TOWARDS A CUSTOMIZABLE TEXT-TO-SPEECH PERSONAL ASSISTANT

Chenyu Shi	**Siwen Tu**	**Shuang Fan**	**Shupei Li**
s3500063	s3631400	s3505847	s3430863

ABSTRACT

The audio assistant technology has been widely applied based on the high development of the deep learning field. In this project, we have used deep learning models to accomplish a customizable text-to-speech personal audio assistant. We trained the neural networks on audio datasets, and fine-tune them on different languages of multi speakers. The final model can produce realistic audio based on input texts, and users can customize the type of languages or speakers according to their choice.

***K*eywords** VITS · TTS · Deep Learning

1 Introduction

With the rapid development of the deep learning field, the audio assistant technology has achieved great success and been widely applied in our daily life. Most of audio assistants are realized by text-to-speech models. In this project, we accomplished a customizable text-to-speech personal audio assistant based on VITS [1], an efficient text-to-speech model with high performance. We trained the model on audio datasets, and fine-tune it on Mandarin and Japanese to test the performance of the model on different languages. Besides, we also fine-tune it on multiple types of speakers to fulfill the requirement of customization. The final model achieves great performance, which can produce realistic audio from text, and the users can make their choice on different languages and types of speakers.

The rest of the reports will be presented in three chapters. In chapter 2, we will show and explain the methodology used in the models. In chapter 3, the experiment results will be displayed and discussed. And in chapter 4, we will have a conclusion of the report and project.

2 Methodology

We describe VITS [1], the backbone of our TTS assistant, from the perspective of variational inference in detail. We also introduce VITS's architecture and training process to provide a comprehensive overview.

2.1 CVAEs

The variational autoencoder (VAE), proposed in [2], is a popular approach widely used in unsupervised learning. It is an effective function approximator and can be optimized by the standard stochastic gradient descent (SGD) method. It has shown promising application value in various fields, such as complex pattern recognition, segmentation, future prediction from static images, etc [3].

Conditional variational autoencoders (CVAEs) are variants of VAEs. Remember that VAEs learn posterior distribution parameters from dataset x without any label information. In CVAEs, we have additional contextual information c when estimating the posterior distribution. The objective function of the CVAE can be written as follows:

$$\max \quad \log p_\theta\left(x|c\right) - \mathcal{D}\left[q_\phi\left(z|x\right) \| p_\theta\left(z|x\right)\right] \tag{1}$$

where $p_\theta(x|c)$ denotes the marginal log-likelihood of the data, $p_\theta(z|x)$ is the posterior distribution and $q_\phi(z|x)$ is its approximator. According to the definition of Kullback-Leibler divergence and Bayes' theorem, maximizing Equation

1 is equivalent to maximizing the following equation:

$$\max \quad \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z) - \log \frac{q_\phi(z \mid x)}{p_\theta(z \mid c)}\right] \tag{2}$$

Notice that the relative KL divergence is always a non-negative number. We can derive the evidence lower bound (ELBO) of the objective function based on Equation 1 and Equation 2:

$$\begin{aligned} \log p_\theta(x \mid c) &\geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z) - \log \frac{q_\phi(z \mid x)}{p_\theta(z \mid c)}\right] \\ &= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right]}_{\text{Reconstruction loss}} - \underbrace{\mathcal{D}\left[q_\phi(z \mid x) \| p_\theta(z \mid c)\right]}_{\text{Regularization loss}} \end{aligned} \tag{3}$$

The ELBO comprises two parts — the reconstruction loss and the regularization loss. The reconstruction loss describes the data distribution given the latent space, while the regularization loss measures the divergence between the true prior distribution and the encoder's approximate distribution. In practice, we usually choose to minimize the negative ELBO because the term $\mathcal{D}\left[q_\phi\left(z|x\right) \| p_\theta\left(z|x\right)\right]$ is hard to calculate. However, a reasonable posterior approximator $q_\phi(z|x)$ can effectively alleviate the impact of the bias.

