Interspeech 2022
18-22 September 2022, Incheon, Korea

End-to-End Text-to-Speech Based on Latent Representation of Speaking Styles Using Spontaneous Dialogue

Kentaro Mitsui[1], *Tianyu Zhao*[1], *Kei Sawada*[1], *Yukiya Hono*[2], *Yoshihiko Nankaku*[2], *Keiichi Tokuda*[2]

[1]rinna Co., Ltd., Japan, [2]Nagoya Institute of Technology, Japan

{kemits,tianyuz,keisawada}@rinna.co.jp, {hono,nankaku,tokuda}@sp.nitech.ac.jp

Abstract

The recent text-to-speech (TTS) has achieved quality comparable to that of humans; however, its application in spoken dialogue has not been widely studied. This study aims to realize a TTS that closely resembles human dialogue. First, we record and transcribe actual spontaneous dialogues. Then, the proposed dialogue TTS is trained in two stages: first stage, variational autoencoder (VAE)-VITS or Gaussian mixture variational autoencoder (GMVAE)-VITS is trained, which introduces an utterance-level latent variable into variational inference with adversarial learning for end-to-end text-to-speech (VITS), a recently proposed end-to-end TTS model. A style encoder that extracts a latent speaking style representation from speech is trained jointly with TTS. In the second stage, a style predictor is trained to predict the speaking style to be synthesized from dialogue history. During inference, by passing the speaking style representation predicted by the style predictor to VAE/GMVAE-VITS, speech can be synthesized in a style appropriate to the context of the dialogue. Subjective evaluation results demonstrate that the proposed method outperforms the original VITS in terms of dialogue-level naturalness.

Index Terms: end-to-end TTS, spontaneous dialogue, speaking style, variational autoencoder, BERT

1. Introduction

Dialogue is a conversation between two or more people. In recent years, the development of natural language processing has greatly improved the quality of text-based dialogue generation resulting in human-computer or computer-computer dialogue [1, 2]. On the other hand, speech is essential for human dialogue. Therefore, TTS has an important role in facilitating communication between humans and computers.

The development of deep learning has resulted in synthesizing speech in a quality comparable to that of humans [3, 4]. However, dialogue speech often has characteristics that are different from those of the recited speech. First, while recited speech has transcript beforehand, dialogue speech is a spontaneous speech. Therefore, dialogue speech is more difficult to model than recited speech because of repetition, fillers, prolongation, and breaths. Second, dialogues are frequently accompanied by backchannels, also known as *aizuchi* [5] in Japanese, and laughter. These factors transcribed in the same way can be uttered in various styles. Thus, it is necessary to appropriately model the one-to-many relationship between text and speech. Finally, several factors of speech such as pitch [6], energy [7], and speech rate [8] can be in sync with the dialogue partner, which is called entrainment [7]. Considering these features, TTS can resemble more natural human-human dialogue.

Several studies have focused on conversational TTS. Yokoyama et al. used Utsunomiya University spoken dialogue database [9] to control paralinguistic information [10]. They utilized paralinguistic information tags and did not consider dialogue history. Guo et al. used the bidirectional encoder representations from Transformers (BERT) [11] to compute encodings of current text and chat history and fed them to the encoder of the acoustic model to improve the naturalness of the synthetic speech [12]. Cong et al. considered the acoustic information of the previous utterance as well as the linguistic information by predicting the Global Style Token [13] of the current utterance from the mel-spectrogram of the previous utterance [14]. These two studies used predefined transcript to record spoken dialogue, which may differ from actual dialogue without transcript, in terms of the frequency of the spontaneous behaviors and the presence or absence of backchannels.

In this study, we record a free-form dialogue on a given topic without preparing a transcript to achieve more human-like dialogue speech synthesis. Of the three aforementioned features of spontaneous dialogue, (1) we use VITS [4], an end-to-end TTS which robustly estimates alignment between text and speech by monotonic alignment search (MAS) and blank tokens. (2) We incorporate an utterance-level latent variable into VITS to facilitate the modeling of one-to-many relationship between text and speech. Following the framework of VAE [15], we propose two methods: VAE-VITS that assumes a standard normal distribution for the prior distribution of the latent variable, and GMVAE-VITS, which assumes a Gaussian Mixture Model (GMM) for the prior. Furthermore, by sharing the latent space among speakers, training is encouraged to make similar speaking styles between speakers close in the latent space. (3) We introduce a style predictor that predicts the speaking style of current speech based on dialogue history to realize an entrainment that is close to actual dialogue. Speech sequences in dialogue history are difficult to handle directly because their length is extremely long. Therefore, we adopt a two-stage training framework: first, VAE/GMVAE-VITS is trained using a single utterance and then style predictor is trained using a sequence of style representations extracted from past utterances.

