AUTOREGRESSIVE VARIATIONAL AUTOENCODER WITH A HIDDEN SEMI-MARKOV MODEL-BASED STRUCTURED ATTENTION FOR SPEECH SYNTHESIS

Takato Fujimoto, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda

Nagoya Institute of Technology, Japan

ABSTRACT

This paper proposes an autoregressive speech synthesis model based on the variational autoencoder incorporating latent sequence representation for acoustic and linguistic features and the structure of a hidden semi-Markov model (HSMM). Although autoregressive models can provide efficient and accurate modeling of acoustic features, they have exposure bias, i.e., the mismatch between training (teacher-forcing) and inference (free-running). To overcome this problem, we introduce an autoregressive latent variable sequence, rather than using autoregressive generation of observations. Latent representation of alignment using HSMM-based structured attention mechanism enables the use of a completely consistent training algorithm for acoustic modeling with explicit duration models. Experimental results indicate that the proposed model outperformed baselines in subjective naturalness.

Index Terms— speech synthesis, variational autoencoder, autoregressive model, attention mechanism, hidden semi-Markov model

1. INTRODUCTION

The rapid development of deep neural networks (DNNs) has greatly improved the speech quality of speech synthesis or text-to-speech (TTS). Autoregressive sequence-to-sequence (seq2seq) models with an attention mechanism, such as Tacotron 2 [1] and Transformer TTS [2], have been reported to generate natural speech comparable to human speech. Despite these successes, problems due to lack of robustness have remained. These problems are mainly due to 1) the exposure bias and error propagation that occurs in autoregressive (AR) generation and 2) the excessive degrees of freedom in alignment between linguistic and acoustic features. Therefore, there have been various works addressing these two causes to improve the robustness of AR attention-based TTS models.

There have been studies on reducing the effect of exposure bias [3, 4, 5, 6] or adding regularization to improve the agreement between forward and backward generation [7, 8]. The basic idea behind each of these is to reduce the mismatch between training (teacher-forcing) and inference (free-running) and leverage future information. However, future information cannot be used directly in an AR manner. We use the encoder and prior with AR structure based on a variational autoencoder (VAE) framework. This allows the encoder to use the future information as an approximate posterior. We also mirror the structure of the generative model in the encoder and share the prior with the encoder. Thus, the mismatch between training and inference is reduced.

Others have proposed attention mechanisms that apply constraints to yield monotonic alignment [9, 10, 11, 12, 13]. However, the errors of the attention mechanism cannot be totally avoided. It is also difficult to control the duration because it is not explicitly represented. Thus, several works [14, 15, 16, 17] have proposed to explicitly incorporate a duration predictor instead of the attention mechanism and expand the phoneme-level hidden states in accordance with the duration. The duration predictor is trained to target an externally given duration or alignment searched within the model. To jointly optimize the duration, EATS [18] and Parallel Tacotron 2 [19] use Soft-DTW to align the predicted spectrogram. A hidden semi-Markov model (HSMM)-based structured attention [20] uses latent alignment based on the VAE framework incorporating statistical generative models (HSMMs) into the encoder, enabling the use of a completely consistent optimization algorithm for acoustic modeling with explicit duration models.

We propose an autoregressive speech synthesis model that integrates the AR VAE and HSMM-based structured attention to improve the speech quality and robustness of model training. The proposed model has a consistent AR structure between training and inference, reducing the exposure bias, and the HSMM structure in the attention mechanism enables appropriate handling of duration in a statistical manner. These structures are integrated as a structured encoder based on the VAE framework. Even though the architecture of the proposed model is complicated, it is automatically derived from a well-defined training algorithm of statistical generative models. The proposed model can also be regarded as a further extension of HMM-based speech synthesis using the flexibility of neural networks from the point that the HSMM-based structured attention achieved.

