MCF-SVC: Zero-shot High-Fidelity Singing Voice Conversion with Multi-Condition Flow Synthesis

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Singing voice conversion is to convert the source singing voice into the target singing voice except for the content. Currently, flow-based models can complete the task of voice conversion, but they struggle to effectively extract latent variables in the more rhythmically rich and emotionally expressive task of singing voice conversion, while also facing issues with low efficiency in voice processing. In this paper, we propose a high-fidelity flow-based model based on multi-condition feature constraints called MCF-SVC, which enhances the capture of voice details by integrating multiple latent attribute encoders. We also use Multi-stream inverse short-time Fourier transform(MS-iSTFT) instead of traditional vocoder to enhance the speed of voice reconstruction. We have compared the synthesized singing voice of our model with those of other competitive models from multiple dimensions, and our proposed model is highly consistent with the current state-of-the-art, with the demo which is available at https://lazycat1119.github.io/MCF-SVC-demo.**

***Index Terms*—Singing voice conversion, Flow model, MS-iSTFT, Multi-Condition.**

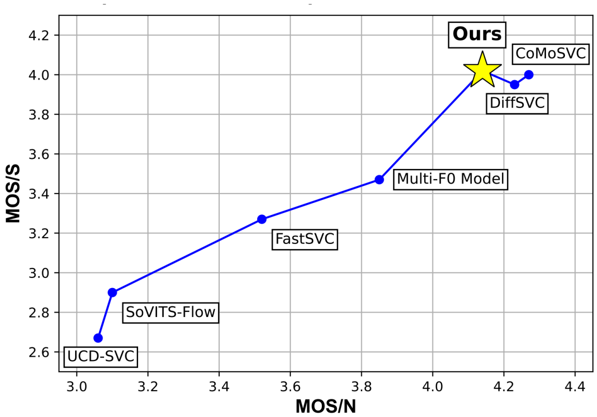

Fig. 1. The performance of competitive models on SVC task, MCF-SVC has the best result in metric of MOS/Similarity

I. Introduction

Singing Voice Conversion (SVC) aims to change a source singer's timbre to that of a target singer while preserving the original singing content, melody, and emotional expression. As an advanced form of Voice Conversion (VC), SVC places additional emphasis on expressive features. With the advancements in deep neural networks, state-of-the-art SVC models such as DDSP-SVC-Diff, So-VITS-SVC, DiffSVC, and CoMoSVC have demonstrated outstanding performance and are widely applied in areas like entertainment, music production, and human-computer interaction [1], [2].

The converted singing voice must fully preserve content information from the source audio. Early approaches employed phonetic posteriorgram (PPG)-based methods [3]–[5] , which extract linguistic features by predicting the posterior probabilities of each phoneme. However, the effectiveness of these methods heavily relies on the performance of Automatic Speech Recognition (ASR) systems, which require large amounts of labeled data for training. To overcome this limitation, unsupervised learning-based representation methods were developed, enabling models to learn from unlabeled or non-parallel speech data and reducing the dependence on annotated datasets. Despite these advancements, unsupervised systems still fall short of supervised ones in terms of quality and intelligibility. To bridge this gap, numerous self-supervised learning (SSL) techniques, such as HuBERT and Encodec, have been introduced. These methods [6] focus on discretizing speech representations by converting continuous waveforms into discrete tokens, thereby accelerating speech processing and integrating semantic information from large language models (LLMs). However, this approach can inadvertently omit some linguistic content, leading to mispronunciations in the converted speech. For instance, when processing fricative sounds, ambiguous frames may be incorrectly assigned to nearby units, resulting in pronunciation errors that compromise the naturalness and accuracy of the singing voice conversion.

Generative models are typically used to implement the encoder-decoder functionality in singing voice conversion tasks. Autoregressive (AR) [7], [8] convert continuous waveforms into discrete tokens using neural audio encoders and decoders, training on discrete speech representations obtained through self-supervised learning. These models excel in zero-shot singing voice conversion, but suffer from slow inference speeds due to the recursive nature of predicting each subsequent token. To address this limitation, diffusion model-based SVC systems have been developed. Diffusion-based SVC systems [9] can generate high-quality audio with excellent fidelity and naturalness. However, their multiple iterative sampling steps also result in slower inference speeds. To improve this, the CoMoSVC model [10] was introduced, which employs a diffusion-based teacher model and further refines a student model under consistency constraints to enable single-step sampling. This approach not only significantly speeds up inference but also enhances the quality of the generated audio. Despite these improvements, CoMoSVC still struggles to achieve sufficient similarity to the target singer in cross-domain conversion tasks. In contrast, flow-based generative models [11]–[16] for SVC offer faster inference speeds compared to diffusion-based systems, because they perform inverse processing without requiring multiple sampling steps. For instance, the widely used So-VITS-SVC model is popular for its rapid inference capabilities. However, its audio quality still lags behind the state-of-the-art diffusion-based SVC systems. Additionally, these flow-based models typically incorporate only speaker identity and content embeddings, which results in generated songs that lacks naturalness and expressiveness. As a result, achieving effective and high-quality singing voice conversion remains challenging with current flow-based approaches.

