Separation of Emotional and Reconstruction Embeddings on Ladder Network to Improve Speech Emotion Recognition Robustness in Noisy Conditions

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

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

Abstract

When *speech emotion recognition* (SER) is applied in an actual application, the system should be able to cope with audio acquired in a noisy, unconstrained environment. Most studies on noise-robust SER require a parallel dataset with emotion labels, which is impractical to collect, or use speech with artificially added noise, which does not resemble practical conditions. This study builds upon the ladder network formulation, which can effectively compensate the environmental differences between a clean speech corpus and real-life recordings. This study proposes a decoupled ladder network, which increases the robustness of the SER system against the influences of non-stationary background noise by decoupling the last hidden layer embedding into emotion and reconstruction embeddings. This novel implementation allows the emotion embedding to focus exclusively on building a discriminative representation, without worrying about the reconstruction task. We introduce a noisy version of the MSP-Podcast database, which contains audio segments collected with a smartphone that simultaneously records sentences from the corpus and non-stationary noise at different *signal-to-noise ratios* (SNRs). We test the effectiveness of our proposed model with this corpus, showing that the decoupled ladder network can increase the performance of the regular ladder network when dealing with noisy recordings.

Index Terms: speech emotion recognition, domain adaptation

1. Introduction

Speech emotion recognition (SER) plays a crucial role in improving human-machine interaction. Recently, its performance has been highly improved with the advance of *deep neural network* (DNN) [1–5] and larger emotional speech datasets [6–9]. Despite the improvements, SER tasks still show poor performance in actual recording conditions, which are often collected with smartphones and wearable devices in unconstrained environments with non-stationary noises.

Several techniques may improve SER performance in noisy conditions by selecting features that are robust to background noise [10–12], transforming the noisy features into clean features [13–15], or augmenting the training data by contaminating clean speech with various types of noises [16, 17]. Another method is to apply domain adaptation techniques to the SER model to compensate for the environmental differences between clean speech corpora and the target recording conditions. Some studies have focused on minimizing the difference between the training and target data by using a small amount of labeled target data [18–22]. However, applying those methods to real-life applications requires knowledge of the target environment or emotional labels of audios collected from the target domain. Such limitations make it impractical in real-life applications.

Semi-supervised domain adaptation is a practical approach to apply to real-life applications, since it can leverage the audios from the target environment without requiring emotional labels. Previous studies [23–25] have demonstrated that ladder networks can be successfully applied in cross-corpus SER by only using unlabeled data from the target corpus. Ladder networks use an autoencoder as its backbone, adding lateral connections between intermediate layer representations in the encoder and decoder. Ladder networks can be trained to predict the emotional label of the input audio, while simultaneously reducing the train/test mismatch by containing the information of the target corpus during the reconstruction of the hidden layer representations, which can be achieved with unlabeled data.

Building upon those studies, we propose an improvement on the ladder network architecture for SER in noisy, unconstrained environments. In these conditions, ladder networks need to deal with the complex background noise in unlabeled audios, which disrupts emotion predictions. The key idea in our approach is that emotional features derived from noise signals should be decoupled from the features needed for the reconstruction task in the ladder network. Exploring this idea, we propose the *decoupled ladder network* (DLN), which can improve the feature representation to be more discriminative for the SER task. Our proposed approach separates the last hidden layer into two different embeddings; emotion embedding and reconstruction embedding. After the separation, the emotion embedding can exclusively focus on the SER task, without worrying about the reconstruction task in the ladder network. Under this implementation, the proposed DLN can eliminate any redundant information that is needed for the reconstruction of the embeddings, but disrupts the emotional predictions.

To evaluate our model, we introduce the noisy version of MSP-Podcast corpus, which simulates recordings from real-life environments. Instead of using manually synthesized noisy speech, we record audio from the corpus in a noisy environment using a smartphone. We played audio from the MSP-Podcast corpus and non-stationary background noise at different *signal-to-noise ratios* (SNRs). The smartphone collects both audios, creating recordings that simulate real-life environments. Our experiments with the noisy version of the MSP-Podcast corpus demonstrate that the DLN can enhance the prediction of arousal by 11.4%, 8.4%, 10.2%, and dominance by 17.1%, 13.2%, 7.0%, in the 10dB, 5dB, 0dB conditions, respectively.

