Singing Voice Conversion via Stochastic Refinement

MDS6224 Final Project

Zhiyuan Yan (222041040)
School of Data Science
Chinese University of Hong Kong, Shenzhen
222041040@link.cuhk.edu.cn

Zhiyou Zhao (222041044)
School of Data Science
Chinese University of Hong Kong, Shenzhen
222041044@link.cuhk.edu.cn

Jiajian Ma (222041049)
School of Data Science
Chinese University of Hong Kong, Shenzhen
222041049@link.cuhk.edu.cn

Abstract

Singing voice conversion (SVC) is a technology aiming to convert the timbre of a song without changing the content and melody. Previous studies have displayed the power of the diffusion model in SVC and other audio signal-processing tasks. However, most of these diffusion-based models sample acoustic features (like Mel-spectrogram) directly from Gaussian noise, leading to longer training time and increased sampling uncertainty, which is unsuitable for generating acoustic features that are sensitive to numerical perturbation, such as MCEP. To address this issue, we propose an alternative framework for applying diffusion models in SVC, which utilizes a diffusion model to refine the output of an initial deterministic acoustic feature predictor. Our results display a significant improvement in the quality of acoustic feature generation and training efficiency compared to the classical diffusion models framework. In the future, we plan to apply this method to generate other acoustic features like Mel-spectrogram and compare it with the stage-of-the-art SVC methods.

1 Workload

Contributions: Zhiyuan Yan (33%) for paper reviewing, model training, report writing, and coding; Zhiyou Zhao (33%) for report writing, model training, and coding; Jiajian Ma (33%) for idea raising, report writing, and coding.

External collaborator: Thank you to Fengyu Yang who finetuned the VQ-VAE pre-train model for us.

Code: `https://github.com/CUHKSZ-RMZHANG/final_project-497662892`

2 Introduction

Singing voice conversion (SVC) is a technology aiming to convert the timbre of a song to another singer without modifying the content and melody. A classical SVC system contains a content encoder, to extract singer-independent features (like words, melody, and rhythm) from a source singing signal, and a conversion model, which is usually a generative model, to transform singer-independent features into either acoustic features or waveforms[1, 2, 3].

Diffusion probabilistic models, namely diffusion models, are one of the most promising and powerful generative models[4, 5]. Previous studies display its state-of-the-art performance in generating natural images with or without different types of conditions [6, 7, 8]. The fundamental structure of a

diffusion model is the denoising network, which is typically a UNet [7, 6] or a ViT model [9] for image generation.

Diffusion models have also been wildly applied in the field of audio signal generation. It can achieve decent performance in the task of raw audio waveform generation[10], text-to-speech synthesis [11], audio editing [12], singing voice synthesis (SVS) [13], and singing voice conversion [14]. However, most of the above models directly sampling acoustic features from Gaussian noise [10, 13, 14], which might be unsuccessful in generating features sensitive to noise and require a longer time for training [14]. At the same time, some models directly use denoising network structures derived from the image generation field[12] or use simple residual network structures[14, 13, 11], which may result in the model can not fully exploiting its performance based on the characteristics of the sequential data.

To explore potential solutions to the above problems, we propose a new framework in this project 1 (b). Rather than using diffusion models to predict features (like MCEP) from scratch, we utilize them to modify the initial prediction result of a deterministic model. Conditioning on the initial prediction, content features, and pitch features, our diffusion model can iteratively restore the MCEP. In addition, we also constructed a latent diffusion model with a Transformer-based denoising network to evaluate whether a network structure designed for sequence data can further improve model performance. In summary, the contribution of our work includes:

- Propose a new framework for the application of diffusion model in SVC task, which outperform the existing diffusion model in MCEP feature generation [14].
- Conducted a preliminary exploration of applying the latent diffusion model and Transformer denoising network in the SVC task.