A vocoder is a crucial component in singing voice conversion systems, responsible for transforming acoustic features such as Mel-spectrograms into audible waveform signals, typically serving as the final stage of the model. In recent years, Generative Adversarial Network (GAN)-based vocoders have gained significant attention due to their ability to efficiently generate high-quality waveforms. MelGAN [19], [20], the first GAN-based vocoder, utilizes a transposed convolution generator along with multi-scale and multi-resolution discriminators to achieve impressive songs synthesis without relying on additional distillation or perceptual losses. However, MelGAN encounters challenges when processing complex audio signals, particularly in accurately reconstructing high-frequency components. To address these limitations, HiFi-GAN [21] was developed with several enhancements. It incorporates Multi-Period Discriminators (MPD) and Multi-Scale Discriminators (MSD), which improve the discriminator's capability to distinguish between synthetic and real audio, thereby enhancing the quality of the generated waveforms. Despite these improvements, HiFi-GAN still exhibits a quality gap compared to autoregressive models in terms of sample fidelity. To further enhance audio quality and decoding speed, researchers proposed the improved RVQGAN model. RVQGAN [22], [23] employs vector quantization techniques and integrates a novel multi-band multi-scale Short-Time Fourier Transform (STFT) discriminator to reduce aliasing artifacts. Nevertheless, the computational complexity of RVQGAN remains high, making real-time applications challenging, especially in resource-constrained environments.

To solve the above problems, we propose a multi-condition based flow model asigned for singing voice conversion task. We use the HuBERT-Soft model [24] to complete content information extraction by modeling the distribution of discrete units ranther than units itself. We also use the flow model to accurately maximize the exact log-likelihood, transform a simple distribution to a complex one. Compared to the multi-step sampling strategy of diffusion models, the flow based approach only need reverse voice, which is faster. Additionally, we introduce timbre encoder, pitch encoder, and emotion encoder as conditions, which improves the information integrity of the generated voice. Finally, we choose MS-iSTFT ranther than traditional vocoder speed up the processing of the decoder module. In this study, we have achieved a fast and high-quality singing conversion model that achieves performance comparable to state-of-the-art (SOTA) models, as shown in Figure. 1. The main contributions of this paper can be summarized as follows:

- We propose to use the HuBERT-Soft model to gain soft singing's units by predicting a distribution over the discrete units, which effectively extract the content information in singing voice conversion.
- We propose a Multi-condition-based Flow, which not only extracts the speaker's timbre but also introduces pitch and emotion as the condition of the flow by extra extractors, which greatly improves the naturalness and expressiveness of singing voice conversion.
- We propose Multi-stream inverse short-time Fourier transform to directly convert from frequency domain features to time domain waveforms, which greatly enhances the speed of synthesis.
- We demonstrate the advantages of conditional generative voice synthesis and the effectiveness of flow model compared to diffusion model in singing voice conversion task.

II. Method

A. Overall pipeline

As shown in Fig.2, our model is based on VITS model [25] . The detailed procedures for training and inference are elaborated in Subsection F. Our model includes an F0 encoder, an emotion encoder, a speaker encoder, a content encoder, a Multi-condition flow, an MS-iSTFT-Decoder, and a discriminator. In the following sections, we focus on describing the aforementioned multiple encoders, the Multi-condition flow, the MS-iSTFT-Decoder, the loss function we used and the strategy in training and inference.

B. Multi-condition encoder

In this section, we will introduce four encoders used to assist voice generation before entering the flow-based model, namely

