Investigating Generative Adversarial Networks based Speech Dereverberation for Robust Speech Recognition

Ke Wang[1,2], Junbo Zhang[2], Sining Sun[1], Yujun Wang[2], Fei Xiang[2], Lei Xie[1]*

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Xiaomi, Beijing, China

{kewang, snsun, lxie}@nwpu-aslp.org, {zhangjunbo, wangyujun, xiangfei}@xiaomi.com

Abstract

We investigate the use of generative adversarial networks (GANs) in speech dereverberation for robust speech recognition. GANs have been recently studied for speech enhancement to remove additive noises, but there still lacks of a work to examine their ability in speech dereverberation and the advantages of using GANs have not been fully established. In this paper, we provide deep investigations in the use of GAN-based dereverberation front-end in ASR. First, we study the effectiveness of different dereverberation networks (the generator in GAN) and find that LSTM leads to a significant improvement as compared with feed-forward DNN and CNN in our dataset. Second, further adding residual connections in the deep LSTMs can boost the performance as well. Finally, we find that, for the success of GAN, it is important to update the generator and the discriminator using the same mini-batch data during training. Moreover, using reverberant spectrogram as a condition to discriminator, as suggested in previous studies, may degrade the performance. In summary, our GAN-based dereverberation front-end achieves 14%∼19% relative CER reduction as compared to the baseline DNN dereverberation network when tested on a strong multi-condition training acoustic model.

Index Terms: Speech dereverberation, robust speech recognition, generative adversarial nets, residual networks

1. Introduction

The performance of automatic speech recognition (ASR) has been boosted tremendously in the last several years and state-of-the-art systems can even reach the performance of professional human transcribers in some conditions [1, 2]. However, room reverberation often seriously degrades the ASR performance, especially in far-field speech recognition where the talker is away from the microphone [3, 4]. Therefore, more attention has been paid recently in the research community to address this issue.

In theory, reverberant speech can be regarded as a room impulse response (RIR) convolving with the clean speech in the time domain [5]. A straightforward approach is called *speech dereverberation*, i.e., remove the reverberation from the contaminated speech. In this track, microphone array and multi-channel signal processing are very helpful [6, 7], but single-channel speech reverberation is still desirable in many real applications in which using multiple microphones may be impractical. Single-microphone speech dereverberation has been intensively studied in the signal processing community and a variety of approaches have been proposed [5, 8, 9, 10, 11]. Another approach to deal with reverberation (and noise) in speech recognition is multi-condition training (MCT), in which speech contaminated with reverberation, either simulated or real-recorded, is added in the acoustic model training set, letting the model learn the reverberation effects automatically. Although the above approaches are reasonably effective, it is still far away from claiming success in the fight against reverberation in speech recognition.

Recently, due to their strong regression learning abilities, deep neutral networks (DNNs) have been used in speech en-

hancement [12] and later in speech dereverberation [3, 4, 13]. The deep structure can be naturally regarded as a dereverberation filter that can learn the essential relationship between the reverberant speech and its counterpart, i.e., the clean speech, through a set of multi-condition data. Various deep structures, e.g., feed-forward [14], recurrent [15] and convolutional [16], have been explored in the field. Either direct spectral mapping [4, 14] or masking [17] can be considered in the dereverberation network. In the typical spectral mapping approach [12], the multi-condition data set used in the network training usually consists of pairs of reverberant and clean speech represented by log-power spectra (LPS). Note that in speech recognition, the output of the dereverberation network can be features like FBanks or MFCCs, which do not need to be inverted back to waveforms.

