DiTVC: One-Shot Voice Conversion via Diffusion Transformer with Environment and Speaking Rate Cloning

Yunyun Wang[1,2], *Jiaqi Su*[2], *Adam Finkelstein*[1], *Rithesh Kumar*[2], *Ke Chen*[2], *Zeyu Jin*[2]

[1]Princeton University [2]Adobe Research

***Abstract*—Traditional zero-shot voice conversion methods typically extract a speaker embedding from a reference recording first and then generate the source speech content in the target speaker's voice by conditioning on that embedding. However, this process often overlooks time-dependent speaker characteristics, such as voice dynamics and speaking rates, as well as environmental acoustic properties of the reference recording. To address these limitations, we propose a one-shot voice conversion framework capable of replicating not only voice timbre but also acoustic properties. Our model is built upon Diffusion Transformers (DiT) and conditioned on a designed content representation for acoustic cloning. Besides, we introduce specific augmentations during training to enable accurate speaking rate cloning. Both objective and subjective evaluations demonstrate that our method outperforms existing approaches in terms of audio quality, speaker similarity, and environmental acoustic similarity, while effectively capturing the speaking rate distribution of target speakers. Audio samples are available at: ditvc.github.io.**

1. INTRODUCTION

Voice conversion is a core task in speech processing that involves transforming a source utterance to sound as if it were spoken by a target speaker, given a reference recording. Recent deep learning advances have greatly improved voice conversion, often producing speech nearly indistinguishable from real recordings. Most current approaches [1]–[5] also support one-shot voice conversion, where the model can mimic a target speaker using only a single short reference sample. Since these methods rely on speaker embeddings, it is also possible to generate novel voices by sampling embeddings without a reference, enabling zero-shot voice conversion.

However, conventional voice conversion settings face several methodological limitations. First, speaker identity is shaped not only by timbre but also by speaking style—such as prosody and rhythm. Yet many models preserve the source speaker's prosody and rhythm entirely, which can reduce perceptual similarity to the target and lead to lower speaker similarity scores in subjective evaluations. Second, most models assume clean source speech and produce clean audio, ignoring the reference's environmental acoustics. As a result, they perform poorly in real-world applications such as dubbing or automated dialogue replacement (ADR), where environmental consistency is crucial. Third, while speaker embeddings aim to capture all aspects of identity, they often fail with unconventional voices—like cartoonish or highly emotional speech—resulting in degraded synthesis quality.

Therefore, we aim to alleviate these constraints: first, by allowing rhythm and prosody to adapt to the target speaker; second, by jointly modeling speaker identity and acoustic environment; and third, by removing the reliance on speaker embeddings to enable one-shot voice conversion directly from a reference sample.

Previous state-of-the-art voice conversion methods have predominantly relied on GAN-based frameworks. Several approaches [2], [4], [6] operate on mel-spectrograms and use a vocoder—such as HiFi-GAN [7] or BigVGAN [8]—to reconstruct the waveform. Alternatively, other methods [3], [5] adopt fully end-to-end architectures with adversarial training to synthesize waveforms directly. However, GAN-based training often introduces artifacts and struggles to model multi-modal distributions, which limits the diversity and expressiveness of generated speech—especially in our setting where rhythm and acoustics are also converted.

To address these limitations, we adopt diffusion models [9], which have recently demonstrated impressive results across image [10], video [11], and audio [12] generation tasks. In voice conversion, DiffVC [13] was one of the first to apply diffusion modeling, aiming to generate prosody instead of copying pitch from the source. However, in practice, DiffVC often produces unnatural prosody [14] and, like many GAN-based methods, still relies on mel-spectrograms and an external vocoder for waveform synthesis. More recently, CoDiff-VC [14] introduced a codec-based diffusion model that integrates a speaker encoder for end-to-end voice conversion. While effective, its dependence on a speaker encoder can reduce expressiveness and fidelity, especially for atypical or emotionally rich speech.

