INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Sequence-to-Sequence Learning for Deep Gaussian Process Based Speech Synthesis Using Self-Attention GP Layer

Taiki Nakamura, Tomoki Koriyama, Hiroshi Saruwatari

Graduate School of Information Science and Technology, The University of Tokyo, Japan
supikiti@g.ecc.u-tokyo.ac.jp, t.koriyama@ieee.org

Abstract

This paper presents a speech synthesis method based on deep Gaussian process (DGP) and sequence-to-sequence (Seq2Seq) learning toward high-quality end-to-end speech synthesis. Feed-forward and recurrent models using DGP are known to produce more natural synthetic speech than deep neural networks (DNNs) because of Bayesian learning and kernel regression. However, such DGP models consist of a pipeline architecture of independent models, acoustic and duration models, and require a high level of expertise in text processing. The proposed model is based on Seq2Seq learning, which enables a unified training of acoustic and duration models. The encoder and decoder layers are represented by Gaussian process regressions (GPRs) and the parameters are trained as a Bayesian model. We also propose a self-attention mechanism with Gaussian processes to effectively model character-level input in the encoder. The subjective evaluation results show that the proposed Seq2Seq-SA-DGP can synthesize more natural speech than DNNs with self-attention and recurrent structures. Besides, Seq2Seq-SA-DGP reduces the smoothing problems of recurrent structures and is effective when a simple input for an end-to-end system is given.

Index Terms: speech synthesis, deep Gaussian process, sequence-to-sequence, Bayesian deep model, sequential modeling

1. Introduction

Text-to-speech (TTS) is a system for synthesizing speech from arbitrary input text sequences. The major studies we show in the followings have investigated TTS based on DNNs. A classical DNN-based statistical parametric speech synthesis (SPSS) [1] system consists of several systems in a pipeline (pipeline TTS): a front-end part that converts a text series into a linguistic feature sequence, a duration predictor that converts a phoneme-level sequence into a frame-level sequence, an acoustic model that converts a linguistic feature sequence into an acoustic feature sequence, and a vocoder that synthesizes speech from the acoustic feature sequence. However, such a classical process requires a high level of expertise, and the quality of synthesized speech is limited. DNN-based TTS using Seq2Seq learning (Seq2Seq-TTS) has been studied to overcome these problems [2, 3, 4]. An encoder-decoder structure is used in attention-based Seq2Seq TTS models. The encoder converts a text series to an intermediate representation. Then, the decoder uses attention architecture [5] to decode the representation into an acoustic representation in an autoregressive manner. Char2Wav [2], Tacotron [3], and DeepVoice3 [4] are renowned DNN-based Seq2Seq-TTS (Seq2Seq-DNN) models using attention architecture. Compared with SPSS methods, Seq2Seq learning, which integrates front-end and acoustic models as a single DNN, can synthesize speech with a particularly high degree of prosody naturalness [3].

As an alternative to neural-network-based TTS systems, a TTS system based on a DGP has been developed, in which the relationship between linguistic and acoustic features is modeled by stacked GPRs instead of a neural network. Koriyama and Kobayashi [6] showed that DGP-based TTS can synthesize more natural speech than a DNN-based one within the feed-forward architecture. In addition to feed-forward DGP, SRU-DGP speech synthesis with a simple recurrent unit (SRU [7]), which can model the entire speech production parallel, has been proposed [8] to capture the continuous temporal changes of speech features. However, the DGP-based TTS requires a longer computation time because DGP requires more complex operations, such as kernel calculation and matrix inversion, than DNN. Although the computation time can be reduced by using a GPU for parallel computing, this is not feasible in the autoregressive architecture in attention-based Seq2Seq-TTS models, making it difficult to incorporate Seq2Seq learning using attention architecture into DGP-based TTS. Another problem of using previous DGP-based TTS, SRU-DGP, is the excessive smoothing of adjacent input linguistic features, which are discrete symbol sequences, caused by recurrent structures.