All the above DNN-based speech dereverberation approaches aim to minimize the mean square error (MSE) between the outputs of network and the ground truth. Hence, there is an underlying hypothesis that the enhanced speech has the minimal value in the MSE loss with the referenced clean speech. However, the MSE objective function has very strong implicit assumptions, e.g., independence of temporal or spatial, equal importance of all signal samples, and inaccurate to describe the degree of signal fidelity [18]. To remedy this deficiency, generative adversarial networks (GANs) [19], which consist of a generator network (G) and a discriminator network (D), learned through a min-max adversarial game, might be a good choice. Specifically, Pascual *et al.* have recently demonstrated the promising performance of GAN in speech enhancement [20] in the presence of additive noise. In the SEGAN approach [20], the generator G tries to learn the distribution of the clean data and generate enhanced samples from noisy speech to fool the discriminator D; while D aims to discriminate between the clean and enhanced samples (generated from G), which captures the essential difference between them. While SEGAN works on the waveform level, which targets to improve the perceptual speech quality, Donahue *et al.* [21] have explored GAN-based speech enhancement for robust speech recognition. Specifically, in [21], GAN works on the log-Mel filter-bank spectra instead of waveforms. The results have shown that GAN enhancement improves the performance of a clean-trained ASR system on noisy speech but it falls short of the performance achieved by conventional MCT. By appending the GAN-enhanced features to the noisy inputs and retraining, a 7% WER improvement relative to the MTR system was achieved.

While the major goal of the above GAN approaches is to remove additive noises, in this paper, we investigate the use of GANs in the mapping-based speech dereverberation for robust speech recognition. Although the same framework can be borrowed from these previous studies, we provide a series of deep investigations in the use of dereverberation front-end in ASR. First, we study the effectiveness of different dereverberation networks (used later as the GAN generator) and find that LSTM dereverberating network can achieve superior speech recognition performance as compared with feed-forward DNN and CNN. Second, further adding residual connections in the deep LSTMs can continuously boost the performance. Finally, we find that it is important to update the generator G and the discriminator D using the same mini-batch data during train-

---

*Corresponding author

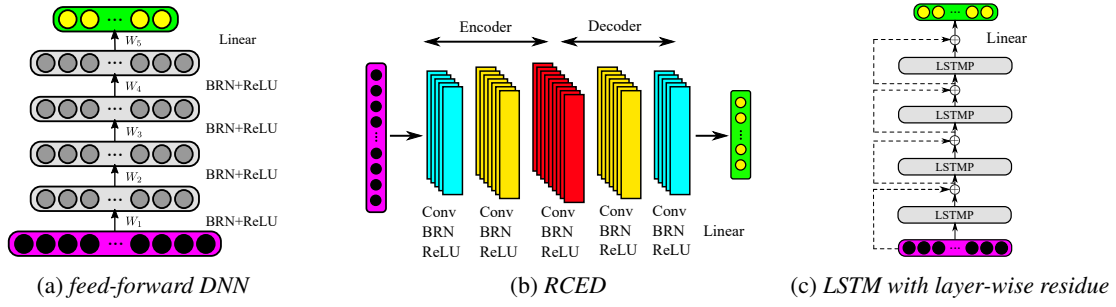(a) *feed-forward DNN* (b) *RCED* (c) *LSTM with layer-wise residue*

Figure 1: *Architectures of different dereverberation networks.*
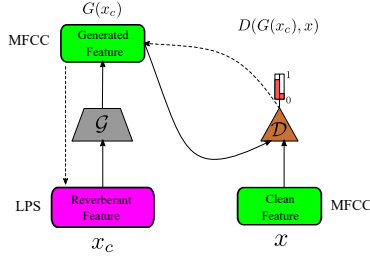

Figure 2: *GAN based speech dereverberation framework.*

ing for the success of GAN. Moreover, we discover that, using reverberant spectrogram as a condition to D, as suggested in [21, 22], may degrade the performance of G. In summary, using the dereverberation GAN can achieve 14%~19% relative character error rate (CER) reduction as compared with the DNN dereverberation baseline when tested on a strong multi-condition training acoustic model.

