LP-WaveNet: Linear Prediction-based WaveNet Speech Synthesis

Min-Jae Hwang*† Frank Soong‡, Eunwoo Song§, Xi Wang‡, Hyeonjoo Kang† and Hong-Goo Kang†
* Search Solution, Seongnam, South Korea
E-mail: min-jae.hwang@navercorp.com
† Yonsei University, Seoul, South Korea, Seongnam, South Korea
E-mail: volleruhe@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr
‡ Microsoft, Beijing, China
E-mail: {frankkps; xwang}@microsoft.com
§ Naver Corporation, Seongnam, South Korea
E-mail: eunwoo.song@navercorp.com

***Abstract*—We propose a linear prediction (LP)-based waveform generation method via WaveNet vocoding framework. A WaveNet-based neural vocoder has significantly improved the quality of parametric text-to-speech (TTS) systems. However, it is challenging to effectively train the neural vocoder when the target database contains massive amount of acoustical information such as prosody, style or expressiveness. As a solution, the approaches that only generate the vocal source component by a neural vocoder have been proposed. However, they tend to generate synthetic noise because the vocal source component is independently handled without considering the entire speech production process; where it is inevitable to come up with a mismatch between vocal source and vocal tract filter. To address this problem, we propose an LP-WaveNet vocoder, where the complicated interactions between vocal source and vocal tract components are jointly trained within a mixture density network-based WaveNet model. The experimental results verify that the proposed system outperforms the conventional WaveNet vocoders both objectively and subjectively. In particular, the proposed method achieves 4.47 MOS within the TTS framework.**

I. Introduction

Waveform generation systems using WaveNet have significantly improved the synthesis quality of deep learning-based text-to-speech (TTS) systems [1]–[5]. Because the WaveNet vocoder can generate speech samples in a single unified neural network, it does not require any hand-engineered signal processing pipeline. Thus, it presents much higher synthetic quality than the traditional parametric vocoders [2].

To further improve the perceptual quality of the synthesized speech, more recent neural *excitation* vocoders take advantages of the merits from both the linear prediction (LP) vocoder and the WaveNet structure [6]–[10]. In this framework, the formant-related spectral structure of the speech signal is decoupled by an LP analysis filter, and the WaveNet only estimates the distribution of its residual signal (i.e., excitation). Because the physical behavior of excitation signal is simpler than the speech signal, the training and generation processes become more efficient.

However, the synthesized speech is likely to be unnatural when the prediction errors in estimating the excitation are propagated through the LP synthesis process. As the effect of LP synthesis is not considered in the training process, the synthesis output is vulnerable to the variation of LP synthesis filter.

To alleviate this problem, we propose an ***LP-WaveNet***, which enables to jointly train the complicated interactions between the excitation and LP synthesis filter. Based on the basic assumption that the past speech samples and the LP coefficients are given as conditional information, we figure out that the distributions of speech and excitation only lies on a constant difference. Furthermore, if we model the speech distribution by using a mixture density network (MDN) [11], then the target speech distribution can be estimated by summing the mean parameters of predicted mixture and an *LP approximation*, which is defined as the linear combination of past speech samples weighted by LP coefficients. Note that the LP-WaveNet is easy to train because the WaveNet only needs to model the excitation component, and the complicated spectrum modeling part is embedded into the LP approximation.

In the objective and subjective evaluations, we verified the outperforming performance of the proposed LP-WaveNet in comparison to the conventional WaveNet-based neural vocoders. Especially, the LP-WaveNet provided 4.47 mean opinion score result in the TTS framework.

II. WaveNet-based speech synthesis systems

A. μ-law quantization-based WaveNet

WaveNet is a convolutional neural network (CNN)-based auto-regressive generative model that predicts the joint probability distribution of speech samples $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ as follows:

$$p(\mathbf{x}|\mathbf{h}) = \prod_n p(x_n|\mathbf{x}_{<n}, \mathbf{h}), \quad (1)$$

where x_n, $\mathbf{x}_{<n}$, and $\mathbf{h}$ denote the n^{th} speech sample, its past speech samples, and the acoustic features, respectively. By stacking the dilated causal convolution layers multiply, the WaveNet effectively extends its receptive field to the thousand of samples.

