Cross-Utterance Conditioned VAE for Non-Autoregressive Text-to-Speech

Anonymous ACL submission

Abstract

Modelling prosody variation is critical for synthesizing natural and expressive speech in end-to-end text-to-speech (TTS) systems. In this paper, a cross-utterance conditional VAE (CUC-VAE) is proposed to estimate a posterior probability distribution of the latent prosody features for each phoneme by conditioning on acoustic features, speaker information, and text features obtained from both past and future sentences. At inference time, instead of the standard Gaussian distribution used by VAE, CUC-VAE allows sampling from an utterance-specific prior distribution conditioned on cross-utterance information, which allows the prosody features generated by the TTS system to be related to the context and is more similar to how humans naturally produce prosody. The performance of CUC-VAE is evaluated via a qualitative listening test for naturalness, intelligibility and quantitative measurements, including word error rates and the standard deviation of prosody attributes. Experimental results on LJ-Speech and LibriTTS data show that the proposed CUC-VAE TTS system improves naturalness and prosody diversity with clear margins.

1 Introduction

Recently, abundant research have been performed on modelling variations other than the input text in synthesized speech such as background noise, speaker information, and prosody, as those directly influence the naturalness and expressiveness of the generated audio. Prosody, as the focus of this paper, collectively refers to the stress, intonation, and rhythm in speech, and has been an increasingly popular research aspect in end-to-end TTS systems (van den Oord et al., 2016; Wang et al., 2017; Stanton et al., 2018; Elias et al., 2021; Chen et al., 2021). Some previous work captured prosody features explicitly using either style tokens or variational autoencoders (VAEs) (Kingma and Welling, 2014; Hsu et al., 2019a) which encapsulate prosody information into latent representations. Recent work achieved fine-grained prosody modelling and control by extracting prosody features at phoneme or word-level (Lee and Kim, 2019; Sun et al., 2020a,b). However, the VAE-based TTS system lacks control over the latent space where the sampling is performed from a standard Gaussian prior during inference. Therefore, recent research (Dahmani et al., 2019; Karanasou et al., 2021) employed a conditional VAE (CVAE) (Sohn et al., 2015) to synthesize speech from a conditional prior. Meanwhile, pre-trained language model (LM) such as bidirectional encoder representation for Transformers (BERT) (Devlin et al., 2019) has also been applied to TTS systems (Hayashi et al., 2019; Kenter et al., 2020; Jia et al., 2021; Futamata et al., 2021; Cong et al., 2021) to estimate prosody attributes implicitly from pre-trained text representations within the utterance or the segment. Efforts have been devoted to include cross-utterance information in the input features to improve the prosody modelling of auto-regressive TTS (Xu et al., 2021).

To generate more expressive prosody, while maintaining high fidelity in synthesized speech, a cross-utterance conditional VAE (CUC-VAE) component is proposed, which is integrated into and jointly optimised with FastSpeech 2 (Ren et al., 2021), a commonly used non-autoregressive end-to-end TTS system. Specifically, the CUC-VAE TTS system consists of cross-utterance embedding (CU-embedding) and cross-utterance enhanced CVAE (CU-enhanced CVAE). The CU-embedding takes BERT sentence embeddings from surrounding utterances as inputs and generates phoneme-level CU-embedding using a multi-head attention (Vaswani et al., 2017) layer where attention weights are derived from the encoder output of each phoneme as well as the speaker information. The CU-enhanced CVAE is proposed to improve prosody variation and to address the inconsistency between the standard Gaussian prior, which the VAE-based TTS system is sampled from, and the true prior of

speech. Specifically, the CU-enhanced CVAE is a fine-grained VAE that estimates the posterior of latent prosody features for each phoneme based on acoustic features, cross-utterance embedding, and speaker information. It improves the encoder of standard VAE with an utterance-specific prior. To match the inference with training, the utterance-specific prior, jointly optimised with the system, is conditioned on the output of CU-embedding. Latent prosody features are sampled from the derived utterance-specific prior instead of a standard Gaussian prior during inference.

The proposed CUC-VAE TTS system was evaluated on the LJ-Speech read English data and the LibriTTS English audiobook data. In addition to the sample naturalness measured via subjective listening tests, the intelligibility is measured using word error rate (WER) from an automatic speech recognition (ASR) system, and diversity in prosody was measured by calculating standard deviations of prosody attributes among all generated audio samples of an utterance. Experimental results showed that the system with CUC-VAE achieved a much better prosody diversity while improving both the naturalness and intelligibility compared to the standard FastSpeech 2 baseline and two variants.

The rest of this paper is organised as follows. Section 2 introduces the background and related work. Section 3 illustrates the proposed CUC-VAE TTS system. Experimental setup and results are shown in Section 4 and Section 5, with conclusions in Section 6.

2 Background

Non-Autoregressive TTS. Promising progress has taken place in non-autoregressive TTS systems to synthesize audio with high efficiency and high fidelity thank to the advancement in deep learning. A non-autoregressive TTS system maps the input text sequence into an acoustic feature or waveform sequence without using the autoregressive decomposition of output probabilities. FastSpeech (Ren et al., 2019) and ParaNet (Peng et al., 2019) requires distillation from an autoregressive model, while more recent non-autoregressive TTS systems, including FastPitch (La'ncucki, 2021), AlignTTS (Zeng et al., 2020) and FastSpeech 2 (Ren et al., 2021), do not rely on any form of knowledge distillation from a pre-trained TTS system. In this paper, the proposed CUC-VAE TTS system is based on FastSpeech 2. FastSpeech 2 replaces the knowledge distillation for the length regulator in FastSpeech with mean-squared error training based on duration labels, which are obtained from frame-to-phoneme alignment to simplify the training process. Additionally, FastSpeech 2 predicts pitch and energy from the encoder output, which is also supervised with pitch contours and L2-norm of signal amplitudes as labels respectively. The pitch and energy prediction injects additional prosody information, which improves the naturalness and expressiveness in the synthesized speech.

