10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Multi-Speaker Modeling for DNN-based Speech Synthesis Incorporating Generative Adversarial Networks

Hiroki Kanagawa and Yusuke Ijima

NTT Media Intelligence Laboratories, NTT Corporation.
hiroki.kanagawa.wk@hco.ntt.co.jp

Abstract

This paper presents a novel DNN-based speech synthesis method we derived from multi-speaker training data. In general, speaker-dependent modeling techniques based on generative adversarial networks (GANs) improve synthesized speech quality. However, they are inadequate for multi-speaker training because conventional discriminators cannot take into account speaker identity, which degrades anti-spoofing performance in GAN discriminators. We introduce two approaches as means to learn GAN speaker characteristics, i.e., auxiliary features and tasks. The first uses speaker codes as additional discriminator input. The second uses speaker identification as a means to verify that anti-spoofing verification methods are effective. Experimental results showed that our proposed techniques outperformed conventional and GAN-based methods.

Index Terms: speech synthesis, multi-speaker modeling, deep neural networks, generative adversarial networks

1. Introduction

Recent studies on DNN-based statistical parametric speech synthesis (SPSS) have produced dramatic improvements [1], but much arbitrary target speaker's data is required to build an acoustic model. Multi-speaker modeling is an effective approach to training acoustic models with little target speaker data. This approach can compensate the target speaker's unseen patterns from other multiple speakers. Considerable multi-speaker modeling has been proposed for both HMM [2, 3, 4] and DNN-based SPSS [5, 6, 7, 8]. In particular, DNN-based methods often perform well when speaker representations such as speaker codes and i-vectors [5, 9] are fed to a DNN. They also perform well when speaker specific DNN output layers [8] are used. Despite these approaches' advantages, using them tends to degrade synthesized speech quality. This degradation occurs because over-smoothed acoustic parameters are produced due to training with the minimum mean squared error (MMSE) criterion.

Two approaches have been presented to alleviate this over-smoothing problem. One is analytical approach and the other is generative adversarial training. In the former, analytic criteria are designed to compensate a difference between natural and synthetic speech [10, 11, 12]. Toda et al. [10] focused on the significant decrease in the synthetic acoustic feature's global variance (GV) and gave a constraint at the parameter generation so as to generate acoustic features with appropriate GV. In contrast, generative adversarial networks (GANs) [13] have attempted to decrease the difference in acoustic features between synthesized and real speech [14, 15, 16]. They focus on the remarkable difference in acoustic feature distribution between natural and a synthetic speech [17]. In this approach, the acoustic model is trained so as to reduce this difference explicitly by using a discriminator network, unlike analytic approaches. However, the conventional GANs are ineffective under multi-speaker conditions because features appearing in natural or synthetic speech differ for each speaker. In fact, Zhao et al. [18] demonstrated that some speakers had lower quality and speaker similarity on GAN-based multi-speaker modeling.

This paper presents a new GAN-based speech synthesis technique that can be used under multi-speaker conditions. We introduce two types of discriminator structures with speaker information into multi-speaker modeling incorporating a GAN. The key idea behind our method is to model speaker characteristics in both generators and discriminators. In the first structure, speaker information is fed to the discriminator and also to a conditional GAN [19]. In this way, we expect speaker information to act as a constraint on training each speaker's characteristics as well as multi-speaker modeling. The second structure introduces speaker identification into the discriminator and optimizes it simultaneously by using multi-task learning. The key idea behind this approach was inspired by a useful multi-class GAN technique [20] for image generation. We expect that speaker-specific characteristics, which cannot be expressed by simply feeding handcraft features such as speaker codes, are automatically acquired. This makes accurate GAN training possible by determining natural or synthetic speech while identifying training speakers. Experimental results showed that merely applying a GAN to multiple speakers doesn't produce significant improvement. We also found that speaker information is important, especially that obtained with discriminators incorporating multi-class GANs.

Even though WaveNet [21] also improves synthetic speech quality, it takes a large computational cost. We focus on a conventional vocoder-based approach to achieve both natural-sounding speech and a low computational cost.

