2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) | 979-8-3503-2372-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/WASPAA58266.2023.10248127

DIFF-PITCHER: DIFFUSION-BASED SINGING VOICE PITCH CORRECTION

*Jiarui Hai, Mounya Elhilali**

Laboratory for Computational Auditory Perception, Johns Hopkins University, Baltimore, USA
{jhai2, mounya}@jhu.edu

ABSTRACT

Pitch correction is the process of adjusting the original pitch of a recording or live performance in order to fit it to specific key or match a target profile. Pitch correction systems typical consist of several stages: original pitch estimation, pitch curve modification, and resynthesis of the audio with the target pitch curve. Unfortunately, the process of resynthesis often leads to significant artifacts that degrade the overall quality of the modified audio, rendering it unnatural and unpleasant. In this work, we introduce Diff-Pitcher[1], a pitch control model that leverages diffusion modeling and source-filter mechanisms to generate high-quality and natural-sounding voice signal matched to a target pitch while ensuring content and timbre consistency. To demonstrate the effectiveness of the proposed method, we evaluate Diff-Pitcher by both subjective and objective experiments in scenarios of pitch shifting and automatic pitch correction. Our results show that Diff-Pitcher outperforms previous pitch control methods in sound-quality and naturalness with great pitch controllability. Furthermore, we apply Diff-Pitcher in template-based and score-based automatic pitch correction systems and explore their application potentials. Meanwhile, for score-based automatic pitch correction, we improve the pitch predictor proposed in KaraTuner to handle variable-length inputs.

Index Terms— Pitch correction, singing voice synthesis, diffusion probabilistic model

1. INTRODUCTION

Pitch correction is an important process of sonic refinement that is integral to the music industry. It involves the delicate adjustment of musical notes to harmonize singing intonation, transforming off-key performances into melodious output that resonates with listeners. This critical system involves several stages, starting with the initial pitch estimation, followed by a detailed pitch curve modification, and culminating in the final resynthesis of the audio using the calibrated target pitch curve.

During the first two stages, there are two primary pitch correction tasks: manual pitch editing and automatic pitch correction (APC). Manual pitch editing, typically performed by a music producer or recording engineer, requires domain expertise and involves the editing of the pitch curve to the desired score. This process, while effective and accurate, is time-consuming and requires professional training. On the other hand, APC can automatically generate the target in-tune pitch curve, making it a more fast and user-friendly option for average users. APC employs three strategies: scale-based, template-based, and score-based. Scale-based approaches, such as those used in Antares Auto-Tune, shift each vocal note to the nearest one in a chosen scale. Deep Autotuner [1], a data-driven model, predicts pitch shifts based on the accompaniment. Template-based methods, like [2], transfer the pitch curve from a professional recording to user's singing, creating natural vocals but requiring temporal alignment. Score-based methods, like [3] and KaraTuner [4], use MIDI sequences to generate target pitch curves. The former employs a rule-based algorithm while the latter uses a vocal-adaptable pitch predictor. Score-based APC creates a balance between scale-based and template-based methods.

The process of resynthesizing the audio signal with the target pitch contour is also a critical component in a pitch correction system. Methods based on phase vocoder and source filter vocoders are widely used in commercial pitch correction software. TD-PSOLA [5] is a phase vocoder which extracts pitch periods from a monophonic voice and shifts copies of the pitch period to change the fundamental frequency. Source filter-based methods such as Linear Predictive Coding (LPC) [6] and WORLD [7] vocoder represent speech or singing signal as a combination of a sound source excitation and an acoustic filter, which can be considered as vocal cord and vocal tract, and control pitch by changing the fundamental frequency of the sound excitation impulse. In recent years, neural network-based pitch controllable vocoders have become increasingly popular to obtain better sound quality and naturalness. LPCNet [8] incorporates LPC in the neural network to synthesize natural speech signals and is capable of pitch-shifting. CLPCNet [9] further improves the capability of pitch-shifting of LPCNet. KaraTuner [4] proposes a pitch-controlled vocoder that replaces the spectrogram input in the Fre-GAN vocoder with the spectral envelope and pitch contour. Source-Filter HiFi-GAN (SiFi-GAN) [10] proposes a hierarchical fusion framework which has a source excitation network and a acoustic filter network based on the HiFi-GAN architecture.

