MODELING-BY-GENERATION-STRUCTURED NOISE COMPENSATION ALGORITHM FOR GLOTTAL VOCODING SPEECH SYNTHESIS SYSTEM

Min-Jae Hwang[1]*, Eunwoo Song*[1,2]*, Kyungguen Byun*[1] *and Hong-Goo Kang*[1]

[1]Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
[2]NAVER Corp., Seongnam-si, Gyeonggi-do, Korea

ABSTRACT

This paper proposes a novel noise compensation algorithm for a glottal excitation model in a deep learning (DL)-based speech synthesis system. To generate high-quality speech synthesis outputs, the balance between harmonic and noise components of the glottal excitation signal should be well-represented by the DL network. However, it is hard to accurately model the noise component because the DL training process inevitably results in statistically smoothed outputs; thus, it is essential to introduce an additional noise compensation process. We propose a modeling-by-generation structure-based noise compensation method that the missing noise component in the generated glottal signal is directly extracted and parameterized during the entire training process. By modeling the noise component using the additional DL network, the proposed system successfully compensates the missing noise component. Objective and subjective test results confirm that the synthesized speech with the proposed noise compensation method is superior to that with conventional methods.

***Index Terms*—** Text-to-speech, speech synthesis, glottal vocoder, glottal excitation model

1. INTRODUCTION

The emergence of the glottal excitation model with the deep learning (DL)-based speech synthesis frameworks has significantly improved the quality of parametric speech synthesis systems [1–4]. In the glottal vocoding system, a pitch-dependent excitation signal is first obtained by applying a linear prediction (LP) inverse filter to an input speech signal [5, 6], and then the temporal variation of the excitation signal is trained and generated via DL techniques. The synthetic speech quality of the glottal excitation model is better than that of conventional band-aperiodicity (BAP)-based approaches [7], however, its synthesized speech is often unnaturally buzzy because of overly smoothed glottal signals.

To address the aforementioned problem, various types of noise compensation algorithms have been introduced. For example, a harmonic-to-noise ratio (HNR) of the generated glottal signal was compared with that of the original glottal signal, and then the noise level was adjusted to have the same HNR values [8]. Alternatively, Airaksinen et al [9] estimated the noise component by subtracting the median filter (MF) output from the original glottal signal with the assumption that the smoothed glottal output caused by the training process can be simulated with an MF in the analysis step . This noise component was also trained together with the glottal signals and added to the generated glottal signal in the synthesis step.

Although both methods are somewhat advantageous in terms of improving the naturalness of the synthesized speech, the inaccurate definition of the missing noise component often produces unwanted noisy artifacts. In the case of HNR-based method, the measurement errors of HNR in both the training and synthesis stages cause inappropriate compensation of missing noise component. In the MF-based approach, there is a mismatch in the smoothing effects that occur with the MF and the training process. Therefore, the generated speech with the conventional glottal vocoding systems tends to be buzzy or noisy when the compensated noise-level is too low or high, respectively, thus, the noise compensation method in the glottal vocoding framework should be carefully considered.

In this paper, we propose a modeling-by-generation (MbG)-structured noise compensation method that directly models the missing noise component during the entire training process. By considering the fact that the generated glottal signal mainly represents the harmonic-like component of the glottal excitation, we define the noise component as the difference between the original glottal signal obtained from the recorded speech and the smoothed glottal signal generated by the trained model. To improve modeling accuracy, the noise component is first parameterized by line spectral frequencies (LSFs) and pulse-wise HNRs to represent its spectral envelope and pulse-wise energy, respectively. Those parameters are then trained/generated by an additional DL network, i.e., a noise model, and used to reconstruct the noise component in the synthesis stage. Finally, the noise component is phase-aligned and added to the generated glottal signal.

2. GLOTTAL VOCODER AND PARAMETRIC SPEECH SYNTHESIS SYSTEM

The left part of Fig. 1 depicts the conventional glottal vocoding system. The glottal feature (GF) vector is composed of a time sequence of the glottal excitation signal[1]; whereas

This research was supported by NAVER Corp., Seongnam-si, Gyeonggi-do, Korea

[1] Before training the glottal model, the 2-pitch-period windowed glottal pulses, having glottal closure instants (GCI) at the middle and at both ends, are centered and zero-padded to have fixed-dimensional network output [2].

