Speech Loss Compensation by Generative Adversarial Networks

Yupeng Shi, Nengheng Zheng, Yuyong Kang, Weicong Rong

The Guangdong Key Laboratory of Intelligent Information Processing, College of Electronic and Information Engineering, Shenzhen University, China

2172262986@email.szu.edu.cn, nhzheng@szu.edu.cn, 1810262077@email.szu.edu.cn, 2172262944@emai.szu.edu.cn

Abstract

Speech loss, including frequency loss and packet loss, can lead to significant speech distortion in many Internet-based speech communication services. In this study, a generative adversarial networks (GANs) structure, which takes deep convolutional neural networks (CNN) as the generator and discriminator components, is adopted as a general framework for speech loss compensation. Network settings are modified for real-time communications. A set of experiments are conducted to evaluate the performance of the GANs-based framework for both bandwidth expansion (BWE) and packet loss concealment (PLC) at several simulated loss conditions. Experimental results demonstrate that the proposed system achieves better performance, with respective to 4 objective metrics, in both BWE and PLC compared to the baseline systems.

Index Terms: Speech loss compensation, bandwidth extension, packet loss concealment, generative adversarial networks.

1. Introduction

Speech loss, from the signal transmission point of view, could happen in either time domain or frequency domain. The time domain loss happens mostly in network congestion situation, where the delay and jitter during speech packet transmission in a "best effort" packet-switched network can result in the packet loss problem [1]. In frequency domain, speech loss happens when recording setup varies [2], e.g., resampling form 16 kHz to 8 kHz will erase the higher band frequency components, band-pass filtering will kill the components outside the passband, etc.

With the arrival of the 5th generation (5G) mobile networks, the number of devices connected to the network will be huge and different speech communication channels, e.g., voice-over-IP (VoIP), voice-over-long-term evolution (VoLTE), WiFi, Bluetooth, etc. might share the same network [3-6]. 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 increasing demand of communication among huge number of devices.

Over the past decades, many techniques have been developed for speech bandwidth extension (BWE) and packet loss concealment (PLC). Conventional approaches (or shallow models) like Gaussian mixture models (GMM), hidden Markov model (HMM), linear prediction analysis and etc. [7-10] have been implemented for BWE or PLC. Besides, many standard codecs such as AMR-WB and Opus [11, 12] embrace the BWE and PLC algorithms. Recently, motivated by the success of deep learning techniques in speech recognition and speech separation [13, 14], many researchers have applied deep neural networks (DNN) to deal with the frequency and/or packet loss problems [15-17]. In [15, 16], BWE was achieved with a DNN, in which a mapping mechanism could be learned and the wideband spectra was reconstructed from the narrowband ones. Results showed that DNN outperforms GMM over objective measures and automatic speech recognition (ASR) accuracy. In [17], a similar DNN-based architecture was adopted to PLC for digital speech transmission. The result showed that DNN outperformed HMMs and AMR-WB algorithm with better speech quality and higher ASR accuracy.

Generative adversarial networks (GANs), trained in an adversarial way between two networks (i.e., the generator G and the discriminator D), have been implemented in image synthesis with significant performance improvement [18,19]. Li et al. [20] proposed a GANs-based system, in which both G and D comprised of DNNs, for BWE with comparable performance to the AMR-WB codec. GANs with G and D comprised of deep convolutional networks have also been successfully implemented in speech enhancement (SE) [21], audio synthesis [22] and BWE [23].

Although 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 SE, BWE and PLC, which could happen simultaneously in speech communication. This paper presents a study of such a general framework. The system takes a similar GANs architecture proposed in [21]. Only its effectiveness on BWE and PLC are investigated in this study since it has already been demonstrated successful for SE in [21]. 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 two DNN-based baselines.

