# SSMamba-VC: Learning Linear-Time Sequence Modeling with Selective State Spaces for Voice Conversion

Ravindrakumar M. Purohit*, Kaustubh S. Wade*, Hemant A. Patil
*Speech Research Lab, Dhirubhai Ambani University (formerly DA-IICT),* Gandhinagar, India
{202321002, 202418024, hemant_patil}@dau.ac.in

***Abstract*—Voice Conversion (VC) modifies the speaker identity of speech while preserving its linguistic content. It has transformed the domains of voice dubbing, voice cloning, and assistive technologies by improving the naturalness of the synthesized speech. Recent advances in generative models, particularly diffusion-based frameworks, have significantly improved perceptual quality but come at a high computational cost and face challenges, including degraded prosody and reduced speech intelligibility. These primarily arise from attention-based sequence modeling, which scales quadratically. In this work, we propose SSMamba-VC, a selective state space (SS)-based hierarchical discussion framework that replaces attention mechanism with linear-time sequence modeling. Compared to the existing state-of-the-art VC systems, SSMamba-VC achieves notable improvements across *subjective*, *objective*, and *quantitative* metrics. In particular, the proposed approach (DiffHier-VC + Ours) achieves the NISQA-MOS of 3.90 (+0.02) and a UTMOS of 2.92 (+0.16), outperforming existing diffusion baselines in perceptual stability. The proposed approach reduces $F_0$ RMSE to 42.44, improving prosodic accuracy, while also improving intelligibility, with WER reduced to 18.14 and CER to 8.73 compared to DiffHier-VC. Furthermore, with 98% fewer computational resources, SSMamba-VC achieves a ∼15.6% improvement in real-time factor (RTF), demonstrating that linear-time SS modeling enables faster inference without degrading the quality of the synthesized VC. SSMamba-VC further reduces the RTF to achieve high-fidelity VC for practical deployment. The demo utterances are publicly available at supplementary website: *https://iamshreeji-copy2.github.io/SSMamba-VC/***

***Index Terms*—Voice Conversion (VC), Diffusion, Speaker Representation, Selective State-Space Models, Adversarial Speaker Regularization.**

## I. Introduction

Despite the increasing capabilities of diffusion-based state-of-the-art (SOTA) frameworks, Voice Conversion (VC) systems still face critical limitations in terms of the *stability* and *fidelity* of synthesized utterances, including speaker diversity and generalization, long-form speech conversion, and real-time applications. Existing architectures use the attention mechanism to align linguistic content (e.g., phonemes, words) between the source and target speech, which directly contributes to misalignment during inference and results in generated utterances that exhibit robotic speech, incorrect intonation, or chopped words/sounds. Furthermore, these systems fail to maintain the prosodic consistency across the long-utterance and struggle with *speaker leakage*, where the source speaker representations inadvertently bleed into the target output during few-shot or zero-shot transfer. The fundamental bottleneck of current SOTA architectures is their reliance on attention-based mechanism, which has *quadratic* complexity. Computational scaling limits the model's context windows. Therefore, models rely on chunk-based processing, which creates problems in processing the long-range prosodic dependencies and introduces audio artifacts, degrading the temporal coherence of the generated speech. In order to improve the quality of VC and trade-off between computational efficiency and perceptual quality of voice, we identified and addressed these issues. The following are main contributions of this paper:

1) We proposed a Mamba-based sequential module for the source-filter and pitch encoder to reduce the quadratic ($\mathcal{O}(n^2d)$) complexity of the attention to a lower-order ($\mathcal{O}(nd)$) using the selective scan mechanism.
2) SSMamba-VC was trained for only 0.2M steps with a batch size of 32 on an NVIDIA TITAN XP (12GB), with adversarial regularization. In the later stage, speaker identity is restricted to the residual source-speaker characteristic in order to improve VC quality of the low-resource, unseen-speaker. Earlier work depends heavily on subjective or objective measures, such as nMOS, WER, and CER. In this work, we thoroughly evaluated using more than 15 *subjective*, *objective*, *quantitative* or performance measures, such as NISQA-MOS, UT-MOS, PESQ, MCD, MSD, $F_0$-RMSE, Frechet Distance, CS, WER, CER, and others. This comprehensive evaluation provides a detailed assessment of baselines *vs.* the proposed approach. The performance evaluation metrics are available at the supplementary website.
3) SSMamba-VC reduces the computations by 98%, and improves the real-time inference by 15.6% (RTF 0.32→0.27), lowers the $F_0$-RMSE to 42.44, and improves intelligibility with 18.15 WER and 8.73 CER, by preserving the speaker similarity (0.99) in the synthesized utterance.