Our method builds upon recent advances in diffusion-based text-to-speech (TTS) models [12], [15]–[17], but is specifically designed for the voice conversion setting. Traditional TTS systems [18], [19] usually rely on text input or speaker embeddings. To incorporate speaker identity and environmental acoustics, we adopt a one-shot TTS approach in which the target voice is generated by continuing a short target audio segment—also known as a prompt. This technique is analogous to outpainting in image generation. For signal representation, we use DAC-VAE, introduced in [17], a variational autoencoder version of the Descript Audio Codec [20], which encodes and decodes raw waveforms at 48 kHz.

Training such a model requires effectively disentangling content from speaker characteristics and prosody. While traditional representations like wav2vec 2.0 [21] and HuBERT [22] are well-suited for speech recognition due to their rich contextual encoding, they often retain speaker-dependent features, which is suboptimal for voice conversion. Therefore, we adopt ContentVec [23], a representation explicitly designed to minimize speaker-specific attributes. In our experiments, ContentVec effectively encodes phonetic content and relative pitch while suppressing speaker identity, making it well-suited for high-quality, expressive synthesis.

To transfer the target speaker's rhythm and prosody, we introduce a conditioning strategy that randomly alters source content length during training and uses cross-attention for dynamic alignment. This enables flexible rhythm control while delivering content accurately.

We summarize our contributions as follows:

- We propose a general diffusion transformer framework, **DiTVC**, for voice conversion that improves speaker similarity and environmental acoustic similarity while maintain high naturalness.
- We introduce targeted training augmentations to develop a variant, **DiTVC-Speed**, which enables flexible control over speaking rate.
- Through both objective metrics and subjective evaluations, we demonstrate that our model delivers high-quality voice conversion with strong speaker similarity and faithful reproduction of environmental acoustic properties.

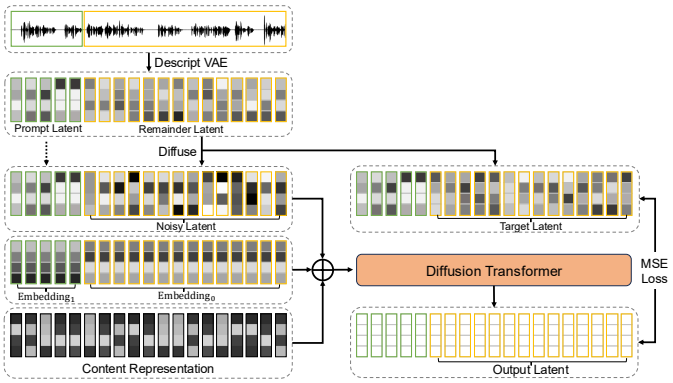

Fig. 1: Framework of **DiTVC**: the input utterance is divided into a prompt and a prediction segment. The diffusion transformer takes in a noisy latent, a mask embedding, a content representation from the prediction part, and generates an output latent using the speaker and acoustic information from the prompt. An MSE loss between the output latent and the target latent of the reparameterized velocity is used for optimization. Lastly, we simulate the diffusion process and use the DAC-VAE model to reconstruct the audio waveform from the latent.

2. METHOD

We propose a general voice conversion framework using a Diffusion Transformer. For general one-shot any-to-any voice conversion, a target utterance is required to capture the target speaker's characteristics, often summarized in a speaking embedding, while a source utterance provides the spoken content. Our model uses the raw target utterance as a prompt to clone all aspects of its speaker characteristics, including timbre, rhythm and prosody, as well as environmental acoustics, maximizing the use of available information.

2.1. Diffusion Transformer (DiT) for Voice Conversion

We illustrate the framework of the synthesis model in Figure 1. We adopt Diffusion Transformer (DiT) [10] as the backbone network for its proven generation quality in the image domain [24] and the audio domain [17]. Similar to the use of Diffusion Transformers in the image domain, where VAE models convert image patches into latent representations, we employ DAC-VAE [20], an audio VAE model, to encode audio waveform into latent representations, enabling effective speech synthesis and transformation. The diffusion process generates the latent for the target audio in a prompt continuation setup while conditioning on the content representation extracted from the source speech audio. The generated latent is later decoded back to an audio waveform using the VAE decoder.