2. DNN acoustic modeling training criteria

In this section, we will give a brief review of DNN-based acoustic modeling criteria.

2.1. Minimum mean squared error (MMSE)

We used a DNN acoustic model as a mapping module from linguistic features $\boldsymbol{l} = [\boldsymbol{l}_1, \dots, \boldsymbol{l}_T]$ obtained from text to acoustic features $\boldsymbol{c} = [\boldsymbol{c}_1, \dots, \boldsymbol{c}_T]$ extracted from speech waveforms, where T is the total number of frames in a speech. For this purpose, the minimum mean squared error (MMSE) criterion is usually used as the loss function in DNN-based speech synthesis [1]. In this framework, the DNN is trained so as to minimize the mean squared error expressed as

$$\mathcal{L}_{\mathrm{MSE}}\left(\boldsymbol{c}, \hat{\boldsymbol{c}}\right) = \mathbb{E}\left[\left(\boldsymbol{c} - \hat{\boldsymbol{c}}\right)^2\right], \tag{1}$$

where $\hat{\boldsymbol{c}} = [\hat{\boldsymbol{c}}_1, \dots, \hat{\boldsymbol{c}}_T]$ denotes the acoustic feature predicted by DNN and $\mathbb{E}\left[\cdot\right]$ is the expectation operator.

10.21437/SSW.2019-8

2.2. Generative adversarial networks (GANs)

GAN-based SPSS attempts to generate synthesized speech, which is difficult to distinguish from natural speech. Saito et al. [14] employed anti-spoofing verification (ASV) [22] as the GAN discriminator. The ASV loss function is defined as

$$\mathcal{L}_{\mathrm{ASV}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \mathcal{L}_{\mathrm{ASV},1}(\boldsymbol{c}) + \mathcal{L}_{\mathrm{ASV},0}(\hat{\boldsymbol{c}}), \tag{2}$$

$$\mathcal{L}_{\mathrm{ASV},1}(\boldsymbol{c}) = -\mathbb{E}\left[\ln D_{\mathrm{ASV}}(\boldsymbol{c})\right], \tag{3}$$

$$\mathcal{L}_{\mathrm{ASV},0}(\hat{\boldsymbol{c}}) = -\mathbb{E}\left[\ln\left(1 - D_{\mathrm{ASV}}(\hat{\boldsymbol{c}})\right)\right], \tag{4}$$

where $D_{\mathrm{ASV}}(\cdot)$ is the ASV's forward calculation operator. The acoustic model is trained so as to deceive the ASV, the loss function being expressed as

$$\mathcal{L}_{\mathrm{G}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \mathcal{L}_{\mathrm{MSE}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) + w_D \frac{E_{\mathrm{MSE}}}{E_{\mathrm{ASV}}} \mathcal{L}_{\mathrm{ASV},1}(\hat{\boldsymbol{c}}), \tag{5}$$

where E_{MSE} and E_{ASV} are respectively the expectation values of $\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{c}, \hat{\boldsymbol{c}})$ and $\mathcal{L}_{\mathrm{ASV},1}(\hat{\boldsymbol{c}})$ calculated by the previous iteration. The hyperparameter w_D controls the effect of adversarial training. Eq. (5) is defined as the weighted sum of the MMSE loss and the ASV one. The second term especially compensates the difference in distributions between natural and synthetic speech.

3. Multi-speaker modeling incorporating GAN

A GAN is a promising way to address the over-smoothing problem described in Section 1. However, it is not revealed 1) whether the synthesized speech quality can also be improved by utilizing a GAN with multi-speaker modeling, and 2) whether speaker information is also effective for a discriminator in the same manner as multi-speaker acoustic modeling. In this study, in order to answer these questions, we introduced a GAN into multi-speaker modeling, and utilized speaker information in the discriminator.

