ICA 2019 AACHEN

PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Restoring Lost Speech Components with Generative Adversarial Networks for Speech Communications in Adverse Conditions

Nengheng ZHENG[1]; Yupeng SHI[2]; Yuyong KANG[3]; Qinglin MENG[4];

[1 2 3]Shenzhen University, China

[2] South China University of Technology, China

ABSTRACT

Speech enhancement has been widely implemented to restore quality of speech in communications between humans or between human and machine. For different speech communication scenarios with specific channel and environmental conditions, the types and degrees of speech distortion could vary significantly and many speech enhancement strategies have been developed accordingly. This study deal with a severe distortion problem, i.e., part of the spectral and/or temporal components of the speech are lost completely. The spectral loss is simulated by a transmission channel with very narrow passing bandwidth (lower than 2 kHz) which results in severely degraded speech quality; the temporal loss is simulated by packet loss up to 50% percent in massive communication which results in poor speech intelligibility. A generative adversarial networks (GAN) based speech enhancement scheme is proposed for restoring the missing spectral and temporal components with different network structure and parameters. A set of experiments have been conducted to evaluate the effectiveness of proposed enhancement scheme and promising results have achieved.

Keywords: speech lost compensation; bandwidth extension; packet loss concealment; generative adversarial networks

1. INTRODUCTION

In speech communication, speech loss could happen in time and/or frequency domains due to the varying characteristics of the recording devices and the transmission channels. For example, changing the sampling frequency from 16 kHz to 8 kHz will erase the higher band frequency components, band-pass filtering will kill the components outside the passband, etc. [1]. The delay and jitter during speech packet transmission in a “best effort” packet-switched network can result in the packet loss problem when network congestion happens [1]. In the coming 5G mobile networks, the massive devices connected to the network and different speech communication channel (e.g., VoIP, VoLTE, WiFi, Bluetooth, etc.) involved will exaggerate the speech loss problems [3][4][5][6]. The effective bandwidth of speech will vary significantly across devices and channels. At the same time, network congestion might be more possible due to the demand of communication between huge number of devices.

Techniques for speech loss compensation in both time and frequency domains have been developed in the past decades. Speech bandwidth extension (BWE) based on Gaussian mixture models (GMM), hidden Markov model (HMM), linear prediction analysis, etc. have been implemented to restore the spectral components[7][8][9][10] . Packet loss concealment (PLC) algorithms are widely implemented in standard codecs such as AMR-WB and Opus [11][12]. Recently, deep neural networks (DNN) based approaches were developed to deal with the frequency and/or packet loss problems with promising results [13][14][15].

More recently, generative adversarial networks (GANs) have demonstrated significant

[1] nhzheng@szu.edu.cn
[2] 2172262986@email.szu.edu.cn
[3] 1810262077@email.szu.edu.cn
[4] mengqinglin@scut.edu.cn

performance improvement in many signal processing tasks including image and speech synthesis, speech enhancement [16][17][18][19]. GANs-based system, in which both G and D comprised of DNNs, have also been proposed for BWE [20].

Even though favorable performances on BWE and/or PLC have been achieved by different neural networks-based systems as abovementioned, there still lacks of a systemic investigation on the effectiveness of a general framework to tackle both BWE and PLC, which could happen simultaneously in speech communication. This paper presents a study of such a general framework. The system takes the same GANs architecture proposed in [18]. Nevertheless, in consideration of real-time speech compensation, the GANs structure was modified to accept shorter waveform chunks (200ms) as network input. A set of experiments were conducted to evaluate the system performance. Results show that the GANs-based framework can obtain comparable or better perceptual quality and intelligibility for both BWE and PLC than DNN-based baselines.

2. SPEECH LOSS COMPENSATION WITH GENERATIVE ADVERSARIAL NETWORKS

2.1 Generative Adversarial Networks

Generative adversarial networks (GANs) were proposed by Goodfellow et al. in 2014 [21]. Generally, the networks consist of two sub-networks, each essentially a deep neural network. As the name implies, GANs are trained in an adversarial way between the two DNNs, the generator G and the discriminator D.