2. Mapping based speech dereverberation

Speech dereverberation can be achieved by a typical mapping approach [12], in which a regression DNN (shown in Fig. 1a) is trained by pairs of reverberant and clean LPS and a linear activation function at the output of DNN is adopted instead of a nonlinear one. Moreover, the target LPS feature is usually normalized globally over all training utterances into zero mean and unit variance (CMVN). In the dereverberation stage, the LPS features of input speech are fed into the well-trained regression DNN to generate the corresponding enhanced LPS features. Finally, the dereverberated waveform is reconstructed from the predicted spectral magnitude and the reverberant speech phase with an overlap-add algorithm.

Besides LPS, the input and output of the dereverberation DNN can be other speech features, e.g., MFCC and FBank. The speech features do not need to be inverted back to waveforms, when used for robust ASR. In [23], results show that the mapping from LPS to MFCC can achieve lower word error rate than the mapping from MFCC to MFCC in a speech recognition task under additive noise conditions. This also indicates that the transformation for different feature domains and nonlinear dereverberation function can be learned by the neural network simultaneously. Furthermore, as shown in Fig. 1b and 1c, CNN and LSTM can be used as enhancers as well. We expect that these more powerful network structures can bring further improvements in the speech dereverberation task. We will elaborate the network configurations and evaluate the performances of different networks later in Section 4.3.

3. Dereverberation GAN

3.1. GAN

Generative adversarial networks (GANs) [19] are generative models implemented by two neural networks competing with each other in a two-player min-max game. Specifically, the generator network G tries to learn a distribution $P_g(x)$ over data x and a prior input noise variables $p_z(z)$. The aim is to match the true data distribution $P_{data}(x)$ to fool the discrimi-

nator D. The discriminator network D serves as a binary classifier which aims to determine the probability that a given sample comes from the real dataset rather than G. Because of the weak guidance, the vanilla generative model cannot generate desirable samples. Hence the conditional GAN (CGAN) [24] was proposed to steer the generation by considering extra information x_c with the following objective function:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{data}(x, x_c)}[logD(x, x_c)]$$
$$+ \mathbb{E}_{x_c \sim p_{data}(x_c), z \sim p_z(z)}[log(1 - D(G(z, x_c), x_c))]. \quad (1)$$

In order to stabilize training and increase the quality of the generated samples in G, least-squares GAN (LSGAN) [25] was further proposed and the objective function changes to

$$\min_D V(D) = \frac{1}{2}\mathbb{E}_{x \sim p_{data}(x, x_c)}[(D(x, x_c) - 1)^2]$$
$$+ \frac{1}{2}\mathbb{E}_{x_c \sim p_{data}(x_c), z \sim p_z(z)}[D(G(z, x_c))^2], \quad (2)$$

$$\min_G V(G) = \frac{1}{2}\mathbb{E}_{x_c \sim p_{data}(x_c), z \sim p_z(z)}[(D(G(z, x_c)) - 1)^2]. \quad (3)$$

3.2. Speech dereverberation with GAN

It is straightforward to use GAN in speech dereverberation and Fig. 2 illustrates such a kind of architecture. It consists of a G and a D, where G, serving as the mapper in conventional methods, tries to learn a transformation from reverberant speech to clean speech and D tries to determine whether the input samples come from $G(x_c)$ or real-data x. Similar to [23], G aims to learning a mapping from the LPS feature input to the MFCC feature output which can be directly used in ASR. In some works [21, 22], the latent code z is excluded from the generator G to learn a direct mapping instead of a diversified translation in the original image-to-image translation task [26]. We borrowed this idea, but we remove the reverberant spectrogram as a condition to D. As we will report in Section 4.5, the added reverberant spectrogram as a condition to D not only increases the parameter size of D, but also degrades the performance of G. Therefore, we learn a generator distribution $P_g(x)$ over the conditional data $P_{data}(x_c)$ with the following proposed objective function:

$$\min_D V(D) = \frac{1}{2}\mathbb{E}_{x \sim p_{data}(x)}[(D(x) - 1)^2]$$
$$+ \frac{1}{2}\mathbb{E}_{x_c \sim p_{data}(x_c)}[D(G(x_c))^2], \quad (4)$$

$$\min_G V(G) = \frac{1}{2}\mathbb{E}_{x_c \sim p_{data}(x_c)}[(D(G(x_c)) - 1)^2]. \quad (5)$$

