



REWIND: Speech Time Reversal for Enhancing Speaker Representations in Diffusion-based Voice Conversion

Ishan D. Biyani¹, Nirmesh J. Shah¹, Ashishkumar P. Gudmalwar¹, Pankaj Wasnik¹, Rajiv R. Shah²

¹Media Analysis Group, Sony Research India, Bangalore

²Indraprastha Institute of Information Technology (IIIT), Delhi, India

{ishan.biyani@sony.com, nirmesh.shah, ashish.gudmalwar1, pankaj.wasnik}@sony.com;
rajivrtn@iiitd.ac.in

Abstract

Speech time reversal refers to the process of reversing the entire speech signal in time, causing it to play backward. Such signals are completely unintelligible since the fundamental structures of phonemes and syllables are destroyed. However, they still retain tonal patterns that enable perceptual speaker identification despite losing linguistic content. In this paper, we propose leveraging speaker representations learned from time reversed speech as an augmentation strategy to enhance speaker representation. Notably, speaker and language disentanglement in voice conversion (VC) is essential to accurately preserve a speaker's unique vocal traits while minimizing interference from linguistic content. The effectiveness of the proposed approach is evaluated in the context of state-of-the-art diffusion-based VC models. Experimental results indicate that the proposed approach significantly improves speaker similarity-related scores while maintaining high speech quality.

Index Terms: Speech Time Reversal, Voice Conversion, Diffusion-based Voice Conversion, Speaker Representation

1. Introduction

Voice Conversion (VC) is the process of altering the perceived speaker identity in a speech signal from a source speaker to that of a target speaker while retaining the original linguistic content [1]. This technology has significant applications in the entertainment industry, where it is used for tasks such as dubbing, character voice transformation during post production in films, video games, advertisements, and animations [2–6]. In many real-world scenarios, voice conversion must be achieved with minimal available data, often relying on just one or even no prior recordings of the target voice, making one-shot and zero-shot VC approaches critical [7–11]. These approaches enable models to adapt to new, unseen voices providing the necessary flexibility. Their importance lies in the ability to rapidly generate high-quality, natural-sounding speech for any target voice, thereby enhancing creative possibilities and streamlining production workflows in the entertainment industry.

In the context of VC, capturing speaker identity is crucial and is typically achieved through speaker encoders. These encoders can be either pre-trained for speaker classification tasks utilizing large datasets to learn robust, speaker specific embeddings [12] or they can be trained jointly with the VC model directly [13, 14]. The resulting speaker embeddings should capture key vocal attributes such as timbre, pitch, and prosody, which are essential for characterizing a speaker's unique identity [15]. However, these embeddings often suffer from linguistic interference, where language related features entangled with the speaker specific characteristics. To address this, language disentanglement strategies are often employed during training

[16–18], ensuring that the learned embeddings remain as language independent as possible. Finally, these embeddings are used to condition the decoder within the VC framework to generate speech that perceptually reflects the target speaker's identity while preserving the original linguistic content.

In this paper, we propose the concept of speech time reversal (STR) as a novel data augmentation strategy to enhance speaker representations. Speech time reversal involves inverting the entire speech signal along the time axis so that it plays backward. Previous studies have utilized short-time speech reversal for noise robust automatic speech recognition (ASR) task [19] because it alters the signal in a controlled manner while still preserving most of the linguistic content. This controlled modification helps models become robust to acoustic variations and distortions without losing the key phonetic cues needed for these tasks. In contrast, complete speech time reversal has been largely overlooked in the literature. Although speech time reversal process makes the signal completely unintelligible by destroying the core phonemic and syllabic structures, it preserves rhythmic and tonal patterns that enable perceptual speaker identification [20]. We initially conducted perceptual studies to determine whether speaker identity is retained in time-reversed speech signals. Based on these findings, we extended our hypothesis to voice conversion by learning speaker embeddings from time reversed speech and fusing them with conventional speaker embeddings. The combined speaker embeddings were subsequently used to condition the diffusion-based decoder in the VC framework. We conducted several subjective and objective evaluations on the LibriTTS [21] and VCTK [22] databases to demonstrate the effectiveness of the proposed strategy against state-of-the-arts (SOTA) VC methods.

