

VibE-SVC: Vibrato Extraction with High-frequency F0 Contour for Singing Voice Conversion

Joon-Seung Choi, Dong-Min Byun, Hyung-Seok Oh, Seong-Whan Lee[†]

Department of Artificial Intelligence, Korea University, Seoul, Korea

js.choi@korea.ac.kr, dm.byun@korea.ac.kr, hs.oh@korea.ac.kr, sw.lee@korea.ac.kr

Abstract

Controlling singing style is crucial for achieving an expressive and natural singing voice. Among the various style factors, vibrato plays a key role in conveying emotions and enhancing musical depth. However, modeling vibrato remains challenging due to its dynamic nature, making it difficult to control in singing voice conversion. To address this, we propose VibE-SVC, a controllable singing voice conversion model that explicitly extracts and manipulates vibrato using discrete wavelet transform. Unlike previous methods that model vibrato implicitly, our approach decomposes the F0 contour into frequency components, enabling precise transfer. This allows vibrato control for enhanced flexibility. Experimental results show that VibE-SVC effectively transforms singing styles while preserving speaker similarity. Both subjective and objective evaluations confirm high-quality conversion.

Index Terms: singing voice conversion, singing style transfer, vibrato control, discrete wavelet transform

1. Introduction

Singing voice conversion (SVC) is a technique that converts a source singer voice into a target singer voice while retaining the source lyrics, melody, and styles. Recently, many singing voice models have been developed using various generative models [1, 2, 3, 4, 5]. Although SVC has improved significantly, several challenges remain. One of the most important challenges is effectively handling pitch. Since the singing voice is more expressive than speech, accurate pitch modeling is essential.

For this reason, several studies [6, 7] have been proposed to improve pitch-related performance. SVCC-T23 [6] extracts multi-scale F0 as an auxiliary input to better capture pitch variation in singing. SPA-SVC [7] introduces a cycle pitch shifting strategy to mitigate voiceless regions and hoarse artifacts caused by narrow pitch range of input dataset.

Some works [8, 9, 10] have focused on handling style characteristics of singing, including pitch styles. Vibrato control is achieved in [8] by extracting vibrato extent from the power spectrogram of the first-order difference. To synthesize and control singing styles, SinTechSVS [10] proposes style recommender and singing technique local score module. TCSinger [9] introduces clustering style encoder to capture singing styles. Unlike these methods which focus on synthesizing singing styles, other approaches [11, 12, 13, 14] focus on transferring singing styles. Since style features are not explicitly defined, these methods disentangle style information implicitly.

Extracting information via signal decomposition is a key focus in various deep learning studies [15, 16]. Discrete wavelet

transform is a method for decomposing an arbitrary signal into functions defined by discretely sampled wavelets. Wavelet transform has been utilized in various methods [17, 18, 19, 20] such as pitch representation [18], singer identification [19], or a downsampling method [20]. We assume that while the overall F0 contour remains consistent across singing styles, style-related variations are reflected in the high-frequency contour. To capture these differences, we use DWT to decompose the F0 contour into low- and high-frequency bands. Most existing methods aim to smooth the F0 contour to eliminate minor fluctuations or singing styles using filters such as median filter [21] or band-pass filter [22]. In contrast, our method employs DWT as a pass filter to disentangle and control singing styles.

In this work, we propose VibE-SVC, which disentangles vibrato style from singing voices and enables style transfer using DWT. By predicting the high-frequency F0 contour, VibE-SVC achieves transfer between straight and vibrato styles. Our approach demonstrates that DWT effectively separates singing styles and allows vibrato extent control without explicitly modeling the vibrato extent feature. Experimental results show that VibE-SVC successfully transfers singing styles. Audio samples are available at <https://castlechoi.github.io/VibE-SVC-demo>.

2. Method

We propose a controllable SVC model that enables style conversion between straight and vibrato. To separate style-related variations, we employ a DWT-based method to decompose the F0 contour into low- and high-frequency contours. The high-frequency contour is then predicted by a pitch style converter to perform singing style conversion. An overview of the VibE-SVC framework is shown in Figure 1.

