Enhancing Automatic Speech Recognition Quality with a Second-Stage Speech Enhancement Generative Adversarial Network

Soha A. Nossier
Dept. Computer Science and Digital Technologies
University of East London
London, UK
soha.abdallah.nossier@gmail.com

Julie Wall
Dept. Computer Science and Digital Technologies
University of East London
London, UK
j.wall@uel.ac.uk

Mansour Moniri
Dept. Computer Science and Digital Technologies
University of East London
London, UK
m.moniri@uel.ac.uk

Cornelius Glackin
Intelligent Voice Ltd
London, UK
neil.glackin@intelligentvoice.com

Nigel Cannings
Intelligent Voice Ltd
London, UK
nigel.cannings@intelligentvoice.com

***Abstract*—Speech enhancement is an essential preprocessing stage for automatic speech recognition in noisy conditions; however, the distortion caused by the denoising process may lead to degradation in automatic speech recognition performance. This paper presents a deep learning-based speech enhancement architecture to overcome this issue by applying a second-stage network that deals with distortion noise. Moreover, a signal-to-noise ratio binary classifier is implemented to activate the speech enhancement network for intrusive noise environments only, which improves the overall performance. The proposed architecture outperforms powerful models in the literature, as it improves a challenging noisy speech test set by 0.8 and 5.9% improvement in the quality and intelligibility scores, respectively. Furthermore, the architecture improves the performance of automatic speech recognition with a 13.8% reduction in the word error rate at 0 dB signal-to-noise ratio. Finally, the second-stage network was proven to improve the performance of first-stage speech enhancement models, not previously seen in the training process.**

***Index Terms*—Automatic speech recognition, deep learning, generative adversarial network, speech distortion, speech enhancement**

I. Introduction

Speech enhancement is the process of improving speech quality and intelligibility by mitigating background noise [1]. Automatic Speech Recognition (ASR) is one of the important applications for speech enhancement. In noisy environments, a frontend speech enhancement network is needed to process the noisy speech signal before performing ASR, to improve performance [2], [3]. Supervised deep learning-based speech enhancement models were proven to be very effective in removing background noise and enhancing speech quality and intelligibility [4]–[6]. However, these models do not always generate a satisfying performance when applied as a separate preprocessing stage to ASR, especially for mismatched noise conditions [7]; conversely, it was shown to degrade the performance in some cases [8], [9].

Much research has been done to understand why speech enhancement models do not improve ASR performance, and the outcome of this research points to the speech distortion issue, a key drawback of the denoising process [8], [10]. While the Deep Neural Network (DNN) tries to eliminate background noise, it removes part of the target speech signal. This introduces a new kind of noise, more specifically distortion noise, which leads to a change in the processed speech characteristics, and makes it not understandable by the ASR model, leading to higher Word Error Rates (WERs) [9].

Consequently, more recently, DNNs for speech enhancement have been designed to minimize distortion, by manipulating the loss function [4] or developing a two-stage speech enhancement network [5], [11], [12]. An alternative solution is to implement a feedback system to the speech enhancement network during training that takes a signal from the ASR model to ensure that it can successfully produce a transcript of the speech processed by the DNN [13].

Another approach that solves this distortion issue is joint training of the speech enhancement network and ASR model, where the ASR model is retrained using the processed speech from the speech enhancement network, to avoid this mismatch problem [9], [14]. However, there are disadvantages to this solution. First, retraining of a running ASR system is required, which is not practical. Moreover, whenever modifications are needed to the speech enhancement network, further retraining will be needed for the ASR model. Second, speech enhancement will be performed in any environmental condition, even for clean speech, which will increase system complexity and

This research is sponsored by University of East London and Intelligent Voice Ltd.

processing time without any gain in ASR performance.

Although the above-discussed solutions help in minimizing the negative effect of speech enhancement processing, further improvement to the WER of ASR can be achieved by adding a switch, to decide whether to perform speech enhancement or not. The decision of this switch can be based on measuring the distortion of the enhanced speech signal generated by the speech enhancement network using Signal-to-Distortion (SDR) ratio, as presented in this work [15]. Alternatively, a deep neural network (DNN) can be trained to measure the distortion added by the speech enhancement processing, and then decide whether to perform SE or not, based on predicting the improvement/deterioration of the WER of the ASR system under testing [16]. However, these solutions are based on making the decision based on the enhanced speech signal, which means speech enhancement processing is always required, and this increases processing time.

