InvoxSVC: Any-to-any Zero-shot Singing Voice Conversion with In-Context Learning in Latent Flow Matching

Wangjin Zhou[1], Tianjiao Du[2], Wenhao Guan[3], Meng Xiao[4], Chenglin Xu[5], Yi Zhao[6], Tatsuya Kawahara[1*]
[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[3]School of Electronic Science and Engineering, Xiamen University, Xiamen, China
[4]Faculty of Innovation Engineering, Macau University of Science and Technology, Macau
[5]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[6]Department of Electronic Engineering, Tsinghua University, Beijing, China
zhou@sap.ist.i.kyoto-u.ac.jp, dtj22@mails.tsinghua.edu.cn, whguan@stu.xmu.edu.cn, 3220002044@student.must.edu.mo,
cxu011@e.ntu.edu.sg, zy.2011@tsinghua.org.cn, kawahara@i.kyoto-u.ac.jp

2025 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3315-9495-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICME59968.2025.11210176

***Abstract*—Recent advancements in singing voice conversion (SVC) have focused on achieving zero-shot, any-to-any voice transformation capabilities. Many approaches attempt to modify voice characteristics by incorporating global timbre variables into acoustic models. However, these methods often depend heavily on the capabilities of timbre extractors and lack an understanding of temporal local information. This limitation poses challenges, particularly in replicating specific voice qualities such as those of children. To address this issue, we introduce InvoxSVC, a latent flow matching model (LFM) designed for rapid and precise singing voice conversion with a particular emphasis on capturing temporal local features. While reducing the residual timbral information in the source singing encoding through singer-guidance, InvoxSVC enhances the model's ability to capture temporal nuances by integrating in-context learning during inference. Additionally, the model employs a pre-trained high-fidelity variational autoencoder (VAE) to improve waveform generation. In comparative evaluations, InvoxSVC outperforms the open-source project So-VITS-SVC in both objective and subjective assessments.**

***Index Terms*—Singing Voice Conversion, Latent Flow Matching, VAE, In-context Learning.**

I. Introduction

Singing Voice Conversion (SVC) is a specialized form of voice conversion that transfers one singer's vocal performance to another while preserving the original melody, lyrics, and musical nuances [1]. The goal of SVC is to generate a new recording that convincingly mimics the target singer's performance without altering the underlying musical composition. This paper presents a zero-shot, any-to-any SVC framework distinguished by its strong generalization capabilities.

So far, a common approach for SVC is based on the following approach [2]–[6]: it involves decomposing a dry vocal signal into three fundamental components: content, pitch, and speaker identity. During the training phase, the model learns to reconstruct the singing waveform from these components using a vocoder. In the inference phase, the speaker identity is replaced with that of the target singer, enabling the transformation of the vocal timbre while maintaining the original melody and lyrics.

Recent advancements have refined the above framework by leveraging acoustic models to extract latent representations of content, pitch, and speaker identity. These representations are then converted back into waveforms using vocoders. For example, the open-source So-VITS-SVC project [1] utilizes a flow-based acoustic model, whereas Diffsvc [3] employs a diffusion model with a WaveNet-based [7] denoiser.

Both these methods rely on a global vector, such as x-vector [8], to represent the speaker's identity and generate the target speaker's waveform by modifying this vector. While these approaches capture certain aspects of vocal timbre, specific temporal characteristics such as pronunciation nuances are challenging to accurately describe with global information alone. This limitation arises because such global speaker vectors have inherent constraints. One possible issue is that these models are restricted by the distribution of the training data used to develop them. As a result, they struggle to represent singers whose voices differ from the data on which they were trained accurately. Another issue arises from the limitations of global vectors themselves, which are not good at representing local temporal details and features. In zero-shot any-to-any voice conversion tasks, this limitation reduces the range of vocal timbres that can be effectively transformed.

