ExcitGlow: Improving a WaveGlow-based Neural Vocoder with Linear Prediction Analysis

Suhyeon Oh*, Hyungseob Lim*, Kyungguen Byun*, Min-Jae Hwang†, Eunwoo Song‡ and Hong-Goo Kang*
* Yonsei University, South Korea
E-mail: {shoh, hyungseob.lim, piemaker90}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr
† Search Solutions, Inc., South Korea
E-mail: min-jae.hwang@navercorp.com
‡ Naver Corp., South Korea
E-mail: eunwoo.song@navercorp.com

***Abstract*—In this paper, we propose ExcitGlow, a vocoder that incorporates the source-filter model of voice production theory into a flow-based deep generative model. By targeting the distribution of the excitation signal instead of the speech waveform itself, we significantly reduce the size of the flow-based generative model. To further reduce the number of parameters, we apply a parameter sharing technique in which a single affine coupling layer is used for several flow layers. To avoid quality degradation, we also introduce a closed-loop training framework to optimize the flow model for both the speech and excitation signal generation processes. Specifically, we choose negative log-likelihood (NLL) loss for the excitation signal and multi-resolution spectral distance for the speech signal. As a result, we are able to reduce the model size from 87.73M to 15.60M parameters while maintaining the perceptual quality of synthesized speech.**

***Index Terms*—Neural vocoder, speech synthesis, excitation modeling, WaveGlow**

I. Introduction

Neural vocoders that directly generate raw speech waveform with deep generative models have been successfully applied in speech synthesis systems [1], [2], [3]. For example, WaveNet [1], which is an autoregressive model, is able to reliably estimate speech samples by conditioning the samples generated at previous time steps. However, its generation speed is very slow due to the sample-by-sample generation method. To accelerate generation speed, probability distillation-based non-autoregressive models have been proposed, such as parallel WaveNet [4] and ClariNet [5]. By adopting a teacher-student framework to improve the training efficiency of the inverse autoregressive flow (IAF) [6], these methods were able to parallelize the generation process in the inference phase. However, training process of these models is not easy, because it requires a well-trained teacher network and a sophisticated distillation process between the teacher and student networks.

Subsequently, motivated by the success of bipartite normalizing flow models [7], [8], [9], the WaveGlow vocoder [3], which can generate speech signals of reasonable quality in parallel with each other, was proposed. In this framework, an invertible neural network is trained to model the relationship between the probability distribution of a simple prior (e.g. isotropic Gaussian) and the complex probability distribution of a speech waveform. However, due to the weak modeling power of the bipartite structure, a large number of transforms is required in order to estimate the distribution of complex speech waveforms compared to the AF or IAF models [3], [10]. In addition, since the WaveGlow model adopts WaveNet-like affine coupling layers, each of which is comparable in size to WaveNet itself, it is much larger than any other neural vocoder.

To address this problem, we propose ***ExcitGlow***, a vocoder that incorporates the source-filter model of voice production theory into a flow-based neural vocoding model. ExcitGlow consists of a WaveGlow-based excitation generator and a parallel spectral filtering module that replaces a linear prediction (LP) synthesis filter. In the training step, the excitation generator learns the relationship between the distribution of excitation signals and Gaussian random variables. To further improve the effectiveness of the model, we apply a weighting filter to the excitation signal for clearer fundamental frequency (F0) trajectory modeling. Additionally, we propose a closed-loop framework in the flow model training process, which uses joint optimization criteria for both the excitation signal and speech. For the excitation signal, negative log-likelihood (NLL) loss is used for training. For the speech signal, a multi-resolution spectral distance metric between the frequency domain representation of the target and that of the generated speech signal is used. At the inference step, we first generate a weighted excitation signal from a Gaussian prior, using the ExcitGlow model, and then synthesize a speech waveform by passing the weighted excitation signal through post and LP synthesis filters.

The ExcitGlow vocoder can be trained more easily than the WaveGlow model because the statistical behavior of an excitation signal is simpler than that of a speech signal [11], [12]. By exploiting this behavior, the network size can be significantly reduced without degrading the quality of synthesized speech.

