Generative approach using the noise generation models for DNN-based speech synthesis trained from noisy speech

Masakazu Une*†, Yuki Saito†, Shinnosuke Takamichi†, Daichi Kitamura‡†,
Ryoichi Miyazaki* and Hiroshi Saruwatari†
* National Institute of Technology, Tokuyama College, Gakuendai Shunan, Yamaguchi 745–8585, Japan
E-mail: {i12une, miyazaki}@tokuyama.ac.jp
† Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo Bunkyo-ku, Tokyo 133–8656, Japan
E-mail: {yuuki_saito, shinnosuke_takamichi, hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp
‡ National Institute of Technology, Kagawa College, 355 Chokushi-cho Takamatsu, Kagawa 761–8058, Japan
E-mail: kitamura-d@t.kagawa-nct.ac.jp

***Abstract*—This paper proposes a generative approach to construct high-quality speech synthesis from noisy speech. Studio-quality recorded speech is required to construct high-quality speech synthesis, but most of existing speech has been recorded in a noisy environment. A common method to use noisy speech for training speech synthesis models is reducing the noise before the vocoder-based parameterization. However, such multi-step processes cause an accumulation of spectral distortion. Meanwhile, statistical parametric speech synthesis (SPSS) without vocoders, which directly generates spectral parameters or waveforms, has been proposed recently. The vocoder-free SPSS will enable us to train speech synthesis models considering the noise addition process generally used in signal processing research. In the proposed approach, newly introduced noise generation models trained by a generative adversarial training algorithm randomly generates spectra of the noise. The speech synthesis models are trained to make the sum of their output and the randomly generated noise close to spectra of noisy speech. Because the noise generation model parameters fit the spectrum of the observed noise, the proposed method can alleviate the spectral distortion found in the conventional method. Experimental results demonstrate that the proposed method outperforms the conventional method in terms of synthetic speech quality.**

I. Introduction

Statistical parametric speech synthesis [1] is a technique that synthesizes natural-sounding speech from the text using statistical models. A variety of techniques has been proposed to improve speech naturalness [2], [3], [4]. Deep neural network (DNN)-based speech synthesis [5] has significantly improved speech quality. In order to construct high-quality speech synthesis, studio-quality speech data is required. However, most of the existing speech data has been recorded in a noisy environment [6]. Besides training using such poor-quality speech is a challenging task in DNN-based speech synthesis. There are several kinds of *noisy environments*, such as narrow frequency bands [7] and poor-quality of communications network [8]. In our research, we focused on speech recorded in a more stationary noise environment such as in a house [9].

The common approach is to perform noise reduction prior to speech synthesis training (as shown in the upper graph Fig. 1. Noise reduction for speech synthesis is different from that for speech recognition because it needs the vocoder parameters (e.g., mel-cepstral coefficients of STRAIGHT [10] and WORLD [11]) to train the acoustic models (speech synthesis models)[1]. Noise reduction for speech synthesis can be classified into two types: 1) direct estimation of vocoder parameters from noisy speech [12] and 2) vocoder-based parameterization after noise reduction. The former trains the statistical models in advance to identify the vocoder parameters from the noisy speech parameters. This method can perform a nonlinear conversion using the DNNs but its resistance to the unseen[2] noise is not guaranteed. The latter type can adopt a signal-processing-based unsupervised noise reduction such as spectral subtraction [13]. It works even when the unseen noise is observed, but the quality of vocoder-based parameterization after the noise-suppressed speech is not guaranteed. Also, the common problem in both types is that the spectral distortion caused by the DNN-based or signal-processing-based advance estimation is accumulated in the speech synthesis training. In our research, we focused on constructing vocoder-free DNN-based speech synthesis from noisy speech [5], [14]. It is a framework that directly estimates the spectral parameters or waveforms, not the vocoder parameters. Avoiding the vocoders makes it possible to apply spectrum-domain or waveform-domain computation used in signal processing research. The noise reduction for speech synthesis described above can be applied in the assessment of the spectral parameters or

[1] Usually, a statistical model that predicts speech from a text is called *acoustic model*, but in this paper we call it *speech synthesis model* to differentiate it from the noise generation model introduced below.