In this paper, we aim to contribute to the above research by presenting a deep learning-based speech enhancement architecture for ASR that minimizes the speech distortion caused by the denoising process. The architecture consists of a Signal-to-Noise Ratio (SNR) classifier, a deep Convolutional Denoising Autoencoder (CDAE) network, and a Least Square Generative Adversarial Network (LSGAN) [17]. The classifier performs binary classification to differentiate between high and low SNR speech. The output signal from the classifier activates the speech enhancement network at low SNRs only, when speech enhancement is essential for ASR. This will avoid the speech distortion caused by speech enhancement processing at high SNRs and also for clean speech; furthermore, it will decrease processing times. At low SNRs, the first stage CDAE-based network performs speech enhancement in the frequency domain to eliminate background noise. Afterwards, the second stage LSGAN acts as a matching network that performs further speech denoising and reconstruction to minimize the mismatch between the processed speech from the first stage speech enhancement network and the input to the ASR model.

The contributions of this paper can be summarized as follows:

- proposing a deep learning-based speech enhancement architecture that minimizes distortion and improves ASR performance, and
- providing a standalone second-stage LSGAN model that acts as a matching network between the speech enhancement model and the ASR model to improve the overall performance.

The rest of this paper is organized as follows. Section II describes the problem under investigation. Section III illustrates the developed architecture. Section IV demonstrates the experimental setup. Results and discussion are presented in Section V. Finally, the paper's conclusions are given in Section VI.

II. Problem Formulation

At low SNRs, where speech enhancement provides crucial preprocessing for ASR, the speech signal, *s*, is affected by noise environment, *n*, and the time domain noisy speech signal,*y*, can be represented as in Eq. 1:

$$y(k) = s(k) + n(k), \tag{1}$$

where *k* is the time index. When processing this noisy signal using a deep learning approach, the DNN performs some nonlinear operations on the input noisy speech to minimize a loss function that aims to generate an estimate of the clean speech signal, $\hat{s}$. As proved in [9], although this estimated clean speech has a higher SNR compared to the noisy speech, it suffers from a new kind of noise originating from the distortion which occurred during the denoising process. As a result, $\hat{s}$ can be expressed as in Eq. 2:

$$\hat{s}(k) = s(k) + \alpha n(k) + n_d(k), \tag{2}$$

where α is a scaling factor that describes the decrease in the noise intensity, and n_d is the added distortion noise. Considering the case that the DNN managed to effectively remove background noise, the speech quality will be improved but the added distortion noise will be significant and results in performance degradation for the backend ASR. In this case, n_d is greater than αn. We hypothesise that when adding a second stage DNN to process $\hat{s}$, the loss function will focus on minimizing this dominant distortion noise, which will ultimately decrease the mismatch issue between the speech enhancement network and the ASR. To achieve this, we used an LSGAN model with a discriminator that learns to differentiate between distorted and clean speech. The loss function of the discriminator (*D*) and the generator (*G*) of this second stage LSGAN can be expressed as in Eqs. 3 and 4, respectively:

$$min_D L_{LSGAN}(D) = \frac{1}{2} E_{s\sim P_{data}(s)}[(D(s,y)-b)^2] + \frac{1}{2} E_{\hat{s}\sim P_{\hat{s}}(\hat{s})}[(D(G(\hat{s},y),y)-a)^2], \tag{3}$$

$$min_G L_{LSGAN}(G) = \frac{1}{2} E_{\hat{s}\sim P_{\hat{s}}(\hat{s})}[(D(G(\hat{s},y),y)-b)^2, \tag{4}$$

where *b* is an all-one vector representing the label for real clean speech, while *a* is an all-zero vector that represents the label for estimated clean speech. *D(s,y)* is the output of the discriminator with concatenated real clean speech and noisy speech as an input, and *D(G(ŝ,y),y)* is the output of the discriminator with concatenated noisy speech and the second stage estimated clean speech from the generator network as an input. The noisy speech is fed to both the generator and the discriminator, as it was found that this improves the learning process because when the noisy signal is seen as a different signal from the clean speech, noise reconstruction will be avoided during the training process.