2.1. Vibrato disentanglement

We use DWT to decompose a signal into approximation and detail coefficients, corresponding to the low- and high-frequency components. Approximation coefficient A and detail coefficient D are defined as follows:

$$A_j[k] = \sum_n x[n]\phi_{j,k}[n], \quad (1)$$

$$D_j[k] = \sum_n x[n]\psi_{j,k}[n], \quad (2)$$

where $x[n]$ denotes source signal. j , k , and n denote the decomposition level of DWT, the position of wavelet function, and the position of signal x , respectively. ϕ and ψ denote the scaling function and the wavelet function, respectively. To reconstruct the low- and high-frequency signals from the corresponding co-

[†]Corresponding author

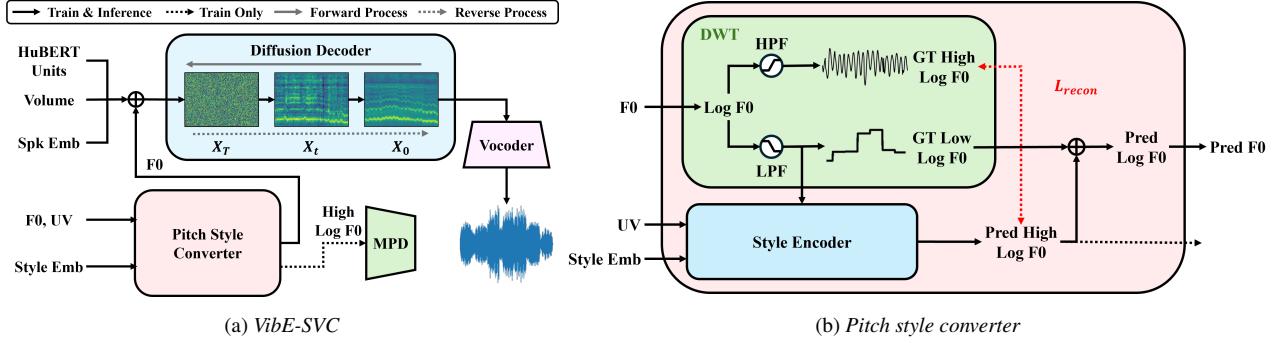


Figure 1: The overview of the proposed VibE-SVC model.

efficients, we use the inverse discrete wavelet transform (iDWT) as follows:

$$x[n] = \sum_k A_L[k] \phi_{L,k}[n] + \sum_{j=1}^L \sum_k D_j[k] \psi_{j,k}[n], \quad (3)$$

where L denotes the lowest frequency DWT level. Based on Equation 3, we reconstruct low- and high-frequency F0 contours from the approximation coefficient and detail coefficients. As shown in Figure 2, high-frequency F0 contour x_{high} and low-frequency F0 contour x_{low} are defined as follows:

$$x_{high}[n] = \sum_k A_L[k] \phi_{L,k}[n], \quad (4)$$

$$x_{low}[n] = \sum_{j=1}^L \sum_k D_j[k] \psi_{j,k}[n]. \quad (5)$$

Based on Equations 4 and 5, the source F0 contour is reconstructed by low- and high-frequency F0 contours as follows:

$$x[n] = x_{low}[n] + x_{high}[n]. \quad (6)$$

We adopt the Daubechies1 (db1) wavelet function [23] to enhance the extraction of consistent vibrato extent and facilitate model training. In DWT, the db1 wavelet function produces a rectangular function that aligns well with the characteristics of the musical instrument digital interface (MIDI).

2.2. Pitch style converter

As shown in Figure 1b, we use log F0 contour to normalize the vibrato scale across frequencies, and decompose it into frequency components using DWT. The low-frequency F0 contour, voice flag vector, and style embedding obtained from the style lookup table are concatenated and used as input to the style encoder to predict the high-frequency F0 contour as the target. After prediction, the log F0 contour converted to the target style is reconstructed by Equation 6. The predicted log F0 contour is then denormalized and fed into the diffusion decoder.

Since the vibrato rate is typically characterized by a frequency range of 5 to 8 Hz [8, 24], we adopt a multi-period discriminator (MPD) [25] to capture the periodic characteristics in the high-frequency F0 contour. The predicted and ground-truth high-frequency log F0 contour are used as the input to the MPD.