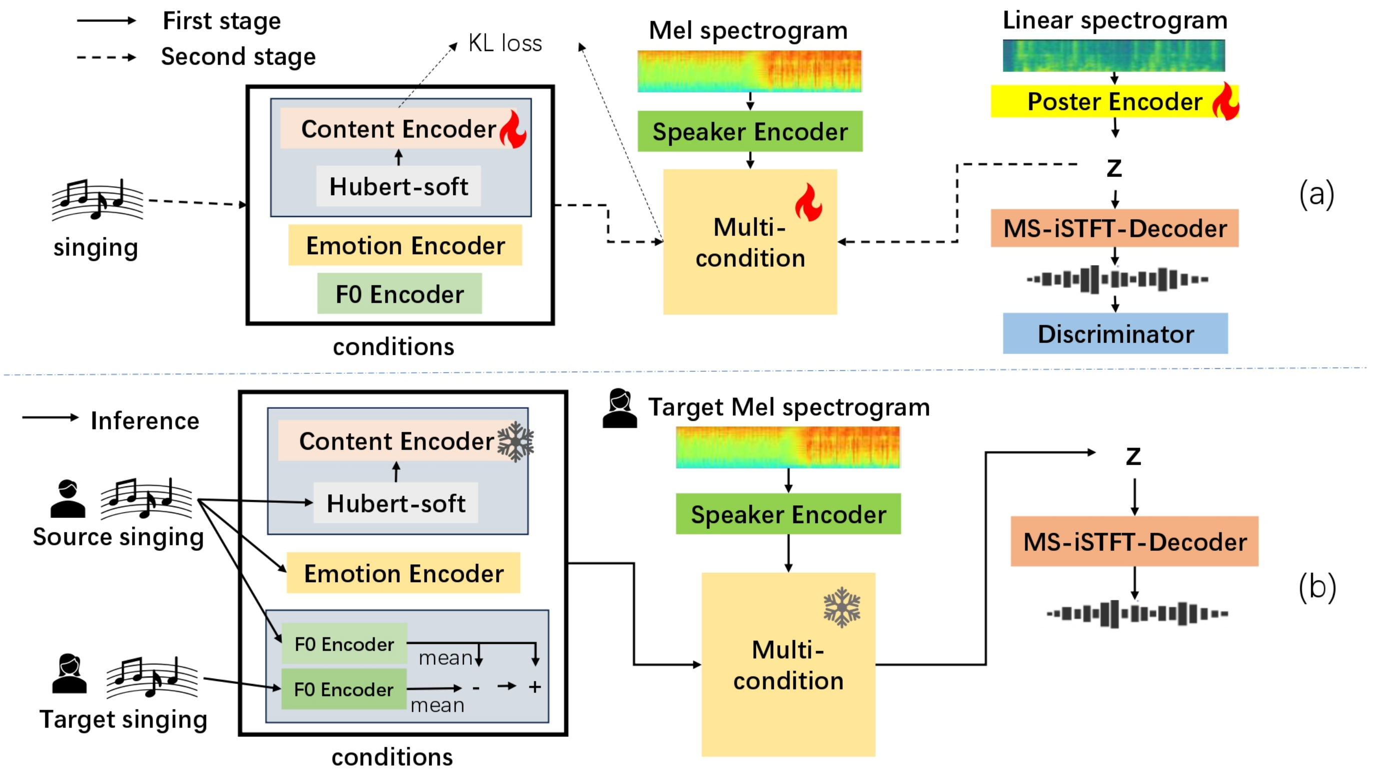

Fig. 2. Overview of **MCF-SVC**. (a) represents the training process of the model. The training process of the model is divided into two stages: the first stage generates reconstructed singing voice, and the second stage constructs a multi-conditional flow model to compute Kullback-Leibler(KL) loss (b) represents the inference process of the model, which utilizes the flow model in reverse to achieve the synthesized singing voice

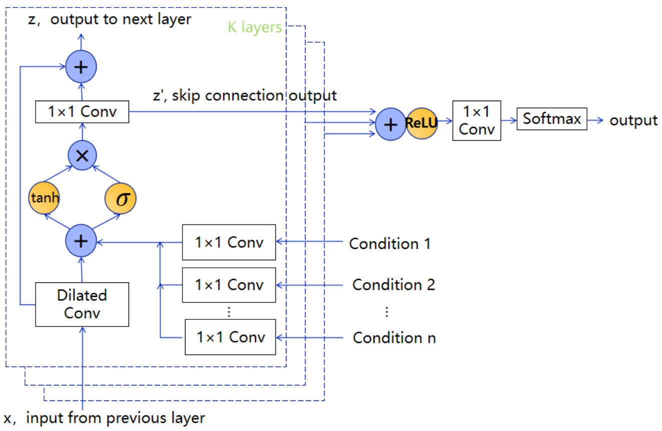

Fig. 3. Multi-condition flow model based WaveNet

content Encoder, speaker encoder, f0 encoder and emotion encoder.

1) Content Encoder: Content encoding includes the Hubert-soft model and a content encoder. A major challenge in self-supervised voice representation learning is that voice contains multiple units and there are typically no discrete words or characters as input. By inputting the source voice into the pre-trained Hubert-soft model, we learn the latent distribution of the discrete variables of songs, which results in an aligned sequence $h(x) = Z = [z_1, \ldots, z_T]$, where each unit has 1024-dimensional features. These features are then fed into the content encoder and transformed into a lower-dimensional content embedding Z_c.