2. Spontaneous dialogue corpus

To record speech that is close to actual human-human conversation, the following method is used for speech recording and post-processing. First, two or more speakers are given a topic and asked to talk freely and their voices are recorded on independent channels. Automatic speech recognition (ASR) automatically transcribes the recorded speech. Transcripts are then manually modified and given time information (start and end time of each utterance) to produce the final transcript with time information. Using this time information, the audio file is split to obtain utterance-level speech. Although it is time-consuming to transcribe and assign time information to free dialogue, the use of ASR can greatly reduce the burden of post-processing. In addition, the lack of predefined transcript allows the speakers to produce more spontaneous speech which contains repetition, fillers, prolongation, and also backchannels. Speech data

10.21437/Interspeech.2022-259

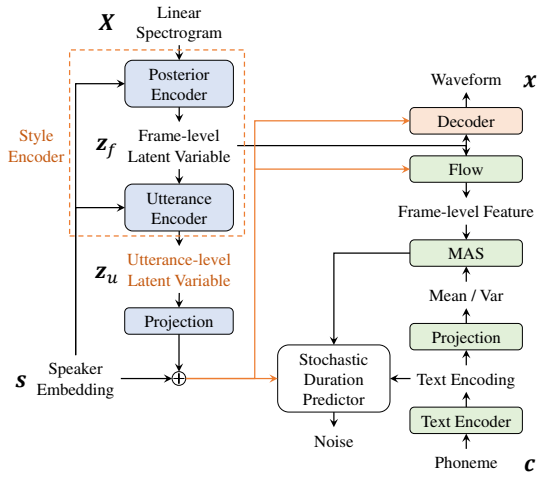


Figure 1: *Training procedure of VITS incorporating utterance-level latent variable.*

recorded in the aforementioned method enables to model the characteristics of actual dialogues more accurately.

3. Two-stage training-based dialogue TTS

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and $s_1, \ldots, s_N$ be the sequence of dialogue speech and speaker ID of each utterance, respectively. The purpose of this study is to synthesize the speech $\boldsymbol{x}_n$ corresponding to the nth text $\boldsymbol{t}_n$ and speaker ID s_n by considering the past dialogue speech $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}$ and speaker ID $s_1, \ldots, s_{n-1}$, that is, to model the distribution $p(\boldsymbol{x}_n | \boldsymbol{t}_n, s_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}, s_1, \ldots, s_{n-1})$. In spoken dialogue, each of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}$ is an extremely long time series, and it is difficult to model them directly when n is large. Therefore, we propose a two-step training framework which is described in section 3.1 and section 3.2.

3.1. Speaking style modeling using VAE/GMVAE-VITS

The first training stage models the utterance-level relationship between text and speech, that is, $p_\theta(\boldsymbol{x}_n | \boldsymbol{t}_n, s_n)$. In this study, we model this relationship using VITS [4]. VITS is an end-to-end TTS model that learns the relationship between the phoneme sequence $\boldsymbol{c}$ and speech waveform $\boldsymbol{x}$ via frame-level latent variable $\boldsymbol{z}_f$. VITS estimates monotonic alignment between $\boldsymbol{c}$ and $\boldsymbol{z}_f$ during training using MAS algorithm [16]. Thus, it can be trained more stably than fully attention-based models such as Tacotron 2 [3].

The proposed method introduces an utterance-level latent variable $\boldsymbol{z}_u$ to represent the speaking style of each utterance. We also introduce an utterance encoder that predicts the mean $\boldsymbol{\mu}_u$ and variance $\boldsymbol{\sigma}_u^2$ of $\boldsymbol{z}_u$ using $\boldsymbol{z}_f$ and speaker embedding $\boldsymbol{s}$. That is, the posterior distribution of $\boldsymbol{z}_u$ is given as follows:

$$q(\boldsymbol{z}_u | \boldsymbol{z}_f, \boldsymbol{s}) = \mathcal{N}(\boldsymbol{z}_u; \boldsymbol{\mu}_u, \mathrm{diag}(\boldsymbol{\sigma}_u^2)). \tag{1}$$

Hereafter, the utterance encoder and the posterior encoder, which predicts $\boldsymbol{z}_f$ from the linear spectrogram $\boldsymbol{X}$ of speech $\boldsymbol{x}$, will be called a style encoder together. We condition the stochastic duration predictor, flow, and decoder on $\boldsymbol{z}_u$ to predict duration, acoustic feature, and waveform considering given speaking style, respectively. Concretely, we apply a linear projection to $\boldsymbol{z}_u$ and add it to the speaker embedding $\boldsymbol{s}$, which is fed to each module. A conceptual diagram of the proposed method is depicted in Fig. 1.