Pre-trained Representation in TTS. It is believed that prosody can also be inferred from language information in both current and surrounding utterances (Shen et al., 2018; Fang et al., 2019; Xu et al., 2021; Zhou et al., 2021). Such information is often entailed in vector representations from a pre-trained LM, such as BERT (Devlin et al., 2019). Some existing work incorporated BERT embeddings at word or subword-level into autoregressive TTS models (Shen et al., 2018; Fang et al., 2019).More recent work (Xu et al., 2021) used the chunked and paired sentence patterns from BERT. Besides, a relational gated graph network with pre-trained BERT embeddings as node inputs (Zhou et al., 2021) was used to extract word-level semantic representations, thus enhancing expressiveness.

VAEs in TTS. VAEs have been widely adopted in TTS systems to explicit model prosody variation. The training objective of VAE is to maximise $p_\theta(\boldsymbol{x})$, the data likelihood parameterised by θ, which can be regarded as the marginalisation w.r.t. the latent vector $\boldsymbol{z}$ as shown in Eq. (1).

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}. \tag{1}$$

To make this calculation tractable, the marginalisation is approximated using evidence lower bound (ELBO):

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \\ - \beta D_{\text{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})\right), \end{aligned} \tag{2}$$

where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is the posterior distribution of the latent vector parameterized by ϕ, β is a hyperparameter, and $D_{\text{KL}}(\cdot)$ is the Kullback-Leibler divergence. The first term measures the expected reconstruction performance of the data from the latent vector and is approximated by Monte Carlo sampling of $\boldsymbol{z}$ according to the posterior distribution. The reparameterization trick is applied to make the sampling differentiable. The second term

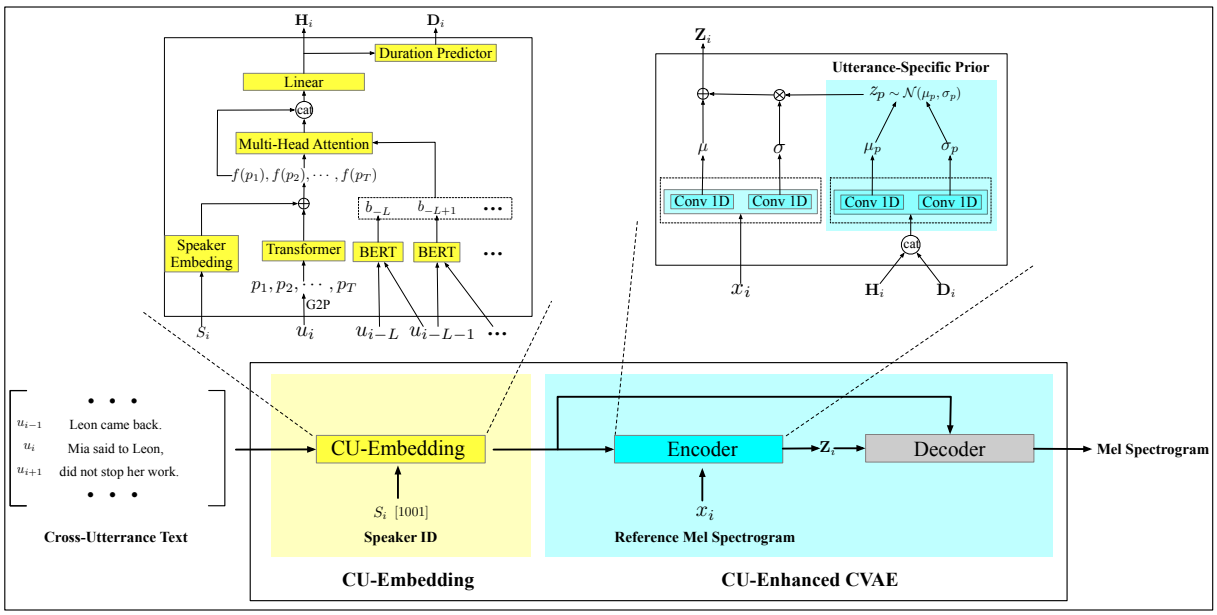


Figure 1: The CUC-VAE TTS system architecture consists of the cross-utterance embedding (CU-embedding) and the cross-utterance enhanced (CU-enhanced) CVAE, which are integrated to into and jointly optimised with the FastSpeech 2 system.

encourages the posterior distribution to approach the prior distribution which is sampled from during inference, and β weighs this term's contribution.

A large body of previous work on VAE-based TTS used VAEs to capture and disentangle data variations in different aspects in the latent space. Works by Akuzawa et al. (2018) leveraged VAE to model the speaking style of an utterance. Meanwhile, Hsu et al. (2019a,b) explored the disentanglement between prosody variation and speaker information using VAE together with adversarial training. Recently, fine-grained VAE (Sun et al., 2020a,b) was adopted to model prosody in the latent space for each phoneme or word. Moreover, vector-quantised VAE was also applied to discrete duration modelling by Yasuda et al. (2021).

