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

Adversarial Latent Representation Learning for Speech Enhancement

*Yuanhang Qiu, Ruili Wang**

School of Natural and Computational Sciences, Massey University, New Zealand
{y.qiu1,ruili.wang}@massey.ac.nz

Abstract

This paper proposes a novel adversarial latent representation learning (ALRL) method for speech enhancement. Based on adversarial feature learning, ALRL employs an extra encoder to learn an inverse mapping from the generated data distribution to the latent space. The encoder builds an inner connection with the generator, and provides relevant latent information for adversarial feature modelling. A new loss function is proposed to implement the encoder mapping simultaneously. In addition, the multi-head self-attention is also applied to the encoder for learning of long-range dependencies and further effective adversarial representations. The experimental results demonstrate that ALRL outperforms current GAN-based speech enhancement methods.

Index Terms: adversarial feature learning, latent space, speech enhancement

1. Introduction

Speech enhancement aims to improve the intelligibility and overall perceptual quality of contaminated speech signals [1]. There are many practical applications such as telephone communications [2], hearing-aid devices [3], and human-computer interactions [4], which regard the speech enhancement as an essential operation for different purposes and processing stages. Obviously, more complicated and critical application scenarios require higher performance of speech enhancement.

Classic signal processing methods of speech enhancement (e.g. Wiener filtering [5], spectral subtraction [6]) perform well in specific additive noise suppressing. However, these methods are difficult to process assorted unknown noise interference satisfactorily. In order to solve this problem, learning appropriate representation of noise data distribution is a key procedure in current data-driven approaches.

Recently, deep learning based methods have shown revolutionary information learning and reconstruction property in many research areas. Profiting from this, a series of neural network based speech enhancement methods such as denoising autoencoder [7, 8], LSTM-based [9], CNN-based methods [10, 11] were also developed for improving speech enhancement performance. Particularly, the generative adversarial networks (GAN) [12, 13], which was originally proposed with artful architecture design for high quality images generation in computer vision, has been applied successfully to speech enhancement [14, 15].

GAN consists of a generator and a discriminator, which are trained adversarially up to the Nash Equilibrium [12]. For speech enhancement as shown in Figure 1, the generator usually takes in noisy speech and extra noise distributions (i.e. latent vectors) as input and exports targeted data distribution. The discriminator is considered as a classifier trained to distinguish generated samples and clean speech as fake or true. With effective representation learning and enhancement performance, GAN-based speech enhancement methods have attracted a good proportion of attention[14, 16] in speech enhancement.

However, no attention has been paid to latent space for representation learning in speech enhancement. Initially, GAN can generate high quality target from latent vectors based on real data distribution. Thus, in speech enhancement, our hypothesis is that the latent vectors play an important part for representation learning conditioned on explicit noisy data distribution.

In this work, we propose a novel GAN-based method named adversarial latent representation learning (ALRL), which employs an extra encoder inversely mapping the generated data distribution to the latent space, for speech enhancement improvement. In particular, the encoder attempts to build an inner connection with the generator and provides relevant representation information for the adversarial features modelling. The new architecture remodels the inner projection from the concatenated input of noisy speech and latent vectors to the clean speech distribution. To implement the encoder mapping, we propose a new encoder mapping loss function, which captures latent representation by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors. Also, we combine the encoder loss with the relativistic loss [17] to further improve the effectiveness of information learning between the generator and discriminator. In the meanwhile, the multi-head self-attention mechanism [18] is also applied to the encoder in our ALRL for long-range dependencies capturing and further effective representation learning.

The remainder of this paper is organized as follows. The related work is given in Section 2. The details of our adversarial latent representation learning are introduced in Section 3. Sections 4 gives the design of the experiments, while the experimental results are presented in Section 5. Finally, the conclusions and future work are shown in Section 6.

