IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 27, 2018, accepted September 20, 2018, date of publication September 24, 2018,
date of current version November 8, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2872060

Wasserstein GAN and Waveform Loss-Based Acoustic Model Training for Multi-Speaker Text-to-Speech Synthesis Systems Using a WaveNet Vocoder

YI ZHAO[1], (Student Member, IEEE), SHINJI TAKAKI[2], (Member, IEEE), HIEU-THI LUONG[2], (Student Member, IEEE), JUNICHI YAMAGISHI[2,3], (Senior Member, IEEE), DAISUKE SAITO[1], (Member, IEEE), AND NOBUAKI MINEMATSU[1], (Member, IEEE)

[1]Department of Electrical Engineering and Information Systems, Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan
[2]Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan
[3]The Centre for Speech Technology Research, The University of Edinburgh, Edinburgh EH89AB, U.K.

Corresponding author: Yi Zhao (zhaoyi@gavo.t.u-tokyo.ac.jp)

This work was supported by MEXT KAKENHI under Grants 16H06302, 17H04687, 18H04120, 18H04112, and 18KT0051.

ABSTRACT WaveNet, which learns directly from speech waveform samples, has been used as an alternative to vocoders and achieved very high-quality synthetic speech in terms of both naturalness and speaker similarity even in multi-speaker text-to-speech synthesis systems. However, the WaveNet vocoder uses acoustic features as local condition parameters, and these parameters need to be accurately predicted by another acoustic model. So far, it is not yet clear how to train this acoustic model, which is problematic because the final quality of synthetic speech is significantly affected by the performance of the acoustic model. Significant degradation occurs, especially when predicted acoustic features have mismatched characteristics compared to natural ones. In order to reduce the mismatched characteristics between natural and generated acoustic features, we propose new frameworks that incorporate either a conditional generative adversarial network (GAN) or its variant, Wasserstein GAN with gradient penalty (WGAN-GP), into multi-speaker speech synthesis that uses the WaveNet vocoder. The GAN generator performs as an acoustic model and its outputs are used as the local condition parameters of the WaveNet. We also extend the GAN frameworks and use the discretized-mixture-of-logistics (DML) loss of a well-trained WaveNet in addition to mean squared error and adversarial losses as parts of objective functions. Experimental results show that acoustic models trained using the WGAN-GP framework using back-propagated DML loss achieves the highest subjective evaluation scores in terms of both quality and speaker similarity.

INDEX TERMS Generative adversarial network, multi-speaker modeling, speech synthesis, WaveNet.

I. INTRODUCTION

A. GENERAL BACKGROUND

In recent years, text-to-speech (TTS) synthesis has gained popularity as an artificial intelligence technique and is widely used in many applications with speech interfaces. There are currently two major categories in the machine learning-based speech synthesis field: a) an end-to-end approach that learns the relationship between text and speech directly and b) the conventional pipeline processing approach that divides text-to-speech conversion into sub tasks such as linguistic feature extraction and acoustic feature extraction. In the latter approach, an acoustic model is trained to learn the relationship between separately extracted linguistic and acoustic features [1]. Previously investigated acoustic models include the hidden Markov model (HMM) [2], the deep neural network (DNN) [3], and the recurrent neural network (RNN) [4], [5]. These are normally trained with the minimum mean squared error (MSE) criterion, and hence, the generated acoustic parameters tend to be over-smoothed regardless of the architectures. Finally, speech waveforms have been reconstructed using a deterministic vocoder based on the acoustic parameters [6]–[8]. However, the generated signals have artifacts and typically sound buzzy. Due to these two major issues, the resultant quality of generated speech sounds obviously worse compared with natural speech.

Very recently, we see emerging solutions for the two issues. To alleviate the over-smoothing problem, Saito *et al.* [9], [10] have incorporated adversarial training into acoustic modeling. The generative adversarial network (GAN) contains a generator as well as a discriminator [11], where the generator aims at deceiving the discriminator and the discriminator is trained to distinguish the natural and generated feature samples. In the framework proposed by Saito *et al.* [10], the generator acts as an acoustic model and is optimized by not only the conventional MSE but also an adversarial loss computed using the discriminator. Experimental results show that GAN can effectively alleviate the over-smoothing effect of the generated speech parameters.

To avoid the artifacts and deterioration caused by deterministic vocoders, WaveNet, which directly models the raw waveform of the audio signal in a non-linear auto-regressive way, has been proposed and dramatically improves the quality of synthetic speech [12], [13]. The original WaveNet model [12] used linguistic features as well as the fundamental frequency (F0) as local conditions. Later, the WaveNet model was used as an alternative to the deterministic vocoders in many studies [14], [15] by conditioning it on acoustic features such as cepstrum, F0, or spectrograms only [14], and results have shown that the sound quality of the WaveNet vocoder outperformed deterministic vocoders and phase recovery algorithms [16].

However, it is also reported that the samples generated from WaveNet occasionally become unstable and generate collapsed speech, especially when less accurately predicted acoustic features are used as the local condition parameters [17]. This would be more critical for the case of multi-speaker acoustic modeling where the same network is used for modeling multiple speakers at the same time, as the prediction accuracy of the multi-speaker model would be worse than well-trained speaker-dependent models.

B. MULTI-SPEAKER ACOUSTIC MODELING

Although deep learning-based methods have significantly advanced the performance of statistical parametric speech synthesis (SPSS), it still suffers from the necessity of a large amount of speech recordings of one speaker to train a high-quality acoustic model. Ideally, a speech synthesis system should be able to generate an arbitrary speaker's voice with a minimum of training data. Multi-speaker speech synthesis is one of the most effective approaches to train such a high-quality acoustic model with a limited amount of speech data of each speaker. Using multiple speakers' data at the same time, we can improve the quality of synthesized speech and can also change the speaker characteristics of synthetic speech flexibly.

Using DNN-based acoustic models as a basis, Fan *et al.* [18] proposed multi-speaker speech synthesis using shared speaker-independent layers as well as a speaker-dependent output layer. They showed that the speaker-dependent output layer can be estimated from a target speaker's data only and that the shared hidden layers can improve the quality of synthesized speech of individual speakers. Wu *et al.* [19] suggested using i-vectors for modeling multiple speakers and controlling the speaker identity of synthetic speech. Hojo *et al.* [20] proposed using speaker codes based on a one-hot vector for modeling multiple speakers and extending the code and associated weights at an input layer for adapting it to unseen speakers. Luong *et al.* [21] proposed estimating code vectors for new speakers via back-propagation and experimented with manually manipulating input code vectors to alter the gender and/or age characteristics of the synthesized speech. Similar work has been extended to Long short-term memory (LSTM)-based acoustic models. Zhao *et al.* [22] examined various speaker identity representations for multi-speaker synthesis and showed that multi-speaker systems trained with less of the target speaker's data can even outperform single speaker speech synthesis, which uses a larger amount of the target speaker's data. Li and Zen [23] investigated multi-speaker modeling with speech data in different languages.

