StableSVC: Latent Diffusion Model for Singing Voice Conversion

MDS6002 Default Project

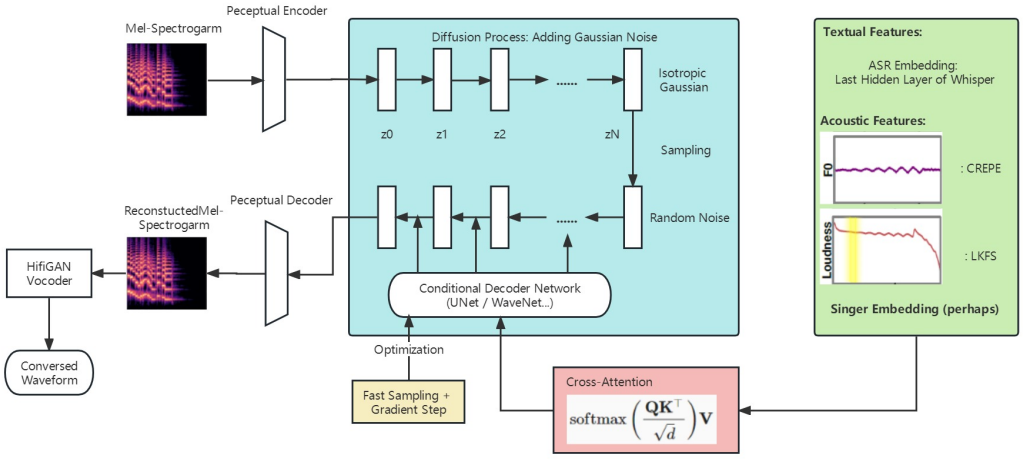
Jiahao Chen (222041038, xx%)
School of Data Science, Chinese University of Hong Kong, Shenzhen
222041038@link.cuhk.edu.cn

1 Overview

- **Source Code**: Repository URL: https://github.com/SLPcourse/MDS6002-222041038-JiahaoChen. Latest Commission: 3e31c844c90a450571c236d146b20e77d46ec1e7.
- **Progress**: Preprocessing Opencpop and M4Singer dataset, extracting and storing features that will be used to train the model, i.e., whisper embedding, Mel-spectrogram, F0 and loudness.
- **Challenges**: Training general-purpose Perceptual VAE and HifiGAN Vocoder may be too heavy a computation burden in our project, since we need to train them across different datasets. To overcome this, we utilize the concept of inductive bias. Since the target singer's mel-spectrogram is all we need, we hope that fitting them into target singer's dataset is enough to generate satisfactory results.

2 Workflow and Architecture

Figure 1: Latent Diffusion Model for SVC



The workflow is comprised of two stages. First, train a Perceptual AutoEncoder/VAE to move Denoising Diffusion Probabilistic Model (DDPM) into latent space (which may be borrowed from [2]). Second, training a decoder network in this latent space within the framework of DDPM [3]. Textual and acoustic features are projected into this latent space via cross-attention mechanism [5]. The last hidden state of Whisper (which is basically a Transformer encoder-decoder) [4] will be used as text embedding to represent textual features (further improvement can be made by training a PPG extracting network on it). F0 is extracted by using CREPE model [1] and the loudness feature

is obtained following [3]. Finally, HifiGAN will be trained as the Vocoder to generate conversed waveform based on the reconstructed Mel-spectrogram.

3 Experiment Design and Results

In this project, what we want to learn is how well will Latent Diffusion Model adapt to the task of singing voice conversion, given that we are essentially dealing with Mel-spectrogram, which is an image itself. LDM is expected to achieve fast training and inference speed as well as conversion quality compared with previous methods based on the result from [2].

Table 1 demonstrates how the experimental settings. It's expected that we compare the performance of the baseline model provided in the project guideline and the model we propose via objective metrics following [3]: Mel-Cepstrum Distortion (MCD) and F0 Pearson Correlation (FPC).

Table 1: Experimental Setup

Name	StableSVC	Baseline
Framework	LDM	DBLSTM Regression
Textual Features	Whisper Embedding (further: PPG)	Whisper Embedding
Acoustic Features	Mel-spectrogram, F0 (CREPE), Loudness	MCEP, F0, SP, AP
Vocoder	HifiGAN	WORLD

References

[1] Jong Wook Kim et al. "Crepe: A convolutional representation for pitch estimation". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 161–165.

[2] Haohe Liu et al. *AudioLDM: Text-to-Audio Generation with Latent Diffusion Models*. 2023. arXiv: 2301.12503 [cs.SD].

[3] Songxiang Liu et al. "Diffsvc: A diffusion probabilistic model for singing voice conversion". In: *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2021, pp. 741–748.

[4] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS].

[5] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.