3 Related Work

Singing Voice Conversion: SVC approaches can be divided into two categories based on the need for parallel training data: parallel SVC and non-parallel SVC[14]. Due to the high cost of collecting parallel data for training, researchers have explored a variety of non-parallel SVC approaches. In [3], an auto-encoder based on the WaveNet was used for unsupervised SVC. This method offers the ability to convert between singers that are present in the training set. GAN-based approaches are also used for non-parallel SVC[15, 16, 15]. Very recently, research has also been conducted on the use of diffusion models [14] for non-parallel SVC.

Diffusion Model: The diffusion model, which originated from [17], employs a diffusion Markov chain to break the data down into Gaussian noise and then applies a reverse process to generate data from the noise. Recently, significant advancements have been made in this field, where the diffusion model achieved state-of-the-art results synthesis in images data [4, 18] as well as audio signals [10, 19]. In a subsequent study [7], Rombach proposed a new model which can perform the diffusion and denoising process in the latent space of an autoencoder, leading to a significant boost in training and inferring speed and improvement in generated image quality.

Diffusion Model in audio signal processing: Recently, lots of diffusion-based models emerge in the field of audio signal processing, including for the task of text-to-speech synthesis [11], audio editing [12], singing voice synthesis (SVS) [13], and singing voice conversion [14] and can achieve really decent performance. However, because diffusion models originated from and rapidly developing in the field of image generation [4, 7, 9, 18], some diffusion model for audio still prefer to follow the model used in the visual domain, including VQ-VAE models based on UNet[12] and denoising networks based on UNet[12] or 2D convolution[14, 13, 11]. Furthermore, since the aforementioned models all start from resampling Gaussian noise, they are mainly suitable for generating acoustic features that are relatively insensitive to numerical perturbations, such as mel-spectrograms[14, 13, 11, 12].

4 Method

The workflow and architecture of our singing voice conversion model are shown in fig1 (a), which includes 3 parts: the singer-independent feature extractor, the MCEP prediction network (green), and the singer-specific synthesizer (red). Our system is similar to the SVC model in Chen's [2] paper, but with a replacement in the MCEP prediction network: we use a 6-layer Transformer as the

initial MCEP predictor and use a diffusion model to iteratively refine this output (fig1, b). The initial predictor and the classical diffusion model (fig1, c) will serve as the baseline models.

4.1 Singer-Independent Feature Extractor

Based on DiffSVC[14], we use phonetic posterior grams (PPG) and f0 as the primary content conditions for the diffusion network. We use the whisper-median to extract PPG features and then project them to 256 dimensions by a linear layer. We use Librosa to extract the f0 of the singing voice and then discretized them to 0-299 after taking the logarithm and embedding them into 256 dimensions. For diffusion models, We concatenate the PPG features and f0 embeddings, and further fuse them with a single layer self-attention.

4.2 Initial MCEP Predictor

We use a 6-layer Transformer as the initial MCEP predictor, which can generate MCEP features based on PPG features input. This model also serves as our baseline.

4.3 Diffusion Model

VQ-VAE model: For the latent diffusion model, we used a VQ-VAE model with the same structure in this paper [7], which is a UNet and can achieve 4x downsampling. It was pre-trained on ImageNet-1K, and we subsequently fine-tuned it on the Opencpop dataset for 100 epochs.

Denoising Network: The detailed architecture of the diffusion model is shown in fig1, as (b) is our proposed framework and (c) is the classical diffusion model. The denoising network could be the 2d convolution in DiffSVC [14] or a Transformer. It can generate MECP conditioned on F0 and PPG features, with (ours) or without (classical) first-stage prediction. The diffusion models are trained by L2 loss [4] and use 1000 steps of DDPM to sample predicted MCEP[4].