We summarize our contributions as follows: (1) we incorporate an LP structure into the WaveGlow model. Thus, the network size, especially the dimension of the condition features, can be significantly reduced; (2) we address the limitation of F0 modeling in targeting the excitation by adding a weighting filter; (3) we verify that using a low dimensional

978-988-14768-8-3/20/$31.00 ©2020 APSIPA

mel-spectrogram is sufficient for conditioning the ExcitGlow model; (4) we successfully combine two loss functions (NLL loss and multi-resolution spectral distance) that are needed to represent the characteristics of the excitation and speech signals, respectively. Experimental results show that our framework enables the model size to be reduced to 17.78% of the original size, while achieving a mean opinion score (MOS) of 3.67.

II. Related works

There have been several attempts to incorporate the source-filter model of voice production theory into neural vocoding systems. For instance, GlotNet, ExcitNet, and LP-WaveNet [13], [11], [12] improved the quality of WaveNet-based neural vocoders by introducing an LP synthesis filter to the waveform generation phase. By dividing a unified speech synthesis model into excitation and spectral modeling parts, noise artifacts that existed in vanilla WaveNet could be alleviated. LPCNet and iLPCNet [14], [15] used the LP-based source-filter model to reduce the complexity of the WaveRNN-based neural vocoding model. By these methods, synthesized speech of reasonable quality was obtained, even with a smaller condition vector dimension, which helped to accelerate synthesis speed and reduced the number of model parameters. In the GELP vocoder [16], a generative adversarial network (GAN)-based parallel waveform generation model was proposed with a parallelized LP synthesis process. In addition to the adversarial loss of GAN, a short-time Fourier transform (STFT) loss-based spectral distance measure was successfully merged into a single framework.

The main objective of this paper is to extend the usage of a source-filter model to the flow-based parallel waveform generation model. Combining a flow-based model with a source-filter model provides several advantages. First, the waveform generation speed is much faster than that of a sample-wise autogressive generative model, because the model structure allows for the adoption of a flow-based parallelized framework. Second, due to the relatively simple probability distribution of excitation signals, the model's memory efficiency can be improved by reducing the conditioning parameters for the flow model. Finally, quality degradation can be minimized by incorporating an LP synthesis process into closed-loop training with effective optimization criteria.

III. WaveGlow

WaveGlow [3] is a type of neural vocoder that uses a flow model. The idea behind WaveGlow is to create a complex target data distribution from a simple distribution prior, such as a Gaussian prior. Using the change of variables formula between two arbitrary distributions, as shown in Eq. (1), it can directly estimate the likelihood of observed data.

$$p_X(x) = p_Z(f(x)) \left| \frac{\partial f(x)}{\partial x} \right|. \tag{1}$$

Since WaveGlow is trained solely to minimize NLL loss, the training process is more stable than those of other types of generative models. To reduce the complexity required to compute the determinant of the Jacobian matrix, a bipartite flow is introduced, where half of the samples are updated by scaling, and adding factors predicted by the coupling layer output from the other half of the samples [7]. The WaveGlow model adopts a trainable invertible 1×1 convolution in the coupling layer, to shuffle efficiently the information of each squeezed channel. Its performance is higher than those of fixed or random permutation methods [9]. The model was extended to a multi-scale architecture through a squeezing operation, and a criterion was introduced to widen the receptive field size of the coupling function [8].

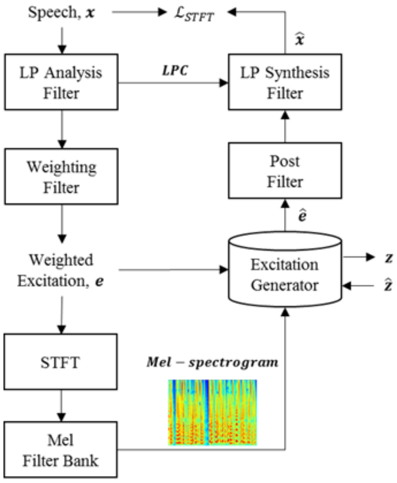


Fig. 1. Overall block diagram of ExcitGlow

WaveGlow is able to synthesize speech dozens of times faster than real time. However, since there can be only one transformation (*i.e.* multiplication and addition) in a single flow chain, a large number of flow chains is required to estimate accurately the complex distribution of the target signals. In addition, each flow chain requires a coupling layer of a size comparable to WaveNet itself. As a result, WaveGlow has a very large number of parameters. The dimension of the condition vector also affects to the model size, because of its effects on the transposed convolution layer.