2.2 Expressing VITS as a CVAE

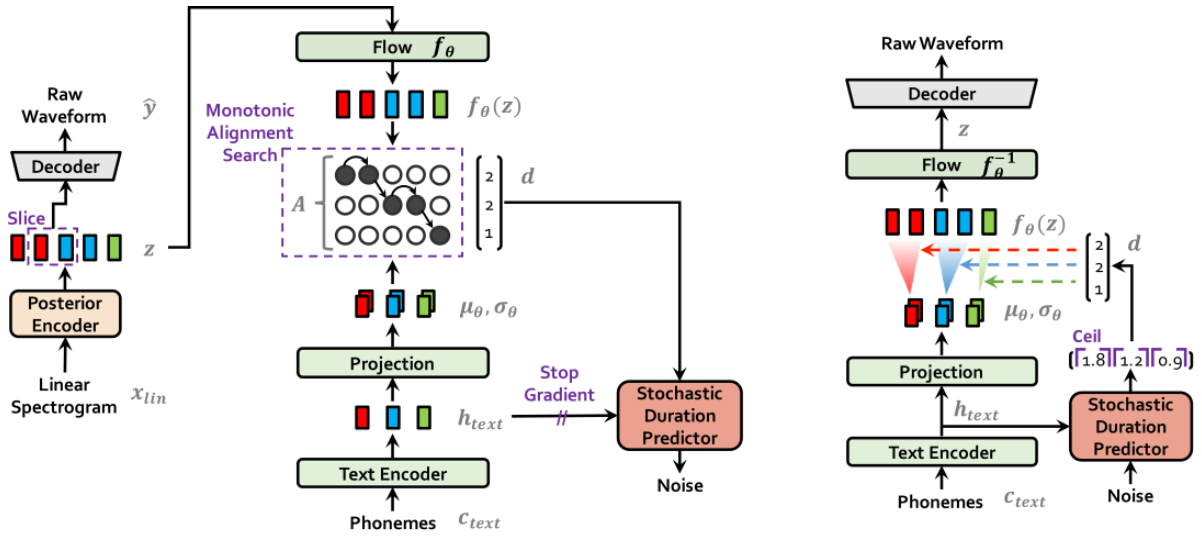

Figure 1: The architecture of VITS. This figure is directly adapted from [1].

The architecture of VITS is shown in Figure 1. VITS model can be expressed as a CVAE. We will illustrate the main idea behind VITS based on the reconstruction loss as well as the regularization loss.

Reconstruction Loss. During the training, we have ground truth soundtracks x and their corresponding text c_{text}. Kim et al. [1] define the reconstruction loss as the L_1 norm between the input's mel-spectrogram and the estimated mel-spectrogram produced by the model:

$$L_{recon} = \|x_{mel} - \hat{x}_{mel}\|_1 \tag{4}$$

It is easy to compute the mel-spectrogram x_{mel} of a specific audio file. As for the approximate mel-spectrogram $\hat{x}_{mel}$, we can upsample the latent space to the waveform domain and then transform it to the mel-spectrogram domain.

Regularization Loss. The text-to-speech task requires an automatic alignment between the text input and the corresponding voice features. VITS adopts the monotonic alignment search algorithm to align the audio and textual embeddings. The monotonic alignment search algorithm (MAS) is proposed in [4]. The intuition behind the MAS is applying the dynamic programming algorithm to maximize the likelihood of the data distribution. It will produce an alignment matrix A. The condition c in VITS is defined as the concatenation of the text input c_{text} and the alignment

matrix A, i.e., $c = [c_{text}, A]$. Given the condition c, we can calculate the regularization loss following the definition of KL-divergence:

$$\begin{aligned} L_{kl} = \mathcal{D}\left[q_\phi(z \mid x) \| p_\theta(z \mid c)\right] = \log q_\phi\left(z \mid x_{lin}\right) - \log p_\theta\left(z \mid c_{\text{text}}, A\right) \\ z \sim q_\phi\left(z \mid x_{lin}\right) = N\left(z; \mu_\phi\left(x_{lin}\right), \sigma_\phi\left(x_{lin}\right)\right) \end{aligned} \tag{5}$$

It is worth mentioning that VITS uses the linear-scale spectrogram of the soundtrack x_{lin} instead of the mel-spectrogram. [1] explains that the linear-scale spectrogram is helpful in enhancing the resolution of generated audio. To improve the model's performance further, VITS introduces a normalizing flow f_θ that connects the conditional prior distribution and the latent space in the following way:

$$p_\theta(z \mid c) = N\left(f_\theta(z); \mu_\theta(c), \sigma_\theta(c)\right) \left|\det \frac{\partial f_\theta(z)}{\partial z}\right| \tag{6}$$

Equation 6 transforms the latent space into some complex distributions that possess a more powerful expressiveness, which is critical for generating realistic outputs.

2.3 Training VITS

The major part of VITS is a CVAE. However, there are more things to consider when handling the text-to-speech task. Firstly, we need to decide the duration of each token in outputs. A straightforward idea is adding a deterministic duration predictor into the model. But this naive method has a big flaw — it can not reflect the features of different speakers, which limits the model's application in multi-speaker scenarios. Alternatively, VITS designs a stochastic duration preditor whose loss function is L_{dur} and optimizes it with other modules during training.

VITS also integrates adversarial training into the architecture to guide the optimization direction. With a discriminator D and a decoder G, VITS combines two kinds of loss in the adversarial training, namely the least-squares loss and the additional feature-matching loss:

$$\begin{aligned} L_{adv}(D) &= \mathbb{E}_{(y,z)}\left[(D(y)-1)^2 + (D(G(z)))^2\right] \\ L_{adv}(G) &= \mathbb{E}_z\left[(D(G(z))-1)^2\right] \\ L_{fm}(G) &= \mathbb{E}_{(y,z)}\left[\sum_{l=1}^{T} \frac{1}{N_l}\left\|D^l(y) - D^l(G(z))\right\|_1\right] \end{aligned} \tag{7}$$

Taken together, we can obtain the final loss for VITS training:

$$L_{vae} = L_{recon} + L_{kl} + L_{dur} + L_{adv}(G) + L_{fm}(G) \tag{8}$$

During the inference stage, we fix all weights and disgard the posterior encoder because we only have the text input. Final outputs are produced by the decoder G.

3 Experiments

4 Conclusion

To conclude, we have built a customizable text-to-speech personal audio assistant based on the VITS model and fine-tune it to achieve optimal performance. Although there still exist some problems, such as equivalent performance. For example, audios from some kind of speaker are of higher quality than others. But overally speaking, the model has achieved the goal and performs well as a personal audio assistant.

References

[1] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR, 18–24 Jul 2021.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[3] Carl Doersch. Tutorial on variational autoencoders, 2021.

[4] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8067–8077. Curran Associates, Inc., 2020.