2. GANs-based Speech Loss Compensation

2.1. The generative adversarial networks

Generative adversarial networks (GANs) were proposed by Goodfellow et al. [24] in 2014, and have achieved significant success in speech and image processing ever since. Given $y \sim p_y$ (i.e., y follows probability distribution function p_y) the target data to be generated from the networks, $z \sim p_z$ the noise data with known distribution p_z to be input to the network, 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. G is trained to learn a mapping function $z \rightarrow \hat{z}$, such to fool D as well as possible, and D is trained to classify 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 speech/image data in speech enhancement and image translation) can be adopted to help the training of GANs, as described in [21, 25]. Thus, the adversarial training of the whole network can be formulated as

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

Recently, Pascual et al. [21] proposed a GANs-based speech enhancement system, i.e., the least square GAN (LSGAN) [26] with L_1 loss. In [21], instead of using Jensen-Shannon divergence as in (1), least square error was adopted for the optimization. Meanwhile, 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. Framework for GAN-based speech loss compensation

The framework in this study is modified from the GANs structure proposed in [21]. As shown in Fig. 1, the G network is structured similarly to an auto-encoder structure, which consists of an encoder and a decoder. The encoder contains 10 convolutional layers with variable depths (16-32-32-64-64-128-128-256-256-512). To simulate the shorter network input (200ms) in speech communication, the fixed strides (2 in [21]) are modified to variable (2-2-1-2-1-2-1-2-1-2). The decoder contains 11 deconvolutional layers almost symmetric to the encoder except for the last 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 impaired speech chunks (3200ms length per chunk and 1600ms overlap) are fed into the encoder of the G network. The encoding output y_e, concatenated with noise vector z, serves as the input to the decoder. Skip connections scheme as in [28] is also adopted in G to improve its performance by passing more useful details from the convolutional layers to the corresponding deconvolutional layers. The output of G and the unimpaired speech, i.e., the compensated speech $\hat{z}$ and y, once again concatenated with the corresponding impaired speech y_c, is 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 G's training.

3. Experiments

3.1. Dataset and preprocessing

Experiments were carried out to evaluate the effectiveness of the proposed GANs-based speech loss compensation system for both BWE and PLC cases. An open access English speech databased designed for evaluating speech enhancement methods [29] was adopted in this study. The database contains parallel clean and noisy speech dataset, in which noisy data is generated from the clean speech by adding different noises.

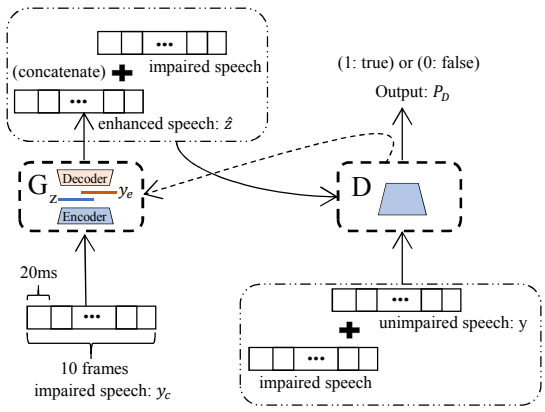

Figure 1: *Diagram of the adversarial training in the GANs-based speech loss compensation framework.*

Each dataset contains a training set (23075 utterances recorded form 56 native English speakers) and a test set (824 utterances recorded form 2 native English speakers). In this study, only the clean dataset was used to generate the training and test data for the experiments. The original data are with 48 kHz sampling rate and were downsampled to 16 kHz in the experiments.

For the BWE experiments, the training data was randomly divided into 4 subsets, 3 of them were passed through low-pass filters with cut-off frequencies at 1.5 kHz, 2.5 kHz and 3.5 kHz, respectively, to generate the high frequency loss data. Similarly, the test data were randomly divided into 5 subsets, 4 of them were low-pass filtered with cut-off frequencies at 1kHz, 1.5 kHz, 2.5 kHz and 3.5 kHz, respectively.

For the PLC experiments, the training data were randomly divided into 5 subsets and packet loss rates of 0, 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.

A packet in the experiments contained a 20ms speech frame and the lost packets were simulated using opus codec demo which is available at https://github.com/xiph/opus.

3.2. Feature extraction

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 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.