To address the limitations of singer embedding extraction, inspired by the second stage of existing zero-shot LM-based TTS systems [9]–[11], we incorporate a transformer-based U-Net [12] as the denoiser for the diffusion process, integrating in-context learning to effectively capture sequential features. Specifically, during inference, we concatenate the sequential features of the target singer's audio (encompassing both content and pitch) with those of the source singing. Flowing

*Corresponding authors.

[1]https://github.com/svc-develop-team/so-vits-svc

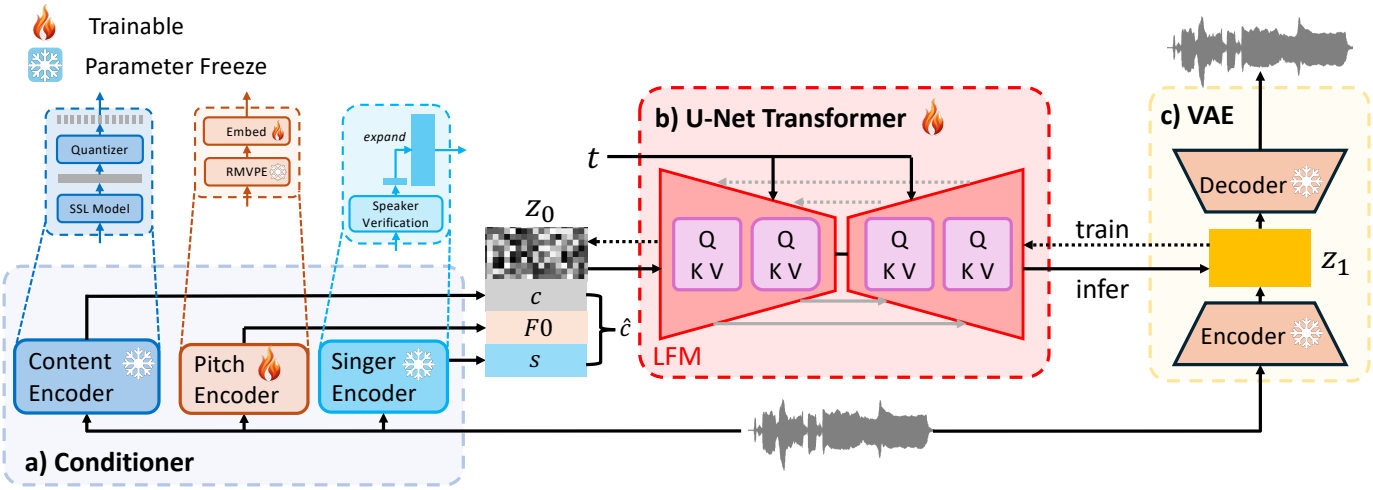

Fig. 1. Overview of InvoxSVC. InvoxSVC comprises 3 key modules: (a) the Conditioner, which encodes content, pitch, and global timbre information, aligns these features temporally, and concatenates them along the channel dimension; (b) the U-Net Transformer, which serves as the denoiser within the Latent Flow Matching (LFM) framework; and (c) the Variational Autoencoder (VAE), which directly encodes and decodes waveforms to enhance training efficiency and achieve higher fidelity waveform reconstruction.

classifier free guidance [13], [14], the final step involves removing the target singer's portion by singer-guidance and producing the transformed waveform.

In summary, this paper makes three key contributions:

- We propose InvoxSVC, a latent flow matching model for Singing Voice Conversion. Along with the guidance of the pre-trained global singer condition, InvoxSVC refines pronunciation details and singer identity information through in-context learning, resulting in enhanced SVC performance.
- We utilize a high-fidelity Variational Autoencoder (VAE) to achieve more precise and detailed waveform reconstruction.
- Our InvoxSVC achieves high-quality song conversion with just ten inference steps, effectively minimizing computational burden while maintaining performance. Audio samples are available at Demo Page [2]

II. Methodology

A. Conditioner

As illustrated in Figure 1, we use a pre-trained content encoder to extract phonetic information and a pitch encoder to capture pitch information (F_0). We also use the embedding extracted by the speaker verification (SV) model to provide singer-specific information (s). These three features are aligned via interpolation to match the longest sequence in the time dimension and are subsequently concatenated along the channel dimension to form the input concatenation condition: $\hat{c} = concat\{(c, F0, s), dim = Channel\}$.