At high SNRs, the effect of speech enhancement processing on ASR is not very significant, because most ASR systems are trained with some noisy speech, so the ASR can deal with non-intrusive noise environments. Moreover, the distortion issue, as discussed above, may outweigh the denoising benefits and have a negative impact on the quality of clean or high

SNR speech, leading to worsened ASR performance. For this reason, we suggest not performing speech enhancement at high SNR conditions.

The decision of performing speech enhancement or not is made in our implementation using an extra SNR binary classifier that processes the noisy speech, and outputs 1 if low SNR is detected, activating the speech enhancement network. This classification is performed based on the average of five audio features that are concatenated together and fed to the classifier network to make the decision. The used input feature vector to the classifier, C_i, can be represented as in Eq. 5:

$$C_i = \overline{y}_{MFCC} \oplus \overline{y}_{Mel} \oplus \overline{y}_{SC} \oplus \overline{y}_{Chroma} \oplus \overline{y}_{T}, \quad (5)$$

where y_{Mel} is the Mel-Spectrogram, y_{MFCC} is the Mel-Frequency Cepstral Coefficients (MFCCs), Y_{SC} is the Spectral Contrast, Y_{Chroma} is the Chromagram, and Y_T is the Tonnetz [18]. The threshold of the classifier decision boundary should be chosen based on the performance of the backend ASR in noisy conditions. In our implementation, we found that the performance of the ASR, used in testing, without speech enhancement is better at SNR values higher than 15 dB, so the classifier was designed to activate the speech enhancement network for noisy speech with SNR 15 dB or less.

III. Architecture Description

The diagram of the fully developed architecture is shown in Fig. 1. The SNR classifier is a one-dimensional (1D) convolution-based network of three convolution layers with Parametric Rectified Linear Unit (PReLU) activations. A dropout layer of 0.2% rate was used to avoid overfitting, and two dense layers were added: one with ReLU activation and the other with Sigmoid activation for prediction.

For the first stage speech enhancement model, we used the single-stage Deep-Encoder Convolutional Autoencoder Denoiser (DE-CADE) network proposed in our previous work [5]. The network consists of several strided-dilated convolution layers in the encoder, and deconvolution and upsampling layers in the decoder. Further details about network hyper-parameters are described in Fig. 1. The network operates in the frequency domain using Short-Time Fourier Transform (STFT) input features of 256 hamming window size with 50% overlap, to estimate the clean magnitude spectrogram. In the second stage speech enhancement network, the LSGAN generator is another DE-CADE but processing is in the time domain using time frames of size 2,048 and 50% overlap as input. The discriminator has nine 1D convolution layers with PReLU activations, and batch normalization was applied after each convolution layer, to ensure training stability. A dropout of 0.2% rate was applied after every three convolution layers. Another 1D convolution layer was used with linear activation before the final two dense layers used for prediction. The discriminator classification is based on the MFCCs input features, because MFCCs are the main feature used in the ASR model, so this will ensure that the processed speech from the second stage keeps the most important features to be correctly interpreted by the ASR model.

IV. Experimental Setup

The training of the architecture is based on the deep noise suppression challenge dataset [19], which has speech data of more than 500 hours and 181 hours of noise data. The speech and noise data were divided into 90% for training and 10% for validation, and then the speech and noise environments were additively mixed at a wide range of SNRs from 0 to 20 dB in steps of 1 dB. For the SNR classifier, the noisy speech of SNR value 15 dB or less is labelled as low SNR speech (binary 1); while 20 dB SNR noisy speech and clean speech data are labelled as high SNR speech (binary 0).