To further improve the ability of the adversarial component, previous CGAN approaches have indicated that it is beneficial to mix the GAN objective function with some numerical loss functions [24]. We follow this approach in the dereverberation GAN approach and the MSE loss is controlled by a new hyperparameter λ. Finally Eq. (5) becomes

$$\min_G V(G) = \frac{1}{2}\mathbb{E}_{x_c \sim p_{data}(x_c)}[(D(G(x_c)) - 1)^2]$$
$$+ \frac{1}{2}\lambda\mathcal{L}_{\text{MSE}}(G(x_c), x). \quad (6)$$

In practice, the generator G can be a feed-forward network, a convolutional network or a LSTM RNN network, as we de-

Table 1: *CERs (%) of Clean and MCT acoustic models.*

AM \ Test	Clean	Real	Simu
Clean	7.86	23.85	20.24
MCT	7.81	16.02	13.99

Table 2: *CERs (%) of different front-end networks.*

Input	Method	Clean AM		MCT AM	
		Real	Simu	Real	Simu
MFCC	DNN	17.86	16.63	16.31	14.72
	RCED	18.28	16.73	16.76	15.09
	LSTM	15.38	13.37	14.21	12.46
LPS	DNN	16.62	15.33	15.35	14.03
	RCED	15.55	14.15	14.15	13.09
	LSTM	15.04	13.16	13.97	12.20

Table 3: *CER (%) comparisons for different layers and residual connection architectures.*

Method	Clean AM		MCT AM	
	Real	Simu	Real	Simu
2-layer LSTM	**15.41**	**13.50**	**14.25**	**12.55**
+ Res-I	16.18	14.41	14.99	13.06
+ Res-L	16.13	13.74	14.61	12.65
4-layer LSTM	15.04	13.16	13.97	**12.20**
+ Res-I	15.81	13.48	14.60	12.47
+ Res-L	**14.99**	**13.13**	**13.90**	12.22
8-layer LSTM		divergence		
+ Res-I	15.53	13.55	14.48	12.49
+ Res-L	**14.67**	**12.75**	**13.62**	**12.04**

scribed in Section 2. Note that the discriminator D is only used in the training and discarded in the dereverberation stage. In our approach, a 2-layer LSTM without residual connection is set to be the architecture of D.

4. Experiments and results

4.1. Datasets

In the experiments, we used a Mandarin corpus as our source of clean speech data, which consists of 103,000 utterances (about 100hrs). The RIRs were from [27], including real-recorded RIRs and simulated RIRs for small, medium and large rooms. We randomly selected 97,000 utterances for network training and 3000 utterances for validation, and convolved with the RIRs (both real-recorded and simulated) to obtain the reverberant utterances. The rest 3000 utterances were used for testing and convolved with the real RIR and the simulated RIRs for small, medium and large rooms. Finally we obtained a testing set named 'Real' that contains 3000 reverberant speech utterances convolved with real RIRs and another testing set named 'Simu' that contains 9000 reverberant speech utterances convolved with simulated RIRs (3000 for small/medium/large). To test the generalization ability of our approach, we ensured the RIRs used for training and testing were totally different. All waveforms were sampled at 16 kHz. We used Kaldi [28] to generate the reverberant speech by convolving the clean signal with the corresponding RIR. As for feature extraction, the frame length was set to 25 ms with a frame shift of 10 ms.

