LP-WAVENET: LINEAR PREDICTION-BASED WAVENET SPEECH SYNTHESIS

Min-Jae Hwang*†*, Frank Soong*†*, Fenglong Xie*†*, Xi Wang*† *and Hong-Goo Kang**

*Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea
†Microsoft Research Asia, Beijing, China

ABSTRACT

We propose a linear prediction (LP)-based waveform generation method via WaveNet speech synthesis. The WaveNet vocoder, which uses speech parameters as a conditional input of WaveNet, has significantly improved the quality of statistical parametric speech synthesis system. However, it is still challenging to effectively train the neural vocoder when the target database becomes larger and more expressive. As a solution, the approaches that only generate the vocal source signal by the neural vocoder have been proposed. However, they tend to generate synthetic noise because the vocal source is independently handled without considering the entire speech synthesis 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 that structurally models the vocal source in the speech training and inference processes. The experimental results verify that the proposed system outperforms the conventional WaveNet vocoders both objectively and subjectively.

Index Terms— Text-to-speech, speech synthesis, statistical parametric speech synthesis, WaveNet, WaveNet vocoder

1. INTRODUCTION

The WaveNet vocoder [1, 2], which uses the acoustic features as a conditional input of WaveNet [3], significantly improves the synthesis quality of conventional deep learning-based statistical parametric speech synthesis (SPSS) systems. 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 quality of synthetic speech than that of the conventional band-aperiodicity (BAP)-based deterministic vocoder [4]. Inspired by this success, many WaveNet-style neural vocoders have been proposed and actively studied [1, 5–8].

However, training neural vocoder is not an easy task, especially when the training database is large and has an expressive characteristic. One effective solution is to model the vocal source signal only, instead of the entire speech waveform [7–9]. In these approaches, the vocal source signal is first estimated by a linear prediction (LP)-based analysis [10–12], then modeled by a neural vocoder. Because the vocal source signal shows physically simpler behavior than the speech signal, its training is relatively easier, too. However, the synthesized speech is likely to be noisy due to the mismatch between vocal source and vocal tract filter. In detail, the synthesis output is vulnerable to the variation of vocal tract filter because the vocal tract filtering effect is not considered in the neural vocoder training.

In this paper, we propose an ***LP-WaveNet***, where both vocal source signal and vocal tract filter are jointly considered during the waveform training and inference processes. Based on the basic assumption that the past speech samples and the LP coefficients are given as conditional information, we figure out that the difference between the random variables of speech and excitation only lies on a constant factor. Furthermore, if we model the speech distribution by a mixture density network (MDN) [13], then the distribution of excitation signal can be converted to the distribution of speech by simply "shifting" the mean components of excitation mixture parameters. In detail, the mixture parameters of excitation signal are predicted first by the WaveNet. Then, the mean parameters of target speech distribution is estimated by summing up the predicted mixture parameters and the *LP approximation*, which presents 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 signal, and complicated spectrum modeling portion is embedded into the LP approximation.

In this study, we only focus on the WaveNet vocoder with the discretized mixture of logistic (MoL) distribution [14, 15]. However, the proposed LP-WaveNet can be extended to any of neural vocoders that utilize an MDN-based auto-regressive generative model using LP coefficients. For example, the sample-RNN vocoder [5] or the FFTNet vocoder [6] can be used instead of the WaveNet vocoder, and the mixture of Gaussian (MoG) [13] can be used instead of MoL.

2. WAVENET-BASED SPEECH SYNTHESIS SYSTEMS

2.1. Mixture Density Network-based WaveNet vocoder

WaveNet is an autoregressive generative model that directly models the joint probability distribution of speech samples

This work was done when the first author was an intern at MSRA.

$\mathbf{x} = \{x_1, x_2, ..., x_N\}$ by the factorized form as follows:

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

where x_n, $\mathbf{x}_{<n}$, and $\mathbf{h}$ are the n^{th} speech sample, its past speech samples, and the acoustic features, respectively.

