Personalized Singing Voice Beautifier

林泳鵬、汪宣甫、陳韋傑

Outline

- Motivation
- Methodology
- Results
- Findings

Motivation

- Our innovation aims to train a **personalized singing voice beautifier**.
- We referred to the relevant paper [1], but identified some limitations in its training data.
    - The training pair data : (professional, amateur) singing voice.
    - We observed that amateur singing surpasses that of ordinary individuals.
- So we want to investigate some methods that can generate data pairs which are more suitable to our scenarios.
- And use those optional datas to train a personalized voice beautifier.

[1] : Liu, Jinglin, et al. "Learning the beauty in songs: Neural singing voice beautifier." arXiv preprint arXiv:2202.13277 (2022).

Methodology

Learning the Beauty in Songs: Neural Singing Voice Beautifier

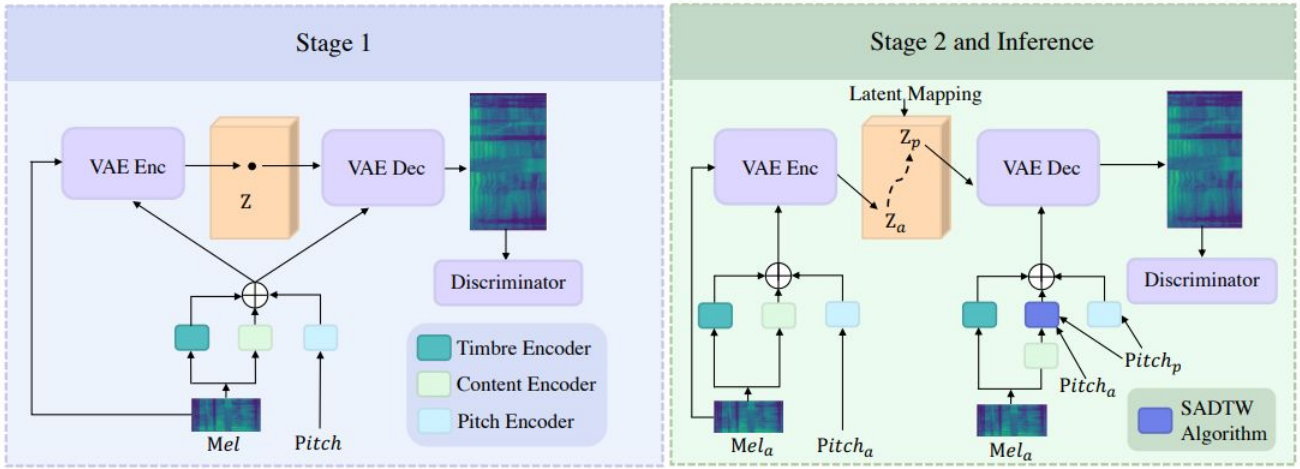
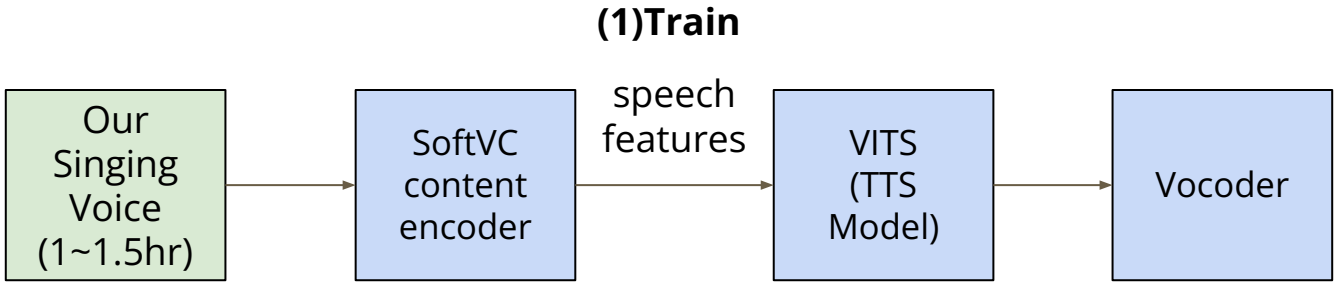


Figure 1: The overview of NVSB. The training process consists of 2 stages, and the second stage shares the same pipeline with the inference stage. "VAE Enc" means the encoder of CVAE; "VAE Dec" means the decoder of CVAE; "Mel" means the mel-spectrogram; "**z**" means the latent variable of the vocal tone; the "a"/"p" subscript means the amateur/professional version.

Methodology (cont.)