Multi-speaker speech synthesis has also been investigated in the recent WaveNet-based approaches and in end-to-end approaches. Hayashi *et al.* [24] attempted WaveNet vocoder-based multi-speaker synthesis using four speakers from the CMU arctic corpus [25]. VoiceLoop [26] involves the data of 109 speakers for acoustic model training, and Deep Voice 3 [27] trained a multi-speaker model using over 2,000 speakers. Wang *et al.* [28] proposed a bank of style embedding vectors and used it for modeling multiple TED speakers. As we can see, very active research on multi-speaker modeling has been carried out.

C. CONTRIBUTION OF THIS PAPER

In this paper, we propose frameworks that incorporate either the conditional GAN [29] or its variant, Wasserstein GAN with gradient penalty (WGAN-GP) [30], into RNN-based speech synthesis systems using the WaveNet vocoder for the purpose of reducing the mismatched characteristics between natural and generated acoustic features and for making the outputs of the WaveNet vocoder better and more stable. We evaluate the proposed frameworks using a multi-speaker modeling task. The generator of GAN is conditioned on both linguistic features and speaker code, and the discriminator aiming at distinguishing the real and predicted mel-spectrograms is also conditioned on speaker information. The WaveNet vocoder is conditioned on both mel-spectrogram and speaker codes, as well.

In addition, we extend the GAN frameworks and define a new objective function using the weighted sum of three kinds of losses: conventional MSE loss, adversarial loss, and discretized mixture logistic loss [31] obtained through the well-trained WaveNet vocoder. Since the third loss will let neural networks consider losses not only in the acoustic feature domain (such as mel-spectrogram) but also in the final

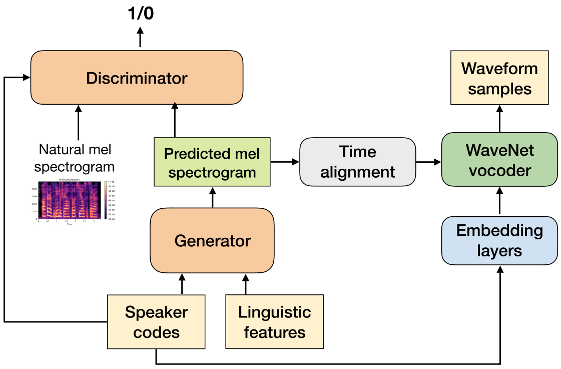

FIGURE 1. Proposed GAN-trained multi-speaker speech synthesis framework using a WaveNet vocoder.

waveform, we hypothesize that it will improve the quality of synthetic speech. In our experiment, simple recurrent units (SRUs) [32] are utilized as basic components since they can be trained faster than the LSTM-based RNN architecture while maintaining a performance as good as or even better than LSTM-RNN.

In Section II, we present the proposed framework for multi-speaker speech synthesis and describe the elements of the structure of the proposed model including SRU, GAN, and WaveNet. The details of the training algorithms are given in Section III. Section IV describes experimental conditions and Section V discusses the results. We conclude in Section VI with a brief summary and mention of future work.

II. MULTI-SPEAKER SPEECH SYNTHESIS INCORPORATING GAN AND WAVENET VOCODER

In this section, we introduce the proposed speech synthesis framework for multi-speaker modeling.

In the conventional SPSS structure, acoustic models and vocoders usually work independently: the acoustic models are trained without any consideration of the speech vocoding process, and vice versa. It was the same in the first versions of end-to-end structures such as Deep Voice [33], where vocoders were usually designed or trained on natural acoustic parameters without considering the divergence between predicted and natural acoustic parameters. This may lead to obvious and unpredictable distortion of the synthesized speech. To alleviate this problem, Tacotron 2 [15] utilized mel-spectrograms predicted beforehand to train the WaveNet vocoder instead of natural mel-spectrograms. Experimental results showed that such a strategy may outperform those that use natural parameters and may achieve a higher evaluation.

A novel idea of the present work is to *minimize the acoustic mismatch of predicted and natural parameters by conducting acoustic model training based on GAN, which also considers vocoder loss*. The proposed multi-speaker speech synthesis framework is shown in Fig. 1. In this framework, a generator part of GAN is adopted to predict acoustic features from linguistic features, and both the generator and discriminator are conditioned on speaker codes and trained with multiple speakers' data. Similar to Tacotron 2, the mel-spectrogram, a low-dimensional representation of the linear-frequency spectrogram, which contains both spectral envelop and harmonics information, is selected as the output of the generator and used to bridge the acoustic model and the WaveNet vocoder. Mel-scale acoustic features have overwhelming advantages in terms of emphasizing the details of audio, especially for lower frequencies, since they are more critical to phonetic information and hence to speech intelligibility in general.

The input of the discriminator is either natural or generated acoustic feature samples. The discriminator is trained to distinguish natural samples from generated ones. Speaker codes are also attached to both the input and hidden layers of the discriminator in order to make a better distinction between different speakers. The discriminator is used to compute the adversarial (ADV) loss, which is expected to alleviate the over-smoothing problem.

In addition to the adversarial (ADV) loss from the discriminator, the average discretized-mixture-of-logistics (DML) loss of a well-trained WaveNet model is also back-propagated to the generator of GAN. This loss corresponds to distortion between natural and generated waveform samples. We hypothesize that this increases the consistency of acoustic features predicted by the acoustic model and utilized in the vocoder since the acoustic model is updated on the basis of gradients directly computed by the pre-trained WaveNet vocoder.

In brief, it is expected that the weighted sum of the conventional MSE loss, the adversarial loss of the discriminator, and the DML from the WaveNet vocoder will improve the accuracy of the predicted acoustic parameters and thus enhance synthesized speech quality. What sets this work apart from other related works is that WaveNet is involved in the process of acoustic modeling training. After extracting acoustic features from a training corpus, the WaveNet vocoder is first trained by utilizing natural mel-spectrograms, and then the trained WaveNet model is directly referenced for acoustic model optimization.

In the following subsections, we review the three major components of the proposed framework, namely, the SRU architecture and the GAN and WaveNet models.