3.1. Basic approach

In multi-speaker modeling using speaker codes [5], both linguistic features and speaker codes are fed into a DNN. The speaker code is represented as a one-hot vector as

$$\boldsymbol{s} = \left[s^1, \ldots, s^M\right]^\top, \tag{6}$$

where M denotes the number of training speakers. To apply a GAN to the multi-speaker, we simply use the concatenated vector $\boldsymbol{v} = \left[\{\boldsymbol{l}_1^\top, \boldsymbol{s}_1^\top\}^\top, \ldots, \{\boldsymbol{l}_T^\top, \boldsymbol{s}_T^\top\}^\top\right]$ instead of the linguistic feature as the acoustic model input described in Section 2.1. Algorithm 1 describes the procedure of multi-speaker modeling incorporating a GAN.

3.2. Speaker information utilization via conditional GAN

We attempted to train the GAN while taking speaker characteristic differences into consideration. In multi-speaker modeling, speaker information is available as an auxiliary feature.

To exploit this, not only acoustic features of natural speech $\boldsymbol{c}$ or synthetic speech $\hat{\boldsymbol{c}}$ but also speaker codes $\boldsymbol{s}$ (Eq. (6) are fed to the discriminator. This method is generally called conditional GAN [19] and its effectiveness has been shown in SPSS [15, 16].

Algorithm 1 Algorithm of multi-speaker modeling incorporating GAN.

```
Input: Linguistic feature $\boldsymbol{l}$, speaker code $\boldsymbol{s}$
  Obtain vector $\boldsymbol{v}$ by concatenating $\boldsymbol{l}$ with $\boldsymbol{s}$
  for the number of iteration times do
    Generate acoustic feature $\hat{\boldsymbol{c}}$ by inputting $\boldsymbol{v}$ to the multi-speaker acoustic model
    Update discriminator by Eq. (2)
    Update acoustic model by Eq. (5)
  end for
Output: Multi-speaker acoustic model, discriminator
```

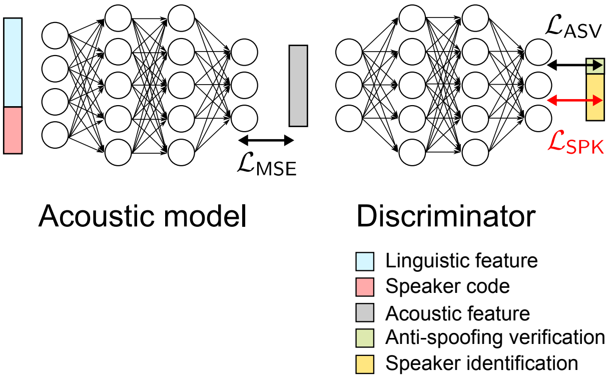

$\mathcal{L}_{\mathrm{ASV}}$
$\mathcal{L}_{\mathrm{SPK}}$
$\mathcal{L}_{\mathrm{MSE}}$
Acoustic model
Discriminator
Linguistic feature
Speaker code
Acoustic feature
Anti-spoofing verification
Speaker identification

Figure 1: *Outline of multi-task learning with speaker identification. ASV and speaker identification are assigned as discriminator outputs. The discriminator is simultaneously optimized using each adversarial loss.*

3.3. Multi-task learning with speaker identification via multi-class discriminator

Conventional GANs train acoustic models so as to only deceive the ASV. However, dealing with multi-speaker data makes it difficult to represent the difference in feature distributions for each speaker. Our proposed discriminator aims at acquiring specific speaker characteristics, which cannot be expressed by simply feeding handcraft features such as speaker codes.

Figure 1 shows the outline of our method. Our discriminator has ASV and speaker identification, which are optimized simultaneously by minimizing the loss function, expressed as

$$\mathcal{L}_{\mathrm{D_{MTL}}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \mathcal{L}_{\mathrm{ASV}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) + \mathcal{L}_{\mathrm{SPK}}(\boldsymbol{c}, \hat{\boldsymbol{c}}), \tag{7}$$

$$\mathcal{L}_{\mathrm{SPK}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \mathcal{L}_{\mathrm{SPK},1}(\boldsymbol{c}) + \mathcal{L}_{\mathrm{SPK},0}(\hat{\boldsymbol{c}}), \tag{8}$$

