INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Using Speech Enhancement Preprocessing for Speech Emotion Recognition in Realistic Noisy Conditions

Hengshun Zhou[1], *Jun Du*[1], *Yan-Hui Tu*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, GA. USA

zhhs@mail.ustc.edu.cn, jundu@ustc.edu.cn, tuyanhui@ustc.edu.cn, chl@ece.gatech.edu

Abstract

In this study, we investigate the effects of deep learning (DL)-based speech enhancement (SE) on speech emotion recognition (SER) in realistic environments. First, we use emotion speech data to train regression-based speech enhancement models which is shown to be beneficial to noisy speech emotion recognition. Next, to improve the model generalization capability of the regression model, an LSTM architecture with a design of hidden layers via simply densely-connected progressive learning, is adopted for the enhancement model. Finally, a post-processor utilizing an improved speech presence probability to estimate masks from the above proposed LSTM structure is shown to further improves recognition accuracies. Experiments results on the IEMOCAP and CHEAVD 2.0 corpora demonstrate that the proposed framework can yield consistent and significant improvements over the systems using unprocessed noisy speech.

Index Terms: speech emotion recognition, speech enhancement, realistic environments, multiple-target learning, LSTM

1. Introduction

Speech emotion recognition, as an important part of human-computer interaction, has been widely investigated [1, 2, 3]. However, the systems that are trained with clean speech often suffer from a huge performance degradation when tested in a noisy environment due to the mismatch between the train and test conditions [4, 5, 6]. Unfortunately, noise pollution is an indiscernible part of our daily life, caused often by various human activities and other background noise. Therefore, in real world applications, speech enhancement (SE) is a necessary module for emotion recognition system.

Despite recent advances in the field of speech emotion recognition [7, 8, 9], recognizing emotions in noisy environments remains an open research topic [10, 11, 12]. The primary concern of SE for emotion recognition is to remove noise effectively and preserve emotional information in noisy speech. Huang et al. [13] have studied the influence of white Gaussian noise on speaker's emotional states based on Gaussian mixture model (GMM), a typical emotion recognition system. By using algorithm based on spectral substraction and masking properties, they showed that the SE algorithms constantly improved the performance of emotion recognition system under various signal-to-noise ratios (SNRs). In [14], noise robust feature selection with k nearest neighbor (KNN) was found to be beneficial to emotion recogniztion in noisy speech. A front-end voice activity detector (VAD) based unsupervised method to select the frames with a relatively better SNR in the spoken utterances was proposed and shown to be effective in [15]. In [16], effects of different feature types and optimization techniques with different noises or microphone positions for automatic speech emotion recognition have been explored. Authors in [6] compared various front-end techniques for their efficacy in emotion recognition. In terms of the intelligibility of expressive speech in noise, researchers in [17] suggested that the intelligibility of emotion speech in noise was simply related to its audibility as conditioned by the effect that the expression of emotion has on its spectral profile. In [18], an interesting research investigated the performance of two enhancement methods in terms of perceptual quality as well as their impacts on emotion recognition. Furthermore, it demonstrated that quality measures can be an important indicator of enhancement model performance towards emotion recognition.

Although the aforementioned studies have shown the benefit of applying denoising algorithms to noisy speech, there are few studies on emotion recognition in realistic noisy environments. The most important reason may be that there are complex environmental noises and interferences to deal with. Researchers in [19] studied how a scalable deep learning (DL) architecture can be trained to enhance audio signals in a large number of unseen environments and benefitf common emotion recognition pipelines in terms of noise robustness. However the tested noisy data in [19] is still simulated.

