INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Enhancement for Noise-Robust Speech Synthesis using Wasserstein GAN

Nagaraj Adiga[†], *Yannis Pantazis*[*], *Vassilis Tsiaras*[†], *and Yannis Stylianou*[†]

[†]Department of Computer Science, University of Crete, Greece
[*]Institute of Applied and Computational Mathematics, FORTH, Greece

nagaraj@csd.uoc.gr, pantazis@iacm.forth.gr, tsiaras@csd.uoc.gr, yannis@csd.uoc.gr

Abstract

The quality of speech synthesis systems can be significantly deteriorated by the presence of background noise in the recordings. Despite the existence of speech enhancement techniques for effectively suppressing additive noise under low signal-to-noise (SNR) conditions, these techniques have been neither designed nor tested in speech synthesis tasks where background noise has relatively lower energy. In this paper, we propose a speech enhancement technique based on generative adversarial networks (GANs) which acts as a preprocessing step of speech synthesis. Motivated by the speech enhancement generative adversarial network (SEGAN) approach and recent advances in deep learning, we propose to use Wasserstein GAN (WGAN) with gradient penalty and gated activation functions to the autoencoder network of SEGAN. We studied the impact of the proposed method on a data set consisting of 28 speakers and different noise types with 3 different SNR level. The effectiveness of the proposed method in the context of speech synthesis is demonstrated through the training of WaveNet vocoder. We compare our method against SEGAN. Both subjective and objective metrics confirm that the proposed speech enhancement approach outperforms SEGAN in terms of speech synthesis quality.

Index Terms: Wasserstein GAN, Speech Enhancement, Gated activation, WaveNet Vocoder, Speech Synthesis.

1. Introduction

Speech enhancement techniques aim to improve the quality of speech by reducing background noise. They are crucial components, either explicitly or implicitly, in automatic speech recognition (ASR) systems. Their main target is to improve the intelligibility of speech and consequently increase the noise robustness of ASR. Besides the standard signal processing approaches such as spectral subtraction and Wiener filter [1, 2], deep neural networks have been widely adopted to either directly reconstruct clean speech [3] or estimate masks [4] from the noisy signals. Different types of networks have also been investigated in the literature for enhancement, such as autoencoders [5], convolution networks [6] and recurrent networks [7]. Motivated by these studies we present a neural network-based speech enhancement technique to assist speech synthesis under noisy conditions.

Typically, the recordings for speech synthesis are performed under controlled conditions with minimum distortions from reverberation or environmental noise. However, there are (at least) two scenarios where the environment cannot be controlled: (a) when recordings are performed using mobile phones and (b) when voices with no high-quality recordings are of interest but new recordings cannot be performed. Therefore, a speech enhancement (SE) method tailored for speech synthesis (SS) and speech adaptation in mid (15-30 dB) signal-to-noise (SNR) conditions is much needed.

The classic approaches like spectral subtraction or Wiener filtering methods work well in high (30-50 dB) SNR conditions. However, their success is limited in mid SNR values. Botinhao *et al.* [8] proposed recently an SE technique for noise robust speech synthesis based on recurrent networks. However, this technique operates in feature domain instead of waveform domain resulting in the implicit introduction of vocoding quality in the enhanced speech. Additionally, this technique was tested in low (0-15 dB) SNR conditions. In a different direction, the transition from feature domain to waveform domain triggered great improvements in tasks like speech synthesis and voice conversion with speech quality almost similar to natural recordings [9, 10, 11, 12].

In [13], authors proposed an SE method at the waveform domain with impressive results in adverse additive noise scenarios. They trained a neural network system using the generative adversarial networks (GANs) framework. The architecture of this method is an autoencoder neural network with skip connections trained on multiple speakers and multiple noise conditions. The speech enhancement using generative adversarial network (SEGAN) model is an end-to-end system without the need for hand-crafted feature extraction. However, it is designed for ASR purposes and when tested on conditions often occurring in SS tasks we observe either minimal or no improvement or even slight deterioration in speech quality. Henceforth, we propose a variation of SEGAN which is dedicated for SS applications. First, we keep the processing at the waveform level but we adopt a different loss function. The original SEGAN used a least-squares type of loss function. We propose to use Wasserstein distance with gradient penalty (WGAN) [14] which has shown to produce superior results. Second, we apply an activation function that is more suitable for speech signals. It has been shown [9] that a data-driven gated linear unit (GLU) activation function is capable of generating very good speech quality. Therefore, we propose GLUs as activation functions.

