Interspeech 2018
2-6 September 2018, Hyderabad

A unified framework for the generation of glottal signals in deep learning-based parametric speech synthesis systems

Min-Jae Hwang[1,2*], *Eunwoo Song*[1,2], *Jin-Seob Kim*[2] *and Hong-Goo Kang*[1]

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

hmj234@dsp.yonsei.ac.kr, eunwoo.song@navercorp.com, paul.jskim@navercorp.com, hgkang@yonsei.ac.kr

Abstract

In this paper, we propose a unified training framework for the generation of glottal signals in deep learning (DL)-based parametric speech synthesis systems. The glottal vocoding-based speech synthesis system, especially the modeling-by-generation (MbG) structure that we proposed recently, significantly improves the naturalness of synthesized speech by faithfully representing the noise component of the glottal excitation with an additional DL structure. Because the MbG method introduces a multistage processing pipeline, however, its training process is complicated and inefficient. To alleviate this problem, we propose a unified training approach that directly generates speech parameters by merging all the required models, such as acoustic, glottal, and noise models, into a single unified network. Considering the fact that noise analysis should be performed after training the glottal model, we also propose a stochastic noise analysis method that enables noise modeling to be included in the unified training process by iteratively analyzing the noise component in every epoch. Both objective and subjective test results verify the superiority of the proposed algorithm compared to conventional methods.

Index Terms: Text-to-speech, speech synthesis, glottal vocoder, modeling-by-generation structure

1. Introduction

With recent developments in deep learning (DL) techniques, glottal vocoder-based speech synthesis systems have significantly improved the quality of synthesized speech [1–3]. In a glottal vocoder, a pitch-dependent excitation signal is first obtained by applying a linear prediction (LP) inverse filter to an input speech signal [4,5], and then the temporal sequence of the excitation signal is trained and generated via DL techniques. The synthetic speech quality of a glottal excitation model is better than that of conventional band-aperiodicity (BAP)-based approaches [6]; however, its synthesized speech is often unnaturally buzzy because of overly smoothed glottal outputs.

To address the aforementioned problem, we proposed the modeling-by-generation (MbG)-structured glottal vocoder that directly models the missing high-frequency component in the generated glottal signal [7]. Using the fact that the difference between the reference and generated glottal signals is regarded as a non-harmonic or noise component, the weighted difference values are used as output noise features (NFs) for an additional ***noise model (NM)***. The glottal excitation in the synthesis stage is constructed by adding the generated outputs of the ***glottal model (GM)*** and the NM. As a result, the perceptual quality of the synthesized speech became much more natural than conventional approaches. However, its training process is highly complicated since the MbG-structured glottal vocoder approach uses a multistage architecture that needs to train three independent models, such as ***acoustic model (AM)***, the GM, and the NM. Moreover, its training process is redundant because similar input features are repeatedly used in each training network.

To alleviate these problems, this paper proposes a unified framework called a ***unified model (UM)*** for the MbG-structured glottal vocoding speech synthesis system. The inputs of the UM are the linguistic features (LFs), and the outputs are a concatenation of acoustic features (AFs), glottal features (GFs), and the NFs. The weights of the UM are optimized to minimize the error between the reference and generated outputs.

Because NF modeling requires an already-trained GM, it cannot be intuitively included in the UM training framework. To include NF modeling in the UM training process, we also propose a stochastic noise analysis method so the GFs and corresponding NFs are concurrently trained and generated in a single UM. At the beginning of the UM training process, the input NFs are filled with a random vector and the network weights are optimized once. After this optimization process, the new NFs are extracted from the "roughly" generated GFs, and used to update the entry of NFs. By repeating this update and optimization process, the GFs and corresponding NFs can be effectively trained in a single unified training network.

As all the output features are generated in a single UM, the proposed method builds a simple but effective glottal vocoding-based speech synthesis system. The objective and subjective test results also confirm that the proposed unified framework provides a much faster synthesis speed with highly qualified synthesized speech than the conventional MbG-structured approach.

2. MbG-structured glottal vocoding system

Figure 1 describes the block diagram of the conventional MbG-structured glottal vocoding speech synthesis system. It consists of the AM, GM, and NM, which is used to generate AFs, GFs, and NFs, respectively.