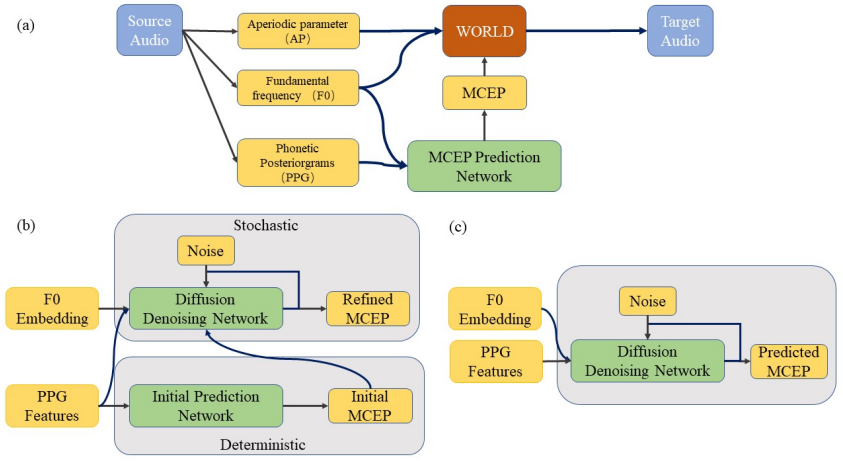

Figure 1: The architecture of our model. (a) The workflow of the SVC system. (b) The structure of our proposed framework. The initial MCEP predictor produces the deterministic candidate, and then the diffusion denoiser network refines it. (c) The structure of the classical diffusion model. It samples the MCEP directly from noise.

4.4 Singer-Specific Synthesizer

We use the World as our vocoder, which can synthesize music with generated spectral envelope, transformed/non-transformed f0, and aperiodic feature input[2].

5 Experiments

5.1 Data

We use Opencpop [20] and M4Singer [21] as our dataset, with the singer in Opencpop as the target singer and other 20 singers in M4Singer as source singers. In the Opencpop dataset, we used 3550 audio segments as training, and the remaining 206 segments as validating. In the M4Singer, we randomly selected 5 segments from each singer, composing a test set with 100 segments.

5.2 Evaluation method

In this project, we use the MSE loss of the predicted MCEP features on the validation and testing set as the primary metric for the quality of SVC.

5.3 Experimental details

The initial deterministic transformer model is trained with lr = $2.6e-5$ and batch size to 8 for 20 epochs. For the diffusion model, We also try 2 types of denoising networks: 1) 12 residual layers of 2d convolution with 256 channels as recommended by [14]; 2) 4 layers of transformer decoder. The total number of diffusion steps is 1000, and the noise schedule β's are set to linearly spaced from 0.0015 to 0.0205. We train the diffusion models with the ADAM optimizer, with a base learning rate of $1e-4$, cosine decay, and linear warmup as mentioned in [7].

5.4 Results and analysis

The results of this project are shown in Table 1. Our proposed framework achieved the best prediction performance on both validation and test sets compared to the initial predictor and classical diffusion model framework, with MCEP MSE Loss leading the way (val: 0.2974 vs 0.3578 vs 0.5977; test: 0.7971 vs 0.8137 vs 0.8998).

Compared to the classical diffusion model framework, our new framework showed significant improvement in any feature space and denoising network (Table 1, #2 vs #6, #3 vs #7, #4 vs #8, #5 vs #9). It also convergences much faster in training, typically requiring only 20-50 epochs to achieve desired results, far superior to 200+ epochs for classical diffusion frameworks. This huge improvement may be caused by the significant reduction in the task difficulty after modifying the framework. Instead of predicting MCEP from scratch, our new framework only requires the model to correct the errors from the initial predictor, which greatly reduces the learning difficulty and reduces randomness on acoustic feature sampling.

In contrast to previous studies[14], our classical diffusion framework did not perform well in predicting acoustic features in SVC and was far inferior to the simple initial predictor (Table1, #1 vs #2). This may be related to the fact that we chose MCEP instead of Mel-spectrogram [14] as the target feature. Since MCEP is derived from Mel-spectrograms, it has fewer dimensions and more concentrated acoustic features, making it more sensitive to numerical perturbations. In the future, we will try to use Mel-spectrogram as our predictive target.

Furthermore, we found that the Transformer network, which is designed for sequence data, performed worse than the 2D convolutional residual network, which is designed for image data, in MCEP generation (val: 0.2974 vs 0.3462; test: 0.7608 vs 0.7971). One possible reason is that the original Transformer decoder is not suitable for handling data with the same size in input and output. After several Transformer blocks, a linear layer is used to project the embedding size to the output size, which greatly limits its predictive ability when compared with the model designed for images.

