StableSVC: Latent Diffusion Model for Singing Voice Conversion

MDS6002 Default Project

Jiahao Chen (222401038)
School of Data Science
Chinese University of Hong Kong, Shenzhen
222041038@link.cuhk.edu.cn

Abstract

Singing Voice Conversion (SVC) is to convert a piece of singing voice from a source singer to another piece of singing voice with its timber resembling other target singers, which essentially belongs to generation task. Recently, a deep generative model framework called Denoising Diffusion Probabilistic Model (DDPM)[5] has gained great success in generation tasks. This work's main contribution is to try to apply DDPM to the task of SVC. And, to the best of my knowledge, this is the first work to try to use conditional UNet[5], which is built on CNN and Multi-head Self Attention (MSA) layer, as the backbone denoising network in the reverse diffusion process in SVC task. Also, this work explores utilizing the latest Automatic Speech Recognition (ASR) system, Whisper[22], to extract textual content in SVC. In addition, several training techniques and SVC task setting are explored through experiments in this work. We have found based on these experiments that. Finally, a more advanced framework of SVC based on Latent Diffusion Model (LDM)[23] is proposed for possible future improvement of the conversion quality and inference speed of SVC task.

1 Key Information to include

- Mentor: Xueyao Zhang, Zhizheng Wu

2 Introduction

SVC, as described in the in the abstract, needs to process a piece of singing voice to another piece of singing voice, with sole difference being the timber of these two singing voice, ideally. The components of a piece of singing voice are divided into textual contents and acoustic properties. In terms of human's perception, the acoustic properties are comprised of volume, pitch, and timbre. What determines the identity of the source of a piece of singing voice most is its timbre, which is actually defined as the remaining voice of a piece of human voice except volume and pitch. These are from the perspective of subjective perception. There are also objective counter-part in the domain of acoustics, e.g., features based on spectral envelope (Mel-Cepstral Coefficients[7], MCEP) for timbre, F0 (fundamental frequency) for pitch.

Therefore, the aim of SVC is to retain the textual contents and acoustic features representing pitch and loudness of the source singing voice, and replace the timbre features of source singer to target singer. This can be achieved by learning a neural network on the target singer's singing voice dataset. Through training, it captures and stores the timbre features of the target singer as its network parameters.

Since we are working on the acoustic features, we may need a model to ensemble the acoustic features we process, receiving some acoustic features (Vocoder features) and generating a piece of singing voice based on it. These kinds of models are called Vocoders. Vocoders facilitate the

manipulation of acoustic features and play an important part in the task of SVC, since an important part of current SVC researches' main pipeline is building a model to replace the timbre features in the Vocoder features with target timbre features. Traditional works utilizing deep models are based on MCEP for WORLD Vocoder[19]. More recently, works based on whole Mel-spetrograms as Vocoder feature achieve more promising results. Therefore, this work will focus on transforming the Mel-spectrograms of source singers to the ones with timbre of target singer. Specifically, HiFi-GAN Vocoder[11] based on the framework of Generative-Adversarial Network (GAN) is adopted in this work.

Since the form of Mel-spectrograms is essentially an image, the works from image generation task can be transferred to SVC. Recently, deep generative framework Denoising Diffusion Probabilistic Model (DDPM)[5] has achieved huge success in image tasks. Most works involving SVC still stops at GAN as their acoustic model to convert timbre features in Mel-spectrograms.Therefore, this work will focus on transferring DDPM to the task of SVC and exploring its conversion capability on SVC task. More recently, a deep generative framework called Latent Diffusion Model (LDM)[23] has beat traditional DDPM with respect to generation quality and inference speed. A SVC framework based on LDM is also proposed in this work without experiments.

Finally, experiments are carried out to explore different training strategies of my model. Quality of the generated audios are compared with different metrics. Specifically, Exponential Moving Average (EMA) of the model is adopted, and the model with or without EMA are compared. Also, effect of validation is investigated into in the context of deep learning. Finally, conversion quality when source singers are female and male are compared, with target singer being female.