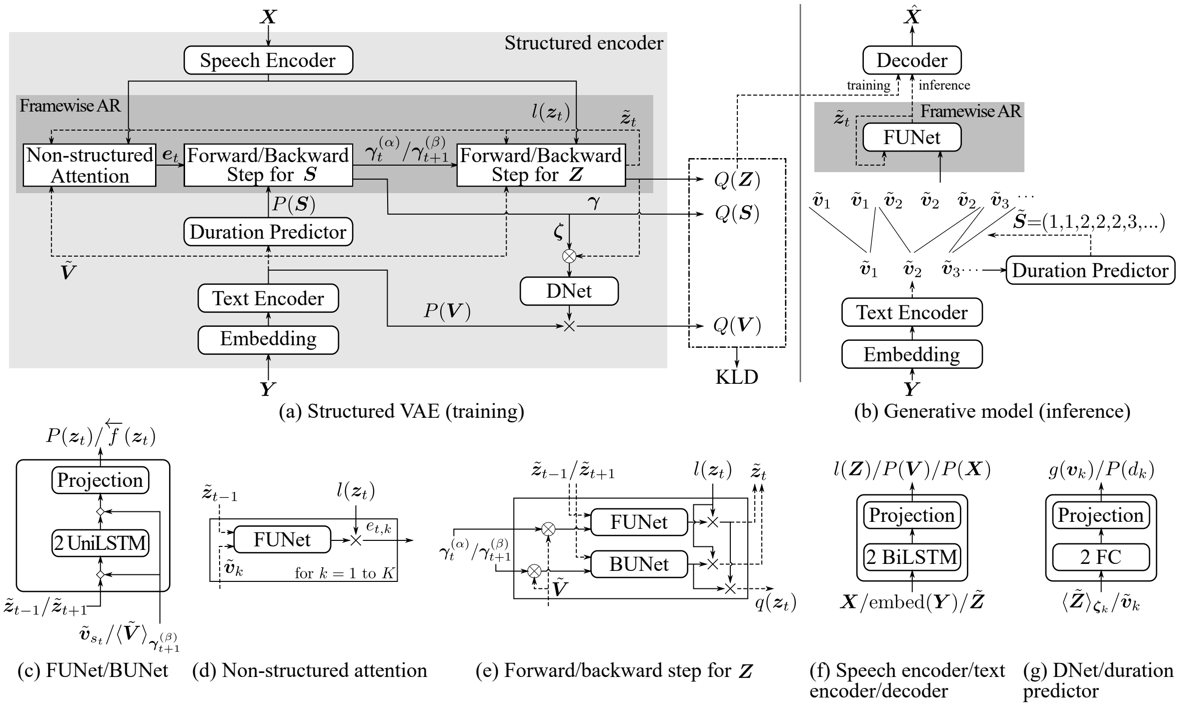
2. RELATED WORK

AR models have the issue of slow inference speed. Non-AR models have significantly faster inference speed than AR models. Although the performance of non-AR models is comparable to AR models under certain conditions, AR still has an advantage in the quality of synthesized speech. In addition, non-AR models usually require more parameters than AR models. Therefore, several works [21, 22] have investigated fast and lightweight AR models.

DurIAN [23] and Non-Attentive Tacotron [24] incorporate the duration predictor into the AR decoder. DurIAN trains the duration predictor using an external alignment. Non-Attentive Tacotron is modeled in an unsupervised manner by using the fine-grained VAE to predict duration and differentiable Gaussian up-sampling. Our proposed model is also trained in an unsupervised manner, but besides reducing the exposure bias with AR VAE, stable alignment learning is expected because the prediction of the duration is restricted by the HSMM structure, and all possible state sequences are counted.

The attention mechanism of our proposed model converts not between observed sequences (linguistic and acoustic features) but between latent variable sequences corresponding to them. A hierarchical generative model for semi-supervised learning that executes

This work was supported by JSPS KAKENHI Grant Number JP19H04136 and JP20K11862, and SCAT Research Grant.


X
Speech Encoder
Structured encoder
Framewise AR
$l(z_t)$
$\tilde{z}_t$
Non-structured Attention
e_t
Forward/Backward Step for S
$\gamma_t^{(\alpha)}/\gamma_{t+1}^{(\beta)}$
Forward/Backward Step for Z
$P(S)$
γ
Duration Predictor
ζ
$\tilde{V}$
Text Encoder
DNet
$P(V)$
Embedding
Y
$Q(Z)$
$Q(S)$
$Q(V)$
KLD
$\hat{X}$
Decoder
training
inference
Framewise AR
$\tilde{z}_t$
FUNet
$\tilde{v}_1$ $\tilde{v}_1$ $\tilde{v}_2$ $\tilde{v}_2$ $\tilde{v}_2$ $\tilde{v}_3$ ···
$\tilde{S}$=(1,1,2,2,2,3,...)
$\tilde{v}_1$ $\tilde{v}_2$ $\tilde{v}_3$···
Duration Predictor
Text Encoder
Embedding
Y
$P(z_t)/\overleftarrow{f}(z_t)$
Projection
2 UniLSTM
$\tilde{z}_{t-1}/\tilde{z}_{t+1}$
$\tilde{v}_{s_t}/\langle\tilde{V}\rangle_{\gamma_{t+1}^{(\beta)}}$
$\tilde{z}_{t-1}$
$l(z_t)$
FUNet
$e_{t,k}$
$\tilde{v}_k$
for $k = 1$ to K
$\tilde{z}_{t-1}/\tilde{z}_{t+1}$
$l(z_t)$
$\tilde{z}_t$
FUNet
BUNet
$\gamma_t^{(\alpha)}/\gamma_{t+1}^{(\beta)}$
$\tilde{V}$
$q(z_t)$
$l(Z)/P(V)/P(X)$
Projection
2 BiLSTM
X/embed(Y)/$\tilde{Z}$
$g(v_k)/P(d_k)$
Projection
2 FC
$\langle\tilde{Z}\rangle_{\zeta_k}/\tilde{v}_k$