In this paper, we investigated deep learning based speech enhancement framework for speech emotion recognition (SER). Specifically, the ideal ratio mask (IRM) estimated by the trained a long short-term memory (LSTM) model was first used for SE. We also find that the SE model trained with emotional corpus could achieve a higher accuracy for SER. To improve the model generalization capability of the regression model, an LSTM architecture with a design of hidden layers via simply densely-connected progressive learning, is adopted for the enhancement model. The proposed architecture further improves the performance of emotion recognition. Finally, considering the complexity of the realistic environment, the proposed improved speech presence probability (ISPP) based post-processing algorithm combined with deep learning by incorporating the estimated progressive ratio mask (PRM) obtained from the progressive learning structure further improves the noise robustness. Synthesized training data pairs generated from the WSJ0 [20] and IEMOCAP databases [21] were used to train SE models. Evaluated on the IEMOCAP and CHEAVD 2.0 databases [22], adopting emotional speech corpus (IEMOCAP) is crucial to SER performance rather than using non-emotional corpus (WSJ0) for both simulated and realistic noisy speech data. Moreover, the progressive learning network combined with ISPP post-processing can yield significant improves for SER on CHEAVD dataset recorded in realistic noisy conditions.

The remainder of this paper is organized as follows. Section 2 gives detailed description of our proposed approach. In Section 3, experimental results and analysis are discussed. Finally, we conclude in Section 4.

http://dx.doi.org/10.21437/Interspeech.2020-2472

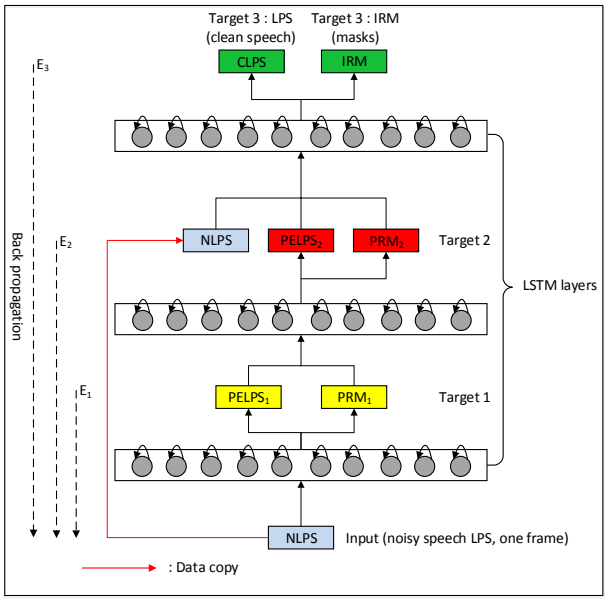


Figure 1: *Architecture of speech enhancement preprocessor.*

2. Speech Enhancement Preprocessing

In daily situations, the speech polluted by background noises may severely destroy the SER performance. Hence, a reliable SE system which can not only effectively suppress the background noise but also retain the emotional information is the key to improve the SER performance. In this study, three strategies used to improve SER performance are investigated.

2.1. Training data for speech enhancement

For automatic speech recognition (ASR), in order to ensure the effectiveness of system, a large number of data must be collected to cover all acoustic changes in speech recognition applications, such as speakers, background noises, different effects of microphone and communication channels, even different recognition tasks, etc. Training data is also important for speech emotion recognition. In [23], generalization involving the target persons speech samples and prior knowledge about their emotional content are investigated. In [24], the authors proposed an adversarial learning framework to alleviate the culture influence on emotion recognition. The effect of gender bias in speech emotion recognition has also been studied in [25]. The preprocessor for speech emotion recognition, which is different from the traditional SE, needs to remove the noise without destroying the emotion clues of speech as much as possible. Inspired by this, we found that when the SE system trained with more matching data (corpus with emotional speech), it is helpful to improve the performance of SER from noisy speech compared to using non-emotional speech corpus. It is verified for the settings of both simulated and realistic noisy data.

2.2. A novel progressive multi-target architecture

Deep neural networks (DNNs) and recurrent neural networks (RNNs) have been widely used in speech enhancement for a long time [26]. However, the conventional RNN can not hold information for a long period and the optimization of RNN parameters via the back propagation through time (BPTT) faces the problem of the vanishing and exploding gradients [27]. The problems can be well alleviated by the invention of LSTM which introduces the concepts of memory cell and a series of gates to dynamically control the information flow. As shown in Figure 1, all LSTM layers consist of memory cells.

