NOT ALL FEATURES ARE EQUAL: SELECTION OF ROBUST FEATURES FOR SPEECH EMOTION RECOGNITION IN NOISY ENVIRONMENTS

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

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

ABSTRACT

Speech emotion recognition (SER) system deployed in real-world applications often encounters noisy speech. While most noise compensation techniques consider all acoustic features to have equal impact on the SER model, some acoustic features may be more sensitive to noisy conditions. This paper investigates the noise robustness of each feature in the acoustic feature set. We focus on *low-level descriptors* (LLDs) commonly used in SER systems. We firstly train SER models with clean speech by only using a single LLD. Then, we rank each LLD with respect to the absolute performance on a development set contaminated with noise, and the relative performance decrease from the results from the models trained with the clean set. Our experiment shows that using all the LLDs leads to worse performance than training the system with a single robust LLD. We propose to select a group of robust features according to their performance and robustness in noisy condition. Without using any compensation method, our feature selection methods improve the performance by 24.4% (arousal), 23.9% (dominance), and 43.2% (valence) in the 10dB noisy condition. Moreover, even though the selection is conducted with the 10dB condition, our selection methods also yield performance improvements in unseen noisy recording conditions.

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

1. INTRODUCTION

Speech emotion recognition (SER) is an useful technology that benefits various domains, including education, entertainment, healthcare, security and defense. By deploying SER systems on ubiquitous devices, such as smartphones, this technology can provide valuable services, improving *human-computer interaction* (HCI). However, in real-world applications, SER systems are highly likely to encounter complex types of background noises in arbitrary recording conditions. Such background noises degrade the quality of the original speech signal, leading to the degradation of SER performance. Therefore, increasing the robustness of the SER system in noisy environments is important to improve its use in real-world applications.

Many studies have tried to improve the noise robustness of SER. Some solutions include adding a speech enhancement module [1–3], augmenting the training set with contaminated speech [4, 5], or applying domain adaptation to the SER model [6, 7]. Although the compensation methods differ, most proposed solutions use a complete acoustic feature set during training, regardless of the environmental condition. However, we expect that noise will differently impact parts of speech productions, resulting in features that are more resilient to noise than others. For example, some features may lead to similar performance, regardless of environmental conditions. Other *good* features in clean conditions may lead to lower performance under noisy conditions. We hypothesize that understanding the robustness of each acoustic feature can lead to more robust SER systems.

Exploring the idea that not all features are equal with regard to noise robustness, we analyze each acoustic feature in the presence of noise, proposing a simple, but novel approach to select a robust acoustic feature group without applying any feature- or model-level compensation method. This paper focuses on the *low-level descriptors* (LLDs) of the feature set introduced in the *Interspeech 2013 computational paralinguistics challenge* (ComParE 2013), generally used in previous SER studies. We train emotion recognition models with a single LLD, reporting the recognition performance for each model. We consider two selection criteria : performance and robustness. For performance, we rank LLDs by their absolute performance in noisy conditions. For robustness, we evaluate the relative performance decrease from clean condition to noisy condition. We also consider both criteria by summing up their ranks. For each ranking method, we cumulatively group them to make new feature sets.

Our experiment with the clean and noisy versions of the MSP-Podcast corpus [7, 8] shows that using some LLDs lead to better performance than using all LLDs in noisy conditions. Moreover, in the noisy conditions, selecting LLDs by their performance and robustness in noisy condition leads to better performance than randomly selecting LLDs or using all the LLDs. We observe this result even when testing in unseen noisy conditions where their *signal-to-noise ratio* (SNR) is different from the dataset used for the feature selection. For example, when selecting LLDs by using the development set for the 10dB condition, grouping features by robustness improves the performance over a system trained with all the LLDs by 29.7% (5dB), 35.5% (0dB) for dominance and 51.3% (5dB), 44.8% (0dB) for valence, and grouping features by joint criterion improves the performance by 49.1% (5dB), 50.5% (0dB) for arousal.