2. Related work

2.1. Ladder Network

Ladder networks were first proposed as a semi-supervised feature representation learning framework [26]. The approach relies on a stacked denoising autoencoder [27] and *denoising*

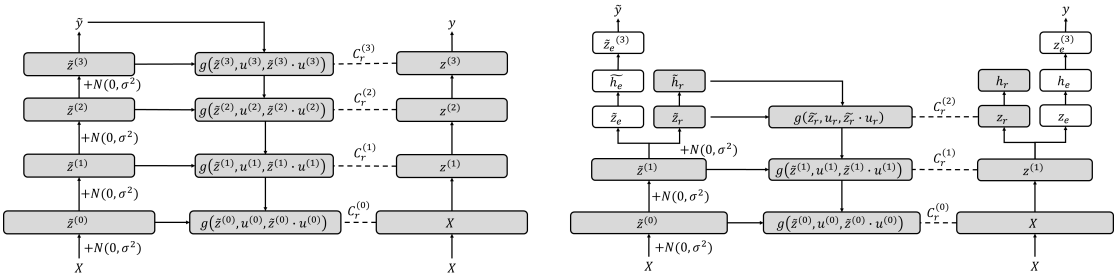


Figure 1: *Model architecture comparison between the ladder network (left) and the decoupled ladder network (DLN) (right). Gray areas indicate the features used in the reconstruction task.*

source separation (DSS) [28]. A ladder network consists of an encoder, decoder, and lateral connection between those two modules. The encoder is trained to predict the label with its noisy representation, and the decoder is trained to reconstruct the clean representation of each hidden layer in the encoder. This model has been further improved by applying batch normalization in each layer [29], and by investigating different types of denoising functions in the decoder [30].

2.2. Use of Ladder Network in SER

Ladder networks have been successfully applied in SER tasks. Huang et al. [23] combined ladder network and *support vector regression* (SVR) to predict categorical emotions. Parthasarathy and Busso [24] used ladder networks, as shown on the left side of Figure 1, to predict the level of arousal, dominance and valence. In their later study [25], they also showed that the ladder network can minimize the domain differences by using unlabeled speech data from the target corpus.

3. Decoupled Ladder Network

By using the ladder network, SER systems can be trained with a large amount of labeled emotional speech data and unlabeled audio samples from its target domain. However, in real-life applications, such unlabeled audio samples are likely to contain a variety of non-stationary background noise, which makes it difficult to reduce train and test mismatches. Even though the ladder network can achieve more relevant features for the prediction by using its lateral connection, due to the high complexity of real-life noise, it needs more constraints to minimize the influences of redundant information induced by noise in its emotion prediction. To explore this idea, we propose the *decoupled ladder network* (DLN) which separates the last hidden layer representation into an embedding needed for the prediction task and an embedding needed for the reconstruction task.

Figure 1 compares the architecture of the DLN with the architecture of the ladder network. In the ladder network, the output of the last hidden layer in the encoder, $\tilde{h}^{(L-1)}$, is directly fed into the output layer. However, in the DLN, $\tilde{h}^{(L-1)}$ is separated into two embeddings: emotion embedding, $\tilde{h}_e$, and reconstruction embedding, $\tilde{h}_r$. We arbitrarily split the embedding into two by assigning the first half of the nodes to $\tilde{z}_e$ and the second half of the nodes to $\tilde{z}_r$. Finally, bias correction and activation functions are applied to each representation, making $\tilde{h}_e$ and $\tilde{h}_r$, respectively. Those two embeddings play different roles in the model. To predict the emotional labels, the output layer only uses the emotion embedding, which is trained exclusively to create a discriminative representation. To reconstruct each clean hidden layer representation of the encoder, only the reconstruction embedding is fed into the highest layer of the decoder. Under this training method, the emotion embedding does not need to contain information needed for the reconstruction, making it easier to extract emotion-related information. Equation 1 shows the total cost function of the DLN C_{DLN}, which adds the prediction loss, C_p, and the summation of the reconstruction loss in layer l, C_r^l, across all layers.