Given y$\sim P_y$ (i.e., y follows probability distribution function P_y) the target data to be generated from the networks and z$\sim P_z$ the noise data with known distribution P_z to be input to the networks, the objective of GANs is to generate a new data $\hat{z}$ from z such that $\hat{z} \sim p_y$. To do so, a generator network G and a discriminator network D are constructed and the networks optimization is done through a minimax two-player game played between G and D. The generator is trained to learn a mapping function $z \to \hat{z}$ such to fool the discriminator that $\hat{z}$ is y. On the other hand, the discriminator is trained to classify the target y as real and $\hat{z}$ as fake. Because of the weak guidance in the vanilla generative model, extra conditional information y_c (e.g., the observed data) can be adopted to help training the GANs, as described in [18][22]. Thus, the adversarial training of the whole network can be formulated as

$$\min_G \max_D V_{GAN}(G,D) = \mathbb{E}_{y,\, y_c \sim p_y}[\log P_D(y, y_c)] + \mathbb{E}_{y_c \sim p_y}\left[\log\left(1 - P_D(\hat{z}, y_c)\right)\right] \quad (1)$$

GANs have achieved significant success in speech and image processing. Recently, Pascual et al. proposed a GANs-based speech enhancement system, i.e., the least square GAN (LSGAN) [23] with L_1 loss [18]. Instead of using Jensen-Shannon divergence as in (1), least square error was adopted for the optimization. And the prior knowledge of y in L_1 loss was adopted to better guide the training of G, i.e.,

$$\min_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{y_c \sim P_y}[(P_D(\hat{z}, y_c) - 1)^2] + \lambda \|\hat{z} - y\|_1 \quad (2)$$

where λ is set to be 100 as in [27]. And the training of D is given by

$$\min_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{y_c \sim P_y}\left[\left(P_D(\hat{z}, y_c)\right)^2\right] + \frac{1}{2}\mathbb{E}_{y, y_c \sim P_y}[(P_D(y, y_c) - 1)^2] \quad (3)$$

2.2 GAN-based Framework for Speech Loss Compensation

The framework in this study follows the GANs structure proposed in [18]. The GANs contains a G and a D network and the adversarial training procedure of the networks for speech loss compensation is given in Fig. 1. As shown, the G network takes it structure similar to an auto-encoder, which consists of an encoder and a decoder. The encoder contains 11 convolutional layers with variable depths (16-32-32-64-64-128-128-256-256-512-1024). To simulate the shorter network input (200ms) in speech communication, the fixed strides of 2 as proposed in [18] are modified to variable ones (2-2-1-2-1-2-1-2-1-2-2). The decoder contains 11 deconvolutional layers almost symmetric to the encoder except for the output layer. The D network consists of an encoder which is the same as the one in G and an activation layer as well.

To train the networks, the degraded speech y_c is first segmented into chunks with 3200ms in length and 1600ms overlap between adjacent chunks and then fed into the encoder of the G network. Within the G, the output from the encoder is concatenated with the noise vector z and serves as the input to

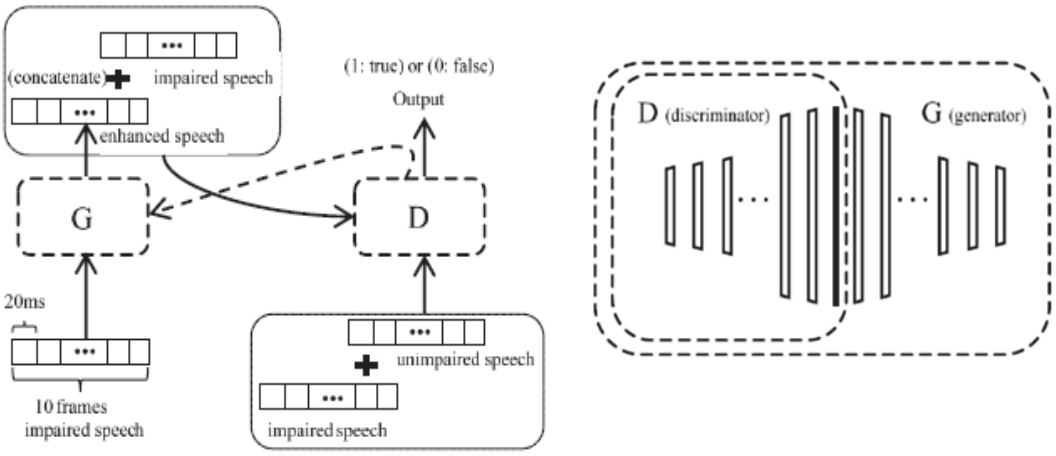

Figure 1: *Diagram of the adversarial training in GANs-based enhancement framework*