2. Related work

Early voice conversion (VC) systems relied on traditional statistical models [1], such as Gaussian Mixture Models (GMMs), to establish a mapping between the acoustic features of source and target speakers [23]. These early approaches [1], while pioneering, struggled with the high-dimensional and nonlinear aspects of speech. With the advent of deep learning, research in VC evolved rapidly, transitioning through deep neural networks (DNNs) [24], sequence-to-sequence [25], variational autoencoders (VAEs) [26], and generative adversarial networks (GANs) [27] to achieve more natural and expressive speech conversion. Recently, diffusion based models have gained traction due to their robust training dynamics and iterative denoising process [28–31], which allows for the generation of high quality speech even in challenging low-data scenarios.

Controlling speaker identity in VC has remained a central challenge, particularly in zero-shot and one-shot settings

[7–10]. To address this, researchers have integrated speaker embedding networks into the conversion framework. These networks, often pretrained on large-scale speaker recognition tasks to produce robust embeddings such as d-vectors or x-vectors, capture essential vocal attributes like timbre, pitch, and prosody [12]. Some approaches use a two-stage training process where the speaker encoder is pretrained independently and later fused with the VC model as a conditioning mechanism. Other methods opt for joint training [32], ensuring that the speaker embeddings are directly optimized for the conversion task. Additionally, techniques such as adaptive normalization and attention-based fusion have been employed to further refine the control over the target speaker’s identity, significantly enhancing speaker similarity and overall conversion quality.

3. Proposed Approach

3.1. Proposed Speech Time Reversal (STR) Strategy

Let $x(t)$ denotes the original signal, then the time reversed speech signal is obtained by flipping $x(t)$ along time axis, i.e., $x(l - t)$, where $t \in [0, l]$. To assess the preservation of speaker identity in the time reversed speech, we designed a controlled perceptual study. In particular, we asked subjects to listen to an original reference sample and then to identify the correct speaker from a set of time-reversed speech signals. We utilized English speech data from total six speakers consisting of three males and three females. Total 25 participants participated in the perceptual study. The goal of this study was to determine whether the unique vocal attributes of a speaker, such as timbre and intonation, remain perceptually recognizable after the complete speech time reversal process. By comparing the reference recording with the altered samples, we aimed to evaluate the effectiveness of time reversal in retaining speaker identity, despite the loss of intelligible linguistic content. In particular, subjects could identify the correct speaker with 80.3% accuracy from the time reversed speech signals. Corresponding confusion matrices for the perceptual study is shown in Figure 1.

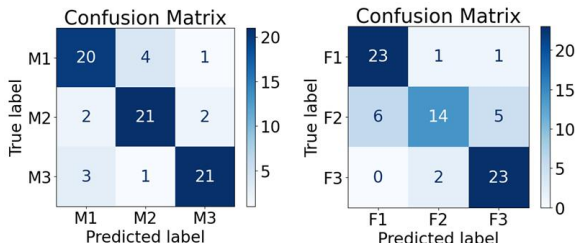


Figure 1: Confusion matrices for the perceptual study of speaker identification from the time reversed speech. Here, $M1$, $M2$, $M3$ and $F1$, $F2$, $F3$ represents three different Male and Female speakers, respectively.

Table 1: WER and speaker similarity (SS) score for different speech time reversal strategies.

	Short Time						Complete Time
	10ms	20ms	50ms	100ms	200ms	500ms	Full length
SS Score	0.89	0.90	0.86	0.91	0.94	0.94	0.96
WER	0%	0%	28%	100%	100%	100%	100%

Additionally, to compare the proposed speech time reversal strategy with the short-time speech reversal approach, we analyzed the spectrographic outputs from both methods. Figure 2 presents the spectrographic visualization of (a) original speech, (b) 20 ms short-time speech reversal and (c) 100 ms short-time speech reversal strategy and (d) complete speech time reversal.