CVAE is a variant of VAE when the data generation is conditioned on some other information $\boldsymbol{y}$. In CVAE, both prior and posterior distributions are conditioned on additional variables, and the data likelihood calculation is modified as shown below:

$$p_\theta(\boldsymbol{x} \mid \boldsymbol{y}) = \int p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y}) p_\phi(\boldsymbol{z} \mid \boldsymbol{y}) d\boldsymbol{z}. \quad (3)$$

Similar to VAE, this intractable calculation can be converted to the ELBO form as

$$\mathcal{L}_{\text{ELBO}}(\boldsymbol{x} \mid \boldsymbol{y}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{y})] - \beta D_{\text{KL}}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y}) \| p(\boldsymbol{z} \mid \boldsymbol{y})\right).$$

To model the conditional prior, a density network is usually used to predict the mean and variance based on the conditional input $\boldsymbol{y}$.

3 CUC-VAE TTS System

The proposed CUC-VAE TTS system, which is adapted from FastSpeech 2 as shown in Fig. 1, aims to synthesize speech with more expressive prosody. Fig. 1 describes the model architecture, which has two components: CU-embedding and CU-enhanced CVAE. The CUC-VAE TTS system takes as input $[\boldsymbol{u}_{i-L}, \cdots, \boldsymbol{u}_i, \cdots, \boldsymbol{u}_{i+L}]$, $\boldsymbol{s}_i$ and $\boldsymbol{x}_i$, where $[\boldsymbol{u}_{i-L}, \cdots, \boldsymbol{u}_i, \cdots, \boldsymbol{u}_{i+L}]$ is the cross-utterance set that includes the current utterance $\boldsymbol{u}_i$ and the L utterances before and after $\boldsymbol{u}_i$. Each $\boldsymbol{u}$ represents the text content of an utterance. Note that $\boldsymbol{s}_i$ is the speaker ID, and $\boldsymbol{x}_i$ is the reference mel-spectrogram of the current utterance $\boldsymbol{u}_i$. In this section, the two main components of the CUC-VAE TTS system will be introduced in detail.

3.1 Cross-Utterance Embedding

The CU-embedding encodes not only the phoneme sequence and speaker information but also cross-utterance information into a sequence of mixture encodings in place of a standard embedding. As shown in Fig. 1, the first L utterances and the last L utterances surrounding the current one, $\boldsymbol{u}_i$, are used as text input in addition to the current utterance and speaker information. Same as the standard embedding, an extra G2P conversion is first performed to convert the current utterance into phonemes $\boldsymbol{P}_i = [p_1, p_2, \cdots, p_T]$, where T is the number of phonemes. Then, a Transformer encoder is used to encode the phoneme sequence into a sequence of phoneme encodings. Besides, speaker

information is encoded into a speaker embedding $\boldsymbol{s}_i$ which is directly added to each phoneme encoding to form the mixture encodings $\boldsymbol{F}_i$ of the phoneme sequence.

$$\boldsymbol{F}_i = [\boldsymbol{f}_i(p_1), \boldsymbol{f}_i(p_2), \cdots, \boldsymbol{f}_i(p_T)], \quad (4)$$

where $\boldsymbol{f}$ represents resultant vector from the addition of each phoneme encoding and speaker embedding.

To supplement the text information from the current utterance to generate natural and expressive audio, cross-utterance BERT embeddings together with a multi-head attention layer are used to capture contextual information. To begin with, $2L$ cross-utterance pairs, denoted as $\boldsymbol{C}_i$, are derived from $2L+1$ neighboring utterances $[\boldsymbol{u}_{i-L}, \cdots, \boldsymbol{u}_i, \cdots, \boldsymbol{u}_{i+L}]$ as:

$$\boldsymbol{C}_i = [c(\boldsymbol{u}_{i-L}, \boldsymbol{u}_{i-L+1}), \cdots, c(\boldsymbol{u}_{i-1}, \boldsymbol{u}_i), \cdots, c(\boldsymbol{u}_{i+L-1}, \boldsymbol{u}_{i+L})], \quad (5)$$

where $c(u_k, u_{k+1}) = \{[\text{CLS}], \boldsymbol{u}_k, [\text{SEP}], \boldsymbol{u}_{k+1}\}$, which adds a special token [CLS] at the beginning of each pair and inserts another special token [SEP] at the boundary of each sentence to keep track of BERT. Then, the $2L$ cross-utterance pairs are fed to the BERT to capture cross-utterance information, which yields $2L$ BERT embedding vectors by taking the output vector at the position of the [CLS] token and projecting each to a 768-dim vector for each cross-utterance pair, as shown below:

$$\boldsymbol{B}_i = [\boldsymbol{b}_{-L}, \boldsymbol{b}_{-L+1}, \cdots, \boldsymbol{b}_{L-1}],$$

where each vector $\boldsymbol{b}_k$ in $\boldsymbol{B}_i$ represents the BERT embedding of the cross-utterance pair $c(\boldsymbol{u}_k, \boldsymbol{u}_{k+1})$. Next, to extract CU-embedding vectors for each phoneme specifically, a multi-head attention layer is added to combine the $2L$ BERT embeddings into one vector as shown in Eq. (6).