We additionally evaluate the benefit of WGAN and GLUs in SE. We show the impact on speech synthesis when background noise takes mid-range SNR values. In order to limit the confounding factors in our study, we use an existing WaveNet model trained on clean speech data to measure the effectiveness of the proposed enhancement method. To quantify the performance in SS, we first train a WGAN model to map simulated noisy speech to the original clean speech. Then, we measure the performance of the WaveNet model on noisy speech before and after enhancement by WGAN. Our experiments indicate that the proposed WGAN approach improves speech synthesis performance under these noise conditions. Using the voice bank corpus for training, we show on both objective metrics and subjective listening tests that both Wasserstein loss function

http://dx.doi.org/10.21437/Interspeech.2019-2648

with gradient penalty and gated activation function significantly improve the quality of speech for both SE and SS tasks under medium SNR conditions.

The rest of this paper is organized as follows. In Section 2, we describe WGAN with GLUs and contrast it with SEGAN. The application of SE in SS tasks as well implementation details are explained in Section 3. In Section 4, results from objective and subjective evaluations are presented. Finally, the paper is concluded in Section 5.

2. Speech Enhancement using Wasserstein GAN

GANs are popular generative models that have been initially used in synthesis and enhancement of images. GANs consist of a pair of two neural networks: the generator (G) which generates output data given some source input and the discriminator (D) which discriminates between real and generated data. GANs have produced impressive results in terms of naturalness and they have been subsequently applied to other fields such as speech and video processing. In the context of SE, Pascual *et al.* [13] proposed a speech enhancement generative adversarial network (SEGAN) approach to enhancing speech at the waveform level. The generator network is inspired by the autoencoder architecture supplemented with skip connections that link the layers of the encoder to the layers of the decoder. The input noisy speech x_n is passed through the encoder of G to get a bottleneck feature vector c, which is then concatenated with the noise latent vector z to fed the decoder of G. SEGAN uses waveforms with the long receptive field (16384 samples at 16 kHz) with pairs of clean x_c and noisy speech x_n and is trained to output clean speech $\tilde{x_c}$. The loss function for training SEGAN is the least squares error for both generator and discriminator networks. An additional l_1 similarity norm regularizes the training of G.

2.1. Wasserstein GAN with Gradient penalty

Wasserstein GAN (WGAN) has been already shown to improve the training in image processing [15]. In WGAN, the discriminator function is required to lie in the space of Lipschitz continuous functions with Lipschitz constant equals to 1. This property can be enforced either through weight clipping or with gradient penalty. Weight clipping has optimization issues, whereas, authors in [14] showed that gradient penalty gave more stable training. Hence, we used the improved WGAN with gradient penalty. The loss function for D with Wasserstein distance and gradient penalty is defined as follows:

$$\begin{aligned} \min_D L_W(D) &= \mathbb{E}_{(x_c,x_n)\sim p_d(x_c,x_n)}[(D(x_c,x_n))] \\ &- \mathbb{E}_{z\sim p_z(z),x_n\sim p_d(x_n)}[(D(G(z,x_n),x_n))] \\ &+ \lambda_1 \mathbb{E}_{z\sim p_z(z),\tilde{x}\sim p_d(\tilde{x})}[(||\nabla_{\tilde{x}} D(\tilde{x})||_2 - 1)^2] \end{aligned} \tag{1}$$

where $\tilde{x} \sim p_d(\tilde{x})$ is defined uniformly along straight lines between pairs of points samples from the data distribution $p_d(x_c)$ and $p_d(\tilde{x_c})$. Parameter λ_1 control the gradient term. Similarly, the loss function for G is given by:

$$\begin{aligned} \min_G L_W(G) &= -\mathbb{E}_{z\sim p_z(z),x_n\sim p_d(x_n)}[D(G(z,x_n),x_n)] \\ &+ \lambda_2 \mathbb{E}_{z\sim p_z(z),(x_c,x_n)\sim p_d(x_c,x_n)}[\| G(z,x_n) - x_c \|_1] \end{aligned} \tag{2}$$

where λ_2 is a hyper-parameter which controls the similarity term in the loss function of G.