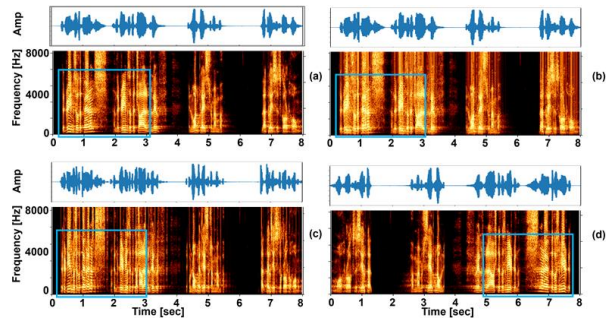


Figure 2: Spectrographic visualization of (a) original speech (b) 20 ms short-time, (c) 100 ms short-time speech reversal, and (d) complete speech time reversal.

Our analysis revealed that, in the case of complete speech reversal, the harmonic structures are prominently visible, which strongly indicates the retention of speaker-specific information. The clear presence of these harmonic patterns suggests that even though the reversed speech is rendered unintelligible, it preserves critical acoustic cues such as timbre and pitch that are unique to the speaker. To further substantiate these findings, we calculated Word Error Rates (WER) and speaker similarity scores based on cosine similarity metric as shown in Table 1. The evaluation showed that the complete reversal method not only maintains a higher WER but also yields higher speaker similarity scores compared to the short-time speech reversal strategy. These results, aligning with [20], confirm that complete speech reversal is more effective in preserving speaker identity with no linguistic interference, making it a promising approach for VC application.

3.2. Proposed Diffusion-based Voice Conversion

Denoising diffusion models have emerged as a powerful class of generative models, significantly advancing task of voice conversion [28–30]. These models work in two stages: a forward diffusion process that incrementally adds noise to the data and a reverse denoising process that gradually recovers the original signal. Traditional diffusion models use a discrete-time Markov chain [33], while score-based generative modeling employs continuous-time stochastic differential equations [34] for greater flexibility and improved sample quality. In particular, decoupled denoising diffusion models (DDDMs) in DDDM-VC leverage multiple disentangled denoisers to control speaking style attributes effectively [30]. Our proposed STR strategy is implemented using DDDM-VC as the backbone architecture.

DDDM-VC features both a source-filter encoder and decoder and utilizes self-supervised speech representations to disentangle various aspects of the speech signal. It employs a prior mixup technique to enhance robustness across different conversion scenarios. The model uses three types of representations for speech disentanglement: (1) Content Representation, which extracts phonetic content using models like XLS-R [35] from Wav2Vec 2.0; (2) Pitch Representation, where the YAPPT algorithm extracts the fundamental frequency (F_0)—normalized per utterance and processed using VQ-VAE—to capture intonation; and (3) Speaker Representation, where a speaker embedding network extracts speaker representation from the target Mel-spectrogram. These features are averaged at the utterance level and integrated across the network to enable robust zero-shot voice style transfer. To further enhance the speaker representation, we propose combining the conventional speaker representation with one derived from the speech time reversal strat-

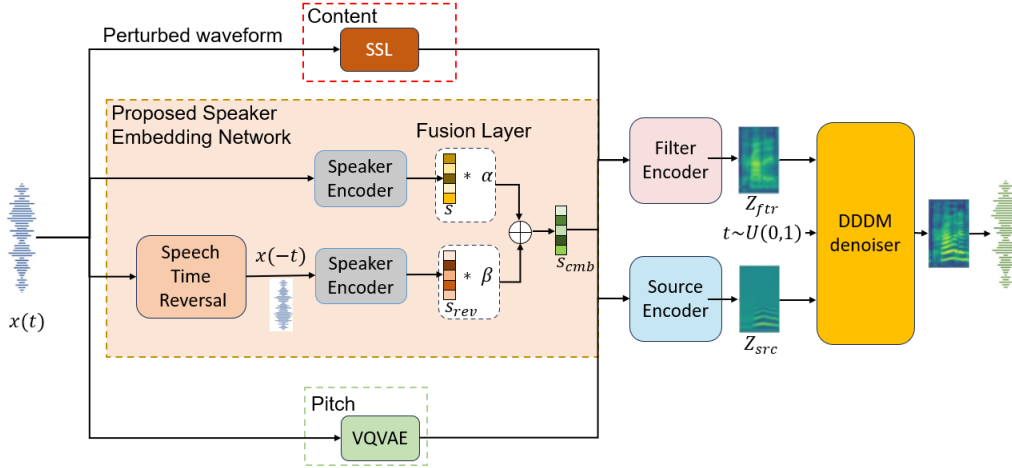


Figure 3: Blockdiagram of the propose approach in Diffusion-based Voice Conversion