[2] Here, *unseen* means noise signals not contained in the training data.

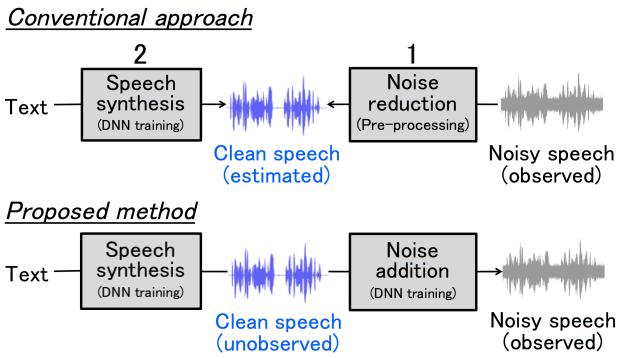

Fig. 1. Speech synthesis training procedures using noisy speech. Conventional method (upper graph) first performs noise reduction, then performs speech synthesis training to predict the noise-suppressed speech parameters. Our method (lower graph) directly predicts parameters of noisy speech using the noise addition process of human speech production.

waveforms. Regarding robustness to unseen noise, the latter method (noise reduction and speech synthesis training) is preferred. However, the accumulation of the spectral distortion explained above remains a critical problem.

In this paper, we propose a generative approach to training high-quality statistical parametric speech synthesis from noisy speech using speech synthesis and noise generation models. The noise generation models are trained by generative adversarial training [15] so that the generated spectral parameters have the natural statistics of the observed noise. In synthesis, the models randomly generate the spectral parameters of the noise. The speech synthesis models are trained so that the sum of the output and the randomly generated noise is close to the spectral parameters of the noisy speech. Namely, the speech synthesis models are trained using the noise addition process as shown in the lower graph in Fig. 1. The proposed method can more efficiently model the statistics of the observed noise compared to the conventional spectral subtraction-based noise reduction. Also, the proposed generative approach can alleviate the spectral distortion observed in the conventional method by using noise reduction. We conducted the evaluation with several settings of the signal-to-noise ratio and noise suppression ratio. The experimental evaluation demonstrated that the proposed method outperformed the conventional method in terms of synthetic speech quality.

II. Conventional Method: Spectral Subtraction and Speech Synthesis Model Training

Using the observed noisy speech, we performed spectral subtraction to suppress the observed noise first. Then we performed speech synthesis model training based on the mean squared error criterion using the suppressed speech.

A. Spectral subtraction

Spectral subtraction [13] approximates the distribution of the power spectrum of the noise signals and subtracts the noise components from the power spectrum of noisy speech. Let the log amplitude sequence of the noise signal be $\boldsymbol{y}_{\mathrm{n}} = [\boldsymbol{y}_{\mathrm{n},1}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{n},t}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{n},T_{\mathrm{n}}}^{\top}]^{\top}$, and that of the noisy speech be $\boldsymbol{y}_{\mathrm{ns}} = [\boldsymbol{y}_{\mathrm{ns},1}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{ns},t}^{\top}, \cdots, \boldsymbol{y}_{\mathrm{ns},T}^{\top}]^{\top}$, where T_{n} and T are the total frame lengths of the noise and noisy speech, respectively. $\boldsymbol{y}_{\mathrm{n},t} = [y_{\mathrm{n},t}(1), \cdots, y_{\mathrm{n},t}(f), \cdots, y_{\mathrm{n},t}(F)]^{\top}$ and $\boldsymbol{y}_{\mathrm{ns},t} = [y_{\mathrm{ns},t}(1), \cdots, y_{\mathrm{ns},t}(f), \cdots, y_{\mathrm{ns},t}(F)]^{\top}$ denote the log amplitude of the noise and noisy speech at frame t, respectively. f is the frequency bin and F is the total number of the frequency bins. Here, $\boldsymbol{y}_{\mathrm{n}}$ indicates the non-speech period of $\boldsymbol{y}_{\mathrm{ns}}$. The log amplitude suppressed by spectral subtraction, $\boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}$, is given as

