ExcitNet Vocoder: A Neural Excitation Model for Parametric Speech Synthesis Systems

Eunwoo Song[1], Kyungguen Byun[2] and Hong-Goo Kang[2]

[1]*NAVER Corp., Seongnam, Korea*
[2]*Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea*
eunwoo.song@navercorp.com

Abstract—This paper proposes a WaveNet-based neural excitation model (ExcitNet) for statistical parametric speech synthesis systems. Conventional WaveNet-based neural vocoding systems significantly improve the perceptual quality of synthesized speech by statistically generating a time sequence of speech waveforms through an auto-regressive framework. However, they often suffer from noisy outputs because of the difficulties in capturing the complicated time-varying nature of speech signals. To improve modeling efficiency, the proposed ExcitNet vocoder employs an adaptive inverse filter to decouple spectral components from the speech signal. The residual component, i.e. excitation signal, is then trained and generated within the WaveNet framework. In this way, the quality of the synthesized speech signal can be further improved since the spectral component is well represented by a deep learning framework and, moreover, the residual component is efficiently generated by the WaveNet framework. Experimental results show that the proposed ExcitNet vocoder, trained both speaker-dependently and speaker-independently, outperforms traditional linear prediction vocoders and similarly configured conventional WaveNet vocoders.

Index Terms—Speech synthesis, WaveNet, ExcitNet

I. INTRODUCTION

Statistical parametric speech synthesis (SPSS) systems are popularly used for various applications, and much research has been performed to analyze the relationship between the accuracy of vocoding techniques and the quality of synthesized speech [1]–[4]. In the typical source-filter theory of speech production [5], the residual signal, i.e. source, is obtained by passing the speech signal through a linear prediction (LP) filter that decouples the spectral formant structure. To reduce the amount of information, the residual signal is approximated by various types of excitation model such as pulse or noise (PoN) [6], band aperiodicity (BAP) [7], [8], glottal excitation [9], [10], and time-frequency trajectory excitation (TFTE) models [11]. As parametric vocoding techniques have become more sophisticated, so the quality of synthesized speech has improved.

Recently, WaveNet-based waveform generation systems have attracted great attention in the speech signal processing community thanks to their high performance and ease of application [12]. In this type of system, the time-domain speech signal is represented as a discrete symbol sequence and its probability distribution is autoregressively modeled by stacked convolutional layers. By appropriately conditioning the acoustic parameters with input features, these systems have also been successfully adopted into neural vocoder structures [13]–[18]. By directly generating the time sequence of speech signals without utilizing parametric approximation, WaveNet-based systems provide superior perceptual quality to traditional linear predictive coding (LPC) vocoders [13].

However, the speech signals generated by a WaveNet vocoder often suffer from noisy outputs because of the prediction errors caused by adopting convolutional neural network (CNN) models. Due to difficulties in capturing the dynamic nature of speech signals, spectral distortion can increase, especially in the high frequency region. Using properties of the human auditory system [19], Tachibana et. al. introduce a perceptual noise-shaping filter as a pre-processing stage in the WaveNet training process [17]. Although this approach improves the perceived quality of the generated speech, its modeling accuracy is relatively low in unvoiced and transition regions. The reason for this can be found from the time-invariant limitation of the noise-shaping filter which is not appropriate for regions where phonetic information varies significantly.

To alleviate the aforementioned problem, we propose *ExcitNet*; a WaveNet-based neural excitation model for speech synthesis systems. The proposed system takes advantage of the merits of both the LPC vocoder and the WaveNet structure. In the analysis step, the LP-based adaptive predictor is used to decouple the spectral formant structure from the input speech signal [20]. The probability distribution of its residual signal, i.e., the excitation, is then modeled by the WaveNet framework. As the spectral structure represented by LP coefficients, or by their equivalents such as line spectral frequencies (LSFs), is changing relatively slowly, it is easy to model with a simple deep learning framework [4]. In addition, because the variation of the excitation signal is only constrained by vocal cord movement, the WaveNet training process becomes much simpler. Furthermore, we significantly improve the WaveNet modeling accuracy by adopting the improved time-frequency trajectory excitation (ITFTE) parameters as conditional features that effectively represent the degree of periodicity in the excitation signal [21]–[23].