The output AFs consist of vocal tract line spectral frequencies (LSF–VT), a voicing flag (VUV), a logarithm energy (Erg), vocal source LSFs (LSF–VS), and a logarithm fundamental frequency (logF0). To extract the LSF-VT, the glottal inverse filtering (GIF) method is applied to the input speech first [4, 5], and then a glottal closure instant (GCI) detection algorithm is used to estimate the GCI, logF0, and VUV [8]. To extract the GFs, the two-pitch-period glottal signals that have GCIs at the middle and both ends are shaped by a cosine window, and they are normalized to have unity energy. Before training the GM,

*Work performed as an intern in Clova Voice, NAVER Corp., Seongnam, Korea.

10.21437/Interspeech.2018-1590

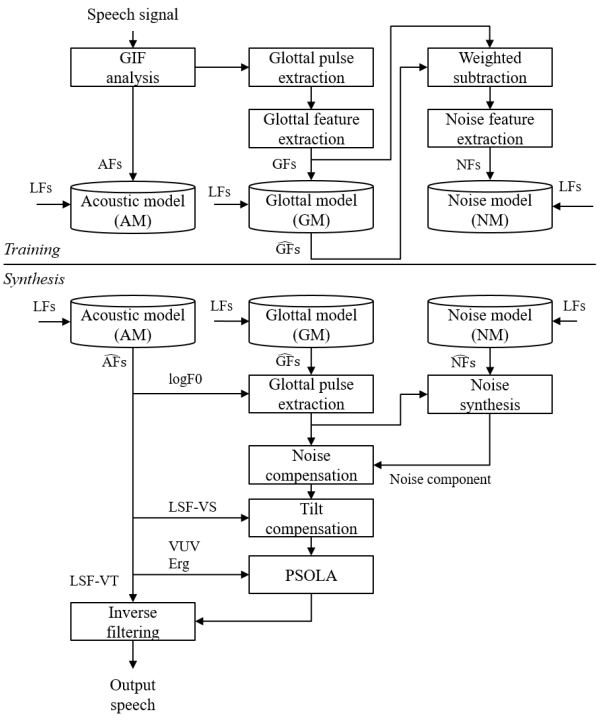

Figure 1: *Block diagram of the conventional MbG-structured glottal vocoding speech synthesis system. Each model constructs the mapping function from linguistic features (LFs) to acoustic features (AFs), glottal features (GFs), and noise features (NFs), respectively. The hat symbol implies the features estimated via DL model.*

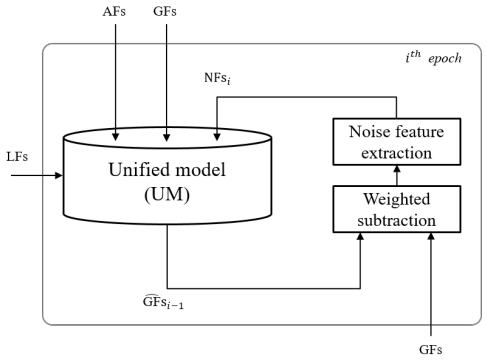

Figure 2: *Training of unified model with a stochastic noise analysis method. The reference NFs in current epoch (NFs_i) are obtained by generated GFs in previous epoch ($\hat{G}Fs_{i-1}$).*

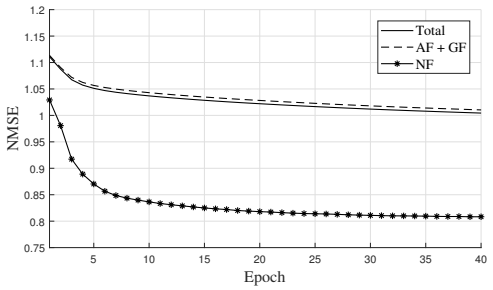

Figure 3: *NMSE curve of the vocoder features*

both ends of the GF are zero-padded to have a fixed dimension. In the NM, the output vector consists of NF vectors parameterized from the missing noise components in the GM outputs. To extract NFs, the noise component is first obtained through a weighted subtraction of the reference glottal pulse extracted from the recorded speech and the smoothed glottal pulse generated from the trained GM [7]. The shape and energy ratio of the noise component is then represented via the LSF (LSF–N) and a harmonic-to-noise ratio (HNR), respectively.