We formalize the diffusion process as follows. Given a raw waveform $\mathbf{y}$, we use DAC-VAE[1] to encode y into its latent representation $\mathbf{x}_0 = \mathcal{E}(\mathbf{y})$. We adopt the shifted cosine noise schedule from SimpleTTS [12] to sample $t \in [0, 1]$ as the timestep in the diffusion process, with α_t, σ_t controlling the variance of signal and noise at time t under shift scale factor $s = 0.5$. The forward diffusion process for noisy latent $\mathbf{x}_t$ is formulated as belows:

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$$

Given $\mathbf{x}_0$ and the sampled t, the model predicts the re-parameterized velocity [25] $\mathbf{v} = \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}_0$, which is optimized as a reconstruction objective during training using an MSE loss:

$$\mathcal{L}\left(f_\theta; p_{\text{data}}\right) = \mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}, t \sim \mathcal{U}(0,1), \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})}[\|\mathbf{v} - f_\theta(\mathbf{x}_t; \mathbf{c}, \mathbf{m}, t)\|_2]$$

where f_θ is the Diffusion Transformer model parameterized by θ, $\mathbf{c}$ represents the content representation, and $\mathbf{m}$ denotes a prompt mask.

[1]https://github.com/innnky/descript-audio-vae.

Unlike many voice conversion methods that rely on pretrained or summarized speaker embeddings, we provide speaker and environment characteristics in the form of prompt directly through the target reference audio. We train the model by replacing a random portion of the diffusion latent with the ground-truth latent for prompt. The corresponding prompt mask in the form of $[1, ..., 1, 0, ..., 0]$ is constructed using a random split ratio $r \sim \mathcal{U}(0.1, 0.25)$. The model learns to predict the remainder using the content representation while ensuring consistency with the prompt.

The Diffusion Transformer model takes in $\mathbf{x}_t, \mathbf{c}, \mathbf{m}, t$ as conditioning inputs. We adopt ContentVec [23] as $\mathbf{c}$ which is a content representation excluding speaker information. By default, the input conditions $\mathbf{m}$ and t, along with $\mathbf{c}$, are embedded, temporally interpolated, and concatenated with $\mathbf{x}_t$. During synthesis, we sample a random Gaussian noise for the diffusion latent, run the inverse diffusion steps, and obtain the final output latent from the predicted re-parameterized velocity. Then, we use the DAC-VAE decoder to decode the output latent back into audio waveform. Finally, as suggested in [26], we apply the original DAC [20] as a post-processing step. Its quantization module helps to refine the audio and reduce artifacts introduced by the VAE and by generation in continuous latent space.

2.2. Speaking Rate Controls

Previous voice conversion models typically inherit the time alignment of the source spoken content, but the mismatch between speaking rate and the target speaker often leaves a perceptual discrepancy in the converted speaker identity, making the results less useful for real-world scenarios. As such, we propose two possible modes of speaking rate controls for voice conversion, one preserving the exact content from the source, while the other enables configurable speaking rate. The former mode is achieved in the default model design discussed above where the content representation is added to the latent to ensure precise time alignment. The latter is achieved by conditioning mechanism using cross-attention as well as training-time augmentation as discussed below.

2.2.1. Speaking rate definition: To quantify how fast or slow a speaker speaks, we define the speaking rate of an utterance as the number of phonemes per second in the non-silent portions of the utterance. In practice, we first detect silences using pydub[2] and remove those longer than 500ms, then apply automatic speech recognition with Whisper [27] and phonemizer[3] to obtain the phoneme sequence. The speaking rate of a specific utterance is calculated by dividing the number of phonemes by its duration. We observe that the typical speaking rate ranges between 8 and 20 phonemes per second.

2.2.2. Speaking rate cloning from the target speech: The model aims to generate outputs with varying durations. A simple modification is to use cross-attention to condition on the content representation so that the model generates outputs of varying lengths based on the given duration budget. We first experimented with a naive augmentation by compressing or stretching the source audio using a random ratio sampled from $\mathcal{U}(0.5, 2.0)$. However, this results in a uniform stretch in the generated output, similar to a simple speed change, which undermines the goal of achieving more natural speech.