3 Related Work

To extract the textual content is a crucial part of SVC. Automatic Speech Recognition (ASR) system is utilized to achieve this. [24] introduced Phonetic PosteriorGrams(PPGs) into neural Voice Conversion (VC) task. It used ASR system to extract PPGs, and then built a Deep Bi-Directional Long-Short-Term Memory (DBLSTM) model to perform regression from PPGs to one of the WORLD Vocoder feature representing timbre: MCEP. [1] further pushed the work into the field of SVC, which utilized PPGs to accomplish nonparallel SVC task (i.e., many-to-one conversion). By regressing to target singer's MCEP and leaving other WORLD Vocoder features unchanged, it managed high-quality SVC.

There are also other works processing the audios directly without using a Vocoder. [21] adopted the idea of self-supervised learning in the space of singer-identity-vector. It used the strategies of Mix-up and Back-translation during training process, managing to achieve Many-to-Many unsupervised cross domain SVC. [2] only extracts the singer embedding and pitch using two networks, and left other higher-level semantic features (i.e., textual contents) for another network to learn. [17] tried to simplify the SVC pipeline to boost inference speed, abandoning PPGs and building light-weighted phoneme recognizer based on Conformer[4].

There are also many recent SVC works focused on processing Mel-spectrogram. And, as DDPM gradually gains popularity as a promising generation model, increasing recent works in the domain of speech processing are exploring on the application of DDPM. [12] introduced DDPM framework into the domain of speech processing, establishing a model upon WaveNet which is capable of conditional generation. Diffusion Vocoder [15] used the shallow version of DDPM to achieve state-of-the-art singing voice synthesis result. [16] followed the path and inherited the DDPM framework, using Mel Spectrogram as target variable for the reversion process of DDPM. During reverse diffusion process, acoustic features of F0 and loudness, textual content of PPGs are sent to the denoising network as conditioning information. This helps to retain other features of the singing voice except for the timbre we want to convert. This work is mainly motivated by [16].

More recently, in the domain of image generation, *Stable Diffusion* model, which is essential LDM, has achieved tremendous success in text-prompted image generation. [14] transferred this work into the domain of speech processing, managed to achieve text-prompted audio generation using LDM.

4 Approach

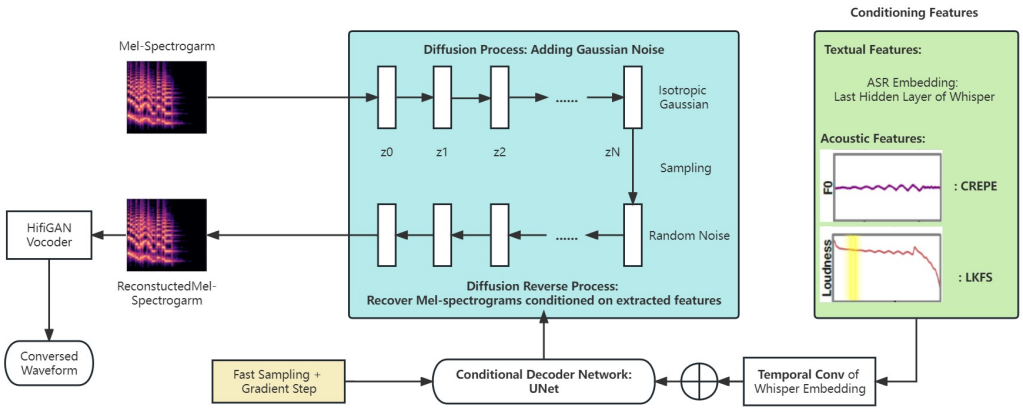
4.1 Framework

First, I propose a simple diffusion framework for SVC task, which is essentially the basic DDPM from [5]. Figure 1 demonstrates this framework. Mel-spectrograms are the target variable during the generation process. Similar to [16], we feed the network in the reverse diffusion process with the features we want to retain: textual contents, pitch and loudness. For textual content, we use the last hidden layer of Whisper model[22] as the ASR text embedding (Whisper embedding). We calculate F0 by the CREPE model[9], which uses CNN to estimate the pitch of a piece of voice. The loudness feature is calculated following the standard of LKFS [1].