A. SRU

For the sake of modeling accuracy as well as time efficiency, we choose SRU [32] as the basic architecture of the acoustic modeling. The SRU architecture was originally designed to speed up the training process of RNN. By utilizing both skip and highway connections, SRU is capable of outperforming RNN, especially on very deep networks. Compared with other recurrent architectures (e.g., LSTM and gated recurrent units), the basic form of SRU includes only a single forget gate f_t to alleviate vanishing and exploding gradient problems instead of using many different gates to control the information flow. In SRU, the forget gate is used to modulate the internal state c_t, which is then used to compute

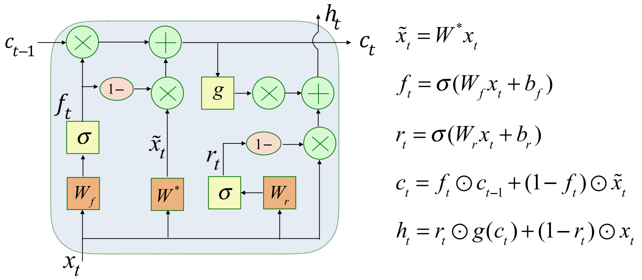
c_{t-1}
h_t
c_t
f_t
1−
g
σ
$\tilde{x}_t$
r_t
1−
W_f
W^*
σ
W_r
x_t
$\tilde{x}_t = W^* x_t$
$f_t = \sigma(W_f x_t + b_f)$
$r_t = \sigma(W_r x_t + b_r)$
$c_t = f_t \odot c_{t-1} + (1 - f_t) \odot \tilde{x}_t$
$h_t = r_t \odot g(c_t) + (1 - r_t) \odot x_t$

FIGURE 2. Details of the SRU cell. $\sigma(\cdot)$ and $g(\cdot)$ represent sigmoid and ReLU activation functions, respectively.

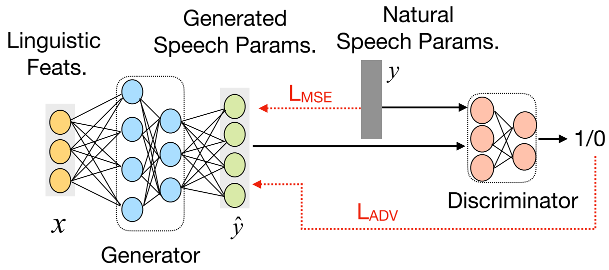

FIGURE 3. GAN-based training of TTS acoustic model. L_{ADV} indicates adversarial loss and L_{MSE} indicates L2 loss.

the output state h_t. Unlike existing RNN architectures that use the previous output state in the recurrence computation, SRU completely drops the connection between the gating computations and the previous states, and this makes SRU computationally efficient and allows us to use parallelization. The complete architecture of SRU is shown in Fig. 2. The reset gate r_t is computed similar to the forget gate f_t and is used to compute the output state h_t, which performs as a combination of the internal state $g(c_t)$ and the input x_t. $g(\cdot)$ represents a Rectified Linear Unit (ReLU) activation function and $\sigma(\cdot)$ is a sigmoid function.

B. GENERATIVE ADVERSARIAL NETWORK

GANs have achieved great success in modeling the distributions of complex data and the predictions of realistic data in many applications. They have also proven beneficial for speaker-dependent speech synthesis [10].

Fig. 3 shows the GAN-based training of acoustic models for TTS systems. The GAN training involves a pair of networks: a generator G aims to produce vivid feature samples that deceive a discriminator D, and the discriminator aims to estimate the probability that a sample y came from the real data set distribution $\mathbb{P}_r$ rather than a generator distribution $\mathbb{P}_g$. For speech synthesis from text, the generator is conditioned on linguistic vectors $x \sim \mathbb{P}_x$. The generator and discriminator are trained like a two-player min-max game objective function, as

$$\min_G \max_D \mathop{\mathbb{E}}_{y \sim \mathbb{P}_r} \left[\log D(y)\right] + \mathop{\mathbb{E}}_{x \sim \mathbb{P}_x} \left[\log\left(1 - D(G(x))\right)\right] \quad (1)$$

This objective function is not easy to optimize. To improve the stability of model training, Wasserstein GAN (WGAN), which minimizes a different distribution divergence called Earth-Mover or Wasserstein-1 distance, has been proposed and achieved a better performance than original GAN in terms of convergence, especially in image processing [34]. The optimization criteria for WGAN is equal to

$$\min_G \max_D \mathop{\mathbb{E}}_{y \sim \mathbb{P}_r} [D(y)] - \mathop{\mathbb{E}}_{x \sim \mathbb{P}_x} [D(G(x))] \quad (2)$$

During the training of WGAN, the updated model parameters of discriminator are clipped into a compact space $[-c, c]$ to enforce a Lipschitz constraint on D. However, the weight clipping may lead to either vanishing or exploding gradients if the clipping threshold c is not carefully tuned, and the resulting discriminator may have a pathological value surface even when optimization performs smoothly [30]. To address this problem, Gulrajani *et al.* [30] proposed penalizing the norm of the gradient deduced from a discriminator with respect to its input. The new objective for WGAN with gradient penalty (WGAN-GP) is shown as follows:

$$\min_G \max_D \mathop{\mathbb{E}}_{y \sim \mathbb{P}_r} [D(y)] - \mathop{\mathbb{E}}_{x \sim \mathbb{P}_x} [D(G(x))] + \lambda \mathop{\mathbb{E}}_{\tilde{y} \sim \mathbb{P}_{\tilde{y}}} [(\left\| \nabla_{\tilde{y}} D(\tilde{y}) \right\|_2 - 1)^2] \quad (3)$$

where λ is a gradient penalty coefficient and $\tilde{y}$ represents samples that are linearly interpolated by the real data y and the fake data generated from the generator $G(x)$:

$$\tilde{y} = \epsilon y + (1 - \epsilon) G(x) \quad (4)$$

where ϵ is a random number that obeys distribution $U[0, 1]$.

The loss function of the generator is also expanded on the basis of the least square errors of y as:

$$L_G(y, \hat{y}) = L_{MSE}(y, \hat{y}) + \gamma_D L_{ADV}(\hat{y}) \quad (5)$$

where $L_{ADV}(\hat{y})$ is the adversarial loss and γ_D controls the weight of the adversarial loss. When $\gamma_D = 0$, the loss function is equivalent to the conventional MSE criteria. In original GAN, $L_{ADV}(\hat{y})$ equals $\mathbb{E}[\log(1 - D(G(x)))]$. In WGAN-GP, $L_{ADV}(\hat{y})$ can be regarded as $-\mathbb{E}[D(G(x))]$.

C. WaveNet

WaveNet is a deep auto-regressive and generative model that models a joint distribution of sequential data as a product of conditional distributions, as

$$p(s) = \prod_t p(s_t | s_{<t}, \theta) \quad (6)$$

where s_t is a variable of s at time t and θ denotes model parameters. The conditional distributions are usually modeled with a neural network that receives all past variables $s_{<t}$ as input and outputs a distribution over possible s_t. The neural network consists of stacked dilated causal convolution layers [12], and each causal convolutional layer can process its input in parallel, making these architectures very fast to train compared to RNNs. It typically uses gated activation functions [35] along with two conditions, global and local, which is another important concept in WaveNet.

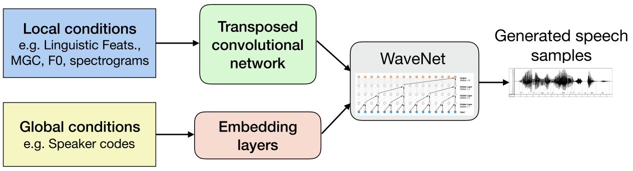

FIGURE 4. Local condition and global condition used in a WaveNet model.