In testing, we used 224 speech audio files for 56 speakers and 224 different speech utterances that were randomly selected from the Voice Bank Corpus dataset [20]. These speech audio files were corrupted with 10 noise environments, taken from the 100 Nonspeech Environmental Sounds dataset [21]. The selected noise environments are a mix of human-generated noise, such as crying and yawning sounds, and other non human-generated noise, such as phone dialling, shower noise, and tooth brushing. The spectrograms of these noise environments are shown in Fig. 2. Four test SNRs were used, two low (0 dB and 5 dB) and two high (15 dB and 20 dB). It should be mentioned that this test set is very challenging based on the fact that the speech dataset is different from the one used in training, the number of speakers is very large, and the noise environments are very intrusive and unseen during the training process [7].

The audio files were down-sampled to 8 KHz, which is the same sampling frequency as the ASR model used for testing. Mean Squared Error (MSE) is the loss function used for the speech enhancement networks. The Adam optimizer is used with a learning rate = 0.0001, β_1 = 0.1 for the first stage DE-CADE network and β_1 = 0.5 for the second stage LSGAN. A batch size of 2 was used in training. The first stage DE-CADE network was trained for 100 epochs, while the second stage LSGAN was trained for 20 epochs, which was enough for the model to converge. For the SNR classifier, the binary cross entropy loss function is used, and the network was trained for 300 epochs.

V. Results and Discussion

The performance of the speech enhancement architecture was evaluated using the well-known speech quality and intelligibility metrics: Perceptual Evaluation of Speech Quality (PESQ) [22] and Short-Time Objective Intelligibility (STOI) [23]. Moreover, we used Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [24] to measure speech distortion. On the other hand, the performance of the ASR system was tested using the WER.

A. Speech Enhancement Performance

A comparison is presented in Table I for the performance of the proposed architecture against other best-performing two-stage and similar GAN models in the literature. As baselines, we used the single-stage Metric-GAN architecture, presented in [25], which is a GAN model designed to optimize the PESQ