The LSTM-based densely connected progressive learning was proposed by [28] and proved to be effective for speech enhancement. To improve the generalization capability of LSTM architecture, a design of hidden layers via densely connected progressive learning and output layer via multiple-target learning is presented (denoted as LSTM-PL-MTL), as illustrated in Figure 1. The log-power spectra (LPS) features are adopted for network inputs and outputs. The input is noisy LPS (NLPS) features and final output is clean LPS (CLPS) and IRM.

A series of progressive ratio masks (PRMs) are concatenated with the progressively enhanced LPS (PELPS) features together as the learning targets. PRM, to perform a trade-off between noise reduction and speech distortion, is defined as:

$$M_{\mathrm{PRM}}(t,f) = \frac{S(t,f) + N_{\mathrm{T}}(t,f)}{S(t,f) + N_{\mathrm{I}}(t,f)} \tag{1}$$

where $S(t,f)$ represents the power spectrum of the speech signal at the time-frequency (T-F) unit (t,f), $N_{\mathrm{T}}(t,f)$ and $N_{\mathrm{I}}(t,f)$ represent the power spectrum of the noise signals in one PRM target and input noise signals at the T-F unit (t,f) respectively. All the target layers are designed to learn intermediate speech with higher SNRs or clean speech. The multi-task error between the output of target layer k and its ground-truth label is

$$\begin{aligned} E_{\mathrm{MTL}}(k) = \sum_{t,f} [& (\hat{x}_{\mathrm{PELPS}}(k,t,f) - x_{\mathrm{PELPS}}(k,t,f))^2 \\ & + (\hat{M}_{\mathrm{PRM}}(k,t,f) - M_{\mathrm{PRM}}(k,t,f))^2] \end{aligned} \tag{2}$$

where $\hat{x}_{\mathrm{PELPS}}(k,t,f)$ and $x_{\mathrm{PELPS}}(k,t,f)$ are predicted and ground-truth PELPS features of the k^{th} target layer, while $\hat{M}_{\mathrm{PRM}}(k,t,f)$ and $M_{\mathrm{PRM}}(k,t,f)$ are predicted and ground-truth PRM features of the k^{th} target layer. Both $\hat{x}_{\mathrm{PELPS}}(k,t,f)$ and $\hat{M}_{\mathrm{PRM}}(k,t,f)$ are nonlinear functions of PELPS and PRM in preceding target layers. $x_{\mathrm{PELPS}}(k,t,f)$ and $M_{\mathrm{PRM}}(k,t,f)$ can be easily calculated with a predefined SNR gain of target layer k. Please note that PELPS and PRM of target layer 3 correspond to clean LPS (CLPS) features and IRM, respectively. The errors of all target layers are computed in the mean squared error (MSE) sense, and added together to optimize the trainable parameters. In our LSTM-PL-MTL, the dimension of both LPS and PRM (IRM) feature vectors is 257, single frame is used for input, the number of LSTM memory cells in each layer is 1024, and we use 3 target layers (with 10dB SNR gain for both target layer 1 and 2).

2.3. Speech post-processing with ISPP

One advantage of the method based on progressive learning is that there are multiple estimated targets that can be obtained from the network. The different targets can provide rich information for post-processing. Meanwhile, a post-processing approach, the improved speech presence probability (ISPP) combining conventional and deep learning techniques [29, 30] by incorporating the estimated PRM obtained from the proposed structure was employed. By incorporating neural network based mask estimation $\hat{M}_{\mathrm{PRM}}(t,f)$ to define an intermediate item

$$\hat{G}(t,f) = \delta \hat{M}_{\mathrm{PRM}}(t,f) + (1-\delta) G_{\mathrm{ISPP}}(t,f) \tag{3}$$

where $G_{\text{ISPP}}(t, f)$ denotes ISPP-based gain function at T-F unit (t, f) and δ is a weighting factor empirically set to 0.5 in our experiments, see [29] for details.

