Analysis by Adversarial Synthesis - A Novel Approach for Speech Vocoding

Ahmed Mustafa[1,2,], Arijit Biswas[2], Christian Bergler[1], Julia Schottenhamml[1], Andreas Maier[1]*

[1]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Erlangen, Germany
[2]Dolby Germany GmbH, Nuremberg, Germany

ahmed.mustafa.ahmed@iis.fraunhofer.de, arijit.biswas@dolby.com
{christian.bergler, julia.schottenhamml, andreas.maier}@fau.de

Abstract

Classical parametric speech coding techniques provide a compact representation for speech signals. This affords a very low transmission rate but with a reduced perceptual quality of the reconstructed signals. Recently, autoregressive deep generative models such as WaveNet and SampleRNN have been used as speech vocoders to scale up the perceptual quality of the reconstructed signals without increasing the coding rate. However, such models suffer from a very slow signal generation mechanism due to their sample-by-sample modelling approach. In this work, we introduce a new methodology for neural speech vocoding based on generative adversarial networks (GANs). A fake speech signal is generated from a very compressed representation of the glottal excitation using conditional GANs as a deep generative model. This fake speech is then refined using the LPC parameters of the original speech signal to obtain a natural reconstruction. The reconstructed speech waveforms based on this approach show a higher perceptual quality than the classical vocoder counterparts according to subjective and objective evaluation scores for a dataset of 30 male and female speakers. Moreover, the usage of GANs enables to generate signals in one-shot compared to autoregressive generative models. This makes GANs promising for exploration to implement high-quality neural vocoders.

Index Terms: speech coding, deep learning, generative adversarial networks, neural vocoder.

1. Introduction

Speech coding is one of the fundamental functionalities of current multimedia communication systems over band limited transmission channels [1]. The conventional approaches for coding speech signals are based on the source-filter model, in which a speech signal is decomposed into its glottal excitation source signal and its vocal tract filter parameters [1]. Linear predictive coding (LPC) is used for implementing such source-filter modelling of speech signals [1]. This is incorporated in many speech coding standards such as AMR-WB [2]. Vocoding is the process of describing speech signals in a fully parametric manner [1]. This provides the ability to build speech synthesizers operating based on the acoustic parameters which represent the target speech signal. The classical vocoding methods are developed based on hand-crafted acoustic parameters that mainly replace the glottal excitation, e.g. F0, VUV, etc. However, the reconstructed speech waveform according to such methods is well known to be synthetic and low in perceptual quality.

Recently, autoregressive (AR) deep generative models have shown a great success in generating raw audio and speech waveforms especially after the emergence of WaveNet [3] and

SampleRNN [4]. Both WaveNet and SampleRNN have been used as neural vocoders for reconstructing speech signals from the hand-crafted parametric representation of the source filter model [5, 6]. The reconstructed speech signals using such neural vocoders are clearly high in their perceptual quality and even outperform the classical speech codecs such as AMR-WB. Unfortunately, this comes with the problem of very slow signal generation due to the sequential sampling of AR deep generative models. Therefore, additional techniques such as probability density distillation are required for running in real time [7].

Generative Adversarial Networks (GANs) provide an alternative approach for very fast generation of realistic data samples [8]. GANs learn implicitly to estimate the PDF underlying the original data in order to directly generate new samples. This is achieved by a minimax adversarial training between a generator network that creates fake data and a discriminator network that compares it to the original data [8]. When the training reaches an equilibrium state, the discriminator becomes fooled by the fake data created by the generator network, which is the target deep generative model.

The usage of GANs for speech vocoding is very recent and was firstly introduced by Bollepalli et al. [9] and Juvela et al. [10]. The main idea of such approaches for GAN-based vocoding is to generate the glottal excitation signal adversarially and then apply synthesis filtering to obtain the speech waveform. However, this incorporates recurrent neural network architectures for predicting the voicing information and building a pulse model before creating the excitation signal with GANs.

