TEXT-TO-SPEECH SYNTHESIS USING STFT SPECTRA BASED ON LOW-/MULTI-RESOLUTION GENERATIVE ADVERSARIAL NETWORKS

Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari

Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Email: {yuuki_saito, shinnosuke_takamichi, hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp

ABSTRACT

This paper proposes novel training algorithms for vocoder-free statistical parametric speech synthesis (SPSS) using short-term Fourier transform (STFT) spectra. Recently, text-to-speech synthesis using STFT spectra has been investigated since it can avoid quality degradation caused by the vocoder-based parameterization in conventional SPSS using a vocoder. In conventional SPSS using a vocoder, we previously proposed a training algorithm for integrating generative adversarial network (GAN)-based distribution compensation. To extend the algorithm to vocoder-free SPSS, we propose low- and multi-resolution GAN-based training algorithms for vocoder-free SPSS. In our algorithm that uses the low-resolution GAN, acoustic models are trained to minimize the weighted sum of the mean squared error between natural and generated spectra in the original resolution and adversarial loss to deceive discriminative models in the lower resolution. Since the low-resolution spectra are close to filter banks and their distribution becomes simpler, GAN-based distribution compensation works well. Furthermore, we propose an algorithm using multi-resolution GANs, which uses both the low-resolution GAN and original-resolution GAN. Experimental results demonstrate that 1) the low-resolution GAN works robustly to the setting of its frequency resolution and hyperparameter, and 2) compared the low-, original-, and multi-resolution GANs, the low-resolution GAN works the best to improve synthetic speech quality.

Index Terms— Text-to-speech synthesis, vocoder-free SPSS, STFT spectra, generative adversarial networks, multi-resolution

1. INTRODUCTION

Text-to-speech synthesis (TTS) [1] is a technique to artificially synthesize human speech from linguistic information. Statistical parametric speech synthesis (SPSS) [2] using vocoder systems has been widely investigated because it can easily control the characteristics of synthetic speech. In conventional SPSS using vocoder systems, several steps are taken to synthesize desired speech. First, linguistic features and vocoder parameters are extracted from a training dataset that includes many pairs of text and speech. Then, acoustic models, which represent the relationship between the linguistic features and vocoder parameters, are trained with several criteria such as the mean squared error (MSE) [3] and minimum generation error (MGE) [4]. Finally, a synthetic speech waveform is synthesized from the predicted vocoder parameters by using high-quality vocoder systems such as STRAIGHT [5] and WORLD [6]. The high-quality vocoders have an important role in SPSS (especially, hidden Markov model-based SPSS [7, 8] and an early stage of deep neural network (DNN)-based SPSS [3]), but the quality degradation by vocoder-based parameterization in state-of-the-art DNN-based speech synthesis has become a critical problem.

One way to avoid this problem is vocoder-free DNN-based SPSS, which directly generates low level features before the vocoder-based parametrization such as short-term Fourier transform (STFT) spectra [9] and speech waveforms [10, 11]. This paper focuses on the vocoder-free SPSS using STFT spectra that we synthesize speech waveform from generated STFT spectra by using Griffin and Lim's phase reconstruction [12], not a vocoding process. This framework can avoid synthesizing buzzy speech caused by the vocoding process. It also provides a way to incorporate signal-processing techniques, such as speech enhancement [13], into speech synthesis training. However, in training the acoustic models, an over-smoothing effect is often observed in generated vocoder parameters or STFT spectra and significantly degrades synthetic speech quality [9, 14]. To address the over-smoothing effect in conventional SPSS with vocoders, we previously proposed a training algorithm [15, 16] incorporating generative adversarial networks (GANs) [17] so that the distribution of generated vocoder parameters is close to that of natural ones. It can effectively alleviate the over-smoothing effect and significantly improve synthetic speech quality without any post-processing methods, which require additional computations to improve speech quality, such as global variance compensation [18], modulation spectrum compensation [19, 20], and GAN-based post-filtering [21]. However, it is difficult to directly apply this algorithm to vocoder-free SPSS because the dimensionality of the STFT spectra is high and the distribution is more complex than that of the vocoder parameters.