1) Content Encoder: We utilize a pre-trained `Hubert Large` [15] model to extract acoustic features and obtain discrete tokens using a pre-trained K-means clustering model with 5,000 clusters. These discrete representations, denoted as c, can be viewed as pseudo-phone representations, capturing specific pronunciations in singing as well as the additional phonetic nuances introduced by various vocal techniques and styles.

2) Pitch Encoder: We utilize `RMVPE` [16] to extract pitch information $F0$, convert it to Log-F0, and input it into an embedding layer. During inference, we adjust the $F0$ of the original singer by mapping it to the target singer's vocal range. Specifically, we first calculate the mean $F0$ values for the target singer and the source singer, denoted as $mean(F0_{\text{src}})$ and $mean(F0_{\text{tgt}})$, respectively. We then adjust the source singer's F0 by multiplying it by the ratio of these mean values, resulting in the modified $F0_{\text{src}}^{\text{tgt}}$ as shown in Equation 1.

$$F0_{\text{src}}^{\text{tgt}} = F0_{\text{src}} \times \frac{mean(F0_{\text{tgt}})}{mean(F0_{\text{src}})} \tag{1}$$

3) Singer Encoder: We directly use a pre-trained Speaker Verification model Resnet34 [3] [17] implemented by WeSpeaker to extract a singer embedding and then expand along the time dimension to form a tensor s of the same length as c and $F0$, enabling concatenation along the channel dimension.

B. Latent Flow Matching

We employ Latent Flow Matching (LFM) [18] as the acoustic model to generate latent features that encode the target singer's information. These features are subsequently passed through the VAE [19] decoder to produce the converted singing voice.

For data x_1 sampled from latent space distribution $p_1(z_1)$, Flow Matching models a probability path x_t from a Gaussian distribution $p_0(z_0)$ to $p_1(z_1)$, as an ordinary differential equation (ODE): $dx_t = v_t(x_t)dt$, where v_t is the time-dependent vector field, and $t \in [0, 1]$. The probability path x_t describes

[2]Demo Page: https://expdemos.github.io/InvoxSVC/

[3]https://wespeaker-1256283475.cos.ap-shanghai.myqcloud.com/models/voxceleb/voxceleb_resnet34.zip

the probability distribution of the latent variable z_1 at time t, such that $\int p_t(x)\,dx = 1$. We adopt the optimal transport path, defined by a linear transformation between the data distribution and the Gaussian distribution as follows:

$$z_t = z_1 * t + (1 - (1 - \sigma_{\min})t)z_0, \quad t \in [0, 1] \tag{2}$$

where $\sigma_{\min}$ is small constant.

During training, a random time t is sampled according to a cosine distribution over the interval $[0, 1]$. And the sample point at time t is computed using Equation 2. The neural network θ is trained to learn the vector field at time t by minimizing the following objective:

$$\hat{\theta} = \arg\min_{\theta} E_{t,z_t} \left\| v_\theta(z_t, \hat{c}, t) - v_t \right\|^2. \tag{3}$$

where $v_\theta(z_t, \hat{c}, t)$ represents the predicted vector field and v_t is the target vector field at time t.

During inference, in our experiments, we generate the desired samples from a Gaussian distribution in just 10 steps using an Euler ODE solver [20], conditioned on $\hat{c}$.

Our proposed InvoxSVC leverages the U-Net transformer architecture from Matcha-TTS [21] as a denoiser, with eight downsampling layers and eight upsampling layers. The transformer's embedding dimension is set to 1024. The condition $\hat{c}$ is directly concatenated along the channel dimension with z_1 (during training) or z_0 (during inference). All other settings are consistent with those specified in Matcha-TTS.