In recent, Diffusion Probabilistic Models (DPMs) [11] have become increasingly prevalent in the field of generation tasks due to their remarkable performance and stable training characteristics, particularly when compared to Generative Adversarial Networks. The fruitful research conducted at the intersection of DPMs and audio signal generation and synthesis tasks has led to significant advances and has demonstrated superior capabilities. Various applications of DPMs have emerged, including neural vocoder [12], voice conversion [13], singing voice synthesis [14], and text-to-audio generation [15].

In this study, we present Diff-Pitcher, a diffusion-based pitch control model designed for pitch correction systems. Our work makes the following contributions: (1) We propose the first diffusion-based pitch control model Diff-Pitcher to the best of our knowledge. The proposed method achieves excellent pitch controllability and can generate natural and pleasant voices. (2) We improve the vocal-adaptable pitch predictor proposed in KaraTuner for score-based APC, enabling it to handle variable-length inputs. (3) We apply Diff-Pitcher in templated-based and score-based APC

*Authors supported by ONR N00014-23-1-2050

[1]Code and audio samples: https://jhu-lcap.github.io/Diff-Pitcher/

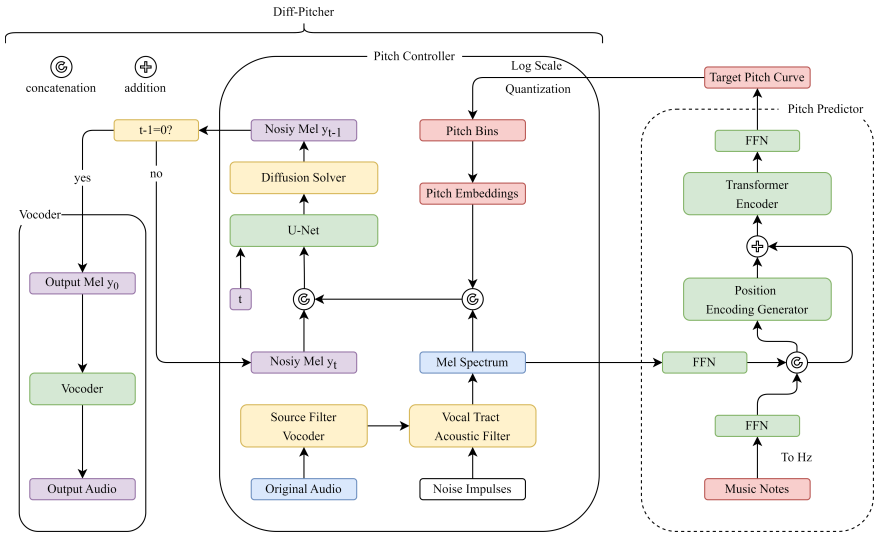

Figure 1: The inference framework of the pitch correction system utilizing Diff-Pitcher. The source filter vocoder is either WORLD or LPC vocoder. The denoising solver is the Stochastic Differential Equation or the Ordinary Differential Equation solver used for diffusion sampling.

and explore their advantages, limitations, and practical applications.

2. METHODOLOGY

As shown in Figure 1, there are three modules in a automatic pitch correction system: a diffusion-based pitch controller for generating spectrograms with target pitch contours; a neural vocoder for waveform reconstruction; and a vocal-adaptive pitch predictor for predicting the desired pitch contour based on out-of-tune vocals and target notes. The pitch controller and vocoder together constitute a pitch-controllable vocoder. The pitch predictor is used in the score-based APC system. Each module is independently trained, and they are combined during the inference process.

2.1. Diffusion-based Pitch Controller

We adapt the diffusion model from DDPMs [11]. The diffusion model models the conditional distribution $p_\theta(y_0|x)$ where y_0 is the target spectrogram and x contains the conditioning information including target pitch and vocal spectrum.