$$C_{DLN} = C_p(y, \tilde{h}_e^{(L)}) + \sum_{l=0}^{L-2} \lambda^l \times C_r^l(\hat{z}_{BN}^{(l)}, z^{(l)}) + \lambda^{L-1} \times C_r^{L-1}(\hat{z}_{BN}^{(L-1)}, z_r), \quad (1)$$

where $\tilde{h}_e^{(L)}$ denotes the last output of the DLN, λ^l denotes a hyperparameter that weighs C_r^l, and $\hat{z}_{BN}^{(l)}$ is a reconstructed representation in the layer l, normalized by the mean and the standard deviation of the clean representation $z^{(l)}$. Note that the decoder does not reconstruct $\tilde{z}_e$ and $\tilde{z}_e^{(L)}$ to make them only related to the primary prediction task.

During training, we alternately present a mini-batch of labeled data and a mini-batch of unlabeled data, as described by Parthasarathy and Busso [25]. We train the model to maximize the *concordance correlation coefficient* (CCC) for the primary task by minimizing $C_p = 1 - CCC$. For C_r^l, we use the *mean square error* (MSE) between the reconstructed representation, $\hat{z}_{BN}^l$, and the clean representation, z^l. When the input is a mini-batch of unlabeled data, the model is only trained to minimize the reconstruction loss. Otherwise, the model is trained to minimize the prediction and reconstruction losses.

Table 1: *Settings for each recording condition and their estimated SNR level. (A) denotes the distance between the smartphone and the speech source. (B) denotes the distance between the smartphone and the noise source.*

Recording condition	**(A)** (inch)	**(B)** (inch)	**Estimated SNR** (dB)
10dB	5	35	11.06
5dB	10	30	4.34
0dB	15	25	0.15

4. Experiment settings

4.1. Noisy version of MSP-Podcast

This study uses the MSP-Podcast corpus (version 1.8) as the labeled speech corpus [7]. This corpus contains clean spontaneous emotional speech segments extracted from various podcast recordings. The recordings are selected when the predicted SNR is above 20dB, which are then formatted at a sampled rate of 16kHz. We use the retrieval approach proposed in Mariooryad et al. [31] to annotate only speech segments that we expect to be emotional as indicated by SER predictions. The annotation is conducted with a modified version of the crowdsourcing protocol proposed by Burmania et al. [32], where we track the performance of the workers in real time. This

study uses emotional attributes for valence (negative versus positive), arousal (calm versus active) and dominance (weak versus strong) collected with a Likert-scale from 1 to 7. Version 1.8 includes 44,879 samples for the train set, 7,800 for the development set and 15,326 for the test set.

A contribution of this study is a noisy version of the MSP-Podcast corpus to simulate real-life recording conditions. Although most of the studies on noise-robust SER have tested their method on artificially-generated noisy speech data by adding noise to clean speech signals, this approach is not enough to resemble the audio recorded from real-life environment. The noise is often fixed and repeated multiple times to match the duration of the clean audio, which is not realistic. Therefore, we directly use a smartphone that simultaneously recorded the clean speech and the noise signal. We use a 13ft $\times$ 13ft *American Speech-Language-Hearing Association* (ASHA) certified single-walled sound booth, which is shown in Figure 2. For the noise source, we used radio shows without copyright, which contains a variety of noise sounds, including human voice, musical sound, and sound effects. We changed the distance between the smartphone and the speech source (denoted as A in Table 1), and the distance between the smartphone and the noise source (denoted as B in Table 1) to change the SNR. For calibration, we recorded 1-minute recordings for the speech and noise. Table 1 shows the distances for A and B, and the estimated SNR for three recording conditions. We refer to these recording conditions as 10dB, 5dB, and 0dB according to the estimated SNR. Since the speech segments from the MSP-Podcast corpus are used in the recordings, we can transfer the same emotional labels for the noisy version of the database.