2. Related Work

In this section, we introduce related GAN-based speech enhancement methods and present a preliminary investigation of latent space with GANs. Based on these works, our ALRL is proposed to learn semantic representation and improve speech enhancement performance.

2.1. GAN-based Speech Enhancement

Recently, the GAN-based models have derived huge progress on semantic representation learning and improved speech enhancement performance significantly. Speech Enhancement GAN (SEGAN) is one of the most famous frameworks proposed for time-domain speech enhancement with improved conditional GAN [14], which combined the conditional GAN with the least-squares GAN (LSGAN) together to further alleviate vanishing gradients. This modification is proved to be ef-

*Corresponding author

http://dx.doi.org/10.21437/Interspeech.2020-1593

fective in performance improvement. Below is the loss function of its discriminator:

$$L_D = \frac{1}{2}E_{x\sim P_x, x_c\sim P_{x_c}}[(D(x,x_c)-1)^2] + \frac{1}{2}E_{z\sim P_z, x_c\sim P_{x_c}}[(D(G(z,x_c),x_c))^2] \quad (1)$$

and its generator:

$$L_G = \frac{1}{2}E_{z\sim P_z, x_c\sim P_{x_c}}[(D(G(z,x_c),x_c)-1)^2] \quad (2)$$

where x_c denotes noisy speech; x denotes clean speech; and z denotes random noise distribution (i.e. latent information).

SEGAN operated on raw speech waveform directly rather than the processed spectral features, which is considered to be able to preserve original sequential information such as phase information effectively. SEGAN worked end-to-end and was trained adversarially based on GAN. The fully convolutional architecture consists of a downsampling and a upsampling modules (i.e. encoder and decoder). The random noise z (i.e. latent vectors) was added to the bottleneck layer for information compensation whereas without further introduction of it. As we know, this work applied conditional GAN to speech enhancement firstly and obtained outstanding performance. In the meanwhile, conditional GANs were also applied to noise-robust speaker verification [19] and speech recognition [20]. However, the latent space still has not been explored thoroughly in speech signal processing.

Speech Enhancement Relativistic GAN (SERGAN) [16] is another framework exploring speech enhancement based on GAN. In the standard GAN, the discriminator is developed to estimate the probability that the original data is real and the generated data is fake, on the contrary, the generator is trained to increase the probability that fake data is real. However, it should simultaneously decrease the probability that real data is real when the generator learns to increase that probability. To accomplish the assumption, the relativistic GAN [17] was proposed for more stable model and higher quality data samples. Below is the loss function of discriminator:

$$L_D = -E_{x\sim P_x, x_c\sim P_{x_c}}[\log(\sigma(C(x,x_c)-C(G(z,x_c),x_c)))] \quad (3)$$

and generator:

$$L_G = -E_{x\sim P_x, x_c\sim P_{x_c}}[\log(\sigma(C(G(z,x_c),x_c)-C(x,x_c)))] \quad (4)$$

where σ is the sigmoid non-linearity, and $C(x)$ denotes the discriminator without the final sigmoid layer. $D(x) = \sigma(C(x))$.

SERGAN built closer information connection between the generator and discriminator for speech enhancement. Moreover, the gradient penalty was also utilized to stabilize model training and improve enhancement performance. The method held a similar architecture with SEGAN but adopted a new loss function to boost information communication between the generator and discriminator. However, this work still did not explore latent space further.

In addition, the improved Speech Enhancement GAN (iSEGAN) [21] conducted a preliminary experiment to explore the impact of the latent vectors on a speech enhancement model by comparing the performance of the model trained with and without latent vector. The results showed that the latent vectors could slightly affect the model performance but were helpful to stabilize model training.

These methods mentioned above attempt to obtain performance gains by modifying model architecture. Also, there are works that explore data space for better model performance.