The difference between the two conditions is shown in Fig. 4. The global condition focuses on conditional vectors irrelevant to time, e.g. a speaker embedding in a TTS model, while the local condition deals with time-series input conditions, such as linguistic and acoustic features. The basic activation function with global conditioning is

$$h_i = \sigma(W_{g,i} * s_i + V_{g,i}c) \odot \tanh(W_{f,i} * s_i + V_{f,i}c) \quad (7)$$

where $*$ denotes a convolution operator and $\odot$ denotes an element-wise multiplication operator. σ is a logistic sigmoid function. c represents a global condition. i is the layer index. f and g denote filter and gate, respectively. W and V are learnable weights. For a case where c denotes the local condition (such as mel-spectrogram), the matrix products $V_{g,i}c$ and $V_{f,i}c$ are replaced by convolutions $V_{g,i} * c$ and $V_{f,i} * c$, respectively.

Oord *et al.* [12] take both linguistic and acoustic features such as F0 as the local conditions. In other studies [14], [15], [27], only acoustic features are used as the local conditions, and the WaveNet model tends to perform as a neural vocoder. In the proposed framework, WaveNet is used as a multi-speaker neural vocoder. It is locally conditioned on mel-spectrograms and globally conditioned on speaker embeddings.

D. DML LOSS

In [12], speech waveform samples were quantized and the cross entropy loss was used for modeling categorical distribution, but if we use additional quantization bits (to reduce the quantization noise), the cost of computations may be exponentially increased. Using discretized mixture of logistics (DML) distribution loss [31] could save memory and improve training efficiency because it just needs to predict parameters for each mixture component instead of all bits. For example, modeling 16-bit quantized bits always requires the training of a 65,536-way categorical distribution, while only ten mixtures of logistic distributions are sufficient to model 16-bit audio samples empirically.

DML distribution assumes that each sample point s is composed of a mixture of continuous uni-variate distributions υ, and each component υ_i obeys logistic distribution, as

$$\upsilon = \sum_{i=1}^{K} \pi_i \upsilon_i, \quad \text{where } \upsilon_i \sim \text{logistic}(\mu_i, \phi_i) \quad (8)$$

where π_i is the mixture weight of component i that satisfies $\sum_{i=1}^{K} \pi_i = 1$. μ is the mean and ϕ is a scale parameter proportional to the standard deviation. The probability on the observed discretized audio sample s excepting the edge cases (e.g., 0 and 65,535 for 16-bit sampling) would be

$$P(s|\pi,\mu,\phi) = \sum_{i=1}^{K} \pi_i \left[\sigma(\frac{s+1-\mu_i}{\phi_i \zeta}) - \sigma(\frac{s-1-\mu_i}{\phi_i \zeta}) \right] \quad (9)$$

$\sigma(\cdot)$ is the logistic sigmoid function. ζ denotes the number of sampling classes and $\zeta = 256$ for 8-bit and 65536 for 16-bit sampling. For the edge case of 0, replace $s-1$ with $-\infty$, and for 255 or 65535, replace $s+1$ with $+\infty$. Finally, the WaveNet model aims at maximizing the average log likelihood of P:

$$L_{DML} = \max_{W} \mathbb{E}[\log P(s|\hat{\pi},\hat{\mu},\hat{\phi})] \quad (10)$$

where $\hat{\pi}, \hat{\mu}, \hat{\phi}$ are predicted mixture component parameters.

III. TRAINING ALGORITHM

A. TRAINING ALGORITHM FOR THE PROPOSED ACOUSTIC MODEL

The overall loss function for training the proposed acoustic model that predicts mel-spectrogram can be written as

$$L_G(y,\hat{y}) = L_{MSE}(y,\hat{y}) + \gamma_D L_{ADV}(\hat{y}) + \gamma_W L_{DML}(y,\hat{y}). \quad (11)$$

In addition to the general MSE loss L_{MSE} and adversarial loss L_{ADV}, the DML loss L_{DML} generated by a well-trained WaveNet model is utilized for updating the model parameters of the generator. Utilizing the DML loss with the generator would integrate the divergence of synthesized speech samples into the acoustic parametric training process. Therefore, the proposed loss function minimizes not only the parametric error of the mel-spectrogram but also the fidelity disparity between predicted and natural audios. γ_W is a hyper-parameter that denotes the weight of L_{DML}. When $\gamma_W = 0$, the loss function is equivalent to the conventional GAN training. Model parameters of the generator θ_G are updated by using the stochastic gradient calculated from $L_G(y,\hat{y})$. Fig. 5 shows the procedure for computing the proposed loss function.

The details of the acoustic model training algorithm are given in Algorithm 1. In the first step, the generator is trained with the MSE criterion for a few epochs. Then, the generator and discriminator are optimized in an iterative way, where one module is being updated while the model parameters of another are fixed. In the final step, the loss of WaveNet $L_{DML}(y,\hat{y})$ is enrolled in the training criterion of the generator. Before this step, the WaveNet vocoder needs to be trained in advance and the optimum model parameters θ_W should be saved. Note that the DML loss does not join the optimization process of the discriminator, and the parameters of the WaveNet model are always kept fixed. In other words,

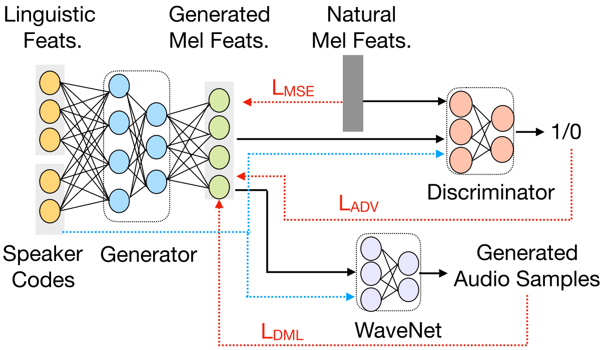

FIGURE 5. Loss functions and gradients for updating acoustic models in the proposed method. Note that neither the model parameters of WaveNet nor the discriminator are updated in this step.

although θ_D and θ_W are included in calculating $L_G(y, \hat{y})$, θ_D is not updated by the back-propagation of L_G in the final step, and neither is θ_W. The WaveNet model is used as a measurement that reflects the divergence between speech samples. In the WGAN-GP-based case, θ_D is first optimized according to Eq. (3) and then θ_G is optimized according to Eq. (11).

B. TIME RESOLUTION ADJUSTMENT

During the training of the multi-speaker acoustic model, there are two instances where we need to pay attention to *time resolution* problems. The first is when the mel-spectrograms are input to the WaveNet vocoder. The other is when DML loss is applied for generator optimization.

When acoustic features are transformed into speech samples, conventional parametric vocoders always use interpolation inside frames to recover the audio sampling points. Since different sampling points may share the same acoustic features, in existing studies related to WaveNet, several approaches have been proposed to align the input conditional features with the speech samples.