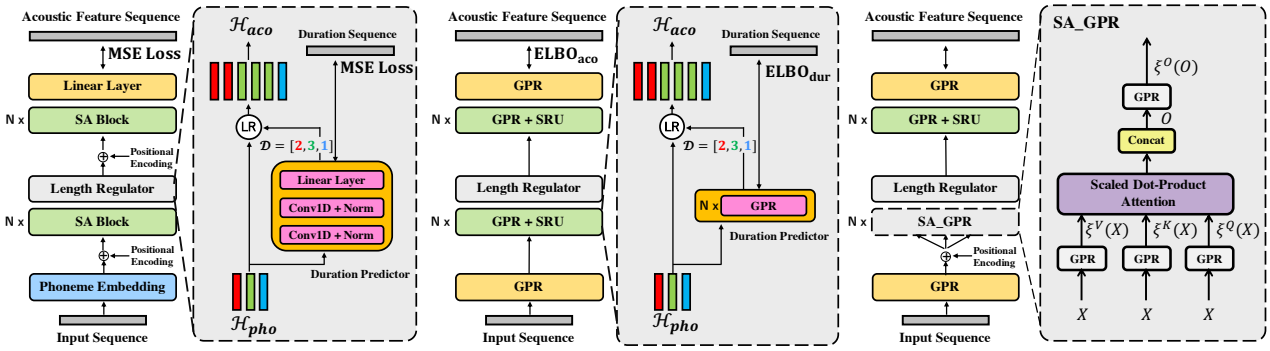
To reduce computation time, we focus on the non-autoregressive Seq2Seq-DNN models such as FastSpeech [9] and AlignTTS [10], which can synthesize acoustic features in parallel unlike autoregressive Seq2Seq-DNN models. Specifically, length regulators [9] are used to align the lengths of input/output features by repeating the input features instead of the attention architecture. Self-attention structures, which can synthesize speech in parallel, are implemented into FastSpeech and AlignTTS to reduce the computation time.

In this paper, we propose a DGP-based Seq2Seq-TTS model (Seq2Seq-SRU-DGP). Seq2Seq-SRU-DGP consists of several SRUs and GPRs in the encoder and decoder, whose effectiveness was verified in the previous DGP-based TTS model, SRU-DGP. A length regulator is used to align the length of input/output features because a length regulator with the low computational cost is compatible with DGP-based TTS. To reduce the smoothing problem, we also propose Seq2Seq-SA-DGP, in which the SRU in the encoder layer of Seq2Seq-SRU-DGP is replaced by SA-GPR, which is a combination of self-attention and GPR. Experimental evaluations show that Seq2Seq-SA-DGP can synthesize more natural synthetic speech than the conventional Seq2Seq-DNN model. The experimental evaluations also show that the self-attention structure is effective in the DGP framework, and Seq2Seq-SA-DGP can synthesize more natural speech than Seq2Seq-SRU-DGP and SRU-DGP.

2. DGP-based Speech Synthesis

DGP is a multi-layered stochastic model of GPR. In speech synthesis using DGP models [6], one DGP model is trained to pre-

http://dx.doi.org/10.21437/Interspeech.2021-896



(a) FastSpeech[9] (b) Proposed Seq2Seq-SRU-DGP (C) Proposed Seq2Seq-SA-DGP

Figure 1: *The architecture of the models discussed in this paper. (a) FastSpeech [9] model: TTS model using self-attention and length regulator. (b). Proposed Seq2Seq-SRU-DGP model: Seq2Seq DGP-based TTS model using length regulator. (c). Proposed Seq2Seq-SA-DGP model: Seq2Seq DGP-based TTS model using SA in encoder and length regulator.*

dict phoneme duration and the other model is trained to predict acoustic features. The transformation from linguistic features $\mathbf{X}$ to acoustic features can be represented by composition of several functions, as shown in (1).

$$f = f^L \circ f^{L-1} \circ \cdots \circ f^1 \tag{1}$$

It is assumed that the latent function f^ℓ is distributed over a Gaussian process. The hidden layer variable at ℓth layer $\mathbf{H}^\ell$ is expressed by $\mathbf{H}^\ell = \left[\mathbf{h}^{\ell,1}, \ldots, \mathbf{h}^{\ell,D_\ell}\right] = f^\ell(\ldots(f^1(\mathbf{X})))$ where D_ℓ is the number of dimension at ℓth layer.