To improve synthetic speech quality of vocoder-free SPSS using STFT spectra, we propose a novel algorithm to train acoustic models that uses a *low-resolution GAN*. Through a pooling layer along with a frequency axis, the STFT spectra are converted into low-resolution spectra. The training criterion of the acoustic model is the weighted sum of the MSE between natural and generated STFT spectra in the original frequency domain and adversarial loss using a discriminator of the low-frequency-domain GAN. The GAN in the low resolution can be regarded as compensating the difference between spectral envelopes of natural and synthetic speech because the low-resolution spectra approximately emulate filter banks. Since the spectral envelopes are dominant features in quality of synthetic speech and the effectiveness of the GAN is particularly noticeable in the case of generating spectral features, we can expect that the GAN-based distribution compensation improve the speech quality better than using the GAN in the original resolution. We also propose an algorithm that uses multi-resolution GANs (the low-resolution GAN and original-resolution GAN). Experimental results indicate that 1) the low-resolution GAN works robustly against the setting of its fre-

quency resolution and hyperparameter to control the weight for the adversarial loss, and 2) comparing among low-, original-, and multi-resolution GANs reveals that the low-resolution one works best to improve synthetic speech quality.

2. CONVENTIONAL ALGORITHMS

2.1. DNN-based TTS using STFT spectra

DNN-based acoustic models, which generate STFT spectral amplitudes from given linguistic features, are trained to minimize a loss function of natural and generated spectra. Let $\boldsymbol{y}$ be a natural spectra sequence $[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ and $\hat{\boldsymbol{y}}$ be a generated spectra sequence $[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$, where t and T denote the frame index and total frame length, respectively. Let $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(F)]^\top$ denote an STFT spectral amplitude vector at frame t, where F indicates the number of frequency bins from 0 Hz to the Nyquist frequency. The loss function for training the acoustic models is defined as the MSE between natural and generated spectral amplitudes as follows:

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

Referring to the study by Takaki et al. [9], we use the MSE loss rather than MGE loss [4]. After the training, $\hat{\boldsymbol{y}}$ is generated from the acoustic models, and its phase information is reconstructed using Griffin and Lim's method [12].

2.2. GAN-based training for SPSS [15]

In our previous study [15], in the same manner as with the algorithm of GANs [17], discriminative models $D(\cdot)$ are incorporated into the training, and the acoustic models and discriminative models are iteratively optimized. First, the discriminative models are updated to minimize the following discrimination loss:

$$L_{\mathrm{D}}^{(\mathrm{GAN})}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = L_{\mathrm{D},1}^{(\mathrm{GAN})}\left(\boldsymbol{y}\right) + L_{\mathrm{D},0}^{(\mathrm{GAN})}\left(\hat{\boldsymbol{y}}\right), \tag{2}$$

$$L_{\mathrm{D},1}^{(\mathrm{GAN})}\left(\boldsymbol{y}\right) = -\frac{1}{T}\sum_{t=1}^{T}\log D\left(\boldsymbol{y}_t\right), \tag{3}$$

$$L_{\mathrm{D},0}^{(\mathrm{GAN})}\left(\hat{\boldsymbol{y}}\right) = -\frac{1}{T}\sum_{t=1}^{T}\log\left(1 - D\left(\hat{\boldsymbol{y}}_t\right)\right), \tag{4}$$

where $L_{\mathrm{D},1}^{(\mathrm{GAN})}\left(\boldsymbol{y}\right)$ and $L_{\mathrm{D},0}^{(\mathrm{GAN})}\left(\hat{\boldsymbol{y}}\right)$ are the loss functions for natural and synthetic speech, respectively. The backpropagation algorithm is used to train $D(\cdot)$ to output 1 for natural speech and 0 for synthetic speech. Then, the acoustic models are updated to minimize the following loss:

$$L_{\mathrm{G}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = L_{\mathrm{MSE}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) + \omega_{\mathrm{D}}\frac{\mathbb{E}_{\hat{\boldsymbol{y}}}\left[L_{\mathrm{MSE}}\right]}{\mathbb{E}_{\hat{\boldsymbol{y}}}\left[L_{\mathrm{ADV}}\right]}L_{\mathrm{ADV}}^{(\mathrm{GAN})}\left(\hat{\boldsymbol{y}}\right), \tag{5}$$

where $L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) = L_{\mathrm{D},1}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}})$ is the adversarial loss to deceive the discriminative models, which makes the distribution of generated speech parameters close to that of natural speech. ω_{D} is a hyperparameter to control the effect of adversarial loss. $\mathbb{E}_{\hat{\boldsymbol{y}}}[L_{\mathrm{MSE}}]$ and $\mathbb{E}_{\hat{\boldsymbol{y}}}[L_{\mathrm{ADV}}]$ are the expectation values of L_{MSE} and L_{ADV}, respectively. Their ratio normalizes the scale of the two losses. Note that, the MSE loss is used in Eq. (5) instead of MGE loss [15].