Specifically, the Whisper embedding is temporally average-pooled to reduce its dimensionality, and is further processed by a one-dimensional convolution layer in the embedding dimension to fit the dimension of Mel-frequency filters number of Mel-spectrograms. The F0 feature and loudness feature are broadcast to fit the shape of Mel-spetrograms as well. Finally, noisy Mel-spectrograms are concatenated with these three conditioning features (Whisper embedding, F0 and loudness) and fed into conditional denoising UNet to recover denoised mel-spectrogram at the previous diffusion step.

My framework is different from previous work [16] in that, instead of using WaveNet[20] as denoising network in the reverse diffusion process, conditional UNet[5] is adopted. Also, Whisper embedding is experimented on to represent textual feature other than PPGs.

Figure 1: Denoising Diffusion Probabilistic Model for SVC



Based on the idea of *Stable Diffusion*, LDM framework for SVC is proposed for possible future work as well. The difference is that we train a perceptual Variational AutoEncoder (VAE)[10][5] to move DDPM into latent space with far smaller dimensionality. This enables higher training efficiency for stronger backbone network and faster inference speed in practice. More detailed illustration can be found in the appendix.

4.2 Denoising Diffusion Probabilistic Model (DDPM)

The idea of DDPM is borrowed from physics and Stochastic Differential Equation. By recovering input from eliminating the noise step by step, the model is expected achieve stable generation result without losing generalization capability.

DDPM can be divided into two stages. The first step is forward diffusion process, which is to gradually add Gaussian noise into the input. It can be expressed as:

$$q\left(x_t|x_{t-1}\right) = N\left(\mu_t = \sqrt{\bar{\alpha}_t}x_0, \sum_t = \left(1-\bar{\alpha}_t\right)I\right). \tag{1}$$

[1] BS.1770 : Algorithms to measure audio programme loudness and true-peak audio level, https://www.itu.int/rec/R-REC-BS.1770-4-201510-I/en

where β_t is the variance schedule rate and t is the diffusion step taking value from $t = 1, ..., T$. By reparameterization trick, it can be rewritten as:

$$q\left(x_t|x_{t-1}\right) = N\left(\mu_t = \sqrt{\bar{\alpha}_t}x_0, \sum_t = \left(1 - \bar{\alpha}_t\right) I\right), \tag{2}$$

where $\alpha_t = 1 - \beta$ and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$.

The second stage in the reverse diffusion process, which aims at letting denoising network learn to eliminate the noise added at step t. By the idea from Bayes Variational Inference, we can convert training the denoising network by optimizing negative log-likelihood of the training data into maximizing the Evidence Lower bound (ELBO):

$$\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, t\right)\right\|^2, \tag{3}$$

where ε_θ is the denoising network taking $\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon$ and t as input.

During inference, we use the trained network to generate desired output $y_{t-1} \sim p_\theta\left(y_{t-1}|y_t\right)$ with given conditions by the following Langevin Dynamics:

$$y_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(y_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\varepsilon_\theta\left(y_t, t\right)\right) + \sigma_t z, \tag{4}$$

where $\sigma_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t$ and z is standard Gaussian noise $N(0, 1)$.

Choices of variance schedule rate β_t can influence the generation quality of DDPM. Here, we use the cosine noise schedule rate with more stable generation capability. It's proposed in [3] and is given by:

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, f(t) = \cos^2\left(\frac{t/T + s}{1 + s} \cdot \frac{\pi}{2}\right). \tag{5}$$

A recommended choice of s is 0.08 as in the original paper.

DDPM algorithm can be concluded as algorithm 1 and algorithm 2.

Algorithm 1 Training of DDPM

Input:

Denoising network $\varepsilon_\theta(\cdot)$; training set $\{(x, Whisper, f_0, loudness)\}_{m=1}^{M}$; diffusion step T; variance schedule rate $\{\beta_t\}_{t=1}^{T}$; number of iterations N.