In the speech synthesis step, an acoustic model designed using a conventional deep learning-based SPSS system first generates acoustic parameters from the given input text [4]. Those parameters are used to compose input conditional features and the WaveNet then generates the corresponding time sequence of the excitation signal. Finally, the speech signal is reconstructed by passing the generated excitation signal through the LP synthesis filter.

We investigate the effectiveness of the proposed ExcitNet vocoder not only in the analysis/synthesis but also in the SPSS frameworks. Our experiments verify the performance of various types of neural vocoders including a plain WaveNet [13], [14], a noise shaping filter-based WaveNet [17], and a proposed ExcitNet trained both speaker-dependently and speaker-independently. The synthesis quality of each system is investigated by varying the amount of the training data set, of which results could be usefully referred to when designing similarly configured WaveNet-based neural vocoding frameworks. Regarding the vocoder itself, in a perceptual listening test, the proposed system shows superiority over the our best prior parametric ITFTE vocoder with the same SPSS model structure.
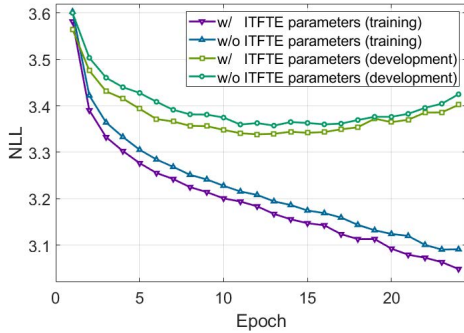
Fig. 1: Negative log-likelihood (NLL) obtained during the training process with (w/) and without (w/o) ITFTE parameters.
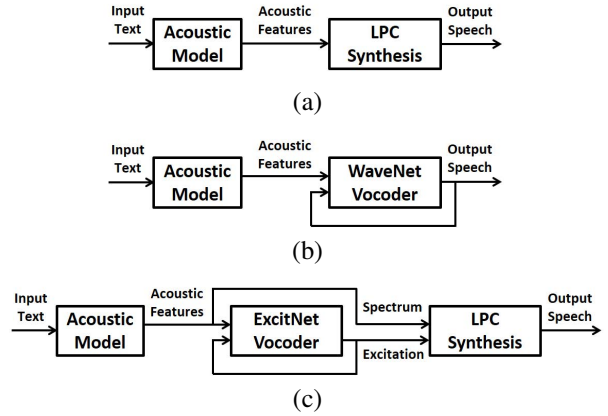

Fig. 2: Speech synthesis frameworks based on a conventional SPSS system with (a) an LPC vocoder, (b) a WaveNet vocoder, and (c) the proposed ExcitNet vocoder.

II. WAVENET VOCODER

The basic WaveNet is an autoregressive network which generates a probability distribution of waveforms from a fixed number of past samples [12]. The joint probability of the waveform is factorized by a product of conditional probabilities as follows:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t | x_1, ..., x_{t-1}), \qquad (1)$$

where $\mathbf{x} = \{x_1, ..., x_T\}$ denotes discrete waveform symbols compressed via μ-law companding transformation. Given an additional input $\mathbf{h}$, defined as the auxiliary features, the WaveNet is able to model the conditional distribution $p(\mathbf{x}|\mathbf{h})$ of the waveform [12]. By conditioning the model on other input variables, the output can be guided to produce waveforms with required characteristics. Typically, the original WaveNet uses linguistic features, fundamental frequency (F0), and/or speaker codes for the auxiliary condition [12]. More recent WaveNet vocoders utilize acoustic parameters directly extracted from the speech such as mel-filterbank energy, mel-generalized cepstrum, BAP, and F0 [13]–[18]. This enables the system to automatically learn the relationship between acoustic features and speech samples which results in superior perceptual quality over traditional LPC vocoders [13], [24].

III. EXCITNET VOCODER

Even though previous studies have indicated the technical possibility of introducing WaveNet-based vocoding, the systems often suffer from noisy outputs because of prediction errors in the WaveNet models [17]. Since it is still challenging for CNNs to fully capture the dynamic nature of speech signals, more sophisticated system architecture that can effectively remove the redundant structures needs to be designed.

In this research, we propose the ExcitNet vocoder; a neural excitation model for speech synthesis systems. In the proposed method, the redundant formant structure of the speech signal is removed using an LP analysis filter and the distribution of its residual signal, i.e., the excitation, is then modeled by a WaveNet framework.