C. VAE

We utilized a VAE to compress mono 44.1 kHz singing wav signals $\boldsymbol{X} \in \mathbb{R}^{1\times L}$ into a smaller latent space $\boldsymbol{z} \in \mathbb{R}^{C\times \frac{L}{r}}$, where C is the channel dimension and r is the downsample factor, to facilitate faster training and generation. We adopted the Variational Autoencoder (VAE) architecture based on the Descript Audio Codec (DAC) [22], [23], which employs fully convolutional layers for both the encoder and decoder. This design accommodates sequences of arbitrary length. Additionally, to enhance audio reconstruction quality, particularly at high compression ratios, we integrated the Snake activation function [24].

Building on the existing open-source DAC [4], we reduced the downsampling factor to 512 and replaced the VQ module with the variational module. In this modified setup, the encoder estimates the distribution of the latent variables, while the decoder reconstructs dense features sampled from this distribution. We trained this VAE on the AudioCaps [25], Common Voice [26], and internal music datasets, optimizing with a combination of reconstruction loss, adversarial loss, and a Gaussian constraint loss.

D. In-Context Learning Inference with Singer-Guidance

To capture the temporal characteristics of the target singer, we introduce an in-context learning inference method inspired by VALL-E [9], in which use target singing as prompt singing.

[4]https://github.com/Stability-AI/stable-audio-tools

Specifically, as illustrated in Figure 2, we extract c_{tgt}, $F0_{tgt}$, and s_{tgt} from the target singer and ensemble them from channel dimension as prompt feature set.

And we extract c_{src} and $F0_{src}$ from the singing of source singer. Further, using Formula 1, we calculate $F0^{tgt}_{src}$ and replace $F0_{src}$ with this adjusted $F0$. To implement voice conversion, the global singer feature of the source singing is replaced with s_{tgt} from the target singer. The c_{src}, $F0^{tgt}_{src}$ and s_{tgt} are concatenate as driven feature set.

The two feature sets are subsequently concatenated along the temporal dimension and input into a U-Net transformer for joint inference. The latent features obtained from this inference process are then cropped to isolate the portion corresponding to the driven feature set, which is denoted as $\overline{\overline{z_t}}$.

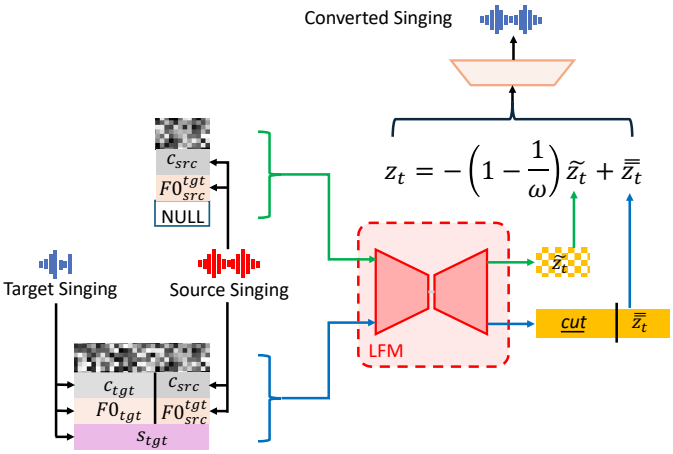

Fig. 2. In-Context Learning Inference with Singer-Guidance.

To further eliminate the residual timbre information in the source content representation c_{src}, we design a singer-guidance inference mechanism during the inference phase, which can be integrated with in-context learning, based on classifier-free guidance [13]. The inference process generates the intermediate VAE representation $\tilde{z}_t$ using only the c_{src}, $F0^{tgt}_{src}$, and an empty singer embedding $NULL$, processed through a U-Net transformer. The final latent representation z_t is obtained by subtracting a controlled portion of $\tilde{z}_t$ from $\overline{\overline{z_t}}$ as defined in Formula 4, where $\omega \in [1, +\infty)$ regulates the amount of information retained in $\tilde{z}_t$. The resulting z_t is then decoded into a waveform.