To improve non-linear time alignment, we address the linear mapping issues caused by direct stretching. We perform enhanced augmentation by dividing the source prediction segment into $k \in [1, 10]$ random segments. Each segment is then stretched using a random ratio sampled from $\mathcal{U}(0.5, 2.0)$. This approach disrupts the linear

[2]https://pypi.org/project/pydub/
[3]https://pypi.org/project/phonemizer/

alignment, forcing the model to rely on the rhythm in the prompt audio for alignment. It enables precise duration control by allowing us to set the length of the latent output. For generation with speaking rate cloning, we compute the source utterance's speaking rate and linearly scale the latent length according to the target speaker's speaking rate to control the output duration.

3. EXPERIMENTS

3.1. Datasets

3.1.1. Training datasets: **DiTVC** is trained on a combination of datasets, including Librivox [28], VoxPopuli [29], and Common-Voice [30], covering the languages *en, es, de, fr, it, pt, sv, da, nl, sv-SE*. These datasets provide a large amount of speech data with a variety of speaker characteristics and environmental acoustics including noise and reverb. This enables the model to effectively replicate these attributes from an unseen prompt audio during synthesis. The utterances are segmented into 10-35 seconds long for training.

3.1.2. Evaluation datasets: We evaluate DiTVC on the VCTK dataset [31] and the DAPS dataset [32]. The VCTK dataset is commonly used for speech synthesis; however, two baseline models—FreeVC and YourTTS—are partially trained on it. Additionally, VCTK is known to contain mild background noise and recording artifacts, thus examining the robustness of voice conversion models to acoustic perturbations. The DAPS dataset provides a clean version of speech recordings (DAPS Clean) and twelve acoustically degraded versions (DAPS Noisy) capturing real-world noise and reverb. For the three evaluation setups (DAPS Clean, DAPS Noisy, VCTK), we select 5 slow and 5 fast speakers with gender ratio $1:1$ from each dataset based on their average speaking rates. Utterances are prepared as 5–10 second clips. For each utterance, a different speaker's utterance from the same dataset is randomly chosen as the target reference speech. In the DAPS Noisy setting, the source is always clean and we randomly select one of the twelve acoustic environments for the target utterance. This yields $1,409$ conversion pairs for VCTK and 884 pairs each for DAPS Clean and DAPS Noisy.

3.2. Training Settings

We train the model with a learning rate of 0.0001, a batch size of 128 on 16 NVIDIA A100 GPUs (80 GB) for up to 600k iterations. We apply gradient accumulation of 4 steps to enable more stable training without exceeding memory constraints. The DAC-VAE model employs a latent size of 64, which is expanded to the model size through a linear layer. The mask embedding is similarly expanded using an embedding layer. The diffusion transformer consists of 16 transformer layers, 8 attention heads, a model size of 768, and a dropout rate of 0.1.

3.3. Baselines

We evaluate two versions of our method: DiTVC and its speed cloning variant, DiTVC-Speed, with the following specifications:

- **DiTVC**: We use ContentVec [23] as the content representation condition in the diffusion transformer.
- **DiTVC-Speed**: We set the length of the diffusion latent to control the output duration and thus the speaking rate of the generated speech. For each conversion pair, we compute the source and target speaking rates and scale the source duration proportionally.

We compare DiTVC and DiTVC-Speed against several state-of-the-art voice conversion baselines to evaluate its performance in terms of audio quality, speaker similarity, and robustness to noise:

- **YourTTS** [2] enables zero-shot voice conversion by leveraging speaker embeddings and adversarial training.
- **FreeVC** [6] utilizes WavLM features and a speaker encoder combined with data augmentation to achieve voice conversion through self-supervised learning.
- **GR0** [4] learns a speaker embedding disentangled from wav2vec features through a reconstruction-based approach while enabling voice conversion.
- **DiffVC** [13] is a diffusion-based voice conversion model that enhances quality by modeling the distribution of natural speech.

	DAPS Clean		VCTK	
Exp.	**WER ↓**	**SSIM ↑**	**WER ↓**	**SSIM ↑**
Source	0.00	0.078	0.00	0.083
Target	120.01	1.000	113.45	1.000
YourTTS	10.11	0.366	10.56	0.373
FreeVC	**3.74**	0.323	**4.13**	**0.505**
GR0	5.56	0.449	6.67	0.292
DiffVC	17.44	0.358	33.78	0.307
DiTVC	**4.15**	**0.564**	**5.44**	**0.426**
DiTVC-Speed	7.06	**0.464**	9.52	0.325

Table 1: The objective scores—Word Error Rate (WER) and speaker similarity (SSIM) — for all baselines. The top 2 scores are highlighted. Note that YourTTS and FreeVC were trained on data that includes the VCTK dataset, while for all other methods, both VCTK and DAPS are unseen during training.