We train the model using reconstruction loss L_{recon} for log F0 contour, adversarial loss L_{adv} [26], and feature matching loss L_{fm} . The training loss functions are defined as follows:

$$L(G) = L_{recon}(G) + \mathcal{L}_{fm}(G) + \mathcal{L}_{adv}(G), \quad (7)$$

$$L(D) = L_{adv}(D), \quad (8)$$

where G denotes the generator and D denotes the discriminator.

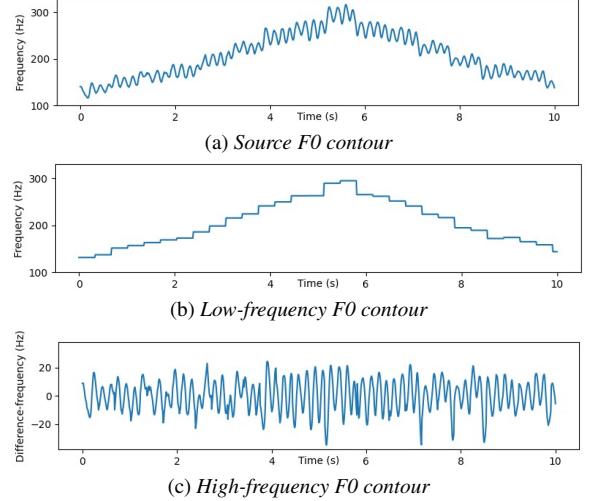


Figure 2: F0 contour disentanglement: (a) Source F0 contour, (b) Reconstructed F0 contour from approximation coefficient, and (c) Reconstructed F0 contour from detail coefficients.

2.3. Overall architecture

As shown in Figure 1, VibE-SVC consists of two main components: the SVC model and the pitch style converter, which are trained separately. The SVC model generates Mel-spectrograms conditioned on HuBERT [27] units, speaker embeddings, the F0 contour, and volume. The HuBERT units are interpolated to align with the F0 contour, while the volume is computed as the squared magnitude of the input audio. The SVC model is optimized with a simple diffusion loss, whereas the pitch style converter is trained according to the method described in Section 2.2. For training, the source speaker embedding and source F0 contour are used as inputs. During inference, we provide the target style embedding to predict the style-transferred F0 contour. Then, we provide the target speaker embedding and the predicted F0 contour as input to the SVC model. To adjust the pitch range, we scale the F0 contour by the ratio of the mean F0 values of the source and target speakers. The mean F0 value for each speaker is computed in advance.

3. Experiments

3.1. Dataset

We use VocalSet [28], a singing voice dataset labeled with 17 different styles. The dataset includes 20 singers, consisting of 11 males and 9 females. For our experiments, we select the

Table 1: Comparison results of style transfer. The MOS and SMOS are presented with 95% confidence intervals.

Method	Style-Only Conversion				Timbre & Style Conversion			
	MOS	SMOS	SECS	Acc	MOS	SMOS	SECS	Acc
GT	4.223 ± 0.08	3.355 ± 0.09	-	-	4.242 ± 0.09	3.549 ± 0.06	-	-
Vocoded	4.221 ± 0.07	3.340 ± 0.09	0.818	0.975	4.290 ± 0.09	3.535 ± 0.06	0.818	0.975
SoVITS w/ Style Emb	4.197 ± 0.08	2.911 ± 0.12	0.800	0.213	4.112 ± 0.11	3.221 ± 0.09	0.768	0.179
SoVITS w/ Style Emb & DWT	4.078 ± 0.09	2.872 ± 0.13	0.798	0.525	3.997 ± 0.12	3.224 ± 0.10	0.765	0.492
SoVITS w/ PST	4.043 ± 0.09	2.897 ± 0.13	0.804	0.425	4.035 ± 0.11	3.224 ± 0.09	0.774	0.434
VibE-SVC (Ours)	4.124 ± 0.08	2.911 ± 0.13	0.814	0.700	4.125 ± 0.10	3.196 ± 0.09	0.774	0.694
w/o MPD	4.016 ± 0.09	2.847 ± 0.13	0.809	0.625	3.997 ± 0.11	3.218 ± 0.09	0.773	0.611
w/o DWT	4.004 ± 0.10	2.887 ± 0.13	0.800	0.475	3.971 ± 0.12	3.213 ± 0.10	0.769	0.535