2. RELATED WORKS

Compensating background noise has been actively studied in the SER field. Some studies have explored feature transformation methods, which are designed to denoise the features from contaminated speech to make it fit the pre-trained model that is only trained with speech without noise. For example, Huang et al. [1] showed that spectral subtraction and perceptual masking-based speech enhancement also showed better performance in predicting arousal and valence levels. Juszkiewicz [2] successfully applied histogram equalization of *Mel-frequency cepstral coefficient* (MFCC) to increase the emotion classification accuracy in noisy audio, which was also shown to be effective in automatic speech recognition [9]. Triantafyllopoulos et al. [3] used a convolutional neural network with residual blocks as a feature enhancement module to improve

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

SER performance.

Data augmentation methods have also been studied to address this problem. Lakomkin et al. [4] augmented the training data by changing the tempo and loudness of the recordings and adding background noise and reverberation by considering room impulse responses. This method showed significant improvements both in emotion classification and the prediction of arousal and valence scores. Tiwari et al. [5] utilized various types of noise from the NOISEX-92 database and modulated white noise to augment the training data.

Domain adaptation approaches can also improve the recognition performance in noisy conditions by compensating environmental mismatch between train and test sets. Leem et al. [7] proposed a semi-supervised domain adaptation approach for SER to compensate for the environmental mismatch between train and test sets. Building on the ladder network backbone [10–12], the study decoupled the emotional and reconstruction embedding to reduce the influence of background noise on emotion predictions.

Unlike previous studies, our method does not change the architecture of the original SER system nor augment its training set. Instead, it only selects robust features against the environmental mismatch from the clean training set. Although Schuller et al. [13] also proposed to selectively use input features based on the performance in noisy condition, our method does not only rely on the performance rank, but we also consider the noise robustness by measuring relative performance decrease, leading to further performance improvement. That method also considered matched conditions where the noise in the train and test conditions were the same. Our approach does not need matched conditions, making it robust to unseen noisy environment. Our study is also different from the study of Pandharipande et al. [14], which discards the noisy frames instead of the features.

3. RESOURCES

3.1. Datasets

To assess the robustness of features in a real-world environment, we use a clean emotional speech corpus and its noisy version, which simulates real-world recording conditions. For the clean speech corpus, we use the MSP-Podcast corpus (version 1.8) [8]. It contains a large amount of natural and diverse emotional speech samples collected from various podcast recordings with over 113 hours. We use the retrieval approach proposed in Mariooryad et al. [15] to choose samples expected to have the target emotions. For the annotation, we use a modified version of the crowdsourcing protocol proposed in Burmania et al. [16] to increase the reliability of the annotations. 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. All samples are selected when the predicted SNR is above 20dB. The recordings do not have background music. They are formatted at a sampling rate of 16kHz. We used 44,879 samples for the train set, 7,800 for the development set, and 15,326 for the test set for the clean condition. The partitions aim to create speaker-independent sets. All models are trained with the train set using the clean condition.

For the noisy speech corpus, we use the noisy version of the MSP-Podcast corpus, which was introduced in Leem et al. [7]. 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, and contain human voices, background music, and various types of sound effects. After simultaneously playing speech and noise sounds with the speakers of two portable devices, those mixed sounds are recorded on a smartphone, replicating daily devices used in real-world applications. All noisy speech samples are collected with three different levels of SNR by changing the distance between the speakers and smartphone. According to their estimated SNR, the conditions are named 10dB, 5dB, and 0dB, respectively. For each noisy speech sample, their emotional labels are directly transferred from the clean version of the MSP-Podcast corpus. We used 7,800 samples of the development set in the 10dB condition to assess the noise robustness of each feature. For the evaluation, we used the same sentences in the test set (15,326 samples) for all the three noisy recording conditions.

Table 1. 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]

3.2. Acoustic features

In our experiments, we use the 65 LLDs from the ComParE 2013 feature set extracted with the openSMILE Toolkit [17], as described in Table 1. To extract each LLD, a 20ms window is applied for the RMSenergy feature, and 60ms is applied for other LLDs with 10ms step size. This approach creates a frame-level representation for each speech signal.