2.2. Gated activation function

We utilize strided convolutional neural networks (CNN) to construct both the discriminator and the generator networks. After each layer, either regular rectified linear units (ReLUs) or parametric ReLUs are used as non-linearities. However, we propose to use gated linear units (GLUs) activation function. GLUs have been reported to capture both long- and short-term dependencies which are present in speech [16]. The output of the l^{th} hidden layer of a gated CNN is described as:

$$H_l = (H_{l-1} * W_l) \otimes \sigma(H_{l-1} * V_l) \tag{3}$$

where σ is the sigmoid function and $\otimes$ represents element-wise product. W_l and V_l are parameters to be learned during training.

2.3. Comparison between WGAN and SEGAN

SEGAN enhances noisy signals when their SNR is low (0-15 dB). Since we are interested in speech synthesis application, the tested SNR values will be higher (15-30 dB). As a first demonstration, we plot in Figure 1 the spectrograms from clean, noisy, SEGAN-enhanced-, and WGAN-enhanced speech using one sentence taken from the Voice bank corpus [17]. The noisy signal is contaminated with additive noise of 20 dB SNR. The spectrogram visualization reveals that WGAN method is closer to the original speech compared to the signal enhanced using SEGAN. The proposed method suppresses background noise in most of the region.

Table 1: *Global mean SNR for both SEGAN and WGAN methods at three different SNR values. Standard error is also reported.*

SNR level (dB)	SEGAN	WGAN
15	17.31 ±0.010	18.27 ±0.004
20	19.10 ±0.011	22.82 ±0.002
25	23.53 ±0.014	26.38 ±0.003

Furthermore, we perform a more systematic assessment of the performance of the WGAN approach relative to SEGAN. We measure the global SNR level after enhancing the noisy signals. The results of WGAN with gradient penalty and GLU activation function are shown in Table 1. The SNR reported in the Table is an average of approximately 250 files. From the Table, we observe that SEGAN fails to enhance speech for SNR level of 20 and 25 dB, whereas, WGAN improves the SNR in all levels. It is also worth noting that we performed an ablation test where we replace the GLUs with parametric ReLUs. We observe that the global SNR for WGAN is decreased by 1 to 1.5 dB. In the next section, we will describe the integration of the proposed speech enhancement module as a pre-filtering algorithm in speech synthesis.

3. Speech Enhancement for Speech Synthesis

In order to evaluate the impact of SE in SS applications, we use the proposed speech enhancement method as a pre-filtering module. Figure 2 shows the integration of the proposed speech enhancement module in speech synthesis. This framework has two stages: firstly, pre-processing the noisy speech using WGAN ans secondly, WaveNet vocoder is trained to produce high quality speech synthesis.