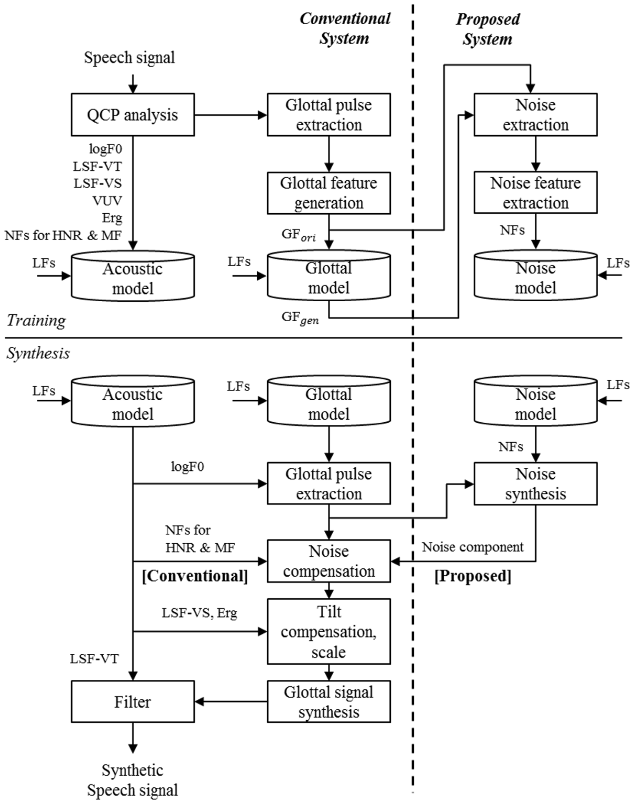

Fig. 1: Block diagram of the speech synthesis system using the glottal vocoder with MbG-based noise compensation.

the acoustic feature (AF) vectors consist of the speech parameters including: (1) quasi-closed phase (QCP) inverse filtering-based LSFs representing the vocal tract (VT) system (LSF–VT) [6], (2) log-fundamental frequency (logF0), (3) frame-level energy (Erg), (4) voicing information (VUV), and (5) low-order vocal source LSFs (LSF–VS) representing the spectral tilt of the glottal excitation signal.

The AFs also include additional parameters that are needed to compensate for the noise portion of the glottal signal, which varies depending on the type of noise compensation method. In the HNR-based noise compensation approach (HNR–NC), the HNRs of several frequency-bands are used for the additional noise-related features. In the MF-based method (MF–NC), the noise component is defined by the residual signal of the MF output, then parameterized into noise LSFs and energy terms to represent the spectral shape and gain of the noise component, respectively.

In the synthesis stage, the 2 pitch period glottal pulses are generated by the trained glottal model and they are weighted with a cosine window. The missing noise component estimated by the corresponding noise compensation approach is then added to the generated glottal pulse. In the HNR–NC approach, the noise component is determined by the gap of HNRs extracted from the glottal pulse and predicted by the acoustic model. In the MF–NC approach, the noise component is reconstructed by applying spectral shaping with the noise LSFs and adjusting the gain to uniformly distributed white noise signals. After performing the noise compensation process, the spectral tilt of the glottal pulse is adjusted to reduce the high-frequency loss that occurs in the glottal pulse generation process. Finally, the glottal excitation signal is constructed by applying a pitch-synchronous overlap-add method, and a single frame of speech signal is synthesized by filtering the glottal excitation signal through the VT filter reconstructed by the generated LSF–VT coefficients.

3. MODELING-BY-GENERATION-BASED NOISE COMPENSATION METHOD

Even though previous studies indicate the technical potentiality of introducing a glottal vocoding synthesis system, it is still challenging to derive the maximum advantages from them because the noise compensation method is typically designed with a heuristic definition. In this section, we propose an MbG-structured noise compensation method that presents a very effective modeling performance for the missing noise component. The right part of Fig. 1 illustrates the proposed glottal vocoding system with the MbG-structured noise compensation method.