The firstly proposed WaveNet, a.k.a., μ-law WaveNet [1], defines the distribution of speech sample as a 256 categorical

class of symbols obtained by an 8-bit μ-law quantized speech samples. To model the distribution of speech sample, the categorical distribution is computed by applying softmax operation to the output of WaveNet. In the training phase, the weights of WaveNet is updated to minimize the cross-entropy loss. In the generation phase, the speech sample is auto-regressively generated in sample-by-sample.

Since the μ-law WaveNet can generate the speech signal in a single unified model, it provides significantly better synthetic sound than the conventional parametric vocoders. However, it is not easy to train the network when the amount of database is larger and its acoustical informations such as prosody, style, or expressiveness are wider. Moreover, the synthesized sound of WaveNet is often suffered from the background noise artifact as the target speech signal is too coarsely quantized.

B. WaveNet-based excitation modeling

One effective solution is to model the excitation signal instead of the speech signal. For instance, in the ExcitNet approach [8], an excitation signal is first obtained by an LP analysis filter, then its probabilistic behavior is trained by the WaveNet framework.

During the synthesis, the excitation signal is generated by the trained WaveNet, then it is passed through an LP synthesis filter to synthesize the speech signal as follows:

$$\begin{aligned} x_n &= e_n + \hat{x}_n, \\ \hat{x}_n &= \sum_{i=1}^{p} \alpha_i x_{n-i}, \end{aligned} \tag{2}$$

where e_n, $\hat{x}_n$, p, and α_i denote the n^{th} sample of excitation signal, the intermediate *LP approximation* term, the order of LP analysis, and the i^{th} LP coefficient, respectively. Note that the LP coefficients are periodically updated to match with the extraction period of acoustic features. For instance, if acoustic features are extracted at every 5-ms, then the LP coefficients are updated at every 5-ms to synchronize the feature update interval.

Because the variation in the excitation signal is only constrained by vocal cord movement, its training is much easier and the quality of finally synthesized speech is much higher, too. However, the synthesized speech often contains unnatural artifacts because the excitation model is trained independently without considering the effect of LP synthesis filter; where it happens mismatch between the excitation signal and LP synthesis filter. To address this limitation, we propose an LP-WaveNet, where both excitation signal and LP synthesis filter are jointly considered for training and synthesis.

III. Linear prediction WaveNet vocoder

A. Fundamental mathematics

Before introducing the proposed LP-WaveNet, a probabilistic relationship between speech and excitation signals have to be clarified. Note that at the moment of n^{th} sample generation process in the WaveNet's synthesis stage, $\hat{x}_n$ shown in (2) can be treated as a given factor since both LP coefficients, a_i, and previously reconstructed samples, $\{x_{n-i}\}$ are already estimated. Hence, we conclude that the difference between two random variables, x_n and e_n, is only a known constant value term of $\hat{x}_n$.

Considering the shift property of second-order random variable, if we define the speech's distribution as a mixture of Gaussian (MoG), the relationship between mixture parameters of speech and excitation distributions can be lie on the only constant difference of mean parameters as follows:

$$p(x_n|\mathbf{x}_{<n}, \mathbf{h}) = \sum_{i=1}^{M} \frac{w_{n,i}}{\sqrt{2\pi} s_{n,i}} \exp\left[-\frac{(x_n - \mu_{n,i})^2}{2s_{n,i}^2}\right], \tag{3}$$

$$\begin{aligned} w_{n,i}^x &= w_{n,i}^e, \\ \mu_{n,i}^x &= \mu_{n,i}^e + p_n, \\ s_{n,i}^x &= s_{n,i}^e, \end{aligned} \tag{4}$$

where M and i denote the number and index of mixture, respectively; w denotes the weights of mixture component; $\mathcal{N}(\mu, s)$ imply the Gaussian distribution having mean of μ and standard deviation of s; the superscripts e and x denote the excitation and the speech, respectively. Based on this observation, we propose an LP-WaveNet vocoder, where the LP synthesis process is structurally reflected to the WaveNet's training and inference processes.