The forward process is a Markov chain with fixed parameters, and it convert y_0 into the latent y_T in T steps:

$$q(y_{1:T}|y_0) := \prod_{t=1}^{T} q(y_t|y_{t-1}) \tag{1}$$

A Gaussian noise is added to y_{t-1} at each diffusion step t:

$$q(y_t|y_{t-1}) := \mathcal{N}(y_t; \sqrt{1-\beta_t}y_{t-1}, \beta_t I) \tag{2}$$

under a variance schedule $\beta = \beta_1, ..., \beta_T$. The diffusion process can be calculated in a closed form for any step t:

$$y_t = \sqrt{\overline{\alpha}_t}y_0 + \sqrt{(1-\overline{\alpha}_t)}\epsilon \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t := 1 - \beta_t$ and $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

The reverse process is also a Markov chain, and the reverse transition distribution can be approximated by a diffusion decoder with parameters θ. To learn the parameters θ, we optimize:

$$\mathbb{E}_{t,\epsilon}[C_t||\epsilon_\theta(\sqrt{\overline{\alpha}_t}y_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, x, t) - \epsilon||_2^2] \tag{4}$$

where C_t is a constant related to β_t. In practice, the C_t term is dropped.

During the sampling process, a Gaussian noise y_T is first sampled, and then the final y_0 is generated by iteratively sampling y_{t-1} for $t = T, T-1, ..., 1$ along the reverse process:

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}}(y_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(y_t, x, t)) + \sigma_t z \tag{5}$$

where $\sigma_t^2 = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$, $z \sim \mathcal{N}(0, I)$ for $t > 1$ and $z = 0$ for $t = 1$.

The diffusion decoder is trained to predict the noise ϵ added in the forward diffusion process given the noisy mel spectrogram y_t, the diffusion step t, the pitch f_0, and the mel spectrum sp. The diffusion step t is specified by adding the sinusoidal position embedding [16] into each block of the decoder. The continuous pitch curve f_0 is converted into the log scale, quantized into pitch bins, and encoded with the sinusoidal position encoding. This approach allows the pitch controller to utilize its generative capacity to adaptively adjust the pitch to better match the vocal spectrum, enhancing naturalness.Assuming that the impulse source is independent of the vocal tract, we apply a source-filter algorithm - either WORLD or LPC vocoder - to extract the vocal spectrum. Vocal spectrum retains content and timbre information while disentangling the pitch information. Following [17], we convert the vocal spectrum into the log-mel spectrum sp to match the shape of the generation target of the diffusion model. As a result, the diffusion decoder can be trained in a self-supervised fashion, taking the original pitch curve and vocal spectrum as conditioning variables. During the inference stage, we can manipulate the vocal pitch by using a customized pitch curve as input. The architecture of the diffusion decoder is based on U-Net. Following [13], we employ linear attention layers within each downsample and upsample block, enabling the model to effectively capture temporal-frequency dependencies.

2.2. Neural Vocoder

Different from KaraTuner and SiFi-GAN, our pitch controller does not directly synthesize waveform signals. Instead, it first produces a mel-spectrogram and subsequently utilizes a neural vocoder to reconstruct the waveform. With this strategy, we significantly reduce the training cost of the pitch controller. For high-fidelity waveform reconstruction, we employ BigVGAN [18], a GAN-based large-scale pre-trained neural vocoder. BigVGAN integrates a periodic activation function and an anti-aliasing representation in its GAN generator, which significantly improve audio quality and robustness to unseen voice. Experimental results in [18] demonstrate that BigVGAN can reconstruct high-quality singing voice.

2.3. Vocal-Adaptable Pitch Predictor

The voice-adaptive predictor, first presented in Karatuner [4], leverages the target musical notes sequence and the out-of-tune vocal spectrum to generate a custom, in-tune pitch contour for score-based APC. The musical score forms the pitch curve's backbone, with the vocal spectrum providing fine-grained details such as voiceless and voiced states, gliding, and vibrato. The original pitch predictor from KaraTuner, based on a Transformer encoder with the absolute position encoding, struggles with input sequences exceeding its training data length. We fix this issue by using the Position Encoding Generator (PEG) [19]. The PEG, a zero-padding convolution layer, extends the Transformer's ability to handle variable-length sequences without performance sacrifices. Since the vocal spectrum may contain pitch information, we introduce a random pitch shift to musical notes and target pitch during training. This forces the model to rely on target notes to predict the shifted pitch curve, and thus generates a scaled pitch contour. The pitch predictor is then trained using L1 loss optimization between the predicted and shifted pitch curves.