In the training stage, the missing noise component is directly obtained by subtracting the generated GF from the original GF estimated by the recorded speech. This component is parameterized into noise features (NFs), and their statistical characteristics are trained by the DL-based noise model. In synthesis stage, the noise component is synthesized and compensated to generate glottal pulse. The detailed descriptions are introduced below.

3.1. Training stage

Let us assume that the original glottal pulse, $\mathbf{g}_o$, contains both harmonic and noise components, and the generated glottal pulse, $\mathbf{g}_g$, only represents noise-removed harmonic component as follows:

$$\mathbf{g}_o = \mathbf{h} + \mathbf{n}, \tag{1}$$

$$\mathbf{g}_g = \alpha\mathbf{h}, \tag{2}$$

where the vector sequence $\mathbf{h}$ and $\mathbf{n}$ denote the harmonic and noise components, respectively, and α denotes a scaling factor to compensate the effect of normalization processes before training and after generating the glottal signals. By assuming that the harmonic and noise components are uncorrelated as follows:

$$\mathbf{n}^T\mathbf{h} = 0, \tag{3}$$

the scaling factor α can be estimated by calculating the cross-correlation between the original and generated glottal pulses as follows:

$$\alpha = \mathbf{g}_g^T\mathbf{g}_g/\mathbf{g}_o^T\mathbf{g}_g. \tag{4}$$

Using Eqs. (1)–(4), the noise component can be extracted as follows:

$$\mathbf{n} = \mathbf{g}_o - \frac{1}{\alpha}\mathbf{g}_g. \tag{5}$$

To improve modeling accuracy, the noise component in Eq. (5) is parameterized by LSFs and a pulse-wise HNR to concentrate the spectral information and gain information, respectively, and then it is used to compose the NF vectors. The pulse-wise HNR is calculated by the energy ratio between harmonic and noise components as follows:

$$HNR = \frac{E[\mathbf{h}^2]}{E[\mathbf{n}^2]}. \tag{6}$$

3.2. Synthesis stage

In synthesis stage, a sequence of random noise is generated, and its frequency response, $N(\omega)$, is shaped to have a magnitude response corresponding to the target spectrum as follows:

$$\widehat{N}(\omega) = \frac{H_t(\omega)}{H_n(\omega)} \cdot N(\omega), \tag{7}$$

where $H_n(\omega)$ denotes an autoregressive (AR) spectrum of the random noise sequence; $H_t(\omega)$ denotes the target AR spectrum obtained by the generated noise LSFs; $\widehat{N}(\omega)$ denotes the shape-adjusted noise spectrum. Additionally, its time-domain sequence, $\widehat{\mathbf{n}}$, is tapered by the cosine window to match the windowing process in the GF extraction as described in Section 2. Then, the noise gain is adjusted to be matched with the generated pulse-wise HNR as follows:

$$\widetilde{\mathbf{n}} = \sqrt{\frac{HNR_n}{HNR_t}} \cdot \widehat{\mathbf{n}}_w, \tag{8}$$

where HNR_n and HNR_t denote the HNR extracted from the generated glottal pulse and that generated by the trained noise model, respectively; $\widehat{\mathbf{n}}_w$ and $\widetilde{\mathbf{n}}$ denote the tapered noise sequence and the target noise component, respectively. To prevent a critical phase mismatch in the generated glottal signal, a high-pass filter with a 2-kHz cut-off frequency is also applied to the target noise component [9].

Note that the noise parametrization and synthesis methods are similar to those of the MF–NC approach. However, since the definition and parameterization processes of the noise component fully depend on the glottal pulses of the original and generated ones, the noise loss modeling based on the proposed approach is highly adaptive to the glottal model compared to the one used in the MF-based approach. Thus, the noise component obtained by the noise model is specialized to capture a stochastic variation of the glottal pulse being lost in the glottal modeling process.

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 16 kHz, and each sample was quantized by 16 bits. In total, 2,500 utterances (about 3 hours) were used for training, 200 utterances were used for validation, and another 200 utterances that were not included in either the training or validation steps were used for testing.