B. Network architecture

The detailed architecture of LP-WaveNet is illustrated in Fig. 1. In the proposed system, the distribution of speech sample is defined as a MoG distribution by following (3), and the LP-WaveNet is trained to generate the MoG parameters, $[\boldsymbol{w}_n, \boldsymbol{\mu}_n, \boldsymbol{s}_n]$ conditioned by the input acoustic features.

In detail, the acoustic features pass through two 1-dimensional convolution layers having kernel size of 3 for explicitly imposing the contextual information of feature trajectory. Then, the residual connection with respect to the input acoustic feature is applied to make the network more focus on the current frame information. Finally, the transposed convolution is applied to upsample the temporal resolution of this features into that of speech signal.

To generate the speech samples, the mixture parameters, i.e., mixture gain, mean and log-standard deviation, of excitation signal are first predicted by WaveNet as follows:

$$[\mathbf{z}_n^w, \mathbf{z}_n^\mu, \mathbf{z}_n^s] = WaveNet(\mathbf{x}_{<n}, \mathbf{h}_n) \tag{5}$$

Then, the LP approximation term, $\hat{x}_n$, is computed by following (2) to generate the MoG parameters of speech sample as follows:

$$\begin{aligned} \boldsymbol{w}_n &= \text{softmax}(\mathbf{z}_n^w) \\ \boldsymbol{\mu}_n &= \mathbf{z}_n^\mu + \hat{x}_n, \\ \boldsymbol{s}_n &= \exp(\mathbf{z}_n^s). \end{aligned} \tag{6}$$

Finally, the likelihood of speech sample $p(x_n|\mathbf{x}_{<n})$ is computed by following (3).

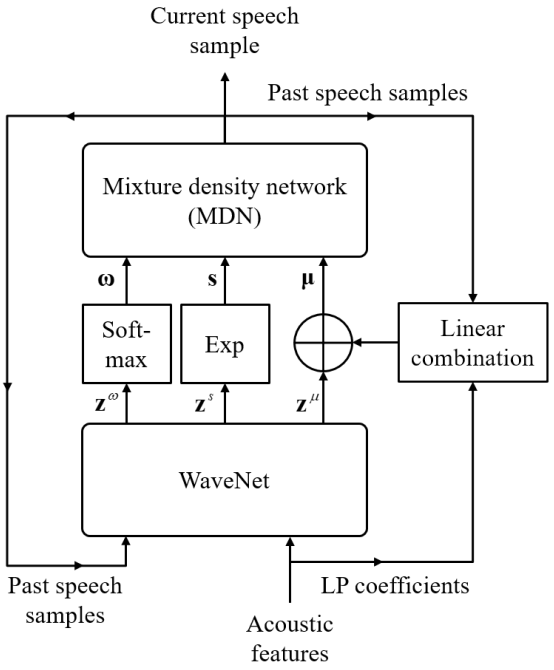

Fig. 1. Block diagram of the LP-WaveNet vocoder.

To train the network, the negative log-likelihood (NLL) of speech signal, $\mathcal{L}$, is computed from the MoG distribution defined at (3) as follows:

$$\mathcal{L} = -\sum_{n} \log p(x_n|\mathbf{x}_{<n}). \tag{7}$$

Then, the weights are optimized to minimize NLL loss.

Because the complicated spectral modeling is now embedded in the mean parameters as depicted in (6), the LP-WaveNet only needs to train an information of excitation signal, which is relatively easy to train. Moreover, because the ultimate training target of LP-WaveNet is speech signal, it is also free from the mismatch problem mentioned in Section II-B As a result, the LP-WaveNet is able to model the both excitation generation and LP synthesis filter processes jointly in a WaveNet structure.