A. Auxiliary Features Employing ITFTE Parameters

Similar to the conventional WaveNet vocoders, the input auxiliary features are composed of the spectral parameters, i.e., LSF, F0, v/uv, and gain. To further improve training efficiency, we also adopt ITFTE parameters [11]. Note that the TFTE represents the spectral shape of excitation along the frequency axis and the

evolution of this shape along the time axis. To obtain the harmonic excitation spectrum, i.e., a slowly evolving waveform (SEW), each frequency component of the TFTE is low-pass filtered along the time-domain axis. Beyond the cut-off frequency, the remaining noise spectrum, i.e., a rapidly evolving waveform (REW), is obtained by subtracting the SEW from the TFTE.

Employing the SEW and REW enables to effectively represent a periodicity distribution of the excitation [21]–[23]. Therefore, adding these parameters to the auxiliary features helps improve the modeling accuracy of the WaveNet. Fig. 1 shows the negative log-likelihood obtained during the training process, of which result confirms that composing auxiliary features with ITFTE parameters enables to reduce both training and development errors as compared to the process without ITFTE parameters.

B. Speech Synthesis Using ExcitNet Vocoder

Fig. 2-(c) depicts a synthesis framework of the ExcitNet vocoder in which the architecture is combined from the conventional LPC vocoder presented in Fig. 2-(a) and the WaveNet vocoder presented in Fig. 2-(b). To obtain the auxiliary features, we adopt our previous SPSS system based on the ITFTE vocoder [4]. From the given text input, an *acoustic model*[1] first estimates the acoustic parameters such as the LSF, F0, v/uv, gain, SEW, and REW, which are then used to compose the auxiliary features. By inputting these auxiliary features, the ExcitNet generates the time sequence of the excitation signal. Finally, the speech signal is reconstructed by passing the excitation signal through the LP synthesis filter formed by the generated LSFs.

IV. EXPERIMENTS

A. Experimental Setup

To investigate the effectiveness of the proposed algorithm, we trained neural vocoding models using two different methods:

- SD: speaker-dependent training model
- SI: speaker-independent training model

Two phonetically and prosodically rich speech corpora were used to train the acoustic model and the SD-ExcitNet vocoder. Each

---

[1]In this framework, the acoustic model consists of multiple feedforward and long short-term memory layers, trained to represent a nonlinear mapping function between linguistic input and acoustic output parameters. More detail about training the acoustic model is provided in the following section.

TABLE I: Objective test results with respect to the different neural vocoders: In the SD systems, the amount of the training database is set to 1, 3, 5, and 7 hours. The systems that returned the smallest errors are in bold font.

| Speaker | System | LSD (dB) | | | | | F0 RMSE (Hz) | | | | |
| | | SD | | | | SI | SD | | | | SI |
		1 h	3 h	5 h	7 h		1 h	3 h	5 h	7 h	
KRF	WN	4.21	4.19	4.18	4.13	4.18	32.61	31.69	31.56	31.49	32.30
	WN-NS	4.15	4.12	4.07	4.01	4.06	31.96	31.75	31.52	31.38	32.23
	ExcitNet	**4.11**	**4.09**	**4.05**	**3.99**	**4.04**	**31.44**	**31.43**	**31.37**	**31.29**	**31.88**
KRM	WN	3.73	3.72	3.69	3.67	3.70	12.60	12.39	12.05	12.05	13.96
	WN-NS	3.54	**3.46**	**3.41**	3.41	3.46	**12.32**	12.16	11.97	11.97	13.34
	ExcitNet	**3.53**	**3.46**	**3.41**	**3.40**	**3.45**	12.72	**12.10**	**11.93**	**11.93**	**12.96**

TABLE II: Number of utterances in different sets.

Speaker	Training	Development	Test
KRF	3,826 (7 h)	270 (30 min)	270 (30 min)
KRM	2,294 (7 h)	160 (30 min)	160 (30 min)

TABLE III: LSD (dB) test results measured in unvoiced and transition regions: The systems that returned the smallest errors are in bold font.