$$\boldsymbol{G}_i = \text{MHA}(\boldsymbol{F_i}\boldsymbol{W}^{\text{Q}}, \boldsymbol{B}_i\boldsymbol{W}^{\text{K}}, \boldsymbol{B}_i\boldsymbol{W}^{\text{V}}), \quad (6)$$

where $\text{MHA}(\cdot)$ denotes the multi-head attention layer, $\boldsymbol{W}^{\text{Q}}$, $\boldsymbol{W}^{\text{K}}$ and $\boldsymbol{W}^{\text{V}}$ are linear projection matrices, and $\boldsymbol{F}_i$ denotes the sequence of mixture encodings for the current utterance which acts as the query in the attention mechanism. For simplicity, we denote Eq. (6) as $\boldsymbol{G}_i = [\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_T]$ from the multi-head attention being of length T and each of them is then concatenated with its corresponding mixture encoding. The concatenated vectors are projected by another linear layer to form the final output $\boldsymbol{H}_i$ of the CU-embedding, $\boldsymbol{H}_i = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T]$ of the current utterance, as shown in Eq. (7).

$$\boldsymbol{h}_t = [\boldsymbol{g}_t, \boldsymbol{f}(p_t)]\boldsymbol{W}, \quad (7)$$

where $\boldsymbol{W}$ is a linear projection matrix. Moreover, an additional duration predictor takes $\boldsymbol{H}_i$ as inputs and predicts the duration $\boldsymbol{D}_i$ of each phoneme.

3.2 Cross-Utterance Enhanced CVAE

In addition to the CU-embedding, a CU-enhanced CVAE is proposed to conquer the lack of prosody variation of FastSpeech 2 and the inconsistency between the standard Gaussian prior distribution sampled by the VAE based TTS system and the true prior distribution of speech. Specifically, the CU-enhanced CVAE consists of an encoder module and a decoder module, as shown in Fig. 1. The utterance-specific prior in the encoder aims to learn the prior distribution $\boldsymbol{z}_p$ from the CU-embedding output $\boldsymbol{H}$ and predicts duration $\boldsymbol{D}$. For convenience, the subscript i is omitted in this subsection. Furthermore, the posterior module in the encoder takes as input reference mel-spectrogram $\boldsymbol{x}$, then model the approximate posterior $\boldsymbol{z}$ conditioned on utterance-specific conditional prior $\boldsymbol{z}_p$. Sampling is done from the estimated prior by the utterance-specific prior module and is reparameterized as:

$$\boldsymbol{z} = \boldsymbol{\mu} \oplus \boldsymbol{\sigma} \otimes \boldsymbol{z}_p, \quad (8)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are estimated from conditional posterior module to approximate posterior distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, $\boldsymbol{z}_p$ is sampled from the learned utterance-specific prior, and $\oplus, \otimes$ are elementwise addition and multiplication operation. Furthermore, the utterance-specific conditional prior module is conducted to learn utterance-specific prior with CU-embedding output $\boldsymbol{H}$ and $\boldsymbol{D}$. The reparameterization is as follows:

$$\boldsymbol{z}_p = \boldsymbol{\mu}_p \oplus \boldsymbol{\sigma}_p \otimes \boldsymbol{\epsilon}, \quad (9)$$

where $\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p$ are learned from the utterance-specific prior module, and $\boldsymbol{\epsilon}$ is sampled from the standard Gaussian $\mathcal{N}(0, 1)$. By substituting Eq. (9) into Eq. (8), the following equation can be derived for the total sampling process:

$$\boldsymbol{z} = \boldsymbol{\mu} \oplus \boldsymbol{\sigma} \otimes \boldsymbol{\mu}_p \oplus \boldsymbol{\sigma} \otimes \boldsymbol{\sigma}_p \otimes \boldsymbol{\epsilon}. \quad (10)$$

During inference, sampling is done from the learned utterance-specific conditional prior distribution $\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p)$ from CU-embedding instead of

a standard Gaussian distribution $\mathcal{N}(0, 1)$. For simplicity, we can formulate the data likelihood calculation as follows, where the intermediate variable utterance-specific prior $\boldsymbol{z}_p$ from $\boldsymbol{D}, \boldsymbol{H}$ to obtain $\boldsymbol{z}$ is omitted:

$$p_\theta(\boldsymbol{x} \mid \boldsymbol{H}, \boldsymbol{D}) = \int p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{H}, \boldsymbol{D}) p_\phi(\boldsymbol{z} \mid \boldsymbol{H}, \boldsymbol{D}) d\boldsymbol{z}, \tag{11}$$

In Eq. (11), ϕ, θ are the encoder and decoder module parameters of the CUC-VAE TTS system.

Moreover, the decoder in CU-enhanced CVAE is adapted from FastSpeech 2. An additional projection layer is firstly added to project $\boldsymbol{z}$ to a high dimensional space so that $\boldsymbol{z}$ could be added to $\boldsymbol{H}$. Next, a length regulator expands the length of input according to the predicted duration $\boldsymbol{D}$ of each phoneme. The rest of Decoder is same as the Decoder module in FastSpeech 2 to convert the hidden sequence into an mel-spectrogram sequence via parallelized calculation.