In our implementation, we train the models with clean data from the MSP-Podcast corpus (train set), evaluating the performance on its development set. We evaluate the results on the test set of the noisy version of the corpus. We create this train-test mismatch to evaluate the generalization of our approach to new conditions. The DLN requires unlabeled data from the target domain (e.g., noisy condition). We create this set by collecting noisy recordings of 43,361 speech segments that have not been annotated yet, matching the number of samples in the train set (10dB, 5dB, and 0dB). We refer to this set as the *unlabeled set*.

4.2. Acoustic Features

For the acoustic feature, we used the 6,373 dimension feature set introduced in the Interspeech 2013 *Computational Paralinguistics Challenge* (ComParE) [33]. All the features are normalized with the mean and standard deviation obtained from the training set. We calculate the mean and variance by considering data within the 0.5 to 99.5 percentiles of the training set to reduce influence of outliers. We clip the value of each feature if they exceed ± 10 after the normalization.

4.3. Settings for the Decoupled Ladder Network

Our proposed DLN has two hidden layers, where each layer has 256 nodes. We split the last hidden layer into a 128 dimension emotion embedding and a 128 dimension reconstruction embedding. The output layer has a single cell to predict the score of the emotional attribute. We used the *rectified linear unit* (ReLU) activation function in the hidden layer, and a linear activation function for the output layer. To regularize the model without disrupting the reconstruction task, we put 10% dropout between the input layer and the first hidden layer. We used a single layer *augmented multilayer perceptron* (AMLP) with 4 cells as a denoising function in the decoder [30].

To test the effectiveness of only using the reconstruction embedding in the reconstruction task, we compare the DLN with the model that has the same architecture of DLN but uses both emotion and reconstruction embeddings in the reconstruction task, which we denoted DLN+h_e. Unlike the DLN, the model is also trained to reconstruct the clean representation in the last hidden layer and the output layer. To reconstruct the output layer, the last output of the encoder is fed into the highest layer of the decoder. To reconstruct the last hidden layer, the decoder receives the concatenation of the emotion embedding and reconstruction embedding from the lateral connection.

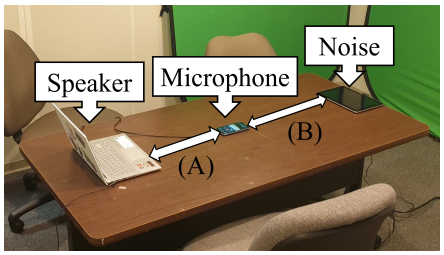


Figure 2: *Settings for collecting the noisy version of MSP-Podcast in the sound booth*

During training, we used the labeled set from the MSP-Podcast corpus, and the *unlabeled set* from the noisy version of the MSP-Podcast. We match the recording condition for the unlabeled training data and test data. For example, when testing the model in the10dB recording condition, we used the unlabeled training data from the 10dB recording condition. To select the model, we checked the performance on the development set of the MSP-Podcast corpus, and select the model that showed the best performance within 100 epochs. We use this model to report the performance in the noisy test set. To train the parameters, we use the Adam optimizer with a learning rate equal to 0.00005, and with a mini-batch of 256 sentences. We set the coefficient of the reconstruction loss in the ladder network to $\lambda^l = 1.0$ for all the layers.

4.4. Baseline models

We compare the proposed DLN with a dense network, and the regular *ladder network* (LN). Both models have the same number of hidden layers and cells as the DLN. We further regularize the dense network by adding a 50% dropout between the input layer and the first hidden layer and between the first hidden layer and the second hidden layer. All the hyperparameters for the training of baseline models are identical to those used in the DLN training. We train the dense network only with the labeled set of the MSP-Podcast corpus, since it is a supervised approach. The LN includes the reconstruction of the last hidden layer and the output layer in its training, as the DLN+h_e model, but without splitting the last layer into two embeddings.

