictions of arousal and valence on speech contaminated with white Gaussian noise by using spectral subtraction and perceptual masking. Zhang et al. [4] used an autoencoder with a neural network implemented with *long-short term memory* (LSTM) layers to enhance *Mel-frequency cepstral coefficient*, yielding performance improvements in speech recordings contaminated with background noise from the CHiME database [5]. Triantafyllopoulos et al. [6] improved the emotion classification performance by enhancing the log magnitude spectrum of noisy speech with a *convolutional neural network* (CNN) implemented with residual blocks. All these studies have aimed to improve SER performance by making noisy features closer to the clean features, using speech enhancement modules that affect all the features. However, the enhancement models, which are often designed to improve speech quality, may affect the emotional discriminative information conveyed on the features. Also, some features may not need enhancement. We discovered that certain acoustic features demonstrate robustness in SER systems, even under noisy conditions [7]. By focusing solely on these robust features, we achieved improved performance compared to using all the features in noisy recording environments. This raises the question: could an enhancement strategy, primarily designed to enhance speech intelligibility, impact the discriminative power of the enhanced features, particularly for these robust acoustic features?

This study proposes to only enhance features that degrade the performance in noisy recording conditions, without modifying the rest of the features, which are robust against background noises. We rank all features based on three criteria to select the least robust feature set: the absolute recognition performance in the target noisy condition (performance), the relative SER model performance degradation from clean to noisy conditions (robustness), and the summation of the ranks assigned by both previous criteria (joint). With these criteria, the features are cumulatively added from the lowest rank to the highest rank to find the weak feature set that needs to be enhanced by the acoustic feature enhancement module. We evaluate the SER performance by only enhancing the selected feature set, increasing the coverage of enhanced features from 10% to 90%. The model with the best performance on the development set is selected for the final model, determining the features to be enhanced.

Our experiment with the clean and noisy version of the MSP-Podcast corpus shows that the selective feature enhancement strategy increases the performance over an SER system

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

where all the LLDs are enhanced by 17.7% (arousal), 21.2% (dominance), and 3.3% (valence) in the 10dB *signal-to-noise ratio* (SNR) condition. Furthermore, our proposed method also yields better performances than using a signal-based speech enhancement. Compared with an SER system trained with speech signal enhanced with the MetricGAN approach [8], the proposed selective feature enhancement method yields relative performance gains of 54.9% (arousal), 63.2% (dominance), and 68.1% (valence) in the 10dB SNR condition.

The rest of the paper is organized as follows. Section II summarizes the previous approaches for enhancing noisy speech and dealing with noisy speech for speech emotion recognition. Section III describes our database and acoustic features to analyze and evaluate our proposed feature enhancement framework. Section IV explains our motivation and the description of the proposed selective feature enhancement framework. Section V describes our experimental settings and the baselines. Section VI presents our experiments with the proposed selective feature enhancement method for SER tasks. The section further analyzes the feature enhancement methods by comparing them with signal enhancement methods. Lastly, Section VII concludes the paper by summarizing our contributions and future research directions.

II. Related works

A. Speech Enhancement

The main goal of a speech enhancement system is to increase the quality and intelligibility of noisy speech contaminated by a low sampling rate, restricted bandwidth, or background noises. To accomplish this goal, the enhancement systems need to suppress the noise while preserving the information of the original speech.

Early studies formulated the speech enhancement problem using classical strategies. For example, some studies formulated a speech enhancement task as an additive noise estimation problem [9], [10]. Their main objective was to estimate a noise spectrum during non-speech activity, then subtract the estimated additive noise from the noisy speech. There are also studies that formulated a speech enhancement task as a filter estimation problem. The approach finds the optimal filter that can minimize the error between clean speech and denoised speech [11]–[13]. Other studies viewed a speech enhancement task as a matrix decomposition problem to find a subspace for clean speech and another for the noises [14]–[16].

Deep learning solutions have emerged as powerful alternatives for SE. The straightforward formulation for an SE task is to train a neural network to map noisy speech into clean speech. In this paradigm, features extracted from noisy speech are fed into the neural network model. Then, the model is trained to make its output to be similar to the corresponding clean speech. Such model includes solutions based on *deep belief network* (DBN) [17], *recurrent neural networks* (RNNs) [18], *convolutional neural networks* (CNNs) [19], [20], and denoising autoencoders [21], [22]. Furthermore, *conditional generative adversarial network* (cGAN) architecture has been adopted to improve the quality of the enhanced speech. Pascual et al. [23] proposed the *speech enhancement GAN* (SEGAN) architecture. The discriminator is trained to classify if the input speech is real or created by the generator. The generator is trained to transform noisy speech into clean speech with the adversarial goal of deceiving the discriminator. Studies have improved this GAN-based speech enhancement model using different strategies. For example, Phan et al. [24] used multiple generators to improve the quality of the enhanced speech. Li et al. [25] investigated the use of a self-attention mechanism to make the enhancement model exploit long-term characteristics in the input features. Fu et al. [8] proposed MetricGAN, which uses common metric scores for speech quality and intelligibility during the training of the enhancement model as the targets of the adversarial training. These metrics include the *perceptual evaluation of speech quality* (PESQ) [26] and the *short-time objective intelligibility* (STOI) [27].

Unlike the aforementioned studies that only enhance the real part of the speech spectrum, some studies investigated the use of the imaginary part. Tan and Wang [28] used *convolutional recurrent network* (CRN) that enhances the phase spectrum as well as the magnitude spectrum. Hu et al. [29] proposed the *deep complex CRN* (DCCRN), which uses the complex convolutional block to simulate the complex-valued operation in the CRN-based speech enhancement model.

B. Speech Emotion Recognition for Noisy Environments

Improving robustness to noise of SER systems has become an active research area. Several studies have tried to solve this problem by directly improving the robustness of the SER model. One of the proposed solutions is to contaminate the clean speech in the training set to augment the data. Lakomkin et al. [30] contaminated the clean training set by adding arbitrary background noises and simulating reverberation. Tiwari et al. [31] utilized various types of noises from the NOISEX-92 database. They modulated white noise to augment the training data. This strategy makes the SER model encounter the target noise during training, building resilient solutions during inference when similar noises are observed.

Another option is to view the noise-robust SER task as a domain adaptation problem. Leem et al. [32] investigated the use of ladder network [33] to improve the performance of SER in the presence of noise. A ladder network is a strong framework for compensating domain mismatches for SER [34]–[37]. They applied the ladder network to the noisy SER task by decoupling the emotional and reconstruction embeddings to reduce the influence of background noise on emotion predictions. Wilf and Provost [38], [39] used *domain separation networks* (DSNs) [40] for SER in noisy speech. The approach simultaneously trains a shared encoder with all the noisy conditions and multiple expert encoders, each with an individual environmental condition. They also applied an adversarial training, which minimizes the difference among the outputs from the expert encoders to generalize the performance in unseen noisy conditions.

Other approaches to improve SER performance in noisy speech are to either use noise-robust features or discard noisy frames or segments in the speech signal. Georgogiannis and

TABLE I
LLDS OF THE COMPARE 2013 FEATURE SET.

Group	LLD	Nomenclature
Energy	Sum of auditory spectrum	Spec-sum
	Sum of RASTA style-filtered auditory spectrum	RASTA-sum
	Root mean square energy	RMSenergy
	Zero-crossing rate	zcr
F0	Fundamental frequency	F0
	Probability of voicing	voicingProbability
Voice Quality	Jitter(local)	jitterLocal
	Jitter(delta)	jitterDDP
	Shimmer(local)	shimmerLocal
	log harmonic-to-noise ratio	logHNR
Spectral	Spectral flux	SpectFlux
	Spectral entropy	SpectEnt
	Spectral variance	SpectVar
	Spectral skewness	SpectSkew
	Spectral kurtosis	SpectKurt
	Spectral slope	SpectSlope
	Spectral harmonicity	SpectHarm
	Spectral Centroid	SpectCent
	Spectral roll-off 0.25	SpectROff25.0
	Spectral roll-off 0.50	SpectROff50.0
	Spectral roll-off 0.75	SpectROff75.0
	Spectral roll-off 0.90	SpectROff90.0
	Spectral energy 250Hz-650Hz	fband250-650
	Spectral energy 1kHz-4kHz	fband1000-4000
	Psychoacoustic sharpness	psySharpness
Cepstral	MFCC	MFCC[1-14]
RASTA	RASTA-style auditory spectrum bands	RASTA-band[1-26]

Digalakis [41] used the Teager energy-based *Mel-frequency cepstral coefficient* (MFCC) to minimize the impact of background noise on the input feature. This feature set has been shown to be robust to noise in other speech tasks. Schuller et al. [42] showed that feature reduction using the information gain ratio improves the SER performance not only in a clean condition, but also in noisy conditions. Leem et al. [7] proposed a feature selection framework for SER by assessing the noise robustness of each acoustic feature in the noisy condition. Pandharipande et al. [43] used a voice activity detection module to discard noisy frames for SER in noisy conditions.