2) Speaker Encoder: We input the Mel-spectrograms feature of the singer into a pre-trained LSTM framework to obtain embeddings represent identity information. LSTM can handle long-term dependencies in singing sequence data, which is crucial for capturing the global voiceprint features of a speaker. Additionally, LSTM can process input sequences of varying lengths, allowing it to flexibly adapt to the different lengths of singing's features from various speakers.

3) F0 Encoder: We employ a monophonic pitch tracker based on a deep convolutional neural network to obtain the continuous fundamental frequency (F0), which is a time-series feature. Considering the differences in fundamental frequencies between male and female voices can lead to unnatural-sounding singing, during the inference phase, we add the average difference over the time dimension in fundamental frequencies between the target and source voices to the source singing's F0.

$$f_0 = \left(\frac{1}{T_1} \sum_{i=1}^{T_1} f_0^{target}(i) - \frac{1}{T_2} \sum_{i=1}^{T_2} f_0^{source}(i) \right) + f_0^{source} \tag{1}$$

This results in a final F0 feature that not only preserves the dynamic characteristics of the original song's vocal track but also retains the vocal traits of the target speaker.

4) Emotion Encoder: We employ emotion2vec [26], a self-supervised and pre-trained universal model for emotional expression, to extract the emotional features of each song. This model was developed through self-supervised pre-training on 262 hours of open-source emotional data, utilizing an online distillation paradigm. It incorporates both sentence-level and frame-level losses to more effectively capture emotional nuances. The embeddings derived from emotion2vec serve as a constraint for our flow model, enhancing the naturalness of song conversion transitions and the expressiveness of the song's artistic conception.

C. Flow with multi-condition attribute constraints

Flow models have demonstrated excellent performance in voice conversion tasks, but they tend to underperform in singing voice conversion, where naturalness and expressiveness are of higher importance. To address this issue, we propose a multi-condition flow model that not only extracts the speaker's timbre but also incorporates pitch and emotion as additional conditions through extra encoders ahead. This enhancement significantly improves the naturalness and expressiveness of the converted singing voice.

Normalizing flows composed of multiple coupled affine layers transform a decomposed simple prior distribution into a more complex one. By applying reversible transformations to the simple prior distribution, they directly maximize the exact log-likelihood.

As shown in Fig.3, we utilize features extracted from multiple encoders as conditions for the normalizing flow based on WaveNet [32]. This approach enhances the expressiveness of the prior distribution, allowing it to better capture the distribution characteristics of real samples. We denote the standardized flow by f_θ. According to the variable transformation theorem, the prior distribution can be rewritten as:

$$p_\theta(z|c) = \mathcal{N}(f_\theta(z); \mu_\theta(c), \sigma_\theta(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right| \quad (2)$$

where c here is the output embedding of the priori encoder. c can be expressed as follows:

$$c = [C_{content}, C_{speaker}, C_{emotion}, C_{f0}] \quad (3)$$

Compared to the multi-step sampling strategy of diffusion models, the multi-condition flow model we proposed only requires reverse voice, which is faster. Additionally, it can synthesize natural and expressive singing voice conversions by incorporating multiple conditions.

D. MS-iSTFT-Decoder

As depicted in Fig.4, we utilize iSTFT to replace certain repetitive network layers in the previous HiFi-GAN vocoder by introducing the computation of phase and amplitude, converting latent embedding into continuous time-domain waveforms. This approach effectively reduces computational load and accelerates the audio synthesis process.

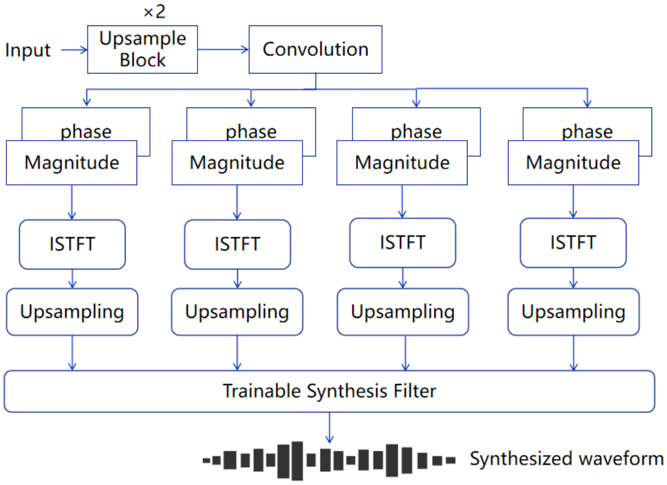

Fig. 4. The architecture of MS-iSTFT-Decoder

E. Multi-loss construction

Similar to VITS, we integrate Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) methodologies into our training process. The overall loss is articulated as follows:

$$L_{\text{total}} = L_{\text{recon}} + L_{\text{kl}} + L_{\text{adv}}(G) + L_{\text{fm}}(G) \quad (4)$$

