Towards a Customizable Text-to-speech Personal Assistant

Project demo

Shupei Li, Chenyu Shi, Siwen Tu, Shuang Fan

Dec.5 2023

CONTENTS

VITS Model

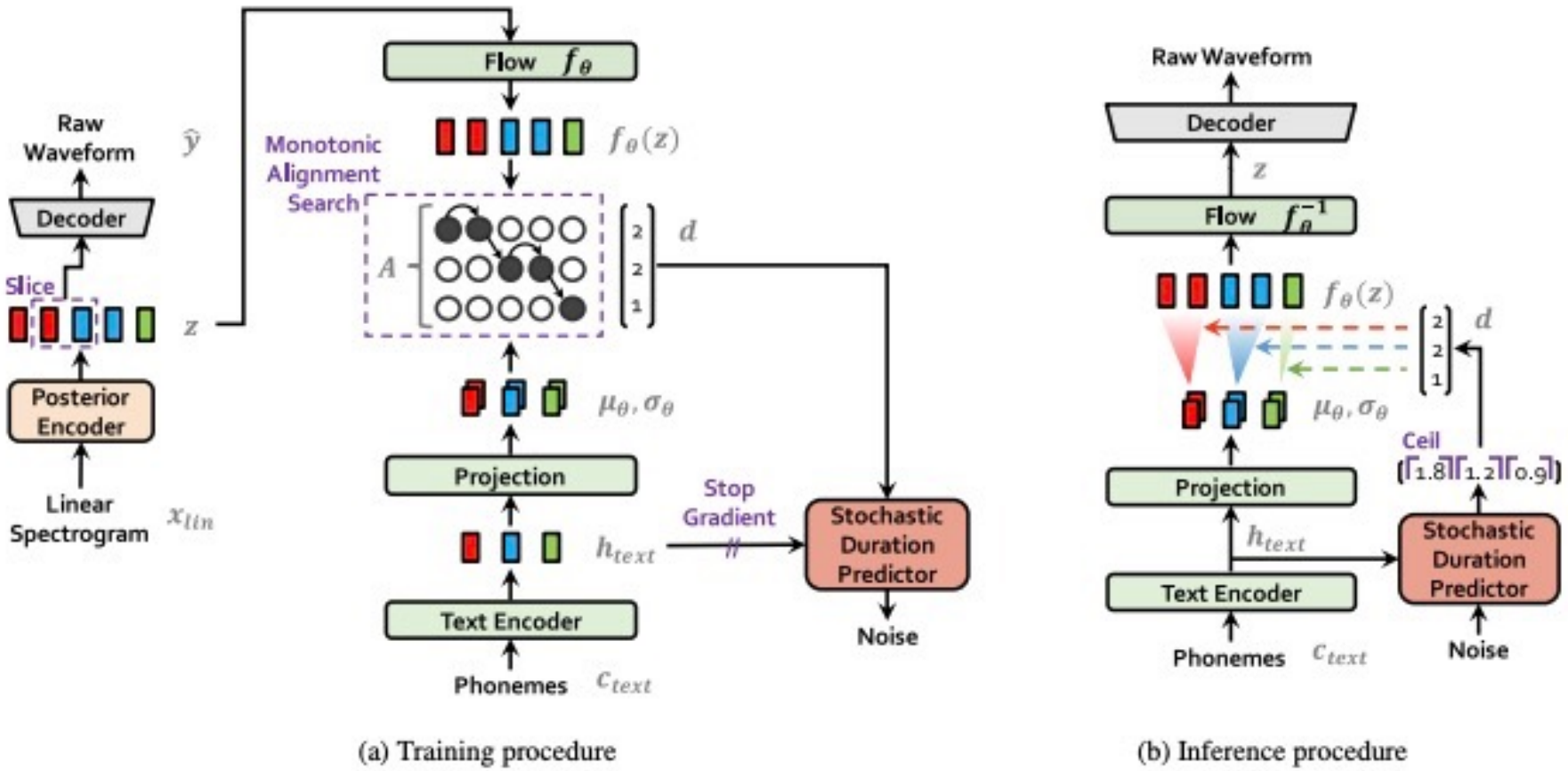
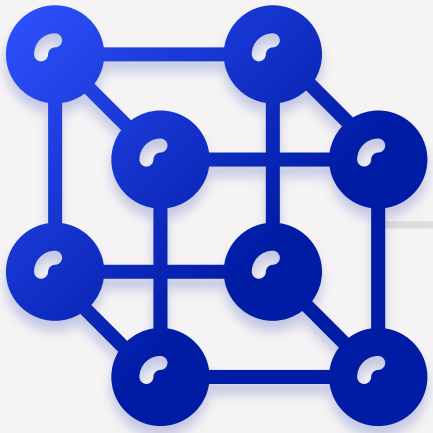


Figure 1. System diagram depicting (a) training procedure and (b) inference procedure. The proposed model can be viewed as a conditional VAE; a posterior encoder, decoder, and conditional prior (green blocks: a normalizing flow, linear projection layer, and text encoder) with a flow-based stochastic duration predictor.

