# StableSVC: Latent Diffusion Model for Singing Voice Conversion

{MDS6002} {Default} Project

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

## 1 Introduction

- **Problem**: This project will be centered around augmenting the conversion quality for mel-spectrogram based Singing Voice Conversion (SVC) using **Denoising Diffusion Probabilistic Model (DDPM)** [9], with idea borrowed from latest **high-resolution image generation** technology. [14]
- **Plan**: This project is to apply **Latent Diffusion Model** in [14] to SVC task, which was proposed originally for high-resolution image generation. I'll first experiment on Opencpop dataset which has only one target singer and then M4Singer dataset which has multiple target singers. An initial conception of the pipeline is to, first, train an AutoEncoder(AE) with perceptual loss proposed in [13] to move SVC process (which is based on mel-spectrogram) into a latent space, and use DDPM in this latent space, conditioning on acoustic and textual features via Cross Attention mechanism [16] during the reverse process of DDPM. In terms of experiments, LDM will be explored on how to adapt to Many-to-One SVC tasks based on a modified **baseline** model pipeline in [9], and I'll further explore inserting singer identity representation vector into LDM utilizing its intrinsic conditioning architecture to perform Many-to-Many SVC if time allows. Another baseline from other parallel research direction can be [8]. For detailed tentative model architecture and pipeline, please refer to Appendix A.

## 2 Milestone

- **Progress**: So far what I have done is reading papers, collecting ideas for possible innovation and trying to conceive some possible new model framework. I'm supposed to start experiment very soon, after reading [8], which is a very recent work applying LDM to the speech processing domain (Text-to-audio, to be specific).
- **Challenges**: The most challenging part of my planned work may lie in the implementation details, for example, taking gradient step when training the denoising network conditioned by cross-attention during the reverse process of DDPM. Possible solution is to learn from the published codes of [8].
- **Source Code**: https://github.com/SLPcourse/StableSVC.
- **Paper Reading (Related Work):** [15] introduced Phonetic PosteriorGrams(PPGs) into neural Voice Conversion (VC) task, using Deep Bi-Directional Long-Short-Term Memory (DBLSTM) to perform regression from PPG to MCEP, which is required by the vocoders back then. [1] further pushed the work into the field of SVC. [13] adopted the means of self-supervised learning in the space of singer-identity-vector (which is achieved by Mix-up and Back-translation) to perform Many-to-Many SVC, using Least Squares GAN (LSGAN) as framework and WaveNet as backbone model. [5] introduced DDPM framework into the domain of speech processing and came up with an improved reversing method named Fast Sampling, establishing a model upon WaveNet which is capable of conditional generation. [9] followed the path and inherited the DDPM framework, using Mel Spectrogram as target

for the reversion process of DDPM, acoustic features (F0, Power Gram) and textual features (PPG) as auxiliary information. Recently, in the domain of computer vision, *Stable Diffusion* model, which is essential LDM, has achieved tremendous success in text-prompted image generation [14], and is also my inspiration of the tentative model framework as well. More papers I am considering to read (especially [8]): [8][2][7][10][4][6][3][11][12][17].

## References


[1] Xin Chen et al. "Singing Voice Conversion with Non-parallel Data". In: *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 2019, pp. 292–296. DOI: 10.1109/MIPR.2019.00059.

[2] Alex Graves et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 369–376.

[3] Haohan Guo et al. "Improving adversarial waveform generation based singing voice conversion with harmonic signals". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 6657–6661.

[4] Rongjie Huang et al. "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis". In: *arXiv preprint arXiv:2204.09934* (2022).

[5] Zhifeng Kong et al. "Diffwave: A versatile diffusion model for audio synthesis". In: *arXiv preprint arXiv:2009.09761* (2020).

[6] Xu Li, Shansong Liu, and Ying Shan. "A Hierarchical Speaker Representation Framework for One-shot Singing Voice Conversion". In: *arXiv preprint arXiv:2206.13762* (2022).

[7] Zhonghao Li et al. "Ppg-based singing voice conversion with adversarial representation learning". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 7073–7077.

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

[9] 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.

[10] Songxiang Liu et al. "Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation". In: *2021 ieee international conference on multimedia and expo (icme)*. IEEE. 2021, pp. 1–6.

[11] Aaron van den Oord et al. "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499* (2016).

[12] Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou. *Speaker Conditional WaveRNN: Towards Universal Neural Vocoder for Unseen Speaker and Recording Conditions*. 2020. arXiv: 2008.05289 [eess.AS].

[13] Adam Polyak et al. "Unsupervised cross-domain singing voice conversion". In: *arXiv preprint arXiv:2008.02830* (2020).

[14] 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.

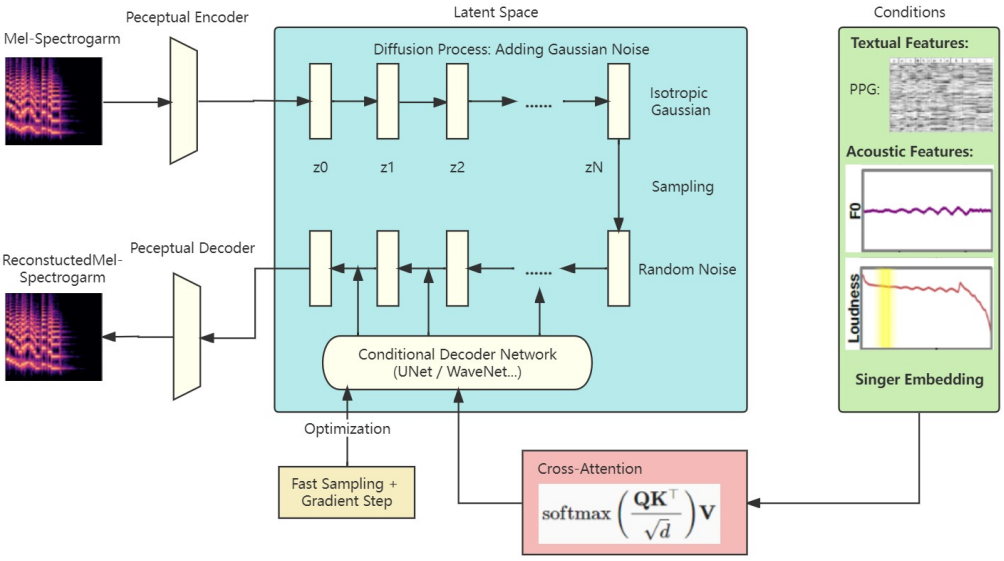
[15] Lifa Sun et al. "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training". In: *2016 IEEE International Conference on Multimedia and Expo (ICME)*. 2016, pp. 1–6. DOI: 10.1109/ICME.2016.7552917.

[16] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[17] Shiliang Zhang et al. "Deep-FSMN for large vocabulary continuous speech recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5869–5873.

# A Appendix: Tentative Model Pipeline

Figure 1: Latent Diffusion Model for SVC




The model pipeline can be divided into two stages. First, train an AutoEncoder with perceptual loss and reconstruction loss[13]. Second, establish a DDPM in the latent space based on the trained AutoEncoder. During the DDPM, a conditional decoder network is used to recover the mel-spectrogram from Gaussian Noise. Acoustic features and textual features (and possibly singer embedding) are projected into the latent space via Cross-Attention Mechanism [16]. The conditional decoder network is trained via the maximizing the ELBO (Evidence Lower Bound) by (fast) sampling reverse process and taking gradient step [5].