3. Experiments and Result Analysis

3.1. Databases

7138 utterances of WSJ0 corpus [20] (about 12 hours of reading style speech) from 83 speakers were used to train LSTM-IRM model, denoted as SI-84 training set.

Interactive emotional dyadic motion capture database (IEMOCAP) corpus [21], one of the widely used standard emotional databases on speech emotion recognition, comprises five sessions, each of which includes labeled emotional speech utterances from recordings of dialogs between two actors. There is no actor overlapping between these sessions.

Chinese natural audio-visual emotion database (CHEAVD) 2.0 [22] was collected by capturing clips from films and TV programs and used for multimodal emotion challenge (MEC) 2017 [31]. These clips are not captured in the controlled studio environment, so there might contain background noises, which are very close to real world scenarios. Each speech utterance has one label among eight emotion categories. The SNR distribution was investigated in [32].

3.2. Implementation details

To train SE models, WSJ0 corpus and IEMOCAP that do not include the same speaker in SER system (about 9 hours) are corrupted with CHiME-4 noise at four SNR levels (-5dB, 0dB, 5dB and 10dB) to build a 36-hour training set respectively, consisting of pairs of clean and noisy utterances. For SER system, we conducted experiments on IEMOCAP in mismatched scenarios, i.e clean-training and noisy-testing. We randomly picked out a session (session 3 was used here) and only added noise to the test set, see [33] for details. Four noise types (BUS, CAF, PED and STR) [34] in CHiME-4 challenge were selected as the noise database for simulation. We investigated the performance of our algorithm at SNR levels ranging from -5dB to 15dB, with an interval of 5dB and used the speech utterances from four emotion categories, i.e., happy, sad, angry and neutral.

MEC 2017 is a more challenging task and the labels of the test set are not available. We randomly selected 700 utterances from the training set with a total of 4917 utterances as the validation set and the rest as the new training set, and the validation set of the competition as the new test set for experiments. Attention based fully convolution network [33] is used as SER system for both IEMOCAP and CHEAVD tasks. Please note that the test set are recorded in realistic noisy conditions. Therefore, there are high mismatches between SE model and SER system, such as speaking style and types of background noise. These mismatches make SER system quite challenging for our proposed enhancement approach.

For front-end configurations, we used pytorch to train the SE network. Each stage consists of 6 epochs and 5 stages are used. The learning rate for the first stage was initialized as 0.25 and then decreased by 20% after each stage. The batch size is 8. For the back-end configurations, the SER systems were trained on TensorFlow, referring to [33] for specific parameter.

3.3. Results on simulated test data using IEMOCAP

Deep learning-based IRM estimation was first used for SE. Under clean conditions, we trained the SER system and achieved an accuracy of 71.90% on the test set. Our results are presented in Table 1. 'Noisy' denotes unprocessed noisy speech. 'IRM-WSJ0' and 'IRM-IEM-2' represent LSTM-IRM model trained by WSJ0 and IEMOCAP respectively, where '2' means the number of hidden layers used in LSTM is 2.

Table 1: *The accuracy (%) comparison of using IRM estimation with different hidden layers (with the corresponding 71.90% for clean speech).*

Enhancement	-5dB	0dB	5dB	10dB	15dB
Noisy	48.54	50.00	52.19	55.47	59.12
IRM-WSJ0	47.81	50.37	56.20	60.95	62.04
IRM-IEM-2	52.92	56.57	61.68	65.33	66.79
IRM-IEM-3	47.81	52.19	59.85	63.87	65.33
IRM-IEM-4	46.35	48.91	57.66	63.50	64.60

Our first observation is that the accuracy decreases to a certain extent as CHiME-4 noise is added to the test set in IEMOCAP, 48.54% at -5dB and 59.12% at 15dB. By using the LSTM-IRM enhancement model trained on WSJ0, the SER systems achieve better performance when using enhanced audio compared to using noisy audio in most cases. The only exception is when the test speech is under -5dB. When the enhanced model trained on the data set of IEMOCAP using the same network structure, the performance of SER has a comprehensive improvement. When the number of hidden layers in LSTM is increased, the performance of the SER system decreases. The reason might be that the deeper structures with limited training data lead to the overfitting problem and emotional information is destroyed, which is also the difficulty of SE for SER.