$$\mathcal{L}_{\mathrm{SPK},1}(\boldsymbol{c}) = -\mathbb{E}\left[\ln D_{\mathrm{SPK}}(\boldsymbol{c})\right], \tag{9}$$

$$\mathcal{L}_{\mathrm{SPK},0}(\hat{\boldsymbol{c}}) = -\mathbb{E}\left[\ln\left(1 - D_{\mathrm{SPK}}(\hat{\boldsymbol{c}})\right)\right], \tag{10}$$

where $D_{\mathrm{SPK}}(\boldsymbol{x}) = Z(\boldsymbol{x}) / (Z(\boldsymbol{x}) + 1)$, $Z(\boldsymbol{x}) = \sum_{k=1}^{M} e^{l_k(\boldsymbol{x})}$. $l_k(\boldsymbol{x})$ denotes the kth speaker identifier output given the input vector $\boldsymbol{x}$[1]. To optimize the acoustic model, the cost function is defined as

$$\mathcal{L}_{\mathrm{G_{MTL}}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = \mathcal{L}_{\mathrm{MSE}}(\boldsymbol{c}, \hat{\boldsymbol{c}}) + w_{\mathrm{D}} \frac{E_{\mathrm{MSE}}}{E_{\mathrm{ASV}} + E_{\mathrm{SPK}}} \left\{\mathcal{L}_{\mathrm{ASV},1}(\hat{\boldsymbol{c}}) + \mathcal{L}_{\mathrm{SPK},1}(\hat{\boldsymbol{c}})\right\}, \tag{11}$$

where E_{SPK} is the expectation value of $\mathcal{L}_{\mathrm{SPK},1}(\hat{\boldsymbol{c}})$ calculated by the previous iteration.

[1] Although an ordinary cross-entropy based speaker identification can be also used, we used a multi-class discriminator in our study.

Table 1: *Methods used for experiments.*

Method	Detail
MMSE	Conventional multi-speaker model [5]
GAN	Multi-speaker model incorporating GAN (Section 3.1)
cGAN	Speaker codes fed to discriminator (Section 3.2)
GAN-SPK	Simultaneous GAN optimization with speaker identification (Section 3.3)

In the image generation field, multi-class discriminators [20] have been reported to be effective when the generator predicts multi-class images. Our method was inspired by this method, as the discriminator distinguishes natural or synthetic speech while identifying the speech for each speaker.

4. Experiments

4.1. Setup

We used a Japanese read speech database comprising 47 speakers. Four speakers, two males and two females, were used as target speakers. We took 30 utterances from each target speaker as evaluation data (total about 5.4 minutes). All other data were used as a training set (39.8 hours), which included target speakers (about 13 minutes per speaker), and a validation one (2.1 hours). The corpus sampling rate was 22.05 kHz. We used a STRAIGHT vocoder [23] to extract forty-dimensional mel-cepstral coefficients, ten band aperiodicities, F0 in log-scale, and voiced-unvoiced flags at five msec steps. Acoustic features comprised 52 dimensional vectors. Linguistic features comprised 305 dimensional vectors including phonemes, accent types, and frame positions in phonemes. In the training phase, acoustic features were normalized to have zero-mean unit-variance by CMVN [24].

The acoustic model is comprised of four feed-forward and two uni-directional long short-term memories (LSTMs) [25]. Their unit sizes were 280. The speaker codes were fed to the first and all hidden layer in the same manner as Hojo et al. [5]. The discriminator is comprised of three feed-forward and one uni-directional LSTM; their unit sizes were 200. Sigmoid was applied to both the acoustic model and discriminator as an activation function. In cGAN, the speaker codes were fed to the first layer. The minibatch size was eight and Adam [26] was used to optimize model parameters of both models. All mel-spectral coefficients except the 0th were fed only to the discriminator.

Because a GAN works robustly against the weight for adversarial loss w_D in Eq. (5) [14], all experiments were performed at $w_D = 1.0$. The acoustic model was initialized by using the MMSE criterion. The discriminator was initialized using natural and synthetic speech obtained from the initialized acoustic model. The number of initial iterations for the acoustic model and the discriminator are respectively 50 and five. Both models were updated alternately with 30 iterations by using GAN's criteria.