1: **for** $i = 1, ..., N$ **do**
2: Sample a batch of $(x, Whisper, f_0, loudness)$ from training set
3: Sample Standard Gaussian noise ε from $N(0, 1)$
4: Sample $t\ fromU(0, T)$ for each sample in the batch $i.i.d.$
5: Take gradient step on $\nabla_\theta\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, t\right)\right\|^2$ in the mini-batch.
6: **end for**
7: **return** trained $\varepsilon_\theta(\cdot)$

4.3 Conditional UNet

Condition UNet, as proposed in [5], is comprised sequentially of several down-sample blocks, several middle blocks to process bottleneck in an intention to extract higher-level features, and several up-sampling blocks. The down-sampling and up-sampling blocks are comprised of two ConvNeXt[18] blocks a linear Multi-head Self-Attention (MSA)[25] layer and one down-sampling/up-sampling layer. The middle block consists of a ConvNeXt block, a MSA layer and a ConvNeXt block sequentially. Residual connection and pre layer normalization is used for each MSA layer. Finally, a initial convolution layer and a final convolution layer with large kernel sizes are added at the start and the end of the network to extract raw information better.

Algorithm 2 Inference of DDPM

Input:
Trained denoising Network $\varepsilon_\theta(\cdot)$; conditions $(Whisper, f_0, loudness)$; diffusion Step T; variance schedule rate $\{\beta_t\}_{t=1}^{T}$.

Sample Standard Gaussian noise y_T from $N(0,1)$
for $t = 1, ..., T$ **do**
  if $t > 0$ **then**
    Sample Standard Gaussian noise z from $N(0,1)$;
  else
    $z = 0$;
  end if
  Denoise as by one step $eq.4$: $y_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(y_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(y_t, t\right)\right) + \sigma_t z$, conditioned on $(Whisper, f_0, loudness)$.
end for
return generation y_0

5 Experiments

5.1 Data

In this work, Opencpop[26] is used as the target singer dataset for training, and M4Singer[28] is used as the source singer dataset for inference. Opencpop contains pieces of singing voice from a professional female singer, while M4Singer contains singing voice from multiple singers, including males and females. Both of the dataset are elaborately annotated. I pick 4 singers from M4Singer, each with 5 singing voice pieces.

5.2 Evaluation method

For evaluation, objective metrics are used. To be specific, we use Mel-Cepstral Distortion (MCD) as descrived in [24] and F0 Cosine Similarity (FCS) to compared the quality of the converted audio from models with different training strategies and SVC settings.

5.3 Experimental details

In this work, sampling rate, number of Mel-filters, size of Short Time Fourier Transform (STFT), window size of STFT and hop size of STFT are uniformly set as 16000, 80, 1024, 1024, 256 respectively. Mel-spectrograms are pad temporally to the length of 496. Whisper model with size of 'base' is used to get textual embedding.

As is suggested in [5], we standardize Mel-spectrograms, F0, loudness to the range of [-1, 1] by Min-Max-Scalar estimated from Opencpop (since it's our target singer's dataset). This is to better fit the standard Gaussian noise in DDPM. After conversion, Mel-spectrograms are scaled back as input of the Vocoder.

I follow the official configuration v1 to train the Hifi-GAN Vocoder on Opencpop dataset. It's trained for 40 epochs with total steps of 8825. The step with best validation Mean Squared Error (MSE) score is used. DDPM model is trained following [27] for a total steps of 31100, nearly 66 epochs.

Different training techniques and SVC settings are experimented on and compared in this work. First, Exponential Moving Average (EMA) of the model is utilized to obtain a better and stabler performance of inference. Models with and without EMA are compared. Second, model from step with best validation score and model from latest step are compared. Finally, conversion quality with female source singer and male source singer are compared.