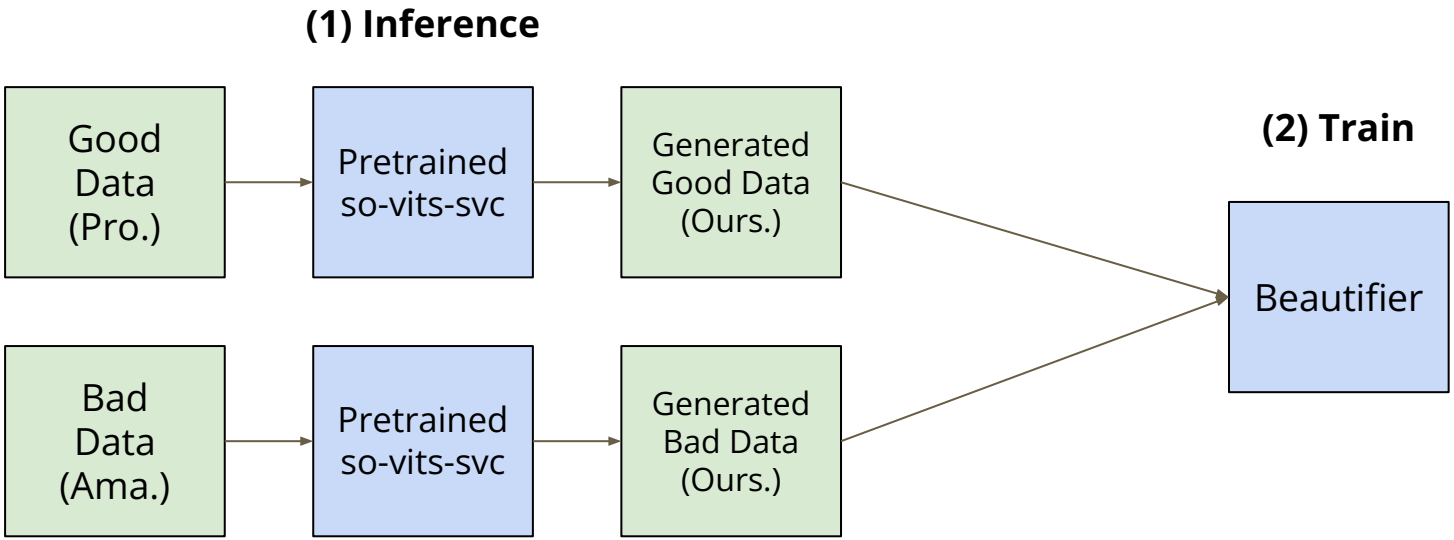
- Singing Voice Conversion
  - Use **so-vits-svc**(SoftVC VITS Singing Voice Conversion)

