Probability Density Distillation with Generative Adversarial Networks for High-Quality Parallel Waveform Generation

Ryuichi Yamamoto[1], Eunwoo Song[2] and Jae-Min Kim[2]

[1]LINE Corp., Tokyo, Japan.
[2]NAVER Corp., Seongnam, Korea

Abstract

This paper proposes an effective probability density distillation (PDD) algorithm for WaveNet-based parallel waveform generation (PWG) systems. Recently proposed teacher-student frameworks in the PWG system have successfully achieved a real-time generation of speech signals. However, the difficulties optimizing the PDD criteria without auxiliary losses result in quality degradation of synthesized speech. To generate more natural speech signals within the teacher-student framework, we propose a novel optimization criterion based on generative adversarial networks (GANs). In the proposed method, the inverse autoregressive flow-based student model is incorporated as a generator in the GAN framework, and jointly optimized by the PDD mechanism with the proposed adversarial learning method. As this process encourages the student to model the distribution of realistic speech waveform, the perceptual quality of the synthesized speech becomes much more natural. Our experimental results verify that the PWG systems with the proposed method outperform both those using conventional approaches, and also autoregressive generation systems with a well-trained teacher WaveNet.

Index Terms: WaveNet, parallel WaveNet, neural vocoder, probability density distillation, generative adversarial network.

1. Introduction

Generative models using WaveNet have significantly improved the quality of synthetic speech signals [1]. In this kind of system, the time domain speech signal is represented as a sequence of discrete symbols, and its distribution is autoregressively modeled by stacked convolutional neural networks. By appropriately conditioning the acoustic features to the input, WaveNet has also been successfully adopted in a neural vocoder structure for statistical parametric speech synthesis systems [2–5], and end-to-end speech synthesis systems [6–10].

However, compared with traditional parametric vocoders [11–14], the WaveNet's inference speed is inherently slow owing to its autoregressive model structure. To address this problem, *teacher-student framework*-based fast waveform generation methods (e.g., parallel WaveNet and ClariNet) have been proposed [15, 16]. In this framework, a bridge defined as probability density distillation (PDD) transfers the knowledge of a well-trained autoregressive teacher WaveNet to an inverse autoregressive flow (IAF)-based student model. As the architecture of feedforward IAFs enables transforming a simple noise signal to a complex distribution in parallel [17], the IAF student can generate speech waveform within a real-time speed.

Typically, conventional PDD methods employ a minimization criterion based on the Kullback-Leibler divergence (KLD) between the output distributions of the student and teacher networks [15]. However, as the objective of this criterion is to guide the student model to learn the teacher's distribution, the best achievable quality of the distilled student cannot be better than that of the teacher network. Although combining auxiliary losses (e.g., a frame-level power loss between recorded and synthetic speech signals) to the KLD criterion helps generating more natural speech segments [16], it often suffers from unexpected artifacts in the synthesis step due to the difficulties to converge the student model.

To further improve synthetic speech quality of WaveNet-based parallel waveform generation (PWG) systems, we propose a generalized optimization criterion for training the IAF students by incorporating generative adversarial networks (GANs) [18]. In the proposed method, a teacher WaveNet is first obtained via maximum likelihood estimation, and an IAF student is incorporated as a generator within the GAN framework. Finally, all the weights in the student model are jointly optimized by the PDD mechanism with an adversarial learning method. Because the adversarial training encourages the IAF student to learn the distribution of realistic speech waveform, the perceptual quality of synthesized speech becomes much more natural. Furthermore, the joint optimization with conventional distillations addresses the difficulties of feedforward GAN to model the long term dependency of the speech signal. Consequently, the performance of the distilled student is effectively improved.

We investigate the effectiveness of the proposed method by conducting subjective evaluations with the PWG systems. The experimental results show that the proposed adversarial training method provides much better perceptual quality than conventional approaches while maintaining the equivalent generation speed; moreover, outperforms even the autoregressive teacher WaveNet.

2. Related work

The idea of using PWG methods in the WaveNet framework is not new. By minimizing the KLD between output distributions of the teacher and student, *parallel WaveNet* successfully achieves to distill the IAF student model from the teacher WaveNet model [15]. By combining regularized KLD distillation with frame-level STFT loss, *ClariNet* has proposed an effective and stable training criterion [16]. As the STFT-based loss function is designed to guide the IAF student to learn the time-frequency characteristics of speech signals, its output quality has been further improved.