3.3. Network setting

As mentioned in section 2.2, the network G is an encoder-decoder convolutional topology containing totally 20 layers, each with a fixed filter size of 31 and a variable stride. Layer weights and bias for the GANs were initialized as in [21]. Activation functions used in G were parametric rectified linear units (PReLUs) in convolutional layers and the hyperbolic tangent (tanh) for the last output layer. Activation functions for D were LeakyReLU nonlinearities with α=0.3. Besides, in order to accelerate the training and avoid overfitting, virtual batch normalization [30] was adopted in the GANs. The GANs-based compensation framework was trained for 50 epochs with a

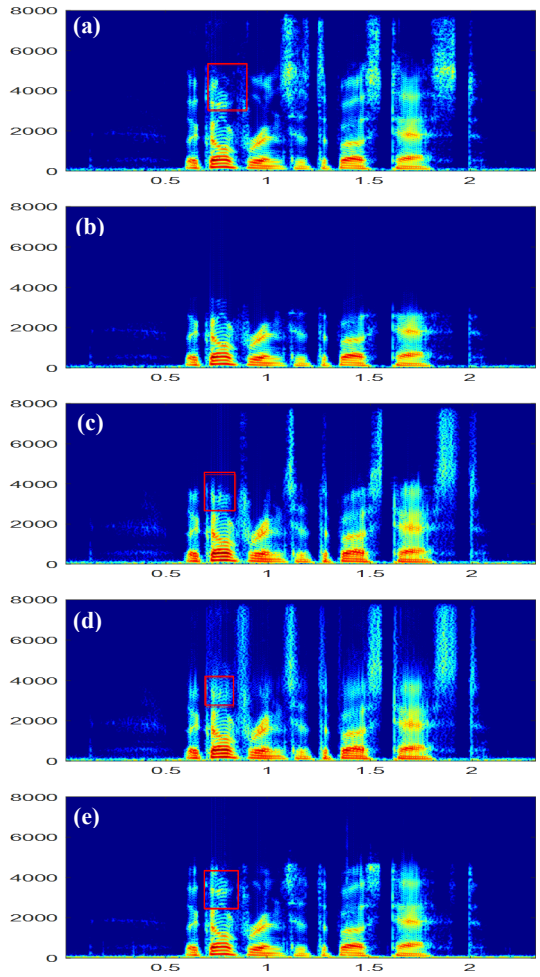

Figure 2: *Spectrograms for high frequency lost and compensated utterances. (a) original, (b) high frequency loss with cut-off frequency at 2.5kHz, (c-e) compensated by DNN1, DNN2 and GANs.*

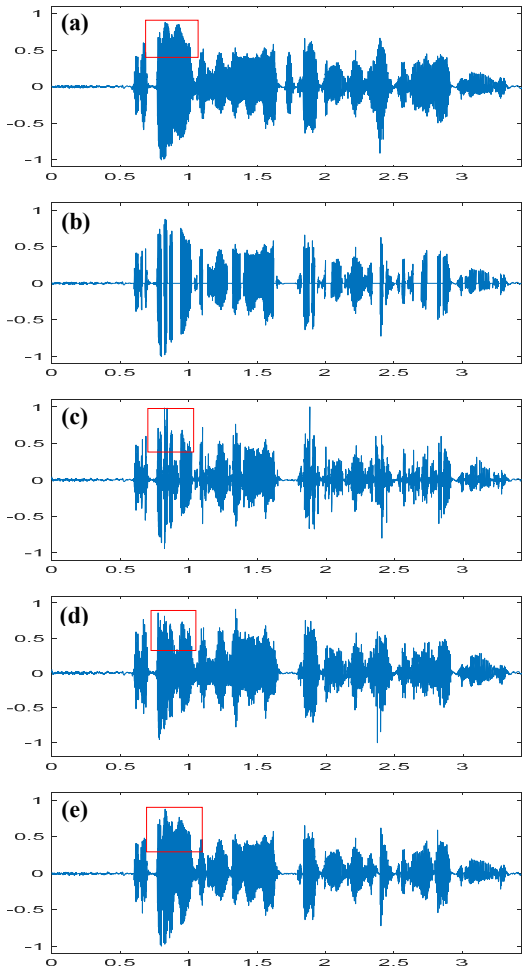

Figure 3: *Waveforms for packet lost and compensated utterances. (a) original, (b) speech with packet loss rate of 30%, (c-e) compensated by DNN1, DNN2 and GANs.*

learning rate of 0.0002 and the gradient descent optimizer was the Adam [31]. The settings are the same for both BWE and PLC cases. The trained model size of G in GANs is about 400M, while the baseline model size is just about 200M. However, GANs perform BWE or PLC in an end-to-end manner without extra signal pre-processing like short time Fourier transform and phase problem.