Therefore, the ELBO objective of the CUC-VAE can be expressed as,

$$\begin{aligned}\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{H}, \boldsymbol{D}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{D},\boldsymbol{H})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}, \boldsymbol{D}, \boldsymbol{H})] \\ - \beta_1 \sum_{n=1}^{T} D_{\mathrm{KL}}\left(q_{\phi_1}\left(\boldsymbol{z}^n \mid \boldsymbol{z}_p^n, \boldsymbol{x}\right) \| q_{\phi_2}\left(\boldsymbol{z}_p^n \mid \boldsymbol{D}, \boldsymbol{H}\right)\right) \\ - \beta_2 \sum_{n=1}^{T} D_{\mathrm{KL}}\left(q_{\phi_2}\left(\boldsymbol{z}_p^n \mid \boldsymbol{D}, \boldsymbol{H}\right) \| p(\boldsymbol{z}_p^n)\right),\end{aligned} \tag{12}$$

where ϕ_1, ϕ_2 are two parts of CUC-VAE encoder ϕ to obtain $\boldsymbol{z}$ from $\boldsymbol{z}_p, \boldsymbol{x}$ and $\boldsymbol{z}_p$ from $\boldsymbol{D}, \boldsymbol{H}$ respectively, β_1, β_2 are two balance constants, $p(\boldsymbol{z}_p^n)$ is chosen to be standard Gaussian $\mathcal{N}(0, 1)$. Meanwhile, $\boldsymbol{z}^n$ and $\boldsymbol{z}_p^n$ correspond to the latent representation for the n-th phoneme, and T is the length of the phoneme sequence.

4 Experimental Setup

4.1 Dataset

To evaluate the proposed CUC-VAE TTS system, a series of experiments were conducted on a single speaker dataset and a multi-speaker dataset. For the single speaker setting, the LJ-Speech read English data (Ito and Johnson, 2017) was used which consists of 13,100 audio clips with a total duration of approximately 24 hours. A female native English speaker read all the audio clips, and the scripts were selected from 7 non-fiction books. For the multi-speaker setting, the train-clean-100 and train-clean-360 subsets of the LibriTTS English audiobook data (Zen et al., 2019) were used. These subsets used here consist of 1151 speakers (553 female speakers and 598 male speakers) and about 245 hours of audio. All audio clips were re-sampled at 22.05 kHz in experiments for consistency.

The proposed CU-embedding in our system learns the cross-utterance representation from surrounding utterances. However, unlike LJ-Speech, transcripts of LibriTTS utterances are not arranged as continuous chunks of text in their corresponding book. Therefore, transcripts of the LibriTTS dataset were pre-processed to find the location of each utterance in the book, so that the first L and last L utterances of the current one can be efficiently obtained during training and inference. The pre-processed scripts and our code are available [1].

4.2 System Specification

The proposed CUC-VAE TTS system was based on the framework of FastSpeech 2. The CU-embedding utilised a Transformer to learn the current utterance representation, where the dimension of phoneme embeddings and the size of the self-attention were set to 256. To explicitly extract speaker information, 256-dim speaker embeddings were also added to the Transformer output. Meanwhile, the pre-trained BERT model to extract cross-utterance information had 12 Transformer blocks and 12-head attention layers with 110 million parameters. The size of the derived embeddings of each cross-utterance pair was 768-dim. Note that the BERT model and corresponding embeddings were fixed when training the TTS system. Network in CU-enhanced CVAE consisted of four 1D-convolutional (1D-Conv) layers with kernel sizes of 1 to predict the mean and variance of 2-dim latent features. Then a linear layer was added to transform the sampled latent feature to a 256-dim vector. The duration predictor which consisted of two convolutional blocks and an extra linear layer to predict the duration of each phoneme for the length regulator in FastSpeech 2 was adapted to take in CU-embedding outputs. Each convolutional block was comprised of a 1D-Conv network with ReLU activation followed by a layer normalization and dropout layer. The Decoder adopted four feed-forward Transformer blocks to convert hidden sequences into 80-dim mel-spectrogram sequence, similar to FastSpeech 2. Finally, HifiGAN (Kong et al., 2020) was used to synthesize waveform from the predicted mel-spectrogram.

[1] https://anonymous.4open.science/r/code-2708

4.3 Evaluation Metrics

In order to evaluate the performance of our proposed component, both subjective and objective tests were performed. First of all, a subjective listening test was performed over 11 synthesized audios with 23 volunteers asked to rate the naturalness of speech samples on a 5-scale mean opinion score (MOS) evaluation. The MOS results were reported with 95% confidence intervals. In addition, an AB test was conducted to compare the CU-enhanced CVAE with utterance-specific prior and normal CVAE with standard Gaussian prior. 23 volunteers were asked to choose the preference audio generated by different models in the AB test.

For the objective evaluation, F_0 frame error (FFE) (Chu and Alwan, 2009) and mel-cepstral distortion (MCD) (Kubichek, 1993) were used to measure the reconstruction performance of different VAEs. FFE combined the Gross Pitch Error (GPE) and the Voicing Decision Error (VDE) and was used to evaluate the reconstruction of the F_0 track. MCD evaluated the timbral distortion, which was computed from the first 13 MFCCs in our experiments. Moreover, word error rates (WER) from an ASR model trained on the real speech from the LibriTTS training set were reported. Complementary to naturalness, the WER metric showed both the intelligibility and the degree of inconsistency between synthetic speech and real speech. The ASR system used in this paper was an attention-based encoder-decoder model trained on Librispeech 960-hour data, with a WER of 4.4% on the test-clean set. Finally, the diversity of samples was evaluated by measuring the standard deviation of two prosody attributes of each phoneme: relative energy (E) and fundamental frequency (F_0), similar to Sun et al. (2020b). Relative energy was calculated as the ratio of the average signal amplitude within a phoneme to the average amplitude of the entire sentence, and fundamental frequency was measured using a pitch tracker. In this paper, the average standard deviation of E and F_0 of three phonemes in randomly selected 11 utterances was reported to evaluate the diversity of generated speech.