This paper proposes a new method for generating speech signal waveforms from a learned-compressed representation of the glottal excitation. The method uses GANs as an end-to-end fully-convolutional generative model that produces raw speech waveforms in one-shot. A simple refinement based on LPC is applied to the generated speech waveform in order to obtain a natural final reconstruction, without considerably affecting the overall complexity. An overview of this method is given in section 2, followed by a description of the experimental setup in section 3. Evaluation results are reported and discussed in section 4, to end up with conclusions in section 5.

2. Analysis by Adversarial Synthesis

Besides the ability of one-shot sample generation, GANs can create realistic data from a totally-abstract noise prior (e.g., Gaussian noise). The adversarial training makes it possible to map a simple prior distribution into complicated real-world distributions in a high-dimensional space. This has been achieved efficiently for image synthesis [11] and recently for general audio synthesis [12].

The speech vocoding task is described by a conditional generation process, so that the noise prior of the GAN model is con-

---

*The author was finishing his thesis with 1 and 2 during the writing of this paper. Currently, he is at Fraunhofer IIS in Erlangen.
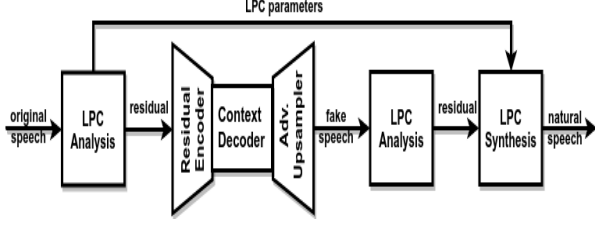
Figure 1: *Illustration of AbAS.*

verted into a parametric representation for the desired speech signal. To accomplish this, the glottal excitation signal, represented by the residual from an LPC analysis filtering of the speech waveform [1], is fed to a neural encoder network. The residual is a noise-like signal as it is uncorrelated and almost spectrally-flat [1]. Thus, it is a good candidate to be compressed by the encoder network. This results in a learned conditional noise prior, which is a characteristic representative for the speech signal.

2.1. Conditional Adversarial Synthesis

Using the learned compressed representation of the residual signal as an input, a fake speech waveform is created using a deep generative model implemented by GANs. The generator network consists of two main blocks. The first block is called the *context decoder*. It maps the conditional prior input (i.e., the context vector) into a multi-channel hidden representation to enable reliable signal generation. While the second block is called the *adversarial upsampler* and it learns to upsample the context decoder output until reaching the desired signal resolution. The cascade of the residual neural encoder with the generator network gives a conditional generator model, which is trained jointly with the discriminator model.

2.2. Cross Synthesis

The fake speech waveform generated with GANs contains the main global and prosodic features of the target signal, especially at the first formants. However, some of phonemes and local details are missed due to the fast progressive generation of the speech waveform. Moreover, the adversarial training procedure does not follow a maximum-likelihood approach as in the AR generative models. This results in a considerable amount of reconstruction artifacts that affect the perceptual quality of the fake signal. To solve this issue, we propose to replace the spectral envelope of the fake speech with the original spectral envelope. This is done by an LPC analysis applied to the fake speech signal to obtain its fake residual. The fake residual is then filtered by the LPC parameters of the original speech signal to obtain a natural signal reconstruction. Hence, we name the whole process *Analysis by Adversarial Synthesis* (AbAS), which is illustrated in Figure 1.

3. Model Configuration and Training Setup

For training and testing the generative model, we used the clean speech signals of the dataset created by Valentini et al. [13]. It is an open source dataset of 15 male and 15 female speakers selected from the Voice Bank corpus introduced by Veaux et al. [14]. The training data is constructed by the speech signals of 28 speakers, divided equally between males and females. While the test data are represented by speech of the remaining two

speakers. The speech signals are downsampled from the original sampling rate of 48 kHz to 16 kHz which is our operating sampling rate. Furthermore, the corresponding glottal excitation signals are created by applying LPC analysis filtering of order 16 to the speech signals, with a frame length of 20 ms.