Meanwhile, GANs have attracted a great deal of attention in the speech signal processing community thanks to their capabilities to learn the distribution of realistic speech signals via adversarial training. The performance of the speech synthesis systems has been also significantly improved by implanting the GAN structure to the acoustic models [19–22], the post-filters [23], and the glottal excitations [24, 25].

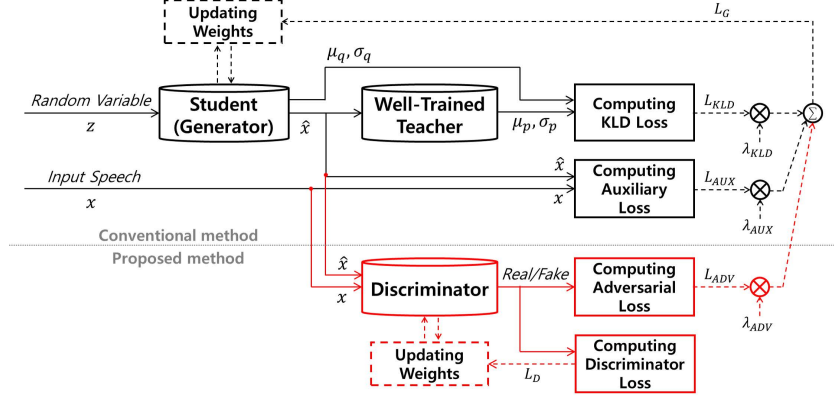Our aim is to incorporate the adversarial learning method

Figure 1: *An illustration of the distillation process for parallel waveform generation, where our proposed teacher-student framework adds an adversarial training process to the conventional methods (upper only).*

into the teacher-student training process to achieve high-quality PWG of speech signals. Although a prior work in using GAN structure in the PWG application has been undertaken [26], our research differs from this study: The GAN in the prior work was not used to distill the student model from the teacher WaveNet, but used to adapt an already-trained student model to a specific speaker (e.g., a speaker adaptation task). On the other hand, we focus on the effect of the adversarial learning method in training the student model itself. We propose a generalized optimization criterion by combining conventional KLD distillation with frame-level STFT loss and the proposed GAN-based adversarial loss.

In addition to above, our experiments seek to verify the superior performance of the proposed method over conventional PWG systems. Furthermore, thanks to the GAN's good capability to represent the nature of speech signals, the quality of the synthesized speech from the student model becomes more natural than even that from the teacher WaveNet.

3. Probability density distillation

3.1. KLD distillation

Conventional teacher-student framework-based systems employ the KLD-based PDD method to transfer the knowledge of a well-trained autoregressive teacher WaveNet to the target IAF student model [15, 16]. As the simplified architecture of the student model enables sampling the speech signal in parallel, the generation speed becomes much faster than that of the autoregressive teacher.

The upper part of Figure 1 depicts a distillation process of the conventional teacher-student framework. During the training process, the student model first transforms the input random variable $\boldsymbol{z}$ to a waveform sample $\hat{\boldsymbol{x}}$, and is evaluated by the corresponding well-trained teacher WaveNet. The entire network of the student model is then optimized to represent the teacher's distribution by minimizing the regularized KLD between the output distributions of the teacher and the student as follows [16]:

$$L_{\mathrm{KLD}}(q,p) = \mathbb{E}_{\boldsymbol{z},\hat{\boldsymbol{x}}}\left[\sum_{t=1}^{T} \mathrm{KL}^{\mathrm{reg}}(q(\hat{x}_t|z_{<t}) \parallel p(\hat{x}_t|\hat{x}_{<t}))\right], \tag{1}$$

where $q(\hat{\boldsymbol{x}}) \sim N(\mu_q, \sigma_q)$ and $p(\hat{\boldsymbol{x}}) \sim N(\mu_p, \sigma_p)$ denote the output distributions of the student and teacher, respectively.