In the synthesis stage, the GM and NM predict their output features to reconstruct the glottal excitation signals. To compensate for the missing noise component, a sequence of random noise is first generated; then, its spectral shape and gain are refined by the generated LSF–N and HNR, respectively. By adding them to the generated glottal pulse and adjusting the spectral tilt, the two-pitch-period glottal pulses are obtained. Finally, the glottal excitation signal is reconstructed by applying a pitch-synchronous overlap-add (PSOLA) method; a single frame of speech signal is synthesized by filtering the glottal excitation signal through the vocal tract filter reconstructed by the generated LSF–VT coefficients.

3. Stochastic noise analysis method based unified model training

Employing the MbG structure in the glottal vocoding system provides significantly better quality than the conventional noise compensation algorithms. However, its training and generating processes are complicated due to the multistage processing pipeline. To construct a simpler and more efficient network, we propose a stochastic noise analysis-based UM training method that is able to capture all the vocoding parameters compactly within a unified network.

Figure 2 describes the training process of the proposed algorithm. In this framework, the LFs are used as inputs, and the concatenations of the AFs, GFs, and NFs are used as the corresponding outputs. Similar to conventional training methods, the pairs of input and output features are used to train the weights of the neural network, but there is a difference in controlling the entry of NFs into the output layer.

Because it is impossible to obtain reference NFs directly at the beginning of the training process, the random sequence fills the entry for NFs, and the weights are optimized once. After this optimization process, the new NFs are extracted from roughly generated GFs, and the NFs' entry is updated by the new ones. The UM in this training step has a better capability of describing GFs than that of UM in the previous training step; thus, the newly extracted NFs more clearly describe the smoothing impact on GF modeling. Consequently, by iteratively updating the entry of NFs with newly extracted ones in every training step, the UM naturally improves the modeling capabilities of GFs and corresponding NFs in a single unified training network.

Figure 3 represents the normalized mean square errors (NMSEs) in the total error, as well as the AFs, GFs, and NFs that are calculated during the UM training process. The smoothly converged AF, GF, and NF curves verify that the proposed stochastic noise analysis successfully converges to the optimal point without any unstable conditions.

Table 1. *Network architectures. All of the networks have FF networks at the input side and an LSTM network at the output side. The merged cell implies the model unification. The subscripts 'a', 'g', 'n', and '/' imply the AM, GM, NM, and their separabilities, respectively. For instance, the system MbG$_{a/gn}$ consists of a separated AM and a unified GM & NM.*

system	Type of layers	Layer architectures (units × layers)			Model size (M)
		AM	GM	NM	
MF$_{a/g}$	FF LSTM	1,024×2 512×1	512×2 256×1	–	5.48
MF$_{ag}$	FF LSTM	1,024×3 512×1		–	5.47
MbG$_{a/g/n}$	FF LSTM	1,024×2 512×1	512×2 256×1	256×2 128×1	5.77
MbG$_{a/gn}$	FF LSTM	1,024×2 512×1	512×3 256×1		5.72
MbG$_{agn}$	FF LSTM	1,024×3 512×1			5.47

Table 2. *Speech features and their dimensions including Δ and $\Delta\Delta$ values for acoustic, glottal and noise features.*

Feature	Component	dim.	Δ dim.
AFs	LSF–VT	30	90
	VUV	1	1
	Erg	1	3
	LSF–VS	10	30
	logF0	1	3
GFs	Glottal pulse	400	400
NFs	LSF–N	15	15
	HNR	1	1

4. Experiments

4.1. Experimental Setup

For all the experiments in this paper, we used a phonetically and prosodically balanced speech corpus recorded by a Korean male professional speaker. 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 included in neither the training nor the validation steps were used for testing.