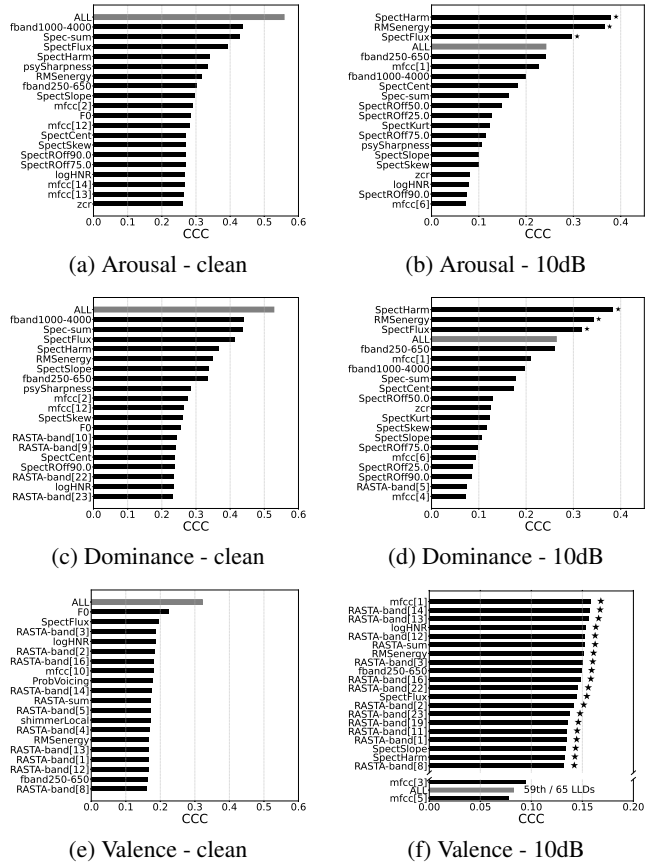
To avoid shifts in the feature distributions due to environmental noise condition, we apply the Z-normalization to the features. We regard the development set of each noisy recording condition as speech samples obtained from the target environment. Then, we use their mean and standard deviation to normalize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

4. METHODOLOGY

4.1. Motivation

Although the ComParE 2013 feature set contains various types of LLDs, the noise robustness of each LLD has not been evaluated in previous studies. Therefore, we compare the emotion recognition performance by only using a single LLD.

We borrow the SER model used as a baseline for the frame-level acoustic features in the study of Parthasarathy and Busso [11]. This model consists of five blocks of 1D convolution layers and



(a) Arousal - clean (b) Arousal - 10dB

(c) Dominance - clean (d) Dominance - 10dB

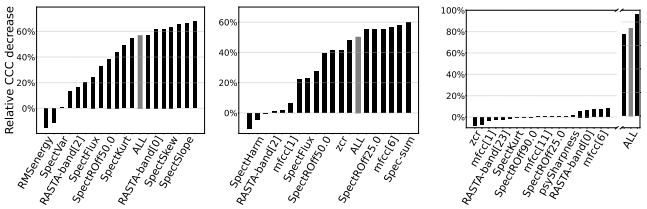
(e) Valence - clean (f) Valence - 10dB

Fig. 1. CCC of models trained with either a single LLD or all the LLDs in clean and noisy conditions (10dB). We mark with a star when the model trained with a single LLD is significantly better than the baseline model trained with all the LLDs.

max-pooling layers, two fully connected layers, each of them implemented with 256 nodes, and an 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 put 10% dropout between the input and first convolution layer and between the last convolution layer and the first fully connected layers. We use the multitask learning approach proposed in Parthasarathy and Busso [18] that showed good performance in predicting emotional attribute scores. Equation 1 illustrates the cost function of our model.

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

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 prediction model, which showed the best performance reported in Parthasarathy and Busso [11]. 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 [19] with a 0.00005 learning rate to optimize the parameters. We train models for 25 epochs with a mini-batch of 512 sentences. While augmenting the database with noisy recording can lead to improvements, this solution is not effective if the noise in the environment is different. We avoid this strategy by training the models with clean speech, aiming to increase the robustness with resilient features.



(a) Arousal (b) Dominance (c) Valence

Fig. 2. Relative CCC decrease for models evaluated in the clean and 10dB conditions.