3.4. Baseline systems

For comparison, DNN-based speech compensation systems were adopted as baselines. The DNN has 3 hidden layers with 2048 nodes in each layer. Layer weights and bias for DNN were initialized as in [15]. Activation functions used in each layer were all "ReLu". For both BWE and PLC experiments, DNN was trained for 100 epochs with a learning rate of 0.001 and an Adm optimizer. To smooth the loss function curve, 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 [15], each signal was segmented into a sequence of frames (20ms frame length and 10ms overlap). To each frame, 512-point short-time Fourier transform (STFT) 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, was used to reconstruct the enhanced speech with the corresponding phase information extracted from the corresponding impaired speech. The baseline is named DNN1 in the experiments. As in [15], another baseline, named DNN2, in which the enhanced speech was reconstructed with the DNN log-magnitude spectra and the ideal phase from the unimpaired speech.

4. Results and Discussion

Figure 2 gives an example of BWE results. It seems that the two DNN based systems recover more high frequency components than the GANs. However, the spectrogram reconstructed by GANs is more closed to the original one than those by DNNs, especially at the frequencies lower than about 4 kHz where the majority of speech power resides. Besides, as to be elaborated

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

BWE		PESQ				LSD				STOI				SNR			
		UP	DNN1	DNN2	GANs	UP	DNN1	DNN2	GANs	UP	DNN1	DNN2	GANs	UP	DNN1	DNN2	GANs
SEEN	unimpaired	4.50	3.58	3.58	**4.06**	0	0.57	0.57	**0.31**	1	0.96	0.96	**0.99**	∞	23.82	23.82	**68.95**
	3500Hz	4.50	3.68	3.69	**3.88**	1.38	0.94	**0.64**	1.06	1.00	0.97	0.97	**0.99**	8.15	7.59	25.47	**41.66**
	2500Hz	4.42	3.64	3.63	**3.89**	1.68	1.08	**0.69**	1.20	0.99	0.96	0.96	**0.99**	1.93	2.17	25.44	**42.25**
	1500Hz	4.01	3.50	3.37	**3.57**	1.96	1.28	**0.77**	1.47	0.95	0.91	0.94	**0.95**	-4.65	-4.42	24.68	**45.53**
UNSEEN	1000Hz	3.74	3.25	3.01	**3.32**	2.08	1.79	**1.19**	1.82	0.91	0.87	0.87	**0.87**	-8.20	-9.49	**17.45**	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	DNN1	DNN2	GANs	UP	DNN1	DNN2	GANs	UP	DNN1	DNN2	GANs	UP	DNN1	DNN2	GANs
SEEN	0%	4.50	3.52	3.52	**4.29**	0	0.52	0.52	**0.30**	1.00	0.96	0.96	**0.99**	∞	21.80	21.80	**69.68**
	10%	2.62	2.88	3.24	**3.79**	0.31	0.62	0.56	**0.45**	0.92	0.91	0.95	**0.96**	24.63	16.46	20.34	**35.98**
	20%	1.90	2.52	**3.03**	3.02	0.61	0.72	0.60	**0.60**	0.84	0.87	**0.93**	0.92	16.77	13.39	19.57	**26.86**
	30%	1.42	2.24	**2.79**	2.32	0.92	0.82	**0.64**	0.77	0.78	0.83	**0.90**	0.87	12.37	10.43	17.41	**20.59**
	40%	0.97	1.92	**2.51**	1.96	1.21	0.91	**0.69**	0.92	0.68	0.77	**0.87**	0.81	9.00	8.04	15.67	**16.52**
UNSEEN	50%	0.71	1.74	**2.29**	1.61	1.49	1.01	**0.76**	1.12	0.61	0.73	**0.84**	0.75	7.06	6.27	13.74	**13.10**

in Table 1, the high-frequency information is not so clear as the ground truth, which seem not to help improve the speech quality obviously. Similarly, an example of PLC results is demonstrated in Fig. 3. It can be observed clearly that the waveform reconstructed by the GANs is more closed to the original one than those by DNNs.