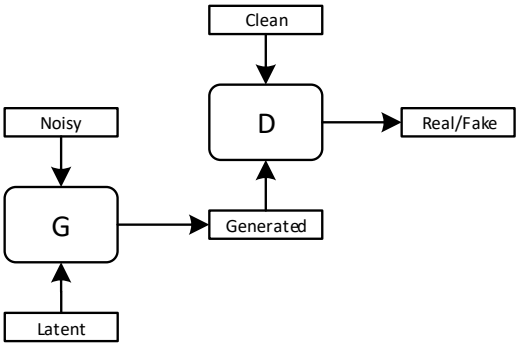

Figure 1: *The basic framework of GAN-based speech enhancement. The generator (G) receives noisy signal and latent vector as input. The discriminator (D) receives generated sample and clean signal as input.*

2.2. Latent Space

It is possible that the latent vectors are used by the generator in a highly entangled way, causing the individual dimensions of latent vectors to not correspond to semantic features of the data [22]. For image generation, Chen et al. proposed to adopt a mutual information strategy for inducing latent vectors. The method decomposed the input noise vectors into a set of semantically meaning factors of variation rather than using single unstructured noise vectors. The work discovered that these latent factors can target salient semantic features of data distribution.

Similarly, Donahue et al. noticed that GAN models could capture semantic variation from latent space, however, have no means of projecting data back into the latent space [23]. This resulted in the architecture ignoring much of the useful information presenting in the structure of the data itself. In addition, interpolations in the latent space of the generator produced smooth and plausible semantic variations and made the model learns associating particular latent directions with specific features. Thus, the Bidirectional Generative Adversarial Networks (BiGAN) was proposed to learn generative mapping from simple latent distributions to arbitrarily complex data distribution [23]. Another similar work about latent space exploration was proposed to map training examples in the data space to the space of latent variables as well [24].

Inspired by mentioned work, we infer that the latent space plays an important role in generative model for semantic representation capturing. Thus, we propose adversarial latent representation learning method for speech enhancement.

3. Adversarial Latent Representation Learning

In this section, we introduce our adversarial latent representation learning (ALRL) for speech enhancement.

3.1. ALRL

The related works show that the latent space can target salient semantic features of data distribution and provide effective guidance for information learning. In our work, an encoder is built for latent representation learning whereas our encoder will be trained for inverse mapping from generated samples to latent space. To implement the encoder mapping, we propose a new encoder loss function, which captures latent representation

by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors.

To further improve the effectiveness of information learning, we combine the encoder loss with the relativistic loss function [17]. The encoder, generator and discriminator will be trained simultaneously. Below is our new loss function for the generator:

$$L_G = - E_{x\sim P_x, x_c\sim P_{x_c}}[\log(\sigma(C(G(z,x_c),x_c) - C(x,x_c)))] \\ - E_{z\sim P_z, x_c\sim P_{x_c}}[||E(G(z,x_c)) - z||_2^2] \quad (5)$$

where E is defined by calculating the squared Euclidean [25] loss. The new loss function improves the semantic representation learning of the generator. To avoid vanishing gradients further, the gradient penalty regularization is also used in discriminator as proposed in [16]. Below is the discriminator:

$$L_D = - E_{x\sim P_x, x_c\sim P_{x_c}}[\log(\sigma(C(x,x_c) - C(G(z,x_c),x_c)))] \\ - \lambda E_{\tilde{x},x\sim P_{(\tilde{x},x)}}[(|| \bigtriangledown_{\tilde{x},x} C(\tilde{x},x)||_2 - 1)^2] \quad (6)$$

where $P_{(\tilde{x},x)}$ is the joint probability of $\tilde{x} = \epsilon x + (1-\epsilon)G(z,x_c)$ and x; ϵ is sampled from a uniform distribution in $[0,1]$; λ is the hyper-parameter that controls the gradient penalty.

Similar setup as SEGAN, the generator receives the noisy speech signal and latent vectors and put them into multi-layers convolutions with the filter (width = 31 and strides N = 2). Before the intermediate layer, a normal 2D convolutional followed by parametric rectified linear units (PReLUs)[26] is used for inherent information capturing from input distributions. Then 2D transposed convolutional, followed again by PReLUs, is used for information reconstruction.