$$z_t = \overline{\overline{z_t}} - (1 - \frac{1}{\omega})\tilde{z}_t \tag{4}$$

III. Experiments

A. Experiment Setup

1) Dataset: We utilized an internal dataset consisting of 200 hours of singing in Chinese from 10,000 non-professional singers, with each singer contributing approximately one minute of singing. To evaluate model performance, we randomly selected 8 out-of-sample voices for timbre evaluation, including 2 adult males, 2 male children, 2 adult females, and 2 female children. Additionally, 20 out-of-sample songs, sung by

TABLE I

OBJECTIVE METRICS (FAD, FPC, SSIM, AND SUBJECTIVE METRICS (MOS, SMOS) UNDER VARIOUS SVC SYSTEMS. "INVOXSVC W/O ICL" REFERS TO INFERENCE WITHOUT THE INTEGRATION OF IN-CONTEXT LEARNING (ICL). THE PARAMETER ω IS USED TO CONTROL THE SINGER-GUIDANCE.

SVC system	ω	Objective Metrics			Subjective Metrics	
		FAD↓	FPC↑	SSIM↑	MOS↑	SMOS↑
DiffSVC	-	3.26	0.985	0.691	2.664 ± 0.113	3.038 ± 0.107
So-VITS-SVC	-	5.08	0.973	0.834	3.924 ± 0.087	3.410 ± 0.115
InvoxSVC w/o ICL	1	**2.59**	0.996	0.879	3.651 ± 0.091	3.733 ± 0.141
InvoxSVC w/o ICL	1.5	2.63	**0.999**	0.885	3.595 ± 0.103	3.762 ± 0.133
InvoxSVC	1	2.63	0.997	0.885	$\mathbf{4.135} \pm 0.097$	4.127 ± 0.124
InvoxSVC	1.5	2.81	0.994	**0.889**	4.122 ± 0.085	$\mathbf{4.254} \pm 0.132$

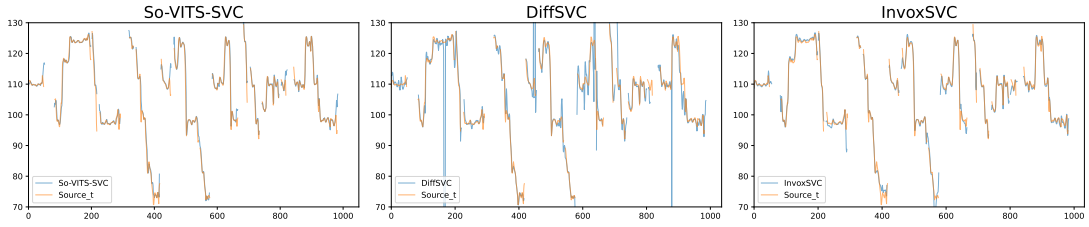

Fig. 3. A case study on the pitch reconstruction capabilities of different SVC models. Source_t stands for $F0_{src}^{tgt}$ which indicates pitch extracted from source singing and transfer by equation 1

different individuals, were randomly chosen to serve as source samples. The singers in the training dataset do not appear in the test phase. All singing waveforms are sampled at a rate of 44.1 kHz.

2) Experiment Details: We trained DiffSVC, So-VITS-SVC, and our proposed InvoxSVC on the same training dataset to ensure a fair comparison of model performance. The model inputs were standardized across all three models, with each utilizing content information c, $F0$ and s as described in Section II-A. Additionally, in Diffsvc, we employed the BigV-GAN [27] with NSF [28], as used in So-VITS-SVC. All models were trained with the AdamW optimizer on 8 A100 GPUs, using a mini-batch size of 80 for 100,000 steps. During training, we applied a learning rate warm-up strategy for the first epoch. The maximum learning rate was set to 1×10^{-5}, and was reduced by weight decay of 1×10^{-3} each epoch.