3. EXPERIMENTS

3.1. Dataset

To ensure our pitch controller is capable of handling various voices, we utilize several speech and singing datasets for training. These include the VCTK multi-speaker English speech dataset [20], the OpenSinger multi-singer Chinese singing dataset [21], and the PopBuTFy multi-singer English singing dataset [22]. The PopBuTFy dataset is particularly interesting as it includes pairs of amateur and professional recordings of the same song, with some amateur versions being off-key by several semitones. This makes the dataset valuable for evaluating the performance of the pitch controller in template-based APC. For validation and evaluation experiments, we randomly select 8 singers from the PopBuTFy dataset.

We employ two singing datasets with paired recordings and MIDI annotations to train our voice-adaptive pitch predictor. These are the CSD dataset [23], comprising 50 Korean and 50 English songs performed by a single professional singer, and the Opencpop dataset [24], consisting of 100 Chinese songs sung by a professional singer. Five songs from the Opencpop dataset are used for validation, while testing is performed on five different songs from the PopCS dataset [14], sung by an unseen artist.

For the assessment of our system's effectiveness in severe out-of-tune situations, we collect a small dataset featuring paired out-of-tune and in-tune vocals from 5 singers across 5 Chinese pop songs. In most instances, the out-of-tune notes deviate randomly by more than 3 semitones. Each in-tune vocal has its own manually annotated MIDI sequence, using the standard music sheet format. In this case, the audio and music notes are not perfectly time-aligned, better reflecting real-world score-based APC scenes.

3.2. Implementation Details

In our experiments, we resample all audio samples to a frequency of 24 kHz. We convert the waveform into mel-spectrograms using the frame size of 1024, the hop size of 256, and 100 mel bins. The same frame size and hop size are used for source-filter algorithms to extract the vocal spectrum. The order of LPC in the LPC Vocoder is set as 16. We employ DIO [25] algorithm to estimate f_0 during training and inference.

The U-Net architecture of the diffusion-based pitch controller comprises 3 downsampling and 3 upsampling blocks. To improve computational efficiency during the training process, we limit the input to 128 frames, corresponding to a 1.36-second audio clip. We set T to 1000 and the β value to constants increasing linearly from $\beta_0 = 0.0001$ to $\beta_T = 0.02$. The pitch controller is trained with the AdamW optimizer, using a constant learning rate of 5×10^{-5}. The voice-adaptive pitch predictor's transformer encoder consists of two self-attention blocks. We train the pitch predictor using the AdamW optimizer with a constant learning rate of 1×10^{-4}. Both the pitch controller and the pitch predictor are trained on a single NVIDIA RTX 3090 GPU.

3.3. Experiment1: Pitch Controller Performance

We use two established pitch-controllable vocoders as baselines, the algorithm-based WORLD Vocoder and the latest neural network SiFi-GAN. We propose two pitch control models, each composed of a diffusion-based pitch controller and a BigVAN vocoder. Diff-Pitcher-LPC uses the vocal spectrum extracted from the LPC vocoder, while Diff-Pitcher-WORLD uses the vocal spectrum extracted via the cheaptrick algorithm in the WORLD vocoder.

We evaluate these models via objective and subjective tests. In the objective experiment, we use 200 clips from the PopBuTFy test set for pitch shifting and reconstruction, assessing pitch controllability with the Root Mean Square Error of log f_0 (RMSE) between the pitch curve of the synthesized audio and the target pitch curve, and content consistency with the Word Error Rate (WER) of lyrics transcribed from original and synthesized audio. For automatic lyric transcription, we employ Whisper [26], a versatile multilingual automatic speech recognition model which shows promising accuracy in lyrics transcription. In the subjective experiment, we apply four models in the template-based APC task. This task offers more variable and challenging target pitch curves than pitch shifting, and it eliminates any unnaturalness in the pitch curve caused by the pitch predictor in score-based APC. Though the in-tune and de-tune recordings are rhythmically aligned, we further align the pitch contour of the reference audio with the target audio using a dynamic time warping algorithm based on mel-cepstral coefficients. Sound quality (noisy vs. clean) and naturalness (robotic voice vs. close to human voice) are estimated with five scaled Mean Opinion Scores (MOS) tests on 15 audio clips, 6 from the PopBuTFy test set and 9 from our custom dataset, with feedback collected from 12 subjects experienced in studio recording or music production, each randomly evaluating 8 clips per method.