3.2. STFT-based auxiliary loss

In addition to KLD minimization, it is well known that incorporating additional auxiliary losses using the ground truth dataset is advantageous to distill the student model well [27]. Note that synthesized speech often contains undesirable artifacts (e.g., whispering voices) when the student IAF is trained with KLD loss alone [16].

To address the aforementioned problem, loss functions that are correlated with the perceptual audio quality should be used to train the student model [28]. In this paper, we adapt a frame-level auxiliary loss between the original and the generated speech samples as follows:

$$L_{\mathrm{AUX}}(q) = \mathbb{E}_{\boldsymbol{x},\hat{\boldsymbol{x}}}\left[L_{\mathrm{SC}}(\boldsymbol{x},\hat{\boldsymbol{x}}) + \lambda_{\mathrm{MAG}}L_{\mathrm{MAG}}(\boldsymbol{x},\hat{\boldsymbol{x}})\right], \tag{2}$$

where $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ denote the target and the estimated speech signal; λ_{MAG} denotes a weight coefficient to balance two losses, *spectral convergence* (L_{SC}) and *log STFT magnitude* loss (L_{MAG}), which is defined as follows [28]:

$$L_{\mathrm{SC}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\parallel |\mathrm{STFT}(\boldsymbol{x})| - |\mathrm{STFT}(\hat{\boldsymbol{x}})| \parallel_F}{\parallel |\mathrm{STFT}(\boldsymbol{x})| \parallel_F}, \tag{3}$$

$$L_{\mathrm{MAG}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \parallel \log|\mathrm{STFT}(\boldsymbol{x})| - \log|\mathrm{STFT}(\hat{\boldsymbol{x}})| \parallel_1, \tag{4}$$

where $\parallel \cdot \parallel_F$ and $\parallel \cdot \parallel_1$ denote the Frobenius and L_1 norms, respectively; $|\mathrm{STFT}(\cdot)|$ denotes the STFT magnitudes. Because the spectral convergence loss emphasizes spectral peaks and the log STFT magnitude loss accurately fits spectral valleys [28], using a linear combination of both losses is helpful to effectively distill the student from the teacher WaveNet.

4. Probability density distillation with generative adversarial networks

The KLD distillation combined with the STFT auxiliary loss has shown the feasibility to enhance the distillation efficiency. To further improve the performance of the student model, we propose to incorporate GAN-based loss into the teacher-student framework.

Figure 1 shows the proposed distillation process. The student model is incorporated as a generator and jointly optimized by minimizing the adversarial loss (L_{ADV}) along with the KLD loss (L_{KLD}) and auxiliary loss (L_{AUX}) as follows:

$$\begin{aligned} L_{\mathrm{G}}(q,p,D) = &\lambda_{\mathrm{kld}}L_{\mathrm{KLD}}(q,p) + \lambda_{\mathrm{aux}}L_{\mathrm{AUX}}(q) \\ &+ \lambda_{\mathrm{adv}}L_{\mathrm{ADV}}(q,D), \end{aligned} \tag{5}$$

where λ_{kld}, λ_{aux} and λ_{adv} denote the normalized weight coefficients for the KLD, STFT auxiliary, and adversarial losses, respectively[1]. The adversarial loss, which represents how the student model learns the speech distribution from the discriminator, is defined as follows:

$$L_{\mathrm{ADV}}(q, D) = \mathbb{E}_{\hat{\boldsymbol{x}} \sim q} \left[(1 - D(\hat{\boldsymbol{x}}))^2 \right], \qquad (6)$$

where D denotes the discriminator[2]. During the training process, the student model tries to deceive the discriminator into recognizing the generated samples as *real* ($D(\hat{\boldsymbol{x}}) \to 1$). On the other hand, the discriminator is trained to correctly classify the generated sample to *fake* while classifying the ground truth to *real* ($D(\boldsymbol{x}) \to 1$) using the following optimization criterion:

$$L_{\mathrm{D}}(q, D) = \mathbb{E}_{\boldsymbol{x} \sim p_{\mathrm{data}}}[(1 - D(\boldsymbol{x}))^2] + \mathbb{E}_{\hat{\boldsymbol{x}} \sim q}[D(\hat{\boldsymbol{x}})^2], \quad (7)$$

where p_{data} denotes the distribution of the speech signals.