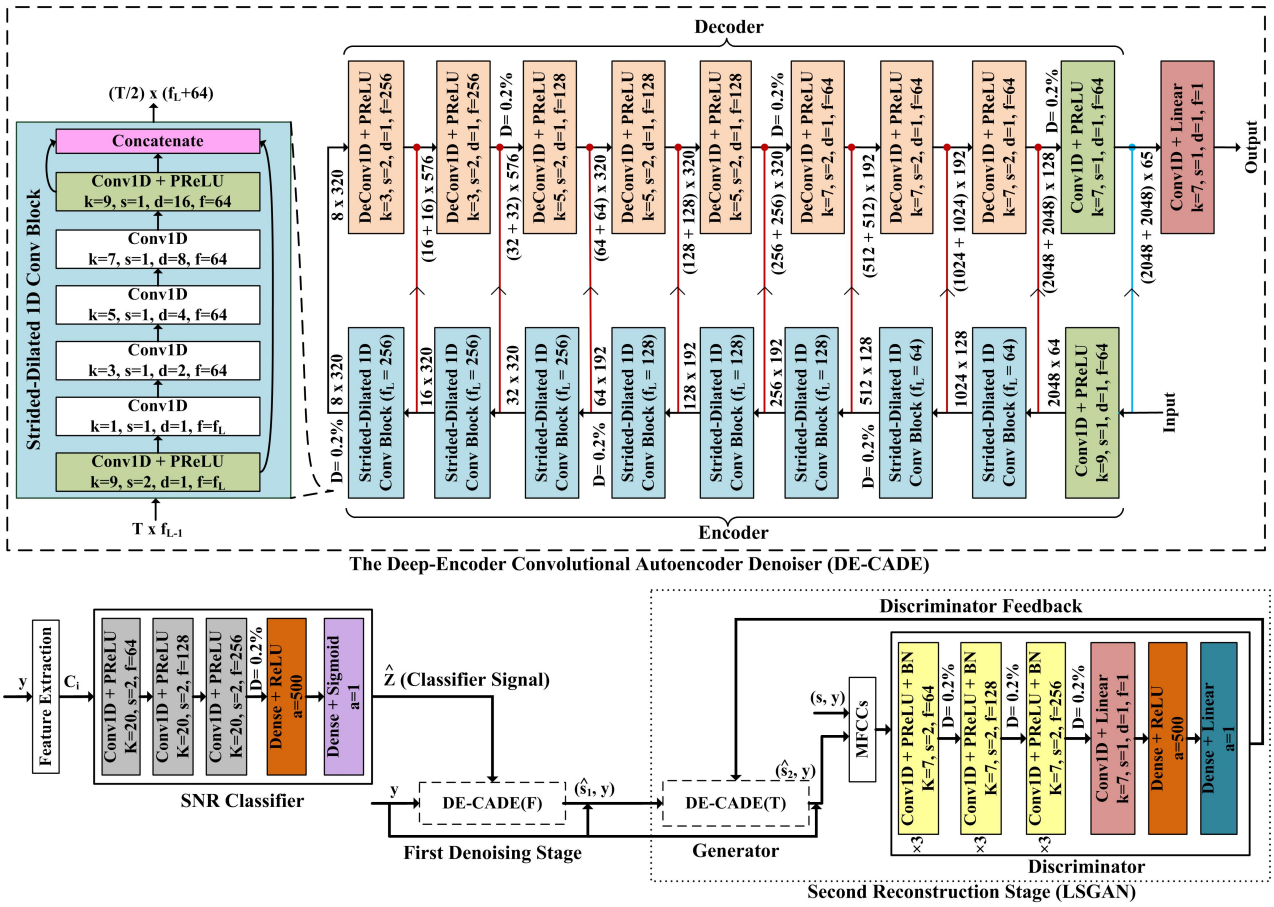

Fig. 1. The proposed speech enhancement architecture. *k*, *d*, *f*, and *L* represent kernel size, dilation rate, number of convolution channels, and layer number respectively; *s* represents stride size in the encoder, and upsampling size in the decoder. *T* is the time samples, and *a* is the number of units. *y* is the noisy speech, C_i is the input feature vector to the SNR classifier, and $\hat{z}$ is the predicted label by the SNR classifier. *s* is the clean speech and $\hat{s}_1$ and $\hat{s}_2$ are the enhanced speech by the first and second stages, respectively.

score; this is denoted by GAN_1. Moreover, our network was compared to the two-stage cascaded GAN model, proposed in [6], which was proven to improve the performance of GANs for speech enhancement; this is denoted by GAN_2. The output from the first stage DE-CADE is also shown in Table I, referred to as *DE-CADE*$_{s1}$. Finally, we compared the architecture to the two-stage DE-CADE network presented in our previous work [5], which performs speech enhancement using cascaded DE-CADE networks, with the first stage operating in the frequency domain and the second stage running in the time domain; this is denoted by *DE-CADE(F-T)*. For a fair comparison, all the models were trained and tested using the same dataset, presented in Section IV. The complexity of all architectures is shown in Fig. 3, where the one-stage GAN that is optimized to improve the PESQ score, GAN1, was used to compare with our first stage DE-CADE$_{s1}$, both have a similar number of parameters of 6.3 million. A two-stage cascaded GAN of 58 million parameters, GAN2, and our previous two-stage DE-CADE network, DE-CADE(F-T) (12.6 million parameters), were used to compare with the two-stage architecture proposed in this work (also 12.6 million parameters).

The presented results are the average of the four test SNRs. The results show that our architecture outperforms in terms of speech quality, intelligibility, and distortion scores. Furthermore, the first stage *DE-CADE*$_{s1}$ shows a better PESQ score than the Metric-GAN model, GAN_1, which is trained to maximize the PESQ score. At the same time, the proposed two-stage architecture performs better than the cascaded GANs, GAN_2, although it is less complex; GAN_2 has 58 million parameters; while ours has 12.6 million parameters.

TABLE I
PERFORMANCE COMPARISON TO THE BEST-PERFORMING SPEECH ENHANCEMENT MODELS.