Table 2: *The accuracy (%) comparison of different targets by using LSTM-PL-MTL (with 71.90% for clean speech).*

Enhancement	-5dB	0dB	5dB	10dB	15dB
Noisy	48.54	50.00	52.19	55.47	59.12
PL-MTL [35]	51.46	56.20	59.85	63.50	67.15
T1-LPS	49.27	54.02	60.58	63.87	67.15
T1-PRM	47.08	48.18	54.02	63.50	64.60
T2-LPS	38.32	47.08	54.02	56.20	57.66
T2-PRM	47.81	48.91	56.57	63.87	66.06
T3-LPS	40.88	48.91	50.73	52.92	55.11
T3-IRM	54.38	57.30	62.77	67.15	68.25

To improve the SE model generalization for SER system, we further investigated the SE model structure based on progressive learning which has been successfully applied to speaker diarization in quite challenging realistic environments [36]. As shown in Table 2, we used the structure in [35]. Interestingly, we find that the best performance can be obtained when decoding with Target 3 IRM. However, the original LSTM-PL-MTL model in [35] underperforms LSTM-IRM model (IRM-IEM-2) in Table 1 for most SNR cases. This might be explained as that the dense connections in LSTM-PL-MTL result in very high dimensional intermediate target layers and the overfitting problem under the setting of limited training data.

Nevertheless, with our simplified architecture as shown in Figure 1, almost all dense connections in original LSTM-PL-MTL model [35] are removed with only one connection from the input layer to the final intermediate target layer. From Ta-

ble 2, we can observe the results with the Target 3 IRM (T3-IRM) in our proposed LSTM-PL-MTL perform better than the system in [35] and IRM-IEM-2, across all SNR levels. Among all the learning targets (T1-LPS, T1-PRM, T2-LPS, T2-PRM, T3-LPS, T3-IRM) in our LSTM-PL-MTL model, T3-IRM achieved the best results. The new architecture helps in all cases and the performance gaps between the highest SNRs (10dB and 15dB) and clean condition are small. In the case of low SNR, there will be more residual noises after enhancement, leading to the poor performance of SER. We compared the enhanced speech spectrograms and observed that more distortions destroying the emotion information appeared at low SNRs. These could explain why performance is still far from clean audio even after being enhanced.

3.4. Results on realistic test using CHEAVD

To verify the effectiveness of our proposed SE approach in more realistic conditions, we conducted the experiments on CHEAVD dataset recorded in realistic noisy conditions, detailed results are presented in Table 3 and 4.

Table 3: *The accuracy (%) comparison of using different speech enhancement methods in real situations.*

Noisy	1000h SE [35]	IRM-WSJ0	IRM-IEM
41.58	41.30	40.88	**42.01**

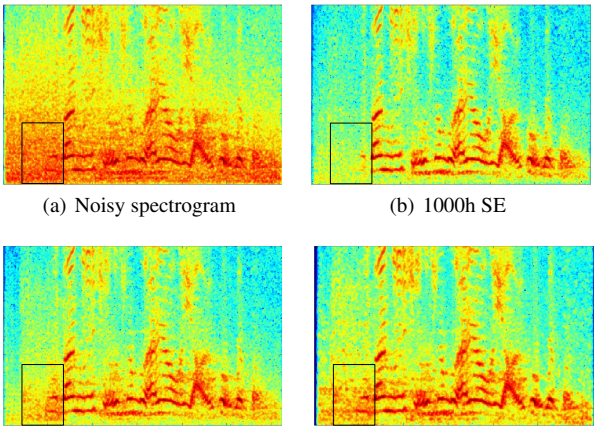
In Table 3, a general SE model trained in corpus of about 1000 hours was first used for comparison [35], and it was found that the performance of SER decreased slightly. Second, IRM-WSJ0 model still degraded the SER performance. The reason may be the high mismatch of speech styles (emotional vs. non-emotional). When IRM-IEM was used, the performance of SER not only exceeded that of 1000h enhanced model, but also exceeds that of unprocessed speech. This is consistent with our observation in the simulation data set.