To assess the robustness of each feature, we train multiple models by changing the type of input feature. Therefore, 65 different models are trained using different types of LLDs for each prediction task. We compare the performance of using a single feature with the baseline model trained with all the LLDs. After training the model, we evaluate the CCC performance on the development sets for the clean and 10dB conditions per emotional attribute. We ran ten experiments for each LLD with different initial weights, reporting the average CCC. To assess the significance of the results, we conduct a two-tailed Welch's t-test between the results achieved when using all LLDs, and when using a single LLD. We assert the significance when p-value ≤ 0.025. Figure 1 illustrates the CCC for this evaluation. We illustrate the top 20 LLDs for each task. The black bar and gray bar demonstrate the performance of using a single LLD and all the LLDs, respectively. In the clean condition, using all the features always shows the best performance for all the tasks. This result evidently indicates the effectiveness of this popular feature set for SER tasks. However, when the model is tested in a noisy condition, some of the single features show better performance than results from model trained with the entire feature set. Moreover, in valence predictions, using all LLDs leads to one of the worst performance. Out of 65 LLDs, 43 features show significantly better performance than using all the features. This result indicates that selectively using acoustic features can improve the performance in noisy conditions.

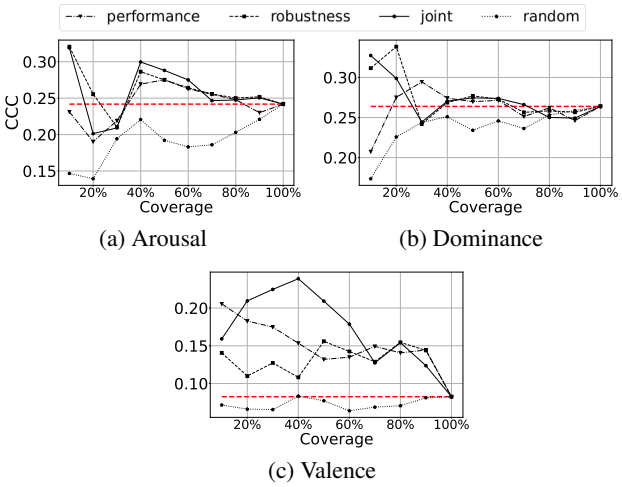
Figure 2 illustrates the relative decrease in CCC from the clean condition to the 10dB condition. We illustrate the top 20 LLDs for each task. Some features show less performance decrease in a noisy condition than using all the features. For valence predictions, most of the features show a smaller decrease than using all the features. This result implies that not all the features are influenced the same by the environmental noise condition. There are robust features against environmental mismatches. There are also features that disrupt the prediction in noisy conditions, which should be avoided.

4.2. Cumulative performance by adding LLD

Section 4.1 showed that some features are more resilient than other against noise. To find a robust subset, we select the feature by cumulatively including LLDs based on their performances (criterion 1) and their robustness (criterion 2). For the performance criterion, we rank features based on the CCC performance on the development set in the 10dB condition. For the robustness criterion, we rank features based on the relative CCC decrease from clean to 10dB conditions. We also combined both criteria by adding those two ranks (joint). Using one of the three criteria (performance, robustness, joint), we add LLDs in increments of 10% from the top 10% to the 100% coverage. As a baseline, we randomly select LLDs to match the target coverage level. We use the same hyper-parameters for training as the one reported in Section 4.1. Although the feature selections are based on the performance of a noisy condition, only the clean training set is used to train the models. We conducted ten experiments,

Table 2. CCC comparison of using subset of features based on performance, robustness, and joint selection criteria. We compare these models with random feature selection (random) and the baseline trained with all the features (all). We use 10%, 20%, and 40% coverage for arousal (Aro.), dominance (Dom.), and valence (Val.) prediction, respectively. We highlight in bold the best performance per condition. The symbols † and * indicate that a feature set shows significant improvement compared to the random and all settings, respectively.