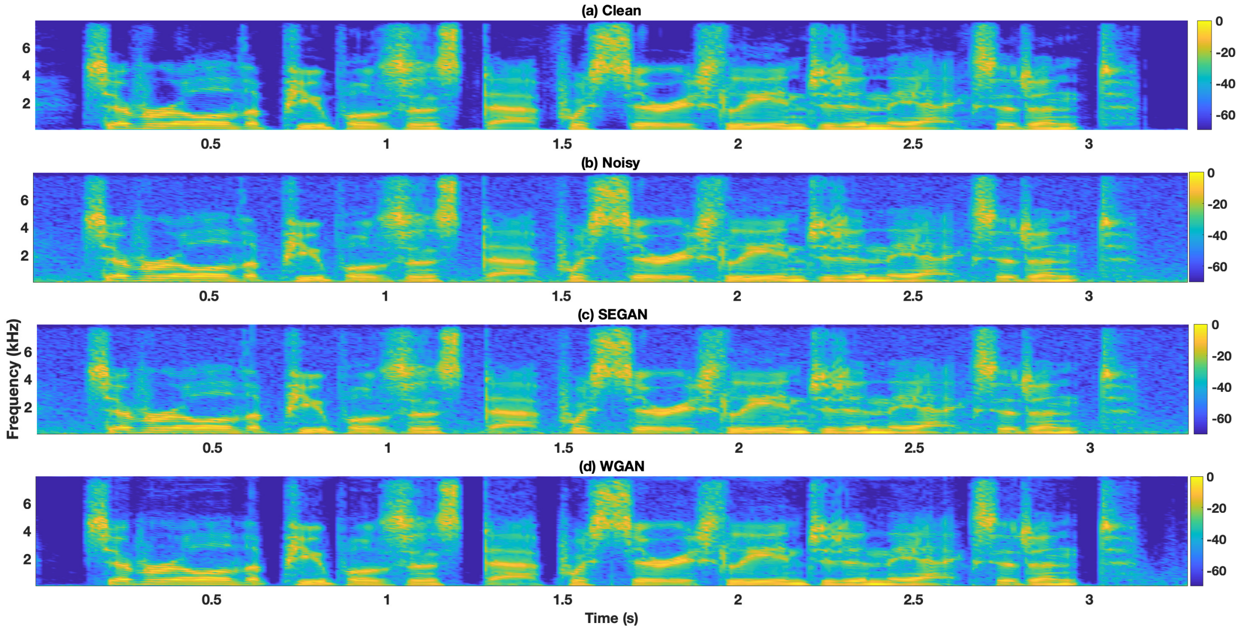

Figure 1: *Spectrogram view of (a) clean, (b) noisy, (c) SEGAN-enhanced, and (d) WGAN-enhanced waveforms.*

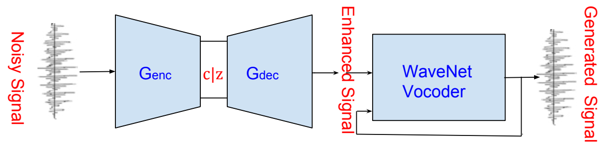

Figure 2: *Proposed neural network-based speech enhancement module for speech synthesis.*

The first stage is the implementation of WGAN network. WGAN generator's encoder (G_{enc}) consists of 11 layers of stride-2 convolution with growing depth, resulting in a feature vector c at the bottle-neck of 8 time-steps with depth 1024. This feature map is concatenated with latent vector z, sampled randomly from uniform noise distribution. The resultant concatenated vector is input to an 11-layer up-sampling decoder (G_{dec}), with skip connections from the corresponding input feature vector to preserve phase and alignment same as that of the original. The architecture of the discriminator network follows an architecture similar to that of generator's encoder. The only difference is that the discriminator takes two input (noisy and clean/enhanced speech) channels with 16384 samples in each batch.

The second stage consists of the neural-network vocoder based on WaveNet. In this work, a non-causal implementation of WaveNet vocoder is used. WaveNet uses a non-linear auto-regressive (AR) model to predict the next signal sample from the past and future samples of signal and local conditioning each sample from the time-varying acoustic features. We have used Mel-filter banks computed from short-time Fourier transform as an acoustic feature. The vocoder architecture has three main modules: a dilated convolution network with increasing dilation factor, which gives long receptive field, a stack of residual-blocks, which acts as a feature extractor, and a post-processing module, which combines the information from residual blocks to predict the next speech sample.