The remaining part of the paper is organized as follows: Section II explores the related work in VC. Section III provides

*Authors contributed equally to this work.

details on the proposed approach of SSMamba-VC, along with problem formulation. Section IV presents the experimental setup. Finally, Sections V and VI include the results, observations, and wrap up the paper with a conclusion, respectively.

## II. Related Work

### A. Voice Conversion

Earlier efforts in non-parallel VC used Generative Adversarial Networks (GANs) to mitigate the spectral oversmoothing caused by Mean Squared Error [1], [2]. However, these frameworks rely on cycle consistency to represent linguistic content without requiring parallel data. The cycle consistency loss is defined as:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{x\sim P_X}\left[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1\right], \tag{1}$$

to minimize the $L_1$ distance between the original input $x$ and its reconstruction through the target domain generator, $G_{X\to Y}$ and source recovery generator, $G_{Y\to X}$. To recover spectral details, an adversarial loss is used to convert the speech in order to match the target data distribution, $P_Y$. The objective function is formulated as:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{y\sim P_Y}\left[\log D_Y(y)\right] + \mathbb{E}_{x\sim P_X}\left[\log\left(1 - D_Y\left(G_{X\to Y}(x)\right)\right)\right], \tag{2}$$

where the discriminator $D_Y$ attempts to distinguish real target utterances $y$ from the generated outputs. However, GAN-based VC faces challenges w.r.t. training instability, and mode collapse, which leads to degraded perceptual quality or unnatural output artifacts in converted speech utterance [3], [4].

To resolve this issue related to generalization and optimization issue, the VC field transitioned towards Denoising Diffusion Probabilistic Models (DDPMs) [5], [6], which formulated the VC as adversarial deception rather than as the reversal of a stochastic process [7]. In the forward pass, it gradually converts the speech signal $\mathbf{x}$ into Gaussian noise via a Stochastic Differential Equation (SDE) of the form:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\, dt + g(t)\, d\mathbf{w}, \tag{3}$$

where $\mathbf{f}(\cdot)$ is the *drift* coefficient and $g(t)$ is the *diffusion* coefficient. The model learns the reverse generative process by training a neural network $s_\theta(\mathbf{x}, t)$ in order to estimate the score function, which pertains to the gradient of the logarithmic probability density function [8]. In particular,

$$s_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x}). \tag{4}$$

On the other hand, the likelihood-based approach guarantees dense distributional coverage, reduces mode collapse, and synthesises realistic high frequency *timbre*. Furthermore, due to the diffusion nature, the iterative process imposes significant computational bottlenecks, and solving recent SDEs requires high real-time inference, which is challenging without distillation [9]. Moreover, the source speaker's intonation (i.e., prosodic features) can leak into the generated output due to the monolithic tensor in the standard diffusion architecture [10].

### B. Sequence Modeling Beyond Attention

The transformer architecture uses the self-attention mechanism to model *global* dependencies. Attention mechanism is able to generate high-fidelity VC at the standard sample rates (e.g., 24 kHz) with quadratic complexity $\mathcal{O}(L^2)$. Furthermore, quadratic complexity affects the modeling of long range supra-segmental features, such as prosody and stress patterns, which are important for source-to-target VC. To address this, Long Sequences with Structured SS [11], Diagonal SS [12] introduced linear-time inference $\mathcal{O}(L)$, using the implicit long convolutions. Regardless, these models operate as Linear Time-Invariant (LTI) systems, with fixed dynamics that remain unchanged regardless of the input signal. Speech signals are highly non-stationary and require adaptive processing to distinguish the transient phonemes and vowels. Mamba overcomes this by adapting the selection mechanism to make the system parameters input-dependent [13]. In the context of VC, *selectivity* helps the model dynamically govern linguistic content by disregarding source-specific acoustic attributes (e.g., pitch ($F_0$) contours). However, this critical gap remains in the literature, as no previous work has successfully coupled the generative capability of hierarchical diffusion with the linear-time inference mechanisms of selective SSMs within a unified VC framework.