4.2. ASR back-end

Our speech dereverberation front-end was used for speech recognition experiments. We used Kaldi to train our back-end ASR system with the similar acoustic model architecture and features in [29]. The original training dataset consists of 1600 hrs Mandarin speech data. We used speed-perturbation and volume-perturbation techniques [30] to do data augmentation. Hence the clean model were trained using 4800 hrs of speech data (1600×3). We also trained an acoustic model (AM) using multi-condition training (MCT) strategy. The training data for the MCT model is 6400 hrs (1600×4), including the above 4800 hrs of clean data and 1600 hrs of reverberant data generated by convolving the clean data with the RIRs in [27] as the dereverberation front-end.

The time delay neural network (TDNN) acoustic model (AM) had 6 layers, and each layer had 850 rectified linear units (ReLUs) with batch renormalization (BRN) [31]. The input contexts of TDNN AM were set to $[-2,2]$-$\{-1,2\}$-$\{-3,3\}$-$\{-7,2\}$-$\{-3,3\}$-$\{0\}$ and the output softmax layer had 5795 units. The notation $[-2,2]$ means we splice together frames $t-2$ through $t+2$ at the input layer and the notation $\{-1,2\}$ means we splice together the input at the current frame minus 1 and the current frame plus 2. The input of the AM was 40-dimensional MFCC. All the speech dereverberation front-ends were tested on both Clean and MCR AMs. A trigram language model (LM), which was trained on about 2 TB scripts with more than 100,000 words in the vocabulary, was used for decoding in the experiments. We also used entropy-based parameter pruning [32] and the threshold was set to be 10^{-8}.

The baseline results of Clean and MCT model are shown in Table 1. We can see a significant increase in CER when speech is contaminated with reverberations. In extending the training data of acoustic model by adding reverberant speech, the MCT AM can greatly reduce CER.

4.3. Mapping-based speech dereverberation

We first investigated the speech dereverberation performances of different networks and input features in the mapping-based

approach. Later we will select the best network as the generator in the GAN-based speech dereverberation. Specifically, we tested three different dereverberation networks, i.e., feed-forward DNN, redundant convolutional encoder decoder (RCED) and LSTM. As shown in Fig. 1, the DNN has 4 hidden layers and each of which contains 1024 ReLU neurons. The structure of the RCED is similar with [16] except the last layer. We changed the last filter CNN layer to a fully connected output layer as shown in Fig. 1b, because our input and target features were not in the same dimension. The input feature contains a context window of 11 frames ($t\pm5$) for the DNN and the RCED. The number of filters and filter width of RCED model were set to 12-16-20-24-32-24-20-16-12 and 13-11-9-7-7-7-9-11-13, respectively. The learning rate was set to 0.001 with a mini-batch size of 256. Moreover, BRN was also used for DNN and RCED training. Instead of using vanilla LSTM, we adopted an LSTM with recurrent projection layer (LSTMP) [33], which means we do not need to add an extra layer to do residual add like sDNN2 in [34] to avoid dimension mismatch. The LSTM has 4 LSTMP layers followed by a linear output layer. Each LSTMP layer has 760 memory cells and 257 projection units and the input to the LSTM is a single acoustic frame. The learning rate was set to 0.0003 and the model was trained with 8 full-length utterances parallel processing.

All the models explored here were optimized with the Adam [35] method and initialized with the Xavier [36] algorithm. We also used exponential decay to decrease the learning rate which was similar with Kaldi nnet3[1] and the terminated learning rate was 5 orders of magnitude smaller than the initial learning rate.

In Table 2, we list all experimental results on both Clean and MCT AMs. Firstly, we observe consistent improvement on all dereverberation networks by replacing MFCC with LPS features as the network input. Here the LPS feature is 257 dimension and the MFCC feature is 40 dimensions. Note that the output of all the dereverberation networks is 40-dimension MFCC which is fed into the ASR system. This conclusion is consistent with that in [23], where LPS performs better than MFCC when used as the input of a denoising network.