B. Evaluation Metrics

1) Objective Metrics: We use three kinds of objective metrics to evaluate the performance of the SVC models. To assess the preservation of pitch after voice conversion, we calculate the F0 Pearson Correlation (FPC) [14], which measures the correlation between the F0 values of the source and converted singing. To evaluate the voice conversion capability, we use Singer Similarity (SSIM), which involves extracting singer embeddings via a speaker verification (SV) model and calculating the cosine similarity between the embeddings of the source and converted singing. We utilized Wespeaker ResNet34 for extracting singer embeddings. Additionally, we use Fréchet Audio Distance (FAD) [29] to assess the realism of the generated audio. FAD is effective for evaluating audio quality in terms of human perception without requiring reference audio. Following the approach in [30], we used VGGish [31] as the audio embedding model, where a lower FAD score indicates better audio quality.

2) Subjective Metrics: For each experimental group, we selected 40 samples for subjective evaluation and invited 25 participants to assess them. The evaluation was conducted across two dimensions: (1) a 5-point Mean Opinion Score (MOS: 1-bad, 2-poor, 3-fair, 4-good, 5-excellent) for naturalness, and (2) a 5-point Similarity Mean Opinion Score (SMOS) for similarity.

IV. RESULTS

The out-of-sample experimental results are shown in Table I.

A. Effectiveness of DAC

Overall, under the same number of training steps, Invox SVC (including w/o ICL) consistently achieves better FAD scores compared to So-VITS-SVC and DiffSVC. Notably, So-VITS-SVC exhibits the lowest audio quality, which is attributed to its vocoder being jointly trained with the acoustic model. This confirms that using the Wav-VAE architecture facilitates achieving relatively high-fidelity waveforms more quickly.

B. Correctness of Content Reconstruction

All systems performed well on the FPC metric, with InvoxSVC achieving an FPC score very close to 1, indicating its superior performance. A case study as shown in Figure revealed that the noise and slight pitch distortions present in DiffSVC samples were not captured by this metric. This

discrepancy is reflected in the subjective MOS ratings, where DiffSVC received a significantly lower score than other systems. A possible reason for this is that the discrete token approach used in this experiment placed greater modeling pressure on DiffSVC.

When inference is performed without in-context learning, the naturalness MOS scores are lower than those of So-VITS-SVC. Upon inspecting the samples, we found that this issue arises from the segmentation and concatenation strategy applied during inference after VAE truncation, which causes timbral discontinuities between segments in some samples. These timbral discontinuities occur particularly when the target singing contains significant noise from the recording environment, negatively affecting the perceived naturalness of the entire song. By integrating in-context learning during inference, we concatenate the previously generated segment after the current prompt as the new prompt singing, effectively mitigating the issue of timbral discontinuity.

C. Similarity of Timbre

In terms of similarity, although the SSIM performance comparison indicates that our proposed InvoxSVC outperforms So-VITS-SVC, the SV model used for singer embedding extraction during training could not distinguish the performance differences between InvoxSVC inferred with or without prompt singing. In fact, noticeable differences in subjective evaluations were observed, with InvoxSVC receiving significantly higher SMOS scores with in-context learning compared with the same system without in-context learning during inference.

By examining specific samples, we found that the improvement in timbre primarily comes from certain articulations, while the global speaker embedding does not reflect this enhancement. Therefore, we believe that the gain brought by in-context learning is primarily related to temporal information. This temporal gain is also indirectly supported by the results of the FAD experiment, where the quality of the generated samples decreased due to the presence of varying levels of noise in the target singing when in-context learning was applied.

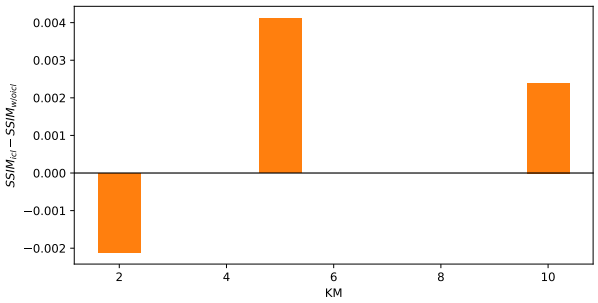

Fig. 4. Difference in timbre similarity of InvoxSVC with or without in-context learning on different center numbers.