5. Results

For the evaluation, we ran 20 experiments for each model with different initial weights, reporting the average scores. We conduct a two-tailed Welch's t-test between the dense network and the other models and between the LN and DLN to assess the effectiveness of our proposed semi-supervised training scheme in noisy conditions. We assert significance at p-value ≤ 0.05.

5.1. Emotional Prediction Performance

Table 2 shows the average CCC value over the 20 trials of the baseline models and our proposed DLN. When testing the model in noisy recording conditions, all the models show lower prediction performance than the performance in the clean

Table 2: *CCC values for each recording condition. Our baseline models, dense network and ladder network, are denoted as $Dense$ and LN, respectively. DLN denotes a decoupled ladder network, and $DLN + h_e$ denotes the model that has same architecture of the DLN, but use both embeddings in its reconstruction task. We highlight in bold the best performance per condition. The symbols † and * indicate that a given model shows significant improvement compared to the $Dense$ and LN models, respectively.*

	Arousal				Valence				Dominance			
	clean	10db	5db	0db	clean	10db	5db	0db	clean	10db	5db	0db
$Dense$	**0.631**	0.248	0.229	0.192	**0.296**	0.151	0.120	0.104	**0.562**	0.253	0.252	0.215
LN	0.627	0.438†	0.424†	0.364†	0.280	0.146	**0.129**	0.111	0.545	0.381†	0.385†	0.339†
DLN	0.625	**0.488**†*	**0.460**†*	**0.402**†*	0.283	**0.160***	0.126	**0.114**	**0.556***	**0.450**†*	**0.436**†*	**0.397**†*
$DLN + h_e$	0.611	0.432†	0.428†	0.372†	0.273	0.146	0.124	0.094	0.541	0.401†	0.407†	0.354†

recording condition. Also, as the level of noise increases, the performance decreases. These results confirm the detrimental effects that background noise has on emotion predictions. The ladder network showed much higher performance than dense network when testing in the noisy condition. The difference is much clearer in the prediction of arousal and dominance. It indicates that this semi-supervised learning approach can adapt a SER model to unconstrained noisy recording environments.

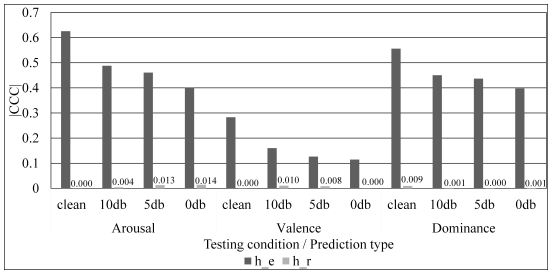
The performance is further improved when using the DLN model that only uses the reconstruction embedding for the reconstruction task. In the prediction of arousal, DLN increases the performance of the ladder network by 11.4%, 8.4%, and 10.2% for the 10dB, 5dB, and 0dB conditions, respectively. In the prediction of dominance, DLN increases the performance by 17.1%, 13.2%, and 7.0% for the 10dB, 5dB, and 0dB conditions, respectively. The model cannot achieve significant improvements for valence in the noisier conditions, leaving it as our future work to improve the decoupled ladder network for this attribute. Including the emotion embedding in the reconstruction task (i.e., $DLN + h_e$ condition) does not show significant improvement, indicating that separating the emotion embedding from the reconstruction task leads to the observed performance improvements.

5.2. Analysis on separating the embedding

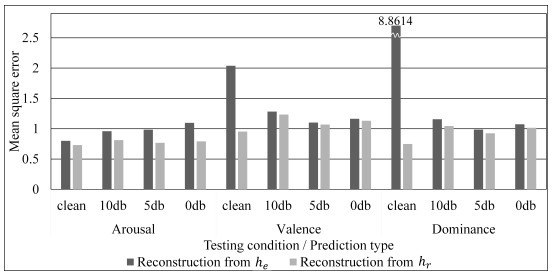
The goal of DLN is to decouple the last hidden layer in two parts, such that one is needed for the prediction and one is needed for the reconstruction. To check if our proposed method achieves this goal, we conduct two analyses on each embedding.