| Speaker | System | SD | | | | SI |
		1 h	3 h	5 h	7 h	
KRF	WN	4.19	4.16	4.15	4.09	4.18
	WN-NS	4.26	4.18	4.12	4.03	4.06
	ExcitNet	**4.15**	**4.10**	**4.06**	**3.98**	**4.04**
KRM	WN	3.95	3.96	3.92	3.92	3.70
	WN-NS	4.41	3.95	3.88	3.88	3.46
	ExcitNet	**3.91**	**3.83**	**3.76**	**3.76**	**3.45**

corpus was recorded by professional Korean female (KRF) and Korean male (KRM) speakers. The speech signals were sampled at 24 kHz, and each sample was quantized by 16 bits. Table II shows the number of utterances in each set. To train the SI-ExcitNet [14], speech corpora recorded by five Korean female and five Korean male speakers were also used. In total, 6,422 (10 h) and 1,080 (1.7 h) utterances were used for training and development, respectively. The speech samples recorded by the same KRF and KRM speakers not included in the SI data set were used for testing.

To compose the acoustic feature vectors, the spectral and excitation parameters were extracted using the ITFTE vocoder. The estimated 40-dimensional LP coefficients were converted into the LSFs for training. To prevent unnatural spectral peaks in the LP analysis/synthesis filter, each coefficient $(a_i, i = 1, ..., 40)$ was multiplied by a linear prediction bandwidth expansion factor (0.981^i) [4]. In contrast, the 32-dimensional SEW and 4-dimensional REW coefficients were extracted for the excitation parameters. The F0, gain, and v/uv information were also extracted. The frame and shift lengths were set to 20 ms and 5 ms, respectively.

In the acoustic modeling step, the output feature vectors consisted of all the acoustic parameters together with their time dynamics [25]. The corresponding input feature vectors included 356-dimensional contextual information consisting of 330 binary features of categorical linguistic contexts and 26 numerical features of numerical linguistic contexts. Before training, both input and output features were normalized to have zero mean and unit variance. The hidden layers consisted of three feedforward layers (FFs) with 1,024 units and one unidirectional long short-term memory (LSTM) layer with 512 memory blocks. The FFs and LSTM were connected to the linguistic input layer and the acoustic output layer, respectively. The weights were initialized by *Xavier* initializer [26] and trained using the *backpropagation through time* (BPTT) algorithm with *Adam* optimizer [27], [28]. The learning rate was set to 0.02 for the first 10 epochs, 0.01 for the next 10 epochs, and 0.005 for the remaining epochs.

In the ExcitNet training step, all the acoustic parameters were used to compose the input auxiliary feature vectors, and they were duplicated from a frame to the samples to match the length of the excitation signals. Before training, they were normalized to have zero mean and unit variance. The corresponding excitation signal, obtained by passing the speech signal through the LP analysis filter, was normalized in a range from -1.0 to 1.0 and quantized by 8-bit μ-law compression. We used a one-hot vector to represent

the resulting discrete symbol. The ExcitNet architecture had three convolutional blocks, each of which had ten dilated convolution layers with dilations of 1, 2, 4, 8, and so on up to 512. The number of channels of dilated causal convolution and the 1×1 convolution in the residual block were both set to 512. The number of 1×1 convolution channels between the skip-connection and the softmax layer was set to 256. The learning rate was 0.0001, the batch size was 30,000 (1.25 sec), the weights were initialized using the Xavier initializer, and Adam optimizer was used.

In the synthesis step, the mean vectors of all acoustic feature vectors were predicted by the acoustic model, and a speech parameter generation (SPG) algorithm was then applied to generate smooth trajectories for the acoustic parameters [29]. Because the acoustic model could not predict the variance used for the SPG algorithm, we used the pre-computed global variances of acoustic features from all training data [30]. By inputting these features, the ExcitNet vocoder generated a discrete symbol of the quantized excitation signal, and its dynamic was recovered via μ-law expansion. Finally, the speech signal was reconstructed by applying the LP synthesis filter to the generated excitation signal. To enhance spectral clarity, an LSF-sharpening filter was also applied to the generated spectral parameters [4].

B. Objective Test Results

To evaluate the performance of the proposed system, the results were compared to those of conventional systems based on a WaveNet vocoder (WN) [13] and on a WaveNet vocoder with a noise-shaping method (WN-NS) [17]. The WaveNet architectures and auxiliary features were the same with those of the proposed system, but the target outputs differed from each other. The target of the WN system was the distribution of the speech signal; that of the WN-NS system was the distribution of a noise-shaped residual signal. A time-invariant spectral filter in the latter system was obtained by averaging all spectra extracted from the training data [31]. This filter was used to extract the residual signal before the training process, and its inverse filter was applied to reconstruct the speech signal in the synthesis step.