3.1. Residual Neural Encoder

This network converts the LPC residual at sampling rate 16 kHz into a learned context vector of 1 kHz. The context vector is the conditional prior required for generating the target fake speech. The network consists of a stack of 4 downsampling convolutional layers. The downsampling is done by a 1D-convolution operation with kernel width 64 and stride 2, so that each layer downsamples its input by a factor of two. Parametric rectified linear unit (PReLU) [15] is used for activation. Reflection padding is used for adjusting the signal length during the learned downsampling process. This results in the following feature maps starting from the input residual until the output of the fourth layer: 16000×1, 8000×32, 4000×64, 2000×64, 1000×128. Finally, the output of the fourth layer is fed to a compressor represented by a convolutional layer with kernel size 65 and one output channel to obtain the context vector.

3.2. Generator Network

3.2.1. Softmax-gated CNN

One important feature of the generator network is the softmax-gated CNN layer. It is defined and implemented similarly to the sigmoid-gated CNN of WaveNet [3]. However, the sigmoid operation is replaced by a softmax along the channel dimension of the gate output. Thus, the output of this gated layer is given as follows:

$$output = \tanh\left(W_f * X\right) \odot softmax(W_g * X)|_c, \quad (1)$$

where X is the input to the gated-CNN layer, W_f are the weights of the 1D-convolutional filter, W_g are the weights of the 1D-convolutional gate, $*$ denotes the convolution operation and $\odot$ denotes the element-wise multiplication. For all gated-CNNs layers in the generator model, a kernel of width 65 is used for both the filtering and gating operations with reflection padding to maintain the signal length.

3.2.2. Context Decoder

This block consists of a stack of 10 identical gated-CNN layers that generate a hidden representation of 64 channels for the context vector learned by the residual encoder. It was found more effective than direct upsampling as it reduces the reconstruction artifacts of the generated fake signal. A 1×1 convolution operation precedes this block to create 64 channels of the context vector ready for manipulation.

3.2.3. Adversarial Upsampler

The adversarial upsampler converts the multi-channel context decoder output of 1 kHz/channel into a single channel fake speech of 16 kHz. This is done by progressive upsampling using 4 layers. Each layer applies 1D-transposed convolution with kernel width 66 and stride 2 in order to obtain an output with doubled length compared to the layer input. Moreover, each layer passes the output of the transposed convolution through a gated-CNN without changing the dimensionality to refine and activate the upsampling. In parallel to this, a Gaussian noise of zero mean and unit variance is independently upsampled and
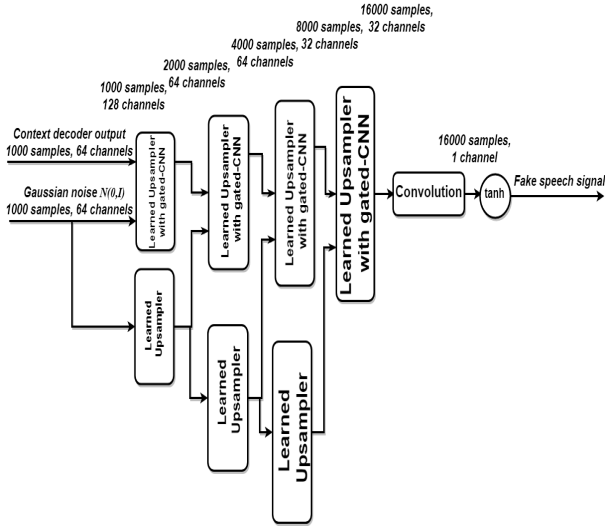
16000 samples,
8000 samples, 32 channels
4000 samples, 32 channels
2000 samples, 64 channels
1000 samples, 64 channels
128 channels

16000 samples,
1 channel

Context decoder output
1000 samples, 64 channels

Gaussian noise $N(0,I)$
1000 samples, 64 channels

Learned Upsampler with gated-CNN

Learned Upsampler with gated-CNN

Learned Upsampler with gated-CNN

Learned Upsampler with gated-CNN

Convolution

tanh

Fake speech signal

Learned Upsampler

Learned Upsampler

Learned Upsampler

Figure 2: *The adversarial upsampler network.*

shaped using transposed convolution without activation. This noise is used for compensating the missing fine details of the speech signal during the residual compression task, e.g. unvoiced speech parts and background noise. It is then concatenated along the channel dimension with the actual signal generation path at every upsampling stage. The upsampler block diagram and the feature maps throughout the signal generation path are illustrated in Figure 2. Note that the Gaussian noise has the same dimensionality as the signal feature maps at every upsampling stage. So that the input channels at each signal upsampling stage are equally divided between the noise channels and the signal channels from the previous stage. The output layer applies 1D-convolution with kernel width 65 and tanh non-linearity.