The discriminator is considered as a binary classifier for judgement to real samples and generated samples. The main component is 2D convolutional layer as well. Differently, the discriminator applies the LeakyReLU function [27] and virtual batch normalization function rather other only PReLUs in the generator. This will greatly improve the discriminable information learning of discriminator and alleviate gradients vanishing.

The main structural ingredients of the encoder are also the 2D convolutional layer. Especially, the multi-head self-attention layer is applied to the encoder for the specific speech information locating and the long-range dependencies learning [18].

3.2. Multi-head Self-attention

For each input sequence, the (Query, Key, and Value) vectors will be created by applying learned linear projection or using feed-forward layers. Then the attention will be applied to all other positions with the three vectors. The procedure can be described as below:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \cdot V \quad (7)$$

where d_k is the dimension of the key vectors. The purpose of this scaling is to improve numerical stability as the dimensions of keys, values, and queries grow. The obtained attention at each position will be used to times the value vector of all other positions including itself. This will produce multiple results called multi-head attention. The sum of all heads will be the final result of the first position input. The same operation will be applied at each subsequent position. Below is the equation of the multi-head calculation:

$$MultiHead(Q,K,V) = (\sum_{i=1}^{h} head_i)W^O \quad (8)$$

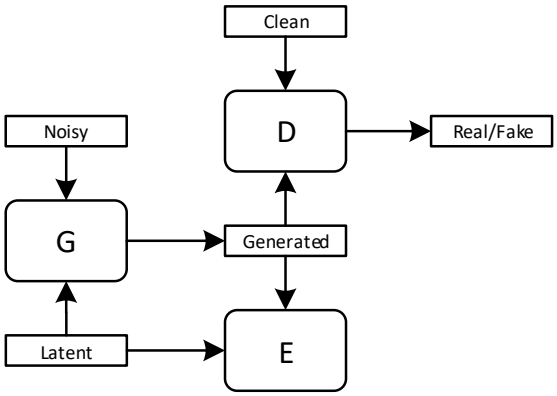

Figure 2: *The framework of adversarial latent representation learning. The encoder (E) processes generated sample and provides latent information for model training.*

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, the matrices W_i^Q, W_i^K, W_i^V, and W^O are the projection weight matrices, respectively. The self-attention module can calculate the response at a specific local position based on the resource collecting from all positions, where the attention vectors are calculated with a small computational cost.

4. Experiments

4.1. Dataset

The selected database is an open and standard resource for the performance evaluation of a speech enhancement system. The original clean speech was selected from Voice Bank corpus[1], including 28 speakers – 14 males and 14 females with the same accent region (England). There are two artificially generated noises (i.e. speech-shaped noise and babble) and eight real noises extracted from the Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) database [28].

For training data, the signal-to-noise (SNR) values are 15dB, 10dB, 5dB and 0dB. That signifies 40 different noisy conditions are produced in this corpus. Each speaker contributes 10 sentences, the corpus will add 400 sentences in total. Each clean speech waveform needs to be normalized and trimmed off silence segments of which are longer than 200ms at the beginning and the end.

Another two speakers (a male and a female), not including in the training data, are picked as the test data from the Voice Bank corpus with the same England accent. Five other noisy types were selected from the DEMAND database. The SNR values are 17.5dB, 12.5dB, 7.5dB and 2.5dB, respectively. Thereby, there are 20 different conditions for each sentence of per test speaker.