the decoder. Skip connections scheme as in [24] is also adopted in G to improve its performance by pass more useful details from the convolutional layers to the corresponding deconvolutional layers. The output of G, i.e., $\hat{z}$, is concatenated with the clean speech y, i.e., speech without any loss, and the corresponding impaired speech y_c, and fed into the network D for computing the probabilities $P_D(\hat{z}, y_c)$ and $P_D(y, y_c)$ as in (3), the former is transmitted in turn to G to guide the training of G.

3. EXPERIMENTS

3.1 Dataset and Preprocessing

An open access English speech database for speech enhancement evaluation as adopted in [25] was also adopted in this study to evaluate the effectiveness of the proposed GANs-based speech restoration system. The database contains parallel clean and noisy speech dataset, in which the clean speeches were recorded from native English speakers and the noisy ones were generated from the clean speeches by adding different noises. Each dataset contains a training set (23075 utterances recorded from 56 speakers) and a test set (824 utterances recorded from 2 speakers). In this study, only the clean data were used to generate the training and test data for the experiments. The original data are with 48 kHz sampling rate and were down-sampled to 16 kHz in the experiments.

For BWE, the training data were randomly divided into four subsets, each for a simulation of a specific spectral loss. Three of the subsets were low-pass filtered with cut-off frequencies at 1.5 kHz, 2.5 kHz and 3.5 kHz, respectively, to generate the high frequency loss data. The remaining one kept unchanged to represent the no loss data. Similarly, the test data were randomly divided into five subsets. One kept unchanged and the other four were low-pass filtered with cut-off frequencies at 1kHz, 1.5 kHz, 2.5 kHz and 3.5 kHz, respectively.

For PLC, the training data were randomly divided into five subsets and packet loss rates of 0 (i.e., no loss), 10, 20, 30 and 40 percent were simulated to generate packet loss speech with the 5 subsets, respectively. Similarly, the packet loss speech data for test were generated from the test data with 6 different packet loss rates of 0, 10, 20, 30, 40 and 50 percent, respectively.

In addition, a packet in the experiments contained a 20ms speech frame and the lost packets were filled with -80dB white Gaussian noise.

3.2 Network Settings

As mentioned in Section 2.2, for training, all signals were segmented into a sequence of 3200ms chunks with 50% overlap. As for test, the utterances were also segmented into chunks of 3200ms but without overlap. For real-time consideration, the 3200ms chunks can be constructed by concatenating the current frame (20ms) with the previous 9 frames in the buffer.

The network G is an encoder-decoder convolutional topology containing 22 layers, each with a fixed filter size of 31 and a variable stride. Layer weights and bias for the GANs were initialized as in [18]. Activation functions used in G were Parametric ReLUs in convolutional layers and the