(a) Structured VAE (training)

(b) Generative model (inference)

(c) FUNet/BUNet

(d) Non-structured attention

(e) Forward/backward step for Z

(f) Speech encoder/text encoder/decoder

(g) DNet/duration predictor

Fig. 1: Block diagram of proposed model. Dashed line, "$\diamond$", "$\otimes$", and "$\times$" denote sampling, concatenation, matrix multiplication, and Gaussian PDF multiplication, respectively. Conditions are omitted. Parameters of FUNet in (b), (d), and (e) are shared. After forward pass is completed, backward pass starts at end of sequence.

a similar conversion has been proposed [25]. Our model can be regarded as a hierarchical generative model consisting of two layers, and extending it for the semi-supervised learning is a promising direction for future work.

3. MODEL ARCHITECTURE

The essential problem of speech synthesis is to generate a sequence of acoustic features $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ conditioned on a sequence of input linguistic features $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K)$. Our proposed model introduces latent variable sequences $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_T)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K)$ corresponding to $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Unlike with typical AR models, an AR dependency on $\boldsymbol{Z}$ not on $\boldsymbol{X}$ is assumed with the proposed model. For simplicity, a non-AR structure is used for $\boldsymbol{V}$ in this paper. The alignment between $\boldsymbol{Z}$ and $\boldsymbol{V}$ is represented by the discrete variable sequence $\boldsymbol{S} = (s_1, s_2, \ldots, s_T)$ corresponding to the state sequence of the HSMM. The marginal likelihood is therefore defined as

$$P(\boldsymbol{X} \mid \boldsymbol{Y}) = \iint \sum_{\boldsymbol{S}} \prod_{t=1}^{T} P(\boldsymbol{x}_t \mid \boldsymbol{Z}) \prod_{t=1}^{T} P(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \boldsymbol{v}_{s_t}) \times \prod_{k=1}^{K} P(d_k \mid \boldsymbol{v}_k) P(\boldsymbol{v}_k \mid \boldsymbol{Y}) d\boldsymbol{V} d\boldsymbol{Z}, \tag{1}$$

where d_k represents the duration in state k in $\boldsymbol{S}$. An overview of the proposed model is shown in Fig. 1. The factorized probabilities $P(\boldsymbol{x}_t \mid \boldsymbol{Z})$, $P(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \boldsymbol{v}_{s_t})$, $P(d_k \mid \boldsymbol{v}_k)$, and $P(\boldsymbol{v}_k \mid \boldsymbol{Y})$ are parameterized as diagonal Gaussian distributions by the decoder, FUNet, duration predictor, and text encoder, respectively, in Fig. 1. Instead of maximizing $\log P(\boldsymbol{X} \mid \boldsymbol{Y})$ directly, we maximize its variational lower bound by introducing $Q(\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{V})$ as follows:

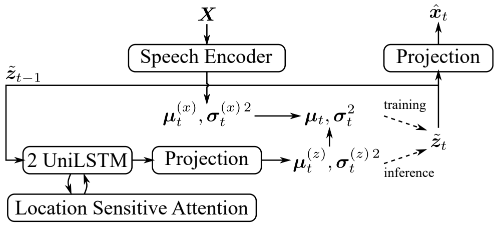


Fig. 2: Block diagram of architecture of AR VAE applied to Tacotron 2. Text encoder, post-net, and stop token are omitted.

$$\mathcal{L} = \langle \log P(\boldsymbol{X} \mid \boldsymbol{Z}) \rangle_{Q(\boldsymbol{Z},\boldsymbol{S},\boldsymbol{V})} - D_{\mathrm{KL}} \left[Q(\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{V}) || P(\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{V} \mid \boldsymbol{Y}) \right] \tag{2}$$

where $\langle\cdot\rangle_Q$ and D_{KL} denote the expectation of Q and the Kullback-Leibler divergence (KLD), respectively, and $Q(\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{V})$ is the approximate posterior distribution for $P(\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{V} \mid \boldsymbol{X}, \boldsymbol{Y})$. To yield the tractable distribution, we assume $Q(\boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{V}) = Q(\boldsymbol{Z})Q(\boldsymbol{S})Q(\boldsymbol{V})$ and derive the update equations for each factorized distribution in this section.