This study focuses on the feature enhancement method, which does not change the original SER model, but adds a speech enhancement module before feeding the noisy speech into the recognition model. Huang et al. [3] showed that spectral subtraction and perceptual masking-based speech enhancement improve the performance of arousal and valence prediction in noisy conditions. Juszkiewicz [44] applied histogram equalization to MFCC, which increases the emotion classification accuracy in noisy audio. Triantafyllopoulos et al. [6] used a convolutional neural network with residual blocks as a feature enhancement module to improve SER performance. All of those studies equally apply the enhancement strategy to the entire feature set. Unlike these studies, our solution is to enhance only the weak features that disrupt the prediction in the presence of noise and keep the features that are not highly affected by background noises.

III. RESOURCES

A. The MSP-Podcast Corpus

SER models need to be evaluated with a dataset that can simulate realistic scenarios, since the main goal of a noise-robust SER task is to improve the SER performance in real-world environments. For this reason, we used the clean [45] and noisy [32] versions of the MSP-Podcast corpus.

The clean version of the MSP-Podcast corpus contains spontaneous emotional speech samples collected from various recordings available in audio-sharing websites. From these recordings, the protocol chooses samples that are expected to have balanced emotions by using the retrieval approach proposed in Mariooryad et al. [46]. Samples that have background music or noisy speech (the estimated SNR is lower than 20dB) are removed. All samples are formatted at a sampling rate of 16kHz. A modified version of the crowdsourcing protocol proposed in Burmania et al. [47] is used to annotate the emotion for each sample. At least five different evaluators annotated each sample with emotional attributes and primary and secondary emotions. We focus on the emotional attribute scores for arousal (calm versus active), dominance (weak versus strong), and valence (negative versus positive) collected with a seven-point Likert-scale. The study uses release 1.8 of the corpus, which has more than 113 hours of annotated emotional speech samples.

The noisy version of MSP-Podcast corpus, which was introduced in our previous work [32], considers unconstrained noises that are highly likely to appear in real-world applications. We directly recorded all speech samples in the MSP-Podcast corpus with non-stationary noise sounds to simulate real-world recording conditions. Noise sounds are collected from traditional radio shows without copyright. Those sounds contain human voices, background music, and various types of sound effects. We simultaneously played speech and noise sounds with the speakers of two portable devices. Then, we recorded those mixed sounds on a smartphone, mimicking the noisy speech collected from real-world applications. We changed the distance between the speakers and the smartphone to achieve different levels of SNR. We collect a one-minute speech recording before the data collection to estimate the SNR of the recording condition. We move the devices until approximately obtaining the following conditions: 10dB, 5dB, and 0dB. We named these settings by their target SNR. The emotional labels for this version of the corpus are directly transferred from the clean version of the MSP-Podcast corpus.

To train the speech enhancement model, we use 43,361 speech segments from clean and noisy speech. These segments have not been annotated yet, so they do not belong to release 1.8 of the corpus. Those samples are not used for training the SER models. We train and evaluate the SER models by using the partitions of the MSP-Podcast corpus. The training set has 44,879 speaking turns. We always train the SER models with clean speech. We use 7,800 clean samples from the development set to select the best model. To assess the robustness of individual features, we use the noisy recordings in the development set (e.g., 7,800 noisy samples for each target SNR condition). For the evaluation, we use four versions of the 15,326 samples from the test set, including the clean and three noisy recording conditions.

B. Acoustic Features

Our study uses the 65 *low-level descriptors* (LLDs) from the 2013 *Computational Paralinguistics Challenge* (ComParE) features set extracted with the OpenSMILE Toolkit [48]. Table I lists the LLDs. We use the standard setting used in OpenSMILE. A 60ms window is applied for the *zcr* feature, *F0* and voice quality feature group. The other LLDs are estimated with a 25ms window. All the features are sampled with a 10ms step size. This approach creates a frame-level representation for each speech signal.

We apply the Z-normalization to the features to avoid shifts in the feature distributions caused by environmental noise conditions. We regard the development set of each noisy recording condition as speech samples obtained from the target environment. Then, we use their mean and standard deviation to normalize the features from the noisy recording conditions. For clean speech, we normalize the features by using the mean and standard deviation of the clean training set, since we already have a training set in the clean condition. We clip the value of each feature if they exceed ± 10 after the normalization to avoid outlier values affecting the training process.

IV. Proposed Approach

A lesson learned from the study in Leem and Busso [7] was that there exist LLDs that are robust against environmental differences. If the goal of a feature enhancement strategy is to improve speech quality, as is commonly the case, it is not guaranteed that the enhanced features will retain emotional information. Using these insights, we propose to selectively enhance only the features that are most affected by noise. This section presents a preliminary analysis that serves as the motivation for our proposed approach (Sec. IV-A). Then, we present the feature enhancement strategy adopted in this study (Sec. IV-B). Then, we describe the proposed strategy to recognize emotions in noisy speech (Sec. IV-C).

A. Motivation

We conduct a preliminary analysis to illustrate the need for a selective enhancement framework that only processes less resilient features. We conduct a controlled experiment to identify which features are resilient to a target noisy environment. We estimate the performance of SER systems trained with individual LLDs with clean features in the target noisy condition (i.e., only one feature per model). We construct a feature probe model set, consisting of multiple SER models where each model is trained with a different LLD. Since our experiment uses LLDs from the ComParE 2013 feature set (Sec. III-B), our feature probe model set has 65 different models each of them trained with one LLD.

We train separate SER models for arousal, dominance, and valence. The SER model is built following the baseline introduced in the study of Parthasarathy and Busso [34]. This model consists of five blocks of 1D convolution layers and max-pooling layers. Then, we add two fully connected layers, each of them implemented with 256 nodes. The final layer is the output layer. We use the *rectified linear unit* (ReLU) as the activation function for the convolution and fully connected layers, and a linear transformation for the output layer. We increase regularization with dropout with a rate set to p=0.1. The dropout is placed between the input and first convolution layer, and between the last convolution layer and the first fully connected layer. We use the multitask learning approach proposed in Parthasarathy and Busso [49], where the model simultaneously predicts the scores for all three emotional attributes during training. Equation 1 illustrates the cost function of our model.