Metric	Noisy	GAN_1	DE-CADE$_{s1}$	GAN_2	DE-CADE(F-T)	Ours
PESQ	2.50	2.81	2.95	3.11	3.20	**3.30**
STOI(%)	83.7	84.8	86.4	87.8	88.2	**88.6**
SI-SDR	6.10	11.16	12.64	12.81	13.98	**15.06**

B. Automatic Speech Recognition Performance

We used a baseline time-delay neural network ASR system provided by Intelligent Voice for research purposes [26], to

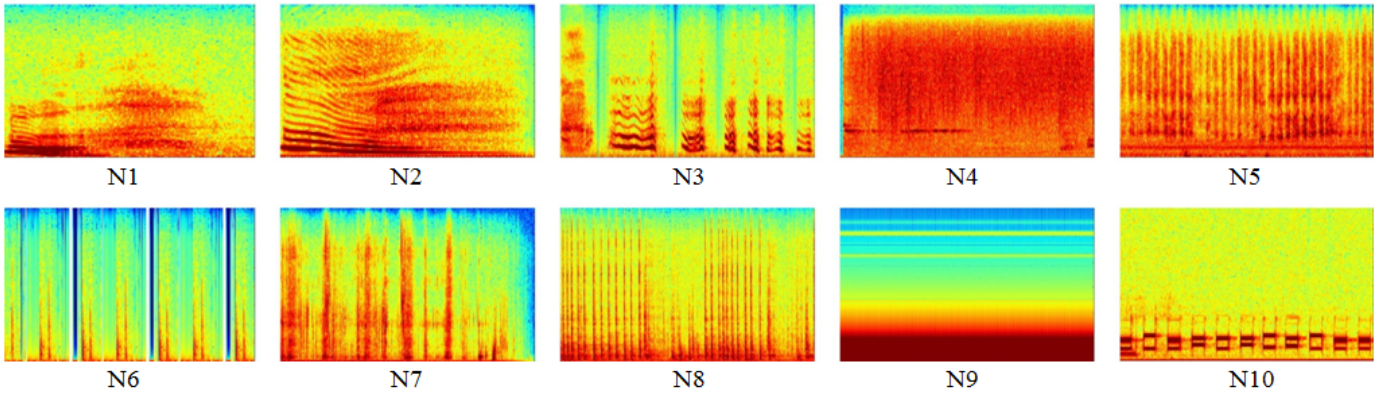

Fig. 2. Testing Noise Environments, N1-N2: yawn sound; N3: Cry; N4: Shower; N5: Toothbrushing; N6-N7: Footsteps; N8: Door moving; N9-N10: Phone dialing

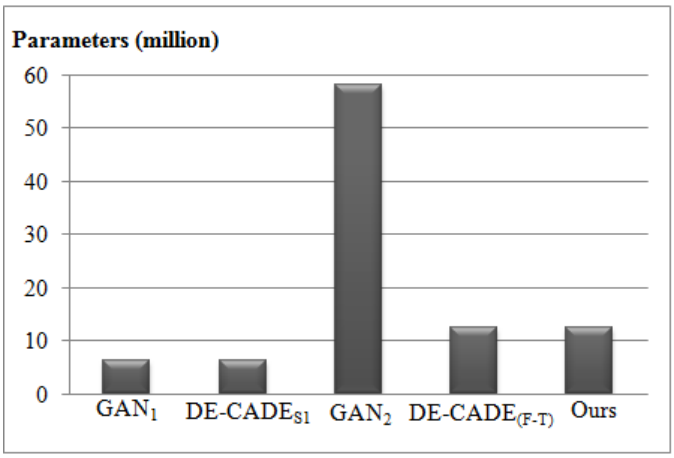

Fig. 3. Comparison of speech enhancement architectures parameters

show the effect of adding the speech enhancement architecture as a preprocessing stage to ASR. The WER of the ASR for the clean test set is 31.9%.

The WER is shown in Table II for the unprocessed speech, WER_{Unproc}, and for the speech processed by the single-stage DE-CADE, WER_{SE1}, the two-stage speech enhancement architecture, WER_{SE2}, and after including the SNR classifier, WER_{C+SE}. The accuracy of the frontend SNR classifier at the testing SNRs and for clean speech is shown in Fig. 4. Although the test data is challenging and highly mismatched, a clear improvement in ASR performance is shown after adding the speech enhancement architecture at very low SNRs 0 dB and 5 dB. The single-stage DE-CADE results in a 6.5% and 5.7% decrease in the WER at 0 dB and 5 dB SNRs, respectively. The second stage significantly improves the performance of the first stage by a further 7.3% and 4.7% decrease in the WER at 0 dB and 5 dB SNRs, respectively. This makes a total of 13.8% and 10.4% WER reduction. SNR classifier accuracy at these very low SNR values is 100%, so the performance is the same after including the classifier.