5 Results

This section presents the series of experiments for the proposed CUC-VAE TTS system. First, ablation studies were performed to progressively show the influence of different parts in the CUC-VAE TTS system based on MOS and WER. Next, the reconstruction performance of CUC-VAE was evaluated by FFE and MCD. Then, the naturalness and prosody diversity using CUC-VAE were compared to FastSpeech 2 and other VAE techniques. At last, a case study illustrated the prosody variations with different cross-utterance information as an example. The audio examples are available on the demo page [2].

5.1 Ablation Studies

Ablation studies in this section were conducted on the LJ-Speech data based on the subjective test and WER. First, to investigate the effect of the different number of neighbouring utterances, CUC-VAE TTS systems built with $L = 1, 3, 5$ were evaluated using MOS scores, as shown in Table 1.

Table 1: The MOS results of CUC-VAE TTS systems on LJ-Speech dataset. MOS was reported with 95% confident intervals. "$L = 1$","$L = 3$","$L = 5$" represented the number of past and future utterances.

Systems	Cross-utterance ($2L$)	**MOS**
CUC-VAE	$L = 1$	2.93 ± 0.12
CUC-VAE	$L = 3$	3.72 ± 0.09
CUC-VAE	$L = 5$	$\mathbf{3.95 \pm 0.07}$

The effect of the different number of neighbouring utterances on the naturalness of the synthesized speech can be observed by comparing MOS scores which is the higher the better. The CUC-VAE with $L = 5$ achieved highest score 3.95 compared to system with $L = 1$ and $L = 3$. Since only marginal MOS improvements were obtained using more than 5 neighbouring utterances, the rest of experiments were performed using $L = 5$.

Then we investigated the influence of each part of CUC-VAE on performance. The baseline was our implementation of Fastspeech 2. For the system denoted as Baseline + fine-grained VAE which served as a stronger baseline, the pitch predictor and energy predictor of FastSpeech 2 were replaced with a fine-grained VAE with 2-dim latent space. Based on the fine-grained VAE baseline, the CVAE was added without the CU-embedding to the system, referred to as Baseline+CVAE to verify the function of CVAE on the system, which conditions on the current utterance. Again, MOS was compared among these systems as shown in Table 2.

As shown in Table 2, MOS progressively increased when fine-grained VAE, CVAE, and CU-embedding were added in consecutively. The proposed CUC-VAE TTS system achieved the highest

[2] https://bit.ly/cuc-vae-tts-demo

Table 2: The MOS results of TTS systems with different modules on LJ-Speech dataset. MOS was reported with 95% confident intervals. Baseline + fine-grained VAE added a fine-grained VAE to baseline. Baseline+CVAE represents a CVAE TTS system without CU-embedding.

Systems	**MOS**
Ground Truth	4.31 ± 0.06
Baseline	3.85 ± 0.07
Baseline+Fine-grained VAE	3.55 ± 0.08
Baseline+CVAE	3.64 ± 0.08
CUC-VAE	$\mathbf{3.95 \pm 0.07}$

MOS 3.95 compared to baselines. The results indicated that CUC-VAE module played a crucial role in generating more natural audio.

To verify the importance of the utterance-specific prior to the synthesized audio, the same CUC-VAE system was used, and the only difference is whether to sample latent prosody features from the utterance-specific prior or from a standard Gaussian distribution. A subjective AB test was performed which required 23 volunteers to provide their preference between audios synthesized from the 2 approaches. Moreover, WER was also compared here to show the intelligibility of the synthesized audio. As shown in Table 3, the preference rate of using the utterance-specific prior is 0.52 higher than its counterpart, and a 4.9% absolute WER reduction was found, which confirmed the importance of the utterance-specific prior in our CUC-VAE TTS system.

Table 3: The subjective listening preference rate between CUC-VAE with or without utterance-specific prior from the AB test. The CUC-VAE without utterance-specific prior was a simplified version of our proposed CUC-VAE where latent samples were drawn from a standard Gaussian distribution instead of utterance-specific prior. WER metric was also reported.

System	utterance-specific prior	**RATE**	**WER**
CUC-VAE	✗	0.24	14.8
CUC-VAE	✓	**0.76**	**9.9**

5.2 Reconstruction Performance

FFE and MCD were used to measure the reconstruction performance of VAE systems. An utterance-level prosody modelling baseline which extract one latent prosody feature vector for an utterance was added for more comprehensive comparison, and is referred to as the Global VAE.

Table. 4 shows the reconstruction performance on the LJ-Speech dataset and LibriTTS dataset, respectively. Baseline had the highest value of FFE and MCD on the LJ-Speech dataset and LibriTTS dataset. The value of FFE and MCD decreased when the global VAE was added and was further reduced when the fine-grained VAE was added to the baseline. Our proposed CUC-VAE TTS system achieved the lowest FFE and MCD across the table on both the LJ-Speech and LibriTTS datasets. This indicated that richer prosody-related information entailed in both cross-utterance and conditional inputs was captured by CUC-VAE.

Table 4: Reconstruction preformance on LJ-Speech and LibriTTS dataset. + Global VAE and + fine-grained VAE represent that the baseline is added the global VAE and the fine-grained VAE, respectively.