As a framework for learning DGP, doubly stochastic variational inference [11] is used, which can be efficiently learned by the stochastic gradient descent for an arbitrary amount of training data. In this method, model parameters are learned to maximize the lower bound of the marginal likelihood, i.e., the evidence lower bound (ELBO). In the ELBO for training the DGP, as shown in (2), the first term represents the data fitting and the second term represents the complexity penalty. Thus, the DGP can be trained while taking complexity and performance into account simultaneously.

The previously proposed SRU-DGP [8] incorporates SRUs [7] into DGP to enable utterance-level modeling of speech synthesis. An SRU is a simple recurrent network consisting of affine transformations that can be parallelized along a time axis. The key concept of SRU-DGP is that these affine transformations are considered to be latent functions that follow a Gaussian process. The ELBO of DGP for the utterance-level modeling is

$$\mathcal{L} = \frac{1}{S}\sum_{s=1}^{S}\sum_{u=1}^{U}\left\{\sum_{d=1}^{D}\mathbb{E}_{q\left(\mathbf{h}_{u,s}^{L,d}|\hat{\mathbf{H}}_{u,s}^{L-1}\right)}\left[\log p\left(\mathbf{y}_u^d|\mathbf{h}_{u,s}^{L,d}\right)\right] - \frac{ST_u}{N}\sum_{\ell=1}^{L}\sum_{d=1}^{D_\ell}\mathrm{KL}(q(\mathbf{u}^{\ell,d})\|p(\mathbf{u}^{\ell,d}|\mathbf{Z}^\ell))\right\} \tag{2}$$

where u, U, and T_u are an utterance index, the number of utterances in the training dataset, and the number of frames in utterance u, respectively. d is a dimension index and D is the number of dimensions. $\mathbf{Z}^\ell$ is a parameter of the DGP model and an inducing variable representing the hidden layer variable at $(\ell-1)$th $\mathbf{H}^{\ell-1}$, and $\mathbf{u}^{\ell,d}$ is an inducing output. S is the number of Monte Carlo samplings obtaining the sample of the $(L-1)$th hidden layer $\hat{\mathbf{H}}_{u,s}^{L-1}$. The inference from the posterior of the hidden layer variable and Monte Carlo samplings is repeatedly performed. This is known as the reparametrization trick in Kingma and Welling's study [12]. $\mathbf{y}_u^d$ and $\mathbf{h}_u^{L,d}$ are the true acoustic feature sequence and the predicted acoustic sequence at the dth dimension, respectively.

3. Seq2Seq DNN Speech Synthesis Using Length Regulator

FastSpeech [9] is a widely known Seq2Seq-DNN speech synthesis model using a length regulator. Figure 1(a) shows the architecture of FastSpeech. The model consists of an embedding layer, self-attention blocks, a length regulator, and a linear layer.

3.1. Self-attention

The FastSpeech model contains self-attention blocks, which use the entire sequence at once to capture the interactions between each phoneme feature. A self-attention block consists of a multihead attention layer [13] and a convolutional layer. The function of the multihead attention at the ℓth layer is expressed as

$$\mathrm{MultiHead}_\ell(X) = \mathrm{Concat}(\mathrm{head}_1, \cdots, \mathrm{head}_\mathrm{h})W^{O,\ell} \tag{3}$$

$$\mathrm{head}_\mathrm{i} = \mathrm{Attention}(XW_i^{Q,\ell}, XW_i^{K,\ell}, XW_i^{V,\ell}) \tag{4}$$