In the widely used MDN-based WaveNet, i.e., MDN-WaveNet [15], a speech signal is assumed to follow the discretized MoL distribution as follows:

$$x_n|\mathbf{x}_{<n}, \mathbf{h} \sim \sum_{i=1}^{N} \pi_i \cdot DistLogistic(\mu_i, s_i), \tag{2}$$

where $DistLogistic(\cdot)$ and N denote the discretized logistic distribution and the number of mixture components, respectively. π_i, μ_i, and s_i are the i^{th} component of the mixture gain, mean, and scale parameters, respectively. WaveNet is used to predict the mixture parameters such as:

$$[\boldsymbol{z}^{\pi}, \boldsymbol{z}^{\mu}, \boldsymbol{z}^{s}] = WaveNet(\mathbf{x}_{<n}, \mathbf{h}), \tag{3}$$

where the vector sequences $\boldsymbol{z}^{\pi}$, $\boldsymbol{z}^{\mu}$, and $\boldsymbol{z}^{s}$ denote the mixture gain, mean, and log-scale parameters, respectively. Then, the mixture parameters can be defined as follows:

$$\begin{aligned} \boldsymbol{\pi} &= \text{softmax}(\boldsymbol{z}^{\pi}), \\ \boldsymbol{\mu} &= \boldsymbol{z}^{\mu}, \\ \boldsymbol{s} &= \exp(\boldsymbol{z}^{s}). \end{aligned} \tag{4}$$

Note that the softmax function, softmax$(\cdot)$ and the exponential function, exp$(\cdot)$ are used to guarantee the unity summed mixture gain and the positive value of mixture scale parameters, respectively.

To train the network, the likelihood of discretized MoL distribution is computed as follows:

$$p(x_n|\mathbf{x}_{<n}, \mathbf{h}) = \sum_{i=1}^{N} \pi_i \cdot \left[\sigma\left(\frac{x + \Delta/2 - \mu_i}{s_i}\right) - \sigma\left(\frac{x - \Delta/2 - \mu_i}{s_i}\right)\right], \tag{5}$$

where $\sigma(\cdot)$ and Δ denote the logistic sigmoid function and the quantization step size, respectively. Note that the quantization step size, Δ, is set to $1/2^{16}$ for matching with that of speech sample. Then, the weights are optimized to minimize the negative log-likelihood (NLL) loss.

2.2. WaveNet-based excitation modeling

Despite of the high quality synthesized speech of MDN-WaveNet, its training is not easy when the amount of database is larger and its acoustical informations such as prosody, style, or expressiveness are wider. One effective solution is to model the vocal source signal instead of the speech signal. In our previous work [9], 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 passes through an LP synthesis filter to synthesize the speech signal as follows:

$$\begin{aligned} x_n &= e_n + \hat{x}_n, \\ \hat{x}_n &= \sum_{i=1}^{p} \alpha_i x_{n-i}, \end{aligned} \tag{6}$$

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 the acoustic features. For instance, if the 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 structure of excitation signal is simpler than that of speech signal, its training is much easier and the quality of finally synthesized speech is much higher, too. However, the synthesized speech becomes often noisy 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.

3. LINEAR PREDICTION WAVENET VOCODER

Before describing the proposed LP-WaveNet, a mathematical relationship between excitation and speech components needs to be clarified. Note that at the moment of n^{th} sample generation process in the WaveNet's synthesis stage, $\hat{x}_n$ given in Eq. (6) can be treated as a known parameter because both LP coefficients, a_i, and previously reconstructed sample, x_{n-i} are estimated already. Thus, 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 that the constant summation to the second-order random variable only shifts with its offset and the shape of distribution remains the same, the relationship between mixture parameters of speech and excitation distributions can then be represented as follows:

$$\begin{aligned} \pi_i^x &= \pi_i^e, \\ \mu_i^x &= \mu_i^e + \hat{x}_n, \\ s_i^x &= s_i^e, \end{aligned} \tag{7}$$

where superscripts e and x denote the excitation and the speech, respectively.

Based on this observation, we propose an LP-WaveNet whose structure is illustrated in Fig. 1. Firstly, the mixture

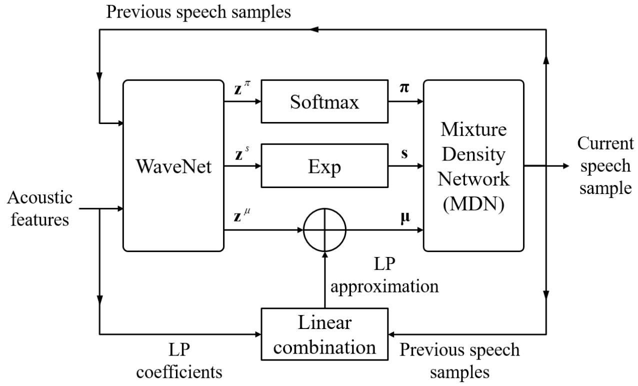

Fig. 1: Block diagram of the LP-WaveNet vocoder.