D. Further Exploration of ICL in InvoxSVC

We trained InvoxSVC using a content encoder with different center numbers of K-means. Figure 4 reveals the difference in timbre similarity of InvoxSVC with or without in-context learning on different center numbers.

We observe that when the center number is around 5000, ICL achieves a relatively better gain. A study has explored the performance of different center numbers within the So-VITS-SVC framework [6]. And found that a larger center number leads to more timbral leakage from the source singing, while the encoding of vocal articulations becomes more detailed. Our study demonstrates that in-context learning requires a certain level of detailed articulation information. However, when the center number becomes sufficiently large, the inevitable timbral leakage from the source singing negatively impacts the effectiveness of in-context learning. Conversely, when the center number is too small, the prompt singing lacks sufficient detailed information, which results in a detrimental effect on in-context learning.

V. Conclusion

In summary, we present InvoxSVC, a novel and efficient zero-shot any-to-any Sing Voice Conversion. By employing an LFM based on a U-Net transformer architecture, we enhance the generation capabilities of the acoustic model, thereby improving the similarity and naturalness of converted singing. Additionally, we further strengthen similarity by incorporating in-context learning and singer guidance during inference to capture local temporal features of the target singing. Moreover, we achieve relatively high-fidelity waveform generation through the use of VAE.

References

[1] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, and T. Toda, "The singing voice conversion challenge 2023," in *Proc. IEEE-ASRu*. IEEE, 2023, pp. 1–8.

[2] E. Nachmani and L. Wolf, "Unsupervised singing voice conversion," *arXiv preprint arXiv:1904.06590*, 2019.

[3] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.

[4] Y. Lu, Z. Ye, W. Xue, X. Tan, Q. Liu, and Y. Guo, "Comosvc: Consistency model-based singing voice conversion," *arXiv preprint arXiv:2401.01792*, 2024.

[5] S. Chen, Y. Gu, J. Cui, J. Zhang, R. Chen, and L. Dai, "Lcm-svc: Latent diffusion model based singing voice conversion with inference acceleration via latent consistency distillation," *arXiv preprint arXiv:2408.12354*, 2024.

[6] W. Zhou, F. Zhang, Y. Liu, W. Guan, Y. Zhao, and H. Qu, "Zero-shot singing voice conversion: Built upon clustering-based phoneme representations," *arXiv preprint arXiv:2407.07728*, 2024.

[7] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, vol. 12, 2016.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[9] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, Y. Liu, J. Li, S. Zhao, X. Wu *et al.*, "Autoregressive speech synthesis without vector quantization," *arXiv preprint arXiv:2407.08551*, 2024.

[10] Z. Zhang, L. Zhou, C. Wang *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.
[11] Z. Du, Q. Chen, S. Zhang *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.
[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
[13] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.
[14] S. Chen, Y. Gu, J. Zhang, N. Li, R. Chen, L. Chen, and L. Dai, "Ldm-svc: Latent diffusion model based zero-shot any-to-any singing voice conversion with singer guidance," *arXiv preprint arXiv:2406.05325*, 2024.
[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
[16] H. Wei, X. Cao, T. Dan, and Y. Chen, "Rmvpe: A robust model for vocal pitch estimation in polyphonic music," *arXiv preprint arXiv:2306.15412*, 2023.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[18] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu *et al.*, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.
[19] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[20] C. Lu, Y. Zhou, F. Bao *et al.*, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 5775–5787.
[21] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-tts: A fast tts architecture with conditional flow matching," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 341–11 345.
[22] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *arXiv preprint arXiv:2306.06546*, 2023.
[23] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Long-form music generation with latent diffusion," *arXiv preprint arXiv:2404.10301*, 2024.
[24] L. Ziyin, T. Hartwig, and M. Ueda, "Neural networks fail to learn periodic functions and how to fix it," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1583–1594, 2020.
[25] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.
[26] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
[27] A. Brock, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.
[28] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5916–5920.
[29] K. Kilgour, M. Zuluaga, D. Roblek *et al.*, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.
[30] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[31] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.