Excitation-by-SampleRNN Model for Text-to-Speech

Kyungguen Byun[1], *Eunwoo Song*[2], *Jinseob Kim*[2], *Jae-Min Kim*[2] and *Hong-Goo Kang*[1]
[1]Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
[2]NAVER Corp. Seongnam, Korea
piemaker90@dsp.yonsei.ac.kr, {eunwoo.song, paul.jskim, kjm.kim}@navercorp.com, hgkang@yonsei.ac.kr

Abstract

In this paper, we propose a neural vocoder-based text-to-speech (TTS) system that effectively utilizes a source-filter modeling framework. Although neural vocoder algorithms such as SampleRNN and WaveNet are well-known to generate high-quality speech, its generation speed is too slow to be used for real-world applications. By decomposing a speech signal into spectral and excitation components based on a source-filter framework, we train those two components separately, i.e. training the spectrum or acoustic parameters with a long short-term memory model and the excitation component with a SampleRNN-based generative model. Unlike the conventional generative model that needs to represent the complicated probabilistic distribution of speech waveform, the proposed approach needs to generate only the glottal movement of human production mechanism. Therefore, it is possible to obtain high-quality speech signals using a small-size of the pitch interval-oriented SampleRNN network. The objective and subjective test results confirm the superiority of the proposed system over a glottal modeling-based parametric and original SampleRNN-based speech synthesis systems.

Keywords: Text-to-speech, SampleRNN

1. Introduction

In conventional deep learning-based statistical parametric based speech synthesis (SPSS) systems, the target parameters for training are pre-determined by traditional vocoding schemes that typically consist of spectral and excitation components [1]. Since it is not easy to accurately infer the dynamic variation of acoustic parameters from linguistic inputs even with a deep learning framework, the quality of the synthesized speech is unnatural. Although some of the researchers tried to address the problem by designing effective vocoding structures such as WORLD [2], time-frequency trajectory excitation [3], and glottal models [4], the quality of synthesized speech was not satisfactory.

Recently, neural vocoders that auto-regressively generate speech waveform in a sample-by-sample manner play a significant role in the text-to-speech (TTS) community thanks to their superior performance on modeling the inherent probabilistic distribution of the speech signal. Oord et. al. proposed a WaveNet framework that utilizes a dilated causal convolution structure to exponentially increase the receptive field size in each hidden layer [5]. Mehri et. al. proposed a SampleRNN framework that utilizes a multiple level of recurrent neural networks that are hierarchically connected to present wider range of past samples [6]. By providing acoustic parameters such as line spectral pairs (LSPs), fundamental frequency (F0), and context features extracted from text input as a conditional vector for the neural vocoder module, they could achieve the state-of-the-art performance in end-to-end based TTS systems [7, 8]. As the neural vocoder-based TTS systems effectively modeled various resolution of time-domain samples with multiple number of hidden layers, they could reduce unwanted artifacts caused by an overlap-and-add processing, phase inconsistency at the frame boundary, and so on.

To further improve the performance of neural vocoder-based TTS systems, Song et. al. and Hwang et. al. have introduced a human voice production system to the neural vocoding frameworks [9, 10]. They first decomposed speech signal into spectral and excitation components, and then applied the neural vocoder model only to the excitation component. Note that the spectral component was separately modeled by an additional deep neural network framework, and used to reconstruct speech signal in the synthesis step. Although they achieved higher performance than the original neural vocoder-based system, however, they could not enjoy the benefits of applying the neural vocoding only to the excitation component. In other words, the size of deep learning network was still the same as the one used for the original waveform generation model making the computation complexity of the system even higher.

In this paper, we verify that the model size of deep learning network, especially the neural vocoder used for modeling the excitation component can be significantly reduced while minimizing the loss of overall synthetic quality. Considering the fact that the excitation signal has smaller dynamics than the speech signal; thus it can be modeled by few parameters similar with low bit rate vocoding technique [11]. Consequently, we significantly reduce the size of hyper-parameters of the neural vocoding systems.