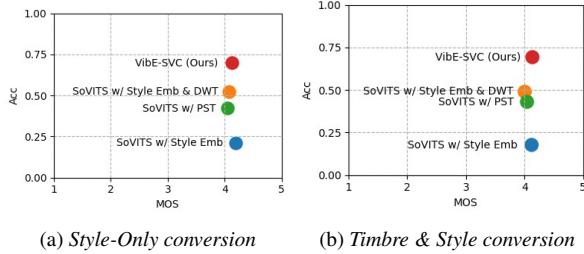


Figure 3: Correlation between MOS and style accuracy.

straight and vibrato styles from the dataset, resulting in a total of 753 samples. All audio recordings are downsampled to 24 kHz. We set the hop size to 256, the FFT size to 1024, and the window size to 1024 for Mel-spectrogram extraction. For training efficiency, each audio clip is segmented into 2-second chunks. For the test set, 4 samples from both straight and vibrato styles are selected per singer, making a total of 160 samples. The remaining data is used for training. The F0 contour and voiced flag vector are extracted using DIO [29].

3.2. Implementation details

VibE-SVC is based on an open-source SVC project², which uses a WaveNet [30]-based diffusion decoder with 1000 steps. The diffusion decoder consists of 20 layers in the residual blocks, 512 output channels in the convolutional layers, a hidden dimension of 256, and an encoder hidden layer size of 256. We use DPM-Solver++ [31] to denoise the diffusion model. We train the SVC model for 600K steps with a batch size of 128 and an initial learning rate of 3×10^{-4} . We use the learning rate schedule with a decay factor of $0.999^{1/8}$ updated at each epoch. We use the AdamW optimizer, with $\beta_1 = 0.8$ and $\beta_2 = 0.99$. We use automatic mixed precision with FP16 for efficient training. We utilize pretrained BigVGAN³ [32] as a vocoder.

The style encoder of the pitch style converter consists of 4 layers of feed-forward transformer (FFT) blocks [33], following the multi-layer perceptron (MLP) layer. The FFT blocks are set an encoder dropout rate of 0.2, a convolution kernel size of 31, and other hyperparameters are same as [33]. The style embedding dimension is set to 256. For the MPD, we follow the settings in [34]. We train the converter for 200K steps with a batch size of 64. We set an initial learning rate of 1×10^{-4} for the converter and 1×10^{-5} for the discriminator. The optimizer settings are the same as those in the SVC model.

²<https://github.com/svc-develop-team/so-vits-svc>
³https://huggingface.co/nvidia/bigvgan_v2_24khz_100band_256x

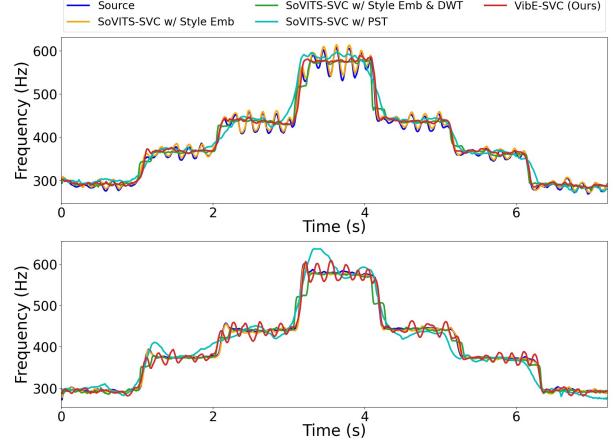


Figure 4: Comparison of F0 contours: vibrato-to-straight conversion (top) and straight-to-vibrato conversion (bottom).

3.3. Baseline models

We construct baseline models with style embedding based on the SoVITS framework. The first baseline, SoVITS with style embedding, incorporates a style embedding directly on the diffusion decoder, enabling the model to learn style information. The hidden dimension is set to 256. The second baseline, SoVITS with style embedding and DWT, shares the same architecture but replaces the source F0 contour with its low-frequency contour. This design allows the model to capture style information by reconstructing the source F0 contour. We include the SoVITS with F0 model, employed in the performance style transfer (PST) [11]. To predict singing styles, we replace the speaker embedding with a style embedding.