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where Q, K, V are the matrices of queries, keys of dimension d_k, and values of dimension d_v, respectively. X is the intermediate features input to self-attention blocks. Scaled dot-product attention [13] (5) can capture the relationship between all phoneme features present in a sentence by computing the dot products of the query with all keys, dividing by $\sqrt{d_k}$, and using a softmax function to obtain the weights on the values. As shown in (3) and (4), queries, keys and values are projected by affine transformations, $W_i^Q \in \mathbb{R}^{d_m \times d_k}$, $W_i^K \in \mathbb{R}^{d_m \times d_k}$, $W_i^V \in \mathbb{R}^{d_m \times d_v}$, and $W_i^O \in \mathbb{R}^{hd_v \times d_m}$ where d_m is the dimension of intermediate features and h is the number of heads.

3.2. Length regulator

The FastSpeech model utilizes a length regulator to align the input/output sequence. The length regulator replicates intermediate features at the phoneme level based on the phoneme durations inferred from the duration predictor consisting of convolutional layers and a linear layer as shown in Fig. 1(a). The duration predictor is trained by minimizing the mean squared

error (MSE) loss between the duration predictor and the duration sequence from a teacher TTS model.

4. Seq2Seq DGP Speech Synthesis

We propose Seq2Seq-SRU-DGP and Seq2Seq-SA-DGP models, the architecture of which is shown in Figs. 1(b) and (c). Our model's multiple layers, length regulator, and duration predictor are based on the FastSpeech model.

4.1. Seq2Seq-SRU-DGP

Each layer of Seq2Seq-SRU-DGP consists of a GPR and SRU-based recurrent architecture [8] instead of the SA blocks in the FastSpeech model. The SRU described here is a parallel computation by affine transformation in SRU, which GP replaces.

4.1.1. Duration predictor using GPR

As with other components, we replace the convolutional layer and linear layer in duration predictor with GPR. There are several methods for training the duration predictor, including using the output from the teacher model [9], automatic alignment from the hidden marcov model (HMM), and the one obtained using the mix density network [10]. Shiga et al. [14] reported that the quality of synthetic speech using the duration from HMM was higher than when the output from the teacher model was used. Thus, we used the automatically aligned phoneme durations from the HMM to train the duration predictor. As in conventional DGP speech synthesis, acoustic features are predicted by propagating the mean predicted distribution obtained by GPR.

4.1.2. Training

The proposed Seq2Seq-SRU-DGP model is trained to maximize the sum of the ELBO related to the acoustic features ($\mathrm{ELBO}_{\mathrm{aco}}$) and the ELBO related to the duration features ($\mathrm{ELBO}_{\mathrm{dur}}$). The following equation gives the ELBO to be maximized in the proposed Seq2Seq-SRU-DGP.

$$\begin{aligned}\mathcal{L} = \frac{1}{S}\sum_{s=1}^{S}\sum_{u=1}^{U}\Bigg\{&\sum_{d=1}^{D}\mathbb{E}_{q\left(\mathbf{h}_{u,s}^{L,d}|\hat{\mathbf{H}}_{u,s}^{L-1}\right)}\left[\log p\left(\mathbf{y}_u^d|\mathbf{h}_{u,s}^{L,d}\right)\right]\\ &+\mathbb{E}_{q\left(\mathbf{h}_{u,s}^{L_{\mathrm{dur}}}|\hat{\mathbf{H}}_{u,s}^{L_{\mathrm{dur}}-1}\right)}\left[\log p\left(\mathbf{d}_u|\mathbf{h}_{u,s}^{L_{\mathrm{dur}}}\right)\right]\\ &-\frac{ST_u}{N}\sum_{\ell=1}^{L}\sum_{d=1}^{D_\ell}\mathrm{KL}(q(\mathbf{u}^{\ell,d})\|p(\mathbf{u}^{\ell,d}|\mathbf{Z}^{\ell}))\Bigg\}\end{aligned} \tag{6}$$

L_{dur} is the last of the layers included in the duration predictor, and $\mathbf{d}_u$ and $\mathbf{h}_u^{L_{\mathrm{dur}}}$ are the true duration sequence and the predicted duration sequence, respectively.