The proposed method can be trained by maximizing the evidence lower bound (ELBO) of the following log-likelihood (for simplicity, we omit $\boldsymbol{s}$ in the equation below):

$$\begin{aligned}
\log p(\boldsymbol{x}|\boldsymbol{c}) \geq\ & \mathbb{E}_{q(\boldsymbol{z}_f|\boldsymbol{x})q(\boldsymbol{z}_u|\boldsymbol{z}_f)} \left[\log p(\boldsymbol{x}|\boldsymbol{z}_f, \boldsymbol{z}_u)\right] \\
& - \mathbb{E}_{q(\boldsymbol{z}_u|\boldsymbol{z}_f)} \left[D_{\mathrm{KL}}(q(\boldsymbol{z}_f|\boldsymbol{x}) || p(\boldsymbol{z}_f|\boldsymbol{c}, \boldsymbol{z}_u)\right] \\
& - D_{\mathrm{KL}}(q(\boldsymbol{z}_u|\boldsymbol{z}_f) || p(\boldsymbol{z}_u)) \qquad (2) \\
\approx\ & \frac{1}{M'M''} \sum_{m'=1}^{M'} \sum_{m''=1}^{M''} \log p(\boldsymbol{x}|\boldsymbol{z}_f^{(m')}, \boldsymbol{z}_u^{(m'')}) \\
& - \frac{1}{M''} \sum_{m''=1}^{M''} D_{\mathrm{KL}}(q(\boldsymbol{z}_f|\boldsymbol{x}) || p(\boldsymbol{z}_f|\boldsymbol{c}, \boldsymbol{z}_u^{(m'')}) \\
& - \frac{1}{M'} \sum_{m'=1}^{M'} D_{\mathrm{KL}}(q(\boldsymbol{z}_u|\boldsymbol{z}_f^{(m')}) || p(\boldsymbol{z}_u)) \qquad (3)
\end{aligned}$$

where M', M'' denote the numbers of Monte Carlo sampling for $\boldsymbol{z}_f, \boldsymbol{z}_u$, respectively. The first and second terms of eq. (3) can be calculated in the same way as in the original VITS. Assuming $p(\boldsymbol{z}_u) = \mathcal{N}(\boldsymbol{z}_u; \boldsymbol{0}, \boldsymbol{I})$, the third term is the KL divergence between two multivariate normal distributions, which can be calculated analytically. We call the proposed method defined by the above model and objective function as VAE-VITS.

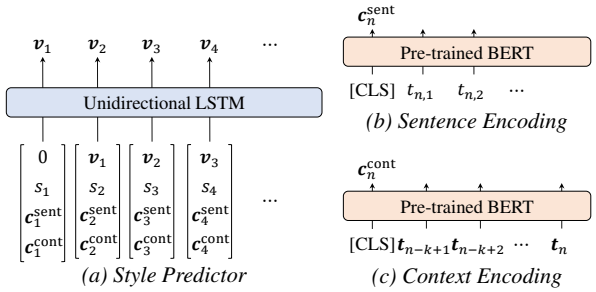
This study further examines the use of a Gaussian mixture model (GMM) with equal mixture weights $p(\boldsymbol{z}_u|\boldsymbol{y}_u) = \mathcal{N}(\boldsymbol{z}_u; \boldsymbol{\mu}_{\boldsymbol{y}_u}, \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{y}_u}))$ for the prior distribution, following GMVAE-Tacotron [17], where $\boldsymbol{y}_u$ denotes the discrete latent class corresponding to $\boldsymbol{z}_u$, and the number of latent classes is defined as K. In this case, the ELBO of the log-likelihood is approximated as follows, instead of eq. (3):