3.1. Autoregressive variational autoencoder

Following the VAE, the approximate posterior is parameterized using DNNs. However, as in previous works [26, 27], the structure of the approximate posterior is chosen to mirror that of the generative model, that is, we choose $Q(\boldsymbol{z}) \propto l(\boldsymbol{z}; \boldsymbol{x})P(\boldsymbol{z})$ as in a generative model $P(\boldsymbol{x}, \boldsymbol{z}) = P(\boldsymbol{x} \mid \boldsymbol{z})P(\boldsymbol{z})$. Here, $l(\boldsymbol{z}; \boldsymbol{x})$ is defined as a parameterized function of the data and can be seen as an approximate likelihood term. Since both $l(\boldsymbol{z}; \boldsymbol{x})$ and $P(\boldsymbol{z})$ are diagonal covariance Gaussian distributions, the approximate posterior can be

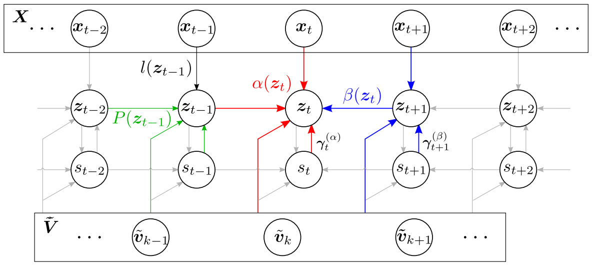


Fig. 3: Illustration of forward-backward algorithm for $Q(\boldsymbol{Z})$

obtained as a Gaussian using the Gaussian probability density functions (PDF) multiplication rule [1]. We applied this approach to the AR structure as $l(\boldsymbol{z}_t;\boldsymbol{X})P(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \boldsymbol{v}_{s_t})$, as shown in Fig. 2. The approximate posterior has the structure by $P(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \boldsymbol{v}_{s_t})$ while being guided by $l(\boldsymbol{z}_t;\boldsymbol{X})$.

3.2. State posterior probability $Q(\boldsymbol{S})$ as HSMM-based structured attention

From the dependency structure of Equation (1), the following approximate posterior for $\boldsymbol{S}$ can be derived:

$$Q(\boldsymbol{S}) \propto \int \prod_{t=1}^{T} l(\boldsymbol{z}_t;\boldsymbol{X})P(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \tilde{\boldsymbol{v}}_{s_t})d\boldsymbol{Z} \prod_{k=1}^{K} P(d_k \mid \tilde{\boldsymbol{v}}_k), \quad (3)$$

where $\tilde{\boldsymbol{v}}_k \sim P(\boldsymbol{v}_k \mid \boldsymbol{Y})$ is used for tractability. Due to the AR structure of $\boldsymbol{Z}$, Equation (3) is still intractable. Therefore, we use the following approximation:

$$\int \prod_{t=1}^{T} l(\boldsymbol{z}_t;\boldsymbol{X})P(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \tilde{\boldsymbol{v}}_{s_t})d\boldsymbol{Z} \approx \prod_{t=1}^{T} e(\boldsymbol{X}, \tilde{\boldsymbol{z}}_{t-1}, s_t),$$
$$e(\boldsymbol{X}, \tilde{\boldsymbol{z}}_{t-1}, s_t) = \mathcal{N}(\boldsymbol{\mu}_t^{(x)}; \boldsymbol{\mu}_{t,s_t}^{(z)}, \boldsymbol{\sigma}_t^{(x)\,2} + \boldsymbol{\sigma}_{t,s_t}^{(z)\,2}) = e_{t,s_t}, \quad (4)$$

where $\tilde{\boldsymbol{z}}_{t-1} \sim \alpha(\boldsymbol{z}_{t-1})$ and $\alpha(\boldsymbol{z}_{t-1})$ is a forward version of the factorized distribution of $Q(\boldsymbol{Z})$ defined in Section 3.3. Since the score $e_{t,k}$ represents the correspondence between frame t and state k as typical attention, the module for computing $e_{t,k}$ is denoted as "non-structured attention" in Fig. 1. Also note that $e_{t,k}$ is a probability density, which can be regarded as the output probability in an HSMM. Therefore, the expectation $\gamma_t(k) = \sum_{\boldsymbol{S}} Q(\boldsymbol{S})\delta(s_t, k)$, i.e., structured attention reflecting duration distribution, can be computed efficiently through the following recursive probabilities by using the generalized forward-backward algorithm [28, 29]:

$$\alpha(s_t) = \sum_{d=1}^{t} \alpha(s_{t-d} = s_t - 1)P(d \mid \tilde{\boldsymbol{v}}_{s_t}) \prod_{t'=t-d+1}^{t} e_{t',s_t}, \quad (5)$$
$$\beta(s_t) = \sum_{d=1}^{T-t} \beta(s_{t+d} = s_t + 1)P(d \mid \tilde{\boldsymbol{v}}_{s_t+1}) \prod_{t'=t+1}^{t+d} e_{t',s_t+1}. \quad (6)$$