TABLE IV: Subjective preference test results (%) of synthesized speech for the KRF speaker: The systems that achieved significantly better preferences ($p < 0.01$) are in bold.

KRF		WN	WN-NS	ExcitNet	Neutral	p-value
		6.8	**64.1**	-	29.1	$< 10^{-30}$
SD		7.3	-	**83.6**	9.1	$< 10^{-49}$
		-	12.7	**58.2**	29.1	$< 10^{-17}$
		12.7	**66.8**	-	20.5	$< 10^{-22}$
SI		8.6	-	**73.6**	27.7	$< 10^{-35}$
		-	19.5	**38.6**	41.8	$< 10^{-3}$

TABLE V: Subjective preference test results (%) of synthesized speech for the KRM speaker: The systems that achieved significantly better preferences ($p < 0.01$) are in bold.

KRM		WN	WN-NS	ExcitNet	Neutral	p-value
		11.8	**60.5**	-	27.7	$< 10^{-19}$
SD		17.3	-	**77.7**	5.0	$< 10^{-24}$
		-	16.4	**73.6**	10.0	$< 10^{-22}$
		27.3	**48.6**	-	24.1	$< 10^{-3}$
SI		13.6	-	**75.5**	10.9	$< 10^{-27}$
		-	17.3	**63.6**	19.1	$< 10^{-15}$

In the test, distortions between the original speech and the synthesized speech were measured by log-spectral distance (LSD; dB) and F0 root mean square error (RMSE; Hz). Table I shows the LSD and F0 RMSE test results, with respect to the different neural vocoders. The findings can be analyzed as follows: (1) As the amount of training database increased in the SD systems, the overall estimation performances gradually improved for both the KRF and KRM speakers; (2) In both the SD and SI systems, the vocoders with spectral filters (WN-NS and ExcitNet) achieved much more accurate speech reconstruction than the WN vocoder, which confirms that decoupling the formant component of the speech signal is beneficial to the modeling accuracy of the remaining signal; (3) Among the vocoders with spectral filters, ExcitNet's adaptive spectral filter helped to reconstruct a more accurate speech signal compared to the conventional system using a time-invariant filter (WN-NS). Since the average spectral filter in the WN-NS vocoder is biased towards voiced components, it is not optimal for unvoiced and transition regions which results in unsatisfactory results in those areas. This was clearly observed in the LSD results measured in the unvoiced and transition regions[2], as shown in Table III.

C. Subjective Test Results

To evaluate the perceptual quality of the proposed system, A-B preference and mean opinion score (MOS) listening tests were performed[3]. In the preference tests, 12 native Korean listeners were asked to rate the quality preference of the synthesized speech. In total, 20 utterances were randomly selected from the test set and were then synthesized using the three types of vocoder. Note that the input auxiliary condition features were obtained by the conventional SPSS system. Table IV and Table V show the preference test results for the KRF and KRM speakers, respectively, and confirm that the perceptual quality of the speech synthesized by the ExcitNet vocoder is significantly better than those of the conventional WaveNet vocoders.

[2]As the F0 does not exists in the unvoiced and transition components, we only compared the LSD results in those regions.

[3]Generated audio samples are available at the following url: https://sewplay.github.io/demos/excitnet/
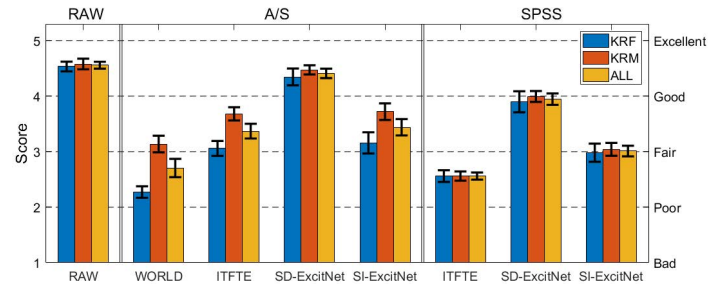

Fig. 3: Subjective MOS test results with 95% confidence interval for previous and proposed systems. In the analysis/synthesis (A/S) and SPSS groups, acoustic features extracted from recorded speech and generated from the acoustic model, respectively, were used to compose the input auxiliary features.