4.2. Setup

ALRL adopts the Adam optimizer [29], a learning rate of 0.0002. The raw speech waveforms preserving the original inherent content of speech signals are used same as SEGAN [14]. About one second speech chunks (16384 samples) is segmented by a sliding window (500ms overlap) during training, however no overlap during test. In addition, a high-frequency

[1] https://datashare.is.ed.ac.uk/

Table 1: *The evaluation results of different methods in quality and intelligibility. The results include four SNR conditions (i.e. 17.5 dB, 12.5dB, 7.5dB, and 2.5 dB) and the overall values. The best results obtained are highlighted in the bold font.*

Strategies		Quality				Intelligibility(%)			
Methods	SNR	PESQ	CSIG	CBAK	CVOL	CSII_{high}	CSII_{mid}	NCM	STOI
SEGAN [14]	17.5dB	2.60	4.88	3.28	3.26	99.67	95.60	99.39	95.43
	12.5dB	2.29	4.76	3.06	2.96	99.12	91.10	98.91	94.25
	7.5dB	2.06	4.51	2.87	2.69	97.68	85.21	97.18	92.91
	2.5dB	1.76	4.03	2.59	2.35	93.10	74.78	92.81	89.03
	Overall	2.16	3.48	2.94	2.80	97.29	86.37	96.98	92.80
SERGAN [16]	17.5dB	2.95	4.95	3.56	3.51	99.75	96.93	99.69	96.14
	12.5dB	2.67	4.92	3.31	3.21	99.36	93.36	99.41	95.70
	7.5dB	2.43	4.77	3.09	2.97	98.21	87.89	98.49	93.73
	2.5dB	2.09	4.39	2.79	2.61	94.63	78.38	95.91	90.19
	Overall	2.52	4.75	3.18	3.06	97.91	88.89	98.33	93.69
Our ALRL	17.5dB	3.00	4.97	3.65	3.60	99.76	96.98	99.71	96.12
	12.5dB	2.73	4.94	3.36	3.32	99.39	93.50	99.40	95.11
	7.5dB	2.47	4.81	3.13	3.06	98.28	88.03	98.61	93.67
	2.5dB	2.14	4.45	2.82	2.70	94.81	78.80	96.20	90.31
	Overall	**2.57**	**4.79**	**3.23**	**3.16**	**97.98**	**89.08**	**98.43**	**93.71**

pre-emphasis filer of coefficient 0.95 to all input samples is applied. The epoch is 80 and the batch size is 100. In this work, the fully convolution is used for distribution modelling during downsampling. For more stable training, the 2D convolutional layers followed by PReLUs [26] are applied to project and compress the input signal. Furthermore, the 2D transposed convolutional layers are designed as the key components of upsampling to reconstruct condensed representations.

5. Results

Many objective evaluation measures can evaluate enhanced speech performance with high correlation. The Perceptual Evaluation of Speech Quality (PESQ: from -0.5 to 4.5) for wide band speech is an effective full-reference speech quality evaluation algorithm [30], Moreover, we also implement the composite evaluation metrics of the enhanced speech including the predicted Mean Opinion Score (MOS) of signal distortion (CSIG: from 1 to 5), background noise distortion (CBAK: from 1 to 5), and overall quality (COVL: from 1 to 5).

The intelligibility of enhanced speech is also implemented in this work. The Coherence-based Speech Intelligibility Index (CSII) measure is computed for the medium-level (CSII_{mid}) and high-level (CSII_{high}) segments of each speech sentence, which can predict the intelligibility of peak-clipping and centering-clipping distortion in the speech signal [31]. In addition, another popular speech intelligibility evaluation metrics the Normalized Covariance Metric (NCM) [31] and the Short-Time Objective Intelligibility (STOI) [32] are also conducted.

The experimental results of different methods are shown in Table 1. We set the SEGAN method as the baseline and its result was described in paper [14]. According to the description of the SERGAN method [16], we retrain the SERGAN model and obtained the results as shown in Table 1. Besides the overall evaluation results, we also split the test data to four respective SNR conditions (i.e. 17.5dB, 12.5dB, 7.5dB, and 2.5dB) and obtain evaluation results.