$$\exp\{y_{\mathrm{ns},t}^{(\mathrm{SS})}(f)\} = \begin{cases} \sqrt{\exp\{y_{\mathrm{ns},t}(f)\}^2 - \beta\bar{y}_{\mathrm{n},t}(f)} \\ \quad \text{if} \quad \exp\{y_{\mathrm{ns},t}(f)\}^2 > \beta\bar{y}_{\mathrm{n},t}(f) \\ 0 \quad \text{otherwise} \end{cases}, \tag{1}$$

$$\bar{y}_{\mathrm{n},t}(f) = \frac{1}{T_{\mathrm{n}}} \sum_{t=1}^{T_{\mathrm{n}}} \exp\{y_{\mathrm{n},t}(f)\}^2, \tag{2}$$

where, β is the noise suppression ratio that adjusts the amount of noise suppression.

B. Speech synthesis model training

Let the speech synthesis model predicting the log amplitude of speech from the input contexts be $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ described as the neural networks [5], [14]. Here, let the input context sequence be $\boldsymbol{x} = [\boldsymbol{x}_1^{\top}, \cdots, \boldsymbol{x}_t^{\top}, \cdots, \boldsymbol{x}_T^{\top}]^{\top}$, the model parameter of $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ is estimated by minimizing mean squared error between the generated log amplitude $\hat{\boldsymbol{y}}_{\mathrm{s}} = \boldsymbol{G}_{\mathrm{s}}(\boldsymbol{x})$ and $\boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}$, which is given as

$$L_{\mathrm{MSE}}\left(\hat{\boldsymbol{y}}_{\mathrm{s}}, \boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}\right) = \frac{1}{T}\left(\hat{\boldsymbol{y}}_{\mathrm{s}} - \boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}\right)^{\top}\left(\hat{\boldsymbol{y}}_{\mathrm{s}} - \boldsymbol{y}_{\mathrm{ns}}^{(\mathrm{SS})}\right). \tag{3}$$

C. Issues

Spectral subtraction causes distortion of the suppressed log amplitude because the statistic of the noise is approximated with the expectation value. Also, it causes the so-called musical noise [16] which is known as a harsh artificial noise. Moreover, this kind of spectral distortion is accumulated during speech synthesis model training, so the quality of the final synthesized speech is significantly lower.

III. Proposed Method: Generative Approach to Train Speech Synthesis and Noise Generation Models

The DNN architecture of the proposed method is shown in Fig. 2. In addition to the speech synthesis model $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ of the conventional method, a noise generation model $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ is introduced here. $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ transforms the prior known distribution to the observed noise distribution and randomly generates the noise spectrum. The speech synthesis model $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ is trained so that the sum of its output and noise spectrum generated by $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ is close to the observed spectrum of the noisy speech. In the preliminary experiment, we tried to simultaneously train the $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ and $\boldsymbol{G}_{\mathrm{n}}(\cdot)$, but the separation performance of speech and noise is limited. Therefore, we first trained the noise generation model $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ to have the natural statistics

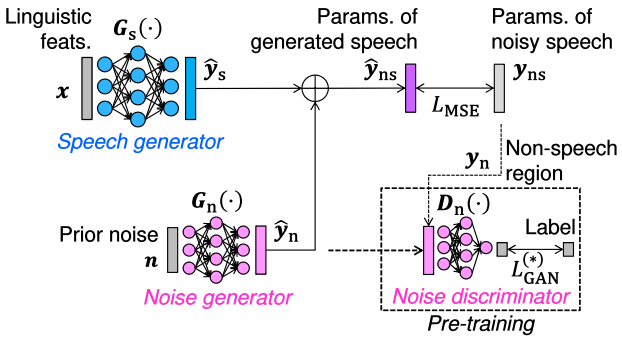

Fig. 2. Architectures of the proposed method. Noise generation model $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ is trained using noise discrimination model $\boldsymbol{D}_{\mathrm{n}}(\cdot)$ and it samples the noise randomly. Speech synthesis model $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ is trained so that the sum of its output and the sampled noise is close to spectra of the noisy speech.