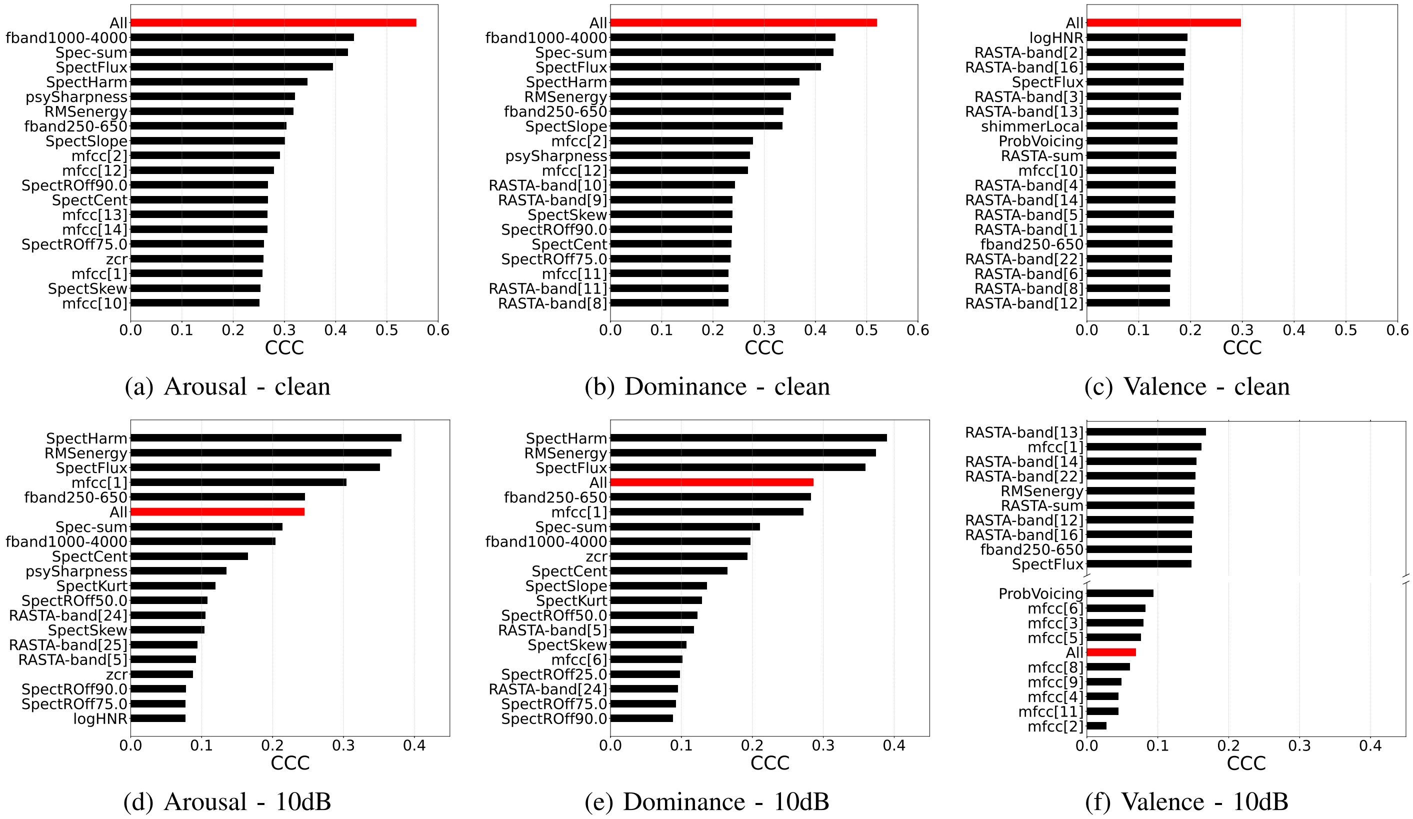
$$\mathcal{L} = \alpha \times \mathcal{L}_{aro} + \beta \times \mathcal{L}_{val} + (1 - \alpha - \beta) \times \mathcal{L}_{dom} \quad (1)$$

where $\mathcal{L}_{aro}$, $\mathcal{L}_{val}$, $\mathcal{L}_{dom}$ denote the loss functions for arousal, valence, and dominance, respectively, and α and β denote the weight of each loss function. We choose $\alpha = 0.7, \beta = 0.3$ for arousal, $\alpha = 0.0, \beta = 0.2$ for dominance, and $\alpha = 0.1, \beta = 0.8$ for valence. These settings showed the best performance reported in Parthasarathy and Busso [34] for both in within-corpus and cross-corpora evaluations. We train the model to maximize the *concordance correlation coefficient* (CCC) by minimizing the term $1 - CCC$ for each loss function. We use the Adam optimizer [50] with a learning rate set to 0.00005 to optimize the parameters. We train models for 25 epochs with a mini-batch of 512 sentences.

To visualize the differences in SER performance among models trained with individual LLDs, we run 10 experiments by changing the initial weights of the emotion recognition models. Figure 1 represents the average CCC values over 10 trials of each feature probe model. The figure also shows the result of an SER model trained with all the LLDs. The models are trained with clean speech but tested with either the clean or the 10dB conditions. The nomenclature of the features is listed in Table I. The figure shows that the model trained with all the LLDs achieves the best performance in the clean condition. For the 10dB condition, however, there are models trained with some LLDs that showed better performance than using all the LLDs (e.g., *SpectHarm* for arousal and valence and *RASTA-band[13]* for valence). This result is particularly clear for valence when only five LLDs out of 65 showed worse performance than the model trained using all the LLDs. The insights from this evaluation indicate that there exist features that degrade the performance when they are combined with other robust features when the models are tested on noisy conditions. Some features, however, are robust enough and show better performance than a model trained with all LLDs. With these ideas, we propose to enhance only weak features and keep the robust features unchanged to increase the SER performance of our model.

B. Feature Enhancement

Before explaining our proposed selective feature enhancement approach, we describe the enhancement method adopted



(a) Arousal - clean

(b) Dominance - clean

(c) Valence - clean

(d) Arousal - 10dB

(e) Dominance - 10dB

(f) Valence - 10dB

Fig. 1. CCC of SER models trained with a single LLD in clean and noisy conditions (10dB). The red bars denote the performance of the SER model trained and tested with all the LLDs. In a noisy condition, training with a single robust feature leads to better performance than training the model with all the LLDs.

in this study. Our enhancement model relies on a *generative adversarial network* (GAN) architecture, which has shown good performance for signal enhancement [8], [23]. In our implementation, the discriminator is trained to determine if the LLDs are extracted from real signals or created from the generator. The generator is trained to transform LLDs extracted from noisy speech into the ones extracted from clean speech. The adversarial loss aims to deceive the discriminator by making its output more realistic. After training the feature enhancement model, the LLDs from the noisy speech are enhanced by the trained generator. Previous studies have shown that adding adversarial loss with a regression loss can improve signal-based speech enhancement models [23], [51]. We follow this approach, training the generator with an adversarial and a simple regression loss. We use the *mean squared error* (MSE) between the enhanced and corresponding clean LLDs for our regression loss.

The generator, G, consists of four layers of 512 bi-directional *gated recurrent unit* (GRU) [52], an output layer with a linear activation function, and a residual connection from the input layer to the output layer. The discriminator, D, consists of three layers of 32 bi-directional GRU, and an output layer implemented with the sigmoid activation function. The generator and the discriminator are implemented with dropout with a rate p=0.2 for all the hidden layers. We adopt the *least square generative adversarial network* (LSGAN) as the cost function to train our feature enhancement model, as illustrated in Equation 2,

$$\begin{aligned}\mathcal{L}_D &= \frac{1}{2}(D(\hat{x})-1)^2 + \frac{1}{2}(D(G(\hat{x}))-0)^2 \\ \mathcal{L}_G &= \frac{1}{2}(D(G(\hat{x}))-1)^2 + (G(\hat{x})-x)^2\end{aligned} \tag{2}$$

where $\mathcal{L}_D$ denotes the loss function for the discriminator, $\mathcal{L}_G$ denotes the loss function for the generator, x denotes the clean feature vector (i.e., LLDs extracted from the clean speech), and $\hat{x}$ denotes the noisy feature vector (i.e., LLDs extracted from the noisy speech). The discriminator is trained to predict 1.0 if the given LLD is from real speech, and 0.0 if the given input is generated by the generator. The generator is trained to deceive the discriminator so that the discriminator predicts 1.0 from the generator's output. This loss function makes the generator create realistic acoustic features. To balance the loss between the generator and the discriminator, we update the generator 10 times more than the discriminator. As mentioned in Section III-A, we use the samples that have not been annotated with emotional labels from the clean and noisy versions of the MSP-Podcast corpus. We use 30,352 sample pairs to train the enhancement model and 13,009 samples to find the best architecture for the feature-based enhancement model.

C. Selective Feature Enhancement Strategy

Figure 2 shows our proposed approach. Instead of enhancing all of the features, we selectively enhance only *weak* features.