Table 1. Speech features and their dimensions including Δ and $\Delta\Delta$ values for acoustic, glottal and noise models.

DL models	Output features	dim.	Δ dim.
Acoustic model	LSF–VT	30	90
	logF0	1	3
	Erg	1	3
	VUV	1	1
	LSF–VS	10	30
Glottal model	Glottal feature	400	400
Noise model	Noise LSFs	15	45
	Pulse-wise HNR	1	3

Table 2. Network architectures.

Type of layers	Acoustic model	Glottal model	Noise model
Input	LFs	LFs	LFs
FF (units × layers)	1024 × 2	512 × 3	512 × 2
LSTM (units × layers)	512 × 2	256 × 1	256 × 1
Output	AFs	GF	NFs

In the analysis step, the frame length was set to 20 ms, and the AFs, GF, and NFs were extracted every 5 ms. Table 1 summarizes the specification of all the features used in the experiments. The 30-dimensional LSF–VT [6], log-fundamental frequency (logF0) [10], frame-level energy (Erg), voicing information (VUV), and 10-dimensional LSF–VS were extracted for the AFs. The 400-dimensional time sequence of the glottal signal was used as the GF; whereas the 15-dimensional noise LSFs and a 1-dimensional pulse-wise HNR were extracted for the NFs.

4.2. DL-based model training and synthesis

Table 2 summarizes the network architectures used in the experiments. In the acoustic model, the AFs with their time-dynamics were composed of 142-dimensional output vectors [11]. The corresponding input feature vectors included 210-dimensional contextual information consisting of 203 binary features for categorical linguistic contexts and 7 numerical features for numerical linguistic contexts. The hidden layers consisted of 2 feed-forward (FF) layers with 1,024 units and 2 long short-term memory (LSTM) layers with 512 memory blocks. The hyperbolic tangent and linear activation functions were used in the hidden and output layers, respectively. The weights were first initialized by using a *Xavier* initializer [12], and then trained by using a *back-propagation through time* procedure with an *Adam* optimizer [13, 14].

In the glottal model, the linguistic input vectors were same as those of the acoustic model[2]; whereas the 400-dimensional

[2] In the glottal model, the linguistic features were used as the input vector to prevent the possible training mismatch or error propagation when employing the generated AF as an input [3, 15].

Table 3. LSD (dB) results for the architectures of the various noise compensation algorithms.

HNR–NC	MF–NC	MbG
8.55	7.93	7.61

GF were composed of the output vectors. There were 3 FF layers with 512 units and a single LSTM layer with 256 memory cells. The initialization and training methods were the same as those of the acoustic model.

The training procedures of the noise model were also similar with those of the acoustic and glottal models, but the 48-dimensional NFs were composed of the output vectors. The hidden layers consisted of 2 FF layers with 512 units and a single LSTM layer with 256 memory cells. For all methods, the training and test procedures were implemented using the *TensorFlow* framework [16].

In the synthesis step, the mean vectors of all the output features were predicted by the trained models. With precomputed global variances of output features from all the training data [17], a speech parameter generation algorithm was applied to generate a smooth trajectory of acoustic and noise features [18]. To synthesize the glottal excitation signal, the two pitch period glottal pulses were first synthesized by the generated GF and logF0, and then the noise compensation and spectral tilt compensation modules were applied to the glottal pulse. By pitch-synchronously constructing the glottal excitation signal, a speech signal was synthesized with the generated LSF–VT and the glottal excitation signals. To enhance spectral clarity, LSF-sharpening and formant enhancement filters were also applied to the generated spectral parameters [19, 20].

4.3. Objective and subjective evaluation results

To objectively evaluate the proposed noise compensation method, i.e., MbG, a log-spectral distance (LSD) of energy-normalized glottal pulses between the original and compensated ones were measured. The additional two glottal vocoding systems with different noise compensation methods such as HNR–NC and MF–NC were also included.

For the HNR–NC approach, the 5-dimensional frequency-band-wise HNRs were used. For the MF–NC approach, the noise component was extracted using a 4-ms median filter and a 2-kHz high-pass filter, and was parameterized into 15-dimensional noise LSFs and energy terms. In total, 142 and 172-dimensional AFs were used in the HNR–NC and MF–NC approaches, respectively.