3.3. Frame-level approximate posterior distributions $Q(\boldsymbol{Z})$ using forward-backward algorithm

As in Equation (3), the approximate posterior for $\boldsymbol{Z}$ is derived as follows:

$$Q(\boldsymbol{Z}) \propto \prod_{t=1}^{T} l(\boldsymbol{z}_t;\boldsymbol{X})P(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \tilde{\boldsymbol{V}}) = \prod_{t=1}^{T} q(\boldsymbol{z}_t). \quad (7)$$

As illustrated in Fig. 3, by introducing the forward and backward distributions $\alpha(\boldsymbol{z}_t)$ and $\beta(\boldsymbol{z}_t)$, a factorized distribution $q(\boldsymbol{z}_t)$ can be represented as

$$q(\boldsymbol{z}_t) \propto \int \prod_{t'=1}^{t} l(\boldsymbol{z}_{t'};\boldsymbol{X})P(\boldsymbol{z}_{t'} \mid \boldsymbol{z}_{t'-1}, \tilde{\boldsymbol{V}})d\boldsymbol{z}_{1:t-1}$$
$$\times \int \prod_{t'=t+1}^{T} l(\boldsymbol{z}_{t'};\boldsymbol{X})P(\boldsymbol{z}_{t'} \mid \boldsymbol{z}_{t'-1}, \tilde{\boldsymbol{V}})d\boldsymbol{z}_{t+1:T}$$
$$\propto \alpha(\boldsymbol{z}_t)\beta(\boldsymbol{z}_t). \quad (8)$$

By applying the Gaussian PDF multiplication rule, $\alpha(\boldsymbol{z}_t)$ and $\beta(\boldsymbol{z}_t)$ are rewritten as a Gaussian as follows:

$$\alpha(\boldsymbol{z}_t) \propto \int \alpha(\boldsymbol{z}_{t-1})l(\boldsymbol{z}_t;\boldsymbol{X})P(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \tilde{\boldsymbol{V}})d\boldsymbol{z}_{t-1}$$
$$\approx l(\boldsymbol{z}_t;\boldsymbol{X})P(\boldsymbol{z}_t \mid \tilde{\boldsymbol{z}}_{t-1}, \tilde{\boldsymbol{V}}), \quad (9)$$
$$\beta(\boldsymbol{z}_t) \propto \int \beta(\boldsymbol{z}_{t+1})l(\boldsymbol{z}_{t+1};\boldsymbol{X})P(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t, \tilde{\boldsymbol{V}})d\boldsymbol{z}_{t+1}$$
$$\approx \overleftarrow{f}(\boldsymbol{z}_t; \tilde{\boldsymbol{z}}_{t+1}, \tilde{\boldsymbol{V}}), \quad (10)$$

where $\tilde{\boldsymbol{z}}_{t-1} \sim \alpha(\boldsymbol{z}_{t-1})$ and $\tilde{\boldsymbol{z}}_{t+1} \sim \beta(\boldsymbol{z}_{t+1})l(\boldsymbol{z}_{t+1};\boldsymbol{X})$. We use soft-selection by weighted sum with the forward and backward occupancy probabilities $\gamma_t^{(\alpha)}(k)$ and $\gamma_{t+1}^{(\beta)}(k)$ as well as traditional attention mechanisms.

$$P(\boldsymbol{z}_t \mid \tilde{\boldsymbol{z}}_{t-1}, \tilde{\boldsymbol{V}}) \approx P(\boldsymbol{z}_t \mid \tilde{\boldsymbol{z}}_{t-1}, \langle\tilde{\boldsymbol{V}}\rangle_{\gamma_t^{(\alpha)}}), \quad (11)$$
$$\overleftarrow{f}(\boldsymbol{z}_t; \tilde{\boldsymbol{z}}_{t+1}, \tilde{\boldsymbol{V}}) \approx \overleftarrow{f}(\boldsymbol{z}_t; \tilde{\boldsymbol{z}}_{t+1}, \langle\tilde{\boldsymbol{V}}\rangle_{\gamma_{t+1}^{(\beta)}}), \quad (12)$$

where $\gamma_t^{(\alpha)}(k) \propto \sum_{d=1}^{t} \alpha(s_{t-d} = k-1) \prod_{t'=t-d+1}^{t} e_{t',k}$ and $\gamma_t^{(\beta)}(k) \propto \sum_{d=1}^{T-1} \beta(s_{t+d} = k+1)e_{t+d,k+1} \prod_{t'=t}^{t+d-1} e_{t',k}$. The backward distribution $\beta(\boldsymbol{z}_t^{(x)})$ is intractable to compute exactly because it requires the inversion of FUNet. Therefore, we introduce BUNet in Fig. 1 (e) as a backward inference model to approximate $P(\tilde{\boldsymbol{z}}_{t+1} \mid \boldsymbol{z}_t, \tilde{\boldsymbol{V}})$ as the function $\overleftarrow{f}(\boldsymbol{z}_t)$. Since the backward pass is not included in the generation process, the generation speed is comparable to that of conventional AR models.

