Towards controllable and high-fidelity Text-to-Speech from latent representation learning with Diffusion-based Refiner

Wenhao Guan[1], **Hukai Huang**[1], **Yishuang Li**[2], **Qi Su**[3], **Yinfei Li**[2]

[1]School of Informatics, Xiamen University
[2]Artificial Intelligence Research Institute, Xiamen University
[3]School of Electronic Science and Engineering, Xiamen University
{31520221154202, 31520221154205, 36920221153095, 23120221150302, 36920221153096}@stu.xmu.edu.cn,

Abstract

Recently, the TTS systems based on diffusion probabilistic models have demonstrated the strong ability to generate state-of-the-art results on many TTS datasets. However, most of them are very slow at the inference phase and can not explore the interpretable representation of latent space. On the other hand, another type of generative models named Variational Autoencoders(VAEs) can manipulate the low-dimensional latent space but often obtain low-quality results. In this paper, we incorporate the VAE framework and diffusion model to an end-to-end Text-to-Speech system, so that it can not only obtain good results but also has the ability to manipulate the latent representations. Finally, the proposed model shows good performance of generated results and style control.

Introduction

With the rapid development of deep learning, End-to-end text-to-speech models which generate speech directly from characters or phonemes have made great progress (Wang et al. 2017; Shen et al. 2018; Ping et al. 2018; Ping, Peng, and Chen 2019; van den Oord et al. 2016; Prenger, Valle, and Catanzaro 2019). These models usually consists of two parts, namely acoustic model and vocoder. The acoustic model transforms normalized text symbols to time-aligned features, such as mel-spectrogram, while the vocoder transforms time-aligned features to audio samples.

The auto-regressive (AR) model, such as Tacotron (Wang et al. 2017) and Tacotron2 (Shen et al. 2018), is a typical framework of autoregressive modelling, which takes character or phoneme sequences as input and generates intermediate representation frame by frame. The AR models can achieve a high speech quality but suffer from a low decoding speed because of the nature of autoregressive modelling. The other family is the non-autogressive model. These models can speed up the inference process by utilizing parallel spectrogram generation. FastSpeech (Ren et al. 2019) uses a well-trained autoregressive teacher model to guide the training process and learn the alignment between text and speech. FastSpeech2 (Ren et al. 2020) utilizes an external force aligner to extract durations. Flow based TTS models are a family of non-autoregressive TTS models, which transform a simple initial density into a complex one by applying a series of invertible transformations. One group of models are based on autoregressive transformations, including autoregressive flow (AF) and inverse autoregressive flow (IAF) (Kingma et al. 2016; Papamakarios, Pavlakou, and Murray 2017; Huang et al. 2018). AF is similar to autoregressive models, which performs parallel density evaluation and sequential synthesis. IAF performs parallel synthesis but sequential density evaluation contrastively. Parallel WaveNet (Oord et al. 2018) and ClariNet (Ping, Peng, and Chen 2019) distill an IAF from a pretrained autoregressive WaveNet, which complicate the training process and increases the cost of development. Another group of flow based models are based on bipartite transformations (Dinh, Sohl-Dickstein, and Bengio 2016; Kingma and Dhariwal 2018), which provide likelihood based training and parallel synthesis. These models can generate speech efficiently.

Another generative framework called denoising diffusion probabilistic model has shown state-of-the-art performance in many fields, such as image generation and audio synthesis (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song et al. 2020; Huang et al. 2022c; Jeong et al. 2021). The denoising diffusion models can be stably optimized according to maximum likelihood and also enjoy the freedom of architecture choices, which is the limited in the flow based TTS models.

In the aspect of improving the expressiveness of synthesized speech, Variational Autoencoder (VAE) (Kingma and Welling 2013), which explicitly models latent variables, is one of the most popular approaches to learn disentangled latent representations of speech. VAE-Loop (Akuzawa, Iwasawa, and Matsuo 2018) and VAE-TTS (Zhang et al. 2019) both ultilize the VAE framework to learn the latent representations of original speech, but use the VoiceLoop (Taigman et al. 2018) and Tacotron2 (Shen et al. 2018) as the decoder respectively in an autoregressive manner. FHVAE (Hsu, Zhang, and Glass 2017) and GMVAE-Tacotron (Hsu et al. 2018) apply the graphical model and VAE to model more fine-grained latent representations of speech under the Bayesian theory. These works demonstrate that VAE has the ability of learning disentanglement of latent representations, and furthermore can interpolate or sampling between latent representations.