$$\begin{aligned}
\log p(\boldsymbol{x}|\boldsymbol{c}) \geq\ & \frac{1}{M'M''} \sum_{m'=1}^{M'} \sum_{m''=1}^{M''} \log p(\boldsymbol{x}|\boldsymbol{z}_f^{(m')}, \boldsymbol{z}_u^{(m'')}) \\
& - \frac{1}{M''} \sum_{m''=1}^{M''} D_{\mathrm{KL}}(q(\boldsymbol{z}_f|\boldsymbol{x}) || p(\boldsymbol{z}_f|\boldsymbol{c}, \boldsymbol{z}_u^{(m'')})) \\
& - \frac{1}{M'} \sum_{m'=1}^{M'} \Bigg\{ D_{\mathrm{KL}}(q(\boldsymbol{y}_u|\boldsymbol{z}_f^{(m')}) || p(\boldsymbol{y}_u)) \\
& + \sum_{\boldsymbol{y}_u=1}^{K} q(\boldsymbol{y}_u|\boldsymbol{z}_f^{(m)}) D_{\mathrm{KL}}(q(\boldsymbol{z}_u|\boldsymbol{z}_f^{(m)}) || p(\boldsymbol{z}_u|\boldsymbol{y}_u)) \Bigg\}. \qquad (4)
\end{aligned}$$

Some utterances such as backchannels are short and uttered in diverse styles, while others are longer and uttered in a relatively consistent style in dialogue speech. By assuming GMM for the prior distribution, the various styles of dialogue speech are expected to be represented more accurately. We call this method GMVAE-VITS.

3.2. Style transition modeling using style predictor

In the second stage of training, a style predictor which predicts the distribution of $\boldsymbol{z}_n$ using sequences of speaking style $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n-1}$ and speaker ID $s_1, \ldots, s_n$ is trained, with the style encoder trained in the first stage fixed. An outline of the proposed style predictor is shown in Fig. 2 (a). A simple unidirectional LSTM is employed to model the transition of speaking styles during a dialogue. When we directly use $\boldsymbol{z}_u$ described in section 3.1 as a speaking style representation, the variation caused by sampling from the posterior distribution hinders the training of style predictor. Therefore, we define $\boldsymbol{v}_u = [\boldsymbol{\mu}_u^\top, \boldsymbol{\sigma}_u^\top]^\top$ as a style vector and used it as an input/output of the style predictor. Hereafter, we replace the subscript u with the index n in the dialogue for simplicity.



(a) Style Predictor *(b) Sentence Encoding* *(c) Context Encoding*

Figure 2: *Conceptual diagram of (a) style predictor, (b) sentence encoding, and (c) context encoding.*

Style predictor takes two sequences as inputs, the style vector sequence $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-1}$ and speaker ID sequence $s_1, \ldots, s_n$, and predicts the style vector $\boldsymbol{v}_n$ of the current utterance. The model is trained to minimize the mean squared error $\|\tilde{\boldsymbol{v}}_n - \boldsymbol{v}_n\|^2$ between the predicted and target style vector $\tilde{\boldsymbol{v}}_n$ and $\boldsymbol{v}_n$, respectively. In addition, following Guo et al. [12], two types of linguistic features extracted using pre-trained BERT [11] are input supplementally: (1) sentence encoding $\boldsymbol{c}_n^{\text{sent}}$, an output vector corresponding to the [CLS] token, where the text $\boldsymbol{t}_n = (t_{n,1}, t_{n,2}, \ldots)$ is prefixed with [CLS] token and input to the BERT, and (2) context encoding $\boldsymbol{c}_n^{\text{cont}}$, an output vector corresponding to the [CLS] token, where a series of k text sequences $\boldsymbol{t}_{n-k+1}, \ldots, \boldsymbol{t}_n$ are concatenated, prefixed with a [CLS] token, and input to the BERT. The procedures of computing these encodings are shown in Fig. 2 (b) and (c).

4. Experiments

4.1. Experimental conditions

4.1.1. Datasets

Japanese dialogues between two females, who can see each other's faces through glass and can hear each other's voice with a headphone, were recorded in isolated soundproof chambers by the method described in section 2 and used for the experiment. The speakers were colleagues working in radio, which made their conversation more friendly. The recorded data contained dialogues of 55 topics, which were transcribed and divided into 18,385 utterances. Azure speech to text was used for ASR. 45 dialogues (15,739 utterances), 5 dialogues (1,284 utterances), and 5 dialogues (1,362 utterances) were used as training, development, and evaluation set, respectively. All the experiments were conducted using 24 kHz/16 bit speech signals. A 186-dimensional linguistic feature extracted using Japanese text frontend, Open JTalk [1] , was used as an input of VITS ($\boldsymbol{c}$ in Fig. 1), instead of phoneme sequences.