5.4 Results

The model failed to convert singing voice from source singer to target singer ideally. After subjective evaluation from a total number of 3 participants, it's found that the model managed to retain the pitch

and loudness features from the source piece and converted to the target timbre successfully, but failed to captured the textual contents and retained it in the output. The results are presented as table 1. The settings of Female and Male both use the best technique of Best, EMA. The low FCS may result from the failure of conversion, and thus great acoustic difference.

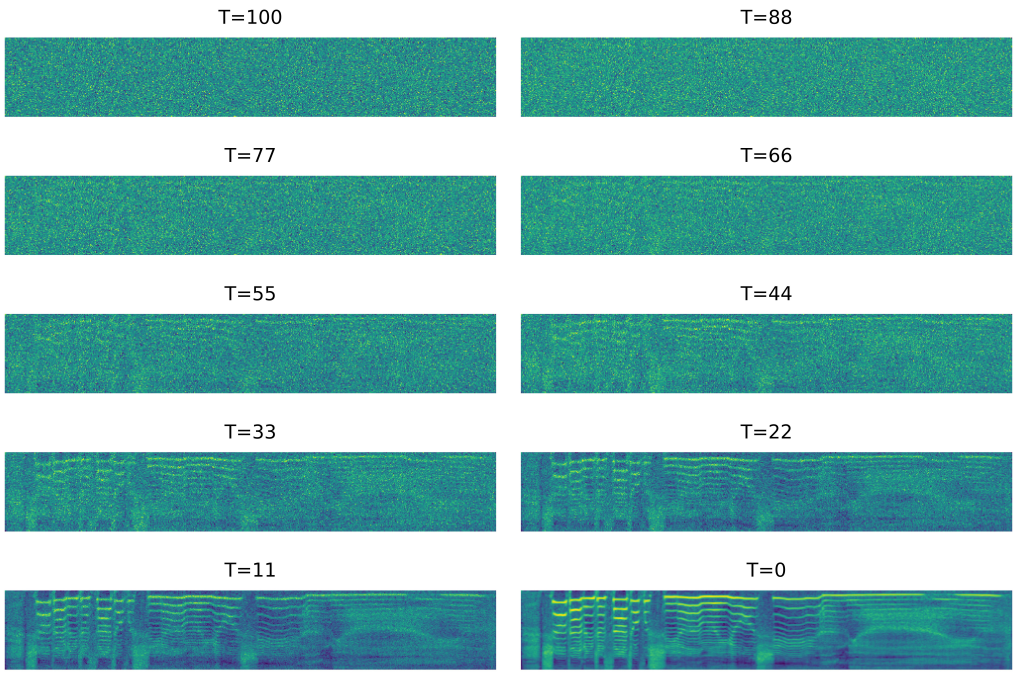
Table 1: Different Training Techniques and SVC Settings

Settings	Best, EMA	Latest, EMA	Best	Latest	Female	Male
MCD	1776	1776	1774	1776	1776	1680
FCS	0.3351	0.3076	0.2763	0.3076	0.3351	0.2999

6 Analysis

Figure 3 shows the process of inference of one sample. This implies that the model works properly. It manages to recover Mel-spectrograms from pure Gaussian noise. However, the recovered Mel-spectrograms is not what's desired. Therefore, the failure of it may lie in the way the conditioning features are processed. This was due the improper way to utilize Whisper embedding and caused the loss of textual contents. Whisper embedding is dense. Even when the audio only has a length of 3 8s, Whisper model will return an embedding that is densely distributed over temporal dimension. For this reason, temporal average-pooling used in this work may lose great amount of information of textual content and caused the model to fail. The model also failed to retain the silence part of the source singing voice. This may also arise from the density of Whisper embedding. Making input temporally length variable or use PPGs may handle this problem.

Figure 2: Example of Inferencing



7 Conclusion

Through this project, I have learned how to deploy DDPM in different generation task. The key finding is that DDPM works well with SVC task. The limitation is the failure of processing textual information. Future work will include more reasonable utilization of textual contents, for example, making the reduction of Whisper embedding learnable with a network, using PPGs as textual contents or training a network from Whisper embedding to PPGs. Future improvement of this work may follow the LDM framework proposed.