The entire training process encourages the student model to learn the distribution of the realistic speech waveform, which enables to generate more natural speech. Furthermore, the joint optimization with conventional distillations can address the limitations of feedforward GAN to capture the sample-level correlations of the speech signal [25]. Consequently, the perceptual quality of synthesized speech generated by the proposed method is effectively improved.

5. Experiments

5.1. Experimental setup

To investigate the effectiveness of the proposed method, we trained student models using the following four different optimization criteria:

- **AX**: Auxiliary loss.
- **AXAD**: Auxiliary and adversarial losses.
- **KLAX**: KLD and auxiliary losses.
- **KLAXAD**: KLD, auxiliary and adversarial losses.

In the experiments, we used a phonetically and prosaically balanced speech corpus recorded by a female professional Japanese speaker. The speech signals were sampled at 24 kHz, and each sample was quantized by 16 bits. In total, 3,299 utterances (7.34 hours) were used for training, 412 utterances (0.89 hours) were used for development, and another 413 utterances (0.92 hours) not included in either the training or development steps were used for evaluation. The leading and trailing silences in the speech signal were trimmed using a pre-processor, and 80-band log-mel spectrograms were extracted for composing the conditioning feature vectors. The frame and shift lengths were set to 25 ms and 5 ms, respectively. Before training, the conditional feature vectors were normalized to have zero mean and unit variance. All the models and experiments were implemented using NAVER smart machine learning (NSML) platform [31].

The teacher model was Gaussian autoregressive WaveNet [16], which consisted of 24 layers of dilated residual convolution blocks with four exponentially increasing dilation cycles.

Table 1: *Normalized weight coefficients for training the different IAF student models.*

Method	λ_{kld}	λ_{aux}	λ_{adv}
AX	-	1.00	-
AXAD	-	0.33	0.67
KLAX	0.09	0.91	-
KLAXAD	0.03	0.32	0.65

The number of residual channels, skip channels were 128 and convolution filter size was 3. The conditioning features were upsampled by nearest neighbor upsampling followed by 2-D convolutions for interpolation [32]. The upsampling was split into five modules. The scales were [2, 2, 2, 3, 5]. The kernel sizes for the 2-D convolutions were set to $2s + 1$, where s denotes the upsampling scale. The teacher model was trained for 1 M steps with an *Adam* optimizer [33]. The initial learning rate was set to 0.001, and it was reduced by half for every 200 K steps. The minibatch size was eight and the length of each audio clip was 12 K time samples.

The student models were based on Gaussian IAFs [16], each consisted of six flows in our settings. Each flow was parameterized by a WaveNet that had ten layers of dilated residual convolution blocks with an exponentially increasing dilation cycle. The number of residual channels, skip channels was 64 and filter size was 3. The architecture of the upsampling network was the same as that of the teacher, and all the weights were initialized by the teacher's. The IAF student models were trained for 500 K steps with an Adam optimizer. The normalized weight coefficients (i.e., λ_{kld}, λ_{aux} and λ_{adv}) for training the different IAF student models are summarized in Table 1. The initial learning rate was set to 0.0001, and it was reduced by half for every 200 K steps. The minibatch size was eight and the length of each audio clip was 20.4 K time samples. The STFT auxiliary loss was computed with a 25 ms Hanning window with 5 ms shift. The weight λ_{MAG} in Equation 2 was set to $1/(N_{\mathrm{frame}} \times F_{\mathrm{freq}})$, where N_{frame} and F_{freq} denote the number of time frames and frequency bins of the STFT magnitude, respectively. The KLD loss was computed as the same way as ClariNet [16].

In the proposed adversarial learning method, the discriminator consisted of ten layers of non-causal dilated 1-D convolutions interleaved with leaky ReLU activation function ($\alpha = 0.2$). The strides for the 1-D convolutions were set to 1 and linearly increasing dilations were applied for the 1-D convolutions[3] starting from 1 to 8 except for the first and last layers. Scalar predictions per-time step were done to better capture sample-level detailed differences between generated and real samples, and then averaged to compute the discriminator loss. The number of channels and filter size were 64 and 3, respectively. The conditioning feature vectors were not used for the discriminator. Because it is impossible to optimize the discriminator directly at the beginning of the training process, the student model as a generator was trained without the adversarial loss during the first 200 K steps. After warmup, the discriminator was sequentially optimized for 50 K with an Adam optimizer and finally entire networks were jointly trained via the adversarial learning method for the remaining 300 K steps. The initial learning rate for the discriminator was set to 0.00005, and it was reduced by half for every 200 K steps.