3. PROPOSED ALGORITHMS

3.1. Training algorithm for STFT spectral amplitudes generation using low-resolution GAN

The method described in Subsection 2.2 can be applied to STFT spectra generation. However, it suffers from a higher dimensionality and complex distribution of the STFT spectra. We introduce low-resolution discriminative models $D^{(\mathrm{L})}(\cdot)$, which distinguish natural and generated STFT spectra in low frequency resolution. Let $\boldsymbol{\phi}(\cdot)$ be an average-pooling function that converts the STFT spectra in the original-frequency resolution $\boldsymbol{y}$ into those in the low-frequency resolution, $\boldsymbol{y}^{(\mathrm{L})}$. The f-th frequency bin of the low-resolution spectra at frame t, $y_t^{(\mathrm{L})}(f)$, is calculated as

$$y_t^{(\mathrm{L})}\left(f\right) = \frac{1}{w}\sum_{i=-p+(f-1)s}^{-p+(f-1)s+w} y_t\left(i\right), \tag{6}$$

where p, w, and s denote the size of zero-padding, width of pooling window, and stride of pooling, respectively. The term $y_t(i)$ takes 0 if $i < 0$ or $i > F$. The total number of frequency bins in the low-frequency resolution $F^{(\mathrm{L})}$ is given as

$$F^{(\mathrm{L})} = \frac{F + 2p - w}{s} + 1. \tag{7}$$

The above processes are similar to conversion from a raw STFT spectra into the filter-bank parameters that represent spectral envelopes of speech. The loss function for training the acoustic models is defined as follows:

$$L_{\mathrm{G}}^{(\mathrm{Low})}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = L_{\mathrm{MSE}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) + \omega_{\mathrm{D}}^{(\mathrm{L})}\frac{\mathbb{E}_{\hat{\boldsymbol{y}}}\left[L_{\mathrm{MSE}}\right]}{\mathbb{E}_{\hat{\boldsymbol{y}}^{(\mathrm{L})}}\left[L_{\mathrm{ADV}}\right]}L_{\mathrm{ADV}}^{(\mathrm{GAN})}\left(\hat{\boldsymbol{y}}^{(\mathrm{L})}\right), \tag{8}$$

where $\hat{\boldsymbol{y}}^{(\mathrm{L})} = \boldsymbol{\phi}(\hat{\boldsymbol{y}}^{(\mathrm{L})})$, and $\omega_{\mathrm{D}}^{(\mathrm{L})}$ is a hyperparameter to control the effect of the second term. This loss function can be regarded as the weighted sum of the MSE in the original resolution and adversarial loss in the lower resolution. Since the distributions of $\boldsymbol{y}^{(\mathrm{L})}$ and $\hat{\boldsymbol{y}}^{(\mathrm{L})}$ are simpler than those of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, we can overcome the difficulties in the training due to the high dimensionality and complex distribution. Also, we can expect the low-resolution GAN to dramatically improve the synthetic speech quality because it can capture the difference between spectral envelopes of natural and synthetic speech, which are dominant features in terms of the speech quality. The low-resolution discriminative models are trained in the same manner as in Eq. (2), but $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are replaced with $\boldsymbol{y}^{(\mathrm{L})}$ and $\hat{\boldsymbol{y}}^{(\mathrm{L})}$, respectively.