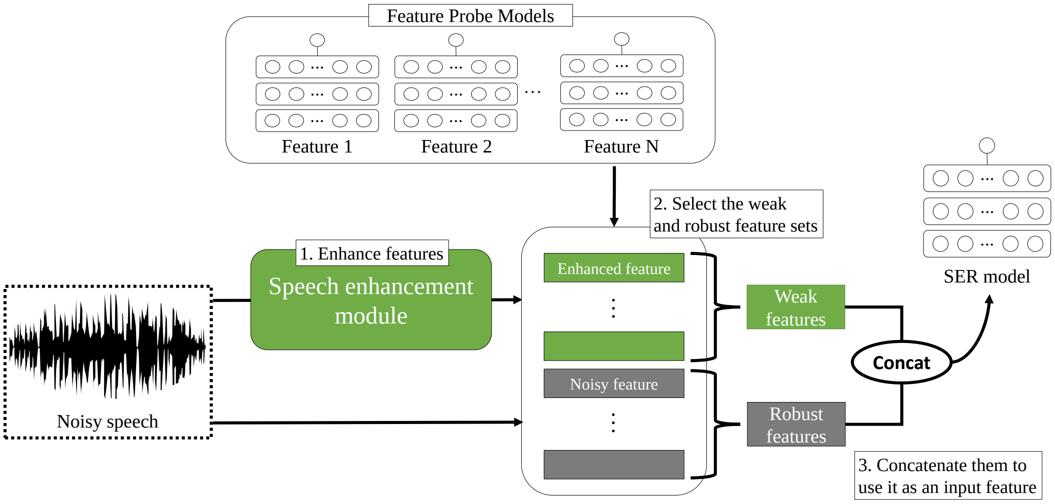


Fig. 2. The diagram of our proposed selective feature enhancement model. With the feature probe models, we select the robust and weak feature sets for the noisy condition. Once they are selected, we enhance only the weak features set, keeping the robust features unchanged. Finally, we concatenate the enhanced and robust features to be used as input for the SER model.

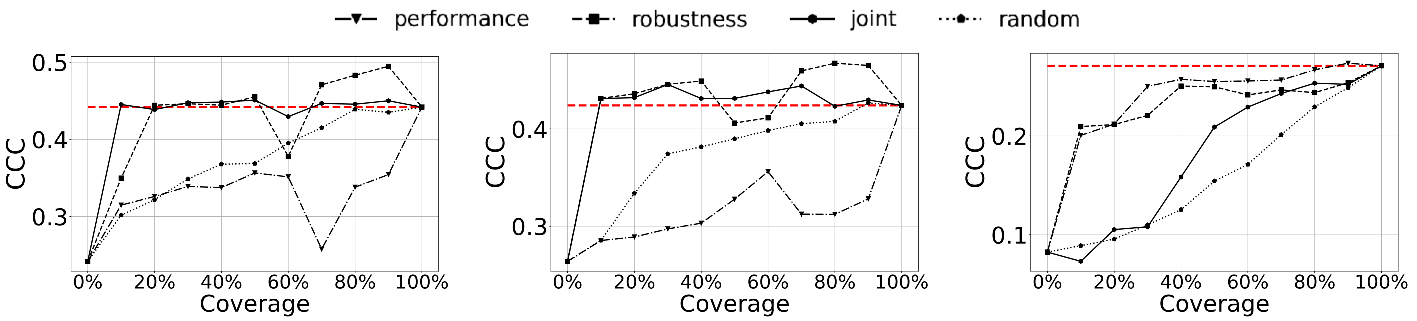
The approach starts by identifying *robust* features. Once the robust feature set is defined, we mask the robust features from the speech enhancement model. We concatenate the unaltered robust features and the enhanced weak features and feed them into the SER model to predict the emotional attribute scores.

A crucial step in our approach involves identifying robust features that remain resilient to SER under noisy conditions. We rely on the classification experiments conducted in Section IV-A, where we train and evaluate SER models with one feature at a time. We test three criteria to rank the LLDs: performance, robustness, and performance+robustness. The performance criterion considers the absolute prediction performance in a noisy environment. We use CCC as the performance metric, ranking the feature that has the highest CCC at the top of the list, and the feature that has the lowest CCC at the bottom of the list. The robustness criterion considers the performance decrease of the SER system from the clean environment to the target noisy environment. In our preliminary experiment, we found that using the relative decrease in performance favored features leading to low CCC in the clean condition. These features are not expected to contain discriminative emotional information. Therefore, we decide to use the absolute performance difference between the clean condition and the target noisy condition. Based on this metric, we rank the feature that has the lowest performance drop at the top of the list, and the feature that has the highest performance drop at the bottom of the list. The performance+robustness criterion, which we refer to as the *joint* criterion, considers the two previous criteria. We combine the two ranks by adding their relative order to the ranked list. Features at the top of the list have good performance in the presence of noise, achieving CCC values that are not too different from the CCC values observed under clean conditions. We create separate lists with the ranking for arousal, dominance, and valence.

Once the LLDs are ranked, we cumulatively include LLDs based on each rank to construct the weak feature subsets from the bottom to the top. Using one of the three criteria, we add LLDs in increments of 10% from the bottom to the top of the list. Since only the low-ranked features in each criterion are enhanced, we can prevent the resilient features from losing their original discriminate emotional information by the enhancement model. The SER model follows the same architecture described in IV-A. The only difference is that the models are trained with 65 LLDs. We only change the input channel size of the first 1D convolution layers from 1 to 65 to accommodate 65 LLDs.

We evaluate the CCC performance on the development sets for the clean and 10dB conditions per emotional attribute. We visualize performance changes using each criterion and each feature coverage by running ten trials with different initialization, reporting the average CCC obtained on the development set in the 10dB condition. Notice that we do not use the test set for this experiment, since selecting the weak and robust features is determined during training. We compare the proposed method with a baseline selection method, where the enhanced features are randomly selected, and with a model trained with a baseline model where all the LLDs are enhanced.

Figure 3 reports the average CCC values as we increase the percentage of enhanced features. For arousal and dominance, we observe important CCC gains using the proposed selective feature enhancement approach. We observe similar performance to the model trained with all the enhanced features even when 10% or 20% of the features are enhanced using the robustness or joint criteria. The best performance is obtained when enhancing 90% of the features for arousal and 80% of the features for dominance using the robustness criterion. For valence, the best performance is obtained when 90% of the features are enhanced, using the performance criterion. This result shows that there exists a robust feature set that does not



(a) Arousal

(b) Dominance

(c) Valence

Fig. 3. CCC in the 10dB condition of SER models trained with different feature coverage. The feature sets are cumulatively created by adding LLDs based on the performance, robustness, joint, and random criteria. The red dashed lines mark the performance of a baseline trained using all the LLDs.

need to be enhanced to improve the prediction. Interestingly, Figure 3 does not show improvements when the features are randomly selected. The improvements are not obtained by enhancing a small number of features. The CCC improvements are observed by enhancing just the weak features.

Based on the results in Figure 3, we use the robustness criterion for arousal and dominance, and the performance criterion for valence. We select the feature group that shows the best performance among all the feature groups. We use 90% (arousal), 80% (dominance), and 90% (valence) feature coverage for testing on the 10dB condition. We replicate the same process for testing on the 5dB and 0dB conditions using the corresponding noisy development set. For the 5dB condition, we use 90% (arousal), 90% (dominance), and 90% (valence) feature coverage. For the 0dB condition, we use 90% (arousal), 80% (dominance), and 90% (valence) feature coverage.

V. Experimental Settings

A. Implementation

In our experiment, each emotion recognition model is trained to predict arousal, dominance, or valence. We use the clean version of the MSP-Podcast corpus to train the SER models. We test them with three different noisy conditions (10dB, 5dB, 0dB) in the noisy version of the MSP-Podcast corpus. We consider matched and mismatched conditions. The matched condition uses the same environmental conditions for the enhancement model (train set), feature selection (development set), and evaluation of SER experiments (test set). The mismatched condition uses one environmental condition for the enhancement model (train set), and feature selection (development set), and another for the evaluation of SER experiments (test set). For mismatched conditions, we use the noisy speech from the 10dB condition for training the feature enhancement model and selecting robust and weak features, and the noisy speech from the 5dB and 0dB conditions for testing the models. We run ten trials to evaluate the significance of our proposed selective feature enhancement approach. For each experiment, the emotion recognition models are initialized with different values. We also variate the enhancement model among trials by saving the model obtained every 50 batches. Then, we select the models achieving the best 10 performances in the development set for the enhancement model. We conduct a two-tailed Welch's t-test to evaluate the methods. We assert significance at p-value ≤ 0.025.