of the noise spectrum using the observed noise spectrum $\boldsymbol{y}_{\mathrm{n}}$. Then, while fixing the model parameter of $\boldsymbol{G}_{\mathrm{n}}(\cdot)$, the $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ was trained using the spectrum of the noisy speech.

A. Generative adversarial training of the noise generation model

Generative adversarial training algorithm [15] is used for efficiently modeling the statistics of the observed noise. An input of $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ is a T_{n}-frame prior noise sequence $\boldsymbol{n} = [\boldsymbol{n}_1^\top, \cdots, \boldsymbol{n}_t^\top, \cdots, \boldsymbol{n}_{T_{\mathrm{n}}}^\top]^\top$ randomly sampled frame by frame from a known probability distribution (e.g., uniform distribution). The $\boldsymbol{n}_t$ indicates the prior noise vector at frame t. The $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ is updated with the noise discrimination models $\boldsymbol{D}_{\mathrm{n}}(\cdot)$ which distinguishes the generated noise $\hat{\boldsymbol{y}}_{\mathrm{n}} = \boldsymbol{G}_{\mathrm{n}}(\boldsymbol{n})$ from the observed noise $\boldsymbol{y}_{\mathrm{n}}$. The loss functions $L_{\mathrm{GAN}}^{(G)}(\cdot)$ for $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ and $L_{\mathrm{GAN}}^{(D)}(\cdot)$ for $\boldsymbol{D}_{\mathrm{n}}(\cdot)$ are

$$L_{\mathrm{GAN}}^{(G)}(\hat{\boldsymbol{y}}_{\mathrm{n}}) = -\frac{1}{T_{\mathrm{n}}}\sum_{t=1}^{T_{\mathrm{n}}} \log D_{\mathrm{n}}(\hat{\boldsymbol{y}}_{\mathrm{n},t}), \tag{4}$$

$$L_{\mathrm{GAN}}^{(D)}(\boldsymbol{y}_{\mathrm{n}}, \hat{\boldsymbol{y}}_{\mathrm{n}}) = -\frac{1}{T_{\mathrm{n}}}\sum_{t=1}^{T_{\mathrm{n}}} \log D_{\mathrm{n}}(\boldsymbol{y}_{\mathrm{n},t}) - \frac{1}{T_{\mathrm{n}}}\sum_{t=1}^{T_{\mathrm{n}}} \log\left(1 - D_{\mathrm{n}}(\hat{\boldsymbol{y}}_{\mathrm{n},t})\right), \tag{5}$$

respectively. The adversarial training minimizes the approximated Jensen-Shannon divergence between $\boldsymbol{y}_{\mathrm{n}}$ and $\hat{\boldsymbol{y}}_{\mathrm{n}}$ distributions. After the training, $\boldsymbol{G}_{\mathrm{n}}(\cdot)$ can randomly sample the noise spectrum that has the statistics of the observed noise.

B. Speech synthesis model training using noise generation model

We assumed that the additivity of speech and noise holds in the amplitude domain and we ignored the phase information. While fixing the estimated parameters of $\boldsymbol{G}_{\mathrm{n}}(\cdot)$, the speech synthesis model $\boldsymbol{G}_{\mathrm{s}}(\cdot)$ was trained to minimize the following loss function:

$$L_{\mathrm{MSE}}(\hat{\boldsymbol{y}}_{\mathrm{ns}}, \boldsymbol{y}_{\mathrm{ns}}) = \frac{1}{T}(\hat{\boldsymbol{y}}_{\mathrm{ns}} - \boldsymbol{y}_{\mathrm{ns}})^\top(\hat{\boldsymbol{y}}_{\mathrm{ns}} - \boldsymbol{y}_{\mathrm{ns}}), \tag{6}$$

$$\hat{\boldsymbol{y}}_{\mathrm{ns}} = \log(\exp\hat{\boldsymbol{y}}_{\mathrm{s}} + \exp\hat{\boldsymbol{y}}_{\mathrm{n}}), \tag{7}$$