3.4. Objective Evaluation

We evaluate the quality of the speech generation using two objective metrics: word error rate (WER) and speaker embedding similarity (SSIM). For WER, we use Whisper [27] to conduct automatic speech recognition (ASR) on the source speech and the generated speech, and compute the word error rate to evaluate the content accuracy. For SSIM, we use the finetuned WavLM-Large model from UniSpeech [33] as the speaker embedding model and compute cosine distance between the embedding extracted from the generated speech and that from the target speech to evaluate speaker identity preservation.

Table 1 presents objective scores for all models. We evaluated only on DAPS Clean and VCTK, as the ASR and speaker embedding models may lack robustness to heavy acoustic degradations as in DAPS Noisy. **DiTVC** and **FreeVC** achieve the best performance. However, **FreeVC** is partially trained on VCTK, leading to its higher performance, but its speaker similarity score drops significantly on DAPS Clean, which demonstrates its overfitting on VCTK dataset as well as its low generalization ability to other datasets. Our proposed **DiTVC** maintains consistently low WER and high speaker similarity. The speed control variant **DiTVC-Speed** shows a higher WER, likely due to duration changes that compress or stretch the content, leading to recognition errors.

3.5. Subjective Evaluation

In addition to objective evaluation, we conduct a subjective listening test on the Prolific platform[4] to collect Mean Opinion Scores (MOS). Each rater listens to recordings with the same linguistic content, generated by all methods in the study, and rates naturalness, speaker similarity, and environmental acoustic similarity on a 5-point Likert scale. The target voice is provided as a reference for speaker and acoustic comparison. To ensure quality responses, participants are pre-screened for English fluency and the absence of hearing impairments, and are compensated at an average rate of $15/hour.

[4]https://www.prolific.co/

	DAPS Clean		DAPS Noisy			VCTK	
Exp.	**Naturalness**	**Speaker Similarity**	**Naturalness**	**Speaker Similarity**	**Acoustic Similarity**	**Naturalness**	**Speaker Similarity**
Source	4.01 ± 0.09	1.94 ± 0.11	3.85 ± 0.10	2.06 ± 0.11	2.67 ± 0.12	3.91 ± 0.09	2.18 ± 0.11
Target	4.54 ± 0.07	4.67 ± 0.07	4.48 ± 0.08	4.63 ± 0.07	4.51 ± 0.07	4.55 ± 0.07	4.74 ± 0.06
YourTTS	2.90 ± 0.12	2.37 ± 0.11	3.21 ± 0.11	2.42 ± 0.10	2.57 ± 0.10	3.09 ± 0.11	2.48 ± 0.10
FreeVC	$\mathbf{3.88} \pm 0.09$	3.12 ± 0.11	$\mathbf{3.81} \pm 0.10$	2.72 ± 0.12	2.90 ± 0.11	$\mathbf{3.82} \pm 0.09$	2.93 ± 0.11
GR0	3.76 ± 0.11	$\mathbf{3.52} \pm 0.11$	3.50 ± 0.11	2.46 ± 0.11	2.71 ± 0.11	3.77 ± 0.10	$\mathbf{3.15} \pm 0.11$
DiffVC	3.28 ± 0.12	2.82 ± 0.12	3.55 ± 0.10	2.71 ± 0.11	2.87 ± 0.11	3.35 ± 0.10	2.75 ± 0.11
DiTVC	$\mathbf{3.85} \pm 0.10$	$\mathbf{3.55} \pm 0.11$	$\mathbf{3.74} \pm 0.09$	$\mathbf{3.35} \pm 0.10$	$\mathbf{3.51} \pm 0.09$	$\mathbf{3.64} \pm 0.09$	$\mathbf{3.03} \pm 0.11$
DiTVC-Speed	3.44 ± 0.11	3.22 ± 0.11	3.39 ± 0.10	$\mathbf{2.85} \pm 0.11$	$\mathbf{3.16} \pm 0.10$	3.31 ± 0.11	2.77 ± 0.11

Table 2: The subjective scores for all baselines with 95% confidence intervals. The top 2 scores are highlighted. The listening tests are conducted in comparison to the **Target** reference including the naturalness test. **DiTVC** performs consistently across all baselines.