In this work, we combine the properties of diffusion model and VAE to make the the final model high-quality,

easy to train and expressive. We propose VAEDiff-TTS, a non-autoregressive TTS model that integrates VAE framework and Diffusion model to generate high quality results and enable manipulation in the latent space. In order to improve the sampling speed of DDPM, we also introduce the accelerated sampling method (Song, Meng, and Ermon 2020) to this system. The contributions of our work are as follows:

- To the best of our knowledge, it is the first time that a VAE framework incorporated with a denoising diffusion probabilistic model was applied to speech synthesis.
- We show that VAEDiff-TTS generates comparable high fidelity audios in terms of Mean Opinion Score (MOS) compared to VAE-TTS, Diff-TTS and ProDiff.
- We analyze the style control and style transfer of VAEDiff-TTS. VAEDiff-TTS can effectively control the latent representations and perform style transfer.

Related work

Text-to-Speech Models. Text-to-Speech (TTS) is an essential component of intuitive human-machine communication system. TTS system can generate an output acoustic sequence given an input text sequence. Concatenative TTS (BLACK 1997) and statistical parametric TTS (Tokuda et al. 2000; Ze, Senior, and Schuster 2013) are the two most successful TTS techniques in the past decades. However, both of them have complex pipelines and and the speech generated often sounds unnatural. With the continuous development of deep learning, a lot of end-to-end TTS systems have been proposed to achieve high-fidelity speech synthesis. These models usually consist of two parts namely acoustic model and vocoder respectively. The acoustic model first generates frame-level intermediate representations given text, while the vocoder generates the audio samples conditioned on the intermediate representations. We focus on the first part, acoustic modelling, in our work.

Diffusion-based Generative Models. Recently, the generative model based on diffusion models have attracted much attention in image and speech generation. Diffusion models can be divided into two categories: discrete time diffusion model, which based on denoising diffusion probabilistic model (DDPM) (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) and continuous time diffusion model, which based on stochastic differential equation (SDE) (Song et al. 2020). This paper mainly discusses on DDPM. In recent years, the TTS system based on DDPM mainly focuses on two aspects, Vocoder and Acoustic model. Vocoder mainly includes FastDiff (Huang et al. 2022a), Diffwave (Kong et al. 2020), etc. Acoustic model mainly includes ProDiff (Huang et al. 2022c), Diff-TTS (Jeong et al. 2021), Grad-TTS (Popov et al. 2021), etc. However, most of these systems mainly utilize the advantages of DDPM itself to make the final results achieve better quality without exploring the latent space. This paper proposes that VAE can be used to model the attributes of data latent representations to mine the late space, and DDPM can be used to refine, thus combining their advantages respectively, generating good results while allowing style control, and analyzing the impact of DDPM on the latent space.

VAE-based Generative Models. Variational autoencoder (VAE) (Kingma and Welling 2013) has been applied for latent representation learning of natural speech for years. It models either the generative process of raw waveform, or spectrograms. In previous work, autoregressive networks are employed as the decoder of VAE (Akuzawa, Iwasawa, and Matsuo 2018; Zhang et al. 2019), but they can be quite slow at synthesis. In this work, we employ FastSpeech2 (Ren et al. 2020) as the decoder of VAE, so that it can not only synthesize speech in parallel, but also attain disentangled representations of speech and control the synthesized speech.

Style Controlling and Transferring. There are already several methods to style controlling and transferring in TTS field. We can classify them to two categories. One is the global style modelling, such as global style token (GST) (Wang et al. 2018) and VAE-TTS (Zhang et al. 2019). The other is based on fine-grained latent variables (Sun et al. 2020a,b). GenerSpeech (Huang et al. 2022b) proposed a multi-level style adaptor and a generalizable content adapter to model style-agnostic and style-specific variations separately. Although the fine-grained latent variables can model more elaborate style and then do style transfer better, they do not have the ability to explore the latent space. So, We utilize VAE to have access to the latent space , and integrate with a diffusion based refiner to obtain high-quality results.