The LSD results shown in Table 3 verify that the compensated glottal pulse with the proposed method had significantly smaller errors than those with conventional methods. As the reconstruction accuracy of the glottal signal is closely related to the quality of synthesized speech, the perceived quality of synthesized speech from the proposed system is expected to be better than that from the baseline systems.

Table 4. Subjective preference test results (%) between the speech samples. The systems that achieved significantly better preference at the $p < 0.01$ level are in bold font.

STR	HNR–NC	MF–NC	MbG	No prefer.	p-value
–	3.3	**87.5**	–	9.2	$< 10^{-79}$
–	–	5.4	**47.1**	47.5	$< 10^{-21}$
25.8	–	**64.6**	–	9.6	$< 10^{-10}$
17.9	–	–	**75.4**	6.7	$< 10^{-23}$
72.1	10.0	–	–	17.9	$< 10^{-34}$
–	1.7	–	**93.3**	5.9	$< 10^{-113}$

Table 5. Subjective MOS test results with a 95% confidence interval for the architectures of the various noise compensation algorithms.

STR	HNR–NC	MF–NC	MbG
2.90 ± 0.10	2.16 ± 0.13	3.20 ± 0.13	**3.72 ± 0.11**

To evaluate the perceptual quality of the proposed system, an A-B preference test and a mean opinion score (MOS) listening tests were performed. In each comparison of the preference tests, 12 native Korean listeners were asked to rate the quality preference using randomly selected 20 synthesized utterances from test set. In addition to HNR–NC and MF–NC, the STRAIGHT-based synthesis system, i.e., STR, was also included as a baseline system [7]. The preference test results shown in Table 4 show that the listeners preferred the proposed system over the conventional systems.

The setups for the MOS test were the same as those for the preference test, except that listeners were asked to make quality judgments about the synthesized speech using the following possible responses: 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent. Table 5 shows the MOS test results, which confirms that the proposed system provides much better perceptual quality than the baseline systems.

5. CONCLUSION

In this paper, we have introduced an MbG-structured noise compensation method for the glottal vocoding speech synthesis system. By directly modeling the smoothing impact to the glottal excitation signal throughout the entire training process, the proposed system successfully compensated the characteristic of noise component caused by statistical averaging in the training process. The experimental results verified that the proposed system was superior to conventional glottal vocoding systems, both objectively and subjectively.

Relationship to prior work - In the conventional glottal vocoding systems, the noise components that are needed to be compensated in the speech synthesis stage were heuristically determined by adopting an HNR or residual signal through an MF process. However, they were inappropriate for accurately compensating noise level, so the synthetic sound was often too buzzy or noisy. By directly modeling the missing noise component from the difference between the glottal pulses of the original and generated ones and by including it in the entire training process, we were able to construct a glottal model-adaptive noise compensation method.

6. REFERENCES

[1] 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.

[2] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Proc. ICASSP*, 2016, pp. 5120–5124.

[3] L. Juvela, X. Wang, S. Takaki, M. Airaksinen, J. Yamagishi, and P. Alku, "Using text and acoustic features in predicting glottal excitation waveforms for parametric speech synthesis with recurrent neural networks," in *Proc. INTERSPEECH*, 2016, pp. 2283–2287.

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

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

[6] 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.

[7] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE trans. Inf. Syst.*, vol. 90, no. 1, pp. 325–333, 2007.

[8] 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.

[9] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN-a full-band glottal vocoder for statistical parametric speech synthesis." in *Proc. INTERSPEECH*, 2016, pp. 2473–2477.

[10] D. Thomas and D. Thierry, "Glottal closure and opening instant detection from speech signals," in *Proc. INTERSPEECH*, 2009.

[11] 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.

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

[13] 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.

[14] 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

[15] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system," in *Proc. INTERSPEECH*, 2017, pp. 1368–1372.

[16] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

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

[18] 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.

[19] 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.

[20] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 1, pp. 59–71, 1995.