As the ASR model is trained on some noisy data, the improvement caused by the speech enhancement network becomes less significant at high SNRs, such as in the case of 15 dB SNR, where a 4.1% decrease in WER is seen after adding the speech enhancement architecture. At the same time, the classifier accuracy drops to 80% for 15 dB SNR noisy speech files, leading to a slightly higher WER compared to the case of processing the speech with the speech enhancement network only. This is because 15 dB SNR is the threshold used by the classifier to differentiate between low and high SNRs; therefore, it is the most challenging SNR value for the classifier to output the correct decision. However, the positive effect of the SNR classifier is shown when processing clean speech and at 20 dB SNR, where the distortion caused by the speech enhancement processing overrides the improvement of the denoising, leading to a higher WER when compared to the WER of unprocessed speech. The classification accuracy is 90% for 20 dB SNR and 94% for clean speech, resulting in a 0.4% and 0.3% WER reduction for 20 dB and clean speech, respectively, for the generated speech by the full architecture, WER_{C+SE}, in comparison to the WER of the processed speech by the speech enhancement network only without the classifier, WER_{SE2}, and this also results in a lower average WER for the full architecture.

TABLE II
AUTOMATIC SPEECH RECOGNITION PERFORMANCE

SNR	Clean	20 dB	15 dB	5 dB	0 dB	Ave
$\mathbf{WER}_{Unproc.}$	31.9	33.2	39.7	53.9	65.7	44.9
$\mathbf{WER}_{SE1}$	32.4	33.9	36.7	48.2	59.2	42.1
$\mathbf{WER}_{SE2}$	32.4	33.8	35.5	43.5	51.9	39.4
$\mathbf{WER}_{C+SE}$	32.1	33.4	35.6	43.5	51.9	**39.3**

C. Second Stage Generalization

An experiment was conducted to show the generalization of the second-stage LSGAN network to other first-stage DNNs, not previously seen in the training process. Two pre-trained DNNs were used in this evaluation: an MLP model [27] and an RNN model [4], available in [28] and [29], respectively.

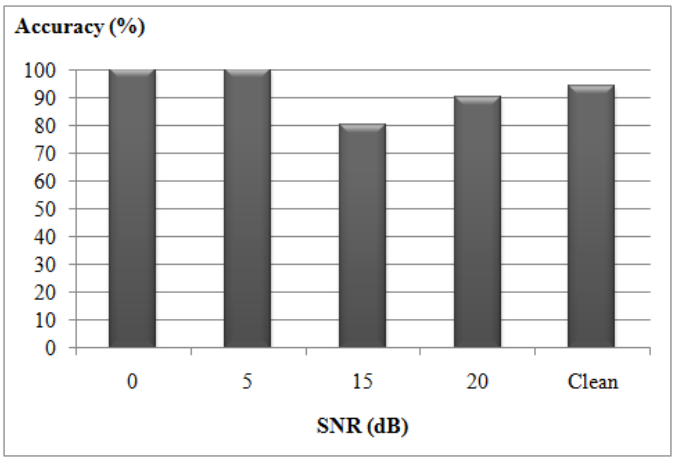

Fig. 4. The accuracy of the frontend SNR classifier at the testing SNRs and for clean speech utterances

Both models are frequency domain-based implementations, but a masking training target is used for these models, which is a different training target than the mapping target used in our first-stage network. This increases the mismatch between the first stage of testing and training DNNs, which ensures a fair assessment of the generalization ability. Additionally, the test set used is seen by the MLP network during the training process. This will show the effect of the proposed second-stage LSGAN even when the test data is not challenging for the first-stage speech enhancement DNN.