---

[1] If the weight λ_{kld} is zero, the optimization criterion is equivalent to adversarial training methods [25, 26]. On the other hand, if the weight λ_{adv} is zero, it is equivalent to conventional PDD with the STFT auxiliary loss [16].

[2] This framework adopts a least-squares GANs thanks to its stability during the training process [24, 26, 29, 30].

---

[3] Our preliminary experiments verified that the linearly increasing dilations performed better than the exponentially increasing receptive fields for the discriminator.
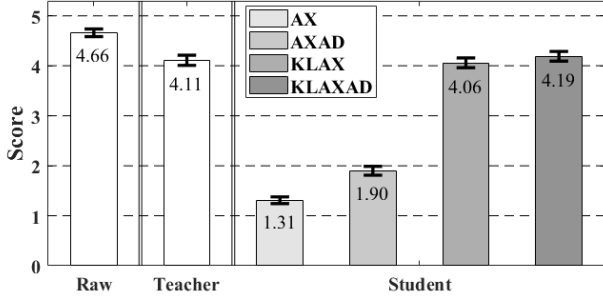
Figure 2: *The MOS results with 95 % confidence intervals.*

5.2. Experiment results

To evaluate the perceptual quality of the proposed system, mean opinion score (MOS)[4] tests were performed. Fourteen native Japanese speakers were asked to make quality judgments about the synthesized speech samples using the following five possible responses: 1 = Bad; 2 = Poor; 3 = Fair; 4 = Good; and 5 = Excellent. In total, 20 utterances were randomly selected from the test set and were then synthesized using the different generation models.

Figure 2 shows the MOS test results with respect to different generation models. The findings can be summarized as follows: (1) The IAF student trained only with the auxiliary loss (i.e., AX) performed worst. Although adding the adversarial loss (i.e., AXAD) proved advantageous to improve the perceptual quality of the synthesized speech, it still scored poorly since it was challenging to capture sample-level correlations of the speech signal without the KLD-based distillation criterion. This can be confirmed by the test results for the KLAX system, where the perceptual quality was significantly improved by using the KLD distillation criterion with the auxiliary loss. (2) Among the IAF students, the proposed adversarial training method (i.e., KLAXAD) achieved the best quality. In particular, the proposed system outperformed even the teacher WaveNet model. This was because the adversarial training guided the IAF student to learn the distribution of realistic speech waveform. Consequently, the proposed system with the adversarial training method achieved 4.186 MOS.

To further verify the effect of the proposed method, we designed additional experiments by refining the normalized weight coefficients of the proposed system (i.e., λ_{kld}, λ_{aux} and λ_{adv} in the KLAXAD system). Note that the previous listening test results verified that it is necessary to use the KLD-based distillation criterion during the training process. However, the best achievable quality of the distilled student model can be limited to the teacher model if the weight for KLD-based distillation (i.e., λ_{kld}) is too large. Therefore, when the IAF student model starts to converge, it is recommended to decrease the λ_{kld} value.

Figure 3 depicts the A/B/X preference test results[5]. Although the normalized weight coefficients were empirically modified (λ_{kld}, λ_{aux}, and λ_{adv} were set to zero, 0.33, and 0.67, respectively), the results confirm that weight-refined system (KLAXAD*) provided better perceptual quality than the one

---

Figure 3: *The test results of A/B/X preference comparison with two proposed methods including the baseline (KLAXAD) and its weight-refined version (KLAXAD*).*

originally proposed (KLAXAD). This implies that, when the student model started to converge, forcing the entire networks to be optimized toward the ground truth speech data rather than the teacher model was advantageous to generate more natural speech signal. Making the normalized weight coefficients learnable during the training process can further improve the general performance, which will be discussed in our future research.