3.4. State-level approximate posterior distributions $Q(\boldsymbol{V})$

The autoregressive generation of $\boldsymbol{Z}$ is conditioned on the state-level latent variable sequence $\boldsymbol{V}$. To feed back $\boldsymbol{Z}$ to $\boldsymbol{V}$ considering the structure of the generative model, the following update equation is used:

$$Q(\boldsymbol{V}) \propto \sum_{\boldsymbol{S}} P(\tilde{\boldsymbol{Z}} \mid \boldsymbol{S}, \boldsymbol{V})P(\boldsymbol{S} \mid \boldsymbol{V}) \prod_{k=1}^{K} P(\boldsymbol{v}_k \mid \boldsymbol{Y})$$
$$\approx \prod_{k=1}^{K} g(\boldsymbol{v}_k; \langle\tilde{\boldsymbol{Z}}\rangle_{\zeta_k})P(\boldsymbol{v}_k \mid \boldsymbol{Y}), \quad (13)$$

where $\tilde{\boldsymbol{Z}} \sim \prod_{t=1}^{T} q(\boldsymbol{z}_t)$ and $\zeta_k(t) = \gamma_t(k)/\sum_{t'=1}^{T} \gamma_{t'}(k)$ is the inverse alignment from $\boldsymbol{Z}$ to $\boldsymbol{V}$. Therefore, $\langle\tilde{\boldsymbol{Z}}\rangle_{\zeta_k}$ represents a k-th segment feature that is a weighted sum of $\boldsymbol{z}_t$ over time. The function $g(\boldsymbol{v}_k)$ is parameterized using the DNet shown in Fig. 1 (a) as the recognition model.

4. EXPERIMENTS

4.1. Experimental conditions

We conducted experiments to demonstrate the effectiveness of the proposed model. The XIMERA [30] dataset was used as a Japanese

[1] $\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_1,\boldsymbol{\sigma}_1^2)\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_2,\boldsymbol{\sigma}_2^2)$
$= \mathcal{N}\left(\boldsymbol{z}; \frac{\boldsymbol{\sigma}_1^{-2}\boldsymbol{\mu}_1+\boldsymbol{\sigma}_2^{-2}\boldsymbol{\mu}_2}{\boldsymbol{\sigma}_1^{-2}+\boldsymbol{\sigma}_2^{-2}}, \frac{1}{\boldsymbol{\sigma}_1^{-2}+\boldsymbol{\sigma}_2^{-2}}\right)\mathcal{N}(\boldsymbol{\mu}_1;\boldsymbol{\mu}_2,\boldsymbol{\sigma}_1^2+\boldsymbol{\sigma}_2^2)$

Table 1: Results of DTW-MCD (dB), DTW-F0RMSE (logHz), and MOS with 95% confidence intervals. Experiments with two different amounts of data were evaluated independently: same speech was used for **AS** only.

Models	9.5 hrs.			0.55 hrs.
	MCD	F0RMSE	MOS	MOS
AS			4.25 ± 0.12	4.53 ± 0.11
Fastspeech 2	5.50	**0.222**	3.91 ± 0.15	3.47 ± 0.13
Tacotron 2	5.63	0.238	3.74 ± 0.13	failed
HSMM-ATTN	5.49	0.245	3.83 ± 0.14	2.95 ± 0.14
AR-VAE	**5.11**	0.231	4.07 ± 0.11	3.29 ± 0.13
ARHSMM-VAE	5.16	0.236	$\mathbf{4.15} \pm 0.12$	$\mathbf{3.51} \pm 0.14$

speech corpus. The dataset was recorded at 48 kHz by a single female speaker. The mel-spectrogram is set to 256 bands and calculated with a 12.5-ms frame shift, 50-ms frame length, and 4096-point fast Fourier transform. The linguistic features were annotated phoneme and accent labels. We used 11,466 utterances for training, 488 for validation, and 133 for testing. The beginning and ending silences were trimmed from the utterances, and the total length of the training set was 9.5 hours. To show the robustness of the training and testing, two proposed models and comparison models were trained on only 450 phonetically balanced sentences (0.55 hours) in the training set. The test set was the same and included out-of-domain utterances (news, novels, words, etc.).