We transform the generated waveform to the mel-spectrogram domain and calculate the L_1 loss against the source song's mel-spectrogram to serve as the reconstruction loss:

$$L_{\text{recon}} = \left| X_{\text{mel}} - \hat{X}_{\text{mel}} \right| \quad (5)$$

KL loss is used to narrow the gap between the prior encoder and the posterior encoder in terms of their distributions. The formula for KL loss is as follows:

$$L_{\text{kl}} = \log q_\phi(z|x_{\text{lin}}) - \log p_\theta(z|c) \quad (6)$$

Among which, x_{lin} represents the linear-scale spectrogram of the songs and the distribution of the posterior encoder can be expressed as:

$$q_\phi(z|x_{\text{lin}}) = \mathcal{N}(z; \mu_\phi(x_{\text{lin}}), \sigma_\phi(x_{\text{lin}})) \quad (7)$$

By introducing a pre-trained discriminator D, we verify the authenticity of the songs generated by the decoder G, leveraging $L_{\text{adv}}(G)$ for supervision. Furthermore, we incorporate an additional feature-matching Loss $L_{\text{fm}}(G)$ to ensure consistency in the reconstruction loss measured within the discriminator's hidden layers between the generated and authentic songs.

TABLE I
MODEL COMPARISON

Method	MOS/Naturalness	MOS/Similarity	Voice similarity	RTF(GPU)
FastSVC	3.52 ± 0.10	3.27 ± 0.22	0.433	0.031
SoVITS-Flow	3.10 ± 0.22	2.90 ± 0.23	0.585	0.008
CoMoSVC	**4.27 ± 0.16**	4.00 ± 0.19	0.585	**0.006**
DiffSVC	4.23 ± 0.19	3.95 ± 0.21	0.598	0.278
Multi-F0 Model	3.85 ± 0.02	3.47 ± 0.04	-	-
UCD-SVC	3.06 ± 0.08	2.67 ± 0.18	0.588	0.103
Ours	4.14 ± 0.17 ↓	**4.02 ± 0.19↑**	**0.603 ↑**	0.048 ↓

Compared with the indexes of the restored songs at 16kHz, many experiments have been done to get the average value, and bold indicates the best result.
↑ and ↓ represent a rise or fall in metrics, the higher the MOS and voice similarity, the better.

F. Training and inference strategy

As shown in Fig.2(a), during the training process, we input the same piece of singing into the network. The linear-scale spectrogram of the singing is sent to the Posterior Encoder to obtain the latent variable Z, which is then fed into the Decoder to produce reconstructed singing. We calculate L_{recon}, $L_{\text{adv}}(G)$, and $L_{\text{fm}}(G)$. Simultaneously, under the constraints of multiple disentangled attributes, the latent variable Z is transformed through a flow and the output is computed with the content embedding derived from the Content Encoder to getL_{kl} .

As shown in Fig.2(b), during the inference process, we input the singing from two different singers into the network. The content and emotion originate from the source singing, the timbre comes from the target singing, and the pitch is derived from a combined calculation of both the source and the target singing. These elements serve as constraints for the flow, resulting in the latent variable Z, which is then fed into the decoder to produce synthesized singing.

III. EXPERIMENT AND RESULTS

A. Datasets

We conduct experiments on four datasets: VCTK, Opensinger, M4singer and NUS-48E. We take the model weights that have been trained on the VCTK dataset (which contains 44 hours of speeches from 107 English speakers with various accents) as the starting point and then continue to train on the Opensinger dataset (a large-scale Chinese singing voice dataset) to learn the distinctive characteristics of Chinese singing voices, and observe the reconstruction results of singing audio. Finally, we use the M4singer dataset (a multi-style and multi-singer Chinese singing voice dataset) to test the evaluation of singing voice Reconstruction and combined the NUS-48E (an English singing voice dataset) to test the evaluation model's capabilities in zero-shot singing voice conversion (SVC) and cross-domain conversion.

B. Detail

In the training stage, the songs resampling frequency are 16khz, and the Mel-spectrogram is extracted by 512-point fast Fourier transform and 512-point window calculation. The model is preloaded with the pre-training weights of QuickVC [27] (trained on VCTK for 2 weeks), then trained on the Opensinger dataset for three days, and tested on M4singer and NUS-48E. At the same time, other experimental schemes such as FastSVC [28] and SoVITS-Flow are compared. All models have been fully iteratively trained on a single NVIDIA 3090ti, with a batch size of 64 and learning rates of 1e-4 and 5e-5 respectively.