When acoustic features are transformed into speech samples in the WaveNet vocoder, Oord *et al.* [12] used a trainable transposed convolutional network to upsample the time resolution of the conditional acoustic features. Deep Voice 2 applied a stack of bidirectional quasi-recurrent neural networks and Tamamori *et al.* [14] simply duplicated the conditional acoustic feature vector of each frame. In our work, we use trainable transposed convolutional layers to align mel-spectrograms and speech samples for the WaveNet vocoder as in [12].

When the well-trained WaveNet vocoder is used for the proposed generator optimization, it would be time-consuming to calculate the DML loss along all the waveform audio samples within the same frame. As shown in Fig. 6, in order to improve computational efficiency, we randomly select a part of the waveform audio points within each frame and back-propagate their averaged DML loss to the generator for acoustic model optimization.

Algorithm 1 Training Algorithm for Acoustic Modeling

```
Require:
  $x$ := linguistic features; $c$ := speaker code; $y$ := mel-spectrogram;
  Initial generator parameter $\theta_G$ and initial discriminator parameter $\theta_D$;
  A well-trained WaveNet model $W$ and $\theta_W$ is fixed;
  batch size $m$, learning rate $\eta$, the gradient penalty coefficient $\lambda$, weight for adversarial loss $r_D$, weight for DML loss $r_W$, generator warming up iterations $n1$, basic adversarial training iterations $n2$, number of total iterations $n3$.
Begin step 1: warming up generator
  for epoch = $1, \cdots, n1$ do
      for training data in $(x, c, y)$ do
          generate $\hat{y}$ from the generator
              $\hat{y} = G(x, c)$
          update $\theta_G$ using MSE criterion:
              $\theta_G \leftarrow \theta_G - \eta_G \nabla_{\theta_G} L_{MSE}(y, \hat{y})$
      end for
  end for
End
Begin step2: adversarial training
  for epoch = $n1, \cdots, n2$ do
      for training data in $(x, c, y)$ do
          for $i = 1, \cdots, m$ do
              $\hat{y} = G(x, c)$
              $\tilde{y} = \epsilon y + (1 - \epsilon)\hat{y}, \quad \epsilon \in U[0, 1]$
              $L_D^{(i)} = D(\hat{y}) - D(y) + \lambda(\left\|\nabla_{\tilde{y}} D(\tilde{y})\right\|_2 - 1)^2$
          end for
          update $\theta_D$ while fixing $\theta_G$:
              $\theta_D \leftarrow \theta_D - \eta_D \nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} L_D^{(i)}$
          update $\theta_G$ using both MSE and adversarial criterion:
              $L_{ADV} = \frac{1}{m} \sum_{i=1}^{m} D(G(x, c))$
              $\theta_G \leftarrow \theta_G - \eta_G \nabla_{\theta_G} (L_{MSE}(y, \hat{y}) + \gamma_D L_{ADV})$
      end for
  end for
End
Begin step 3: fine tuning the generator by utilizing WaveNet loss.
  for epoch = $n2, \cdots, n3$ do
      for training data in $(x, c, y)$ do
          generate $\hat{y}$ and update $\theta_D$ following step 2.
          upsampling $\hat{y}$.
          generate $\hat{s}$ from the well-trained WaveNet model:
              $\hat{s} = W(\hat{y}, c)$
          update $\theta_G$ with DML loss from WaveNet:
              $\theta_G \leftarrow \theta_G - \eta_G \nabla_{\theta_G} (L_{MSE}(y, \hat{y}) + \gamma_D L_{ADV} + \gamma_W L_{DML}(s, \hat{s}))$
      end for
  end for
End
```

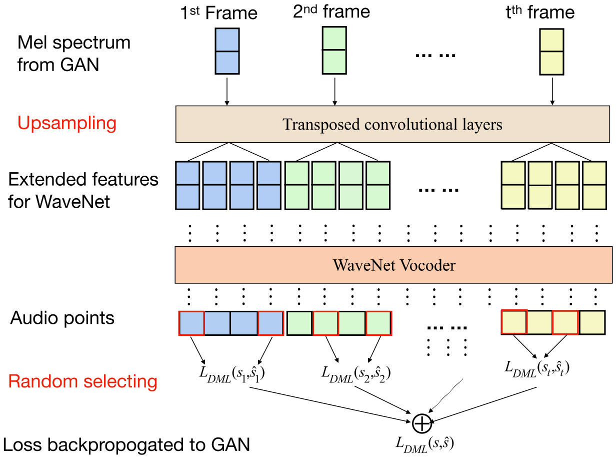

FIGURE 6. Time resolution adjustment of conditional acoustic features. One frame includes four waveform audio points. Transposed convolutional layers are used to upsample the conditional acoustic features. The DML loss was computed using randomly selected waveform audio points within each frame.

IV. EXPERIMENTAL SETUP

We used six speakers (awb, bdl, clb, ksp, rms, and slt) from the CMU-ARCTIC database for multi-speaker training. Two speakers (clb and slt) are female and the others are male. For each speaker, 1000 utterances were used for training. Their speech waveforms have a sampling frequency of 16 kHz and a 16-bit PCM format. The six speakers read out the same set of utterances. Linguistic labels were generated by Festival TTS and consist of 376-dimensional binary vectors and 5-dimensional duration information. The linguistic features are normalized by the min-max rule. Speaker codes consist of seven dimensions, where six dimensions represent speaker identity difference in one-hot format and the other dimension denotes gender. The speaker codes are input to the first layer of both the generator and discriminator as auxiliary features. For the WaveNet vocoder, the speaker codes are first input to a fixed-size embedding layer and then converted to an input format compatible with WaveNet. None of the utterances in the testing set appear in either the training or development sets.

As acoustic features, 80-dimensional static mel-spectrograms are adopted in our experiment. To compute mel-spectrograms, we first perform a short-time Fourier transform (STFT) on audios using a 15-ms frame size, 5-ms frame shift, and a Hann window function. Then we transform the STFT magnitude spectrum to the mel scale using an 80-channel mel-filterbank that ranges from 125 Hz to 7.6 kHz, followed by log dynamic range compression. Prior to the log compression, the filterbank output magnitudes are clipped to a minimum value of 0.01 in order to limit the dynamic range in the logarithmic domain. The mel-spectrograms are then normalized to have zero-mean unit variance.

We used six bidirectional SRU layers for acoustic modeling and three feed-forward layers for the discriminator. In the generator, each layer has 512 hidden nodes, and in the discriminator, each layer has 128 hidden nodes. The ReLU activation function is utilized in the SRU cell. A stochastic gradient descent (SGD) optimizer was used as the optimizer for both the generator and discriminator. Learning rate was initialized to 0.01 for the generator and 0.001 for discriminator along with exponential decays corresponding to the number of training epochs.