Note that the sequence length of $\hat{\boldsymbol{y}}_{\mathrm{n}}$ here is T. In synthesis, the log amplitude of the synthesized speech is given as $\hat{\boldsymbol{y}}_{\mathrm{s}} = \boldsymbol{G}_{\mathrm{s}}(\boldsymbol{x})$. The final synthesized speech is obtained by applying Griffin-Lim's phase reconstruction algorithm [17]. In this paper, the speech synthesis and the noise generation model predict the log amplitude, and the predicted outputs are summed up in the linear amplitude domain. A simpler implementation is one in which the models output the linear amplitude. However, our method did not work well in this implementation.

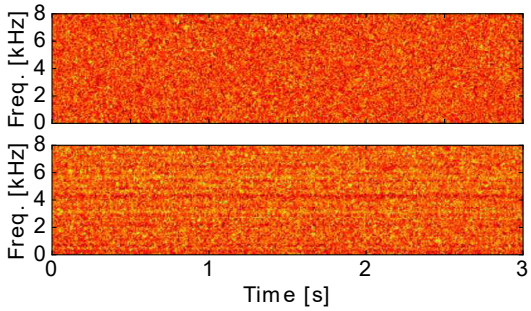

Fig. 3. Spectrograms of observed noise (upper graph) and generated noise (lower graph). The generated noise is randomly sampled frame by frame independently. We can see some temporal stripes in the generated noise, but its overall tendency is similar to the observed noise.

C. Discussion

The proposed method does not define the explicit probability distribution and expresses only the empirical distribution using generative adversarial network (GAN). Therefore, the proposed method reduces the spectral distortion that causes musical noise. The proposed method currently captures only stationary noise in the amplitude domain but it can be extended to recurrent architectures to capture non-stationary noise, conditional GAN [18] to capture context-dependent noise (e.g., pop noise), and WaveNet [19] for waveform-domain calculation [20]. Also, our GAN-based approach is expected to extend to GAN-based model adaptation [21]. Pre-recorded clean speech data will be used to build prior models.

IV. Experimental Evaluation

A. Experimental conditions

We used approximately 3000 Japanese utterances (clean speech) included in the JSUT corpus [22]. We artificially generated Gaussian noise as an observed noise and added it to the clean speech. The evaluation data was 53 sentences of subset J form ATR Japanese database [23]. The training data was sampled at 16 kHz. The window length was 400 samples (25 ms), the frame shift length was 80 samples (5 ms), and the size of the fast Fourier transform (FFT) length was 512. The window function was the Hamming window. The speech

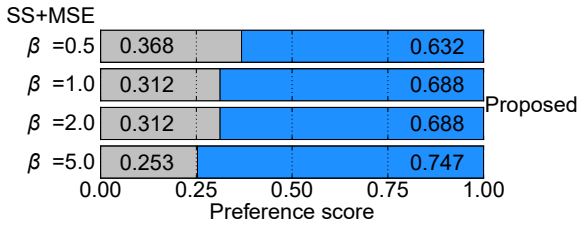

Fig. 4. Preference scores on synthetic speech quality (SNR = 0 dB).

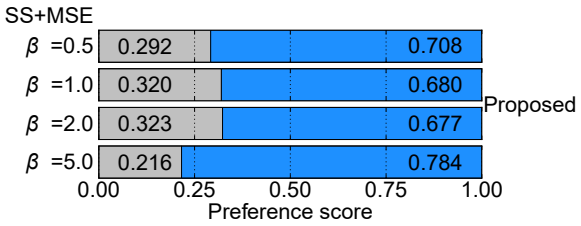

Fig. 5. Preference scores on synthetic speech quality (SNR = 5 dB).