IV. Proposed method

To reduce the model complexity of WaveGlow, we propose ExcitGlow, a model that changes the target signal from speech to excitation signals. The rationale is as follows. Firstly, the transformation from a Gaussian prior to a probability distribution of excitation signals is simpler than a transformation to speech signals. Secondly, we can reduce the dimension of the condition vector by using a smaller dimensional mel-spectrogram. We also introduce a closed-loop structure for training that simultaneously uses the optimization criteria in both the excitation and speech signal domains.

Fig. 1 is the overall block diagram of the proposed vocoder system. To prepare features for training, we need to estimate linear prediction coefficients (LPCs) to represent the spectral information of the speech signal and extract an excitation signal. An additional weighting filter is applied to the extracted excitation signal for better F0 modeling in the Waveglow-based excitation generator. Furthermore, the condition vectors for the flow model, (*i.e.* the mel-spectrogram) also need to be estimated. In the training step, the weighted excitation signal is used as the target of the generator and the extracted excitation features as a conditioning vector for the excitation generator. Since the generator is an invertible model, the random variable $\hat{z}$ is sampled from a Gaussian distribution to simulate the inference stage. It is transformed into $\hat{e}$ by passing through the excitation generator in the reverse direction. The estimated excitation signal then goes through the post and a parallel LP synthesis filter to synthesize the speech signal. We use two kinds of loss functions for training: NLL and spectral distance. In the inference phase, the excitation and speech samples are generated in the same way as in the simulation scenario used for the training steps.

A. Weighting filter for F0 modeling

The excitation signal is obtained by subtracting the output of the LP filter from the reference speech, as follows:

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k], \tag{2}$$

where p and a_k represent the LPC order and the k-th order LPC, respectively.

Since the LP filter uses only short-term (20-40 samples) relationships within input speech, long-term information such as the F0 trajectory tends to remain in the excitation signal. However, in some voiced regions with a simple periodic form, excitation signals approach to nearly zero because of too accurate linear prediction results. In those regions, the target excitation barely contains any F0 information, so the excitation generator fails to generate a clear F0 trajectory. To solve this problem, we apply a perceptual weighting filter to the excitation signal; it emphasizes the null regions and de-emphasizes the formant regions. With this filter, we can obtain the prominent F0 trajectory in weighted excitation signals. The weighting filter of the t-th frame, Q_t, can be obtained as follows:

$$\boldsymbol{Q_t} = \frac{1}{1 - \sum_{k=1}^{p} a_k \alpha^{-k} z^{-k}}, \tag{3}$$

where α is the perceptual weight coefficient.

B. Closed-loop training with parallel LP synthesis filter

There are two reasons to use closed-loop training with a parallel LP synthesis filter in our model. First, due to the autoregressive nature of the LP synthesis filter, we cannot make full use of the fast generation speed of WaveGlow. Second, if the flow model is trained to estimate only the excitation signal, without considering the LP synthesis step (*i.e.* open-loop training), unavoidable errors in excitation estimation can be amplified due to the feedback nature of the AR filtering process.

To address this problem, we use the parallel filtering method proposed in the GELP vocoder [16] that includes the LP filtering part in the training framework. In the parallel LP filter, speech samples are synthesized in the frequency domain by using an element-wise multiplication. The frequency response of the t-th frame, H_t, is retrieved from the LPC as follows:

$$\boldsymbol{H_t} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{4}$$

By including this parallel filter in the proposed framework, it is possible to create a closed-loop training scheme that eventually targets a speech signal.