Model

In this section, we first review Variational Autoencoder and Denoising Diffusion Probabilistic Model. And then, we discuss the architecture design of our VAEDiff-TTS and the entire training pipeline in detail. Overall, we design a non-autoregressive two-phase neural network architecture for VAEDiff-TTS, which first trains a VAE based controllable TTS and secondly trains the DDPM by refining the results from first phase. We also analyzed the influence of VAE and DDPM latent space respectively. By combining the advantages of variational autoencoder and diffusion probabilistic model, the final system is controllable and high-fidelity.

Variational Autoencoder

The goal of generative machine learning approaches is to model the data distribution $p(x)$. The Variational Autoencoder (VAE) does so by learning to reconstruct input data from a compressed latent code. An underlying idea of the model is that real world data can be represented by a relatively small set of higher level features. It is assumed that the observed data distribution $p(x)$ is generated by some random process from a random latent variable z. The true posterior distribution $p_\theta(z|x)$ is intractable because of the indifferentiable marginal likelihood $p_\theta(x)$. To address this problem, a recognition model $q_\phi(z|x)$ is introduced as an approximation to the true posterior distribution $p_\theta(z|x)$. Finally, we can get the formulation of $logp_\theta(x)$ as shown in equation (1).

$$
\begin{aligned}
logp_\theta(x) &\geq E_{q_\phi(z|x)}[log\frac{p_\theta(x,z)}{q_\phi(z|x)}] \\
&= E_{q_\phi(z|x)}[logp_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z))
\end{aligned} \tag{1}
$$

The first term in Equation (1) measures the reconstruction likelihood of the decoder from the variational distribution; this ensures that the learned distribution is modelling effective latents that the original data can be regenerated from. The second term measures how similar the learned variational distribution is to a prior belief held over latent variables, minimizing this term encourages the encoder to actually learn a distribution rather than collapse into a Dirac delta function. The encoder of the VAE is commonly chosen to model a multivariate Gaussian with diagonal covariance, and the prior is often selected to be a standard multivariate Gaussian:

$$
q_\phi(z|x) = N(z; \mu_\phi(x), \sigma^2_\phi(x)I) \tag{2}
$$

$$
p(z) = N(z; 0, I) \tag{3}
$$

In practice, $\mu(x)$ and $\sigma^2(x)$ are learned from observed dataset via neural network which can be viewed as an encoder. Because each z is generated by a stochastic sampling procedure, which is generally non-differentiable, the reparameterization trik is introduced to VAE framework. Thus, each z is computed as a deterministic function of input x and auxiliary noise variable ϵ, where $\odot$ represents an element-wise product.

$$
z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon \tag{4}
$$

After training a VAE, generating new data can be performed by sampling directly from the latent space and then running it through the decoder.

Denoising Diffusion Probabilistic Models

The concept of diffusion was first defined in (Sohl-Dickstein et al. 2015) and then researchers proposed DDPM (Ho, Jain, and Abbeel 2020) which greatly promotes the development of generative models. Diffusion process and reverse process are given by diffusion probabilistic models, which could be used for the denoising neural networks θ to learn data distribution. With the predefined fixed noise schedule β and diffusion step t, we compute the corresponding constants respective to diffusion and reverse processes:

$$
\alpha_t = 1 - \beta_t, \quad \bar{\alpha_t} = \prod_{t=1}^{t} \alpha_t, \tag{5}
$$

Similar as previous work (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020), we define the data distribution as $q(x_0)$. The diffusion process is defined by a fixed Markov chain from data x_0 to the latent variable x_T:

$$
q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{6}
$$

$$
q(x_1, \cdots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{7}
$$

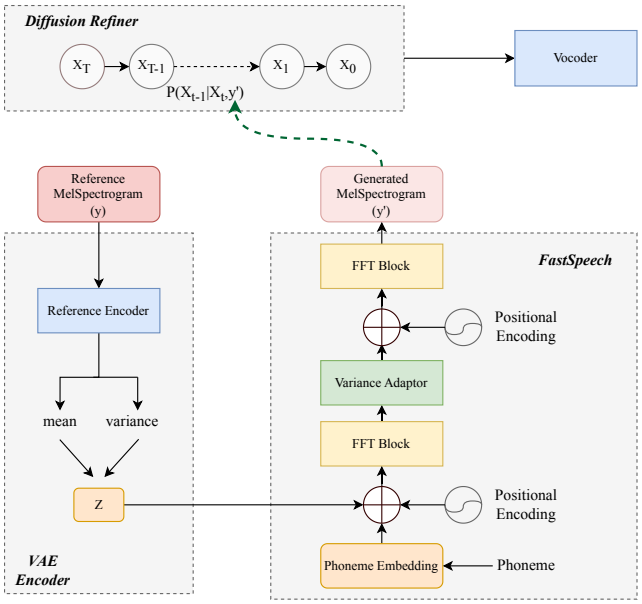

Figure 1: The structure of VAEDiff-TTS.

The reverse process aims to recover samples from Gaussian noises, which is Markov chain from x_T to x_0 parameterized by shared θ:

$$
p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \tag{8}
$$

$$
p_\theta(x_0, \cdots, x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{9}
$$

Finally, we can get the training objective as follows:

$$
L_{DDPM} = E_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||_2^2] \tag{10}
$$

For sampling phase, the sampling formulation is computed as follows:

$$
\begin{aligned}
x_{t-1} &= \mu_\theta(x_t, t) + \sigma_t z \\
&= \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z,
\end{aligned} \tag{11}
$$

where $p(z) = N(z; 0, I)$ and $\sigma_t = \sqrt{\frac{1-\bar{\alpha_{t-1}}}{1-\bar{\alpha_t}}\beta_t}$. As a result, the final data distribution $p(x_0)$ is obtained through iterative sampling over all of the time steps.

Proposed VAEDiff-TTS

Our proposed VAEDiff-TTS inherits all the advantages of FastSpeech, including fast speech synthesis speed and few words skipping problems. In order to explore the latent space of speech and control style unsupervisedly, we introduce VAE framework to FastSpeech to have access to the latet space. By introducing the diffusion based refiner, the final system of VAEDiff-TTS could generate comparable results with some other good systems.

The structure of VAEDiff-TTS is shown in Fig.1. The proposed architecture is divided into three parts. First, we describes the VAE framework of our VAEDiff-TTS. Here ,we

adopt a recurrent reference encoder followed by two fully connected layers to model the values of mean and variance. We use the same architecture for reference encoder first proposed in (Skerry-Ryan et al. 2018) which consists of six 2-D convolutional layers followed by a GRU layer. The output, which denotes some embedding of the reference audio, is then passed through two separate fully connected layers to generate the mean and variance of latent variables z. Then z is derived by reparameterization trick. And then, z should be passed through a fully connected layer to make sue the dimension equal to text encoder. Second, we describes the Acoustic model of our system. We utilize FastSpeech (Ren et al. 2019) as our acoustic model which generate mel-spectrograms from characters or phonemes. The difference is that we have to receive the embedding of z from the first part, we add z to the embedding of text encoder, and we use Montreal forced aligner (McAuliffe et al. 2017) (MFA) instead of the attention based alignment teacher in FastSpeech. The FFT block and Variance adapter have the same architecture as FastSpeech. Third, we describes the proposed diffuison based refiner. The model architecture of diffusion based refiner is mainly to model the reverse process of DDPM, we utilize the U-Net model from (Nichol and Dhariwal 2021) to model our diffusion based refiner. Finally, the results generated from diffusion based refiner are passed through FastDiff Vocoder (Huang et al. 2022a) to reconstruct the waveform. As for the training objective, inspired by (Pandey et al. 2022), the total loss of our proposed model is shown in equation (12):

$$\begin{aligned} Loss = & E_{q_\psi(z|x_0)}[p_\theta(x_0|z)] - D_{KL}(q_\psi(z|x_0)||p(z)) \\ & + E_{q(z|x_0)}[E_{q(x_{1:T}|\hat{x}_0,x_0)}[\frac{p_\phi(x_{0:T}|\hat{x}_0)}{q(x_{1:T}|\hat{x}_0,x_0)}]] \end{aligned} \tag{12}$$

where, x_0 represents the original data, $\hat{x}_0$ represents the data generated from VAE framework. ψ and ϕ represents the parameters of VAE encoder and reverse process of DDPM respectively.