B. Signal-Based Enhancement Baselines

We compare our selective feature enhancement models with SER models trained where all the features are enhanced. We use two types of signal-based enhancement models: MetricGAN [8] and DCCRN [29]. MetricGAN uses a generative adversarial network that is the same as our feature enhancement model (Sec. IV-B). However, MetricGAN enhances the magnitude spectrum instead of the acoustic feature, as our approach which directly enhances the LLDs extracted from the noisy signal. After using MetricGAN, we extract the LLDs from the enhanced signal. When training MetricGAN, the discriminator is trained to predict the normalized PESQ and STOI metrics. The scores of PESQ and STOI range from -0.5 to 4.5 and from 0 to 1, respectively. We apply min-max normalization to the PESQ so that the best score is 1 and the worst score is 0, matching the range for STOI. The discriminator is trained to predict the normalized PESQ and STOI scores from the generator's output. For example, when clean speech is fed into the discriminator, its output should be 1. The generator is trained to produce clean speech by making the output of the discriminator to be 1, which means that the input of the discriminator is clean. We follow the same architecture and training procedure as described in the study of Fu et al. [8]. The only difference is that we update the generator 10 more times than the discriminator to balance the loss between the generator and the discriminator. The noisy version of the MSP-Podcast corpus contains various non-stationary noises. In contrast, the original study of Fu et al. [8] used speech contaminated with stationary noise including machinery noise and water sounds. Therefore, we assume that enhancing our noisy corpus is a harder task than enhancing the speech in the original study, which makes the discriminator learns faster than the generator.

When using DCCRN, we also enhance the signal, before extracting the LLDs from the enhanced signal. However, DCCRN also enhances the phase spectrum (MetricGAN only

TABLE II
PERFORMANCE OF THE SIGNAL-BASED ENHANCEMENT MODELS USING THE PESQ AND STOI METRICS. WE EVALUATE THE DCCRN AND METRICGAN METHODS FOR NOISY SPEECH USING THE 10DB, 5DB, AND 0DB CONDITIONS.

	10dB		5dB		0dB	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
Noisy signal	2.493	0.707	2.305	0.726	2.170	0.590
MetricGAN	2.588	0.790	2.370	0.755	2.101	0.630
DCCRN	2.712	0.803	2.498	0.769	2.190	0.650

TABLE III
CCC OF MODELS USING ROBUST FEATURES AND EACH ENHANCEMENT METHOD FOR THE 10DB, 5DB, AND 0DB CONDITIONS. ALL THE ENHANCEMENT MODELS USE NOISY SPEECH SAMPLES WHERE THE ENVIRONMENTAL CONDITION IS MATCHED WITH THE ENVIRONMENTAL CONDITION ON THE TEST SET. WE HIGHLIGHT IN BOLD THE BEST PERFORMANCE PER CONDITION AND PREDICTION TASK. THE SYMBOLS * AND † INDICATE THAT A GIVEN ENHANCEMENT METHOD SHOWS SIGNIFICANT IMPROVEMENT COMPARED TO THE ORIGINAL NOISY SPEECH AND THE ENHANCED LLDS USING FEATURE-BASED ENHANCEMENT, RESPECTIVELY.

	Arousal	Dominance	Valence
		10dB	
Model without Enhancement	0.278	0.288	0.097
Only using robust features	0.364*	0.385*	0.159*
DCCRN	0.151	0.138	0.140*
MetricGAN	0.342*	0.297*	0.110*
Feature enhancement	0.450*	0.400*	0.179*
Selective feature enhancement	**0.530***†	**0.485***†	**0.185***†
		5dB	
Model without Enhancement	0.228	0.262	0.076
Only using robust features	0.302*	0.370*	0.139*
DCCRN	0.111	0.087	0.081*
MetricGAN	0.227	0.247	0.111*
Feature enhancement	0.412*	0.403*	0.177*
Selective feature enhancement	**0.467***†	**0.457***†	**0.178***
		0dB	
Model without Enhancement	0.194	0.214	0.058
Only using robust features	0.268*	0.321*	0.117*
DCCRN	0.083	0.068	0.081*
MetricGAN	0.168	0.135	0.073*
Feature enhancement	0.393*	0.376*	0.147*
Selective feature enhancement	**0.397***†	**0.392***†	**0.151***†

enhances the magnitude spectrum). By using complex convolution operation, DCCRN is trained to generate a clean spectrogram from the noisy one by simultaneously enhancing the magnitude and phase of the spectrum. Unlike MetricGAN, it does not exploit the adversarial training procedure, but only uses the *scale-invariant source-to-noise ratio* (SI-SNR) loss by using a clean spectrum and the enhanced spectrum (i.e., the output of DCCRN). We follow the same architecture and training procedure as described in the original work [29].

For our experiment, we want to avoid the data mismatch problem in training the enhancement model. Therefore, we train and test this module on the MSP-Podcast corpus. This strategy leads to a better result (PESQ score = 2.190) than using a separate corpus such as the DNS-3 challenge database [53] (PESQ score = 1.739), which includes additive ambient noise sounds without speech and synthetic reverberation. We train both signal-based enhancement models with the three conditions of the noisy version of the MSP-Podcast: 10db, 5db, and 0db. In our preliminary study, we observe that training the enhancement model only with a low SNR condition does not yield better sound quality even when the test set has the same condition used during training. For this reason, we borrow the curriculum learning strategy. We first train the model with the higher SNR condition (easier task). Starting with this model, we train with the desired lower SNR condition (harder task). For example, when training the enhancement model with the 5dB condition, we first train the model with the 10dB condition, before retraining the model with the 5dB condition.

To make a fair comparison, we need to assure that both signal-based enhancement models improve the quality and the intelligibility of the noisy signal. For this reason, we measure the PESQ [26] and STOI [27] metrics with the clean, enhanced and noisy speech signals. Table II shows the sound quality from both enhancement models. Our analysis shows that both models improve the signal quality and intelligibility of the noisy speech, suggesting that both enhancement models are well-trained.

VI. EXPERIMENTAL EVALUATION

A. Matched Conditions

First, we analyze the performance of our approach where the noise level is the same for training the enhancement model, setting the feature selection criterion and testing the model. Table III shows the average CCC value over the ten trials for each enhancement method. Both signal-based enhancement baselines fail to improve the performance for arousal and dominance in the low SNR condition. In contrast, the feature-based enhancement always shows better performance than the baselines for all types of emotional attributes and environmental conditions. This result shows that enhancing the quality of speech does not always lead to performance improvements for SER tasks. Instead of extracting the LLDs from the enhanced signal, it is better to directly enhance the noisy LLDs to increase the performance. We will further analyze why the feature-based enhancement method performs better than the signal-based enhancement method in Section VI-E.

Table III shows that our selective feature enhancement method further improves the performance compared to a system that enhances all the features. In the 10dB condition, our proposed method improves the performance of the approach that enhances all the features by 17.7% for arousal, 21.2% for dominance, and 3.3% for valence. Our result indicates that there exist features that are already resilient to the noise and can deteriorate the recognition performance if they are enhanced. This result highlights the importance of assessing the noise robustness of each feature before the feature enhancement module. According to the result, our performance and robustness criteria can be good options for selecting the discriminative features that should be kept and the ones that should be enhanced. Table III also shows that while using only robust features without the feature enhancement module performs better than the baselines, it is not better than enhancing all the features or our selective feature enhancement approach which achieves the best performance. This result indicates that it is important to combine robust features with

TABLE IV

CCC of models using each enhancement method for the 5dB and 0dB conditions, using the enhancement model trained with 10dB condition. In our selective feature enhancement method, the LLDs are selected using the 10dB condition. We highlight in bold the best performance per condition and prediction task. The symbols * and † indicate that a given enhancement method shows significant improvement compared to the original noisy speech and the enhanced LLDs using feature-based enhancement, respectively.