Before performing the filtering step, we compute and concatenate the frequency response of each analysis frame, denoted as $H = [H_1, H_2, ..., H_L]$. The excitation signals, estimated from the WaveGlow-based generator, are then converted into the frequency domain using an STFT. Finally, two 2-D vectors are multiplied element-wise in the frequency domain and transformed into the time domain by an inverse STFT (ISTFT) to obtain the final speech output, as follows:

$$\hat{\boldsymbol{x}} = ISTFT\{STFT\{\hat{e}\} \odot \boldsymbol{H}\} \tag{5}$$

C. Dimensional reduction of condition features

In order to efficiently reduce the dimension of the condition vector of the WaveGlow-based excitation model, we devise a method to extract the condition vector by considering the signal characteristics. In the low-frequency band of the excitation signal, a relatively high frequency resolution is required to represent F0 information. Conversely, high-frequency information is relatively unimportant because the LP synthesis filter employed after the excitation generator can reconstruct this information. For this reason, we can reduce the dimension of the weighted excitation's mel-spectrogram while maintaining the frequency resolution by changing the maximum frequency of the mel-filterbanks.

With this approach, we are able to reduce the number of mel-spectrogram channels, compared to the conventional WaveGlow model. This also has the effect of significantly reducing the number of parameters without severe quality degradation. We conducted experiments to find the optimal setting for determining the maximum frequency and the dimension of the mel-spectrogram; the results are provided in Section V.

D. Loss functions

Two loss terms are used in the proposed framework as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{NLL} + \lambda \mathcal{L}_{SD}, \tag{6}$$

where L_{NLL} denotes NLL loss, which is used to train the WaveGlow-based weighted excitation generator. L_{SD} is the multi-resolution spectral distance loss between the reference and target speech. The total loss of the entire framework is a weighted sum of the two loss terms, where λ is the weighting factor determined by experiments.

1) Negative log-likelihood loss: L_{NLL} is measured by the change of variables theorem, as follows:

$$\mathcal{L}_{NLL} = -\log p_z(f(e)) - \log \left| \frac{\partial f(e)}{\partial e} \right|, \tag{7}$$

where e is the reference weighted excitation signal and p_z is a Gaussian prior.

2) Multi-resolution spectral distance loss: We use STFT loss, which measures the logarithmic difference between the STFT magnitudes of the ground truth and the estimated speech signal, to improve the perceptual quality of synthesized speech. Additionally, we apply multi-resolution STFT loss to complement the trade-off between time and frequency resolutions, as follows:

$$\mathcal{L}_{SD} = \frac{1}{N} \sum_{k=1}^{N} (\log |STFT_k\{\boldsymbol{x}\}| - \log |STFT_k\{\hat{\boldsymbol{x}}\}|)^2, \tag{8}$$

where N denotes the number of STFT loss terms and k represents the index of the STFT loss.

V. Experiments

A. Experimental settings

1) Database: We used the LJ Speech Dataset [17] to evaluate the performance of the proposed framework. The data was recorded by a professional female English speaker at a sampling rate of 22.05 kHz. There are a total of 13,100 utterances, totaling approximately 24 hours of speech. We used 13,000 utterances for training and the remainder for testing. In every iteration, a randomly segmented set of 16,384 samples from each utterance were cropped for training.

To prepare the weighted excitation signals, training utterances were passed through an LP analysis filter and a perceptual weighting filter. In our experiments, we used 24th order LP filter coefficients and set the perceptual weight α to 0.8. To prepare the mel-spectrogram needed for conditioning the flow model, we first applied a STFT to the weighted excitation signals. We used 1024/256 samples for analysis frame and shift length for STFT computation, which approximately correspond to 50/12.5 ms in 22.05 kHz sampling rate. As explained in Section IV, we extracted mel-spectrograms by changing the maximum frequency and the number of mel-filterbanks. Specifically, the maximum frequency was set to 2000, 4000, or 8000 Hz, and the dimension of the mel-spectrogram was reduced to 24.