4.1.2. Model and training details

We trained three models: the original VITS [4], which does not explicitly consider speaking styles, and the proposed VAE-VITS and GMVAE-VITS described in section 3.1. The hyperparameters of VITS were set to be the same as in the previous study. Following GMVAE-Tacotron [17], the utterance encoder was composed of two 1D-convolutional layers with 512 filters and a kernel size of three, two bidirectional LSTM layers with 256 cells at each direction, and a mean pooling layer followed by a linear projection layer. The number of Monte Carlo sampling was set to 1 and the dimension of $\boldsymbol{z}_u$ was set to 16. For GMVAE-VITS, the number of latent classes K was set to 10, and the initial value and lower bound of $\boldsymbol{\sigma}_{\boldsymbol{y}_u}$ were set to e^{-1} and e^{-2}, respectively. All the models were trained for 200k steps using AdamW optimizer [18] with $\beta_1 = 0.8, \beta_2 = 0.99$ and weight decay $\lambda = 0.01$. The batch size was set to 48 and the learning rate was scheduled in the same manner as in the previous study [4]. KL annealing [19] was introduced for training VAE/GMVAE-VITS: KL weights of terms newly introduced by the proposed method were increased from 0 to 1 by cosine annealing over the initial 50k steps.

[1] `http://open-jtalk.sourceforge.net/`

Table 1: *RMSE between the style vector predicted using the style predictor and one extracted from the target speech.*

Method	**None**	**S**	**C**	**S+C**
VAE-VITS	0.329	0.256	0.291	0.250
GMVAE-VITS	0.713	0.484	0.610	0.470

Table 2: *Objective evaluation results.*

Method	MCD	MSD	DUR
VITS	7.70	9.50	0.50
VAE-oracle	7.06	8.18	0.44
GMVAE-oracle	7.04	8.17	0.41
VAE-predicted	7.54	9.22	0.50
GMVAE-predicted	7.51	9.18	0.47

Three unidirectional LSTM layers with 256 cells and dropout [20] rate 0.5 were used as the style predictor. The target style vector was obtained using the style encoder of trained VAE/GMVAE-VITS. We trained BERT [11] from scratch on approximately 400 GB of Japanese text and used it to compute 1024-dimensional sentence encoding $\boldsymbol{c}^{\text{sent}}$ and context encoding $\boldsymbol{c}^{\text{cont}}$. The text sequence length k for obtaining $\boldsymbol{c}^{\text{cont}}$ was set to 10. While the dialogues in the training set consisted of 160–693 utterances, we (1) randomly selected a sequence length l from the range $[10, 30]$, and (2) randomly extracted consecutive l utterances from each dialogue. Thereby, we avoided overfitting caused by memorizing the entire series. The model was trained up to 2,000 steps with batch size 32 using the same AdamW optimizer used in training VITS and the checkpoint with the smallest validation loss was used.