As shown in Table 1, our adversarial latent representation learning (ALRL) method outperforms the SEGAN and SERGAN methods and achieves the highest scores in both speech quality and intelligibility. Specifically, our method improves PESQ by 1.98% and 19.0%, improves STOI by 0.213‰ and 9.81‰ over SERGAN and SEGAN, respectively. Moreover, our method also obtains outstanding enhancement performance in each SNR condition. Our method improves PESQ by 1.69% and 15.4% in 17.5dB, 2.39% and 21.6% in 2.5dB, improves STOI by -0.208‰ and 7.23‰ in 17.5dB, 1.33‰ and 14.4‰ in 2.5dB over SERGAN and SEGAN. Our ALRL can effectively improve the intelligibility and quality of noisy speech, especially for low SNR scenarios.

6. Conclusions and Future Work

In this paper, we propose a novel adversarial latent representation learning (ALRL) method for speech enhancement. An extra encoder model is built in our ALRL to learn the semantic representation by inverse mapping from the generated samples to the latent space. The encoder greatly improves the effectiveness of adversarial training and the complex data distribution learning. To accomplish the inverse mapping, we propose a new loss function, which captures latent representation by calculating the squared Euclidean distance from the inverse mapped generator samples to the latent vectors. In addition, the multi-head self-attention mechanism, applied to the encoder, is also effective for long-range dependencies capturing and further semantic representation learning. The experimental results demonstrate that ALRL outperforms current existing methods in both speech quality and intelligibility, especially for low signal-to-noise ratio scenarios. Our experiments have shown that the latent space is effective to learn semantic representation with adversarial training and our ALRL is effective for speech enhancement performance improvement.

In future work, we will improve our encoder architecture for latent representation learning. Also, our improved speech enhancement method can also be applied to noise-robust speaker identification and speech recognition.

7. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. USA: CRC Press, Inc., 2013.

[2] S. Y. Low, "Compressive speech enhancement in the modulation domain," *Speech Communication*, vol. 102, pp. 87–99, 2018.

[3] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 1, pp. 99–113, 2019.

[4] 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 (TASLP)*, vol. 27, no. 12, pp. 2080–2091, 2019.

[5] Y. Yang and C. Bao, "Dnn-based ar-wiener filtering for speech enhancement," in *Proceedings of the 43th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2901–2905.

[6] D. Cao, Z. Chen, and X. Gao, "Research on noise reduction algorithm based on combination of lms filter and spectral subtraction." *Journal of Information Processing Systems*, vol. 15, no. 4, 2019.

[7] F.-K. Chuang, S.-S. Wang, J.-w. Hung, Y. Tsao, and S.-H. Fang, "Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 3173–3177.

[8] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 86–90.

[9] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *Proceedings of the 43th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5054–5058.

[10] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5756–5760.

[11] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *arXiv preprint arXiv:2004.04098*, 2020.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[13] P. Shamsolmoali, M. Zareapoor, R. Wang, D. K. Jain, and J. Yang, "G-ganisr: Gradual generative adversarial network for image super resolution," *Neurocomputing*, vol. 366, pp. 140–153, 2019.

[14] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proceedings of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 3642–3646.

[15] J. Lin, S. Niu, Z. Wei, X. Lan, A. J. van Wijngaarden, M. C. Smith, and K.-C. Wang, "Speech enhancement using forked generative adversarial networks with spectral subtraction," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 3163–3167.

[16] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 106–110.

[17] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31th Annual Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.

[19] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proceedings of the 18th Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2008–2012.

[20] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proceedings of the 43th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5024–5028.

[21] D. Baby, "isegan: Improved speech enhancement generative adversarial networks," *arXiv preprint arXiv:2002.08796*, 2020.

[22] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016, pp. 2172–2180.

[23] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[24] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[25] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2016, pp. 658–666.

[26] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[28] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[29] I. Bello, B. Zoph, V. Vasudevan, and Q. V. Le, "Neural optimizer search with reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 459–468.

[30] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2007.

[31] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

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