egy. The two embeddings are then integrated using a weighted fusion layer (as shown in Figure 3). The combined speaker representation (S_{cmb}) is defined as follows:

$$S_{cmb} = \alpha * S + \beta * S_{rev} \quad (1)$$

where S and S_{rev} denotes the speaker embeddings learned from the original and time-reversed speech, respectively. The weighting coefficients α and β are selected from the range $[0, 1]$.

Finally model uses source-filter theory [36] for speech resynthesis. While the source encoder processes pitch and speaker information, the filter encoder deals with content and speaker information. A data-driven prior in the diffusion process can successfully direct the beginning point of the reverse process. This is improved by using fully reconstructed source and filter Mel-spectrograms, Z_{src} and Z_{ftr} , respectively, to provide a more thorough prior. These are regularized against the target Mel-spectrogram X_{mel} with the following loss function:

$$\mathcal{L}_{rec} = ||X_{mel} - (Z_{src} + Z_{ftr})||_1, \quad (2)$$

where $Z_{src} = E_{src}(pitch, s)$, $Z_{ftr} = E_{ftr}(content, S_{cmb})$.

It is also important to note that the disentangled source and filter Mel-spectrograms are converted using different speaker representations. The entire model is jointly trained end-to-end, including the style encoder, source-filter encoder, and decoder, without pre-trained XLS-R and F_0 VQ-VAE.

4. Experimental Analysis

4.1. Experimental Setup

We trained our model on LibriTTS (1,100 speakers, 10 held out) and evaluated it on the held-out speakers as well as the VCTK dataset. For self-supervised speech representations, audio was downsampled from 24 kHz to 16 kHz and fed into XLS-R (0.3B). We generated 80-D log-scale Mel spectrograms for the target voice and speaker encoder using a hop size of 320 a window size of 1280, and a 1280 point Fourier transform in order to match the time frames between the self-supervised representation and the Mel spectrogram without interpolation. We trained DDDM-VC using the AdamW optimizer (Loshchilov & Hutter, 2019) with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and a weight decay of $\lambda = 0.01$. The learning rate started at 5×10^{-5} and decayed according to a schedule of $0.999^{\frac{t}{3}}$. Training was conducted on two NVIDIA L40 GPUs (32GB each) for 1000 epochs with a

batch size of 64, taking approximately 2 days. The proposed Speaking Style Encoder uses two multi-head attention layers and a 0.1 dropout rate to process log-scale Mel-spectrograms from the input waveform with an 80 dimensional input, extracting speaker embeddings of size 256-D. Similarly, reverse speaker embeddings are extracted from the time-reversed waveform. Both embeddings are fused through a weighted fusion layer to produce the final speaker representation. We employed HiFi-GAN V1 [37] for waveform synthesis, substituting multi-scale STFT-based discriminators (MS-STFTD) from EnCodec [38] to improve vocoder quality. The Encoder and Decoder have 6.36M and 60M parameters, respectively.

4.2. SOTA Baseline VC Methods

The following is a list of the SOTA VC baseline methods used to validate the effectiveness of our proposed strategy.

Diff-VC [28]: A diffusion-based VC baseline that iteratively refines noisy speech signals through reverse diffusion, producing high-fidelity conversions with strong speaker adaptation.

Diff-HierVC [29]: A hierarchical VC system that uses dual diffusion models DiffVoice for speech quality and DiffPitch for target F_0 generation along with a source-filter encoder and masked precursor to enhance speaker adaptation and pitch accuracy.

LM-VC [7]: A two-stage language modeling approach for zero-shot VC that employs a masked prefix language model and window-attentive external LM to disentangle speaker information and recover masked content.