	Arousal	Dominance	Valence
	5dB		
Model without Enhancement	0.228	0.262	0.076
DCCRN	0.101	0.070	0.064
MetricGAN	0.207	0.192	0.078
Feature enhancement	0.441*	0.405*	0.149*
Selective feature enhancement	**0.485***†	**0.454***†	**0.150***
	0dB		
Model without Enhancement	0.194	0.214	0.058
DCCRN	0.078	0.068	0.060
MetricGAN	0.142	0.131	0.058
Feature enhancement	0.390*	0.365*	**0.121***
Selective feature enhancement	**0.422***†	**0.397***†	0.120*

TABLE V

CCC of models using multiple SNR levels for training feature enhancement method in 10dB, 5dB, and 0dB conditions. We compare the performance of enhancing all features and using our proposed selective feature enhancement method. We highlight in bold the best performance per condition and prediction task. The symbols * and † indicate that a given enhancement method shows significant improvement compared to the model without enhancement and the feature enhancement with enhancing all the LLDs, respectively.

	Arousal	Dominance	Valence
	10dB		
Model without Enhancement	0.278	0.288	0.097
Feature enhancement	0.410*	0.423*	0.175*
Selective feature enhancement	**0.527***†	**0.444***†	**0.182***†
	5dB		
Model without Enhancement	0.228	0.262	0.076
Feature enhancement	0.441*	**0.426***	0.175*
Selective feature enhancement	**0.503***†	0.417*	**0.177***
	0dB		
Model without Enhancement	0.194	0.214	0.058
Feature enhancement	0.405*	**0.396***	0.148*
Selective feature enhancement	**0.457***†	0.353*	**0.150***

the enhanced weak features.

B. Mismatched Conditions

We also evaluate the models in mismatched conditions. The enhancement model is trained with the 10dB condition. We also select the coverage and best feature selection criterion using the 10dB condition. Then, the models are tested with either the 5dB or 0dB condition. Table IV shows that even when the enhancement model is trained with mismatched SNR conditions, the feature-based enhancement model shows better performance than other methods. Both signal-based enhancement baselines fail to achieve significant performance improvements over a model that does not incorporate any enhancement method. In fact, these approaches result in lower performance for most conditions. In contrast, the performance of our selective feature enhancement approach is significantly better than the model without enhancement and the signal-based enhancement methods. This result shows the strength of the feature-based enhancement approach in mismatched SNR conditions. Even when there is no information about the SNR level in the target environment, the feature-based enhancement model can improve the recognition performance.

Even though the feature selection is performed with a mismatched dataset, our selective feature enhancement can further improve the performance of a model trained by enhancing all features for arousal and dominance. When we compare our model with the feature enhancement approach, we observe performance gains of 9.9% (5dB) and 8.2% (0dB) for arousal, and 12.1% (5dB) and 8.7% (0dB) for dominance. The performances for valence are very similar. Therefore, enhancing only weak features is still an effective approach even when the features are assessed with speech collected in an environment with mismatched conditions.

C. Multi-Noise Condition Training

In Sections VI-A and VI-B, we use only one SNR level to train the enhancement model for our experiments to provide the results when SNR is totally matched or mismatched between training and testing conditions. However, it is common to use multiple SNR levels to increase the generalization of real-world conditions. Therefore, we also compare the performance of using multiple SNR levels to train the enhancement model to validate the generalization ability of our proposed selective feature enhancement method.

We train the single feature enhancement network with a range of SNR levels. The enhancement model is trained with 10dB, 5dB, and 0dB conditions. We randomly select SNR levels among those three SNR levels for each noisy speech sample during training. We also select the robust features by using multiple SNR levels. We first define the best coverage for each SNR level and emotional attribute. We randomly select the best coverage for each attribute among those three SNR levels. We test this framework with the 10dB, 5dB, and 0dB conditions.

Table V illustrates the result of enhancing all the features and our proposed selective feature enhancement with multiple SNR levels. Consistent with the results in matched conditions, using our proposed selective feature enhancement method generally shows better performance than enhancing all the features in the multiple SNR condition. For example, our selective feature enhancement method yields the best performance for the arousal prediction task across all conditions. This result shows that our approach works well in predicting arousal, even when training the model with multiple SNR levels. However, our proposed method does not yield significantly better performance for dominance and valence in the 5dB and 0dB conditions, which have low SNR levels. This finding is different from the result in mismatched conditions, where the robust features are consistently selected by the 10dB condition. This result shows that the robust features need to be selected by fixing the SNR level, even when the SNR level is mismatched from the testing condition.

Compared with training in matched conditions (Table III), the training of the feature enhancement model with multi-SNR

TABLE VI
CCC OF MODELS IN THE NOISY VERSION OF THE IEMOCAP CORPUS. WE COMPARE THE PERFORMANCE OF ENHANCING ALL FEATURES AND USING OUR PROPOSED SELECTIVE FEATURE ENHANCEMENT METHOD. WE HIGHLIGHT IN BOLD THE BEST PERFORMANCE PER CONDITION AND PREDICTION TASK. THE SYMBOLS * AND † INDICATE THAT A GIVEN ENHANCEMENT METHOD SHOWS SIGNIFICANT IMPROVEMENT COMPARED TO THE MODEL WITHOUT ENHANCEMENT AND THE FEATURE ENHANCEMENT WITH ENHANCING ALL THE LLDS, RESPECTIVELY.

	Arousal	Dominance	Valence
Model without Enhancement	0.530	0.417	0.083
Feature enhancement	0.605*	0.494*	0.126*
Selective feature enhancement	**0.636*†**	**0.533*†**	**0.183*†**

conditions shows lower performance in the 10dB condition, but better performance for arousal and dominance in the 5dB and 0dB conditions. We assume that training the feature enhancement model with a low SNR level condition makes the training difficult, which could be alleviated by introducing multiple SNR level conditions where some samples have higher SNR. Interestingly, the multi-SNR condition does not show better performance for valence across all the conditions. This result indicates that including a high SNR level in the enhancement model's training set does not lead to improved performance for valence in noisy conditions.

D. Experiments in Different Data Distribution

Previous sections provides our experimental results using the clean and noisy version of the MSP-Podcast corpus. To assess our model in different data distributions, we test our model with a different emotional speech corpus and different types of noise sounds. We use the IEMOCAP corpus [54] for our clean speech and contaminate it with the noise sounds from the DNS-3 challenge dataset [53]. Each clean speech sample in the IEMOCAP corpus is contaminated with a noise sound in the DNS-3 challenge dataset, where the SNR level ranges between –5dB and 20dB. We also add real and synthetic room impulse responses to contaminate the IEMOCAP speech samples. We use this clean and noisy version of the IEMOCAP corpus to train the emotion recognition model, feature probe model, and feature enhancement model. The IEMOCAP corpus has five sessions in its corpus. We use three sessions for the training set, one session for the development set, and the remaining one session for the testing set. We do not overlap the noise sounds among the training, development, and testing sets when contaminating the speech samples of the IEMOCAP corpus. The best coverages for the noisy version of the IEMOCAP corpus are 70% with the performance criterion for arousal, 20% with the performance criterion for dominance, and 50% with the joint criterion for valence. We compare the model trained without using feature enhancement, with enhancing all features, and with our proposed selective feature enhancement model. We report the average CCC of five trials with different feature enhancement models.

Table VI compares the SER performance of these models in a noisy condition for the IEMOCAP dataset. Our proposed selective feature enhancement method yields the best performance for all the emotional attributes in the noisy version of the IEMOCAP corpus. Our proposed framework improves the performance by 5.1% (arousal), 7.9% (dominance), and 45.2% (valence) compared with the approach that enhances all the LLDs. This result demonstrates that our proposed selective feature enhancement framework can be applicable to the different emotional speech data distribution and noise types.

TABLE VII
MSE AND CORRELATION COEFFICIENT BETWEEN THE ENHANCED FEATURES USING EITHER THE METRICGAN APPROACH OR THE FEATURE-BASED ENHANCEMENT METHOD AND THE LLDS EXTRACTED FROM THE CLEAN AND 10DB CONDITIONS.