4.2. Seq2Seq-SA-DGP

Since the proposed Seq2Seq-SRU-DGP uses SRUs, which are recurrent structures, adjacent features are smoothed in time axis because the next state is predicted sequentially from the current state. This smoothing is incompatible in Seq2Seq-SRU-DGP, which uses linguistic features as input.

4.2.1. Self-attention using GPR

To prevent the excessive smoothing, we propose Seq2Seq-SA-DGP, in which partial functions in the self-attention structure are replaced with the functions over GPs in the encoder. Fig. 1(c) shows the architecture of Seq2Seq-SA-DGP. Unlike conventional self-attention, the affine transformations at the ℓth layer described in Sect. 3.1 are generalized by the latent functions, $\xi^{V,\ell}, \xi^{K,\ell}, \xi^{Q,\ell}, \xi^{O,\ell}$, which are distributed over GPs as

$$\mathrm{MultiHead}_\ell(X) = \mathrm{Concat}(\mathrm{head}_1, \cdots, \mathrm{head}_{\mathrm{h}})\xi_i^{O,\ell}(O) \tag{7}$$

$$\mathrm{head}_{\mathrm{i}} = \mathrm{Attention}(\xi_i^{Q,\ell}(X), \xi_i^{K,\ell}(X), \xi_i^{V,\ell}(X)). \tag{8}$$

We refer to the self-attention architecture described in (7) as SA_GPR.

5. Experimental Evaluations

5.1. Experimental conditions

We used a Japanese female speech dataset, JSUT corpus [15], for experimental evaluations. We used 4500, 250, and 250 sentences from JSUT's basic5000 set as the training, development, and test sets, respectively. The acoustic features were extracted using the WORLD vocoder [16] from the audio signal with a sampling rate of 48 kHz. We used 199-dimensional acoustic features as output features, which consisted of the 0–59th mel-cepstrum, log f_{o} , 5-dim code aperiodicity, and their Δ and Δ^2 features, and a voiced/unvoiced flag. We used 531-dimensional context vectors of binary-valued linguistic information (FU) as input to evaluate the model's performance in situations where rich linguistic information is given. We also used 137-dimensional vectors consisting of phoneme embeddings and accent embeddings (PA) to investigate the performance of the model in a similar situation to an end-to-end one where little or no linguistic information is given. We used phoneme durations, which were automatically aligned using a HMM, as the teacher data. The input vectors were normalized so that the values were in the range of $[-1, 1]$, and the output vectors were normalized so that they had a mean of 0 and a variance of 1, respectively.

The proposed Seq2Seq-SRU-DGP with 11 layers had feed-forward layers in the top and bottom, four SRU layers in the encoder, four SRU layers in the decoder, andone GPR layer of the duration predictor. Continuous-valued relative frame positions were added to the output of the intermediate features from the length regulator. The kernel function was the ArcCos kernel [17] with normalization terms based on the results of the prior study [6]. The dimensionality of the hidden-layer variables was 256, and each layer contained 1024 inducing points. The optimization of the model parameters was based on Adam [18] with a learning rate of 10^{-2}, and one utterance was used as a minibatch.