Main Idea, VITS and Our Contributions

Main idea

1. Use the VITS model as the backbone.
2. A parallel end-to-end TTS method.
3. Adopt variational inference augmented with normalizing flows.
4. Improve the expressive power of generative modeling.



New language support

Train the VITS model on Chinese datasets from scratch.

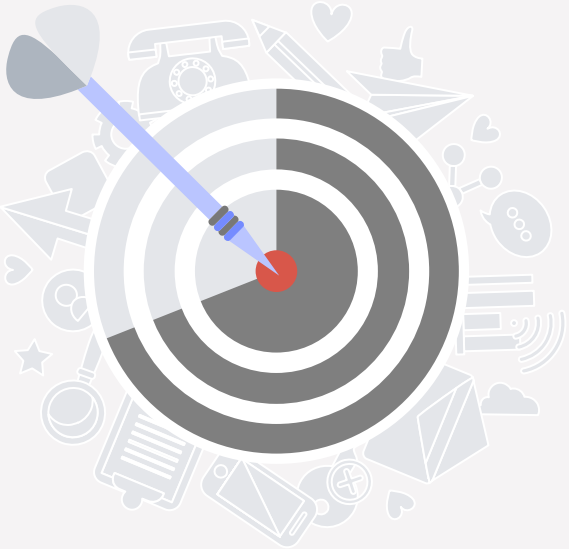
New speaker support

Fine-tune the VITS model on Chinese and Japanese with pre-trained model weights.

Training

Chinese support

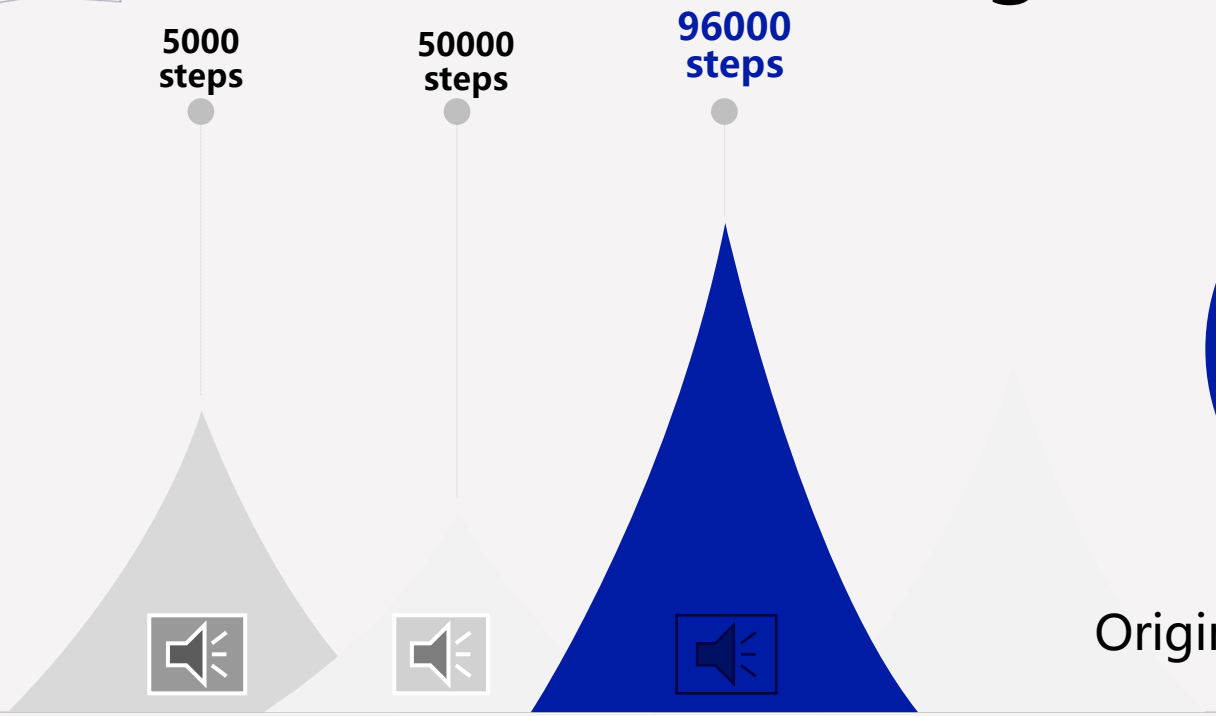
· Dataset: Chinese Standard Mandarin Speech Corpus（10000 Sentences）
· Hardware: 1 Titan Xp GPU.
· Train the model for 96,000 steps. (~ 48 hours)
· DEMO: "我爱计算机科学！"