In the analysis step, the vocoding parameters were extracted every 5-ms with a 20-ms frame length. Table 2 describes all the vocoder features used in this experiment. At the beginning of AF extraction, the logF0, VUV, and GCI were estimated using the SEDREAM algorithm [8]. Then, the quasi-closed phase GIF method was applied to estimate the 30-dimensional LSF–VT and the pitch-dependent glottal excitation signal [4, 5]. Additionally, a 10-dimensional LSF–VS was obtained by applying an LP analysis of the glottal pulse. The 400-dimensional time sequence of the glottal signal was used as the GF; meanwhile, the 15-dimensional noise LSFs and a 1-dimensional pulse-wise HNR were extracted for the NFs.

In the training step, the input LF vectors included 210-dimensional contextual information were used. The LF vectors consist of 203-dimensional binary features (e.g. identity of quinphone), 6 numerical features (e.g. the number and position of phonemes, syllables, and words), and one additional numerical feature for duration of current segment. The corresponding output AF feature vectors contained 127-dimensional acoustic parameters, including their time dynamics [9], whereas the GF and NF vectors contained 400- and 16-dimensional static parameters, respectively. Before training, both the input and output features were normalized to have zero-mean and unit-variance. The hidden layers consisted of multiple feedforward (FF) and long short-term memory (LSTM) layers, which were connected to the input layer and the output layer, respectively. Table 1 summarizes the number of layers, the number of units, and the corresponding model size among the different architectures of neural networks.

The conventional glottal vocoding system with a median filter (MF)-based noise compensation algorithm was also included as a baseline system [10]. In this system, the noise component is defined by the residual signal of the MF output, then parameterized into 15-dimensional noise LSFs and 1-dimensional energy terms to compose the NF vectors. In addition to 127-dimensional AF vectors, total 143-dimensional output features were trained via the AM. The rectified linear unit (reLu) and linear activation functions were used on the hidden and output layers, respectively. The weights were first initialized using a *Xavier* initializer [11], and then trained using a *back-propagation through time* procedure with an *Adam* optimizer [12, 13]. The training and test procedures were implemented using the *TensorFlow* framework [14].

In the synthesis step, the mean vectors of all the output features were predicted by the trained models. With the pre-computed global variances of output features from all the training data [15], a speech parameter generation algorithm was applied to generate a smooth trajectory of the AFs [16]. 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 and spectral tilt compensation modules were applied to the glottal pulse. By constructing the glottal excitation signal pitch-synchronously, a speech signal was synthesized with the generated LSF–VT and glottal excitation signals. To enhance spectral clarity, LSF-sharpening and formant enhancement filters were also applied to the generated spectral parameters [17, 18].

4.2. Objective and subjective evaluation results

In the objective test, distortions in speech parameters obtained from the original speech and estimated from various DL models were evaluated. The metrics for measuring distortion were

Table 3. *Objective evaluation results for the various speech synthesis systems*

System (model size; M)	LSD–VT (dB)	F0 RMSE (Hz)	VUV error (%)	LSD–GP (dB)	LSD–N (dB)	Generation time (ms)	Relative speed
$MF_{a/g}$ (5.48)	3.82	13.89	6.04	6.43	3.10	68.28	1.51
MF_{ag} (5.47)	3.80	13.66	6.16	6.40	3.10	45.31	1.00
$MbG_{a/g/n}$ (5.77)	3.82	14.38	6.19	6.45	3.09	87.86	1.93
$MbG_{a/gn}$ (5.72)	3.81	14.74	6.22	6.38	3.11	67.71	1.49
MbG_{agn} (5.47)	3.81	13.81	6.24	6.40	3.10	45.44	1.00

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

Test index	$MF_{a/g}$	MF_{ag}	$MbG_{a/g/n}$	$MbG_{a/gn}$	MbG_{agn}	Neutral	p-value
Test 1	43.3	28.7	–	–	–	28.0	0.03
Test 2	**18.0**	–	–	–	**58.0**	**24.0**	$\mathbf{< 10^{-8}}$
Test 3	–	**18.7**	–	–	**60.0**	**21.3**	$\mathbf{< 10^{-8}}$
Test 4	–	–	36.7	30.7	–	32.7	0.37
Test 5	–	–	40.7	–	32.0	27.3	0.21
Test 6	–	–	–	28.0	38.7	33.3	0.11

the log-spectral distance of the spectral parameters from the vocal tract, glottal pulse, and noise component (LSD–VT, LSD–GP, and LSD–N, respectively) in dB, the root mean square error (RMSE) for F0 in Hz, and the error rate of voicing flag (VUV error) in %. To evaluate the synthesis efficiency, we also measured the generation time (s) for evaluating all the output parameters' synthesis speeds in the test sets.