Systems	**LJ-Speech**		**LibriTTS**	
	MCD	**FFE**	**MCD**	**FFE**
Baseline	6.70	0.58	6.32	0.58
Baseline+Global VAE	6.50	0.41	6.27	0.45
Baseline+Fine-grained VAE	6.34	0.26	6.28	0.35
CUC-VAE	**6.27**	**0.24**	**6.04**	**0.34**

5.3 Sample Naturalness and Diversity

Next, sample naturalness and intelligibility were measured using MOS and WER respectively on both LJ-Speech and LibriTTS datasets. Complementary to the naturalness, the diversity of generated speech from the conditional prior was evaluated by comparing the standard deviation of E and F_0 similar to (Sun et al., 2020b).

LJ-Speech experiments were shown in left part of Table. 5. Compared to the global VAE and fine-grained VAE, the proposed CUC-VAE received the highest MOS and achieved the lowest WER. Although both F_0 and E of the CUC-VAE TTS system were lower than the baseline + fine-grained VAE, the proposed system achieved a clearly higher prosody diversity than the baseline and baseline + global VAE systems. The fine-grained VAE achieved the highest prosody variation as its latent prosody features were sampled from a standard Gaussian distribution, which lacks the constraint of language information from both the current and the neighbouring utterances. This caused extreme prosody variations to occur which impaired both the naturalness and the intelligibility of synthesized audios. As a result, the CUC-VAE TTS system was able to achieve high prosody diversity without hurting the naturalness of the generated speech. In fact, the adequate increase in prosody diversity im-

Table 5: Sample naturalness and diversity results on LJ-Speech and LibriTTS datasets. Three metrics are reported for each dataset, namely MOS, WER, and Prosody Std. The Prosody Std. includes standard deviations of relative energy (E) and fundamental frequency (F_0) in Hertz within each phonene.

	LJ-Speech				LibriTTS			
	MOS	WER	Prosody Std.		MOS	WER	Prosody Std.	
			F_0	E			F_0	E
Ground Truth	4.31 ± 0.06	8.8	-	-	4.10 ± 0.07	5.0	-	-
Baseline	3.85 ± 0.07	10.8	1.86×10^{-13}	6.78×10^{-7}	3.53 ± 0.08	6.0	2.13×10^{-13}	7.22×10^{-7}
Baseline+Global VAE	3.82 ± 0.07	10.4	1.46	0.0004	3.59 ± 0.08	10.8	2.01	0.0054
Baseline+Fine-grained VAE	3.55 ± 0.08	12.8	49.60	0.0670	3.43 ± 0.08	5.6	63.64	0.0901
CUC-VAE	$\mathbf{3.95 \pm 0.07}$	**9.9**	26.35	0.0184	$\mathbf{3.63 \pm 0.08}$	**5.5**	30.28	0.0217

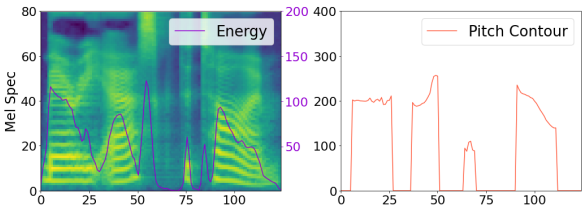
proved the expressiveness of the synthesized audio, and hence increased the naturalness.

The right part of Table. 5 showed the results on LibriTTS dataset. Similar to the LJ-Speech experiments, the CUC-VAE TTS system achieved the best naturalness measured by MOS, the best intelligibility measured by WER, and the second-highest prosody diversity across the table. Overall, consistent improvements in both naturalness and prosody diversity were observed on both single-speaker and multi-speaker datasets.

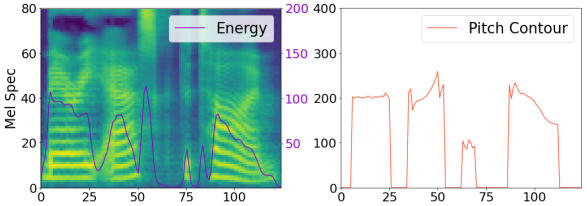
5.4 A Case Study

To better illustrate how the utterance-specific prior influenced the naturalness of the synthesized speech under a given context, a case study was performed by synthesizing an example utterance, "Mary asked the time", with two different neighbouring utterances: "Who asked the time? Mary asked the time." and "Mary asked the time, and was told it was only five." Based on the linguistic knowledge, to answer the question in the first setting, an emphasis should be put on the word "Mary", while in the second setting, the focus of the sentence is "asked the time". The model trained on LJ-Speech dataset was used to synthesize the utterance and the results were shown in Fig. 2.

Fig. 2 showed the energy and pitch of the two utterance. Energy of the first word "Mary" in Fig. 2(a) changed significantly (energy of "Ma-" was much higher than "-ry"), which reflected an emphasis on the word "Mary", whereas in Fig. 2(b), energy of "Mary" had no obvious change, i.e., the word was not emphasized. On the other hand, the fundamental frequency of words "asked" and "time" stayed at a high level for a longer time in the second audio than the first one, reflecting another type of emphasis on those words which was also coherent with the given context. Therefore, the difference of energy and pitch between the two utterances demonstrated that the speech synthesized by our model is sufficiently contextualized.



(a) Who asked the time? **Mary asked the time.**



(b) **Mary asked the time,** and was told it was only five.

Figure 2: Comparisons between the energy and pitch contour of same text "Mary asked the time" but different neighbouring utterances, generated by CUC-VAE TTS trained on LJ-Speech.