There is also an issue that which type of neural vocoder is more proper for excitation modeling. In terms of modeling performance, both SampleRNN and WaveNet algorithms show satisfactory results. However, in terms of controllability, SampleRNN is better thanks to its RNN structure. In the WaveNet case, a dilated convolution network captures the pattern of the past samples, but there is no clear one-to-one relationship between the condition vectors and the generated signals. Its receptive field constantly moves one sample at a time whenever it predicts new sample, which results in the boundaries between adjacent acoustic features being dimmed. In contrast, SampleRNN consists of hierarchical RNNs that can update their hidden states at different sample interval. Typically, the update interval of the highest tier is synchronized to the acoustic features' extraction interval.

Contents of this paper are as follows: Section 2 describes SampleRNN algorithm and its features. Section 3 and 4 describes the proposed method in detail and experimental results are provided respectively. In Section 5, conclusions are drawn.

2. Background

The SampleRNN algorithm is originally proposed to auto-regressively generate audio waveforms. To train the model parameters, they introduced a criterion to minimize the Kullback–Leibler (KL) divergence between the model's softmax output as shown in eq.1 and the output quantized index of speech signal.

$$p(X)=\sum_{i=0}^{T-1} p(x_{i+1} \mid x_1, \dots, x_i) \qquad (1)$$

In a Char2Wav algorithm [7], the SampleRNN is utilized as a neural vocoder to generate speech signals from the acoustic condition vectors. They are delivered to the top tier of the SampleRNN network, and then merged with the past sample sequences by using a single layer of linear projection.

As depicted in Figure 1, the architecture of SampleRNN is composed with the multiple level of tiers: The frame-level tiers cover a wide range of sample relations and the sample-level tier models the sample-by-sample distribution. Various types of RNNs such as long short-term memory (LSTM) or Gated Recurrent Unit (GRU) are used for composing frame-level tiers, and multi-layer perceptron (MLP) are used for composing the sample-level tier.

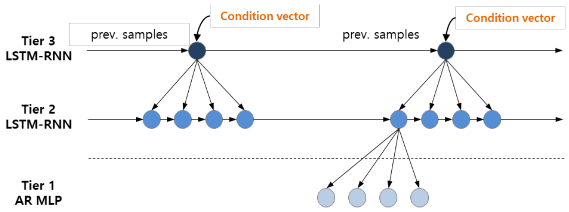

Figure 1: A framework of the conditional SampleRNN model

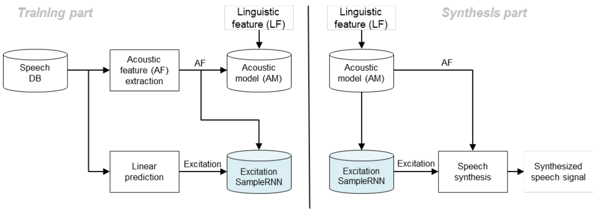

Figure 2: Training and synthesis process of the proposed Excitation-by-SampleRNN model

Airaksinen, et. al. proposed a method that directly estimates the time-domain glottal features in the GlottDNN framework [4]. The glottal features were sophisticatedly extracted by a glottal inverse filtering method that removes a lip radiation effect and residual spectral information from the excitation signal in a pitch-synchronous manner [12, 13]. The glottal feature is then normalized by its power and length since it must have a fixed dimension to be applied to a neural network-based regression model (*i.e.* glottal model). However, since the glottal model employs a mean-square error-based training criterion, generated outputs became inevitably over-smoothed. Thus, the noisy component of the generated glottal feature should be compensated by the post-filters [4, 14].