SEF-VC [8]: A speaker embedding-free VC model that leverages position-agnostic cross-attention to directly model and integrate speaker timbre from reference speech, enhancing adaptability and stability.

StyleVC [10]: A VC model that mitigates training-inference mismatches using adversarial training, pitch prediction, and style generalization with unpaired samples to improve acoustic quality and speaker adaptation.

StableVC [9]: A style controllable zero-shot VC system that disentangles linguistic content, timbre, and style using a conditional flow matching module with dual attention, ensuring fast and high-quality synthesis.

VALLE VC [39]: A model that utilizes a large scale generative speech framework and neural audio codecs to achieve expressive, high-fidelity voice conversion while preserving target speaker identity.

4.3. Experimental Results

The voice conversion performance is evaluated in the following section by comparing the proposed approach with baseline and state-of-the-art approaches. Demo samples are available at URL ¹. Since improving speaker representation is the main goal, the speaker similarity score serves as the primary objective metric for evaluating effectiveness. This metric helps in determining how effectively the proposed method improve speaker characteristics during the voice conversion task. Table 2 provides a summary of the results, encompassing both subjective and objective assessments. Speaker embeddings are extracted from both generated and reference audio samples using Resemblyzer² in order to calculate the Objective Speaker Similarity Score. The 256-dimensional summary vector that Resemblyzer produces captures the characteristics of the speaker. The similarity between the generated and reference speech is then measured by computing the cosine distance between their respective embeddings. The results presented in Table 2 indicate an improvement in speaker similarity metrics for diffusion-based voice conversion baselines, particularly in the cases of DiffHierVC and DDDM-VC. Additionally, DiffVC demonstrates comparable performance. Overall the results shows an improvement of 4.16% using proposed approach with respect to baseline highlighting its effectiveness.

Another objective evaluation metric that was used was the AutoPCP score, which measures the prosody-level similarity across speech samples using a model-based estimator of Prosody Conversion Precision (PCP) [40]. Better prosodic alignment is reflected in a higher AutoPCP score. The findings show that, in comparison to the baseline techniques, the suggested strategy often produces either improvements or comparable performance in AutoPCP scores. We used two Mean Opinion Score (MOS) predictions methods to evaluate MOS objectively. The first, WV-MOS, is an absolute objective speech quality measure based on a fine-tuned Wave2Vec 2.0 model for direct MOS score prediction. Since winning several categories in the 2022 VoiceMOS challenge, the second, UTMOS, is an ensemble-based MOS prediction system that has been widely used to evaluate the top TTS systems. By evaluating audio quality on a scale of 1 to 5, with 5 denoting the highest quality and 1 the lowest, UTMOS determines the mean opinion score (MOS) for a given speech sample. The proposed approach shows improvement of 0.5% to 5.5% in both metrics compared to baseline methods.

For subjective evaluation, the MUSHRA-based MOS metric was used to assess speech similarity. A group of 25 participants, all between the ages of 25 and 35, having no known hearing impairments, evaluated the samples on a scale of 0 to 100, where 100 is the greatest quality, depending on how similar the speaker was to the reference. Statistical significance is computed via 95% confidence intervals in subjective evaluations, with MOE ranging from ± 1.66 to ± 2.96 . According to the evaluation’s findings, the proposed approach outperforms baseline techniques by 2.67% w.r.t. average gain in subjective speaker similarity score.

4.4. Ablation Study

Our ablation study explored the impact of combining the conventional speaker embedding with its time-reversed counter-