C. Evaluation Metrics

We conducted subjective and objective experiments to comprehensively evaluate the model. In the subjective experiment, we used Mean Opinion Score(MOS) to evaluate the synthesized songs. We invited more than 100 people to score the similarity and naturalness of the songs, with five grades ranging from 0 to 5. In the objective metrics, we use Perceptual evaluation of speech quality(PESQ) to evaluate the quality of the reconstructed songs, which are also divided into five grades. At the same time, we use the advanced and trained ASR model to calculate the similarity before and after the singing voice conversion, and finally give the reasoning speed RTF of the model, which refers to the several seconds that the model can process per second. The model is carried out on a single NVIDIA GeForce RTX 3090ti GPU.

D. Results and analysis

1) Evaluation of singing voice Reconstruction: In the training process, the source and target vocals are from the same song, and the aim is to reconstruct the song, we use different competitive models to reconstruct several songs in the validation set, and we use metrics, such as PESQ and MOS , to assess the quality of the reconstructed songs, and the validation set consists of unseen songs from 40 singers from the M4singer dataset.

TABLE II
RECONSTRUCTION EXPERIMENT

Method	MOS/N	MOS/S	PESQ
FastSVC	3.98± 0.10	3.87± 0.09	-
SoVITS-Flow	4.15 ± 0.21	3.15 ± 0.17	2.486
CoMoSVC	**4.67 ± 0.12**	4.32 ± 0.21	**2.948**
DiffSVC	4.37 ± 0.02	4.01 ± 0.20	2.917
Multi-F0 Model	4.03 ± 0.19	3.53 ± 0.18	-
UCD-SVC	3.52 ± 0.13	3.01 ± 0.13	-
Ours	4.56 ± 0.07 ↓	**4.54 ± 0.21↑**	2.834

2) Evaluation of singing voice conversion: In the inference period of experiment, we use both objective and subjective evaluation metrics to compare our model with FastSVC, SOVITS-Flow[1], CoMoSVC [29], DiffSVC [30], Multi-F0 Model [31] and UCD-SVC [2] models. The evaluation objects are the singing voice conversion of different models in the dataset M4Singer → NUS-48E and the singing voice conversion across languages. Table 1 shows that, in terms of subjective metrics, the MOS naturalness of our proposed MCF-SVC model reaches 4.14, which exceeds all baseline models except CoMoSVC and DiffSVC, proving that the multi-condition strategy we joined did not reduce naturalness, but made the emotion and pitch of singing more accurate and rich; In terms of MOS similarity, our model scored higher in the target timbre and the converted timbre, surpassing all baseline models, which proves that MCF-SVC can be more consistent in the conversion processing. In terms of subjective metrics, our model maintains the highest voice similarity in cross-domain and cross-language singing voice conversion. At the same time, our model reasoning speed has also maintained a good level.

3) t-SNE visualization of converted songs: To verify how well our model maintain the identity information of the singers, we use the t-SNE, a dimensionality reduction algorithm, to plot identity information embeddings of the singers in two-dimensional space. As shown in Fig. 5, we randomly selected seven different singers , and performed multiple song conversions for a number of their songs within them. The results show that the identity information of the seven singers before and after the conversion is still concentrated, which indicates that our multi-conditional embedding does not destroy the singers' identity information, but also effectively enhances the timbre similarity of songs, bridges the timbres distance difference even between different songs, and learns reasonable representational identity information.

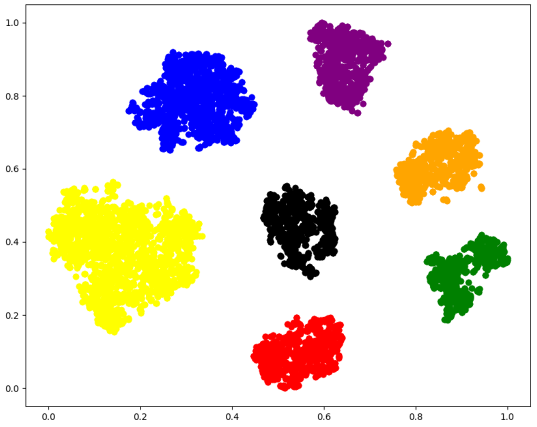

Fig. 5. t-SNE visualization of songs before and converted songs

4) Visualization analysis of Spectrogram: From Fig. 6, it can be seen that the resonance peaks move upward and the spacing of the vertical stripes is obviously widened when the male voice shifts to the female voice; when the female voice shifts to the male voice, the resonance peaks move downward and the spacing of the vertical stripes is obviously narrowed. Meanwhile, the main spectral features of the original singing voice (e.g., the overall energy distribution of the spectrogram) are preserved after both transformations. It can be seen that our model can not only accurately adjust the fundamental frequency (f0) in the transformation of male and female voices, but also effectively preserve the singing content.