IV. Effective training and generation methods

A. Waveform generation via conditional distribution sharpening

During waveform generation, a random sampling that follows the probability distribution of waveform is commonly used. However, its synthetic sound is noisy due to the stochastic sampling process. In this study, we control the noisiness by adjusting the sharpness of waveform distribution by reducing the scale parameters generated by the WaveNet. Because the buzziness and the hiss of synthetic speech are sensitive to the sharpness of distribution, the scale parameters have to be carefully adjusted. After several trials, we concluded that reducing the scale by factor of 0.85 at only voiced region presents the best performance.

B. Upper bound limitation on the generated log-scale parameters

During the waveform generation process, we figured out that the generated waveform can be often unstable when the generated log-scale parameters are too high. This problem could be prevented by clipping the upper bound of scale parameter value. If the clipping was set too low, then the unvoiced region was not sufficiently modeled, resulting in a dry synthetic sound though the waveform could be stably generated. If the clipping was set too high, then the possibility of waveform explosion became higher, but the synthetic sound became more lively than the lower clipping value case. Based on experiments, we limited the scale parameter to -4.0 natural logarithm.

V. Experiments

A. Speech database and features

In the experiments, phonetically and prosodically riching speech corpus recorded by a professional Korean female speaker was used for the experiments. The speech signals were sampled at 24-kHz with 16-bits quantization. The randomly selected 4,976 utterances (9.9 hours) were used for training, 280 utterances were used for validation, and another 140 utterances were used for test, respectively. The acoustic features were obtained by the ITFTE vocoder [12] at every 5-ms interval; 40-dimensional line spectral frequencies (LSFs), logarithmic fundamental frequency (F0), logarithmic energy, voicing flag, 32-dimensional slowly evolving waveform, and 4-dimensional rapidly evolving waveform, all of which composed a total 79-dimensional feature vector.

B. WaveNet vocoders

Total three WaveNet vocoding systems were tested.

- WN$_S$: μ-law WaveNet vocoder that directly models the speech signal [2].
- WN$_E$: ExcitNet vocoder that models the excitation signal with explicit LP synthesis filter [8].
- WN$_{LP}$: Proposed LP-WaveNet vocoder.

For a fair comparison with similar computing resource, the same WaveNet architecture was used to all systems. Firstly, the dilations were set to $[2^0, 2^1, ..., 2^9]$ and repeated three times, resulting in 30 layers of residual blocks and 3,071 samples of the receptive field. In the residual blocks and the post-processing module, the 128 channels of convolution layers were used. The number of mixture was set to 10, resulting in 30 channels of output layer. For the LP-WaveNet, the single Gaussian distribution was assumed, and the weight normalization technique, which normalizes the weight vectors to have unit-length, is applied to stabilize a training process of LP-WaveNet [13]. Moreover, the scale parameter was clipped by the lower bound of -10.0 natural logarithm when calculating a negative log-likelihood (NLL) loss to stabilize the training of mixture density network (MDN) [14]. The weights were firstly

TABLE I
OBJECTIVE EVALUATION RESULTS OF THE VARIOUS WAVENET VOCODERS WITH ANALYSIS AND SYNTHESIS (A/S) AND PARAMETRIC TTS SYSTEMS. THE SYSTEM WITH HIGHEST PERFORMANCE IS REPRESENTED IN BOLD TYPEFACE.

	System	VUV (%)	F0 RMSE (Hz)	LSD (dB)	F-LSD (dB)
A/S	WN_S	4.09	3.76	2.01	9.90
	WN_E	3.77	3.17	2.32	8.80
	WN_{LP}	**2.28**	**2.70**	**1.67**	**8.47**
TTS	WN_S	5.06	13.67	4.45	12.81
	WN_E	4.84	13.61	4.43	**12.30**
	WN_{LP}	**4.12**	**13.54**	**4.41**	12.37

TABLE II
SUBJECTIVE MEAN OPINION SCORE (MOS) TEST RESULT WITH A 95% CONFIDENCE INTERVAL FOR VARIOUS SPEECH SYNTHESIS SYSTEMS. THE SYSTEM WITH HIGHEST SCORE IS REPRESENTED IN BOLD TYPEFACE. THE MOS RESULT OF RECORDED SPEECH WAS 4.75.