¹<https://research.sri-media-analysis.com/interspeech25-rewind-vc/>

²<https://github.com/resemble-ai/Resemblyzer>

Table 2: Subjective and objective evaluation analysis

	Subjective SS Score	Objective SS Score	AutoPCP	WV-MOS	UT-MOS
SOTA VC Baselines					
LMVC	51.28 \pm 2.48	0.74	2.75	3.32	2.9
SEFVC	43.17 \pm 1.84	0.71	2.7	3.3	2.47
StyleVC	51.97 \pm 2.91	0.76	2.93	3.85	3.75
StableVC	62.11 \pm 2.26	0.67	2.89	4.21	3.78
VALLE-VC	55.61 \pm 1.83	0.72	2.87	3.96	3.63
Diffusion-based VC baselines					
DiffVC	50.12 \pm 1.90	0.75	2.45	3.57	3.4
Diff VC+Ours	53.42 \pm 2.09	0.71	2.43	3.56	3.46
DiffHierVC	75.76 \pm 2.96	0.76	2.88	3.99	3.68
DiffHier-VC + Ours	75.72 \pm 1.66	0.8	2.89	3.99	3.83
DDDM-VC	77.46 \pm 2.07	0.7	2.87	3.84	3.21
DDDM-VC + Ours	78.61 \pm 1.66	0.79	2.94	3.91	3.55

part using different weighting schemes and the integration of cross-attention. Here, we again used MUSHRA-based subjective score to calculate speaker similarity from 25 subjects as mentioned in the previous subsection. We observed that an equal weighting of 0.50 for both embeddings consistently produced superior speaker similarity scores (as shown in Figure 4). This finding suggests that both the standard and time-reversed embeddings contribute equally valuable speaker-specific information, and the use of cross-attention further refines their integration, leading to enhanced overall performance in our voice conversion framework.

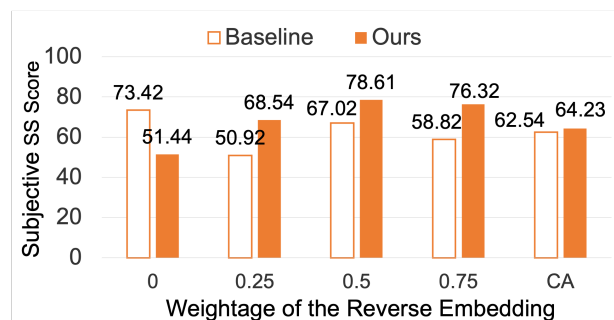


Figure 4: Ablation analysis for subjective speaker similarity scores w.r.t. Baseline(DiffVC)

5. Summary and Conclusions

We leverage full-utterance speech time reversal (STR) as a targeted data-augmentation strategy within a diffusion-based voice-conversion pipeline to strengthen speaker embeddings. Whereas earlier efforts have applied reversal only to short segments or portions of the signal, our method inverts the entire utterance. This destroys intelligible linguistic content yet preserves global rhythmic and tonal contours, which, as our perceptual studies show, are alone sufficient to maintain speaker-specific information. This finding supports the fusion of conventional speaker embeddings with those derived from time-reversed speech, providing a robust means to disentangle speaker identity from linguistic content. Our experimental evaluations on the LibriSpeech and VCTK databases, using a diffusion-based voice conversion framework, reveal that the proposed approach significantly improves speaker similarity scores while maintaining high speech quality. These results underscore the potential of STR to overcome data limitations in zero-shot VC scenarios and pave the way for future research to further refine and integrate such unconventional signal transformations into voice conversion systems.