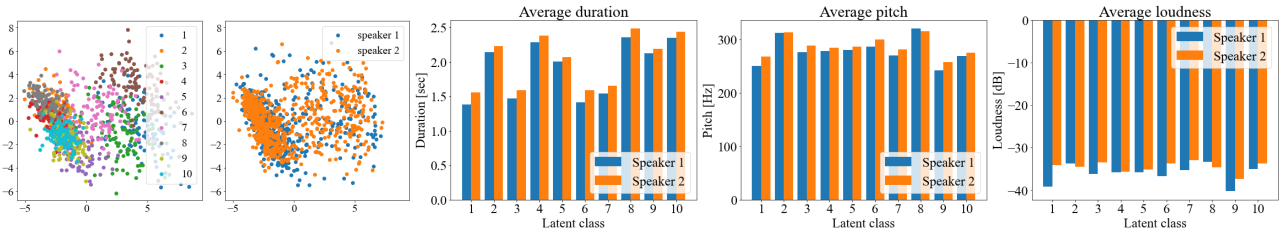
4.2. Results

4.2.1. Objective evaluation of style predictor

To evaluate the effectiveness of providing additional linguistic information to the style predictor, we trained the following four models for each of VAE/GMVAE-VITS: (1) **None**: neither $\boldsymbol{c}^{\text{sent}}$ nor $\boldsymbol{c}^{\text{cont}}$ was used, (2) **S**: only $\boldsymbol{c}^{\text{sent}}$ was used, (3) **C**: only $\boldsymbol{c}^{\text{sent}}$ was used, and (4) **S+C**: both $\boldsymbol{c}^{\text{sent}}$ and $\boldsymbol{c}^{\text{cont}}$ were used. The root mean squared error (RMSE) between predicted and target style vectors is presented in Table 1. The RMSE of **S** was significantly smaller than **None** for both VAE and GMVAE, which indicates the effectiveness of $\boldsymbol{c}^{\text{sent}}$ in style prediction. In addition, by comparing **None** and **C**, or **S** and **S+C**, we can see that $\boldsymbol{c}^{\text{cont}}$ also contributed to improved prediction accuracy. These results suggest that it is effective to use not only acoustic but also linguistic history in predicting transition of speaking styles during a dialogue. In the following experiments, we used the style predictor trained in the **S+C** condition.

4.2.2. Objective evaluation of overall system

We conducted an objective evaluation to compare the performance of the proposed methods with baseline VITS. For the proposed methods, we evaluated two cases: one is to use $\boldsymbol{z}_u$ obtained from the target speech using the style encoder (VAE/GMVAE-oracle) and the other is to use $\boldsymbol{z}_u$ sam-



(a) Prior distribution *(b) Posterior distribution* *(c) Duration* *(d) Pitch* *(e) Loudness*

Figure 3: *Analysis of latent space learned by GMVAE-VITS in terms of (a) prior distribution, (b) posterior distribution, (c) average duration, (d) average pitch, and (e) average loudness.*

Table 3: *Results of MOS evaluation on utterance-level and dialogue-level naturalness with 95% confidence intervals.*

Method	Utterance	Dialogue
VITS	3.38±0.14	3.34±0.12
GMVAE-oracle	3.51±0.12	3.59±0.12
GMVAE-predicted	3.56±0.12	3.53±0.11

pled from the distribution defined by the predicted style vector $\tilde{\boldsymbol{v}}_u$ (VAE/GMVAE-predicted). The performance was evaluated in terms of following metrics: (1) mel-cepstral distortion (MCD) [21], the RMSE of a 60-dimensional mel-cepstrum extracted from synthetic and target speech, (2) mel-spectral distortion (MSD), the RMSE of 80-dimensional mel-spectrogram extracted from synthetic and target speech, and (3) total duration error (DUR), the error of speech length of synthetic and target speech. Since the series length differs between synthetic and target speech, dynamic time warping [22] was used to align them before calculating MCD and MSD.

The results are presented in Table 2. Both VAE/GMVAE-oracle showed significant improvement in MCD and MSD compared to baseline VITS. DUR was also slightly improved, suggesting that $\boldsymbol{z}_u$ represents duration-related features as well as acoustic features. VAE/GMVAE-predicted also showed improvement in MCD and MSD relative to baseline VITS. This indicates that the style predictor was able to predict speaking styles that are close to those of target speech. The performance of GMVAE-VITS was slightly better than VAE-VITS for both oracle and predicted. This is probably because the richer prior of GMVAE-VITS could represent the various speaking styles of dialogue speech more appropriately. In the next section, we further compare the proposed GMVAE-VITS with baseline VITS.

4.2.3. *Subjective evaluation of overall system*

We conducted two mean opinion score (MOS) tests to evaluate the subjective quality of the synthetic speech[2]. In the utterance-level evaluation, raters were presented only one utterance and asked to evaluate its naturalness. In the dialogue-level evaluation, raters were presented a short dialogue consisting of 6 utterances (approximately 10–20 sec) and asked to evaluate its naturalness as a spoken dialogue (whether natural entrainment occurred, whether the speaking style was suitable for the context, etc.). We used ground truth timing of each utterance to construct dialogue samples because our spontaneous dialogue corpus contained numerous overlaps and simply playing the synthesized speech alternatively resulted in unnatural dialogue. For dialogue-level evaluation, we also computed text-speech alignment over recorded speech using MAS and used the alignment information for synthesis to align speech length with the original one. The evaluation was conducted on a 5-point scale from 1 (bad) to 5 (excellent). Thirty raters participated in the evaluation, and each rater evaluated thirty speech samples.