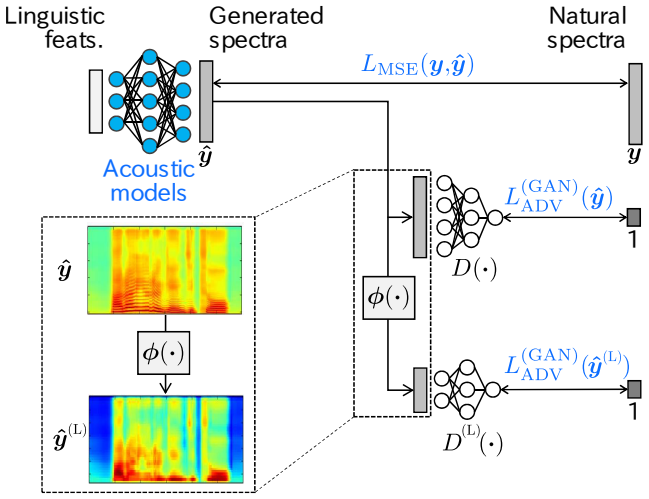

Fig. 1. Loss functions for updating acoustic models in proposed algorithm using multi-resolution GANs. $\boldsymbol{\phi}(\cdot)$ is average-pooling function to convert STFT spectra into low-resolution spectra.

3.2. Training algorithm for STFT spectral amplitudes generation using multi-resolution GANs

The proposed algorithm that uses the low-resolution GAN described in Subsection 3.1 can be extended to use multi-resolution GANs,

which introduces not only the low-resolution discriminative models $D^{(\mathrm{L})}(\cdot)$ but also original-resolution discriminative models $D(\cdot)$. The loss function for training the acoustic models is defined as follows:

$$L_{\mathrm{G}}^{(\mathrm{Multi})}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}} \frac{\mathbb{E}_{\hat{\boldsymbol{y}}}[L_{\mathrm{MSE}}]}{\mathbb{E}_{\hat{\boldsymbol{y}}}[L_{\mathrm{ADV}}]} L_{\mathrm{ADV}}^{(\mathrm{GAN})}(\hat{\boldsymbol{y}}) + \omega_{\mathrm{D}}^{(\mathrm{L})} \frac{\mathbb{E}_{\hat{\boldsymbol{y}}}[L_{\mathrm{MSE}}]}{\mathbb{E}_{\hat{\boldsymbol{y}}^{(\mathrm{L})}}[L_{\mathrm{ADV}}]} L_{\mathrm{ADV}}^{(\mathrm{GAN})}\left(\hat{\boldsymbol{y}}^{(\mathrm{L})}\right). \tag{9}$$

When $\omega_{\mathrm{D}} = 0$, this loss function is the same as that in Eq. (8). Figure 1 illustrates the computation procedure of the loss function. Note that the discriminative models are trained separately.

3.3. Discussion

Kaneko et al. [22] proposed a GAN-based post-filter for STFT spectra. As explained in Section 1, this post-filter-based approach requires additional computation in synthesis, but our algorithms do not. Also, because the previous work splits the STFT spectra into several sub-frequency bands and applies GANs to each band *independently*, it ignores the overall spectral structures (i.e., spectral envelope) and their correlation. On the other hand, our algorithms can effectively capture them.

By shifting our research from vocoder-level GANs [15] to STFT-level GANs (this study), we expect that it will become easier to extend the GAN-based algorithm, e.g., waveform-level GANs, in the future.

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

We used speech data of a Japanese female speaker who uttered 4007 sentences. The number of utterances included training and evaluation data were 3808 and 199, respectively. Speech signals were sampled at a rate of 16 kHz. The frame length, shift length, and FFT length were set to 400 (25 ms), 80 (5 ms), and 1024 samples, respectively. We used the hamming window for FFT analysis. In the training phase, linguistic features, which have a real value, and log spectral amplitudes were normalized to have zero-mean unit-variance. We removed 90% of the silence frames from the training data to improve training accuracy.

The DNN architectures for acoustic and discriminative models were Feed-Forward networks. The input of the acoustic models were 444-dimensional vectors including 439-dimensional linguistic features, 3-dimensional duration features, continuous log F_0, and U/V. The F_0 was extracted from speech data by using STRAIGHT vocoder systems [5]. We constructed DNNs, which predicted duration and F_0 features from linguistic features, in advance. The architecture for the acoustic models included 3 $\times$ 1024-unit rectified linear unit (ReLU) [23] hidden layers and a 513-unit linear output layer. The architecture for the discriminative models in the original resolution included 3 $\times$ 512-unit ReLU hidden layers and one unit sigmoid output layer. The architectures for the discriminative models in the lower resolution were almost same as that in the original resolution; that is, the activation functions used in the hidden and output layers were ReLU and sigmoid, the number of hidden layers was 3, but the number of input and hidden units varied in accordance with the parameters of the pooling function $\boldsymbol{\phi}(\cdot)$. In the following experiments, we fixed $p = 6$ and $s = w/2$ in Eq. (7). w was set to 14, 30, and 70. Accordingly, the number of input units F^{L} was set to 74, 34, 14, and the number of hidden units was set to 128, 64, 32, respectively.