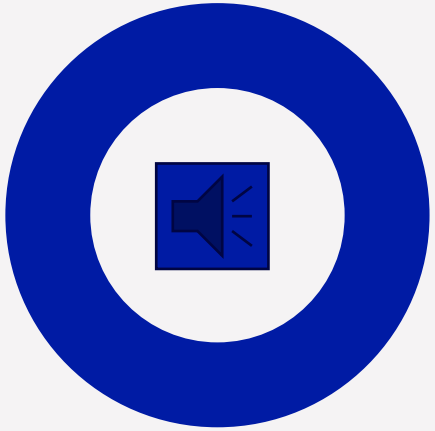


Fine-tuning: Chinese and Japanese

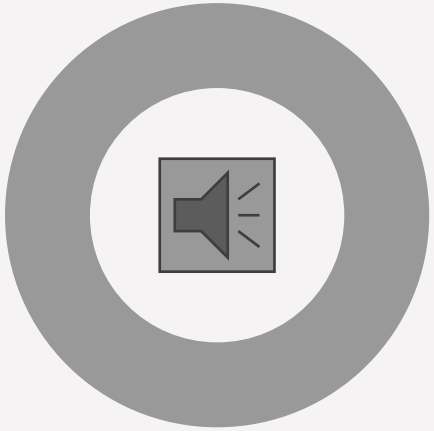
• Use a pretrained model.
• Dataset: AISHELL-3 for Chinese, JVS for Japanese.
• Fine-tuning the model for 84,500 steps in total. (~ 16 hours)
• Add five speakers(Chinese): child voice, young female voice, middle-aged female voice, young male voice, middle-aged male voice.
• Add six speakers(Japanese).
• DEMO

Results for Training





Original speaking in dataset



Generated speaking

Train from scratch on Chinese dataset for 96000 steps
“我爱计算机科学”
“ I love computer science”

Then apply the model to generate specific sentence
“沉鱼落雁，闭月羞花”
“fish sink and geese fall, eclipse the moon and flowers blush”

Our Website and Demo Shows

Customizable Text-to-speech Personal Assistant

2023 fall API final project, finished by

Students: Shupei Li, Chenyu Shi, Siwen Tu, Shuang Fan

Introduction

With the development of deep learning, text-to-speech models have achieved great progress in many applications, such as audiobook reading or audio assistant. In this project, we built a customizable text-to-speech personal assistant based on VITS (original repo: link). We trained and fine-tuned the model on different languages of different speakers, which allows us to build a text-to-speech personal assistant of different options.

Training of the model

We trained the model for 96000 epochs in total. And we record the performance during the training. Here you can choose the number of epoch to hear a clip of the audio of the sentence "我爱计算机科学" ("I love computer science", the original speaking language is Mandarin). We can hear that at the beginning, the audio clip is just some noise, but after training , this audio clip gradually becomes more realistic.

- ○ 0/96000 epoch
- ◉ 5000/96000 epoch
- ○ 10000/96000 epoch
- ○ 30000/96000 epoch
- ○ 50000/96000 epoch
- ○ 70000/96000 epoch
- ○ 96000/96000 epoch

0:01 / 0:01

https://sd12321sd.github.io/api_project.github.io/

Goals for Demo Day(Finished)

01

Accomplish the text-to-speech task based on VITS model.

02

New language support. Train the VITS model on a Chinese dataset from scratch.

03

New speaker support. Fine-tune the model on multi-speakers Chinese datasets and Japanese datasets.

Work Division

Text-to-speech model construction: all group members

Multi-language fine-tuning task: Shupei Li, Siwen Tu

Multi-speaker fine-tuning task: Chenyu Shi, Shuang Fan

Report writing and presentation preparation: all group members

Thanks!