	Clean			10dB			5dB			0dB		
	Aro.	Dom.	Val.	Aro.	Dom.	Val.	Aro.	Dom.	Val.	Aro.	Dom.	Val.
Performance	0.401	0.399	0.165	0.265†	0.298	0.109†	0.288*†	0.305*	0.096*†	0.236*†	0.258*†	0.083*†
Robustness	0.379	0.429	0.151	0.316†	**0.357***†	**0.139***†	0.252†	**0.34***†	**0.115***†	0.201†	**0.29***†	**0.084***†
Joint	0.414	0.413	0.192	**0.346***†	0.319*	0.115*†	**0.34***†	0.302*	0.109*†	**0.292***†	0.257*	0.076*†
Random	0.376	0.405	0.181	0.157	0.239	0.074	0.141	0.221	0.063	0.116	0.183	0.048
All	**0.572**	**0.505**	**0.212**	0.278	0.288	0.097	0.228	0.262	0.076	0.194	0.214	0.058



(a) Arousal (b) Dominance

(c) Valence

Fig. 3. CCC performance for each feature coverage, and selection criterion in the 10dB condition. The red dashed line marks the performance of the baseline using all the LLDs.

reporting the average CCC on the development set for each coverage and selection criterion.

Figure 3 illustrates the CCC as a function of number of features for each selection criterion. Our three selection methods show better performance than the random selection across all feature coverage, showing the merit of the criterion. The performance of the random selection criterion leads to worse performance than using all the features. However, in the other three selection criteria, there exist feature subsets that show better performance than the baseline using all the features. In the development set, the robustness criterion reaches the best performance for arousal (10% coverage) and dominance (20% coverage). For the valence, the joint criterion leads to the best performance (40% coverage). Selecting a feature group based on the noise robustness of the feature can improve the performance, even without using any compensation method.

4.3. Proposed approach

According to the observations described in Sections 4.1 and 4.2, we propose to use a subset of robust acoustic features to improve the recognition performance of a SER system in noisy conditions. Based on the previous experiments, we select 10%, 20%, and 40% coverage for the arousal, dominance, and valence prediction tasks, respectively. With these feature coverage, we evaluate the recognition performance in the test set for each of the three selection criteria: performance, robustness, and joint. We compare those performances with the baseline model trained with all the features. We also consider a second baseline trained by randomly selecting features until reaching the target coverage. To assess the effectiveness of the feature selection in unseen noisy conditions, we also evaluate the performance on the test set in the 5dB and 0dB conditions.

5. RESULTS

We run ten experiments to evaluate the significance of our selection criterion. We conduct a two-tailed Welch's t-test to evaluate the methods. We assert significance at p-value ≤ 0.025.

Table 2 shows the average CCC value over the ten trials for each method. When testing in clean condition, none of the feature subsets improves the performance of the baseline model trained with all the features. However, in noisy conditions, the joint criterion always show significantly better performance than using all features for all prediction tasks. Even though we select the features only based on the development set in the 10dB condition, those feature selection methods can yield performance improvement in the 5dB and 0dB conditions, which are unseen in the feature selection stages. For arousal prediction, the joint selection criterion shows the best performance, with relative improvements over the baseline equal to 24.4% (10dB), 49.1% (5dB), and 50.5% (0dB). In the dominance and valence prediction tasks, only considering the robustness criterion leads to the best performance. The robustness criterion improves the performances over a system trained with all the features by 23.9% (10dB), 29.7% (5dB), and 35.5% (0dB) in dominance prediction, and 43.2% (10dB), 51.3% (5dB), and 44.8% (0dB) in valence prediction. Although the performance criterion also shows significant improvement in the 5dB and 0dB conditions, it does not reach the performance of the robustness selection method. Our rank-based criteria show significantly better performance than random selection for most conditions. Simply using a small number of features is not the reason for the reported improvements. Instead, the robustness of the selected features leads to the reported improvements.

6. CONCLUSION

We proposed to select features based on the performance and robustness of individual LLDs in noisy conditions to improve the SER performance. Even though we did not apply compensation methods or leverage a noisy training set while training the model, our rank-based selection criteria yielded significant improvement over a system trained with all the features. We also determined that such improvement cannot be achieved by only randomly selecting a small number of features. We must select robust features that are resilient to noisy conditions.

Based on this study, we plan to selectively enhance the features that are highly influenced by background noises. We also plan to generalize this study by explicitly determining the features affected by types of noises (e.g. babble noise), including reverberation with different room impulse response. Our feature selection is very simple, but it does not consider interaction between LLDs. A more sophisticated feature selection may lead to further improvements.

7. REFERENCES

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 2871–2875.

[8] 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.

[9] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, May 2005.

[10] 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., pp. 143–171. Academic Press, London, UK, May 2015.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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