Table 1 shows the four methods used in the experiments. When testing, original phoneme durations were used for all methods.

4.2. Objective evaluations

We examine the effectiveness of GAN from various aspects because improvements are not necessarily confirmed by using a single objective measure.

4.2.1. Global variance (GV)

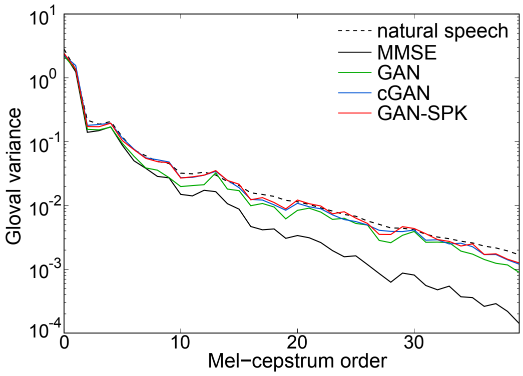
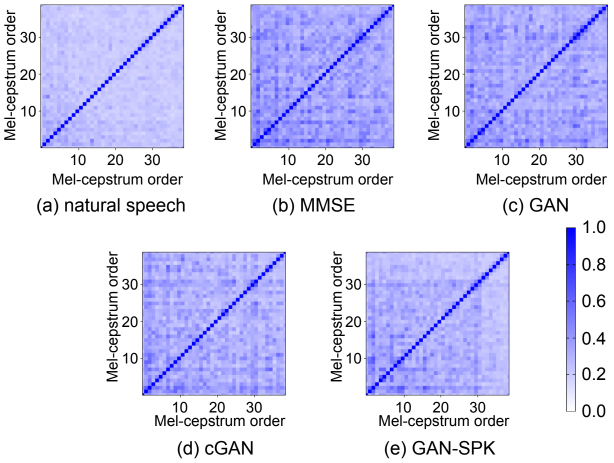


Figure 2: *Average GV of mel-cepstrum obtained from all evaluation utterances.*



(a) natural speech (b) MMSE (c) GAN (d) cGAN (e) GAN-SPK

Figure 3: *MICs between dimensions of mel-cepstrum coefficients obtained from one target female speaker's evaluation utterance.*

To confirm the improvement regarding over-smoothing problem, we calculated GV. Figure 2 shows the average GV obtained from all evaluation data utterances. The GV of natural speech was greater than that for other methods in terms of overall dimensions, whereas that of *MMSE* significantly decreased in terms of high orders of the mel-cepstral coefficients. The GV of *GAN* keeps close to natural ones unlike *MMSE*. Moreover, we can see that the methods with speaker information (*cGAN* and *GAN-SPK*) obtained slightly larger GV than *GAN*'s, and both methods are comparable.

4.2.2. Maximum information coefficients (MIC)

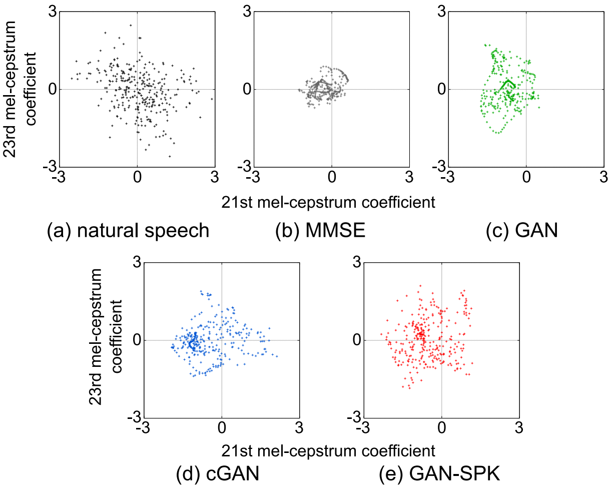
We also calculated maximum information coefficients (MICs) [27]. MICs are correlated with subjective evaluation scores of naturalness [17]. Figure 3 shows the results obtained from one target female speaker's evaluation utterance. The MIC of natural speech has a weak correlation whereas the MIC of *MMSE* has a strong one. GAN-based methods have weaker correlations than *MMSE*. We quantitatively evaluated the difference between natural speech and synthetic speech by using the distance defined as