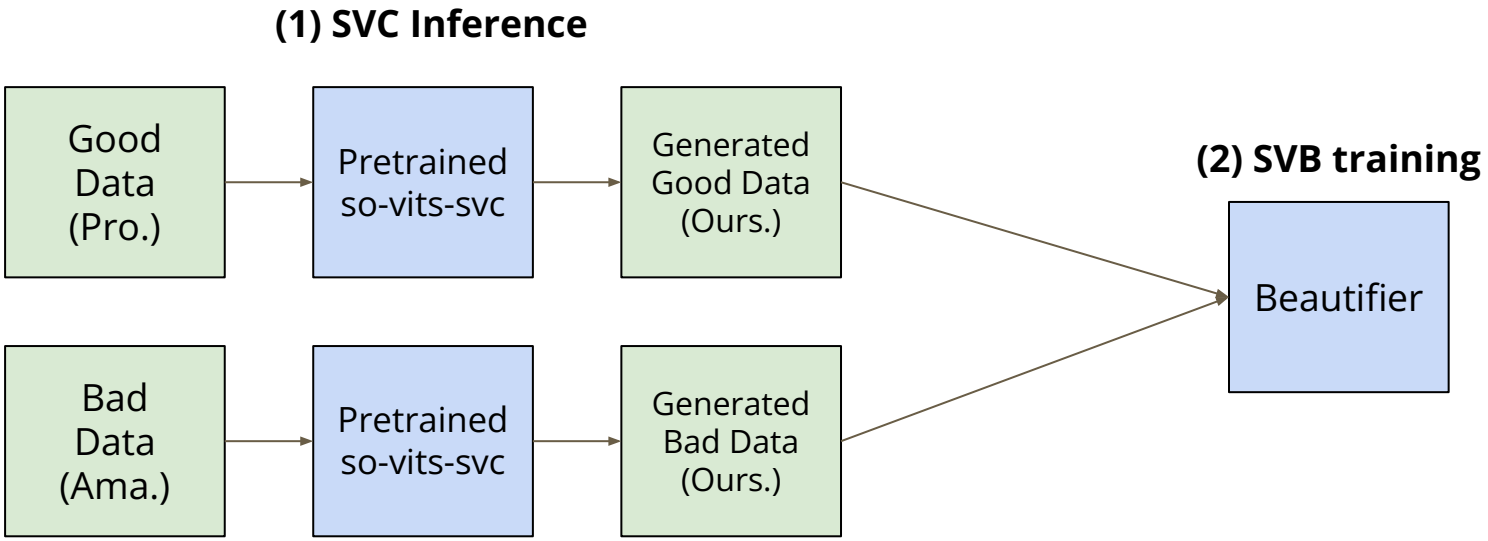


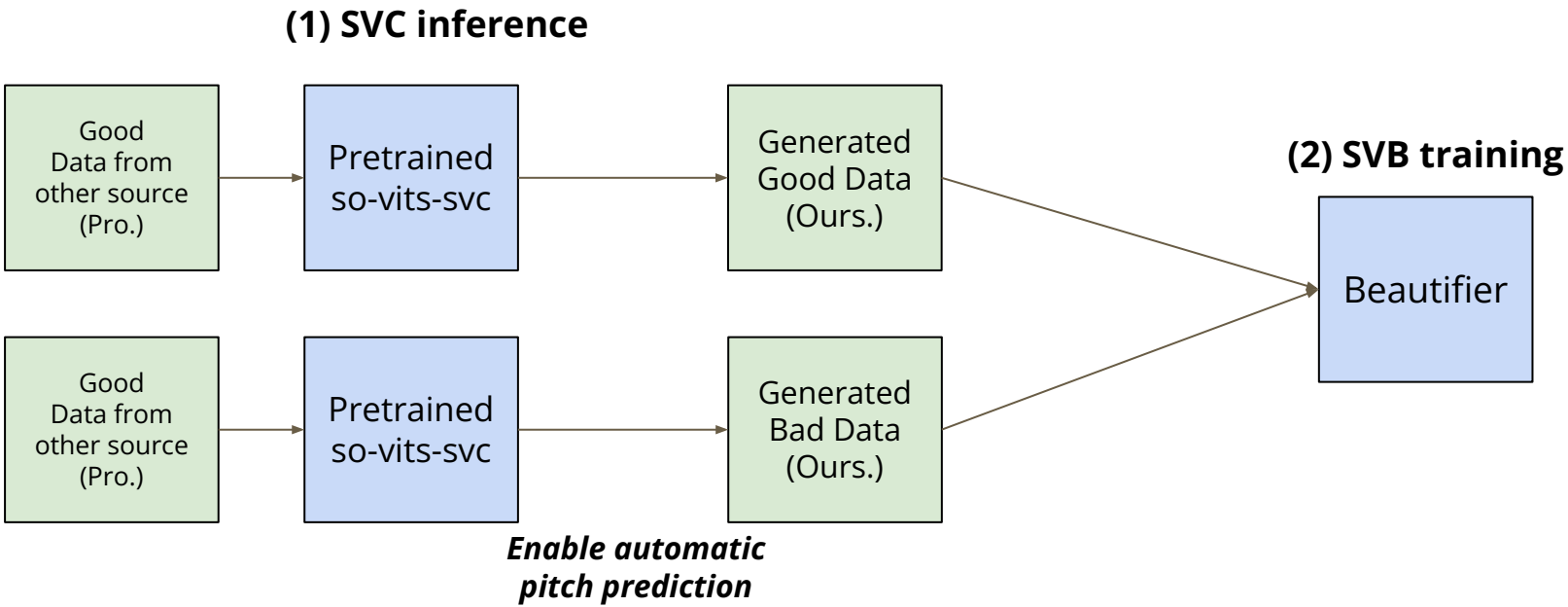
Reference : https://github.com/svc-develop-team/so-vits-svc

Methodology (cont.)