3. Proposed Method

As described in Figure 2, our proposed method consists of two sequentially connected models such as an acoustic model and an SampleRNN-based excitation model. Before training both models, the pair of acoustic features and excitation signals should be prepared. The acoustic features contain the spectral parameters represented by the LSPs and the excitation parameters such as a log-F0, a v/uv flag, and a log-energy. For the preparation of the corresponding excitation signal, frame-wise excitation signals are obtained by inverse filtering the speech signals via linear prediction (LP) coefficients. Finally, the excitation signal is concatenated by an overlap-add process.

In the training phase, the bi-directional LSTM-based acoustic model learns a complex mapping function between the pre-processed context features and the prepared acoustic features. In contrast, the Excitation-by-SampleRNN model learns the distribution of the target excitation signal with the conditional acoustic features.

In the synthesis phase, the acoustic features are first predicted and a parameter generation algorithm is then applied to the generated features to relieve the over-smoothing effect. By inputting those parameters, a time-domain excitation signal is generated by the SampleRNN based excitation model. Finally, a speech signal is reconstructed by the convolution of the excitation signal with the generated LP coefficients.

4. Experiments and Results

The experiments used a phonetically and prosaically balanced speech corpus recorded by a Korean male professional speaker. The speech signals were sampled at 24 kHz, and each sample was quantized by 16 bits. In total, ten hours of speech database were used for training, one hour's database were used for validation, and another one hour's database not included in either the training or validation steps were used for evaluation.

To compose the acoustic feature vector, we extracted the 40-dimensional LSPs, the log-F0, the v/uv flag, and the log-energy by using the STRAIGHT vocoder [15]. 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 derivatives. The corresponding input feature vectors included 356-dimensional context features extracted from the input text. The hidden layers consisted of two feedforward layers with 1,024 units and two BLSTM layers with 512 memory blocks.

In the Excitation-by-SampleRNN (EbS) training step, all the acoustic parameters were used to compose the input condition feature vectors. The corresponding excitation signal was normalized in a range from -1.0 to 1.0 and quantized by 8-bit mu-law compression. The model architecture consisted of three-tier SampleRNN of which update interval were set to 120, 10, and 1, respectively. Because the update interval of the highest tier was the same with the frame rate (5 ms), the generated acoustic features could be directly used to compose the condition vector of the Excitation-by-SampleRNN model. The number of the GRU nodes was set to 256 and all the weights were optimized via a truncated backpropagation through time technique. The length of the truncated sequences was set to 6,000 samples. The batch size was sixteen and 300,000 iterations were done to sufficiently converge the model.

To evaluate the objective performance of the proposed system, the results were compared to that of conventional system based on the SampleRNN vocoder. In the test, distortions between the original speech and the synthesized speech were measured by log-spectral distance (LSD; dB), F0 root mean square error (RMSE; Hz), and v/uv error rate (ER; %). Table 1 shows the LSD and F0 RMSE test results, with respect to the different neural vocoders. Comparing with the baseline SampleRNN system, the proposed Excitation-by-SampleRNN model achieved better estimation performance in terms the LSD and the F0 RMSE.

Table 1: Objective test results with respect to the different neural vocoders: The conventional SampleRNN and the proposed Excitation-by-SampleRNN (EbS). The systems that returned the smallest errors are in bold font.

System	LSD	F0 RMSE	v/uv ER
SampleRNN	4.67	15.24	**5.13**
EbS	**4.16**	**13.42**	5.19

Table 2: Subjective preference test results (%) for different types of vocoders.

Test	GlottDNN	EbS	Neutral
Preference	4.2	73.7	22.1

Table 3: Subjective MOS test results for different types of vocoders.

Test	Record	SampleRNN	EbS
MOS	4.86±0.05	3.53±0.19	3.80±0.16

Table 4: Generation (Gen.) speed depending on the model size for single (s) and multiple (m) sentence generation.