3.3. Conditional Adversarial Training

A conditional generative adversarial network (CGAN) is trained with the same technique used for image-to-image translation [16]. The conditional discriminator D receives a 2-channel concatenation of the residual and the corresponding original/fake speech signals. The network of D comprises 6 1D-CNN layers with stride 2 and kernel width 32 per each. LeakyReLU [15] with a leakage factor of 0.2 is used for activating all layers, except the last one where only the convolution operation is applied. The channel depths starting from the input until the output of D are: 2, 16, 16, 32, 32, 64 and 32. Spectral normalization [17] is applied to all convolutional layers of D to ensure the Lipschitz continuity that is required for stable adversarial training using distance-based loss functions [18]. The conditional generator G is the cascade of the residual encoder, the context decoder and the adversarial upsampler networks which are trained jointly. We have also applied spectral normalization to all convolutional layers of G as this was found helpful for better training stability [19]. The training of D is driven by the adversarial hinge loss [19]:

$$L_D = -\mathbb{E}_{x \sim p_{speech}, r \sim p_{residual}}[\min(0, -1 + D(x,r))] \\ -\mathbb{E}_{z \sim p_z, r \sim p_{residual}}[\min(0, -1 - D(G(z,r),r))], \quad (2)$$

where L_D is the total conditional loss of D, p_{speech} denotes the original speech data, $p_{residual}$ denotes the residual data, p_z

denotes the Gaussian noise used during the adversarial upsampling and $G(z, r)$ denotes the fake speech data. For training G, the total loss function L_G is given by the following convex form:

$$L_G = \gamma \mathbb{E}_{x \sim p_{speech}, z \sim p_z, r \sim p_{residual}}[|| G(z,r) - x ||_1] \\ -(1 - \gamma)\mathbb{E}_{z \sim p_z, r \sim p_{residual}}[D(G(z,r),r)], \quad (3)$$

with regularization factor γ of 0.00015. Both D and G are optimized alternately with equal number of training iterations. Adam optimizer with AMSG$_{RAD}$ [20] is used, with lr = 0.0006 for D / 0.00015 for G and $\beta = [0.5, 0.99]$. Xavier algorithm [21] is used for initializing the weights for both D and G. The batch size is 32.

4. Results

The main outcome of this work is the ability of CGANs to create realistic speech waveforms in one-shot from a highly compressed representation of the glottal excitation. This is enhanced by the cross synthesis step in order to obtain a natural reconstruction, as illustrated in Figure 3.
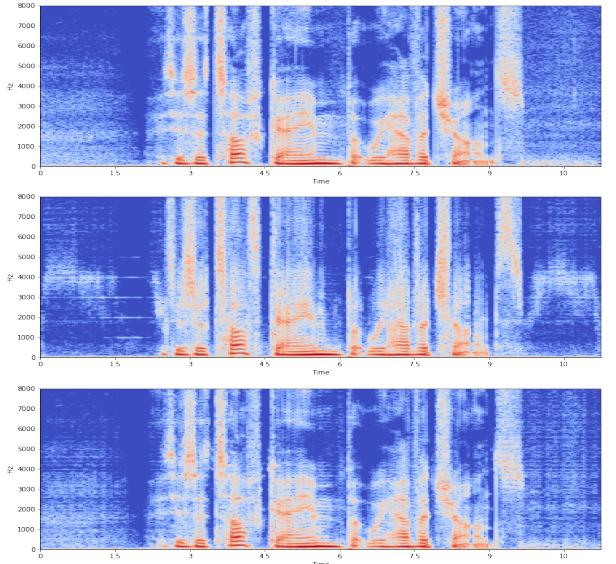