The objective experiment results shown Table 1 can be summaried as follows: (1) Diff-Pitcher-WORLD is comparable to SiFi-GAN in terms of RMSE, but slightly inferior to the algorithm-based

Table 1: Objective evaluations results of pitch controllers.

Model	−6 Semitones		−3 Semitones		Reconstruction		+3 Semitones		+6 Semitones	
	RMSE ↓	WER ↓	RMSE ↓	WER ↓	RMSE ↓	WER ↓	RMSE ↓	WER ↓	RMSE ↓	WER ↓
WORLD	0.03	2.00%	0.03	2.68%	0.03	3.23%	0.03	3.04%	0.03	2.17%
SiFi-GAN	0.05	2.98%	0.04	2.85%	0.04	3.35%	0.05	2.98%	0.04	3.61%
Diff-Pitcher-LPC	0.05	4.12%	0.04	4.33%	0.04	4.64%	0.07	5.21%	0.16	6.51%
Diff-Pitcher-WOLRD	0.04	2.57%	0.03	2.41%	0.04	2.88%	0.04	2.64%	0.06	2.23%

Table 2: Mean Opinion Score evaluation results with their 95% confidence intervals of pitch controllers. MOS-Q measures sound quality and MOS-N measures naturalness.

Method	MOS-Q ↑	MOS-N ↑
WORLD	2.97 ± 0.18	3.10 ± 0.19
SiFi-GAN	3.51 ± 0.20	3.33 ± 0.20
Diff-Pitcher-LPC	3.43 ± 0.17	3.35 ± 0.17
Diff-Pitcher-WOLRD	$\mathbf{3.59 \pm 0.18}$	$\mathbf{3.65 \pm 0.17}$

Table 3: Pitch prediction RMSE of pitch predictors.

Method	256 frames (1x)	512 frames (2x)
Baseline (KaraTuner)	0.09	Unsupported
Improved Pitch Predictor	0.08	0.09

WORLD Vocoder. However, this RMSE is at a low level, proving its ability to control pitch effectively. (2) Diff-Pitcher-LPC performs well in downshifting and reconstruction, but struggles with upshifting, particularly for high-pitched female voices. (3) All three methods except Diff-Pitcher-LPC show low WERs, indicating promising content consistency. Diff-Pitcher-LPC, however, occasionally mispronounces similar consonants.

The subjective experiment results presented in Table 2 indicate that Diff-Pitcher-WORLD achieves the highest MOS-Q and MOS-N scores, showing its superior performance on audio quality and naturalness as the pitch controller. Specifically, we find that the WORLD Vocoder usually generates noticeable and undesirable artifacts. SiFi-GAN, while exhibiting good sound quality, sometimes produces robotic voices. While Diff-Pitcher-LPC can generate natural-sounding voices, the audio it generates lacks clarity, leading to a decrease in sound quality.

3.4. Experiment2: Pitch Predictor Performance

We compare our enhanced pitch predictor with the one from KaraTuner, both trained on 256-frame (2.72s) audio clips. As shown in Table 3, they perform comparably when tested on data with the same length as the training data. The baseline model cannot handle longer segments limited by the absolute position encoding. In contrast, the performance of the improved model remains stable, demonstrating its capacity of handling variable-length inputs.

3.5. Experiment3: Template-based APC and Score-based APC

To explore the advantages, limitations, and practical application potentials of template-based and score-based APC systems utilizing

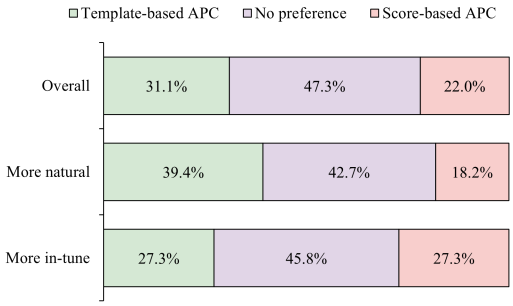

Figure 2: Preference between score-based and template-based automatic pitch correction.