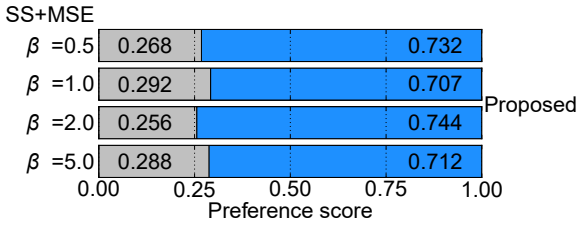

Fig. 6. Preference scores on synthetic speech quality (SNR = 10 dB).

and noise generation models predicted 257-dimensional log amplitude which did not include the dynamic feature. The synthesized speech waveform was synthesized by applying Griffin-Lim's phase reconstruction [17] to the predicted log amplitude. Because we confirmed that the residual noise was included in the synthesized speech in the conventional and proposed methods, we applied weak spectral subtraction to the generated spectra so that their speech components were not perceptually distorted. No post-emphasis method, such as cepstrum [24], global variance [25], and modulation spectrum [3]-based methods, was used. Contextual features consisted of 439-dimensional linguistic features, 3-dimensional duration features, continuous log F_0 and voiced/unvoiced label [14]. In practice, the duration feature, continuous log F_0 and voiced/unvoiced labels must be extracted from the noisy speech. However, we used these ones extracted from clean speech to avoid speech quality degradation caused by extraction of these features from the noisy speech [26]. In the training phase, the context $\boldsymbol{x}$ and noisy speech log spectrum $\boldsymbol{y}_{\rm ns}$ were normalized to have zero-mean unit-variance. In the synthesizing phase, $\hat{\boldsymbol{y}}_{\rm s} = \boldsymbol{G}_{\rm s}(\boldsymbol{x})$ was un-normalized using the statistics of $\boldsymbol{y}_{\rm ns}$. Note that the un-normalization process is an ill-posed problem, i.e., only the scale of $\boldsymbol{y}_{\rm ns}$ scale is known, but we need to un-normalize the component $\boldsymbol{y}_{\rm s}$. The implementation of un-normalization is one aspect of the future work. The input of the noise generation model $\boldsymbol{n}_t$ was 100-dimensional vector randomly generated by the uniform distribution. 90% of the silence frames were removed in the speech synthesis model training phase. The DNN architecture for the speech synthesis model, noise generation model, and noise discrimination model were feed-forward neural networks, and the conventional and proposed methods used the same architectures of the speech synthesis models. The architecture for the speech and noise generation models included 3 $\times$ 512-unit leaky rectified linear unit (leaky ReLU) [27] hidden layers and a 257-unit linear output layer. Also, the architecture for the noise discrimination model included 3 $\times$ 512-unit leaky ReLU hidden layers and a one-unit sigmoid output layer. We initialized each model parameter using a random number and adoptive gradient algorithm AdaGrad [28] as the optimization algorithm.

B. Results of subjective evaluation

We compared two synthetic speech samples of the following methods:

- **SS+MSE**: the conventional method where the speech synthesis model is trained based on minimum squared error criterion after the spectral subtraction.
- **Proposed**: the proposed method where the speech synthesis model is trained so that the sum of its output and generated noise is close to spectra of noisy speech.

Note that vocoder-based approach was not used in this evaluation because the purpose of this research was to compare the vocoder-free DNN-based speech syntheses. The signal-to-noise ratio (SNR) was set as 0, 5, and 10 dB. Because the tradeoff between the amount of speech distortion (or the residual noise) and speech recognition accuracy was empirically known in DNN-based speech recognition, we believe that the same applies in DNN-based speech synthesis. Therefore, we used several kinds of noise suppression ratio of spectral subtraction, i.e., $\beta = 0.5, 1.0, 2.0,$ and 5.0. The smaller β makes less distortion of speech (but larger amount of residual noise) and the larger β makes stronger distortion. Preference AB tests were conducted to evaluate the naturalness of the synthetic speech in every pair of SNR and noise suppression ratio. The tests were done using our cloud-sourcing evaluation system. Twenty five listeners participated in each test; the total number of listeners was 300.