2) Model: We used the conventional WaveGlow model with 12 flow chains as our baseline. Additionally, we trained a WaveGlow model with 8 flow chains to make the number of parameters comparable to ExcitGlow, in an effort to validate the effectiveness of the ExcitGlow model. We also examined the efficiency of the parameter sharing model, which has one-quarter of the number of parameters of the ExcitGlow model, by comparing the synthesized signal quality of each system.

These models were trained using a Gaussian distribution with a standard deviation of 1.0, which follows the vanilla WaveGlow setup. As an affine coupling layer, we adopted a non-autoregressive WaveNet-like architecture with 8 dilated convolutions and 256/512 channels of skip/residual connections. At every 4 flow layers, the early output technique was applied to directly connect the two input channels to the network output. 80-dimensional mel-spectrograms, which contained the speech signal's information under 8000 Hz, were used for the condition features.

TABLE I
STFT Settings for Multi-resolution Spectral Distance Loss

STFT index	FFT size	Window size	Shift size
$STFT_1$	1024	550 (25ms)	110 (5ms)
$STFT_2$	2048	1100 (50ms)	220 (10ms)
$STFT_3$	512	220 (10ms)	44 (2ms)

TABLE II
Conventional WaveGlow systems

Model	Vocoder	# of flows	# of params.	MOS
A	WaveGlow	12	87.73M	3.72 ± 0.09
B	WaveGlow	8	60.68M	2.85 ± 0.14
-	Reference	-	-	4.99 ± 0.02

Besides reducing the dimension of the condition features, we also implemented a parameter sharing method to further reduce our model size without significant quality degradation. Unlike the baseline model, we trained the ExcitGlow vocoder by using a Gaussian distribution with a standard deviation of 0.1 to reflect the smaller magnitude range of excitation signals. We pre-trained the model with only a NLL loss for the first 100,000 iterations, to maintain stability in the training process. After pre-training, the STFT loss was included in the training criteria. For multi-resolution spectral distance loss, three window and shift length settings were used. They are summarized in Table I.

All neural vocoders were trained for 400,000 iterations with a batch size of 8 and a learning rate of 10^{-4}. The weights were normalized with a weight normalization method [18] and the Adam optimizer [19] was used for training. We performed a MOS test to evaluate all models. Specifically, we employed 15 participants to score the perceptual quality of 15 randomly chosen test sentences from each model.

B. Effect of maximum frequency in reduced mel-spectrograms

To reduce the number of model parameters, we reduced the dimension of the mel-spectrograms from 80 to 24. We also tried 16 and 32 dimensions; but, quality degradation was severe in the 16 dimension case and there were only minimal differences between the 24 and 32 dimensions. The MOS test results of synthesized speech from the two conventional WaveGlow systems and reference speech are summarized in Table II. Model A, and B denote the baseline WaveGlow model with 12 and 8 flow layers, respectively. Model B resulted in a significantly lower MOS than model A (2.85 vs. 3.72), which implies that simply reducing the number of flow layers results in severe quality degradation.

TABLE III
COMPARISON OF SYSTEM COMPLEXITY AND MOS DEPENDING ON THE MAXIMUM FREQUENCY OF CONDITION FEATURES

Model	Vocoder	Maximum frequency	# of params.	MOS
C	WaveGlow	8000 Hz	59.75M	3.22 ± 0.12
D	WaveGlow	4000 Hz	59.75M	3.15 ± 0.19
E	WaveGlow	2000 Hz	59.75M	2.28 ± 0.18
F	ExcitGlow	8000 Hz	59.75M	3.43 ± 0.15
G	**ExcitGlow**	**4000 Hz**	**59.75M**	**3.85 ± 0.18**
H	ExcitGlow	2000 Hz	59.75M	3.45 ± 0.21