To implement the WaveNet model, we referenced [36] and adopted a modified version of the WaveNet architecture. Instead of predicting discretized buckets with a softmax layer, we followed Tacotron 2 and Parallel WaveNet and used a 10-component mixture of logistic distributions to generate 16-bit samples at 16 kHz. To compute the logistic mixture distribution, the WaveNet stack output was passed through a ReLU activation, followed by a linear projection to predict parameters (mean, log scale, mixture weight) for each mixture component. We adopted 24 dilated convolution layers grouped into four dilation cycles. The dilation rate of the k-th layer was set to $2^{k \pmod 6}$, where $k \in [0, 1, 2 \cdots 23]$. Finally, 24 residual blocks were connected. The number of channels of (dilated) causal convolution and 1 × 1 convolution in the residual block were set to 512. The number of 1 × 1 convolution channel between skip-connection and output layer was set to 256. We used three transposed convolutional layers for up-sampling. The Adam algorithm [37] was used for the optimization, and its learning rate was initialized to 0.001 and scheduled carefully with a scheme similar to [38]. Other parameters in the Adam optimizer were set as $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1.0e^{-8}$. We also maintained an exponentially weighted moving average of the network parameters over update steps with a decay of 0.9999. A GeForce GTX 1080 was used for training. It took about a week to train a high-quality multi-speaker WaveNet vocoder and eight minutes to synthesize ten seconds of speech. When updating the generator using the DML loss back-propagated from the trained WaveNet Vocoder, we randomly chose half of the sampling points in each frame to efficiently calculate the DML loss. γ_D was set equal to $E(L_{MGE})/E(L_{ADV})$, and $E(\cdot)$ represented expectation value. γ_W was fixed as 0.0001.

V. EXPERIMENTAL EVALUATION

We compared the performance of the following configurations based on a listening test:

1) Baseline: Acoustic model trained using $L_{MSE}(y, \hat{y})$ as a criterion.
2) GAN: Acoustic model trained using $L_{MSE}(y, \hat{y}) + \gamma_D L_{ADV}(\hat{y})$ as a criterion.
3) GAN^{W}: Acoustic model trained using $L_{MSE}(y, \hat{y}) + \gamma_D L_{ADV}(\hat{y}) + \gamma_W L_{DML}(y, \hat{y})$ as a criterion.
4) WGAN-GP: Acoustic modeling trained using $L_{MSE}(y, \hat{y}) + \gamma_D L_{ADV}(\hat{y})$ as a criterion. WGAN-GP was also used.
5) WGAN-GP^{W}: Acoustic model trained using $L_{MSE}(y, \hat{y}) + \gamma_D L_{ADV}(\hat{y}) + \gamma_W L_{DML}(y, \hat{y})$ as a criterion. WGAN-GP was also used.

TABLE 1. Statistical significance analysis using *t*-tests with Holm-Bonferroni correction in terms of quality judgment.

System	Baseline	GAN	GAN$^{\text{W}}$	WGAN-GP	WGAN-GP$^{\text{W}}$	AbS
GAN	<2e-16	-	-	-	-	-
GAN$^{\text{W}}$	< 2e-16	0.05206	-	-	-	-
WGAN-GP	< 2e-16	0.00028	0.19850	-	-	-
WGAN-GP$^{\text{W}}$	< 2e-16	1.2e-06	0.01916	0.24092	-	-
AbS	< 2e-16	< 2e-16	< 2e-16	<2e-16	<2e-16	-
Natural	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16

6) Analysis by Synthesis (AbS): Synthetic speech generated by a WaveNet vocoder using ground-truth mel-spectrograms.
7) Natural: Natural speech.

Note that systems 1 to 5 are TTS systems and use SRU as basic architectures for acoustic models, as described earlier. Also note that all the above TTS systems and analysis by synthesis use the same WaveNet vocoder. The differences are how the local condition parameters of the WaveNet vocoder, that is, mel-spectrogram, are predicted.

A. EVALUATION METHODOLOGY

For the listening test, we selected 20 utterances from the testing set of each speaker and generated sets of synthetic speech corresponding to the above experimental systems.[1] Each experimental system had 20 utterances, so 20 utterances × 6speakers × 7 = 840 samples that needed to be evaluated in total. Crowdsourced perceptual evaluation was carried out to evaluate naturalness as well as speaker similarity of generated speech. In the crowdsourcing test, we evaluated each sample ten times to alleviate personal bias. The testing samples were divided into different evaluation sets. Each set consisted of three utterances generated by seven different systems. Therefore, there were 42 utterances to be evaluated in each set: 21 for naturalness and 21 for similarity. We then collected 400 sets to cover all 840 samples (400 = 840 × 10/21). This guarantees at least 40 unique listeners, since we limited the maximum number of sets per crowdsourced participant to ten. The actual number of listeners who participated in our test was 42.

To evaluate naturalness, listeners were asked to ignore the meaning of the sentence and concentrate only on rating how natural the speech sounded on a five-point scale:

1) completely unnatural
2) mostly unnatural
3) equally natural and unnatural
4) mostly natural
5) completely natural

For speaker similarity, listeners were asked to ignore the meaning of the sentence and concentrate only on rating the speaker identity. Synthetic speech samples and the corresponding natural sound were presented in pairs at every turn and listeners were asked to judge whether the two samples were from the same or different speaker(s). The scale for speaker similarity was judged on a four-point scale:

1) same speaker, absolutely sure
2) same speaker, not sure
3) different speaker, not sure
4) different speaker, absolutely sure

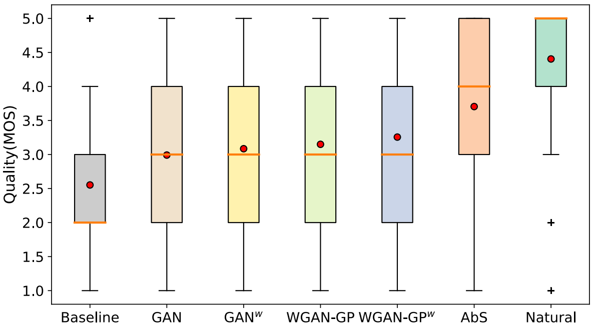

FIGURE 7. Box plots on naturalness evaluation results. Red dots represent the mean of each group averaged across all speakers.

B. EVALUATION RESULTS AND ANALYSIS

Fig. 7 shows the box plots for the naturalness evaluation results averaged across all speakers. Table 1 shows statistical significance. From these, we can see that four GAN-based experimental groups (GAN, GAN$^{\text{W}}$, WGAN-GP, WGAN-GP$^{\text{W}}$) outperform the baseline significantly. Upper quartiles and mean opinion scores of the four GAN-based groups are much higher than those of the baseline, although their lower quartiles are quite similar to the baseline. Note that all the systems (apart from natural speech) use the same WaveNet vocoder. Hence, this also indicates that the quality of WaveNet synthetic speech is affected by the local condition parameters and that the ones predicted by the GAN-based acoustic models sound more natural than those by the baseline. We also see that WGAN-GP systems (WGAN-GP, WGAN-GP$^{\text{W}}$) are better than the original GAN system. The use of DML loss alone did not bring statistically significant improvements, but it obviously reduced p-values (see Table 1), and hence a combination of WGAN-GP and the DML loss resulted in the highest scores among the TTS systems and was significantly better than GAN and GAN$^{\text{W}}$ ($p < 0.05$).