- Task 1 : SVC + SVB
  - Use **PopBuTFy** Dataset(English)

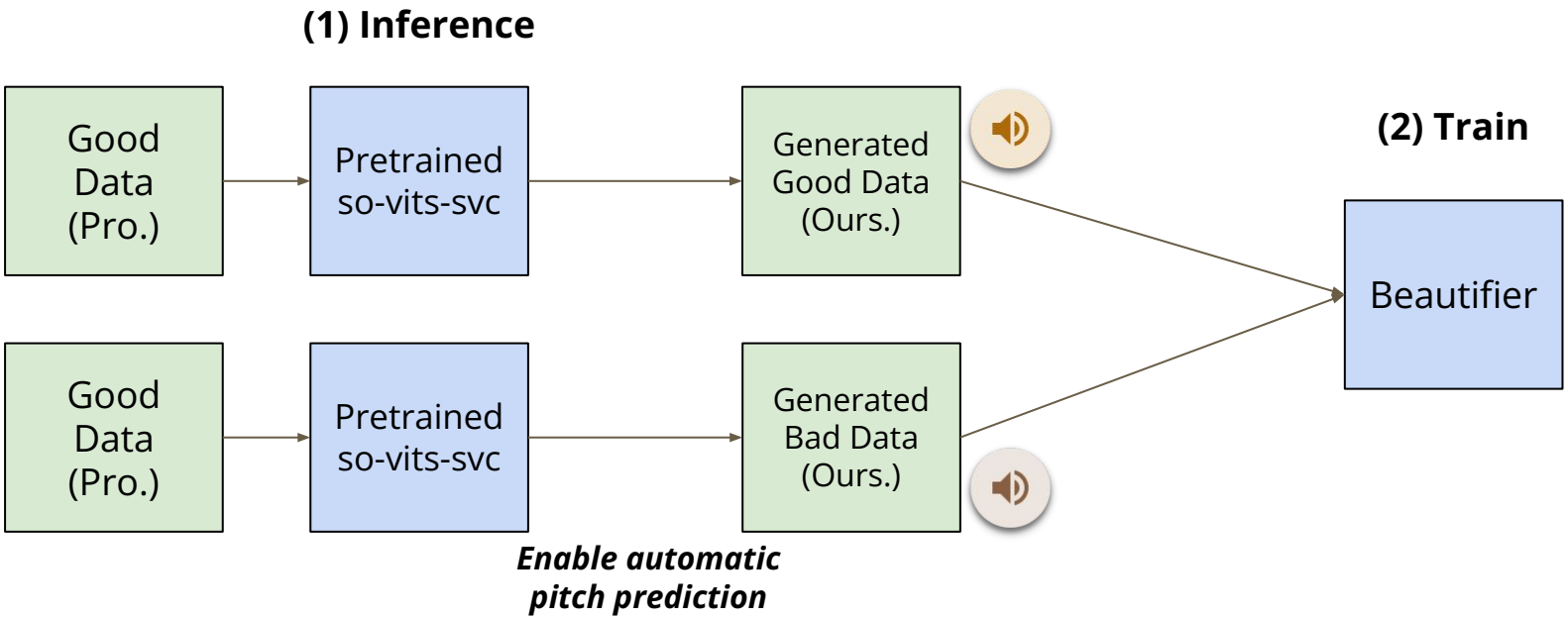



(1) SVC Inference
Good Data (Pro.)
Pretrained so-vits-svc
Generated Good Data (Ours.)
Bad Data (Ama.)
Pretrained so-vits-svc
Generated Bad Data (Ours.)
(2) SVB training
Beautifier


(1) SVC inference
Good Data from other source (Pro.)
Pretrained so-vits-svc
Generated Good Data (Ours.)
Good Data from other source (Pro.)
Pretrained so-vits-svc
Generated Bad Data (Ours.)
(2) SVB training
Beautifier
Enable automatic pitch prediction

Methodology (cont.)

- Task 2 : SVC + SVB + extend to chinese public dataset
  - Use **OpenSinger** dataset(Chinese)



Results

- F0 Root Mean Square Error (F0 RMSE)
  - We use F0 RMSE to estimate the pitch correction performance between the resulting audio and recordings of professional singers.

Method	F0 RMSE	
	泳鵬	宣甫
Baseline(Ours.)	122.44	152.58
Paper(SVB)	109.69	148.20
Task1(SVC+SVB)	110.15	146.85
Task2(SVC+SVB+CH)	**108.32**	**146.15**

Results

- Demo
  - Our Singing Voice
    - 泳鵬 / 宣甫
  - Paper(SVB)
    - 泳鵬 / 宣甫
  - Task1(SVC+SVB)
    - 泳鵬 / 宣甫
  - Task2(SVC+SVB+CH)
    - 泳鵬 / 宣甫

Findings

- In inference, we encounter some instability due to the **necessity of precise pitch information from the same singer**, despite relying solely on pitch templates from professional singers.
- For better results in the inference, **aligning singing data** with professional audio is still necessary, as the original paper's alignment algorithm has limitations.
- We discovered that employing **auto-predicted f0** in the voice conversion system can somehow **simulates pitch mismatches.**
    - Thus, we use it to generate amateur data.
- The **timbre of the output is influenced** by utilizing different speakers' template pitch.