The dimensions of phoneme and accent embeddings and the hidden sizes of bidirectional LSTM (BiLSTM), unidirectional LSTM (UniLSTM), and fully connected (FC) layers in the proposed model were all set to 256. The latent variables $\boldsymbol{z}$ and $\boldsymbol{v}$ were configured to be 256-dimensional variables. A HSMM had one state per phoneme, structured left-to-right with no skips. We shared the parameters of the priors and the corresponding components of the encoder (i.e., FUNet, duration predictor, and text encoder) in the proposed model. The proposed model was trained using the Adam optimizer [31] with a learning rate of 10^{-4} until 20k steps then with the learning schedule following [32] for 20k warm-up steps. All models were combined with the same pre-trained WaveGrad [33] vocoder.

4.2. Evaluation

We conducted objective and subjective evaluation experiments to demonstrate the effectiveness of our proposed model. Two metrics, mel-cepstral distortion (MCD) and fundamental frequency root mean square error (F0RMSE), were used for objective evaluation. We used a dynamic time warping (DTW) algorithm to find the corresponding frame pairs between the recorded and synthesized speech. Mean opinion score (MOS) tests were also conducted for subjective evaluation. [2] Fifteen utterances were chosen at random from the test set, and the subjects were 10 native Japanese speakers.

To demonstrate the effectiveness of the AR VAE described in Section 3.1, we applied it to Tacotron 2. In that model, called **AR-VAE**, the backward path was omitted and a projection was added between the LSTM and projection to output the mean and variance of the latent variable, and a speech encoder with a structure similar to that of the encoder of Tacotron 2 was used, removing the pre-net (see Fig. 2). We compared two versions of the proposed model (**AR-VAE** and **ARHSMM-VAE**) with other baseline models, including Analysis by Synthesis (**AS**), **Fastspeech 2** [14], **Tacotron 2** [1], and the HSMM attention-based model (**HSMM-ATTN**) [20]. The proposed model and **HSMM-ATTN**, which use HSMM-based attention, and the other models were set to batch sizes of 1 and 16, respectively.

Table 2: Results of DMAE (frames), MCD (dB), and F0RMSE (logHz) in ablation studies.

Models	DMAE	MCD	F0RMSE
ARHSMM-VAE	1.36	**6.09**	**0.307**
w/ $q(\boldsymbol{z}_t) = \alpha(\boldsymbol{z}_t)$	1.38	**6.09**	0.322
w/ $q(\boldsymbol{z}_t) = l(\boldsymbol{z}_t; \boldsymbol{X})$	1.61	6.21	0.327
w/o sharing params.	1.53	6.52	0.336
w/o $\boldsymbol{V}$	**1.34**	6.22	0.330

For **HSMM-ATTN**, we shared the parameters of the DNN-HSMM, duration model, and pre-net and used BiLSTM layers instead of FC layers. For **Fastspeech 2**, the Montreal forced aligner (MFA) [34] was used following [14]. The MFA trained on the 9.5-hour training set was used for all experiments. The hidden sizes were set to 512 and 256 for the 9.5-hour and 0.55-hour training sets, respectively. For models except **Fastspeech 2**, two frames were treated as one unit. Guided attention loss [35] was used for **Tacotron 2** and **AR-VAE** with location-sensitive attention on the 0.55-hour training set. However, training **Tacotron 2** failed.

The results are shown in Table 1. **ARHSMM-VAE** subjectively outperformed all other models on both training sets. This shows that the proposed model is effective. The speech quality of **ARHSMM-VAE** and **Fastspeech 2** was comparable on the 0.55-hour training set. **ARHSMM-VAE** has only 6.7 million parameters, which is 4.5 times smaller than **Fastspeech 2**, and is trained without external duration, pitch, and energy, unlike **Fastspeech 2**. **AR-VAE** and **ARHSMM-VAE** significantly improved in MCD and MOSs. In contrast, **Tacotron 2** and **HSMM-ATTN** with the conventional AR decoder obtained lower MOSs. These results seem to be due to the exposure bias.

4.3. Ablation study

We conducted ablation studies to verify the effectiveness of several components of **ARHSMM-VAE** using three objective metrics: duration mean absolute error (DMAE), MCD, and F0RMSE. The target duration was the Viterbi path obtained from the structured encoder of each model without using the prior duration distribution (i.e., the duration predictor). Then, instead of using DTW, the speech was generated using that duration for calculating MCD and F0RMSE. The results are shown in Table 2. Removing the backward path from the approximate posterior of $\boldsymbol{Z}$ made F0RMSE worse, then removing both the forward and backward paths also made DMAE and MCD worse. This is like a vanilla VAE with an AR prior, indicating that a consistent structure between priors and posteriors is important for VAEs. Next, the model without sharing the parameters of the prior and posterior of $\boldsymbol{Z}$ (i.e., FUNet parameters) achieved the worst MCD. This shows the validity of using the structures of statistical generative models in the proposed model. Finally, the model without $\boldsymbol{V}$, i.e., treating $\boldsymbol{V}$ as constant, increased MCD and F0RMSE. This means that the interaction between $\boldsymbol{Z}$ and $\boldsymbol{V}$ is effective in adapting it to given speech samples in the structured encoder.