Figs. 4, 5, and 6 show the results of our experiments. Our method outperformed the conventional method in all cases. The p-values between the methods were smaller than 10^{-6}, meaning that the results were statistically significant. In Fig. 4, we can see that the score of the conventional method became worse as β increased in 0 dB SNR. We found that the speech distortion caused by strong spectral subtraction was significantly higher in speech synthesis training with the worse SNR. We think that this observation is an important factor in vocoder-free DNN-based speech synthesis trained from noisy speech.

V. Conclusion

In this paper, we introduced a generative approach to DNN-based speech synthesis trained from noisy speech using the proposed noise generation model. The noise generation model was trained to have natural statistics of the observed noise and to randomly generate the noise spectra. The speech synthesis model (acoustic model) was trained so that the sum of its output and generated noise was close to noisy speech spectra. We compared the proposed method and conventional method using noise reduction and standard speech synthesis training in synthetic speech quality. The results demonstrated that our method outperformed the conventional method under several signal-to-noise ratios and noise suppression ratios. As future work, we will introduce an activation matrix of non-negative matrix factorization [29] and compare it with the vocoder-based methods.

Acknowledgment

Part of this work was supported by SECOM Science and Technology Foundation, and JSPS KAKENHI Grant Number 18K18100.

References

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] S. Takamichi, K. Tomoki, and H. Saruwatari, "Sampling-based speech parameter generation using moment-matching network," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017.

[3] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.

[4] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[5] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*. Vancouver, Canada, May 2013.

[6] S. A.-E.-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," vol. abs/1609.08675, 2016.

[7] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, "Statistical bandwidth extension for speech synthesis based on Gaussian mixture model with sub-band basis spectrum model," *IEICE Transactions on Information and Systems*, vol. E99-D, no. 10, pp. 2481–2489, 2016.

[8] A. Saeb, R. Menon, H. Cameron, W. Kibira, J. Quinn, and T. Niesler, "Very low resource radio browsing for agile developmental and humanitarian monitoring," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 2118–2122.

[9] S. Takamichi and H. Saruwatari, "CPJD corpus: crowdsourced parallel speech corpus of japanese dialects," in *Proc. LREC*, Miyazaki, Japan, May 2018, pp. 434–437.

[10] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[11] 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.

[12] C. V.-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proc. INTERSPEECH*, Sep. 2016, pp. 352–356.

[13] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[14] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proc. NIPS*, pp. 2672–2680, 2014.

[16] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2080–2094, Sep. 2012.

[17] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2015.

[19] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," vol. abs/1609.03499, 2016.

[20] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. H.-Johnson, "Speech enhancement using Bayesian Wavenet," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 2013–2017.

[21] M. Sato, H. Manabe, H. Noji, and Y. Matsumoto, "Adversarial training for cross-domain universal dependency parsing," in *Proc. the CoNLL 2017 Shared Task*, Vancouver, Canada, Aug. 2017, pp. 71–79.

[22] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *arXiv preprint, 1711.00354*, 2017.

[23] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara, "A large-scale Japanese speech database," in *ICSLP90*, Kobe, Japan, Nov. 1990, pp. 1089–1092.

[24] 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*, Budapest, Hungary, Apr. 1999, pp. 2347–2350.

[25] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 816–824, 2007.

[26] P. Baljekar and A. W. Black, "Utterance selection techniques for TTS systems using found speech," in *Proc. SSW9*, Sunnyvale, CA, USA, Sep. 2016, pp. 199–204.

[27] L. A. Maas, Y. A. Hannun, and Y. A. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30.

[28] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *EURASIP Journal on Applied Signal Processing*, vol. 12, pp. 2121–2159, 2011.

[29] F. Cedric and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, Aug. 2011.