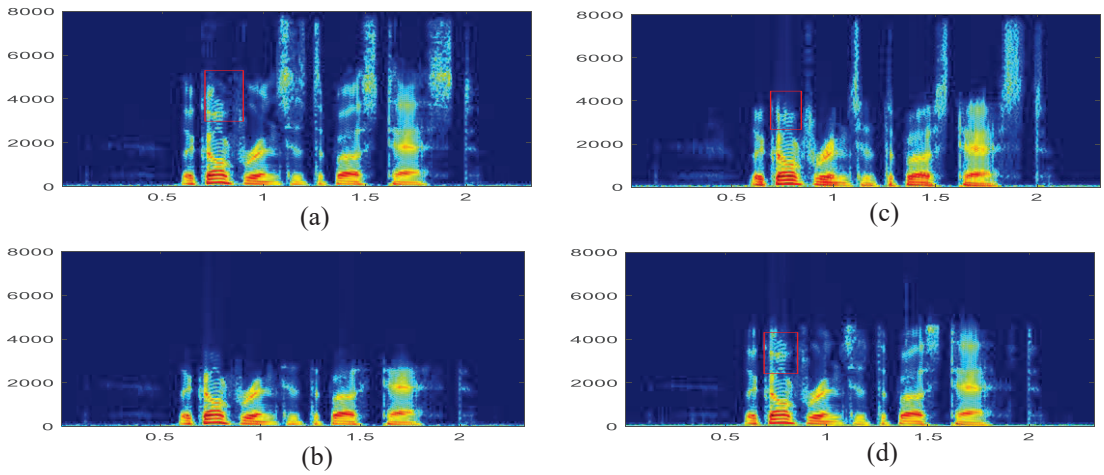

Figure 2: *Spectrograms for high frequency lost and compensated utterances. (a) original, (b) high frequency loss with cut-off frequency at 2.5kHz, (c-d) compensated by DNN and GANs.*

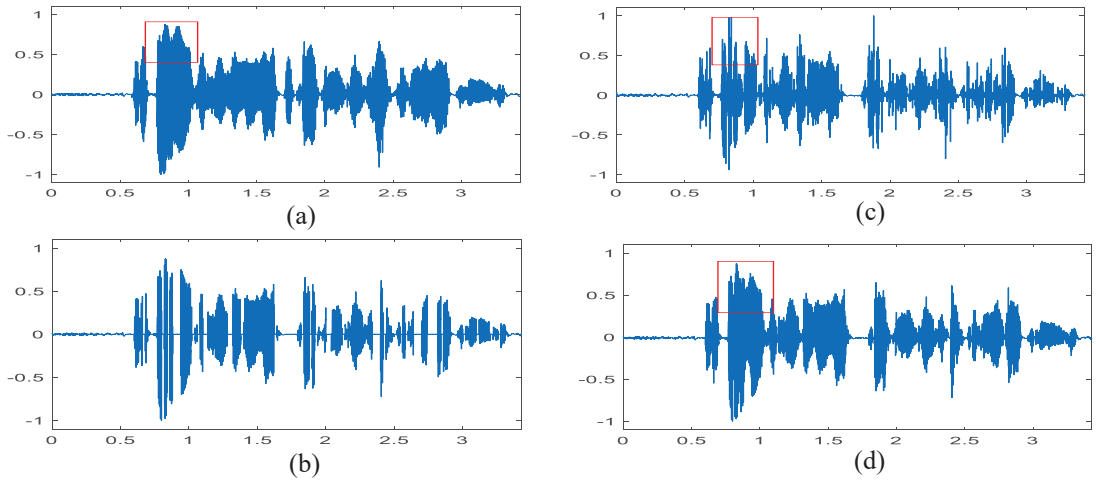

Figure 3*: Waveforms for packet lost and compensated utterances. (a) original, (b) speech with packet loss rate of 30%, (c-d) compensated by DNN and GANs*

hyperbolic tangent (tanh) for the output layer. Activation functions for D were Leaky ReLU nonlinearities with α=0.3. Besides, in order to accelerate the training and avoid overfitting, virtual batch normalization [26] was adopted in the GANs. The GANs-based compensation framework was trained for 50 epochs with a learning rate of 0.0002 and the gradient descent optimizer was the Adam [27]. The settings are the same for both BWE and PLC cases.

For comparison, DNN-based systems were adopted as baselines. The DNN has 3 hidden layers with 2048 nodes in each layer. Layer weights and bias were initialized as in [13]. Activation functions used in each layer were all "ReLu". For both BWE and PLC, DNN was trained for 100 epochs with a learning rate of 0.001 and an Adm optimizer. Learning rate decreased in exponential decay rate (initialized to 0.9) per epoch. Batch normalization was applied to stabilize the network training.

The features to the DNN were the same as in [13], each signal was segmented into a sequence of frames (20ms per frame and 10ms overlap). To each frame, 512-point short-time Fourier transform was implemented and the log-magnitude of the first 257 frequency components was calculated to compose a 257-dimensional vector. To the t^{th} frame, the input to the network was a 9*257-dimensional feature vector consisting of 9 such 257-dimentional vectors computed from frames t-4 ~ t+4.

In test, the output from the DNN, i.e., the compensated log-magnitude spectra, were used to reconstruct the enhanced speech with the corresponding phase information extracted from the impaired speech.