	MSE		Correlation Coef.	
	clean	10dB	clean	10dB
Signal-based (MetricGAN)	11.479	1.143	0.788	0.962
Feature enhancement	8.428	11.072	0.775	0.858

TABLE VIII
AVERAGE PERFORMANCE OF SER MODELS TRAINED WITH ONE LLD WHEN EVALUATED WITH THE TOP FIVE LLDS FOR EACH OF THE CONDITIONS (NOISY SPEECH, SIGNAL-BASED ENHANCEMENT, AND FEATURE-BASED ENHANCEMENT).

	Arousal	Dominance	Valence
Noisy speech	0.244	0.203	0.154
Signal-based (MetricGAN)	0.217	0.187	0.153
Feature enhancement	**0.250**	**0.267**	**0.158**

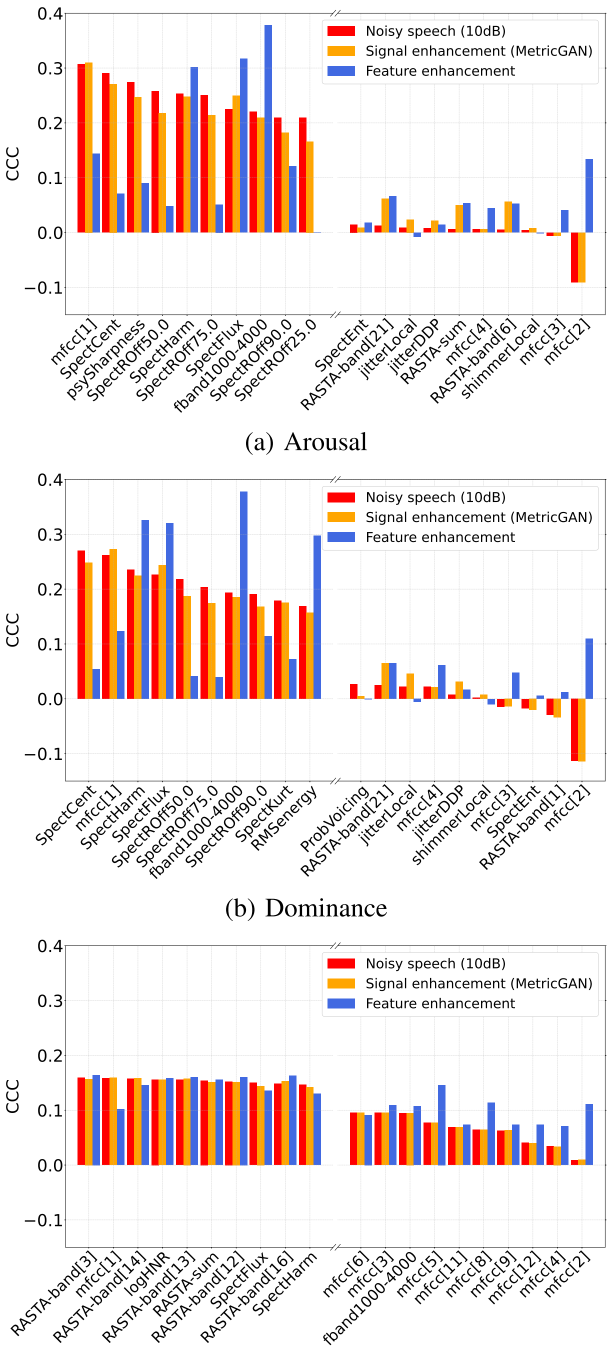
E. Feature-Based or Signal-Based Enhancement

As we can see in Sections VI-A and VI-B, using signal-based enhancement usually performs worse than using feature-based enhancement. Moreover, the use of signal-based enhancement sometimes degrades the performance of a system trained without any enhancement method. For this reason, this section analyzes in more detail why feature-based enhancement is better than signal-based enhancement for SER tasks in a noisy environment. For signal-based enhancement, we select the MetricGAN approach, which is one of the baselines that we used in the previous evaluations. For feature-based enhancement, we use the GAN-based feature enhancement model. For the analyses in this section, we enhance the noisy speech from the 10dB condition of the noisy version of the MSP-Podcast corpus.

We first analyze which enhancement method yields better features. We expect that the new features after the enhancement process will be close to the features from clean speech, and far from the features from noisy speech (10dB) before the enhancement. We calculate the *mean squared error* (MSE) and correlation coefficient between the LLDs generated from each enhancement method and the LLDs extracted from either the clean or noisy speech signals. For the signal-based enhancement, we first enhance the noisy speech signals and then extract the LLDs from the enhanced signals. For the feature-based enhancement, we first extract the LLDs from the noisy signal and then enhance the extracted LLDs. Table VII shows the results. The LLDs extracted from the speech enhanced by the MetricGAN approach have a shorter distance to the noisy LLDs than to the clean LLDs, indicating that the enhancement process was not very successful. In contrast, the LLDs enhanced with the feature-based enhancement approach have a shorter distance to the clean LLDs than to the noisy LLDs. These results are also supported by the correlation results. Using signal-based enhancement leads to LLDs that are very correlated to noisy LLDs ($\rho = 0.962$). This correlation is reduced when using feature-based enhancement ($\rho = 0.858$). Even if the speech quality and intelligibility are improved with the signal-based enhancement method, it does not help to improve the acoustic features needed for the SER task.

We conduct a discriminative analysis per feature of the LLDs extracted from the noisy and enhanced speech signal, and the LLDs enhanced by our feature enhancement model. Each SER model is trained with a single LLD using the same approach described in Section IV-A using the clean version of the MSP-Podcast corpus. The single-feature models are then evaluated using LLDs from these three feature sets (noisy speech, signal-enhanced LLDs, and feature-enhanced LLDs). We train 10 different single-feature models for each LLD, reporting the average performances in the test set. Figure 4 shows the performance for some of the LLDs. Interestingly, the performances of single-feature models tested with LLDs extracted from either the enhanced signal or noisy speech are very similar. However, the performances using feature-based enhancement are very different from the performance obtained with LLDs from noisy speech. Although feature-based enhancement decreases the performance for some features, we observe that this approach leads to the highest performance for some other features (e.g., for arousal SpectHarm, SpectFlux, fband1000-4000, and mfcc[2]). We hypothesize that these high-performing features can compensate for the low performance of other features when all LLDs are combined. We quantify this hypothesis by averaging the SER performance for the top five LLDs extracted from either the noisy speech, signal-enhanced speech, or feature-enhanced method. Table VIII shows the performance. Compared with using noisy speech and signal-based enhancement, feature-based enhancement leads to the highest average performance using single features for all the emotional attributes. We conclude that feature-based enhancement can lead to higher improvements for the top features than signal-based enhancement and that not all the features must be enhanced.

In addition, feature-based enhancement frequently shows clear improvements in the weak features identified in the analysis of Section IV-C, which is not the case for the signal-enhancement approach. Table IX reports the average performance obtained only with weak features. For this analysis, we adopt the weak features selected under the 10dB SNR condition. When using weak features, the feature enhancement leads to better performance than models tested using either noisy speech or signal-based enhancement for arousal and dominance. This result shows the benefits of combining the feature enhancement approach and our proposed robust feature selection method.



(a) Arousal

(b) Dominance

(c) Valence

Fig. 4. Performance of models trained with one feature evaluated with LLDs enhanced with the feature-based enhancement approach, and with LLDs extracted from the signal-based enhanced speech (MetricGAN), and noisy speech (10dB condition). We illustrate 20 LLDs corresponding to the top 10 and bottom 10 CCC performance in the 10dB noisy condition.

VII. Conclusions

Instead of enhancing all the features, this study proposed to enhance only the features that disrupt the SER prediction due to noise and to keep the features that are resilient. To select those features, we train multiple single-feature probe models, ranking the LLDs based on the performance (i.e., features that lead to good performance) and robustness (i.e., features that lead to a similar performance in noisy and clean conditions) criteria. We trained an emotion recognition model with features extracted from clean speech. Our selective feature enhancement approach can improve the prediction of

TABLE IX
AVERAGE CCC PERFORMANCE ACHIEVED BY TESTING THE SER MODELS TRAINED WITH ONE LLD WITH THE WEAK FEATURES IDENTIFIED IN SECTION IV-C. WE DEFINE THE ROBUST AND WEAK FEATURES BY USING THE DEVELOPMENT SET UNDER THE 10DB CONDITION.