3.4. Evaluation metrics

For objective evaluation, we evaluate speaker similarity using speaker encoder cosine similarity (SECS) with Resemblyzer⁴. To evaluate style transfer quality, we use MERT [35], a self-supervised model designed for music information retrieval tasks. We use a pretrained checkpoint⁵ and fine-tune an additional MLP layer to classify singing styles as either straight or vibrato. For subjective evaluation, we conduct 5-point mean opinion score (MOS) and 4-point similarity mean opinion score (SMOS) tests via Amazon MTurk where at least 20 participants evaluate the naturalness and speaker similarity of 50 samples per model.

⁴<https://github.com/resemble-ai/Resemblyzer>

⁵<https://huggingface.co/m-a-p/MERT-v1-330M>

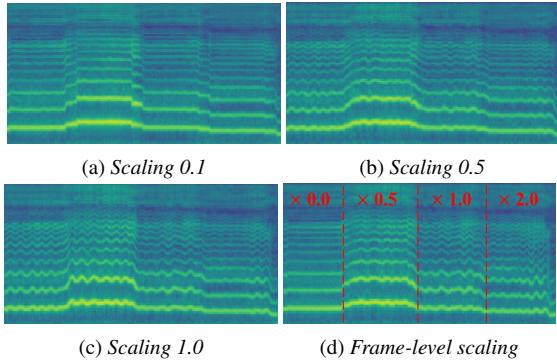


Figure 5: (a), (b), and (c) show global-level vibrato scaling; (d) shows frame-level vibrato scaling.

4. Results

4.1. Style transfer

4.1.1. Objective evaluation

We conduct two experiments to evaluate style transfer performance. The first focuses on style-only conversion, aiming to assess style transfer without considering speaker conversion quality. The second experiment evaluates pitch style transfer performance in a joint timbre and style conversion setting to assess the model’s adaptation ability for the SVC task. As shown in Table 1, our method outperforms baseline models in SECS and style accuracy, demonstrating superior timbre and style conversion in both experiments. The experimental results confirm that our model converts singing styles more effectively than baseline model while preserving speaker similarity. The consistent style accuracy in both experiments shows that our model preserves style information after conversion.

4.1.2. Subjective evaluation

As shown in Table 1, our method achieves comparable results in subjective evaluations in terms of speaker similarity and naturalness. In both experiments, our model shows slight difference results with 95% confidence intervals in naturalness and speaker similarity across baseline models. Figure 3 shows the trade-off between naturalness and style transfer performance among baseline models in both experiments. SoVITS with style embedding achieves the highest naturalness but has the lowest style accuracy, whereas other baseline models exhibit higher style transfer performance at the cost of reduced naturalness. Although our model shows slightly lower naturalness in style-only conversion and slightly lower speaker similarity in timbre and style conversion, the model shows highest style accuracy, balancing effectively this trade-off. These results indicate that the model successfully converts the styles without affecting naturalness and speaker similarity performance.

4.1.3. Comparison of the F0 contour

Figure 4 compares the converted F0 contours for vibrato-to-straight and straight-to-vibrato conversions. In the vibrato-to-straight conversion, the baseline models fail to fully remove vibrato or produce unnatural contours. In contrast, our method generates a relatively flat and natural F0 contour, effectively removing the vibrato. For the straight-to-vibrato conversion, the baseline models struggle to produce noticeable vibrato patterns. However, VibE-SVC successfully generates a clear and natural vibrato, demonstrating superior style conversion performance.

Table 2: The style accuracy with various vibrato scaling.

Scaling Factor	0.1	0.3	0.5	0.7	1.0	2.0
Style Acc	0.066	0.093	0.217	0.421	0.690	0.928

Table 3: Ablation study for disentanglement level of DWT.