## III. Proposed Method

### A. Problem Formulation

Let $\mathbf{X}^s \in \mathbb{R}^{T\times F}$ denote a source utterance represented using a Mel spectrogram, where $T$ and $F$ are the time frames and frequency bins, respectively. $\mathbf{e}^t \in \mathbb{R}^D$ represents a target speaker ($t$) embedding vector of dimension $D$, derived from a reference utterance of the target speaker, $\mathbf{X}^t$. The objective of VC is to generate a converted Mel spectrogram $\mathbf{Y} \in \mathbb{R}^{T\times F}$, while preserving the linguistic essence of the source $\mathbf{X}^s$ and embracing the *timbral* identity of the intended speaker ($t$, $\mathbf{e}^t$). This can be formulated using the conditional distribution: $p(\mathbf{Y} \mid \mathbf{X}^s, \mathbf{e}^t)$. To facilitate the disentanglement of *content* and *speaker* information, we employ a speaker-invariant content encoder $E_{\text{cnt}}$ that extracts the content representation $\mathbf{C}$ from the source utterance such that: $\mathbf{C} = E_{\text{cnt}}(\mathbf{X}^s)$. Consequently, the generation process is conditioned on a disentangled content representation and the target speaker embedding, $\mathbf{e}^t$. Therefore, the global objective function is:

$$p(\mathbf{Y} \mid \mathbf{X}^s, \mathbf{e}^t) \approx p(\mathbf{Y} \mid \mathbf{C}, \mathbf{e}^t). \tag{5}$$

The distribution of the conditional diffusion probabilistic model parameterized using $\theta$ as Eq. 6, where global objective of VC model training is to minimize the negative log-likelihood of the target data distribution $\mathcal{D}_t$ using conditioning variables. In particular,

$$\min_\theta \ \mathbb{E}_{\mathbf{Y},\mathbf{C},\mathbf{e}^t}\left[-\log p_\theta(\mathbf{Y} \mid \mathbf{C}, \mathbf{e}^t)\right], \tag{6}$$

where optimisation can be achieved using the iterative denoising approach, and the model learns to reconstruct $\mathbf{Y}$ from Gaussian noise conditioned on $\mathbf{C}$ and $\mathbf{e}^t$.

### B. Proposed Approach: SSMamba-VC

We propose **SSMamba-VC**, a selective-scan-based hierarchical diffusion framework for many-to-many voice conversion that explicitly disentangles linguistic content, prosody, and speaker identity while maintaining linear-time sequence modeling. The overall architecture is illustrated in Fig. 1 and consists mainly of the following two key components:

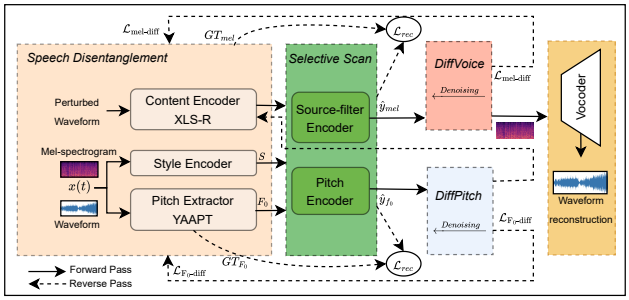


Fig. 1. Overall architecture of the SSMamba-VC. After [14].