6. References

- [1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.
- [2] W. Brannon *et al.*, "Dubbing in practice: A large scale study of human localization with insights for automatic dubbing," *Trans. of the Assoc. for Comput. Ling.*, vol. 11, pp. 419–435, 2023.
- [3] C. Hu *et al.*, "Neural dubber: Dubbing for videos according to scripts," *Advances in neural information processing systems*, vol. 34, pp. 16 582–16 595, 2021.
- [4] N. Sahipjohn *et al.*, "Dubwise: Video-guided speech duration control in multimodal llm-based text-to-speech for dubbing," in *INTERSPEECH*, Kos Island, Greece, 2024.
- [5] S. Mhaskar, N. Shah, M. Zaki, A. Gudmalwar, P. Wasnik, and R. Shah, "Isometric neural machine translation using phoneme count ratio reward-based reinforcement learning," in *NAACL Findings*, Mexico, 2024, pp. 3966–3976.
- [6] A. Gudmalwar, N. Shah, S. Akarsh, P. Wasnik, and R. R. Shah, "Vccl-tts: Voice identity and emotional style controllable cross-lingual text-to-speech," in *INTERSPEECH*, 2024.
- [7] Z. Wang, Y. Chen, L. Xie, Q. Tian, and Y. Wang, "Lm-vc: Zero-shot voice conversion via speech generation based on language models," *IEEE Signal Processing Letters*, 2023.
- [8] J. Li *et al.*, "Sef-vc: Speaker embedding free zero-shot voice conversion with cross attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 296–12 300.
- [9] J. Yao, Y. Yan, Y. Pan, Z. Ning, J. Ye, H. Zhou, and L. Xie, "Stablevc: Style controllable zero-shot voice conversion with conditional flow matching," *arXiv preprint arXiv:2412.04724*, 2024.
- [10] I.-S. Hwang, S.-H. Lee, and S.-W. Lee, "Stylevc: Non-parallel voice conversion with adversarial style generalization," in *ICPR*. IEEE, 2022, pp. 23–30.
- [11] N. Shah, M. Singh, N. Takahashi, and N. Onoe, "Nonparallel emotional voice conversion for unseen speaker-emotion pairs using dual domain adversarial network & virtual domain pairing," in *ICASSP*, 2023, pp. 1–5.
- [12] M. Chen and Z. Duan, "Controlvc: Zero-shot voice conversion with time-varying controls on pitch and speed," *arXiv preprint arXiv:2209.11866*, 2024.
- [13] S. Seki, H. Kameoka, K. Tanaka, and T. Kaneko, "Jsv-vc: Jointly trained speaker verification and voice conversion models," *ICASSP*, 2023.
- [14] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, "Using joint training speaker encoder with consistency loss to achieve cross-lingual voice conversion and expressive voice conversion," *arXiv preprint arXiv:2307.00393*, 2023.
- [15] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Interspeech*, 2017, pp. 1497–1501.
- [16] S. H. Mun, M. H. Han, M. Kim, D. Lee, and N. S. Kim, "Disentangled speaker representation learning via mutual information minimization," *arXiv preprint arXiv:2208.08012*, 2022.
- [17] K. Nam *et al.*, "Disentangled representation learning for multilingual speaker recognition," *arXiv preprint arXiv:2211.00437*, 2022.
- [18] Y. Tu, M.-W. Mak, and J.-T. Chien, "Contrastive speaker embedding with sequential disentanglement," *arXiv preprint arXiv:2309.13253*, 2023.
- [19] F.-A. Chao, S.-W. F. Jiang, B.-C. Yan, J.-w. Hung, and B. Chen, "Tenet: A time-reversal enhancement network for noise-robust asr," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 55–61.
- [20] K. Saberi and D. R. Perrott, "Cognitive restoration of reversed speech," *Nature*, vol. 398, no. 6730, pp. 760–760, 1999.
- [21] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.
- [22] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.
- [23] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.
- [24] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.
- [25] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.
- [26] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.
- [27] Z. Zhang, B. He, and Z. Zhang, "Gazev: Gan-based zero-shot voice conversion over non-parallel speech corpus," *arXiv preprint arXiv:2010.12788*, 2020.
- [28] V. Popov *et al.*, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," *arXiv preprint arXiv:2109.13821*, 2021.
- [29] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation," *arXiv preprint arXiv:2311.04693*, 2023.
- [30] H. Y. Choi, S.-H. Lee, and S.-W. Lee, "Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 862–17 870.
- [31] A. P. Gudmalwar, I. D. Biyani, N. J. Shah, P. Wasnik, and R. R. Shah, "Emoreg: Directional latent vector modeling for emotional intensity regularization in diffusion-based voice conversion," in *AAAI*, vol. 39, no. 22, 2025, pp. 23 960–23 968.
- [32] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only auto-encoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.
- [33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [34] Y. Song, J. Sohl-Dickstein, D. P. Kingma, K. Abhishek, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.
- [35] A. Babu, C. Wang, A. Tjandra, K. Lakhota, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.
- [36] G. Fant, "Acoustic theory of speech production (no. 2)," 1970.
- [37] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.
- [38] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.1343*, 2022.
- [39] C. Wang *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.
- [40] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenhaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haahheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.