TABLE IV
COMPARISON OF SYSTEM COMPLEXITY AND MOS DEPENDING ON THE TYPE OF STFT LOSS

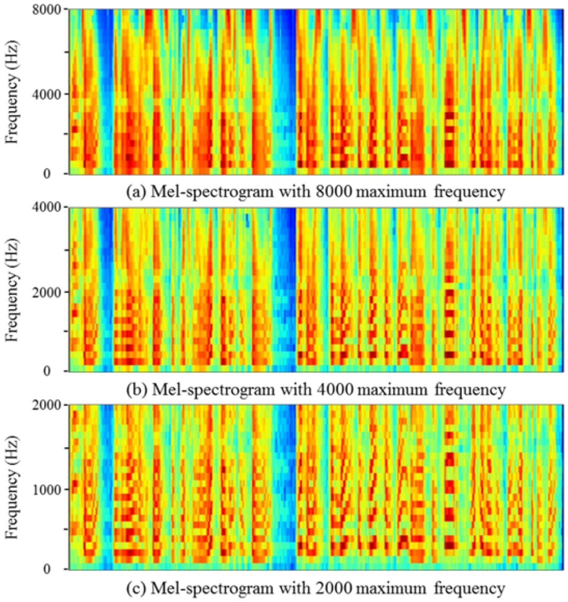
Model	Vocoder	# of params.	MOS	Remarks
I	ExcitGlow	15.60M	3.67 ± 0.16	Single resolution
J	ExcitGlow	15.60M	3.66 ± 0.22	Multi-resolution

TABLE V
COMPARISON OF INFERENCE SPEED

Model	Vocoder	Real time factor
A	12-WaveGlow	0.024
I, J	Shared ExcitGlow	0.023 (exc) / 0.12 (total)

Models C-H were tested to analyze the effects of varying the maximum frequency of the mel-spectrogram under both the WaveGlow and ExcitGlow frameworks. All parameter settings without the type of condition vector were set to be the same as for baseline model A. The results are summarized in Table III. Model C shows the effect of reducing the dimension of the mel-spectrograms; due to a less informative condition features, MOS decreases to 3.22. In WaveGlow models with lower maximum frequencies (models C-E), synthesized speech quality degrades significantly because high-frequency information cannot be retrieved from the condition features. However, in the ExcitGlow cases (models F-H), quality degradation is not as severe because the LP synthesis filter can represent the formant structure of full-band speech.

When we reduced the dimension of the mel-spectrograms without maximum frequency reduction, the pitch contours became unrecognizable in many regions, as shown in Fig. 2(a). By decreasing the maximum frequency of the mel-filterbanks while keeping the order, the F0 trajectory is more clearly observed due to the increased frequency resolution of the filterbanks. However, when the maximum frequency is set to 2000 Hz, the condition vector cannot provide proper information for signal generation because high order harmonics still exist in frequency regions higher than 2000 Hz. We therefore, set the maximum frequency to 4000 Hz in the ExcitGlow model because it maintained the best performance. With the maximum frequency reduction, ExcitGlow does not show harmonic artifacts that are present in WaveGlow. Because of this, we see that model G achieves better performance than the conventional WaveGlow model (model A).



(a) Mel-spectrogram with 8000 maximum frequency

(b) Mel-spectrogram with 4000 maximum frequency

(c) Mel-spectrogram with 2000 maximum frequency

Fig. 2. Mel-spectrogram depending on the maximum frequency

C. *Parameter sharing & multi-resolution STFT Loss*

To further reduce the number of model parameters, we applied a parameter sharing technique to the affine coupling layer of the flow model. To prevent severe modeling power degradation, sharing was applied at every 4 flow layers, where early stopping occurs. With parameter sharing, the number of model parameters is reduced to one-fourth of that of the reduced condition feature case and one-sixth of the original model. Even with the smaller number of parameters, we were able to obtain MOS of 3.67, which is comparable to the original WaveGlow model.

To verify the effectiveness of the multi-resolution STFT loss, we compared the performances of models I, and J. Model J was trained with multi-resolution STFT loss, depicted in Table I. On the other hand, model I was trained only with single resolution STFT loss, using the $STFT_2$ setting in Table I. The efficiency of applying multi-resolution STFT loss to training criteria was proved by [20], [21] when the model targets speech signals. However, the synthesized quality did not dramatically improve in ExcitGlow because the LP synthesis filter reliably provides spectrum-related information.