3.1. Experimental set-up

To train WGAN, we used 30 speakers (15 male and 15 female speakers) from voice bank corpus [18]. Each speaker utters approximately 400 sentences. We used 28 speakers for training and 2 speakers for testing. The original utterances are sampled at $48kHz$, which we down-sample to $16kHz$ for faster processing. We used noise signals from the BUTReverbDB database [19] to contaminate clean speech utterances. The noise files consist of recordings from silent office and conference rooms. The clean-noisy pairs are created with SNR levels of 15, 20 and 25 dB. Once the pairs have been created we train the SE models. Both WGAN and SEGAN are trained for 150 epochs with Adam optimizer. The learning rate used for WGAN's generator and discriminator network is 0.00005 and 0.000025, respectively. During training, waveform windows of 1 sec length (16384 samples) are feed to the networks. A shift of 0.5 sec is used. During testing, non-overlapping samples are fed to the generator network whose outputs are concatenated to obtain the enhanced utterance. The penalty terms λ_1 and λ_2 is set to 10 and 100, respectively. We experimented with several combinations of values for both hyper-parameters, however, we did not observe any significant change in the l_1-norm of the generator's loss function.

In order to verify the impact of the enhancement techniques in speech synthesis, we trained four WaveNet vocoder systems using clean, noisy, SEGAN-enhanced, and WGAN-enhanced testing utterances. WaveNet training requires at least 1 hr ($\sim$1000 utterances) of data to get good quality synthesis [20]. However, only 400 utterances for each speaker in the voice bank corpus are available. Hence, we used four additional speakers from CMU-ARCTIC database [17]; SLT, BDL, CLB, and RMS for training the four WaveNet systems. From the two testing speakers of the voice bank corpus, we used 300 utterances for training and 50 sentences for validation while the remaining 50 sentences were used for testing. The selected WaveNet architecture is similar to the one mentioned in [20] with the addition of non-causality function. Acoustic features were extracted every $5ms$ after applying a window of $25ms$. We used 45 Mel-filter-bank features computed from standard short-time Fourier magnitude spectrum for local conditioning the WaveNet. All signals were up-sampled to the same sampling rate ($16kHz$).

4. Results and Discussion

4.1. Objective evaluation

To know the quality of the enhanced speech by WGAN method different objective metrics is computed. Perceptual evaluation of speech quality (PESQ) (from -0.5 to 4.5) is a wideband version recommended in ITU-T P.862.2 [21] is measured to know the quality of enhanced speech. The short-time objective intelligibility (STOI) that records the improvement in speech intelligibility [22] is computed. The gain in SNR through the enhancement processing is being evaluated by segmental SNR (SSNR) [23]. The composite measure for signal distortion (CSIG) predicts the mean opinion score of the signal (MOS) distortion (from -0.5 to 4.5). The composite measure for background interferences (CBAK) and the overall composite measure (COVL) (from 1 to 5) predicts the extent of background noise suppression in the speech and the overall effect, respectively [24].

The objective metrics for SE task are evaluated on two testing speakers of voice bank corpus (1 male and 1 female speaker) with a total of 824 utterances. The results are shown as an average in Table 2. The relative standard error is of order 10^{-4} for all objective metrics. The objective performance gain of the proposed WGAN model can be observed when compared to SEGAN method. The PESQ score got increased by 0.2 while STOI with 0.02 factor. This indicates that the quality is being improved further without losing the intelligibility. The CSIG gain from 3.57 to 3.69 is a sign of clean speech restoration by the WGAN method compared to SEGAN. At the same time, noise suppression (CBAK) has been slightly better from 3.46 to 3.48 in the proposed method. Hence, the overall quality of the output speech (COVL) is got improved by 0.10. This is even more clear when we look into the segmental SNR gain through the processing. The SSNR has been increased around 1 dB by processing in WGAN comparison to the SEGAN method. This is a clear confirmation that the model with WGAN performed better for speech enhancement task in mid SNR conditions.

Table 2: *Objective evaluation results comparing the noisy signal and SE using SEGAN- and WGAN-based methods.*

Metric	Noisy	SEGAN	WGAN
PESQ	2.65	2.61	2.81
STOI	0.97	0.97	0.99
SSNR	13.32	14.30	15.87
CSIG	3.46	3.57	3.69
CBAK	3.02	3.46	3.48
COVL	3.32	3.54	3.64