	Arousal	Dominance	Valence
Noisy speech	0.051	0.045	0.113
Signal-based (MetricGAN)	0.078	0.068	0.115
Feature enhancement	0.082	0.079	0.115

emotional attribute scores under matched and mismatched environmental conditions. This observation remains consistent, even when the environmental conditions utilized for training the feature enhancement model and selecting weak and robust features do not align with the target environment for testing the models. Our analysis revealed that employing feature-based enhancement results in superior performance compared to using signal-based enhancement. We analyze the performance of SER systems trained with clean speech using a single LLD. We evaluated these single-feature models with LLDs extracted from noisy speech, signal-enhanced speech, and feature-based enhancement models. Signal-based enhancement does not clearly improve the performances using individual LLDs. In contrast, the feature-based enhancement approach leads to clear improvements for the top-performing features, which compensate for other features when all the LLDs are combined. Our analysis also showed that some features lead to lower SER performance after they are enhanced by the feature-based enhancement model, implying the importance that not all the features need to be enhanced.

A limitation of our feature selection method is that it requires training multiple feature probe models for each target environment, which consumes computational resources as we adapt the SER model to multiple environments. We plan to investigate how to optimize our feature selection procedure to simultaneously deal with multiple noisy environments. Moreover, we also plan to study if our feature enhancement method is applicable to SER models built using self-supervised speech representations, such as Wav2Vec2.0 [55] or HuBERT [56], which have led to good performance in recent SER studies [57]–[59].

REFERENCES

[1] R. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C.Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, November 2021.

[3] 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.

[4] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3593–3597.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015)*, Scottsdale, AZ, USA, December 2015, pp. 504–511.

[6] 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.

[7] 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.

[8] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds. Long Beach, CA, USA: PMLR, June 2019, vol. 97, pp. 2031–2041.

[9] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, April 1979.

[10] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1979)*, Washington, DC, USA, April 1979, pp. 208–211.

[11] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. Cambridge, MA: MIT press, March 1964.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[13] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, October 1992.

[14] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, February 1991.

[15] D. W. Tufts and A. A. Shah, "Estimation of a signal waveform from noisy data using low-rank approximation to a data matrix," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1716–1721, April 1993.

[16] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[17] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, January 2015.

[18] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Latent Variable Analysis and Signal Separation (LVA/ICA 2015)*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds. Liberec, Czech Republic: Springer International Publishing, August 2015, vol. 9237, pp. 91–99.

[19] S. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1993–1997.

[20] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 6875–6879.

[21] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech 2013*, Lyon, France, August 2013, pp. 436–440.

[22] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM 2017)*, Donostia, Spain, May 2017, pp. 1–5.

[23] S. Pascual, A. Bonafonte, and J. Serrá, "SEGAN: Speech enhancement generative adversarial network," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 3642–3646.

[24] H. Phan, I. McLoughlin, L. Pham, O. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, September 2020.

[25] L. Li, Z. Lu, T. Watzel, L. Kürzinger, and G. Rigoll, "Light-weight self-attention augmented generative adversarial networks for speech enhancement," *Electronics*, vol. 10, no. 13, p. Electronics, June 2021.

[26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *IEEE International*

Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001), vol. 2, Salt Lake City, UT, USA, May 2001, pp. 749–752.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, September 2011.

[28] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 6865–6869.

[29] D. Hu, Z. Wang, H. Xiong, D. Wang, F. Nie, and D. Dou, "Curriculum audiovisual learning," *ArXiv e-prints (arXiv:2001.09414v1)*, pp. 1–10, January 2020.

[30] 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.

[31] 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.

[32] 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.

[33] H. Valpola, "From neural PCA to deep unsupervised learning," in *Advances in Independent Component Analysis and Learning Machines*, E. Bingham, S. Kaski, J. Laaksonen, and J. Lampinen, Eds. London, UK: Academic Press, May 2015, pp. 143–171.

[34] 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.

[35] J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and J. Yi, "Speech emotion recognition using semi-supervised learning with ladder networks," in *Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia 2018)*, Beijing, China, May 2018, pp. 1–5.

[36] 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.

[37] L. Goncalves and C. Busso, "Learning cross-modal audiovisual representations with ladder networks for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[38] 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.

[39] ——, "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.

[40] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems (NIPS 2016)*, vol. 29, Barcelona, Spain, December 2016, pp. 1–9.

[41] A. Georgogiannis and V. Digalakis, "Speech emotion recognition using non-linear Teager energy based features in noisy environments," in *European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, August 2012, pp. 2045–2049.

[42] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *ISCA Speech Prosody*. Dresden, Germany: ISCA, May 2006.

[43] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "An unsupervised frame selection technique for robust emotion recognition in noisy speech," in *European Signal Processing Conference (EUSIPCO 2018)*, Rome, Italy, September 2018, pp. 2055–2059.

[44] Ł. Juszkiewicz, "Improving noise robustness of speech emotion recognition system," in *Intelligent Distributed Computing VII*, ser. International Symposium on Intelligent Distributed Computing (IDC 2013), F. Zavoral, J. Jung, and C. Badica, Eds. Prague, Czech Republic: Springer International Publishing, 2014, vol. 511, pp. 223–232.

[45] 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.

[46] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[47] 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.

[48] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[49] 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.

[50] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.

[51] Z.-X. Li, L.-R. Dai, Y. Song, and I. McLoughlin, "A conditional generative model for speech enhancement," *Circuits, Systems, and Signal Processing*, vol. 37, pp. 5005–5022, March 2018.

[52] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder - decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, October 2014, pp. 1724–1734.

[53] C. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 deep noise suppression challenge," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 2796–2800.

[54] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

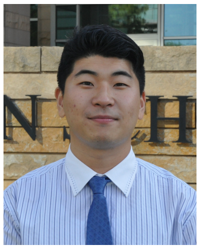
[55] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, Virtual, December 2020, pp. 12 449–12 460.

[56] 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.

[57] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Adapting a self-supervised speech representation for noisy speech emotion recognition by using contrastive teacher-student learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[58] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, September 2023.

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



Seong-Gyun Leem (S'21) received his B.S. and M.S. degree in Computer Science and Engineering at Korea University, Seoul, South Korea in 2018 and 2020, respectively. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas. His current research interests include speech emotion recognition, noisy speech processing, and machine learning.

Daniel Fulford is Associate Professor of Occupational Therapy, Rehabilitation Sciences, and Psychological & Brain Sciences at Boston University. He holds a PhD in Clinical Psychology from the University of Miami and a BA in psychology from UCLA. His research includes laboratory-based and ambulatory studies to better understand motivation and emotion in psychopathology, using smartphones as tools for experience sampling and behavioral sensing. Much of this research informs the development and testing of digital interventions to support psychosocial functioning in individuals with schizophrenia and other serious mental illness.

Jukka-Pekka "JP" Onnela is Associate Professor of Biostatistics in the Department of Biostatistics at the Harvard T.H. Chan School of Public Health of Harvard University. After completing his doctorate in Finland, he completed a junior research fellowship at the University of Oxford, was a Fulbright Scholar at Harvard University, and a postdoctoral fellow at Harvard Medical School. His main interest is in developing quantitative methods in two areas: statistical network science and digital phenotyping.

David Gard is a Professor of Psychology and the Director of the Motivation and Emotion Research Lab at San Franscico State University (SFSU). He received his Ph.D. in clinical psychology from the University of California at Berkeley in 2005 and started as an Assistant Professor at SFSU that year. His research interests are broadly in the area of emotion and motivation dysfunction in various mental health disorders including schizophrenia, bipolar disorder, and depression. He also researches novel treatments in severe mental illness.

Carlos Busso (S'02-M'09-SM'13-F'23) is a professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). His research interest is in human-centered multimodal machine intelligence and applications. His current research includes the broad areas of affective computing, nonverbal behaviors for conversational agents, and machine learning methods for multimodal processing. He is an IEEE Fellow.