**Mamba-based Content Encoder**: Conventional attention-based encoders suffer from quadratic complexity ($\mathcal{O}(n^2d)$) and temporal over-smoothing when modeling long speech sequences. To address this, we adopt an **SS Mamba encoder** with linear complexity ($\mathcal{O}(nd)$), enabling long-range dependency modeling without attention (Fig. 2). Given input sequence $\mathbf{x} = \{x_t\}_{t=1}^{T}$, the selective scan updates a latent state:

$$\mathbf{h}_t = \alpha \mathbf{h}_{t-1} + \beta x_t, \tag{7}$$

$$\mathbf{y}_t = \gamma \mathbf{h}_t, \tag{8}$$

where $\alpha, \beta$, and $\gamma$ are *learnable* SS parameters. Stability is enforced via a softplus-constrained parameterization:

$$\mathbf{A} = -\text{softplus}.(\tilde{\mathbf{A}}). \tag{9}$$

To preserve both past and future prosodic cues, we employ a **non-causal, bidirectional scan**:

$$\mathbf{y}_t = \mathbf{y}_t^{\rightarrow} + \mathbf{y}_t^{\leftarrow}, \tag{10}$$

which avoids the prosody collapse observed in strictly causal encoders. The final content representation is:

$$\mathbf{c}_t = \text{Linear}(\text{SiLU}(\mathbf{y}_t)). \tag{11}$$

This design preserves pitch ($F_0$) contours and rhythm while removing the speaker identity, a critical requirement for the high-fidelity VC task.

**Speaker-aware conditioning using diffusion-based reconstruction**: We inject the target speaker embedding $s$ through adaptive modulation:

$$\hat{\mathbf{c}}_t = \mathbf{c}_t \odot \gamma(s) + \beta(s), \tag{12}$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ are learned projections. This **early conditioning** strategy allows the diffusion model to generate speaker-consistent trajectories throughout the denoising process.

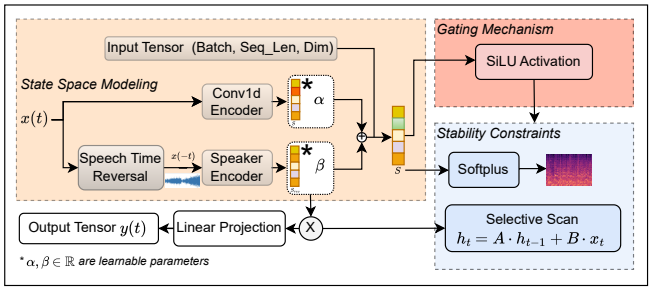


Fig. 2. Proposed architecture of the Source-Filter and Pitch encoder from SSMamba-VC.

TABLE I
COMPARISON OF VCTK AND LIBRITTS DATASETS WITH KEY CHARACTERISTICS AND BIASES. AFTER [15], [16]

| Attribute | CSTR VCTK | LibriTTS |
|---|---|---|
| Total Speakers (Duration) | 110 (≈ 44 hours) | 2,456 (≈ 585 hours) |
| Gender Split | ≈ 63 Male / 47 Female | ≈ 1,300 Male / 1,150 Female |
| Sampling Rate | 48 kHz, 16-bits | 24 kHz, 24-bits |
| Type | Read, studio speech | Read audiobook speech |
| Text Content | Newspaper text, Rainbow Passage, Elicitation prompts | Sentences with original and normalized transcripts |

### C. Training Objective

The model is trained using a score-matching objective over all diffusion scales:

$$\mathcal{L}_{diff} = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{y}_t, t, \hat{\mathbf{c}}, s)\|^2\right]. \tag{13}$$

The total loss is:

$$\mathcal{L} = \mathcal{L}_{diff}^{(f)} + \mathcal{L}_{diff}^{(s)} + \mathcal{L}_{diff}^{(u)} + \lambda \mathcal{L}_{rec}, \tag{14}$$

where $\mathcal{L}_{rec}$ enforces waveform reconstruction consistency via the vocoder. This formulation enables stable many-to-many voice conversion with improved long-term coherence and prosody preservation.

## IV. Experimental Setup

### A. Datasets and Preprocessing

In this work, we used two datasets, namely CSTR VCTK [15] dataset and LibriTTS [16]. Both corpora are multi-speaker English corpora that include clean studio-recorded speech and its transcripts. The VCTK and LibriTTS corpora contain British and North American accents, respectively. Both corpora are used in various downstream speech tasks, such as building ASR systems, multi-speaker TTS, VC, voice cloning, and vocoders. Additional details of both corpora are summarised in Table I. In the preprocessing, we used resampling, channel downmixing, segmentation, pitch ($F_0$) extraction using the pYAAPT algorithm[1], and feature normalization to better learn the pitch data. Dataset splitting was performed, resulting in training and validation sets, with 95% and 5% of the total dataset, respectively.