Table 1. Preference scores of speech quality with their p-values (original-resolution GAN)

	Score	p-value	
Baseline	**0.700** vs. 0.300	$< 10^{-10}$	$\omega_{\mathrm{D}} = 0.5$
$\omega_{\mathrm{D}} = 1.0$	0.280 vs. **0.720**	$< 10^{-10}$	Baseline
$\omega_{\mathrm{D}} = 0.5$	0.496 vs. **0.504**	8.6×10^{-1}	$\omega_{\mathrm{D}} = 1.0$

In the training phase, we initialized the acoustic models by minimizing the MSE between natural and generated STFT spectra described in Subsection 2.1 with 25 iterations. Iteration means using all the training data (3808 utterances) once for training. The discriminative models in the original and lower resolution were initialized using natural speech and generated spectra after the initialization of the acoustic models. The number of iterations for the initialization was 5. The proposed training algorithms were used with 25 iterations. The expectation values for scaling the loss functions were estimated at each iteration step. We used AdaGrad [24] as the optimization algorithm, setting the learning rate to 0.01.

4.2. Subjective evaluations

We conducted subjective evaluations on the quality of the synthetic speech with various hyperparameter settings. A preference test (AB test) was conducted to evaluate the quality of speech produced from several algorithms. 25 listeners participated in each of the following evaluations by using our crowd-sourced evaluation systems, and each listener evaluated 10 samples. The total number of listeners was 375. In the following evaluations, "Baseline" denotes the method that trains the acoustic models using conventional MSE loss [9], i.e., both hyperparameters, ω_{D} and $\omega_{\mathrm{D}}^{(\mathrm{L})}$ in Eq. (9), were set to 0.

4.2.1. Evaluation of original-resolution GAN

First, to investigate the effect of GAN-based training in the original resolution (i.e., the same as [15]), we fixed $\omega_{\mathrm{D}}^{(\mathrm{L})} = 0$, and set $\omega_{\mathrm{D}} = 0.5$ or 1.0. We compared the quality of "Baseline" and our proposed algorithm using original-resolution GAN with "$\omega_{\mathrm{D}} = 0.5$," and "$\omega_{\mathrm{D}} = 1.0$." Table 1 shows the experimental results. Compared with "Baseline," the methods using the original-resolution GAN significantly degraded synthetic speech quality regardless of the hyperparameter settings. Therefore, we can confirm that simply applying the GAN-based training algorithm, which is effective in conventional SPSS with vocoders [15], does not improve STFT spectra generation.

4.2.2. Evaluation of low-resolution GAN

Next, to investigate effect of w, we fixed $\omega_{\mathrm{D}} = 0$ and set $\omega_{\mathrm{D}}^{(\mathrm{L})} = 1$. We compared the quality of generated speech samples using "Baseline" and our algorithm using the low-resolution GAN win "$w = 14$," "$w = 30$," and "$w = 70$." Table 2 shows the experimental results. From the results shown in Table 2(a), we can see that the proposed algorithm using the low-resolution GAN always achieved better scores than "Baseline," regardless of its parameter settings of the pooling function, which demonstrates the effectiveness of this algorithm. We set w to 30 in the following evaluation because Table 2(b) shows that "$w = 30$" was the best, although there were no significant differences among the scores.