All baseline methods and our proposed models are evaluated across three datasets: DAPS Clean, DAPS Noisy (sampled from all non-clean subsets), and VCTK. For the listening test, we randomly sample 100 voice conversion pairs from the full evaluation set used in the objective analysis. Each dataset is evaluated by a pool of 240 participants.

Table 2 presents the subjective scores from the listening test. **YourTTS** shows lower performance in both naturalness and speaker similarity. **FreeVC** and **GR0** perform relatively well among the baselines. **FreeVC** consistently achieves high naturalness, but its speaker similarity remains subpar. We also observe that, regardless of the target reference's noise level, **FreeVC** consistently outputs clean speech, resulting in low environmental acoustic similarity on the DAPS noisy set. **GR0** achieves high speaker similarity on the clean sets but performs poorly on the noisy set, where the generated samples contain many artifacts. **DiffVC** produces samples with unnatural prosody, as noted in [14], leading to a lower naturalness score. All these baselines share a common limitation: they struggle with noisy targets, either producing outputs with artifacts or failing to reproduce the target environmental acoustics altogether.

Our method **DiTVC** delivers stable results across all evaluation sets. The score on VCTK is slightly lower than on DAPS, likely due to the specific noise artifacts present in the VCTK dataset. Unlike other methods, our methods successfully convert to a noisy target while maintaining high speaker similarity scores. Our methods also show a clear advantage in the environmental acoustic similarity scores. The speaking rate cloning variant, **DiTVC-Speed**, shows reduced quality in the evaluation. We hypothesize that: 1) training with augmentation and cross-attention is more challenging than time-aligned addition and requires more iterations to reduce artifacts and match the quality; 2) generating speech with a different, especially faster, speaking rate can lead to loss of content details, negatively impacting the scores. We believe the framework is well-suited for more controllable voice conversion, as well as higher generation quality.

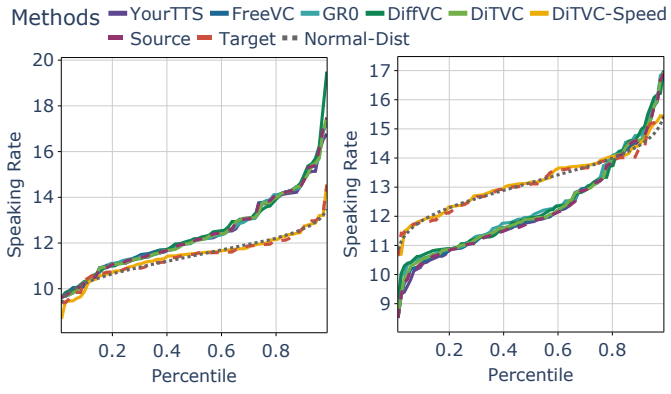

Fig. 2: This figure shows the speaking rate distribution for conversions to speakers m5 (left) and f1 (right) in DAPS. The normal distribution is derived from the mean and variance of the target speaker's speaking rate. Most methods align with **Source**, while our method, **DiTVC-Speed**, aligns with **Target**.

3.6. Speaking Rate Evaluation

Evaluating the matching of speaking rates through subjective scores is challenging due to the subtlety of perceived speaking rate differences. We conducted a listening test to assess speaking rate similarity, but listeners struggled to reliably distinguish the differences. Therefore, we choose to directly compare the speaking rate distributions. Figure 2 visualizes the speaking rate distribution by percentile for two target speakers from DAPS. We also include a curve representing the normal distribution computed from the target speaker's speaking rate. The target speaking rate follows the normal distribution well. As shown in the figure, only **DiTVC-Speed** aligns well with the **Target** distribution, while all other methods tend to follow the **Source** distribution. This confirms that the speaking rate control in **DiTVC-Speed** is effective.