Figure 3: *GANs for speech vocoding: A fake speech signal is generated by CGAN (middle) at 16 kHz from the 1 kHz learned compression of the residual signal. This fake signal preserves the main spectral and prosodic features of the original speech (top) especially at the low frequency bands. However, it is more challenging to accurately reconstruct the high frequency details and the background noise of the original signal. That is why the cross synthesis step is incorporated to refine the fake speech signal using the original LPC parameters, which results in a natural final reconstruction (bottom).*

The gated activation was found more robust than the ReLU-based one for penalizing the reconstruction artifacts during the signal generation. Furthermore, using the softmax along the channel dimension of feature maps of the gate output is more effective than the element-wise sigmoid. The L1 loss curve of the conditional generator shows a faster decay with softmax than the sigmoid case, as illustrated by Figure 4. A possible reason for this is that the softmax along the channel dimension models the relationship between the frequency bins of the signal at

every time instant. This leads to a probability mask that gives a higher weight to the frequency bins which are more relevant to the desired samples of the target signal, while penalizing the artifacts with lower weights to their frequency components.
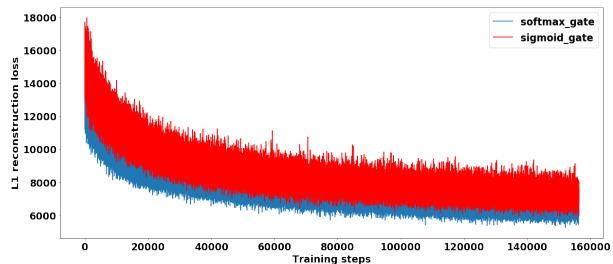

Figure 4: *Outperformance of the proposed softmax gating over the sigmoid one in terms of the L1 reconstruction loss.*

The proposed AbAS approach is assessed by objective and subjective perceptual evaluation measures. This is done in comparison with the classical vocoder introduced by Hedelin [22] and refined by Klejsa et al. [6]. There is no quantization applied to the compressed representation of signals for both the classical vocoder and AbAS. We only focus on evaluating the signals reconstructed from the non-quantized parametric representation.

4.1. Objective Evaluation

We resort to the 5 objective measures used by Pascual et al. for evaluating the SEGAN [23]: PESQ-WB, CSIG, CBAK, COVL, SSNR. In addition, we use the ViSQOL perceptual objective score [24]. All of these measures give their results in mean opinion score (MOS), except for the SSNR which is given in dB. This ensures a precise evaluation of the proposed approach in terms of perceptual quality and robustness against the reconstruction artifacts. Table 1 shows how AbAS outperforms the classical vocoder reconstructions.

Table 1: *Objective evaluation results, 40 speech signals are randomly selected from both the training and testing datasets for assessment.*

Metric	ValidationSet		TestSet	
	Vocoder	*AbAS*	*Vocoder*	*AbAS*
ViSQOL[MOS]	2.9062	3.1355	2.9161	3.1065
PESQ-WB[MOS]	2.5138	2.8234	2.3065	2.5977
CSIG[MOS]	4.1234	4.4876	3.9771	4.3143
CBAK[MOS]	2.4697	3.0317	2.3760	2.8720
COVL[MOS]	3.2941	3.6596	3.1146	3.4534
SSNR[dB]	-2.1973	3.1438	-2.0200	2.5865

4.2. Subjective Evaluation

A MUSHRA listening test [25] is performed by 7 subjects to evaluate the perceptual quality of 20 reconstructed speech signals for males and females. The CGAN model is trained with 600 epochs. Most of the AbAS reconstructions are more perceptually preferred than the classical vocoder ones. It is worth mentioning that longer training with more data should give better results due to the better approximation of data modalities by

CGAN. But we just emphasize here the proof of the concept. It was also found that the perceptual quality can be scaled by increasing the cross synthesis parameters as this compensates the degradation of the missing phonemes which are not well reconstructed with CGAN due to the sub-optimal distribution modelling.
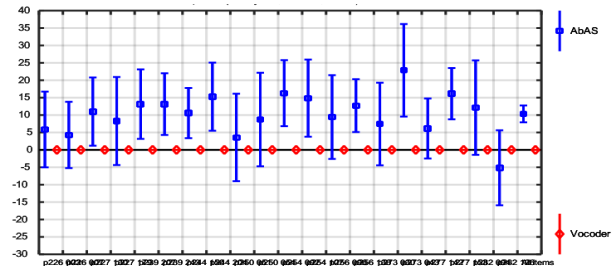