parameters of excitation signal are predicted by the WaveNet, and the LP approximation term, $\hat{x}_n$ is computed by the linear combination of previous speech samples weighted by LP coefficients. Then, the mixture parameters are defined as follows:

$$\begin{aligned} \boldsymbol{\pi} &= \text{softmax}(\boldsymbol{z}^{\pi}), \\ \boldsymbol{\mu} &= \mathbf{z}^{\mu} + \hat{x}_n, \\ \boldsymbol{s} &= \exp(\boldsymbol{z}^{s}). \end{aligned} \tag{8}$$

Since the constant summation guarantees the linearity, the weights of WaveNet can be successfully trained by a normal back-propagation process, e.g., *Adam* optimization [16]. To train the network, the discretized MoL distribution is computed, then the weights are optimized to minimize the NLL loss with respect to the speech signal.

Because the complicated spectrum modeling is embedded into the internal LP synthesis structure, the LP-WaveNet only need to model the excitation signal, which is easy to train. Moreover, because the ultimate training target is speech signal, it is free from the mismatch problem mentioned in Section 2.2. As a result, the LP-WaveNet is able to model the both excitation generation and LP synthesis filter processes jointly in a WaveNet structure.

4. EXPERIMENTS

4.1. Speech database and features

A phonetically and prosodically balanced speech corpus recorded by a Korean male professional speaker was used for the experiments. The speech signals were sampled at 16kHz with 16 bits quantization. In total, 2,500 utterances (about 3.2 hours) were used for training, 200 utterances were used for validation, and another 200 utterances were used for test, respectively.

The acoustic features included 40-dimensional line spectral frequencies (LSFs), logarithmic fundamental frequency (logF0), logarithmic energy, voicing flag, and 5-dimensional BAP. The BAP was estimated by the WORLD analysis [17]. The RAPT algorithm [18] was used to extract the logF0 and the voicing flag. The length of analysis window to extract the LSFs and the energy was set to 35-ms. All features were extracted in every 5-ms.

4.2. WaveNet vocoders

Total three WaveNet vocoding systems were tested.

- WN$_S$: MDN-WaveNet that models the speech signal.
- WN$_E$: MDN-WaveNet that models the excitation signal. The additional LP synthesis filter was applied to synthesize the speech waveform.
- WN$_{LP}$: Proposed LP-WaveNet.

For a fair comparison, 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. The weights were firstly initialized by the *Xavier* initializer [19], 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 4 GPUs, resulting in 80,000 samples per mini-batch. The networks were trained in 600,000 iterations.

4.3. Acoustic model

To test the performance of WaveNet vocoders in an SPSS system, a long-short term memory (LSTM) network-based acoustic model was used. In detail, the 3 feed-forward (FF) layers with 1,024 hidden nodes were used at the input side, and the 1 LSTM layer with 512 memory cells was used at the output side. The ReLu and linear activation functions were used at the hidden and output layers, respectively. The input vectors were composed of 210-dimensional linguistic features, and the output vectors were composed of 142-dimensional acoustic features, including their dynamic features (except the voicing flag). During synthesis, the maximum likelihood parameter generation (MLPG) [20] and the LSF-sharpening [21] algorithms were used as a post processing. All of WaveNet vocoders and acoustic models were implemented using the *PyTorch* framework [22].

4.4. Waveform generation via 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. When the WaveNet's output layer is a softmax layer, this noise can be controlled by adjusting the sharpness of waveform distribution [6, 23]. However, there's no

Table 1. Objective evaluation results of the various WaveNet vocoders with analysis and synthesis (A/S) and statistical parametric speech synthesis (SPSS) systems. The system with highest performance is represented in bold typeface.

	System	VUV	F0 RMSE	LSD	F-LSD
		(%)	(Hz)	(dB)	(dB)
A/S	WN$_S$	3.62	3.98	2.22	7.7
	WN$_E$	3.29	3.31	**1.98**	6.97
	WN$_{LP}$	**3.15**	**3.30**	2.05	**6.87**
SPSS	WN$_S$	**6.33**	15.55	5.01	11.35
	WN$_E$	6.35	15.23	**4.94**	11.39
	WN$_{LP}$	6.56	**15.17**	4.95	**11.28**

prior studies that using this solution on the MDN-WaveNet. In this study, we adjust the sharpness of waveform distribution by controlling the scale parameter 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 2 at only voiced region presents the best performance.