Additionally, we did not find that denoising in the latent space improved the optimal performance of our model (Table 1, #2 vs #4, #6 vs #8, #7 vs #9). This phenomenon may be related to the VQ-VAE model we used, which was mainly pre-trained on natural images[7] and only fine-tuned on MCEP data in the Opencpop dataset. Given the significant difference between MCEP and natural images, the current VQ-VAE is naturally challenging to extract features from MCEP and reconstruct them well. We believe that using a VQ-VAE pre-trained on Mel-spectrograms[22] will help improve the performance of the latent diffusion model.

Table 1: The result of our experiments

ID	Model	Denoising Network	MCEP MSE (val)	MCEP MSE (test)
#1	Initial Predictor	NA	0.3578	0.8137
#2	Classical (MCEP space)	2d conv	0.5977	0.8998
#3	Classical (MCEP space)	Transformer	5.6569	4.6916
#4	Classical (Latent space)	2d conv	0.8414	1.0937
#5	Classical (Latent space)	Transformer	3.6998	2.8176
#6	Ours (MCEP space)	2d conv	**0.2974**	**0.7608**
#7	Ours (MCEP space)	Transformer	0.3462	0.7971
#8	Ours (Latent space)	2d conv	0.4052	0.9912
#9	Ours (Latent space)	Transformer	0.4491	0.8428

6 Conclusion

In this project, we proposed a new application framework for diffusion models in SVC, which utilize it to refine the initial predicted MCEP of a deterministic model, rather than predicting it from scratch. Our proposed framework outperforms any diffusion model with the classical framework, in the quality of acoustic feature generation and training efficiency.

However, this study still has some limitations, such as using MCEP, which is more sensitive to numerical perturbations, as the target acoustic feature instead of the more commonly used Mel-spectrogram; using a VQ-VAE model that is not pre-trained on MCEP data; without exploring the performance of larger models such as UNet and ViT; and lacking subjective metrics for SVC quality. In the future, we will address these limitations to improve our model.

References

[1] Xueyao Zhang. Singing Voice Conversion Tutorial. *RMSnow's Blog*, 2023.

[2] Xin Chen, Wei Chu, Jinxi Guo, and Ning Xu. Singing voice conversion with non-parallel data. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 292–296. IEEE, 2019.

[3] Chengqi Deng, Chengzhu Yu, Heng Lu, Chao Weng, and Dong Yu. Pitchnet: Unsupervised singing voice conversion with pitch adversarial network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7749–7753. IEEE, 2020.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems*, 2020.

[5] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj "orn Ommer. High-resolution image synthesis with latent diffusion models. 2023.

[8] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[9] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. *arXiv preprint arXiv:2304.03283*, 2023.

[10] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv: Audio and Speech Processing*, 2020.

[11] Zhijun Liu, Yiwei Guo, and Kai Yu. Diffvoice: Text-to-speech with latent diffusion. *arXiv preprint arXiv:2304.11750*, 2023.

[12] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, and Sheng Zhao. Audit: Audio editing by following instructions with latent diffusion models. *arXiv preprint arXiv:2304.00830*, 2023.

[13] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11020–11028, 2022.

[14] Songxiang Liu, Yuewen Cao, Dan Su, and Helen Meng. Diffsvc: A diffusion probabilistic model for singing voice conversion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 741–748. IEEE, 2021.

[15] Haohan Guo, Zhiping Zhou, Fanbo Meng, and Kai Liu. 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)*, pages 6657–6661. IEEE, 2022.

[16] Junchen Lu, Kun Zhou, Berrak Sisman, and Haizhou Li. Vaw-gan for singing voice conversion with non-parallel training data. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 514–519. IEEE, 2020.

[17] Jascha Sohl-Dickstein, Eric L. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv: Learning*, 2015.

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2023.

[19] Nanxin Chen, Yu Zhang, Heiga Zen, Ron Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. 2021.

[20] Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. *arXiv preprint arXiv:2201.07429*, 2022.

[21] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35:6914–6926, 2022.

[22] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.