	ITFTE	WN_S	WN_E	WN_{LP}
A/S	2.85±0.20	3.40±0.19	4.11±0.16	**4.58±0.12**
TTS	2.32±0.06	3.57±0.11	4.04±0.16	**4.47±0.09**

TABLE III
SUBJECTIVE PREFERENCE TEST RESULTS (%) BETWEEN VARIOUS WAVENET VOCODING SYSTEMS. THE SYSTEMS THAT ACHIEVED SIGNIFICANTLY BETTER PREFERENCE AT THE $p < 0.01$ LEVEL ARE IN BOLD TYPEFACE.

Index	System	WN_S	WN_E	WN_{LP}	Neutral	p-value
Test 1	A/S	**9.4**	**71.3**	–	**19.3**	$\mathbf{< 10^{-21}}$
Test 2		**2.6**	–	**82.7**	**14.7**	$\mathbf{< 10^{-45}}$
Test 3		–	**12.0**	**52.0**	**36.0**	$\mathbf{< 10^{-10}}$
Test 4	TTS	**8.0**	**57.3**	–	**34.7**	$\mathbf{< 10^{-16}}$
Test 5		**1.4**	–	**79.3**	**19.3**	$\mathbf{< 10^{-46}}$
Test 6		–	**12.0**	**33.3**	**54.7**	$\mathbf{< 10^{-4}}$

initialized by the *Xavier* initializer [15], and then trained using an *Adam* optimizer [16]. The learning rate was set to 10^{-4}. The mini-batch size was 20,000 samples with 8GPUs, resulting in 160,000 samples per mini-batch. The networks were trained in 600,000 iterations.

C. TTS acoustic model

To evaluate the performance of WaveNet vocoders in the TTS system, we implemented a simple acoustic model by using multiple feed-forward (FF) and long-short term memory (LSTM) layers. In detail, the network consisted of three FF layers with 1,024 units and one LSTM layer with 512 memory cells. The ReLu activation and linear functions were used at the hidden and output layers, respectively.

The input vector was composed of 356-dimensional linguistic features including 330 binary features of categorical linguistic contexts and 26 numerical features of numerical linguistic contexts. The corresponding output vector consisted of all the acoustic parameters together with their time dynamics [17]. Before training, both input and output features were normalized to have zero mean and unit variance. The weights were trained using a backpropagation through time algorithm with Adam optimization [18].

In the synthesis step, the means of all acoustic features were predicted by the acoustic model first, then a speech parameter generation algorithm was applied with the pre-computed global variances [19]. To enhance spectral clarity, an LSF-sharpening filter was also applied to the spectral parameters [12]. Finally, the generated acoustic features were used to compose the input features of the WaveNet vocoders.

D. Objective and subjective evaluation results

In the objective test, distortions in acoustic features extracted by the original speech and synthesized speech were evaluated. Firstly, the analysis and synthesis (A/S) system, which synthesizes the speech with the ground truth acoustic features was tested to evaluate the vocoder's performance itself. Then, the TTS system, which uses the acoustic features predicted by the LSTM-based acoustic condition model was tested in a real application scenario.

The metrics for the distortion measuring were the error rate of voicing flag (VUV) in %, the root mean square error (RMSE) for F0 in Hz, the log-spectral distance (LSD) for LSFs in dB, and the LSD for speech magnitude response in frequency domain (F-LSD) in dB. All the features needed for the metrics were extracted with 35-ms window at every 5-ms interval, then all the measures were averaged. The F0 RMSE and F-LSD were measured in only voiced region. To estimate the F-LSD, by computing phase mismatch, we compensated a lag to have maximum correlation between two speech frames within a 5-ms sample shift interval.