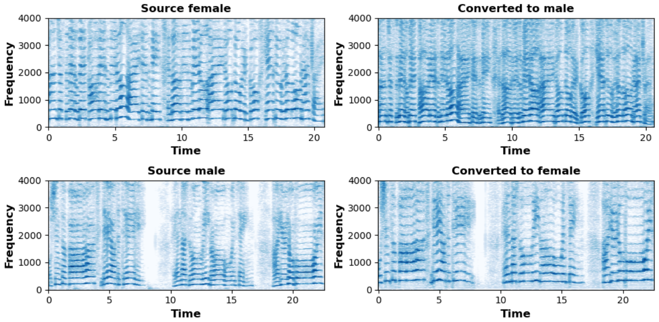

Fig. 6. Visualization analysis of Spectrogram

5) Contrast visualization of timbre similarity: We let the human judges listen to three songs first, namely: the target singer's song, the source singer's song, and the converted song. After that, we changed the songs sung by the target singer and the source singer and asked the judges to compare the similarities between the three songs and the source singer one by one, and then compare the similarities with the target singer. There are four kinds of ratings: same absolute certainty, same uncertainty, different uncertainty, and different absolute certainty. The results of the similarity comparison are shown in Fig. 7, and the evaluation results show that our model can transform the singing voice similarly to the target singer.

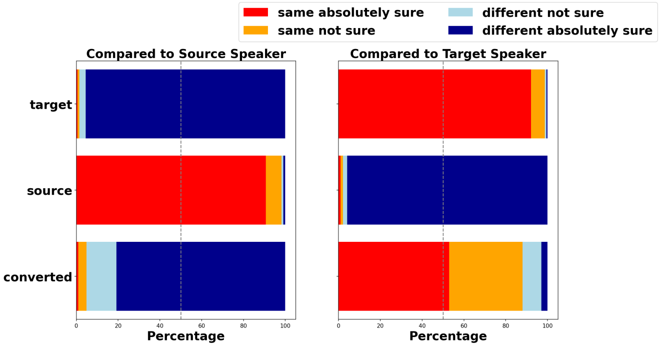

Fig. 7. visualization of timbre similarity

E. Ablation experiment

To demonstrate the effectiveness of the multi-conditional flow model, we conducted an ablation study and the demo is available in the link within the abstract. Specifically, we first

[1] https://github.com/svc-develop-team/so-vits-svc?tab=readme-ov-fle#sovits-model

individually removed the F0 encoder, speaker encoder, and emotion encoder from the model. Subsequently, we removed these encoders in pairs to further investigate their individual and combined contributions to the model's performance. Finally, we still used the subjective and objective metrics to evaluate them. According to Table 3, we can see that each condition we added contributes to the naturalness and similarity of the converted songs, with F0 encoder contributing the most to the naturalness and speaker Encoder contributing the most to the similarity.

TABLE III
ABLATION EXPERIMENT

Method	MOS/N	MOS/S	Voice similarity
Ours	4.14	4.02	0.603
- *F0 Encoder*	3.31	3.35	0.538
- *speaker Encoder*	3.45	3.20	0.522
- *emotion Encoder*	3.88	3.81	0.572
- *F0 and speaker Encoder*	2.97	2.62	0.361
- *speaker and emotion Encoder*	3.22	3.08	0.493
- *F0 and emotion Encoder*	2.82	3.00	0.472
- *F0 , emotion, speaker Encoder*	2.65	2.55	0.350

IV. CONCLUSION

In this paper, we propose a high-fidelity flow-based model based on multi-decoupling feature constraints. This model uses timbre, pitch, content, and emotion to assist the flow model in completing the song synthesis and conversion, and it is also used in diversified singing scenes. The inverse Fourier transform is also applied to the decoder to improve the conversion efficiency. The experimental results show that the naturalness and similarity of the songs after the conversion of our proposed model are 4.14 and 4.02. Finally, we have given a demo and will release a Pytorch trainer for singing voice conversion to promote further research in this field.

REFERENCES

[1] W.-C. Huang, H.-T. Hsu, X. Wang, and Y. Lee, "The singing voice conversion challenge 2023," in "Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)," IEEE, 2023.

[2] Polyak, A., Wolf, L., Adi, Y., and Taigman, Y. (2020). Unsupervised cross-domain singing voice conversion. unpublished.

[3] Z. Li, B. Tang, X. Yin, Y. Wan, L. Xu, C. Shen, and Z. Ma, "PPG-based singing voice conversion with adversarial representation learning," in "Proc. ICASSP," pp. 7073–7077, IEEE, 2021.