Test	EbS	Light EbS
Model size	254 M	17 M
Gen. speed (s)	0.86 k	2.53 k
Gen. speed (m)	40.29 k	244.29 k

To evaluate the perceptual quality of the proposed system, A-B preference test was performed. In the preference tests, ten native Korean listeners were asked to rate the quality preference of the synthesized speech. In total, fifteen utterances were randomly selected from the test set and were then synthesized using the two types of vocoders: The baseline parametric GlottDNN and the proposed Excitation-by-SampleRNN. The A-B test results is shown in Table 2, which confirms that a generative model-based excitation modeling method provided much better perceptual quality than the traditional parametric approach.

To further evaluate the perceptual quality of the proposed system, perceptual mean opinion score (MOS) listening tests were performed. The 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. The test results shown Table 3 confirm that the proposed Excitation-by-SampleRNN performed significantly better than the conventional SampleRNN, which implies that decoupling the formant component of the speech signal via the LP inverse filter significantly improves the modeling accuracy.

To verify the effectiveness of the proposed system, we conducted additional experiments by reducing the model size. Note that the excitation signal has much smaller dynamics than the speech signal; thus it can be modeled by few parameters. To compose a light Excitation-by-SampleRNN, we changed the number of bits for the mu-law quantization from eight to six (*i.e.*

64-level representation of the excitation signal) and also reduced the number of nodes in GRU in frame-level tiers.

Table 4 represents the number of model parameters and its corresponding generation speed defined as the number of generated samples per second. Note that the generation was performed by using a GeForce 1080 Ti GPU. The test results confirm that the light version of the Excitation-by-SampleRNN vocoder significantly reduced the number of model parameters, which enabled to reconstruct speech signal more than three times faster. For the multiple sentence generation case, 300 sentences are generated. The number of generated samples per second increased since sentences were generated in parallel. The ratio of the generation speed between two systems was even higher (6:1), which exceeds the ratio of the single sentence case. Although this process resulted in degrading the perceptual quality of the synthesized speech (the MOS was decreased to 3.20), more effective quantization methods can alleviate this problem, which will be discussed in our future research.

5. Conclusion

In this paper, we proposed an effective Excitation-by-SampleRNN vocoder that exploited a generative model to predict the excitation component of speech signal. Because the prediction quality of the excitation signal deeply affected to the synthetic quality and bypassing the quantization error problem in the direct waveform prediction structure, the proposed model achieved significant better performance than the conventional GlottDNN and SampleRNN models. Also, we investigated the impact of overall quality by manipulating the number of model parameters, the effects of parameter reduction, and the number of quantization bit settings.

6. Acknowledgements

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

References

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

[2] M. Morise, F. Yokomori, and K. Ozawa, ``WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877-1884, 2016.

[3] 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*, pp. 4949–4953, 2015.

[4] M. Airaksinen, B. Bollepalli, L. Juvela, and Z. Wu, “GlottDNN-a full-band glottal vocoder for statistical parametric speech synthesis.” in *Proc. INTERSPEECH*, pp. 2473-2477, 2016.

[5] 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:1609.03499*, 2016.

[6] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, “SampleRNN: An unconditional end-to-end neural audio generation model,” in Proc. *ICLR*, 2017.

[7] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, “Char2Wav: End-to-End Speech Synthesis,” in *Proc. ICLR,* 2017.

[8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomygiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” *Proc. INTERSPEECH,* pp. 4006-4010, 2017.

[9] E. Song, K. Byun, and H. Kang, “ExcitNet Vocoder: A neural excitation model for parametric speech synthesis systems,” *arXiv*: 1811.03311, 2018

[10] M. Hwang, F. Soong. F. Xie, X. Wang, and H. Kang, “LP-WAVENET: Linear Prediction-based WaveNet Speech Synthesis,” *arXiv*:1811.11913, 2018.

[11] A. Mcree, and T. Barnwell, “A Mixed Excitation LPC vocoder Model for Low Bit Rate Speech Coding,” *IEEE Trans. Speech and Audio Process.*, pp.242-250, 1995.

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

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

[14] 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.

[15] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” in Proc. *MAVEBA*, pp. 13-15, 2001.