When we compare Table 2 with Table 1, we can find that the mapping-based dereverberation works quite well. When tested on the Clean AM, all the dereverberation networks are effective with significant CER reduction; when tested on the MCT

---

[1]egs/wsj/s5/steps/libs/nnet3/train/common.py(get_learning_rate)

Table 4: *CERs (%) comparisons by previous mapping based networks and our proposed framework. "DB" means we use different mini-batch data to update the parameters of GAN and "CD" means we add the conditional information to input of D. Relative improvements are given in parentheses w.r.t. the corresponding DNN model.*

Method	Clean AM		MCT AM	
	Real	Simu	Real	Simu
SEGAN	32.98 (−98.44)	37.14 (−142.27)	30.18 (−96.61)	32.37 (−130.72)
DNN	16.62 (0.00)	15.33 (0.00)	15.35 (0.00)	14.03 (0.00)
LSTM	15.04 (9.51)	13.16 (14.16)	13.97 (8.99)	12.20 (13.04)
+ Res	14.99 (9.81)	13.13 (14.35)	13.90 (9.45)	12.22 (12.90)
+ GAN	**14.07 (15.34)**	12.02 (21.59)	13.15 (14.33)	11.42 (18.60)
+ GAN+Res	14.10 (15.16)	**11.96 (21.98)**	**13.14 (14.40)**	**11.40 (18.75)**
+ GAN+Res (DB)	15.72 (5.42)	13.95 (9.00)	14.60 (4.89)	12.83 (8.55)
+ GAN+Res+CD	14.27 (14.14)	12.19 (20.48)	13.38 (12.83)	11.43 (18.53)

AM, the dereverberation networks with the LPS input are still effective with apparent CER reduction. Comparing different model structures, we discover that LSTM achieves the best performance. For instance, the LPS-LSTM dereverberation network reduces the CER from 23.85% (real-reverberation added) to 15.04% for the Clean AM and reduces the CER from 16.02% (real-reverberation added) to 13.97% for the MCT AM. We believe that the superior peformance is because of the LSTM's ability to model long-term contextual information that is essential is the speech dereverberation task. We also find RCED-CNN is not good when MFCC is used as the input. We will use LSTM as our network in the rest of the experiments.

4.4. Adding ResNet

Table 3 shows the results of different residual connection architectures. The layer-wise residual connection (Res-L) structure can be seen in Fig. 1c; while the input residual connection (Res-I) structure is similar with Res-L and more details can be found in [37]. As we expected, it's not necessary to add residual connections to shallow networks. Performances degrade when residual connections are used in a 2-layer LSTM. Res-L always performs better than Res-I. This is reasonable because Res-L tries to learn the residue of the high-level abstract feature while Res-I just learns the residue of the input feature. When the LSTM is as deep as 4 layers, Res-L starts to work and the lowest CERs are achieved when the LSTM has 8 layers. As training a 8-layer LSTM is time-consuming, we perform the GAN experiments with a 4-layer LSTM generator in the following.

4.5. Speech dereverberation with GAN

We finally investigated the ability of GAN in mapping-based speech dereverberation. We also reproduced the SEGAN approach [20] with the open-source codes[2] as a comparison. As shown in Table 4, SEGAN degrades the ASR performance within our expectation, which is consistent with the reported results in [21]. We believe this is because SEGAN aims to improve the perception of noisy speech and time-domain enhancement may be not appropriate for reverberant speech recognition.

In the proposed GAN-based methods, the architecture of G is consistent with that in Fig. 1c with 4 hidden LSTMP layers. The architecture of D is similar with G but contains only 2 LSTMP layers and the cell number and the projection dimension are set to 256 and 40, respectively. The hyper-parameter λ in Eq. (6) was set to 200 and the learning rate of G and D were set to 0.00008 and 0.0003, respectively. In each iteration, we updated the parameters of G twice and the parameters of D once. To stabilizing GAN training, we also add instance Gaussian noise to the MFCC input of D[3]. In Table 4, we demonstrate that using GAN (in LSTM+GAN) is not only viable but also outperforms the LSTMs.