Figure 5: *MUSHRA differential scores of AbAS w.r.t classical vocoder.*

Instead of AbAS, we tried to generate a fake residual from a very compressed representation with CGAN and hence apply an LPC synthesis using the original LPC parameters to reconstruct the speech signal. However, this gave poor results compared to AbAS. That is because the discriminator is stronger in rejecting fake uncorrelated signals (i.e., residuals) than correlated ones, which makes it harder to generate realistic residuals from a compressed representation. Figure. 6 illustrates this finding.
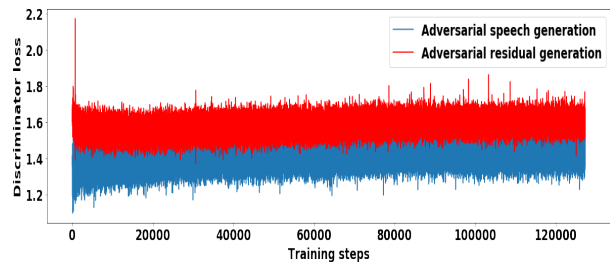

Figure 6: *Higher discriminator loss for generating fake residual compared to fake speech, which indicates a lower quality for the generated residual samples.*

5. Conclusions

This paper introduces a new method for neural speech vocoding, with much faster generation than autoregressive generative models and higher perceptual quality than classical vocoding. The method, which is called analysis by adversarial synthesis (AbAS), starts with generating a fake speech signal from a neurally-learned parametric representation of the glottal excitation using conditional GANs. This is accompanied by an LPC cross synthesis step, using the spectral envelope parameters of the original speech, to obtain a natural reconstruction. A possible future work is to explore better convolutional architectures for the generator model to reduce the reconstruction artifacts. Further work can be done to investigate the possibility of predicting the cross synthesis parameters from the fake speech. This makes it promising to optimize this approach for competing with advanced classical speech codecs at considerably lower coding rates.

6. References

[1] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[2] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (amr-wb)," *IEEE transactions on speech and audio processing*, vol. 10, no. 8, pp. 620–636, 2002.

[3] 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," *CoRR abs/1609.03499*, 2016.

[4] 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.

[5] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "Wavenet based low rate speech coding," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 676–680.

[6] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin, and L. Villemoes, "High-quality speech coding with sample rnn," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7155–7159.

[7] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 3918–3926.

[8] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[9] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proc. of Interspeech*, 2017, pp. 3394–3398.

[10] L. Juvela, B. Bollepalli, X. Wang, H. Kameoka, M. Airaksinen, J. Yamagishi, and P. Alku, "Speech waveform synthesis from mfcc sequences with generative adversarial networks," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5679–5683.

[11] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=B1xsqj09Fm

[12] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial neural audio synthesis," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=H1xQVn09FX

[13] 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*, 2016, pp. 146–152.

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

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=B1QRgziT-

[18] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. of the International Conference on Machine Learning*, 2017, pp. 214–223.

[19] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 7354–7363. [Online]. Available: http://proceedings.mlr.press/v97/zhang19d.html

[20] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *Proc. of the 6th International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=ryQu7f-RZ

[21] 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.

[22] P. Hedelin, "A sinusoidal lpc vocoder," in *Proc. of the IEEE Workshop on Speech Coding*. IEEE, 2000, pp. 2–4.

[23] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proc. of INTERSPEECH*, 2017, pp. 3642–3646.

[24] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Visqol: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 13, May 2015. [Online]. Available: https://doi.org/10.1186/s13636-015-0054-9

[25] R. B. ITU-R, "1534-1, method for the subjective assessment of intermediate quality levels of coding systems (mushra)," *International Telecommunication Union*, 2003.