Compared with the natural speech and AbS, all TTS methods have obvious gaps. There is also a gap between the

[1] Audio samples of generated synthetic speech are available at https://nii-yamagishilab.github.io/TTS-GAN-WN-MultiSpeaker/

TABLE 2. Statistical significance analysis using *t*-tests with Holm-Bonferroni correction in terms of speaker similarity judgment.

Systems	Baseline	GAN	GAN^W	WGAN-GP	WGAN-GP^W	AbS
GAN	0.43810	-	-	-	-	-
GAN^W	0.28401	1.00000	-	-	-	-
WGAN-GP	0.00426	0.31593	0.47565	-	-	-
WGAN-GP^W	0.00044	0.11438	0.28401	1.00000	-	-
AbS	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	-
Natural	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	<4.2e-06

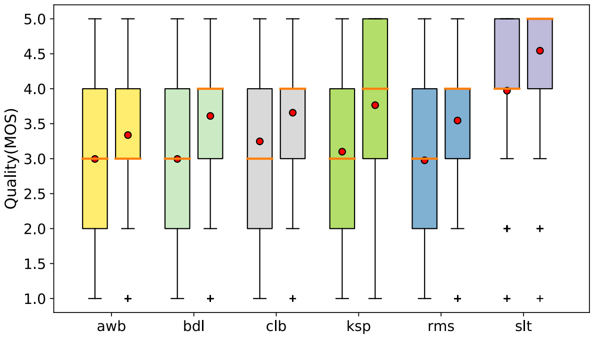

FIGURE 8. Box plots of the MOS scores of six speakers. Left: WGAN-GP^W system. Right: AbS system.

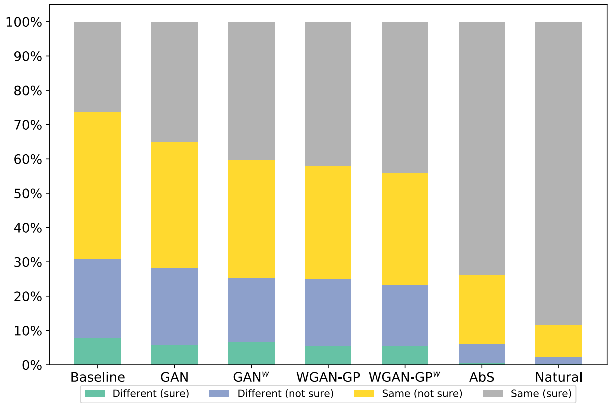

FIGURE 9. Similarity results averaged across all speakers.

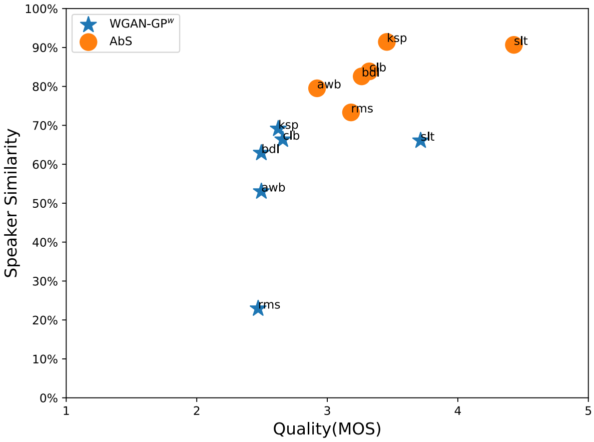

FIGURE 10. Scatter plot matching naturalness and similarity scores for each speaker in system WGAN-GP^W and AbS. The similarity score is defined as the added percentage of 'same (not sure)' and 'same (sure)' scores.

AbS samples and natural speech. This indicates that our multi-speaker TTS systems do not sound as good as natural speech yet, and the multi-speaker WaveNet vocoder itself does not sound as good as natural speech either, even if it uses the ground-truth mel-spectrogram. In other words, both the neural vocoder and the acoustic model have room for further improvement.

Through our experiments, we found that the quality of our synthetic speech varied speaker by speaker. Fig.8 shows box plots of the MOS scores of the best WGAN-GP^W system and the AbS system of the six speakers. The left box plot shows the results of the WGAN-GP^W system and the right box plot shows those of the AbS system for each speaker. Interestingly, the quality of synthetic speech varied speaker by speaker, and there is a very large gap between speaker SLT and the other speakers. This implies that we need a more generalized model that can handle multiple speakers better and can reproduce the differences between speakers more precisely.

The similarity evaluation results are shown in Fig.9. The WGAN-based systems outperform the baseline, and we can clearly see that the portions of "Same" (yellow and gray) have been increased. The proposed systems using a combination of WGAN-GP and DML loss achieved more apparent preference in terms of "Same, absolutely sure". Likewise in the quality evaluation, we can see a gap between TTS systems and WaveNet analysis-by-synthesis systems as well as between WaveNet analysis-by-synthesis systems and natural speech. The t-test results for similarity are shown in Table 2. Compared with Table 1, Table 2 shows less significant differences overall. This suggests that although the proposed GAN-based groups behave obviously better than the baseline for the quality evaluation, only WGAN-GP and WGAN-GP^W show significantly better performance than baseline in terms of similarity.

Fig. 10 shows a scatter plot matching naturalness and similarity scores of the best WGAN-GP^W system and AbS system of six speakers. Interestingly, the speaker similarity scores also significantly varied speaker by speaker, and speaker RMS had a very low speaker similarity score. Our next step is to investigate why a few speakers had lower speaker similarity.

VI. CONCLUSION

This paper investigated how we should train the acoustic model that predicts the local condition parameters to be used by neural vocoders. Specifically, we looked into conditional GANs or WGAN-GP to reduce the mismatched characteristics between natural and generated acoustic features. We also extended the GAN frameworks and used the discretized mixture logistic loss of a well-trained WaveNet along with mean squared error and adversarial losses as parts of the objective functions. These new objective functions were evaluated in multi-speaker speech synthesis that uses the WaveNet vocoder. Experimental results show that acoustic models trained with the WGAN-GP framework using back-propagated DML loss achieved the highest subjective evaluation scores in terms of both quality and speaker similarity.

Our future work will investigate why some speakers have lower quality of synthetic speech or lower similarity. We will also perform larger scale experiments using more speakers.

ACKNOWLEDGMENT

The authors thank Mr. Lauri Juvela from Aalto University, Finland for his suggestions on WGAN-GP.

REFERENCES

[1] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 4, Apr. 2007, pp. IV-1229–IV-1232.

[2] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 7962–7966.

[3] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3844–3848.

[4] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1964–1968.

[5] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4470–4474.

[6] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 184–194, Apr. 2014.

[7] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4230–4234.

[8] F. Espic, C. Valentini-Botinhao, and S. King, "Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis," in *Proc. Interspeech*, Stochohlm, Sweden, 2017, pp. 1–5.