4. RESULTS AND DISCUSSIONS

Figure 2 gives an example of BWE results. It seems that the DNN based systems recover more high frequency components, especially for those of unvoiced speech, than the GANs. However, the spectrogram reconstructed by GANs is more closed to the original one than those by DNNs, especially at frequency lower than about 4 kHz where the majority of speech power resides. Figure 3 gives an example of PLC results. It is clear that the waveform reconstructed by the GANs is more closed to the original one than those by DNN.

Table 1 gives the BWE results obtained from three systems, i.e., no processing (UP), DNN baseline (DNN) and the proposed GANs system (GAN). The results are measured in four evaluation metrics, i.e., PESQ, LSD, STOI and SNR [28][29][30][31]. In the table, SEEN means that the cut-off frequencies of the test data are the included in the training data while UNSEEN means that the loss conditions of the test data are not included in training. As shown, the speech quality degrades more significantly as the cut-off frequency decreases, i.e., more high frequency component loss. GANs outperform DNN for all metrics except LSD. This is because that the DNN baseline was trained to minimize the Euclidean distance between the compensated log-magnitude spectra and the target ones, on the other hand, the GANs-based system was an end-to-end framework with waveform input.

The PLC results are given in Table 2. As shown, the more packets lost, the worse the speech quality with respect to all metrics. The superiority of GANs over DNN in PLC is more significant than in BWE. Even for LSD, GANs lost to DNN1 only at 40% and 50% packet loss cases.

Table 1 : *Mean scores for PESQ, LSD, STOI and SNR obtained from three bandwidth extension systems*

BWE		PESQ			LSD			STOI			SNR		
		UP	DNN1	GANs	UP	DNN	GANs	UP	DNN	GANs	UP	DNN	GANs
SEEN	unimpaired	4.50	3.58	4.06	0	0.57	0.31	1	0.96	0.99	∞	23.82	68.95
	3500Hz	4.50	3.68	3.88	1.38	0.94	1.06	1.00	0.97	0.99	8.15	7.59	41.66
	2500Hz	4.42	3.64	3.89	1.68	1.08	1.20	0.99	0.96	0.99	1.93	2.17	42.25
	1500Hz	4.01	3.50	3.57	1.96	1.28	1.47	0.95	0.91	0.95	-4.65	-4.42	45.53
UNSEEN	1000Hz	3.74	3.25	3.32	2.08	1.79	1.82	0.91	0.87	0.87	-8.20	-9.49	13.81

Table 2: *Mean scores for PESQ, LSD, STOI and SNR obtained from three packet loss concealment systems*

PLC		PESQ			LSD			STOI			SNR		
		UP	DNN	GANs	UP	DNN	GANs	UP	DNN	GANs	UP	DNN	GANs
SEEN	0%	4.50	3.52	4.29	0	0.52	0.30	1.00	0.96	0.99	∞	21.80	69.68
	10%	2.62	2.88	3.79	0.31	0.62	0.45	0.92	0.91	0.96	24.63	16.46	35.98
	20%	1.90	2.52	3.02	0.61	0.72	0.60	0.84	0.87	0.92	16.77	13.39	26.86
	30%	1.42	2.24	2.32	0.92	0.82	0.77	0.78	0.83	0.87	12.37	10.43	20.59
	40%	0.97	1.92	1.96	1.21	0.91	0.92	0.68	0.77	0.81	9.00	8.04	16.52
UNSEEN	50%	0.71	1.74	1.61	1.49	1.01	1.12	0.61	0.73	0.75	7.06	6.27	13.10

5. CONCLUSIONS

This study investigated the effectiveness of a GANs as a general framework for restoring lost speech lost compensation. A set of experiments were carried out to evaluate the performance of the GANs-based system in comparison with a DNN baseline. Different frequency and packet loss conditions were simulated for evaluation. Results show that GANs obtained better speech quality and intelligibility than DNN for both seen and unseen loss conditions.

ACKNOWLEDGEMENTS

This work is jointly support by National Natural Science Foundation of China (Project 61771320, 11704129) and Shenzhen Science and Technology Foundation (Project JCYJ20170302145906843).