The first analysis evaluates if the reconstruction embedding has information for the emotion prediction task. We check the performance of using the reconstruction embedding as the input of the output layer, instead of using the emotion embedding. Figure 3a compares the prediction performance when using the reconstruction embedding for the SER task with the performance of the proposed DLN. The results demonstrate that the CCC values are nearly 0 when using the reconstruction embedding for the emotional prediction. This result indicates that the output layer cannot extract emotional information from the reconstruction embedding.

The second analysis evaluates if the emotion embedding has information needed to reconstruct the input feature. In this analysis, we put the emotion embedding into the decoder, instead of feeding the reconstruction embedding. Figure 3b shows the MSE of the input reconstruction when using the emotion embedding, compared to the ones using the reconstruction embedding. Our results show that the reconstruction losses using the emotion embedding are much higher than the losses using the reconstruction embedding. For valence, the difference in the reconstruction losses between embeddings is smaller than the differences for arousal and dominance. This result suggests that more improvements for valence can be achieved by imposing more strict constraints to decouple the two embeddings.



(a)



(b)

Figure 3: *Analyses on emotion and reconstruction embeddings. (a) Comparison of the absolute CCC values for using emotion embedding and reconstruction embedding in the DLN. (b) Mean square error of input reconstruction for using emotion embedding and reconstruction embedding in the DLN.*

6. Conclusions

This paper proposed the *decoupled ladder network* (DLN) to improve the robustness of a SER system in noisy environments. The approach decouples the emotion embedding from the reconstruction embedding in the last layer, so it can exclusively focus on improving the SER performance. To simulate more realistic recording conditions, we collected a noisy version of the MSP-Podcast corpus using a smartphone that simultaneously recorded the clean speech and non-stationary radio noise. Our experiments showed that the DLN can improve the performance of the ladder network in unconstrained recording conditions by decoupling the representation into emotion and reconstruction embeddings. We found that the DLN can separate the information for the prediction and reconstruction tasks. As future work, we will investigate stronger constraints on the DLN that can further reduce the dependencies between the emotional and reconstruction embeddings, which we expect to improve the model. We also plan to test our model using *multi-task learning* (MTL) framework, which has showed better performance than single-task learning for dense network [34], and ladder network [25].

7. Acknowledgement

Study supported by NIH under grant 1R01MH122367-01.

8. References

[1] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, Singapore, September 2014, pp. 223–227.

[2] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *IEEE International Conference on Multimedia and Expo (ICME 2017)*, Hong Kong, China, July 2017, pp. 583–588.

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

[4] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 815–826, April 2019.

[5] ——, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 951–955.

[6] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[7] 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.

[8] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 1823–1827.

[9] C. Busso *et al.*, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[10] 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.

[11] 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.

[12] 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.

[13] Y. Ghanbari and M. R. Karami-Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, vol. 48, no. 8, pp. 927–940, August 2006.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 511–516.

[19] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 257–261.

[20] Q. Mao, W. Xue, Q. Rao, F. Zhang, and Y. Zhan, "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 2608–2612.

[21] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[22] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, "Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition," *Speech Communication*, vol. 93, pp. 1–10, October 2017.

[23] 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)*, Beijing, China, May 2018, pp. 1–5.

[24] 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.

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

[26] 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.

[27] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, December 2010.

[28] J. Särelä and H. Valpola, "Denoising source separation," *Journal of Machine Learning Research*, vol. 6, pp. 233–272, December 2005.

[29] A. Rasmusi, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 3546–3554.

[30] M. Pezeshki, L. Fan, P. Brakel, A. Courville, and Y. Bengio, "Deconstructing the ladder network architecture," in *International Conference on Machine Learning (ICML 2016)*, New York, NY, USA, June 2016, pp. 2368–2376.

[31] 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.

[32] 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.

[33] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[34] 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.