$$d = \frac{1}{R}\sum_{r=1}^{R}\left\|\boldsymbol{M}_{\mathrm{natural}}^{(r)} - \boldsymbol{M}_{\mathrm{synthetic}}^{(r)}\right\|_2, \tag{12}$$

where r and R are respectively the utterance index and the number of utterances. $\|\cdot\|_2$ denotes the Frobenius norm and $\boldsymbol{M}_{\mathrm{natural}}^{(r)}$, $\boldsymbol{M}_{\mathrm{synthetic}}^{(r)}$ are respectively MIC of natural speech and that of synthetic speech.

Table 2: *Average MIC distance d between natural and synthetic speech obtained from all evaluation utterances (Eq.* (12)*).*

Method	*d*
MMSE	5.20
GAN	4.87
cGAN	4.05
GAN-SPK	**3.75**



(a) natural speech (b) MMSE (c) GAN

(d) cGAN (e) GAN-SPK

Figure 4: *Distributions between 21st and 23rd mel-cepstrum coefficients obtained from one target female speaker's evaluation utterance.*

Table 2 shows the average MIC distances *d* obtained from all evaluation utterances. The *GAN*'s distance is smaller than *MMSE*'s. *cGAN* has further small distance, especially *GAN-SPK*'s distance is minimum. These results revealed that our proposed methods can generate mel-cepstrum coefficients which have correlations close to that of natural speech.

4.2.3. Distribution of mel-cepstram coefficients

In this section, we confirmed whether the synthetic speech parameter distribution became close to that of natural speech. Figure 4 shows scatter plots of the 21st and 23rd mel-cepstrum coefficients obtained from one target female speaker's utterances. The horizontal and vertical axes indicate respectively the 21st and 23rd order mel-cepstrum coefficients. We can see natural speech has a wide distribution, whereas that of *MMSE* is degenerated. Although *GAN* makes it possible to avoid distribution degeneration, its shape is significantly different from that of natural speech. On the other hand, *cGAN* and *GAN-SPK* can reproduce shapes closer to that of natural speech as compared to *MMSE* and *GAN*. These results also show that utilizing speaker information is effective for multi-speaker modeling.

Next, we used Jensen-Shannon (JS) divergence to quantify the difference in mel-cepstram distribution between natural and synthetic speech for all dimensions. Figure 5 shows the average JS divergence obtained from all evaluation utterances. The JS divergence of *MMSE* increases in terms of high coefficient order. Although *GAN* gives smaller JS divergences than that of *MMSE* for overall dimensions, even smaller JS divergences are obtained from *cGAN* and *GAN-SPK*. These objective evaluations led us to ascertain that *GAN* with speaker information appropriately compensates the synthetic feature distribution.

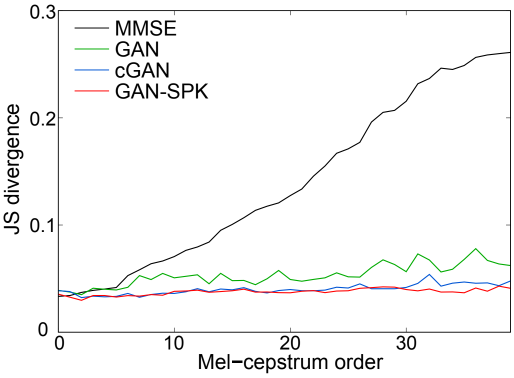


Figure 5: *Average JS divergence of mel-cepstrum coefficients obtained from all evaluation utterances.*