Diff-Pitcher, we conduct a subjective A/B test comparing pitch accuracy, naturalness, and overall quality of these two methods. We select 12 audio pairs for this experiment. The same subjects from Experiment1 evaluate all audio pairs. The Diff-Pitcher-WORLD is employed to synthesize vocal with the target pitch curve.

The experimental results, depicted in Figure 2, highlight that: (1) Both methods can produce in-tune vocals. (2) Template-based APC system usually yields more natural sounds than score-based APC, but in most cases the gap between them is not significant. (3) Despite the score-based method sometimes offering less natural sounds, the overall listening experiences of two methods are closed in most scenarios. Specifically, we notice that the lack of naturalness in score-based APC is primarily due to the challenges in representing glides and vibratos with discrete musical notes, and the pitch predictor's inability to predict them from the formant spectrogram when they are absent in out-of-tune vocals. However, as Diff-Pitcher employs pitch quantization and is able to adaptively adjust fine-grained pitches, the overly smooth pitch curves occasionally predicted by the pitch predictor do not result in distinctly robotic sounds. In summary, with Diff-Pitcher, score-based APC remains a viable choice when pitch templates are unavailable or when there is a desire to preserve the original singing techniques of the voice.

4. CONCLUSION

This paper presents a novel pitch correction solution that employs a diffusion-based pitch control model Diff-Pitcher. Given Diff-Pitcher's excellent ability in controlling pitch and generating high-quality and natural singing voices, it can be applied to various pitch correction scenarios, such as pitch correction plugins in DAWs and Auto-Tune modules in Karaoke Apps. In future work, we plan to further enhance the naturalness of the score-based APC and optimize the diffusion model's sampling process for faster inference.

5. REFERENCES

[1] S. Wager, G. Tzanetakis, C.-i. Wang, and M. Kim, "Deep autotuner: A pitch correcting network for singing performances," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 246–250.

[2] Y.-J. Luo, M.-T. Chen, T.-S. Chi, and L. Su, "Singing voice correction using canonical time warping," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 156–160.

[3] O. Perrotin and C. d'Alessandro, "Target acquisition vs. expressive motion: dynamic pitch warping for intonation correction," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 23, no. 3, pp. 1–21, 2016.

[4] X. Zhuang, H. Yu, W. Zhao, T. Jiang, and P. Hu, "KaraTuner: Towards End-to-End Natural Pitch Correction for Singing Voice in Karaoke," in *Proc. Interspeech 2022*, 2022, pp. 4262–4266.

[5] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11. IEEE, 1986, pp. 2015–2018.

[6] B. E. Caspers and B. S. Atal, "Changing pitch and duration in lpc synthesized speech using multipulse excitation," *The Journal of the Acoustical Society of America*, vol. 73, no. S1, pp. S5–S5, 1983.

[7] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[8] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[9] M. Morrison, Z. Jin, N. J. Bryan, J.-P. Caceres, and B. Pardo, "Neural pitch-shifting and time-stretching with controllable lpcnet," *arXiv preprint arXiv:2110.02360*, 2021.

[10] R. Yoneyama, Y.-C. Wu, and T. Toda, "Source-filter hifi-gan: Fast and pitch controllable high-fidelity neural vocoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[12] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," in *International Conference on Learning Representations*, 2020.

[13] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *International Conference on Learning Representations*, 2021.

[14] 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.

[15] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] S. Nercessian, "Differentiable world synthesizer-based neural vocoder with application to end-to-end audio style transfer," in *Audio Engineering Society Convention 154*. Audio Engineering Society, 2023.

[18] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," in *The Eleventh International Conference on Learning Representations*, 2022.

[19] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," in *The Eleventh International Conference on Learning Representations*, 2022.

[20] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[21] R. Huang, F. Chen, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3945–3954.

[22] J. Liu, C. Li, Y. Ren, Z. Zhu, and Z. Zhao, "Learning the beauty in songs: Neural singing voice beautifier," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7970–7983.

[23] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, "Children's song dataset for singing voice research," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[24] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, "Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 4242–4246.

[25] 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 *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.