The objective evaluation of A/S and TTS results are summarized in Table I. The experimental results verify that (1) In all matrices, the proposed WN_{LP} showed significantly better performance than the conventional WN_S and WN_E when the acoustic features are ground truth. (2) All the performances are degraded in the TTS system as the prediction error of acoustic features. However, the performance of LP-WaveNet is still signifcantly better than the other systems.

To evaluate the perceptual quality of the proposed system, the mean opinion score (MOS) listening test A-B preference test were performed[1]. Total 11 native Korean listeners were asked to score the randomly selected 15 synthesized utterances from the test set using a following possible 5-point MOS responses: 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent. In addition to the WaveNet vocoding systems, the ITFTE-based vocoding system [12], i.e., ITFTE, having the same acoustic model with the WaveNet vocoding system was also included as a reference system.

The MOS test results are summarized in Table II. In the A/S system, all of WaveNet vocoders showed better quality than the parametric ITFTE vocoder. Specifically, the proposed WN_{LP} showed the best quality among the WaveNet vocoders with the only 0.17 lower MOS score than the recorded speech. Even though the MOS result of WN_E was higher than 4.0, its quality was significantly worse that the proposed WN_{LP}.

[1]Generated audio samples are available at the following URL: https://min-jae.github.io/apsipa2020/

In the TTS system, all systems presented worse synthetic quality than the A/S system due to the prediction error of acoustic features. However, their relative tendency was same with the results of A/S system. Even though the prediction error of acoustic features, the proposed WN_{LP} showed very high quality of synthesized speech with 4.47 MOS.

The setup for the A-B preference test was the same as that for the preference test, except the listeners were asked to rate the randomly selected 15 synthesized utterances from the test set by a quality preference. The preference results shown in Table III verified that the perceptual quality of the proposed WN_{LP} was significantly better than the conventional WN_E and WN_S in both of A/S and TTS systems (Test 2, 3, 5, and 6). Also, the WN_E verified that its performance was better than the plain WN_S (Test 1 and 4).

VI. Conclusion

In this paper, we proposed an LP-WaveNet vocoder. By utilizing the causality of WaveNet and the linearity of LP synthesis filtering process, we structurally merged the LP synthesis filter into the WaveNet framework. The experimental results verified that the proposed system outperformed the conventional WaveNet systems both objectively and subjectively. Future works include to extend the idea of LP-WaveNet to the non-autoregressive waveform models for achieving real-time waveform generation.

VII. Acknowledgements

The work was supported by Clova Voice, NAVER Corp., Seongnam, Korea, and partially performed when the first author was an intern at Microsoft Research Asia.

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," *arXiv preprint arXiv:1609.03499*, 2016.

[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] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[5] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of WaveNet as a statistical vocoder," in *Proc. ICASSP*, 2018, pp. 5674–5678.

[6] L. Juvela, V. Tsiaras, and B. Bollepalli, "Speaker-independent raw waveform model for glottal excitation," in *Proc. INTERSPEECH*, 2018, pp. 2012–2016.

[7] Y. Cui, X. Wang, L. He, and F. K. Soong, "A new glottal neural vocoder for speech synthesis," in *Proc. INTERSPEECH*, 2018, pp. 2017–2021.

[8] E. Song, K. Byun, and H. Kang, "ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems," in *Proc. EUSIPCO*, 2019, pp. 1179–1183.

[9] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. ICASSP*, 2019, pp. 5916–5920.

[10] J.-M. Valin and J. Skoglund, "LPCNet: Improving Neural Speech Synthesis Through Linear Prediction," in *Proc. ICASSP*, 2019, pp. 5891––5895.

[11] C. M. Bishop, "Mixture density networks," Tech. Rep., 1994.

[12] 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.

[13] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. NIPS*, 2016, pp. 901–909.

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

[15] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[17] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986.

[18] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for online training of recurrent network trajectories," *Neural computat.*, vol. 2, no. 4, pp. 490–501, 1990.

[19] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.