We compared the proposed Seq2Seq-SRU-DGP model with a pipeline DGP model (SRU-DGP [8]) and two TTS models based on FastSpeech. The model architectures of SRU-DGP with six layers had feed-forward layers in the top and bottom and four SRU layers between the feed-forward layers, as in the Seq2Seq-SRU-DGP. The other detailed settings were the same as in the proposed method. The first FastSpeech-based model used the automatic alignment of phoneme durations from HMMs instead of using the output of the teacher model (Seq2Seq-SA-NN), which had four self-attention blocks in encoder and decoder. In the second one, the self-attention of Seq2Seq-SA-NN was replaced with SRU. The basic structure and parameters of Seq2Seq-SA-NN and Seq2Seq-SRU-NN were identical to those in Ren et al.'s work [9]. The DGP and

Model	MCD	MOS
SRU-DGP	5.87	3.41
Seq2Seq-SRU-NN	6.29	2.08
Seq2Seq-SA-NN	6.03	2.83
Seq2Seq-SRU-DGP	6.15	2.71
Seq2Seq-SA-DGP	6.01	2.96
ORIG	-	4.42

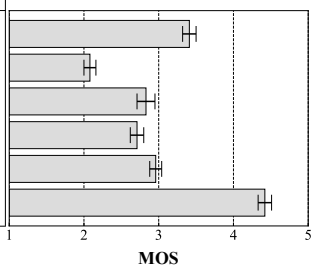


Figure 2: *MOS and MCDs results compared with pipeline DGP model, FastSpeech-based model using FU input. Error bar indicates the 95 % confidence intervals*

NN models were trained using PyTorch and FastSpeech [1]. We used the 199-dimensional acoustic features described above instead of the melspectrograms used by Ren et al. [9] and the WORLD vocoder instead of the WaveGlow vocoder [19], to experiment under the same conditions as the proposed method. We compared the following models with our Seq2Seq-SRU-DGP model:

SRU-DGP: Pipeline DGP model. SRU-DGP is the conventional model using GPR and SRU.

Seq2Seq-SA-NN: FastSpeech-based model. FastSpeech uses phoneme durations from HMM instead of using the output of the teacher model.

Seq2Seq-SRU-NN: FastSpeech-based model. The SA structure of Seq2Seq-SA-NN is replaced with SRU.

Seq2Seq-SRU-DGP: Proposed model.

Seq2Seq-SA-DGP: Proposed model. Seq2Seq DGP-based TTS contains self-attention in the encoder and SRU in the decoder.

5.2. Results

5.2.1. Subjective Evaluation

We conducted a mean opinion score (MOS) test to measure the quality of synthetic speech. Sixty participants listened to 24 utterances (four sentences, six methods) and rated the naturalness on a five-point scale: 5: excellent, 4: good, 3: fair, 2: poor, and 1: bad. Fig. 2 lists the results of the MOS tests and shows mel-cepstral distortions (MCDs). As seen in the Figure, the quality of the proposed Seq2Seq-SA-DGP using FU as input was lower than that of the pipeline DGP-based TTS model, SRU-DGP. In contrast, the quality of the proposed Seq2Seq-SA-DGP model was higher than that of Seq2Seq-SRU-NN and the Seq2Seq-SA-NN.

We also conducted preference AB tests on the naturalness of the synthesized speech to evaluate the performance of Seq2Seq-SA-DGP in the situation where a wealth of linguistic information is not given. Thirty listeners participated in each evaluation, and each listener evaluated 10 samples randomly extracted from the evaluation data. Each test is conducted by using our crowdsourcing evaluation system. Fig. 3 lists the results of AB tests of comparison between SRU-DGP and Seq2Seq-SA-DGP under two input conditions, FU input and PA input. Although the quality of the Seq2Seq-SA-DGP using FU input was lower than that of the SRU-DGP model using FU input as well

[1] https://github.com/xcmyz/FastSpeech

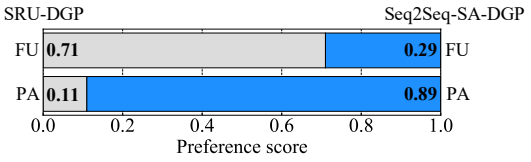


Figure 3: *Results of subjective evaluation for comparing SRU-DGP, and proposed Seq2Seq-SA-DGP using two types of input, FU and PA.*