Level	MOS	SMOS	SECS	Acc
3	3.905 ± 0.09	2.797 ± 0.09	0.775	0.163
4	3.777 ± 0.10	2.796 ± 0.10	0.774	0.694
5	3.646 ± 0.11	2.792 ± 0.09	0.772	0.694

4.2. Vibrato scaling control

We conduct an experiment to evaluate the fine-grained vibrato control capability of VibE-SVC. As shown in Figure 5, the model adjusts vibrato extent by multiplying the high-frequency F0 contour with a constant value before adding the low-frequency F0 contour during inference. Frame-level vibrato scaling is also possible by applying scaling at specific target indices. To verify that the high-frequency F0 contour contains style information, we perform straight-to-vibrato conversion experiments with different scaling factors using the finetuned MERT [35]. As shown in Table 2, the decrease in style accuracy with scaling factors confirms our assumption. VibE-SVC is also able to emphasize vibrato by scaling the high-frequency F0 contour to 2, as shown in Figure 5d.

4.3. Ablation studies

We conduct ablation studies to evaluate the effectiveness of each component in VibE-SVC. As shown in Table 1, removing the MPD decreases naturalness and style accuracy, confirming its importance for singing style modeling. Replacing the high-frequency F0 prediction with the source F0 prediction further reduces both metrics, highlighting the effectiveness of the DWT-based disentanglement method.

To analyze the effect of DWT decomposition levels on performance, we conduct additional experiments. Table 3 presents the results for different DWT levels. Level 3 shows the lowest style accuracy due to the absence of vibrato information in high-frequency F0 contour. Although level 5 captures more high-frequency information, it slightly underperforms compared to level 4 because it contains unrelated information beyond vibrato styles. Based on these results, level 4 is chosen as the optimal decomposition level.

5. Conclusions

In this paper, we propose a controllable SVC model that designed to adjust singing styles such as vibrato. To disentangle singing style from the F0 contour, we propose a method that decomposes it into low- and high-frequency components using DWT. By predicting the high-frequency F0 contour, the model convert arbitrary inputs into either straight or vibrato styles. Experimental results from both objective and subjective evaluations demonstrate the effectiveness of our approach. The current model focuses on controlling vibrato, serving as a starting point for exploring a wider range of singing styles. In future work, we plan to extend the model to capture additional styles related to timbre and aperiodic pitch variations, further improving its applicability.

6. Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (Artificial Intelligence Graduate School Program (Korea University) (No. RS-2019-II190079), Artificial Intelligence Innovation Hub (No. RS-2021-II212068), AI Technology for Interactive Communication of Language Impaired Individuals (No. RS-2024-00336673), and Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents (IITP-2025-RS-2025-02304828)).