Experiments and Analysis

Experiment Setup

As for dataset, We used VCTK Corpus (Yamagishi et al. 2019) (VCTK), which was recorded by 109 native English speakers with various accents. There are 44081 audio clips through preprocessing, we use 43001 samples from it for training and remaining 1080 utterances for testing and half of test set for valid set. 80 dimensional mel-spectrograms were extracted with frame length 64 ms and frame shift 16 ms.

We used GST and VAE-FastSpeech as our baseline model for style control and transfer and FastSpeech as our baseline model for generated quality comparison, where VAE-FastSpeech simply modifies the acoustic model to FastSpeech from VAE-Tacotron. The hyperparameters are set according to (Ren et al. 2019), the noise schedule in the DDPM forward process was set to a linear schedule between $\beta_1 = 10^{-4}$ and $\beta_2 = 0.02$ during training, the dimension of latent variable z in VAE was set to 32.

At inference stage, in evaluation of style control, we directly manipulate z without going through the VAE encoder, and then feed manipulated z to the remaining model. With regard to evaluation of style transfer, we feed audio clips as reference and go through the whole model. Both parallel and non-parallel style transfer audios are generated and evaluated. And we analyzed the latent space of VAE and DDPM respectively.

Speech Synthesis Quality

For the subjective evaluation of audio fidelity, we performed a 5-scale Mean Opinion Score (MOS) test with 30 audio examples per model and 12 participants. The audio examples were randomly selected from the test dataset. The second phase of VAEDiff-TTS, which is also called diffusion based refiner (Refiner), was trained with 100 time-step and synthesised samples by accelerating the sampling speed by DDIM (Song, Meng, and Ermon 2020). The performance is as good as Diff-TTS abd VITS (Kim, Kong, and Son 2021). It indicates that accelerated sampling is a practical method without significantly sacrificing speech quality.

Table 1: Comparison with other text-to-speech models in terms of synthesized quality

Method	***MOS***
GT	4.51
GT(Mel+FastDiff)	4.33
FastSpeech	4.06
VITS	4.38
Diff-TTS	4.35
ProDiff	4.17
VAEDiff-TTS(T=100)	5
VAEDiff-TTS(T=50)	5
VAEDiff-TTS(T=10)	5

Style Control

It is known from (Bowman et al. 2015) that VAE supports smmothly interpolation and continuous sampling between latent representations. And DDPM can be considered as a special form of VAE (Luo 2022). Thus, the proposed VAEDiff-TTS model consists of two types of latent representations: the low-dimensional VAE latent code z_{vae} and the DDPM intermediate representations $x_{1:T}$ associated with the DDPM reverse process. We analyze the effects of manipulating both z_{vae} and z_T.

We first do interpolation in the VAE latent space z_{vae}. We got two VAE latent code z^1_{vae} and z^2_{vae} by feeding two different audios to the VAE encoder. We then perform linear interpolation between z^1_{vae} and z^2_{vae} to obtain intermediate VAE latent codes $\widetilde{z}_{vae} = \lambda z^1_{vae} + (1-\lambda) z^2_{vae} (0 < \lambda < 1)$, which are then fed into the remaining model to generate corresponding controlled samples.

We then do interpolation in the DDPM latent space with fixed z_{vae}. We got a VAE latent code z_{vae} the same way as above. With a fixed z_{vae}, we then sample two initial DDPM representations x^1_T and x^2_T from $p(x_T)$. We then perform