[4] H. Guo, H. Lu, N. Hu, C. Zhang, S. Yang, L. Xie, and D. Yu, "Phonetic posteriorgrams based many-to-many singing voice conversion via adversarial training," unpublished.

[5] K. W. Kim, S. W. Park, J. Lee, and M. C. Joe, "Assem-VC: Realistic voice conversion by assembling modern speech synthesis techniques," in "Proc. ICASSP," pp. 6997–7001, IEEE, 2022.

[6] B. Sha, X. Li, Z. Wu, Y. Shan, and H. Meng, "Neural concatenative singing voice conversion: Rethinking concatenation-based approach for one-shot singing voice conversion," in "Proc. ICASSP," pp. 12577–12581, IEEE, 2024.

[7] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, Y. Liu, J. Li, S. Zhao, X. Wu, and others, "Autoregressive speech synthesis without vector quantization," unpublished.

[8] Meng, L., Zhou, L., Liu, S., Chen, S., Han, B., Hu, S., Liu, Y., Li, J., Zhao, S., and Wu, X. (2024). Autoregressive speech synthesis without vector quantization. unpublished.

[9] Chen, S., Gu, Y., Zhang, J., Li, N., Chen, R., Chen, L., and Dai, L. (2024). LDM-SVC: Latent Diffusion Model Based Zero-Shot Any-to-Any Singing Voice Conversion with Singer Guidance. unpublished.

[10] Lu, Y., Ye, Z., Xue, W., Tan, X., Liu, Q., and Guo, Y. (2024). CoMoSVC: Consistency Model-based Singing Voice Conversion. unpublished.

[11] T. Merritt, A. Ezzerg, P. Biliński, M. Proszewska, K. Pokora, R. Barra-Chicote, and D. Korzekwa, "Text-free non-parallel many-to-many voice conversion using normalizing flow," in Proc. ICASSP, pp. 6782–6786, IEEE, 2022.

[12] Y. Guo, C. Du, Z. Ma, X. Chen, and K. Yu, "VoiceFlow: Efficient text-to-speech with rectified flow matching," in Proc. ICASSP, pp. 11121–11125, IEEE, 2024.

[13] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," unpublished.

[14] Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. "Density indication using real nvp." In Proc. ICLR 2017

[15] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," "Advances in Neural Information Processing Systems," vol. 33, pp. 8067–8077, 2020.

[16] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based articulatory-to-acoustic mapping with WaveGlow speech synthesis," unpublished.

[17] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Y. Liu, "MultiSpeech: Multi-speaker text to speech with transformer," unpublished.

[18] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Transformer-based text-to-speech with weighted forced attention," in Proc. ICASSP, pp. 6729–6733, IEEE, 2020.

[19] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in Proc. IEEE Spoken Language Technology Workshop (SLT), pp. 492–498, IEEE, 2021.

[20] Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., De Brebisson, A., Bengio, Y., and Courville, A. C. (2019). MelGAN: Generative adversarial networks for conditional waveform synthesis.

[21] Kong, J., Kim, J., and Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*, vol. 33, pp. 17022–17033.

[22] Lu, Y., Ye, Z., Xue, W., Tan, X., Liu, Q., and Guo, Y. (2024). CoMoSVC: Consistency Model-based Singing Voice Conversion. unpublished.

[23] Shechtman, S., and Dekel, A. (2024). Low bitrate high-quality RVQGAN-based discrete speech tokenizer. unpublished.

[24] B. Van Niekerk, M. A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in Proc. ICASSP, pp. 6562–6566, IEEE, 2022.

[25] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in Proc. Int. Conf. Machine Learning (ICML), pp. 5530–5540, PMLR, 2021.

[26] Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., and Chen, X. (2023). emotion2vec: Self-supervised pre-training for speech emotion representation. unpublished.

[27] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, "QUICKVC: A lightweight VITS-based any-to-many voice conversion model using ISTFT for faster conversion," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, IEEE, 2023.

[28] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, "FastSVC: Fast cross-domain singing voice conversion with feature-wise linear modulation," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), pp. 1–6, IEEE, 2021.

[29] Y. Lu, Z. Ye, W. Xue, X. Tan, Q. Liu, and Y. Guo, "CoMoSVC: Consistency model-based singing voice conversion," unpublished.

[30] S. Liu, Y. Cao, D. Su, and H. Meng, "DiffSVC: A diffusion probabilistic model for singing voice conversion," in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 741–748, IEEE, 2021.

[31] H. Cuesta, B. McFee, and E. Gómez, "Multiple F0 estimation in vocal ensembles using convolutional neural networks," unpublished.

[32] Van Den Oord, "Wavenet: A generative model for raw audio," unpublished.