Setups for testing the MOS were the same as for the preference tests except that listeners were asked to make quality judgments of the synthesized speech using the following five possible responses: 1 = Bad; 2 = Poor; 3 = Fair; 4 = Good; and 5 = Excellent. To verify vocoding performance, speech samples synthesized by conventional vocoders such as ITFTE and WORLD (D4C edition [32]) were also included. The test results shown in Fig. 3 confirm the effectiveness of each system in several ways. First, the SI-ExcitNet performed similarly to the ITFTE vocoder but performed much better than the WORLD system in analysis/synthesis. Across all systems, the SD-ExcitNet provided the best perceptual quality (4.35 and 4.47 MOS for the KRF and KRM speakers, respectively). Next, owing to the difficulty of representing high-pitched female voices [23], the MOS results for the KRF speaker were worse than those for the KRM speaker in the SI vocoders (WORLD, ITFTE, and SI-ExcitNet). On the other hand, the results for the KRF speaker in the SD-ExcitNet were similar to those for the KRM speaker, which implies that modeling speaker-specific characteristics is necessary to represent high-pitched voices effectively. Lastly, in terms of SPSS, both the SD- and SI-ExcitNet vocoders provided much better perceptual quality than the parametric ITFTE vocoder. Although the acoustic model generated overly smoothed speech parameters, ExcitNet was able to alleviate the smoothing effect by directly estimating time-domain excitation signals. Consequently, the SPSS system with the proposed SD-ExcitNet vocoder achieved 3.78 and 3.85 MOS for the KRF and KRM speakers, respectively; the SI-ExcitNet vocoder achieved 2.91 and 2.89 MOS for the KRF and KRM speakers, respectively.

V. CONCLUSION

This paper proposed the ExcitNet vocoder, built with hybrid architecture that effectively combined the merits of WaveNet and LPC vocoding structures. By decoupling the spectral formant structure from the speech signal, the proposed method significantly improved the modeling accuracy of the excitation signal using the WaveNet vocoder. The experimental results verified that the proposed ExcitNet system, trained either speaker-dependently or speaker-independently, performed significantly better than traditional LPC vocoders as well as similarly configured conventional WaveNet vocoders. Future research includes integrating the ExcitNet vocoder into speech synthesis systems that use an end-to-end approach.

REFERENCES

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

[2] 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.

[3] Q. Hu, Z. Wu, K. Richmond, J. Yamagishi, Y. Stylianou, and R. Maia, "Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning," in *Proc. INTERSPEECH*, 2015, pp. 854–858.

[4] 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.

[5] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, 1999, pp. 2347–2350.

[7] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. ICASSP*, 1997, pp. 1303–1306.

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

[9] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 1, pp. 153–165, 2011.

[10] M. Airaksinen, B. Bollepalli, J. Pohjalainen, and P. Alku, "Glottal vocoding with frequency-warped time-weighted linear prediction," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 446–450, 2017.

[11] E. Song, Y. S. Joo, and H.-G. Kang, "Improved time-frequency trajectory excitation modeling for a statistical parametric speech synthesis system," in *Proc. ICASSP*, 2015, pp. 4949–4953.

[12] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

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

[14] 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.

[15] Y.-J. Hu, C. Ding, L.-J. Liu, Z.-H. Ling, and L.-R. Dai, "The USTC system for blizzard challenge 2017," in *Proc. Blizzard Challenge Workshop*, 2017.

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

[17] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation," in *Proc. ICASSP*, 2018, pp. 5664–5668.

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

[19] M. R. Schroeder, B. S. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of Acoust. Soc. of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[20] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 27, no. 3, pp. 247–254, 1979.

[21] E. Song and H.-G. Kang, "Deep neural network-based statistical parametric speech synthesis system using improved time-frequency trajectory excitation model," in *Proc. INTERSPEECH*, 2015, pp. 874–878.

[22] E. Song, F. K. Soong, and H.-G. Kang, "Improved time-frequency trajectory excitation vocoder for DNN-based speech synthesis," in *Proc. INTERSPEECH*, 2016, pp. 874–878.

[23] ——, "Perceptual quality and modeling accuracy of excitation parameters in DLSTM-based speech synthesis systems," in *Proc. ASRU*, 2017, pp. 671–676.

[24] 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.

[25] 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.

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

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

[28] 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

[29] 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.

[30] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[31] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.

[32] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech commun.*, vol. 84, pp. 57–65, 2016.