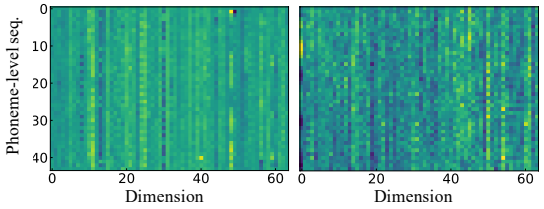


Figure 4: *Comparison of the output from encoder using SRU (left) and self-attention (right). 64 of the 256 dimensions are displayed.*

as the MOS test result, the proposed Seq2Seq-SA-DGP could synthesize more natural speech than SRU-DGP for the case of input with less linguistic information. This would be an effect of Seq2Seq learning, which can extract linguistic information more effectively than pipeline TTS.

5.2.2. Comparison of intermediate features

A comparison of the phoneme-level intermediate feature between Seq2Seq-SRU-DGP that uses SRU in the encoder and Seq2Seq-SA-DGP that uses self-attention in encoder is shown in Fig. 4. In Seq2Seq-SRU-DGP, the intermediate features of neighboring phonemes had similar values, suggesting that excessive smoothing occurs, while no such excessive smoothing is observed in the intermediate features of Seq2Seq-SA-DGP; as a result of this improvement, the synthesized speech from Seq2Seq-SA-DGP could be more natural than the one synthesized from Seq2Seq-SRU-DGP.

6. Conclusions

In this work, we proposed a DGP-based TTS framework, in which Seq2Seq learning is carried out by using length regulator. We also proposed Seq2Seq-SA-DGP, which introduces self-attention into the encoder to mitigate excessive smoothing between adjacent frames, a problem that arises when using a recurrent structure. The experimental evaluation results show that the proposed Seq2Seq-SA-DGP outperformed the TTS model based on FastSpeech in the situation where a wealth of linguistic information is given, and the pipeline DGP-based TTS in the near end-to-end situation. In future work, we will continue to improve the quality of the synthesized speech, and compare our model with other DNN-based TTS models such as AlignTTS [9] and FastSpeech2 [20].

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19K20292.

8. References

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

[2] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *Proc. ICLR workshop*, 2017.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[4] W. Ping, K. Peng, A. Gibiansky, S.O.Arik, A. Kannan, S.Narang, J.Raiman, and J.Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. ICLR*, 2018, pp. 214–217.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[6] T. Koriyama and T. Kobayashi, "Statistical parametric speech synthesis using deep Gaussian processes," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 948–959, 2019.

[7] T. Lei, Y. Zhang, S. I. Wang, H. Dai, and Y. Artzi, "Simple recurrent units for highly parallelizable recurrence," in *Proc. EMNLP*, 2018, pp. 4470–4481.

[8] T. Koriyama and H.Saruwatari, "Utterance-level sequential modeling for deep gaussian process based speech synthesis using simple recurrent unit," in *Proc. ICASSP*, 2020, pp. 7244–7248.

[9] Y.Ren, Y.Ruan, X.Tan, T.Qin, S.Zhao, Z.Zhao, and T.-Y.Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, 2019, pp. 3165–3174.

[10] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "Aligntts: Efficient feed-forward text-to-speech system without explicit alignment," in *Proc. ICASSP*, 2020, pp. 6741–6718.

[11] H. Salimbeni and M. Deisenroth, "Doubly stochastic variational inference for deep Gaussian processes," in *Proc. NIPS*, 2017, pp. 4591–4602.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, and L. Jones, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.

[14] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Transformer-based text-to-speech with weighted forced attention," in *Proc. ICASSP*, 2020, pp. 6729–6733.

[15] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "Jsut and jvs: Free japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[16] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[17] Y. Cho and L. K. Saul, "Kernel methods for deep learning," in *Proc. NIPS*, 2009, pp. 342–350.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[19] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, 2019, pp. 3617–3621.

[20] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.