Note that for speaking rate cloning, the total duration is scaled based on the ratio between the source and target speaking rates. **DiTVC-Speed** can naturally adjust the speed of each sub-segment of the speech within a ratio range of $[0.5, 2.0]$ to fulfill the total duration with a plausible rhythm, rather than applying a uniform scaling. This effect is particularly noticeable at higher speeds (e.g., near 2×), where some phonemes naturally get omitted – mimicking human fast speech. In contrast, uniform speed-up tends to articulate all phonemes clearly at an unnaturally fast pace, resulting in less realistic output.

4. CONCLUSIONS

In this work, we introduce a one-shot voice conversion approach based on a diffusion transformer that relieves the need of speaker embeddings, The model clones speaker characteristics and environmental acoustics directly from the target reference utterance that is provided as a prompt audio for generation, while conditioning on the source speech content. We show that our diffusion framework for voice conversion works effectively with the off-the-shelf content representation ContentVec. We further incorporate augmentation during training to enable speaking rate control during synthesis. Both objective and subjective evaluation results demonstrate that our approach achieves high-quality voice conversion with strong speaker and environmental acoustic similarity. The synthesized audio with speaking rate cloning closely matches the target speaker's speaking rate distribution.

REFERENCES

[1] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *Proc. ICASSP*. IEEE, 2020, pp. 6284–6288.

[2] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proc. ICML*. PMLR, 2022, pp. 2709–2720.

[3] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[4] Y. Wang, J. Su, A. Finkelstein, and Z. Jin, "Gr0: Self-supervised global representation learning for zero-shot voice conversion," in *Proc. ICASSP*, 2024.

[5] Y. Guo, Z. Li, J. Li, C. Du, H. Wang, S. Wang, X. Chen, and K. Yu, "vec2wav 2.0: Advancing voice conversion via discrete token vocoders," *arXiv preprint arXiv:2409.01995*, 2024.

[6] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[7] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. NeurIPS*, vol. 33, pp. 17 022–17 033, 2020.

[8] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[10] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. ICCV*, 2023, pp. 4195–4205.

[11] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[12] J. Lovelace, S. Ray, K. Kim, K. Q. Weinberger, and F. Wu, "Simple-tts: End-to-end text-to-speech synthesis with latent diffusion," 2023.

[13] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," *arXiv preprint arXiv:2109.13821*, 2021.

[14] Y. Li, X. Zhu, H. Li, J. Yao, W. Tian, Y. Chen, Z. Li, and L. Xie, "Codiff-vc: A codec-assisted diffusion model for zero-shot voice conversion," *arXiv preprint arXiv:2411.18918*, 2024.

[15] R. Huang, Z. Zhao, H. Liu, J. Liu, C. Cui, and Y. Ren, "Prodiff: Progressive fast diffusion model for high-quality text-to-speech," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2595–2605.

[16] K. Lee, D. W. Kim, J. Kim, and J. Cho, "Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer," *arXiv preprint arXiv:2406.11427*, 2024.

[17] Y. A. Li, R. Kumar, and Z. Jin, "Dmdspeech: Distilled diffusion model surpassing the teacher in zero-shot speech synthesis via direct metric optimization," *arXiv preprint arXiv:2410.11097*, 2024.

[18] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *Advances in neural information processing systems*, vol. 36, pp. 14 005–14 034, 2023.

[19] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," *arXiv preprint arXiv:2304.09116*, 2023.

[20] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Proc. NeurIPS*, vol. 36, 2024.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[22] 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 Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[23] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *Proc. ICML*. PMLR, 2022, pp. 18 003–18 017.

[24] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Proc. ICML*, 2024.

[25] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[26] H. R. Guimarães, J. Su, R. Kumar, T. H. Falk, and Z. Jin, "Ditse: High-fidelity generative speech enhancement via latent diffusion transformers," *arXiv preprint arXiv:2504.09381*, 2025.

[27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICLR*, 2023.

[28] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Reviews*, 2014.

[29] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[30] 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.

[31] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," https://datashare.ed.ac.uk/handle/10283/3443, 2019.

[32] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Process. Lett.*, 2014.

[33] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *Proc. ICML*. PMLR, 2021, pp. 10 937–10 947.