REFERENCES

[1] Y. X. Li and S. Kang, “Artificial bandwidth extension using deep neural network-based spectral envelope estimation and enhanced excitation estimation,” IET Signal Processing, vol. 10, no. 4, pp. 422-427, 2016.
[2] M. Yang and N. G. Bourbakis, “An efficient packet loss recovery methodology for video streaming over IP networks,” IEEE Transactions on Broadcasting, vol. 55, no. 2, pp. 190–201, 2009.
[3] S. Karapantazis and F. Pavlidou, “VoIP: A comprehensive survey on a promising technology,” Comput. Netw., vol. 53, no. 12, pp. 2050–2090, 2009.
[4] H. Kim, D. Kim, M. Kwon, H. Han, Y. Jang, D. Han, T. Kim and Y. Kim, “Breaking and fixing VoLTE: exploiting hidden data channels and mis-implementations,” Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 328–339, 2015.
[5] S. Song and B. Issac, “Analysis of WiFi and WiMAX and wireless network coexistence,” International Journal of Computer Networks & Communications (IJCNC), vol. 6, no. 6, pp. 63-78, 2014.
[6] K. V. S. S. S. S. Sairam, N. Gunasekaran, and S. R. Redd, “Bluetooth in wireless communication,” IEEE Communications Magazine, vol. 40, no. 6, pp. 90-96, 2002.
[7] K.-Y. Park and H. S. Kim, “Narrowband to wideband conversion of speech using GMM based transformation,” in Proc. ICASSP, vol. 3, pp. 1843–1846, 2000.
[8] P. Jax and P. Vary, “Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model,” in Proc. ICASSP, vol. 1, pp. 680-683, 2003.
[9] C. A. Rodbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, “Hidden Markov model-based packet loss concealment for voice over IP,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 5, pp. 1609-1623, 2006.
[10] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, “Speech enhancement via frequency bandwidth extension using line spectral frequencies,” in Proc. ICASSP, pp. 665-668, 2003.
[11] A. Takahashi, H. Yoshino, and N. Kitawaki, “Perceptual QoS assessment technologies for VoIP,” IEEE Commun. Mag., vol. 42, no. 7, pp. 28–34, 2004.
[12] K. Vos, K. V. Sorensen, S. S. Jensen, and J.-M. Valin, “Voice coding with Opus,” in the 135th AES Convention, 2013.
[13] K. H. Li and C.-H. Lee, “A deep neural network approach to speech bandwidth expansion,” in Proc. ICASSP, pp. 4395-4399, 2015.
[14] K. H. Li, Z. Huang, and C.-H. Lee, “DNN-Based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech,” in INTERSPEECH, pp. 2578-2582, 2015.
[15] B.-K. Lee and J.-H. Chang, “Packet loss concealment based on deep neural networks for digital speech transmission,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 2, 378-387, 2016.
[16] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in ICLR, 2016.
[17] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” arXiv preprint arXiv:1711.11585, 2017.
[18] S. Pascual, A. Bonafonte, and J. Serrà, “SEGAN: Speech enhancement generative adversarial network,” In INTERSPEECH, 2017.
[19] C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” arXiv preprint arXiv:1802.04208, 2018.
[20] S. Kim and V. Sathe, “Bandwidth extension on raw audio via generative adversarial networks,” arXiv preprint arXiv: 1903.09027, 2019.
[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” arXiv preprint arXiv: 1406.2661v1, 2014.
[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” arXiv preprint arXiv: 1611.07004, 2016.
[23] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” arXiv preprint arXiv:1611.04076, 2016.
[24] X. J. Mao, C. H. Shen, and Y. B. Yang, “Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections,” arXiv preprint arXiv:1603.09056, 2016.
[25] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech”, in 9th ISCA Speech Synthesis Workshop, pp. 159-165, 2016.
[26] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs,” arXiv preprint arXiv: 1606.03498v1, 2016
[27] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 2, pp. 749–752, 2001.
[29] J. Du and Q. Huo, “A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions,” in INTERSPEECH, pp. 569–572, 2008.
[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in ICASSP, pp. 4214–4217, 2010.
[31] V. Kuleshov, S. Z. Enam, and S. Ermon, “Audio super resolution using neural networks,” arXiv preprint arXiv:1708.00853, 2017.