7. References

- [1] E. Nachmani and L. Wolf, “Unsupervised singing voice conversion,” in *Proc. Interspeech*, 2019, pp. 2583–2587.
- [2] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, “Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 7237–7241.
- [3] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, “Diffssinger: Singing voice synthesis via shallow diffusion mechanism,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 11 020–11 028.
- [4] S. Liu, Y. Cao, D. Su, and H. Meng, “Diffsvc: A diffusion probabilistic model for singing voice conversion,” in *Proc. IEEE Autom. Speech Recognit. Underst. Workshop (ASRU)*, 2021, pp. 741–748.
- [5] D.-M. Byun, S.-H. Lee, J.-S. Hwang, and S.-W. Lee, “Midi-voice: Expressive zero-shot singing voice synthesis via midi-driven priors,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 12 622–12 626.
- [6] Z. Ning, Y. Jiang, Z. Wang, B. Zhang, and L. Xie, “Vits-based singing voice conversion leveraging whisper and multi-scale f0 modeling,” in *Proc. IEEE Autom. Speech Recognit. Underst. Workshop (ASRU)*, 2023, pp. 1–8.
- [7] B. Bai, F. Wang, Y. Gao, and Y. Li, “Spa-svc: Self-supervised pitch augmentation for singing voice conversion,” in *Proc. Interspeech*, 2024, pp. 4353–4357.
- [8] R. Liu, X. Wen, C. Lu, L. Song, and J. S. Sung, “Vibrato learning in multi-singer singing voice synthesis,” in *Proc. IEEE Autom. Speech Recognit. Underst. Workshop (ASRU)*, 2021, pp. 773–779.
- [9] Y. Zhang *et al.*, “TCSinger: Zero-shot singing voice synthesis with style transfer and multi-level style control,” in *Proc. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2024, pp. 1960–1975.
- [10] J. Zhao, L. Q. H. Chetwin, and Y. Wang, “Sintechsvs: A singing technique controllable singing voice synthesis system,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 2641–2653, 2024.
- [11] Y.-T. Hsu, J.-Y. Wang, and J.-S. R. Jang, “Many-to-many singing performance style transfer on pitch and energy contours,” *IEEE Signal Process. Lett.*, pp. 166–170, 2025.
- [12] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, “Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 3277–3281.
- [13] P.-W. Chen and V.-W. Soo, “A few shot learning of singing technique conversion based on cycle consistency generative adversarial networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 1–5.
- [14] C. Wang *et al.*, “Towards high-fidelity singing voice conversion with acoustic reference and contrastive predictive coding,” in *Proc. Interspeech*, 2022, pp. 4287–4291.
- [15] Y.-K. Lim, S.-H. Choi, and S.-W. Lee, “Text extraction in mpeg compressed video for content-based indexing,” in *Proc. 15th Int. Conf. Pattern Recognit.*, 2000, pp. 409–412.
- [16] J. Kim, J. Schultz, T. Rohe, C. Wallraven, S.-W. Lee, and H. H. Bühlhoff, “Abstract representations of associated emotions in the human brain,” *J. Neurosci.*, pp. 5655–5663, 2015.
- [17] S.-W. Lee, C.-H. Kim, H. Ma, and Y. Y. Tang, “Multiresolution recognition of unconstrained handwritten numerals with wavelet transform and multilayer cluster neural network,” *Pattern Recognit.*, pp. 1953–1961, 1996.
- [18] Y. Ren *et al.*, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [19] V. D. Noyum, Y. P. Mofenjou, C. Feudjio, A. Göktug, and E. Fokoué, “Boosting the predictive accuracy of singer identification using discrete wavelet transform for feature extraction,” *arXiv:2102.00550*, 2021.
- [20] S.-H. Lee, J.-H. Kim, K.-E. Lee, and S.-W. Lee, “Fre-gan 2: Fast and efficient frequency-consistent audio synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 6192–6196.
- [21] T. Kim, C. Cho, and Y. H. Lee, “Period singer: Integrating periodic and aperiodic variational autoencoders for natural-sounding end-to-end singing voice synthesis,” in *Proc. Interspeech*, 2024, pp. 1875–1879.
- [22] Y. Song *et al.*, “Singing voice synthesis with vibrato modeling and latent energy representation,” in *Proc. IEEE Intl. Workshop on Multimedia Signal Processing (MMSP)*, 2022, pp. 1–6.
- [23] I. Daubechies, “Ten lectures on wavelets,” *Society for industrial and applied mathematics*, 1992.
- [24] T. Nakano, M. Goto, and Y. Hiraga, “An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features,” in *Proc. Interspeech*, 2006, pp. 1706–1709.
- [25] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 17 022–17 033.
- [26] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2794–2802.
- [27] 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 Trans. Audio, Speech, Lang. Process.*, p. 3451–3460, 2021.
- [28] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, “Vocalset: A singing voice dataset,” in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 468–474.
- [29] M. Morise, H. Kawahara, and H. Katayose, “Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech,” in *Proc. AES 35th International Conference*, 2009, pp. CD-ROM.
- [30] A. Van Den Oord *et al.*, “Wavenet: A generative model for raw audio,” *arXiv:1609.03499*, 2016.
- [31] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, “Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models,” *arXiv:2211.01095*, 2022.
- [32] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “BigVGAN: A universal neural vocoder with large-scale training,” in *Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [33] Y. Ren *et al.*, “Fastspeech: Fast, robust and controllable text to speech,” in *Neural Inf. Process. Syst. (NeurIPS)*, 2019.
- [34] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 5530–5540.
- [35] Y. Li *et al.*, “MERT: Acoustic music understanding model with large-scale self-supervised training,” in *Int. Conf. Learn. Represent. (ICLR)*, 2024.