This evaluation is based on 0 dB SNR, and its results are shown in Table III, where subscripts 1 and 2 denote running the model as a single stage and after adding the second stage of our architecture, respectively. The results show that adding the second stage LSGAN results in a speech enhancement performance gain in terms of both speech quality and intelligibility. On the other hand, a remarkable reduction in the WER of the ASR is shown for both models after adding the LSGAN. Moreover, the ability of the LSGAN to solve the mismatch problem between speech enhancement and ASR is clear for the MLP model. Although the MLP generated speech with better quality and intelligibility than the noisy speech, it fails to improve the ASR performance, leading to a worse WER than the noisy speech. However, the MLP managed to improve the WER of the ASR after adding the second stage LSGAN.

TABLE III
SECOND STAGE NETWORK GENERALIZATION TO OTHER SPEECH ENHANCEMENT MODELS

Metric	Noisy	MLP_1	MLP_2	RNN_1	RNN_2
PESQ	1.92	2.84	**2.92**	2.48	**2.53**
STOI(%)	73.8	82.7	**82.9**	79.9	**80.1**
WER	65.7	65.9	**59.3**	60.1	**54.2**

VI. CONCLUSION

This paper presents a speech enhancement architecture that minimizes distortion, for integration with an ASR model to improve the ASR model's performance. The architecture performs speech enhancement for low SNR environments only based on the decision of an SNR classifier as a first processing step. If a low SNR is detected, a two-stage speech enhancement processing is applied using a first-stage CDAE-based network for denoising, and a second-stage LSGAN architecture to deal with the distortion caused by the first enhancement stage. The architecture shows better speech enhancement performance when compared to the best-performing models in the literature. Additionally, it improves the performance of the ASR model for highly challenging noisy test data. Furthermore, the results show that the second-stage LSGAN can be used as a standalone speech enhancement network to improve first-stage DNNs for speech enhancement, not seen in the training process. Future work will be done to improve the classification accuracy of the SNR classifier for SNR values near the decision boundary.

REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC, 2013.
[2] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Snr-based progressive learning of deep neural network for speech enhancement." in *Interspeech*, 2016, pp. 3713–3717.
[3] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *ICASSP*. IEEE, 2018, pp. 5609–5613.
[4] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *International Conference on Speech and Computer*. Springer, 2020, pp. 79–86.
[5] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Two-stage deep learning approach for speech enhancement and reconstruction in the frequency and time domains," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–10.
[6] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.
[7] A. Pandey and D. Wang, "On cross-corpus generalization of deep learning based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2489–2499, 2020.
[8] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.
[9] P. Wang, K. Tan *et al.*, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2019.
[10] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," *Computer Speech and Language*, 2016.
[11] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, 2020.
[12] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 239–243.
[13] Y.-L. Shen, C.-Y. Huang, S.-S. Wang, Y. Tsao, H.-M. Wang, and T.-S. Chi, "Reinforcement learning based speech enhancement for robust speech recognition," in *ICASSP*. IEEE, 2019, pp. 6750–6754.

[14] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[15] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, T. Moriya, and N. Kamo, "Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition," in *INTERSPEECH*, 2021, pp. 1149–1153.

[16] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, N. Kamo, and T. Moriya, "Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6287–6291.

[17] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[18] F. Alías, J. C. Socoró, and X. Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Applied Sciences*, vol. 6, no. 5, p. 143, 2016.

[19] C. K. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," *arXiv*, 2020.

[20] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *O-COCOSDA/CASLRE*. IEEE, 2013, pp. 1–4.

[21] G. Hu, "100 nonspeech environmental sounds." [Online]. Available: http://www.cse.ohiostate.edu /pnl/corpus/HuCorpus.html, 2014.

[22] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation, p. 862.*, 2001.

[23] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[24] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP*. IEEE, 2019, pp. 626–630.

[25] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[26] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[27] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *INTERSPEECH*, 2015, pp. 1508–1512.

[28] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Deep Noise Suppression (DNS) Challenge." [Online]. Available: https://github.com/microsoft/DNS-Challenge, 2020.

[29] X. Yong, D. Jun, H. Zhen, D. Li-Rong, and L. Chin-Hui, "DNN based Speech Enhancement Demo." [Online]. Available: https://github.com/yongxuUSTC/DNN-Speech-enhancement-demo-tool, 2015.