We also investigated the effect of the hyperparameter in the low-resolution GAN. We fixed $\omega_{\mathrm{D}} = 0$ and set $\omega_{\mathrm{D}}^{(\mathrm{L})} = 0.5$ or 1.0. We compared the quality of generated speech using "Baseline" and our algorithm using the low-resolution GAN with "$\omega_{\mathrm{D}}^{(\mathrm{L})} = 0.5$," and "$\omega_{\mathrm{D}}^{(\mathrm{L})} = 1.0$." Table 3 shows the experimental results. From the

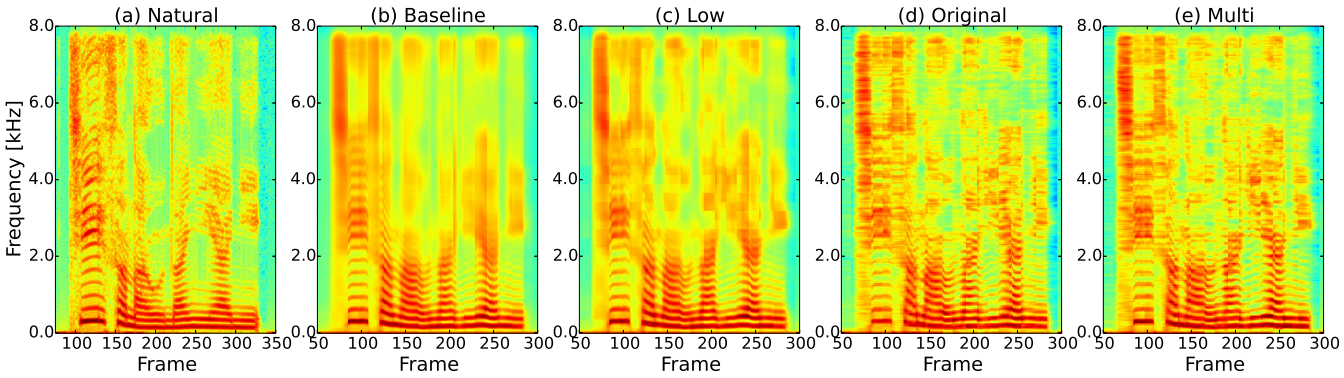

Fig. 2. STFT spectral magnitudes of natural and synthetic speech. We modified the ranges of temporal axis in (a) for clear illustration.

Table 2. Preference scores of speech quality with their p-values (low-resolution GAN with various pooling-parameter settings)

(a) Results of comparing "Baseline" with using low-resolution GAN

	Score	p-value	
$w = 14$	**0.568** vs. 0.432	2.3×10^{-3}	Baseline
$w = 30$	**0.572** vs. 0.428	1.2×10^{-3}	Baseline
$w = 70$	**0.528** vs. 0.472	2.1×10^{-1}	Baseline

(b) Results of proposed algorithms using low-resolution GANs

	Score	p-value	
$w = 14$	0.488 vs. **0.512**	5.9×10^{-1}	$w = 30$
$w = 30$	**0.532** vs. 0.468	1.5×10^{-1}	$w = 70$
$w = 70$	0.472 vs. **0.528**	2.1×10^{-1}	$w = 14$

Table 3. Preference scores of speech quality with their p-values (low-resolution GAN with various hyperparameter settings)

	Score	p-value	
Baseline	0.456 vs. **0.544**	4.9×10^{-2}	$\omega_{\mathrm{D}}^{(\mathrm{L})} = 0.5$
$\omega_{\mathrm{D}}^{(\mathrm{L})} = 1.0$	**0.588** vs. 0.412	7.6×10^{-5}	Baseline
$\omega_{\mathrm{D}}^{(\mathrm{L})} = 0.5$	**0.504** vs. 0.496	8.6×10^{-1}	$\omega_{\mathrm{D}}^{(\mathrm{L})} = 1.0$

results, we can conclude that the proposed algorithm using the low-resolution GAN successfully improved synthetic speech quality regardless of its hyperparameter settings.