6 Conclusion

In this paper, a non-autoregressive CUC-VAE TTS system was proposed to synthesize speech with better naturalness and more prosody diversity. CUC-VAE TTS system estimated the posterior distribution of latent prosody features for each phone based on cross-utterance information in addition to the acoustic features and speaker information. The generated audio was sampled from an utterance-specific prior distribution, approximated based on cross-utterance information. Experiments were conducted to evaluate the proposed CUC-VAE TTS system with metrics including MOS, preference rate, WER, and the standard deviation of prosody attributes. Experiment results showed that the proposed CUC-VAE TTS system improved both the naturalness and prosody diversity in the generated audio samples, which outperformed the baseline in all metrics with clear margins.

References

K. Akuzawa, Yusuke Iwasawa, and Y. Matsuo. 2018. Expressive speech synthesis via modeling expressions with variational autoencoder. *ArXiv*, abs/1804.02135.

Liping Chen, Yan Deng, Xi Wang, F. Soong, and Lei He. 2021. Speech bert embedding for improving prosody in neural tts. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6563–6567.

Wei Chu and A. Alwan. 2009. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3969–3972.

Jian Cong, Shan Yang, Na Hu, Guangzhi Li, Lei Xie, and Dan Su. 2021. Controllable context-aware conversational speech synthesis. *ArXiv*, abs/2106.10828.

Sara Dahmani, Vincent Colotte, Valérian Girard, and S. Ouni. 2019. Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. In *INTERSPEECH*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Isaac Elias, H. Zen, Jonathan Shen, Yu Zhang, Jia Ye, R. Skerry-Ryan, and Yonghui Wu. 2021. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *ArXiv*, abs/2103.14574.

Wei Fang, Yu-An Chung, and J. Glass. 2019. Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models. *ArXiv*, abs/1906.07307.

Kosuke Futamata, Byeong-Cheol Park, Ryuichi Yamamoto, and Kentaro Tachibana. 2021. Phrase break prediction with bidirectional encoder representations in japanese text-to-speech synthesis. *ArXiv*, abs/2104.12395.

Tomoki Hayashi, Shinji Watanabe, T. Toda, K. Takeda, Shubham Toshniwal, and Karen Livescu. 2019. Pre-trained text embeddings for enhanced text-to-speech synthesis. In *INTERSPEECH*.

Wei-Ning Hsu, Y. Zhang, Ron J. Weiss, H. Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Z. Chen, Jonathan Shen, P. Nguyen, and Ruoming Pang. 2019a. Hierarchical generative modeling for controllable speech synthesis. *ArXiv*, abs/1810.07217.

Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James R. Glass. 2019b. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5901–5905.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Ye Jia, H. Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. 2021. Png bert: Augmented bert on phonemes and graphemes for neural tts. *ArXiv*, abs/2103.15060.

Panagiota Karanasou, S. Karlapati, A. Moinet, Arnaud Joly, Ammar Abbas, Simon Slangen, Jaime Lorenzo-Trueba, and Thomas Drugman. 2021. A learned conditional prior for the vae acoustic space of a tts system. *ArXiv*, abs/2106.10229.

Tom Kenter, Manish Sharma, and R. Clark. 2020. Improving the prosody of rnn-based english text-to-speech synthesis by incorporating a bert model. In *INTERSPEECH*.

Diederik P. Kingma and M. Welling. 2014. Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *ArXiv*, abs/2010.05646.

R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1:125–128 vol.1.

Adrian La'ncucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP*.

Younggun Lee and Taesu Kim. 2019. Robust and fine-grained prosody control of end-to-end speech synthesis. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5911–5915.

Kainan Peng, Wei Ping, Z. Song, and Kexin Zhao. 2019. Parallel neural text-to-speech. *ArXiv*, abs/1905.08459.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. *ArXiv*, abs/2006.04558.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, M. Schuster, Navdeep Jaitly, Zongheng Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. Skerry-Ryan, R. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *NIPS*.

Daisy Stanton, Yuxuan Wang, and R. Skerry-Ryan. 2018. Predicting expressive speaking style from text in end-to-end speech synthesis. *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 595–602.

G. Sun, Y. Zhang, Ron J. Weiss, Yuan Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Yonghui Wu. 2020a. Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6699–6703.

G. Sun, Y. Zhang, Ron J. Weiss, Yuanbin Cao, H. Zen, and Yonghui Wu. 2020b. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6264–6268.

Aäron van den Oord, S. Dieleman, H. Zen, K. Simonyan, Oriol Vinyals, A. Graves, Nal Kalchbrenner, A. Senior, and K. Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *SSW*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Yuxuan Wang, R. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Y. Xiao, Z. Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, R. Clark, and R. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *INTERSPEECH*.

Guanghui Xu, Wei Song, Zhengchen Zhang, C. Zhang, Xiaodong He, and Bowen Zhou. 2021. Improving prosody modelling with cross-utterance bert embeddings for end-to-end speech synthesis. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6079–6083.

Yusuke Yasuda, Xin Wang, and J. Yamagishi. 2021. End-to-end text-to-speech using latent duration based on vq-vae. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5694–5698.

H. Zen, Viet-Trung Dang, R. Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Z. Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *INTERSPEECH*.

Zhen Zeng, Jianzong Wang, Ning Cheng, Tian Xia, and Jing Xiao. 2020. Aligntts: Efficient feed-forward text-to-speech system without explicit alignment. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6714–6718.

Yixuan Zhou, C. Song, Jingbei Li, Zhiyong Wu, and H. Meng. 2021. Dependency parsing based semantic representation learning with graph neural network for enhancing expressiveness of text-to-speech. *ArXiv*, abs/2104.06835.