5. CONCLUSIONS

We proposed a VAE-based AR speech synthesis model to address the problems of conventional AR models with an attention mechanism. The proposed model integrates AR VAE to reduce the exposure bias and HSMM-based structured attention to avoid attention errors and control duration. Our experimental results indicate that the proposed model outperformed the baselines. Future work includes speeding up training and inference, simplifying the HSMM-based structured attention, and extending it to the hierarchical generative model.

[2] Audio samples: https://www.sp.nitech.ac.jp/~taka19/demos/icassp22/.

6. REFERENCES

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *AAAI*, 2019.

[3] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *NeurIPS*, 2015.

[4] H. Guo, F. K. Soong, L. He, and L. Xie, "A new GAN-based end-to-end TTS training algorithm," in *Proc. Interspeech*, 2019, pp. 1288–1292.

[5] Q. Dou, J. Efiong, and M. J. Gales, "Attention forcing for speech synthesis," in *Proc. Interspeech*, 2020, pp. 4014–4018.

[6] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust Tacotron-based TTS," in *Proc. ICASSP*, 2020, pp. 6274–6278.

[7] Y. Zheng, J. Tao, Z. Wen, and J. Yi, "Forward-backward decoding sequence for regularizing end-to-end TTS," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2067–2079, 2019.

[8] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," in *Proc. ICML*, 2019, pp. 5410–5419.

[9] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *Proc. ICASSP*, 2018, pp. 4789–4793.

[10] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS," in *Proc. Interspeech*, 2019, pp. 1293–1297.

[11] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. ICASSP*, 2020, pp. 6194–6198.

[12] Q. Tian, C. Liu, Z. Zhang, H. Lu, L. Chen, B. Wei, P. He, and S. Liu, "FeatherTTS: Robust and efficient attention based neural TTS," in *Proc. SSW*, 2021, pp. 200–204.

[13] S. Mehta, E. Szekely, J. Beskow, and G. E. Henter, "Neural HMMs are all you need (for high-quality attention-free TTS)," in *Proc. ICASSP (accepted)*, 2022.

[14] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," *arXiv:2006.04558*, 2020.

[15] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *NeurIPS*, 2020.

[16] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[17] Y. Yasuda, X. Wang, and J. Yamagishi, "End-to-end text-to-speech using latent duration based on VQ-VAE," in *Proc. ICASSP*, 2021, pp. 5694–5698.

[18] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *ICLR*, 2021.

[19] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Skerry-Ryan, and Y. Wu, "Parallel Tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling," in *Proc. Interspeech*, 2021, pp. 141–145.

[20] Y. Nankaku, K. Sumiya, T. Yoshimura, S. Takaki, K. Hashimoto, K. Oura, and K. Tokuda, "Neural sequence-to-sequence speech synthesis using a hidden semi-Markov model based structured attention mechanism," *arXiv:2108.13985*, 2021.

[21] D. Wang, L. Deng, Y. Zhang, N. Zheng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Fcl-Taco2: Towards fast, controllable and lightweight text-to-speech synthesis," in *Proc. ICASSP*, 2021, pp. 5714–5718.

[22] S. Wang, Z. Ling, R. Fu, J. Yi, and J. Tao, "Patnet: A phoneme-level autoregressive transformer network for speech synthesis," in *Proc. ICASSP*, 2021, pp. 5684–5688.

[23] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, "DurIAN: Duration informed attention network for speech synthesis," in *Proc. Interspeech*, 2020, pp. 2027–2031.

[24] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-Attentive Tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling," *arXiv:2010.04301*, 2020.

[25] T. Fujimoto, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Semi-supervised learning based on hierarchical generative models for end-to-end speech synthesis," in *Proc. ICASSP*, 2020, pp. 7644–7648.

[26] T. Salimans, "A structured variational auto-encoder for learning deep hierarchies of sparse features," *arXiv:1602.08734*, 2016.

[27] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *NeurIPS*, 2016.

[28] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE transactions on information and systems*, vol. 90, no. 5, pp. 825–834, 2007.

[29] K. Tokuda, K. Hashimoto, K. Oura, and Y. Nankaku, "Temporal modeling in neural network based statistical parametric speech synthesis," in *Proc. SSW*, 2016, pp. 106–111.

[30] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *Proc. SSW*, 2004, pp. 179–184.

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

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[33] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *ICLR*, 2021.

[34] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[35] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. ICASSP*, 2018, pp. 4784–4788.