Table 4: *The accuracy (%) comparison of using ISPP post-processing. "Fusion" means that the score fusion of SER systems with enhanced speech obtained from T1-PRM and corresponding ISPP post-processing.*

T1-PRM-ISPP	T2-PRM-ISPP	T3-IRM-ISPP	Fusion
43.00	42.29	42.72	**44.13**

Considering no significant performance improvements in Table 3, we add to two LSTM layers for each target learning in LSTM-PL-MTL and use the ISPP post-processing in Section 2.3. The results are shown in Table 4 and remarkable improvements of SER performance could be achieved by using the proposed LSTM-PL-MTL structure and post-processing. When using post-processing based on T1-PRM, 43.00% accuracy can be obtained. By score fusion of SER systems with enhanced speech from T1-PRM and its post-processing, the best accuracy of 44.13% is achieved, which yields an absolute 2.55% improvement over the unprocessed noisy speech. This also shows the effectiveness of proposed method.

Finally, to illustrate why the proposed speech preprocessing can help emotion recognition. Figure 2 gives an utterance example from the real test set of CHEAVD 2.0. Figure 2(a) plots the spectrogram of the unprocessed noisy utterance. The girl's



(a) Noisy spectrogram (b) 1000h SE

(c) LSTM-IRM (d) LSTM-PL-MTL + ISPP

Figure 2: *The comparison of SE results of different approaches for an utterance from the real test set of CHEAVD 2.0.*

sad voice was concealed to some extent by the environmental noise, and wrongly classified as "neutral". By using the trained enhancement model, a lot of background noise was removed, and it was also correctly classified as "sad". But it also brings some non-linear distortions to speech, as shown in the spectrogram in the black rectangle of Figure 2(b). A listening inspection on enhanced speech showed that for SE model trained with non-emotional corpus, in addition to removing the noise, sounds like slight "ha ha" and sighs were also destroyed or removed to some extent. Considering that emotion recognition is sensitive to these changes resulting in performance degradations, training SE models using emotional speech data can help recovering these key speech emotion cues as shown in Figure 2(c) and Figure 2(d). Moreover, our proposed LSTM-PL-MTL with T1-PRM and ISPP post-processing made the better trade-off between noise reduction and speech emotion preservation over LSTM-IRM method.

4. Conclusion

In this paper, we study the effects of speech enhancement as a preprocessor on speech emotion recognition in challenging noisy environments. We first find that speech enhancement models trained with emotion speech is more effective than non-emotion speech. We also observe that important cues, such as low-energy signs and laughters, are often masked by noises and distorted by some enhancement models. We propose training SE models with emotion speech corpora to achieve a higher accuracy for speech emotion recognition. We also present a novel LSTM-PL-MTL architecture with ISPP-based post-processing that proves to be effective in enhancing speech for emotion recognition, achieving a considerable performance improvement over unprocessed noisy speech.

5. Acknowledgement

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Huawei Noah's Ark Lab.

6. References

[1] X. Wu, S. Liu, Y. Cao, X. Li, and H. M. Meng, "Speech emotion recognition using capsule networks," *IEEE ICASSP2019*, 2019.

[2] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," *Interspeech 2019*, 2019.

[3] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," *Interspeech 2018*, Sep 2018. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1883

[4] F. Chenchah and Z. Lachiri, "Speech emotion recognition in noisy environment," *International Conference on Advanced Technologies for Signal & Image Processing*, 2016.

[5] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Robust front-end processing for emotion recognition in noisy speech," *International Symposium on Chinese Spoken Language Processing*, 2018.