The results are presented in Table 3. With regard to utterance-level naturalness, although the scores of GMVAE-oracle/predicted were slightly higher than VITS, there was no significant difference between them. Though VITS does not utilize explicit style representation, the synthetic speech was evaluated as natural because various speaking styles exist that sound natural when heard as a single utterance. Regarding dialogue-level naturalness, the score of GMVAE-oracle was significantly higher than VITS ($p = 0.003$ in Student's t-test), confirming that using appropriate speech styles contributed to the naturalness of dialogue. Furthermore, GMVAE-predicted also achieved a significantly higher score than VITS ($p = 0.021$), indicating that style predictor was able to predict the appropriate speaking style when heard as a dialogue.

4.2.4. *Analysis of latent space*

Fig. 3 (a) and (b) illustrates the prior and posterior distribution of trained GMVAE-VITS, respectively, where dimensionality reduction was applied using principal component analysis. We observed that the latent representations of the two speakers were mixed, indicating that the learned latent space was speaker-independent. This is because the speaker embedding $\boldsymbol{s}$ was explicitly used, allowing the latent variable $\boldsymbol{z}_u$ to represent only speaker-independent speaking styles. We also synthesized all texts in the evaluation set with different speaker IDs and latent classes and described the average duration, pitch of voiced segments, and loudness of synthetic speech in Fig. 3 (c), (d), and (e). We confirmed that each latent class had different characteristics and they were common across speakers. With these characteristics, the learned prior distribution can be applied to modify speaking style to the desired one.

5. Conclusions

In this study, we aimed to synthesize spoken dialogue that is close to human spontaneous dialogue and proposed (1) recording and transcription of free-form dialogues without transcripts, (2) VAE/GMVAE-VITS to model various speaking styles, and (3) a style predictor that predicts speaking styles using linguistic and acoustic features from past dialogues. The combination of GMVAE-VITS and the style predictor achieved higher naturalness than conventional VITS in a dialogue-level evaluation. The latent space acquired by GMVAE-VITS was speaker-independent and had different characteristics for each latent class. This study assumed that transcriptions of past utterances and timing of each utterance were available; however, actual applications will require estimating these as well. Future work will include introducing a mechanism to automatically estimate them and unifying the proposed two-stage training framework into a single end-to-end training framework.

[2]Speech samples are available at : `https://rinnakk.github.io/research/publications/DialogueTTS`.

6. References

[1] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DialoGPT: Large-scale generative pre-training for conversational response generation," in *Proc. ACL*, online, Jul. 2020, pp. 270–278.

[2] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, "LaMDA: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, Jan. 2022.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, May 2018, pp. 4779–4783.

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

[5] S. Kita and S. Ide, "Nodding, aizuchi, and final particles in japanese conversation: How conversation reflects the ideology of communication and social relationships," *Journal of Pragmatics*, vol. 39, no. 7, pp. 1242–1254, 2007.

[6] A. Gravano, Š. Beňuš, R. Levitan, and J. Hirschberg, "Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement," in *Proc. SLT*, California, U.S.A., Dec. 2014, pp. 578–583.

[7] R. Levitan and J. B. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 3081–3084.

[8] R. Street, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[9] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, Jan. 2011.

[10] M. Yokoyama, T. Nagata, and H. Mori, "Effects of dimensional input on paralinguistic information perceived from synthesized dialogue speech with neural network," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 3053–3056.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, U.S.A., Jun. 2019, pp. 4171–4186.

[12] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end TTS for voice agents," in *Proc. SLT*, online, Jan. 2021, pp. 403–409.

[13] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style Tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, Stockholm, Sweden, Jul. 2018, pp. 5180–5189.

[14] J. Cong, S. Yang, N. Hu, G. Li, L. Xie, and D. Su, "Controllable context-aware conversational speech synthesis," in *Proc. INTERSPEECH*, online, Sep. 2021, pp. 4658–4662.

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

[16] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, vol. 33, online, Dec. 2020, pp. 8067–8077.

[17] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. ICLR*, New Orleans, U.S.A., May 2019.

[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, New Orleans, U.S.A., May 2019.

[19] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proc. CoNLL*, Berlin, Germany, Aug. 2016, pp. 10–21.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[21] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE PACRIM*, Victoria, Canada, May 1993, pp. 125–128.

[22] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. KDD Workshop*, vol. 10, Seattle, U.S.A., 1994, pp. 359–370.