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] Chengqi Deng et al. "Pitchnet: Unsupervised singing voice conversion with pitch adversarial network". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7749–7753.

[3] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: 2105.05233 [cs.LG].

[4] Anmol Gulati et al. *Conformer: Convolution-augmented Transformer for Speech Recognition*. 2020. arXiv: 2005.08100 [eess.AS].

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[6] Jonathan Ho and Tim Salimans. "Classifier-free diffusion guidance". In: *arXiv preprint arXiv:2207.12598* (2022).

[7] S. Imai. "Cepstral analysis synthesis on the mel frequency scale". In: *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 8. 1983, pp. 93–96. DOI: 10.1109/ICASSP.1983.1172250.

[8] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[9] 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.

[10] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17022–17033.

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

[13] Anders Boesen Lindbo Larsen et al. *Autoencoding beyond pixels using a learned similarity metric*. 2016. arXiv: 1512.09300 [cs.LG].

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

[15] Jinglin Liu et al. *DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism*. 2022. arXiv: 2105.02446 [eess.AS].

[16] 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.

[17] 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)*. 2021, pp. 1–6. DOI: 10.1109/ICME51207.2021.9428161.

[18] Zhuang Liu et al. "A convnet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11976–11986.

[19] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications". In: *IEICE TRANSACTIONS on Information and Systems* 99.7 (2016), pp. 1877–1884.

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

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

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

[23] 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.

[24] 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.

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

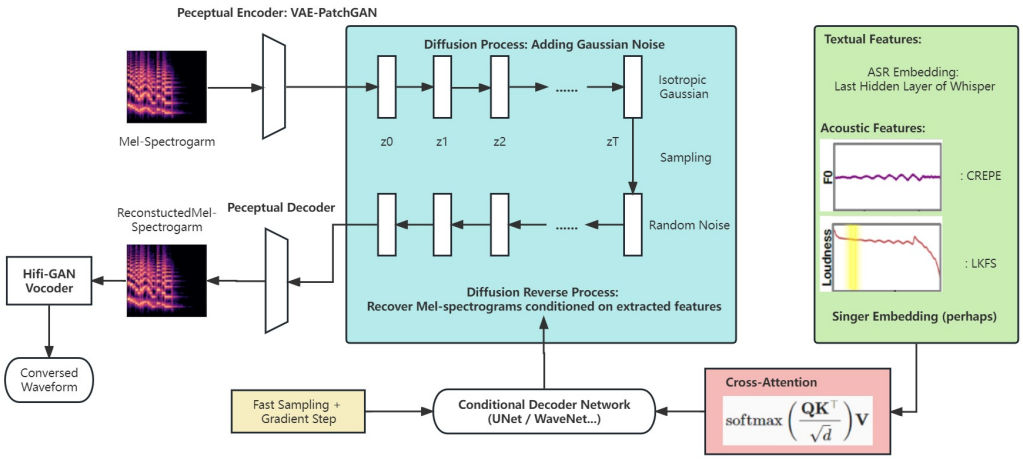
[26] Yu Wang et al. “Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis”. In: *arXiv preprint arXiv:2201.07429* (2022).

[27] Yuancheng Wang et al. *AUDIT: Audio Editing by Following Instructions with Latent Diffusion Models*. 2023. arXiv: 2304.00830 [cs.SD].

[28] Lichao Zhang et al. “M4Singer: A Multi-Style, Multi-Singer and Musical Score Provided Mandarin Singing Corpus”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 6914–6926.

A StableSVC: Latent Diffusion Model for SVC

Figure 3: Latent Diffusion Model for SVC



In Latent Diffusion Model, conditional DDPM is also utilized. Specifically, VAE-PatchGAN combined from [13] and [8] can be utilized. The three conditioning features are projected into the latent space of Mel-spectrogram via cross-attention mechanism[25], which may enhance the conditioning effect than mere concatenation. Then, Classifier-Free Guidance[6] mechanism is utilized to train the conditional denoising network.