[9] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4900–4904.

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

[11] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[12] A. van den Oord *et al.* (2016). "WaveNet: A generative model for raw audio." [Online]. Available: https://arxiv.org/abs/1609.03499

[13] A. van den Oord *et al.* (2017). "Parallel WaveNet: Fast high-fidelity speech synthesis." [Online]. Available: https://arxiv.org/abs/1711.10433

[14] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.

[15] J. Shen *et al.* (2017). "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions." [Online]. Available: https://arxiv.org/abs/1712.05884

[16] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi. (2018). "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis." [Online]. Available: https://arxiv.org/abs/1804.02549

[17] Y.-C. Wu, K. Kobayashi, T. Hayashi, P. L. Tobing, and T. Toda. (2018). "Collapsed speech segment detection and suppression for WaveNet vocoder." [Online]. Available: https://arxiv.org/abs/1804.11055

[18] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4475–4479.

[19] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 879–883.

[20] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proc. INTERSPEECH*, Sep. 2016, pp. 2278–2282.

[21] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4905–4909.

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

[23] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Proc. INTERSPEECH*, 2016, pp. 2468–2472.

[24] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 712–718.

[25] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Workshop Speech Synthesis*, 2004, pp. 223–224.

[26] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "VoiceLoop: Voice fitting and synthesis via a phonological loop," in *Proc. 6th Int. Conf. Learn. Represent.*, 2018, pp. 1–14.

[27] W. Ping *et al.*, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. 6th Int. Conf. Learn. Represent.*, 2018, pp. 1–16.

[28] Y. Wang *et al.* (2018). "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis." [Online]. Available: https://arxiv.org/abs/1803.09017

[29] M. Mirza and S. Osindero. (2014). "Conditional generative adversarial nets." [Online]. Available: https://arxiv.org/abs/1411.1784

[30] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5769–5779.

[31] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. (2017). "PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications." [Online]. Available: https://arxiv.org/abs/1701.05517

[32] T. Lei and Y. Zhang. (2017). "Training RNNs as fast as CNNs." [Online]. Available: https://arxiv.org/abs/1709.02755v1

[33] S. O. Arik *et al.* (2017). "Deep voice: Real-time neural text-to-speech," [Online]. Available: https://arxiv.org/abs/1702.07825

[34] M. Arjovsky, S. Chintala, and L. Bottou. (2017). "Wasserstein GAN." [Online]. Available: https://arxiv.org/abs/1701.07875

[35] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves, "Conditional image generation with PixelCNN decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4790–4798.

[36] T. Le Paine *et al.* (2016). "Fast WaveNet generation algorithm." [Online]. Available: https://arxiv.org/abs/1611.09482

[37] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[38] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

YI ZHAO received the B.E. degree in communication engineering from the Beijing University of Posts and Telecommunications, China, in 2011, and the M.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014. She is currently pursuing the Ph.D. degree with the Graduate School of Engineering, The University of Tokyo, Japan. Her current research interests include statistical machine learning and speech synthesis.

SHINJI TAKAKI received the B.E. degree in computer science, and the M.E. and Ph.D. degrees in scientific and engineering simulation from the Nagoya Institute of Technology, Nagoya, Japan, in 2009, 2011, and 2014, respectively. From 2013 to 2014, he was a Visiting Researcher with The University of Edinburgh. Since 2014, he has been a Project Researcher with the National Institute of Informatics. His research interests include statistical machine learning and speech synthesis. He is a member of the Acoustical Society of Japan and the Information Processing Society of Japan.

HIEU-THI LUONG received the B.E. degree in computer science from Vietnam National University, Vietnam, in 2014, and the M.E. degree in computer science from the Ho Chi Minh City University of Science, Vietnam, in 2016. He is currently pursuing the Ph.D. degree in statistical speech synthesis and machine learning with the National Institute of Informatics, Tokyo, Japan. He was a recipient of the Japanese Government (Monbukagakusho: MEXT) Scholarship in 2017.

JUNICHI YAMAGISHI received the Ph.D. degree from the Tokyo Institute of Technology in 2006. His Ph.D. dissertation pioneered speaker-adaptive speech synthesis. He is currently an Associate Professor with the National Institute of Informatics, Tokyo, Japan, and also a Senior Research Fellow with The Centre for Speech Technology Research, The University of Edinburgh, Edinburgh, U.K. Since 2006, he has authored or co-authored over 180 refereed papers in international journals and conferences. He is a member of the Speech and Language Technical Committee. He was a recipient of the Tejima Prize as the best Ph.D. thesis of the Tokyo Institute of Technology in 2007. He received the Itakura Prize from the Acoustic Society of Japan in 2010, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan in 2013, the Young Scientists' Prize from the Minister of Education, Science and Technology in 2014, and the JSPS Prize from Japan Society for the Promotion of Science in 2016. He was one of organizers for special sessions on Spoofing and Countermeasures for Automatic Speaker Verification at Interspeech 2013, ASVspoof Evaluation at Interspeech 2015, Voice Conversion Challenge 2016 at Interspeech 2016, and 2nd ASVspoof Evaluation at Interspeech 2017. He was an Associate Editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing and a Lead Guest Editor of the IEEE Journal of Selected Topics in Signal Processing Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification.

DAISUKE SAITO received the B.E., M.S., and Dr. Eng. degrees from The University of Tokyo, Tokyo, Japan, in 2006, 2008, and 2011, respectively. He is currently a Lecturer (Senior Assistant Professor) with the Graduate School of Engineering, The University of Tokyo. He is interested in various areas of speech engineering, including voice conversion, speech synthesis, acoustic analysis, speaker recognition, and speech recognition. From 2010 to 2011, he was a Research Fellow (DC2) with the Japan Society for the Promotion of Science. Dr. Saito is a member of the International Speech Communication Association (ISCA), the Acoustical Society of Japan (ASJ), the Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers, and the Institute of Image Information and Television Engineers. He was a recipient the ISCA Award for the Best Student Paper of INTERSPEECH 2011. He received the Awaya Award from the ASJ in 2012 and the Itakura Award from ASJ in 2014.

NOBUAKI MINEMATSU (M'08) was born in Nishinomiya, Japan in 1966. He received the Dr. Eng. degree from The University of Tokyo in 1995. From 1995 to 2000, he was a Research Associate with the Toyohashi University of Technology. Since 2000, he has been an Associate Professor with The University of Tokyo, where he has been a Full Professor since 2012. He has a wide interest in speech communication covering the areas of speech science and speech engineering; especially, he has expert knowledge on computer-aided language learning. He is a member of ISCA, SLaTE, IPA, APSIPA, IEICE, IPSJ, ASJ, and PSJ. He received paper awards from RISP, JSAI, ICIST, O-COCOSDA, and IEICE in 2005, 2007, 2011, 2014, and 2016, respectively, and the Encouragement Award from PSJ in 2014. He was a Distinguished Lecturer of APSIPA from 2015 to 2016.