D. *Inference speed*

We also measured the real time factor (RTF), which represents the ratio between the amount of time to process the input and the duration of the input [22]. For the generation process, an NVIDIA TITAN RTX was used; the results are summarized in Table V. When considering only the excitation generation part, models I, and J generate faster than model A due to their smaller model size. But if we take the LP synthesis part into account, the generation speed becomes slower than the baseline WaveGlow model. Nevertheless, our proposed model is still able to generate speech much faster than real time.

VI. Conclusion

We proposed ExcitGlow, a vocoder designed to reduce the model complexity of WaveGlow. To improve the effectiveness of the conventional WaveGlow model, we targeted the excitation signal instead of the speech signal. By representing complicated speech spectrum information through an LP synthesis filter, we were able to reduce the dimension and the maximum frequency of the mel-spectrogram, which was necessary to condition the WaveGlow-based excitation generator. In addition, we proposed a closed-loop training framework with multi-resolution loss to prevent quality degradation in the proposed model. As a result, we were able to synthesize speech signals of comparable quality while using only one-sixth of the number of parameters of the conventional WaveGlow model.

VII. Acknowledgements

The work was supported by Clova Voice, NAVER Corp., Seongnam, Korea.

References

[1] A. v. d. 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.

[2] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient neural audio synthesis,” *arXiv preprint arXiv:1802.08435*, 2018.

[3] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[4] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, “Parallel wavenet: Fast high-fidelity speech synthesis,” *arXiv preprint arXiv:1711.10433*, 2017.

[5] R. Levy, D. Blaauw, G. Braca, A. Dasgupta, A. Grinshpon, C. Oh, B. Orshav, S. Sirichotiyakul, and V. Zolotov, “Clarinet: A noise analysis tool for deep submicron design,” in *Proceedings of the 37th Annual Design Automation Conference*. ACM, 2000, pp. 233–238.

[6] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, “Improved variational inference with inverse autoregressive flow,” in *Advances in Neural Information Processing Systems*, 2016, pp. 4743–4751.

[7] L. Dinh, D. Krueger, and Y. Bengio, “Nice: Non-linear independent components estimation,” *arXiv preprint arXiv:1410.8516*, 2014.

[8] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real nvp,” *arXiv preprint arXiv:1605.08803*, 2016.

[9] D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[10] W. Ping, K. Peng, K. Zhao, and Z. Song, “Waveflow: A compact flow-based model for raw audio,” *arXiv preprint arXiv:1912.01219*, 2019.

[11] E. Song, K. Byun, and H.-G. Kang, “Excitnet vocoder: A neural excitation model for parametric speech synthesis systems,” in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[12] M.-J. Hwang, F. Soong, F. Xie, X. Wang, and H.-G. Kang, “Lp-wavenet: Linear prediction-based wavenet speech synthesis,” *arXiv preprint arXiv:1811.11913*, 2018.

[13] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, “Glotnet—a raw waveform model for the glottal excitation in statistical parametric speech synthesis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1019–1030, 2019.

[14] J.-M. Valin and J. Skoglund, “Lpcnet: Improving neural speech synthesis through linear prediction,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[15] M. Hwang, E. Song, R. Yamamoto, F. Soong, and H. Kang, “Improving lpcnet-based text-to-speech with linear prediction-structured mixture density network,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7219–7223.

[16] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, “Gelp: Gan-excited liner prediction for speech synthesis from mel-spectrogram,” *arXiv preprint arXiv:1904.03976*, 2019.

[17] K. Ito, “The lj speech dataset,” https://keithito.com/LJ-Speech-Dataset/, 2017.

[18] T. Salimans and D. P. Kingma, “Weight normalization: A simple reparameterization to accelerate training of deep neural networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 901–909.

[19] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[20] X. Wang, S. Takaki, and J. Yamagishi, “Neural source-filter-based waveform model for statistical parametric speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5916–5920.

[21] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[22] A. V. Ivanov, P. L. Lange, D. Suendermann-Oeft, V. Ramanarayanan, Y. Qian, Z. Yu, and J. Tao, “Speed vs. accuracy: Designing an optimal asr system for spontaneous non-native speech in a real-time application.”