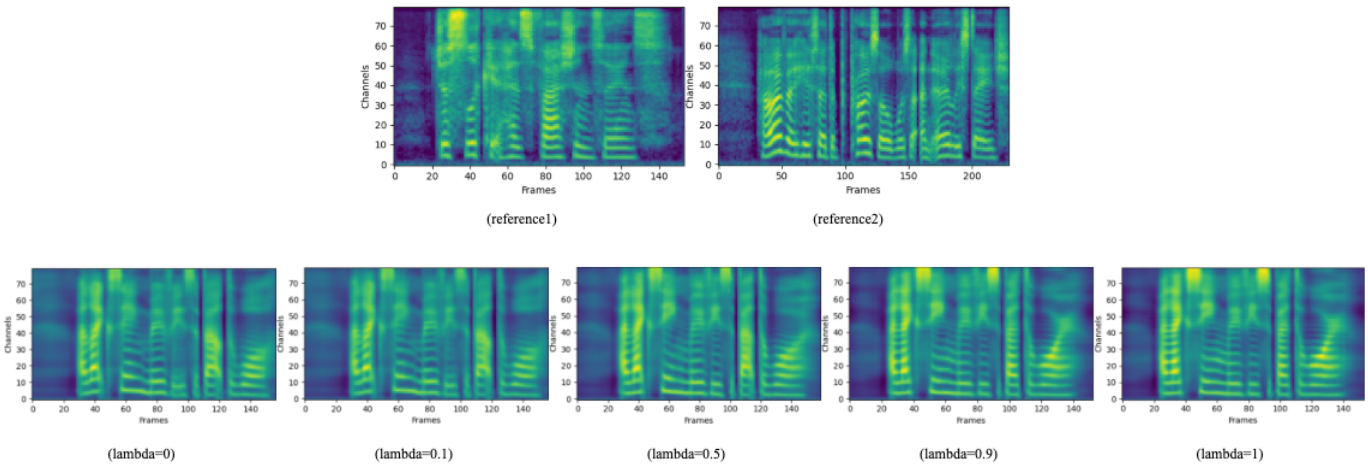

Figure 2: The results of Style control. two z were obtained by reference 1 and reference 2 in the first row. The second row shows that mel-spectrograms generated by interpolation by two z in different lambda.

linear interpolation the same way between x_T^1 and x_T^2 with a fixed z_{vae} to generate interpolated samples.

We also infer that VAE and Diffusion both have the potential of exploring latent space, and VAE control the global information as well as Diffusion control the local details from the experiment results.

Style Transfer

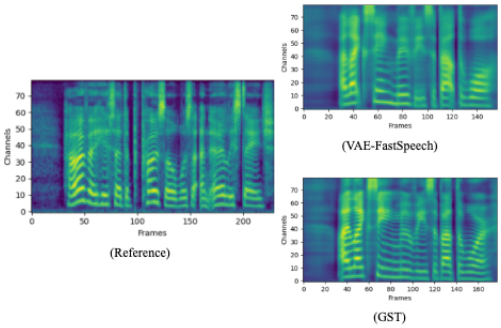

Figure 3: The results of Style transfer. The left picture describes the mel-spectrogram of reference speech. The right pictures describe the mel-spectrogram referenced on the left reference speech in VAE-FastSpeech and GST respectively.

As for style transfer, we could conduct our experiments for two types: Parallel and Non-Parallel style transfer, which are categorized by the text consistency between reference and generated speech samples.

We evaluated the performance of style transfer subjectively by conducting a crowd-sourcing ABX preference tests on parallel and non-parallel transfer. For parallel style transfer, 30 audio clips with their texts are randomly selected from test set. For non-parallel style transfer, 30 sentences of text and 30 other reference audio clips are randomly selected to generate speech. The baseline voice is generated from the GST model and VAE-FastSpeech model we have built. Each case in ABX test is judged by 12 juders. The criterion in rating is "which one's speaking style is closer to the reference style", for each reference, the listeners were asked to choose a prefered one among the samples synthesized by baseline models and proposed VAEDiff-TTS.

Table 2: The ABX preference test results for parallel style transfer and non-parallel style transfer.

Method	Parallel	Non-Parallel
GST	100%	100%
Neutral	100%	100%
VAEDiff-TTS(T=100)	100%	100%
VAE-FastSpeech	100%	100%
Neutral	100%	100%
VAEDiff-TTS(T=100)	100%	100%

The results show that our proposed VAEDiff-TTS can better model the latent representations, which results in better style transfer.

Conclusion

In this work, we presented a speech synthesis model that integrates the properties of VAE and diffusion models, enabling the system to generate samples of better quality and provide a DDPM latent space except for the VAE latent space, which gives us the way to control the latent space by using diffusion models. We have demonstrated the latent space interpolation and style transfer. The proposed model shows good performance in synthesized quality, style control and style transfer, which outperforms GST model and VAE-FastSpeech via ABX test.