6. Conclusions

This paper proposed an effective probability density distillation algorithm with generative adversarial networks (GANs) for WaveNet-based parallel waveform generation (PWG) systems. Within a teacher-student framework, the proposed method incorporated an inverse autoregressive flow (IAF)-based student model as a generator in the GAN framework. Using novel optimization criteria based on adversarial learning method, the perceptual quality of the synthesized speech became much more natural. The experimental results verified that the PWG system using the proposed GAN-based training method performed better than the systems with conventional approaches. Despite the fact that the IAF student model was distilled from the teacher WaveNet, the merits of GAN to represent the nature of speech waveform enabled the student model to generate more natural speech than even the well-trained teacher model.

7. Acknowledgements

8. References

[1] 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," in *Arxiv*, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499

[2] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. INTERSPEECH*, 2017, pp. 1118–1122.

[3] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, 2017, pp. 712–718.

[4] E. Song, K. Byun, and H.-G. Kang, "Excitnet vocoder: A neural excitation model for parametric speech synthesis systems," in *Proc. EUSIPCO (in press)*, 2019.

[5] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. ICASSP*, 2018, pp. 4804–4808.

[6] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proc. ICML*, 2017, pp. 195–204.

---

[4]Generated audio samples are available at the following URL: https://r9y9.github.io/demos/projects/interspeech2019/

[5]The setups for the test were the same as for the MOS tests except that listeners were asked to rate the quality preference of the synthesized speech samples.

[7] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. NIPS*, 2017, pp. 2962–2970.

[8] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. ICLR*, 2018.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[10] O. Kwon, E. Song, J.-M. Kim, and H.-G. Kang, "Effective parameter estimation methods for an excitnet model in generative text-to-speech systems," *arXiv preprint arXiv:1905.08486*, 2019.

[11] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[12] Y. Agiomyrgiannakis, "VOCAINE the vocoder and applications in speech synthesis," in *Proc. ICASSP*, 2015, pp. 4230–4234.

[13] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. EUSIPCO*, 2014, pp. 2290–2294.

[14] E. Song, F. K. Soong, and H.-G. Kang, "Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 11, pp. 2152–2161, 2017.

[15] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. 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 *Proc. ICML*, 2018, pp. 3915–3923.

[16] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR (in press)*, 2019.

[17] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Proc. NIPS*, 2016, pp. 4743–4751.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[19] Y. Zhao, S. Takaki, H.-T. Luong, J. Yamagishi, D. Saito, and N. Minematsu, "Wasserstein GAN and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a WaveNet vocoder," *IEEE Access*, vol. 6, pp. 60 478–60 488, 2018.

[20] S. Yang, L. Xie, X. Chen, X. Lou, X. Zhu, D. Huang, and H. Li, "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework," in *Proc. ASRU*, 2017, pp. 685–691.

[21] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 26, no. 1, pp. 84–96, 2018.

[22] J. Y. Lee, S. J. Cheon, B. J. Choi, N. S. Kim, and E. Song, "Acoustic modeling using adversarially trained variational recurrent neural network for speech synthesis," in *Proc. INTERSPEECH*, 2018, pp. 917–921.

[23] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," in *Proc. INTERSPEECH*, 2017, pp. 3389–3393.

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

[25] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Waveform generation for text-to-speech synthesis using pitch-synchronous multi-scale generative adversarial networks," in *Proc. ICASSP (in press)*, 2019.

[26] Q. Tian, B. Yang, J. Chen, B. Tang, and S. Liu, "Generative adversarial network based speaker adaptation for high fidelity WaveNet vocoder," in *Arxiv*, 2018. [Online]. Available: https://arxiv.org/abs/1812.02339

[27] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *EMNLP*, 2016, pp. 1317–1327.

[28] S. Ö. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Procees. Letters*, vol. 26, no. 1, pp. 94–98, 2019.

[29] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, 2017, pp. 2794–2802.

[30] S. Pascual, A. Bonafonte, and J. Serr, "SEGAN: Speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.

[31] H. Kim, M. Kim, D. Seo, J. Kim, H. Park, S. Park, H. Jo, K. Kim, Y. Yang, Y. Kim *et al.*, "NSML: Meet the MLaaS platform with a real-world case study," in *Arxiv*, 2018. [Online]. Available: https://arxiv.org/abs/1712.05902

[32] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.