4.2.3. Evaluation of multi-resolution GANs

Finally, we examined the effects of the proposed algorithm using the multi-resolution GAN. We generated speech samples using the following algorithms:

Original: $(\omega_{\mathrm{D}}, \omega_{\mathrm{D}}^{(\mathrm{L})}) = (1.0, 0.0)$

Low: $(\omega_{\mathrm{D}}, \omega_{\mathrm{D}}^{(\mathrm{L})}) = (0.0, 1.0)$

Multi: $(\omega_{\mathrm{D}}, \omega_{\mathrm{D}}^{(\mathrm{L})}) = (1.0, 1.0)$

Table 4 shows the results, Obviously, the proposed algorithm using the low-resolution GAN achieved a much higher score than the others. To investigate this reason, we plotted the STFT spectral magnitudes of synthetic speech used for the evaluations illustrated in Fig. 2. We can see that high randomness observed in natural spectra

Table 4. Preference scores of speech quality with their p-values (multi-resolution GANs)

	Score	p-value	
Low	**0.808** vs. 0.192	$< 10^{-10}$	Multi
Multi	0.492 vs. **0.508**	7.2×10^{-1}	Original
Original	0.192 vs. **0.808**	$< 10^{-10}$	Low

(Fig. 2(a)) was excessively smoothed in synthetic speech of "Baseline" (Fig. 2(b)), while the proposed three algorithms reproduced the randomness by using GANs, However, there were some temporal discontinuities in the spectra generated by using original- and multi-resolution GANs (Figs. 2(d) and (e)), which might considerably degrade the synthetic speech quality. One can address the quality degradation by using recurrent architectures such as long-short term memory [25, 26] for the acoustic and discriminative models to make them capture the temporal dependency of the STFT spectra. Further improvements also can be achieved by conditioning the GANs with the specific information of the utterance such as the phonetic contents, and U/V [27].

5. CONCLUSION

We proposed two training algorithms to incorporate generative adversarial networks (GANs) into vocoder-free speech synthesis using short-term Fourier transform (STFT) spectra. In the proposed algorithm using a low-resolution GAN, acoustic models are trained to minimize the mean squared error between natural and generated STFT spectral amplitudes at the original resolution and the distribution differences of their distributions at low resolution. This algorithm can be extended to one using multi-resolution GANs, which also minimizes the distribution differences of natural and generated STFT spectra at the original resolution. Experimental results indicated that the algorithm using the original-resolution GAN and our proposed algorithm using multi-resolution GANs degraded synthetic speech quality, but the proposed algorithm using the low-resolution GAN successfully improved it. In the future, we will further investigate the effects of the hyperparameters of the proposed algorithms and adopt a conditional GAN [28] for training.

Acknowledgements: Part of this work was supported by SECOM Science and Technology Foundation, and JSPS KAKENHI Grant Number 16H06681 and 17H06101.

6. REFERENCES

[1] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

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

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

[4] Z. Wu and S. King, "Improving trajectory modeling for DNN-based speech synthesis by using stacked bottleneck features and minimum trajectory error training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, Jul. 2016.

[5] 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, Apr. 1999.

[6] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," 2016, vol. E99-D, pp. 1877–1884.

[7] H. Zen and T. Toda, "An overview of nitech hmm-based speech synthesis system for blizzard challenge 2005," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sep. 2005, pp. 93–96.

[8] K. Tokuda, Y Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, Apr. 2013.

[9] 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, pp. 1128–1132.

[10] A. 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," *arXiv*, vol. abs/1609.03499, 2016.

[11] 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," *arXiv*, vol. abs/1612.07837, 2016.

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

[13] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[14] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. INTERSPEECH*, California, U.S.A., Sep. 2016, pp. 1632–1636.

[15] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," in *Proc. ICASSP*, New Orleans, U.S.A., Mar. 2017, pp. 4900–4904.

[16] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *arXiv (preprint of IEEE/ACM Transactions on Audio, Speech, and Language Processing)*, vol. abs/1709.08041, 2017.

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

[18] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[19] 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, Apr. 2016.

[20] Z.-H. Ling, X.-H. Sun, L.-R. Dai, and Y. Hu, "Modulation spectrum compensation for hmm-based speech synthesis using line spectral pairs," in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5595–5599.

[21] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. ICASSP*, New Orleans, U.S.A., Mar. 2017, pp. 4910–4914.

[22] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3389–3393.

[23] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, Lauderdale, U.S.A., Apr. 2011, pp. 315–323.

[24] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, Jul. 2011.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4470–4474.

[27] S. Yang, L. Xie, X. Chen, X. Lou, X. Zhu, D. Huang, and H. Li, "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework," *arXiv*, vol. abs/1707.01670, 2017.

[28] M. Mirza and S. Osindero, "Conditional generative adversarial networks," *arXiv*, vol. abs/1411.1784, 2014.