Notice

Due to time, the experiment was not completed completely, so the rest of Table 1 and Table 2 could not get the exper-

imental results. There is also a lack of some comparative experiments on style control and transfer.

References

Akuzawa, K.; Iwasawa, Y.; and Matsuo, Y. 2018. Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder. *Proc. Interspeech 2018*, 3067–3071.

BLACK, A. 1997. Automatically clustering similar units for unit selection in speech synthesis. *Proc. EUROSPEECH, Sep 1997*.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hsu, W.-N.; Zhang, Y.; and Glass, J. 2017. Unsupervised learning of disentangled and interpretable representations from sequential data. *Advances in neural information processing systems*, 30.

Hsu, W.-N.; Zhang, Y.; Weiss, R. J.; Zen, H.; Wu, Y.; Wang, Y.; Cao, Y.; Jia, Y.; Chen, Z.; Shen, J.; et al. 2018. Hierarchical Generative Modeling for Controllable Speech Synthesis. In *International Conference on Learning Representations*.

Huang, C.-W.; Krueger, D.; Lacoste, A.; and Courville, A. 2018. Neural autoregressive flows. In *International Conference on Machine Learning*, 2078–2087. PMLR.

Huang, R.; Lam, M. W.; Wang, J.; Su, D.; Yu, D.; Ren, Y.; and Zhao, Z. 2022a. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. *arXiv preprint arXiv:2204.09934*.

Huang, R.; Ren, Y.; Liu, J.; Cui, C.; and Zhao, Z. 2022b. GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech Synthesis. *arXiv preprint arXiv:2205.07211*.

Huang, R.; Zhao, Z.; Liu, H.; Liu, J.; Cui, C.; and Ren, Y. 2022c. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2595–2605.

Jeong, M.; Kim, H.; Cheon, S. J.; Choi, B. J.; and Kim, N. S. 2021. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*.

Kim, J.; Kong, J.; and Son, J. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, 5530–5540. PMLR.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Luo, C. 2022. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*.

McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; and Sonderegger, M. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, volume 2017, 498–502.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Oord, A.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; Driessche, G.; Lockhart, E.; Cobo, L.; Stimberg, F.; et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, 3918–3926. PMLR.

Pandey, K.; Mukherjee, A.; Rai, P.; and Kumar, A. 2022. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*.

Papamakarios, G.; Pavlakou, T.; and Murray, I. 2017. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30.

Ping, W.; Peng, K.; and Chen, J. 2019. ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech. In *International Conference on Learning Representations*.

Ping, W.; Peng, K.; Gibiansky, A.; Arik, S. Ö.; Kannan, A.; Narang, S.; Raiman, J.; and Miller, J. 2018. Deep Voice 3: 2000-Speaker Neural Text-to-Speech.

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, 8599–8608. PMLR.

Prenger, R.; Valle, R.; and Catanzaro, B. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3617–3621. IEEE.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2020. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4779–4783. IEEE.

Skerry-Ryan, R.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.; Clark, R.; and Saurous, R. A. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, 4693–4702. PMLR.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

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

Sun, G.; Zhang, Y.; Weiss, R. J.; Cao, Y.; Zen, H.; and Wu, Y. 2020b. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 6264–6268. IEEE.

Taigman, Y.; Wolf, L.; Polyak, A.; and Nachmani, E. 2018. VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop. In *International Conference on Learning Representations*.

Tokuda, K.; Yoshimura, T.; Masuko, T.; Kobayashi, T.; and Kitamura, T. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, 1315–1318. IEEE.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop*, 125–125.

Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Wang, Y.; Stanton, D.; Zhang, Y.; Ryan, R.-S.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; and Saurous, R. A. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, 5180–5189. PMLR.

Yamagishi, J.; Veaux, C.; MacDonald, K.; et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

Ze, H.; Senior, A.; and Schuster, M. 2013. Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, 7962–7966. IEEE.

Zhang, Y.-J.; Pan, S.; He, L.; and Ling, Z.-H. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6945–6949. IEEE.