4.5. 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 SPSS 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 SPSS results are summarized in Table 1. Findings in the experimental results are: (1) The WN$_{LP}$ showed a better F0 contour modeling and waveform distribution modeling accuracies than the conventional WN$_S$ and WN$_E$ (F0 RMSE and F-LSD).; (2) The WN$_E$ showed the best performance on the spectrum modeling task, but its speech prediction error was worse than the WN$_{LP}$ (LSD and F-LSD). This is due to the fact that the process of generating final synthesized speech has not been considered in its training procedure.

Table 2. 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 4.81.

	STR	WN$_S$	WN$_E$	WN$_{LP}$
A/S	2.83±0.19	4.78±0.08	4.58±0.08	**4.84±0.11**
SPSS	2.80±0.12	**4.14±0.16**	3.67±0.20	4.04±0.12

To evaluate the perceptual quality of proposed system, the mean opinion score (MOS) listening test were performed. Total 12 native Korean listeners were asked to score the randomly selected 20 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 STRAIGHT-based synthesis system, i.e., STR, having same acoustic model with WaveNet vocoding systems was also included as a baseline system [4].

The MOS test results are summarized in Table 2. In the A/S system, all of WaveNet vocoders presented high quality synthetic sound by showing the MOS value over 4.00, whereas STR was not. Specifically, the proposed WN$_{LP}$ showed the best quality among the WaveNet vocoders. In the SPSS system, the quality of the WN$_S$ was better than the proposed WN$_{LP}$, but they still presented high quality synthetic sound exceeding 4.0 MOS. Moreover, the quality degradation of WN$_E$ by using predicted acoustic features was more perceptible than the other two WaveNet vocoders. We conjecture that this is due to the prediction mismatch between the excitation signal and the LP coefficients, which makes the system vulnerable to the prediction error.

5. 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 outperforms the conventional WaveNet systems both objectively and subjectively. Our future work will be a testing the idea of LP-WaveNet to other neural vocoders such as sample-RNN vocoder [5] or FFTNet vocoder [6].

Relationship to prior work - In the conventional systems, the vocal source is modeled by the neural vocoder without considering the final synthetic speech. Because of the mismatch between vocal source and vocal tract filter, their synthetic speech can create noisy artifacts. In the proposed system, the vocal tract filter is merged into the WaveNet structure, and the vocal source and vocal tract filter are jointly considered. Thus, we were able not only to improve the performance of WaveNet system, but also to solve the mismatch problem happened in the vocal source modeling systems.

6. REFERENCES

[1] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *INTERSPEECH*, 2017.

[2] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2017, pp. 712–718.

[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," in *Arxiv*, 2016.

[4] H. Kawahara, "STRAIGHT, exploration of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 5, pp. 349–353, 2006.

[5] Y. Ai, H. Wu, and Z. Ling, "Sample-RNN-based neural vocoder for statistical parametric speech synthesis," in *ICASSP*. IEEE, 2018, pp. 5659–5663.

[6] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFT-Net: a real-time speaker-dependent neural vocoder," in *The 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018.

[7] L. Juvela, V. Tsiaras, and B. Bollepalli, "Speaker-independent raw waveform model for glottal excitation," in *INTERSPEECH*, 2018.

[8] Y. Cui, X. Wang, L. He, and F. K. Soong, "A new glottal neural vocoder for speech synthesis," in *INTERSPEECH*, 2018.

[9] E. Song, K. Byun, and H.-G. Kang, "A neural excitation vocoder for statistical parametric speech synthesis systems," *Submitted to Pattern Recogn. Lett.*, 2018.

[10] T. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*. Prentice Hall Press, 2001.

[11] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering." *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[12] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 3, pp. 596–607, 2014.

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

[14] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications," *CoRR*, vol. abs/1701.05517, 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," *CoRR*, 2017.

[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] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions*, vol. 99-D, pp. 1877–1884, 2016.

[18] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Klein and K. K. Palival, Eds. Elsevier, 1995.

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

[20] 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.

[21] 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.

[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," 2017.

[23] W. Xin, L.-T. Jaime, T. Shinji, J. Lauri, and Y. Junichi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. ICASSP*, 2018.