At the early stage of our experiments, we updated the parameters of G and D using different mini-batch data like the ways they do in image tasks. In other words, the parameters

of D were updated using one mini-batch data and then the parameters of G were updated using a new mini-batch data. We found that this training strategy—LSTM+GAN+Res (DB) in the second-to-last row of Table 4—was quite unstable in our experiments and we always achieved results worse than the non-adversarial training (e.g., LSTM+Res) as shown in Table 4. Instead, when we tried to update the parameters of G and D using the same mini-batch data, we achieved consistently better results (LSTM+GAN+Res in Table 4). We believe that this strategy is essential in making our GAN approach performing well. Adding residual connections works for most cases. LSTM+GAN+Res lowered the MCT AM CER from 15.35% down to 13.14% with 14.4% relative CER reduction for the Real set and lowered the MCT AM CER from 14.03% down to 11.40% with 18.75% relative CER reduction for the Simu set. Finally, we also find the performance of LSTM+GAN+Res+CD is worse than LSTM+GAN+Res. This means that adding the reverberant spectrogram as a condition to D is useless to the dereverberation performance.

5. Summary

In this paper, we provide a deep investigation of GAN in mapping-based speech dereverberation for robust speech recognition. In the selection of the generator network, we find that LSTM achieves superior performance, while adding residual connections (ResNets) in deep LSTMs can further boost the performance. In the use of GAN, we find that it is essential to update the generator and the discriminator using the same mini-batch data during model training; and using reverberant spectrogram as a condition to the discriminator may degrade the performance. With the above findings, we are able to achieve 14%∼22% relative CER reduction in ASR as compared with a DNN baseline, while the SEGAN baseline even does not work on the ASR task. In the future, we plan to further explore the use of GAN in more adverse conditions (both reverberant and noisy) and try to combine the framework with joint-training strategy to further improve the ASR performance.

6. Acknowledgements

7. References

[1] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," *arXiv preprint arXiv:1708.06073*, 2017.

[2] G. Kurata, B. Ramabhadran, G. Saon, and A. Sethy, "Language modeling with highway lstm," *arXiv preprint arXiv:1709.06436*, 2017.

[3] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani,

---

[2]https://github.com/santi-pdp/segan

[3]http://www.inference.vc/instance-noise-a-trick-for-stabilising-gan-training/

B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[4] B. Wu, K. Li, M. Yang, C.-H. Lee *et al.*, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 1, pp. 102–111, 2017.

[5] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, 1979.

[6] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1791–1801, 2007.

[7] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[8] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[9] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[10] S. Mosayyebpour, M. Esmaeili, and T. A. Gulliver, "Single-microphone early and late reverberation suppression in noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 322–335, 2013.

[11] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 2, pp. 276–289, 2016.

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

[13] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Interspeech*, 2017, pp. 384–388.

[14] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[15] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4623–4627.

[16] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[17] D. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[18] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[20] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[21] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," *arXiv preprint arXiv:1711.05747*, 2017.

[22] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *arXiv preprint arXiv:1709.01703*, 2017.

[23] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[25] 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.

[26] 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.

[27] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition." ICASSP, 2017.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[29] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts." in *INTERSPEECH*, 2015, pp. 3214–3218.

[30] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *INTERSPEECH*, 2015, pp. 3586–3589.

[31] S. Ioffe, "Batch renormalization: Towards reducing mini-batch dependence in batch-normalized models," *arXiv preprint arXiv:1702.03275*, 2017.

[32] A. Stolcke, "Entropy-based pruning of backoff language models," *arXiv preprint cs/0006025*, 2000.

[33] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[34] M. Tu and X. Zhang, "Speech enhancement based on deep neural networks with skip connections," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5565–5569.

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

[36] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[37] Z. Chen, Y. Huang, J. Li, and Y. Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection," in *Proc. Interspeech*, 2017.