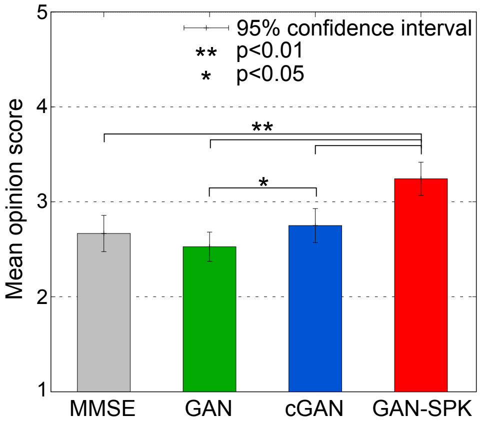


Figure 6: *Mean opinion scores of synthetic speech on naturalness.*

4.3. Subjective evaluation

We subjectively evaluated synthetic speech naturalness by using mean opinion score (MOS) on a five-point scale ranging from 5: very natural to 1: very unnatural. 20 utterances are randomly selected for each method. Seven listeners participated in the test, each of them giving scores for synthetic speech.

Figure 6 shows the subjective evaluation results. *GAN* underperformed *MMSE* because *GAN* raised unexpected noises due to incorrectly expanded distribution as mentioned in Section 4.2.3. On the other hands, *cGAN* slightly outperformed *MMSE* and *GAN*, and *GAN-SPK* achieved significant improvement compared to other methods. In our proposed methods, unexpected noises as seen in *GAN* were not found in evaluation data. From these results, we found that speaker information is effective for multi-speaker GAN modeling, and especially a multi-class discriminator incorporating speaker identification can take account of the difference among training speakers precisely.

5. Conclusion

We have introduced a generative adversarial network (GAN) into multi-speaker modeling and investigated the relationship between the quality of synthesized speech and how to consider speaker information in discriminators. Evaluation results demonstrated that simply using a GAN does not produce significant improvement. However, we achieved better naturalness by using speaker information. Particularly the best performance was obtained by multi-class discriminator optimization with speaker identification.

6. References

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

[2] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals*, vol. E86-A, no. 8, pp. 1956–1963, 2003.

[3] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," *Proc. ICASSP*, pp. 1611–1614, 1997.

[4] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," *Proc. ICASSP*, pp. 805–808, 2001.

[5] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-based speech synthesis using speaker codes," *IEICE Trans. on Information and Systems*, vol. E101-D, no. 2, pp. 462–472, 2018.

[6] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," *Proc. INTERSPEECH*, pp. 879–883, 2015.

[7] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for CNN-based speech synthesis," *Proc. ICASSP*, pp. 6950–6954, 2019.

[8] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," *Proc. ICASSP*, pp. 4475–4479, 2015.

[9] Y. Zhao, D. Saito, and N. Minematsu, "Speaker representations for speaker adaptation in multiple speakers' BLSTM-RNN-based speech synthesis," *Proc. INTERSPEECH*, pp. 2268–2272, 2016.

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

[11] T. Nose, V. Chunwijitra, and T. Kobayashi, "A parameter generation algorithm using local variance for HMM-based speech synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 8, no. 2, pp. 221–228, 2014.

[12] 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 Trans. Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.

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

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

[15] 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," *Proc. ASRU*, pp. 685–691, 2017.

[16] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on generative adversarial networks," *Proc. ICASSP*, pp. 6955–6959, 2019.

[17] Y. Ijima, T. Asami, and H. Mizuno, "Objective evaluation using association between dimensions within spectral features for statistical parametric speech synthesis," *Proc. INTERSPEECH*, pp. 337–341, 2016.

[18] Y. Zhao, S. Takaki, H.-T. Luong, J. Yamagishi, D. Saito, and N. Minematsu, "Wasserstein gan and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a wavenet vocoder," *IEEE Access*, vol. 6, no. 1, pp. 60 478–60 488, 2018.

[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[20] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *Proc. NIPS*, pp. 2234–2242, 2016.

[21] A. van den 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," *arXiv preprint arXiv:1609.03499*, 2016.

[22] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection – the SJTU system for ASVspoof 2015 Challenge," *Proc. INTERSPEECH*, pp. 2097–2101, 2015.

[23] H. Kawahara, I. Masuda-Katsuse, and A. de 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, pp. 187–207, 1999.

[24] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

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

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.