Table 1 and 2 gives the overall results of PESQ (ranging -0.5 to 4.5), LSD, STOI (ranging from 0 to 1) and SNR [32-35] for the impaired speech and the compensated outputs from the three systems in BWE and PLC cases, respectively. In both tables, UP, DNN1, DNN2 and GANs stand for the unprocessed speech, speech processed by the two baselines DNN1 and DNN2, and by the proposed GANs system, respectively. SEEN means that the cut-off frequencies or the packet loss rates of the test data are the same as those in the training data while UNSEEN means that the cut-off frequencies or the packet loss rates of the test data are not included in the training set.

One can see from Table 1 that, the speech quality degrades more significantly as the cut-off frequency of the low-pass filter decreases (i.e., more high frequency component loss). In general, GANs outperform DNN1 for all metrics except LSD. This is because that the DNN-based systems were 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. Comparing the results by DNN1 and DNN2, one can see that the phase information is very important in improving the speech quality with respect to LSD and SNR while ignorable in PESQ and STOI. It may be because that the phase information is an important component to compute LSD and affects the energy distribution in frequency domain. However, even with the ideal phase information for reconstruction, the DNN2 performs worse than GANs in most cases except LSD.

As illustrated in Table 2, the more packets lost, the worse the speech quality with respect to all metrics. The superiority of GANs over DNN1 in PLC is more significant than in BWE. Even for LSD, GANs lost to DNN1 only at 40% and 50% packet loss cases. Furthermore, the packet loss compensation achieved by GANs is comparable to or better than DNN2 at low packet loss rates (0%~20%).

5. Conclusions

This study investigated the effectiveness of a GANs as a general framework for speech loss compensation including both BWE and PLC tasks. A set of experiments were carried out to evaluate the performance of the GANs-based system in comparison with two DNN-based systems. Different frequency and packet loss conditions were simulated in the experiments. Results show that the GANs obtained better speech quality and intelligibility than the DNN1 system for both seen and unseen speech loss conditions. Furthermore, the GANs achieved comparable performances to the DNN2 system in which the ideal phase information was assumed known for reconstruct the compensated speech.

6. Acknowledgements

This work is jointly supported by NSF of China (Grant No. 61771320), Shenzhen Science and Innovation Funds (Grant No. JCYJ 20170302145906843) and Guangdong Province with Grant No. 2018B030338001. Nengheng Zheng is the corresponding author.

7. References

[1] 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.

[2] 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.

[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] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*, pp. 4835-4839, 2017.

[14] D. L. Wang and J. T. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, 2018.

[15] K. H. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. ICASSP*, pp. 4395-4399, 2015.

[16] 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.

[17] 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.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.

[19] 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.

[20] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, "Speech bandwidth extension using generative adversarial networks," in *Proc. ICASSP*, pp. 5029-5033, 2018.

[21] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," In *INTERSPEECH*, 2017.

[22] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv*:1802.04208, 2018.

[23] S. Kim and V. Sathe, "Bandwidth extension on raw audio via generative adversarial networks," *arXiv preprint arXiv*: 1903.09027, 2019.

[24] 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.

[25] 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.

[26] 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.

[27] A. Pandey and D. L. Wang, "On adversarial training and loss functions for Speech Enhancement," in *ICASSP*, pp. 5414-5418, 2018.

[28] 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.

[29] 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.

[30] 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

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv*:1412.6980, 2014.

[32] 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.

[33] 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.

[34] 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.

[35] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks," *arXiv preprint arXiv*:1708.00853, 2017.