4.2. Subjective evaluation

We conducted two listening experiments. The first experiment is to know the quality of the enhanced speech from the proposed WGAN based-SE method in mid SNR conditions. The second experiment is to know the impact of the presence of noise in the data used for training neural network-based vocoder. We evaluated the sound quality of the enhanced speech using a mean opinion score (MOS). The subjects rated the sound quality (naturalness) of the speech using a 5-point scale: "5" for perfect, "4" for very good, "3" for good, "2" for bad, and "1" for very bad. 20 non-native subjects participated in the listening experiment. Subjects have to listen to 10 sentences in each method and give their opinion score. Table. 3 shows the results of the MOS test. From the table, we first observe that WGAN-based SE is better than SEGAN-based SE system. However, there is a still scope to improve quality as clean speech has higher MOS score. The MOS of SS task under the mid SNR condition with WaveNet vocoder is given in the third column of Table 3. It is clear that WaveNet vocoder results are degrading when training data having background noise. This is reflected in dropping MOS from 3.57 in the case of clean speech to 2.01 in case of noisy speech. Further, when we use the proposed WGAN based SE algorithm MOS increased to 2.86. This indicates that the proposed neural network-based SE algorithm using WGAN can be used as a pre-filtering algorithm to remove additive noise present in recording data and to improve overall synthesis quality. Examples of speech files from both SE and SS experiments can be found in [1].

Table 3: *Naturalness MOS for SE and SS using SEGAN- and WGAN-based methods.*

Type of signal	SE	SS
Clean	4.51 ± 0.14	3.57 ± 0.20
Noisy	2.91 ± 0.18	2.01 ± 0.18
SEGAN	3.08 ± 0.19	2.20 ± 0.20
WGAN	4.06 ± 0.17	2.86 ± 0.19

5. Conclusion

In this work, we proposed a WGAN-based speech enhancement method for medium SNR conditions. The proposed model uses gradient penalty to stabilize the training of WGAN and gated activation functions. Both features resulted in significant improvements in speech enhancement. The objective metrics confirm that the proposed WGAN model is advantageous for speech enhancement in medium SNR values. Furthermore, the added value of the proposed SE approach is shown in speech synthesis task. Using WaveNet vocoder, we showed that WGAN model can be used as a pre-filtering step in a speech synthesis system. The pre-processing of the recorded speech files will decrease the background noise. The enhanced files allowed us to produce good quality speech using WaveNet vocoder for both male and female speakers. The proposed WGAN approach was supported by objective metrics of intelligibility and sound quality as well as by subjective listening test results. Future work includes moving from a parallel data set-up to a non-parallel set-up for speech enhancement in waveform domain.

[1] `https://www.csd.uoc.gr/~nagaraj/IS19.html`

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," *J. Acoust. Soc. A.*, vol. 125, no. 1, pp. 360–371, 2009. [Online]. Available: https://doi.org/10.1121/1.3023069

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

[4] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7092–7096.

[5] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

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

[7] A. Maas, Q. V. Le, T. M. O'neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," 2012.

[8] C. Valentini-Botinhao and J. Yamagishi, "Speech Enhancement of Noisy and Reverberant Speech for Text-to-Speech," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 8, pp. 1420–1433, Aug 2018.

[9] A. van den oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499*, 2016.

[10] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *arXiv preprint arXiv:1705.08947*, 2017.

[11] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[12] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model," *arXiv preprint arXiv:1703.10135*, 2017.

[13] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Interspeech*, 2017.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[15] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[16] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. on Machine Learning*, vol. 70. JMLR. org, 2017, pp. 933–941.

[17] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224. [Online]. Available: http://festvox.org/cmu_arctic/index.html

[18] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Int. Conf. Oriental COCOSDA held jointly with 2013 Conf. on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pp. 1–4.

[19] I. Szoke, M. Skacel, L. Mosner, J. Paliesek, J. Cernocky *et al.*, "Building and evaluation of a real room impulse response dataset," *arXiv preprint arXiv:1811.06795*, 2018.

[20] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of wavenet as a statistical vocoder," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Process.*, 2018, pp. 5674–5678.

[21] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. European Congress on Acoust.*, 2005, pp. 2725–2728.

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

[23] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.

[24] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *9th Int. Conf. on Spoken Lang. Process.*, 2006.