[6] R. Chakraborty, A. Panda, M. Pandharipande, S. Joshi, and S. K. Kopparapu, "Front-end feature compensation and denoising for noise robust speech emotion recognition," *Interspeech*, 2019.

[7] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," *IEEE ICASSP 2019*, 2019.

[8] G. Paraskevopoulos, E. Tzinis, N. Ellinas, T. Giannakopoulos, and A. Potamianos, "Unsupervised low-rank representations for speech emotion recognition," *Interspeech 2019*, 2019.

[9] B. Wang, M. Liakata, H. Ni, T. Lyons, and K. Saunders, "A path signature approach for speech emotion recognition," *Interspeech 2019*, 2019.

[10] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–16, 2011.

[11] R. Xia and Y. Liu, "Using denoising autoencoder for emotion recognition," *INTERSPEECH*, 2013.

[12] E. Parada-Cabaleiro, A. Baird, A. Batliner, N. Cummins, and B. Schuller, "The perception of emotions in noisified nonsense speech," *Interspeech 2017*, 2017.

[13] 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.

[14] T. L. Pao, W. Y. Liao, Y. T. Chen, J. H. Yeh, and C. S. Chien, "Comparison of several classifiers for emotion recognition from noisy mandarin speech," *International Conference on Intelligent Information Hiding & Multimedia Signal Processing*, 2007.

[15] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "An unsupervised frame selection technique for robust emotion recognition in noisy speech," *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[16] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," *IEEE International Conference on Acoustics*, 2007.

[17] C. Davis, C. S. Chong, and J. Kim, "The effect of spectral profile on the intelligibility of emotional speech in noise," *Interspeech*, 2017.

[18] A. R. Avila, M. J. Alam, D. D. O'Shaughnessy, and T. H. Falk, "Investigating speech enhancement and perceptual quality for speech emotion recognition," *INTERSPEECH*, 2018.

[19] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," *Interspeech 2019*, 2019.

[20] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[21] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources & Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[22] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: a chinese natural emotional audiovisual database," vol. 8, no. 6, pp. 913–924, 2017.

[23] L. Chao, J. Tao, M. Yang, and Y. Li, "Improving generation performance of speech emotion recognition by denoising autoencoders," *The 9th International Symposium on Chinese Spoken Language Processing*, pp. 341–344, 2014.

[24] J. Liang, S. Chen, J. Zhao, Q. Jin, H. Liu, and L. Lu, "Cross-culture multimodal emotion recognition with adversarial learning," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4000–4004, 2019.

[25] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," *Interspeech*, 2019.

[26] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[27] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, 2013.

[28] Y.-H. Tu, J. Du, T. Gao, and C.-H. Lee, "A multi-target snr-progressive learning approach to regression based speech enhancement," *Audio, Speech, and Language Processing (accepted), IEEE/ACM Transactions on*, 2020.

[29] Y. Tu, I. Tashev, S. Zarar, and C. Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2531–2535, 2018.

[30] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech enhancement based on teacher-student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. P-P, no. 99, pp. 1–1, 2019.

[31] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, "Mec 2017: Multimodal emotion recognition challenge," *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1–5, 2018.

[32] F. Tao, G. Liu, and Q. Zhao, "An ensemble framework of voice-based emotion recognition system for films and TV programs," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pp. 6209–6213, 2018. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461617

[33] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2018, Honolulu, HI, USA, November 12-15, 2018*, pp. 1771–1775, 2018. [Online]. Available: https://doi.org/10.23919/APSIPA.2018.8659587

[34] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[35] L. Sun, J. Du, X. Zhang, T. Gao, X. Fang, and C. Lee, "Progressive multi-target network based speech enhancement with snr-preselection for robust speaker diarization," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7099–7103, 2020.

[36] L. Sun, J. Du, T. Gao, Y. Lu, Y. Tsao, C. Lee, and N. Ryant, "A novel lstm-based speech preprocessor for speaker diarization in realistic mismatch conditions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5234–5238, 2018.