The objective results are summarized in Table 3. The findings verify the advantages of the proposed unified framework (MF_{ag} and MbG_{agn}) as follows: (1) It achieved a performance equivalent to the separated training cases ($MF_{a/g}$ and $MbG_{a/g/n}$). This means that if the network is "well" optimized, then the accuracy of feature estimation does not critically depend on the types of output features. (2) The stochastic noise analysis methods ($MbG_{a/gn}$ and MbG_{agn}) were as effective as the conventional MbG-structured method ($MbG_{a/g/n}$), even though they did not have fixed-reference NFs during the training process. They show an LSD-N performance similar to the separated case, just allowing for a difference lower than 0.02 dB. (3) The unified framework significantly reduced generation time, despite having a similar number of parameters. Firstly, the unification of AM and GM (MF_{ag}) showed a synthesis speed 1.5 times faster than the separated one ($MF_{a/g}$). Secondly, the unification of GM and NM ($MbG_{a/gn}$) showed a synthesis speed 1.3 times faster than the separated one ($MbG_{a/g/n}$). Consequently, the UM merging the AM, GM, and NM (MbG_{agn}) showed a synthesis speed about two times faster than the separated training case ($MbG_{a/g/n}$).

To evaluate the perceptual quality of the proposed system, an A-B preference test and the mean opinion score (MOS) listening test were performed. In the preference test, 10 native Korean listeners were asked to rate the randomly selected 15 synthesized utterances from the test set by quality preference. The preference results shown in Table 4 verify that the perceptual quality of the unified framework is indistinguishable from that of the separated training cases (Test 1, 4, 5, and 6). Because the estimated noise component is quite different between the MF and MbG approaches, the listeners preferred the proposed MbG approach to the conventional MF approach (Test 2 and 3).

The setup for the MOS test was the same as that for the preference test, except listeners were asked to make quality judgments about the synthesized speech using the following possible responses: 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, and 5 = Excellent. In addition to the glottal vocoding system, the STRAIGHT-based speech synthesis system was also included as a baseline system [6]. For fair comparison, the 3 FF layers with 1,024 units and a single LSTM layer with 512 memory cells having a model size of 5.27 M was used as AM for the modeling of STRAIGHT features. Table 5 shows the MOS test results, which confirm that the unified framework achieved a performance similar to that of the separated models, and it provided a much better perceptual quality than the baseline STRAIGHT system.

Table 5. *Subjective MOS test results with a 95% confidence interval for various speech synthesis systems.*

STRAIGHT	$MbG_{a/g/n}$	$MbG_{a/gn}$	MbG_{agn}
2.91 ± 0.13	3.79 ± 0.21	3.71 ± 0.18	3.68 ± 0.17

5. Conclusion

In this paper, we introduced a unified training framework for a glottal vocoding system with a stochastic noise analysis method. By including the modeling of a smoothing impact on the glottal signal in every optimization step, the proposed system successfully simplified the training and generation processes. The experimental results verified that the proposed framework showed an equivalent modeling accuracy and perceptual quality to conventional systems; whereas the generation speed was two times faster. Consequently, the proposed framework successfully constructed a simple and compact speech synthesis system by removing the unnecessarily redundant multistage processing pipelines.

6. Acknowledgements

This research was supported by Search & Clova, NAVER Corp., Seongnam, Korea.

7. 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] 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.

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

[5] 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.

[6] 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.

[7] M.-J. Hwang, E. Song, and H.-G. Kang, "Modeling-by-generation-structured noise compensation algorithm for glottal vocoding speech synthesis system," in *Proc. ICASSP*, 2018.

[8] D. Thomas and D. Thierry, "Glottal closure and opening instant detection from speech signals," in *Proc. INTERSPEECH*, 2009, pp. 2891–2894.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.