[1]https://github.com/mcraig2/pyaapt (Last accessed January 31, 2026)

TABLE II
COMPARISON OF THE TRAINING INFRASTRUCTURE, COMPUTATIONAL COST, AND RTF OF SOTA AND DIFFUSION-BASED VC MODELS ON CSTR VCTK [15] DATASET. A DASH (-) INDICATES INFORMATION NOT DISCLOSED BY THE AUTHORS.

| Model | #GPUs (↓) | GPU Used (VRAM) (↓) | FLOPs (↓) | Batch Size (↓) | Epochs / Steps (↓) | Training Time (Days) (↓) | RTF (↓) |
|---|---|---|---|---|---|---|---|
| **SOTA VC Baselines (Pre-trained Models)** | | | | | | | |
| Auto-VC [17] | – | – | – | 32 | 100 K steps | – | – |
| Style-VC [18] | – | – | – | 24 | 100 K steps | – | 0.14 |
| StarGAN-VC2 [19] | – | – | – | 10 | 150 epochs | – | – |
| **Diffusion-based VC Baselines** | | | | | | | |
| Diff-VC [20] | – | – | – | 128 | 500 epochs | – | 0.15 |
| Diff-VC + Ours | 1 | NVIDIA GTX 1080 (8 GB) | 8.87 | 16 | **∼0.2 M steps** | **2** | **0.11** |
| DiffHier-VC [14] | 2 | NVIDIA A100 (40 GB) | 624 | 64 | 2M steps | 5 | 0.32 |
| DiffHier-VC + Ours | 1 | NVIDIA TITAN XP (12 GB) | 12.1 | 32 | **∼0.2 M steps** | **2** | **0.27** |

### B. Simulation Setup

Training experiments for the SSMamba-VC were performed on an Ubuntu-based workstation powered by a Core i5-14 Gen CPU and an NVIDIA GeForce GTX Titan Xp with 12 GB of VRAM, with 32GB of RAM. Model evaluation and inference were performed on a laptop equipped with an NVIDIA GTX 1650 GPU with 4 GB of VRAM. Results are reported for models trained over $1.5 \times 10^5$ iterations.

### C. Model Parameters

SSMamba-VC training performed with initial learning rate $5\times10^{-5}$with exponential decay with the factor 0.9999875. The batch size and the training steps varied across experiments, as shown in Table II. Audio processed at the sampling rate of 16 kHz and log-scale Mel spectrogram extracted using the 1280 samples FFT, with a window length of 1280 and hop size of 320. Each Mamba block uses the hidden dimension equal to the encoder hidden size, with an expansion factor of 2 and a selective SS size of 16. Local temporal patterns captured using a depthwise convolution with a kernel size of 4. Discretisation parameters were estimated using a low-rank projection with rank $\lceil d_{\text{model}}/16 \rceil$ for input-friendly dynamics, and the continuous-time state matrix was parameterised in log-space to ensure stability during sequential training.

## V. EXPERIMENTAL RESULTS

The comprehensive evaluation results of all the considered VC systems are summarized in Table III, which reports subjective, objective, and quantitative metrics for SOTA VC baselines, by considering the Auto-VC [17], Style-VC [18], and StarGAN-VC2 [19], which are mainly inferred using the pretrained weights released by the authors, and diffusion-based VC models, e.g., Diff-VC [20], DiffHier-VC [14] and their variants (+ Our) are re-trained with custom settings on our hardware to evaluate the effectiveness of the proposed approach.

### A. Performance Evaluation

To assess the computational efficiency and practical deployability of the proposed approach, we compare training costs, hardware requirements, and real-time inference performance with SOTA and diffusion-based VC systems. Table II shows that the proposed method delivers substantial efficiency improvements over diffusion baselines while using significantly fewer resources. Auto-VC [17], Style-VC [18], and StarGAN-VC2 [19] baselines, pretrained models adapted in this work, for which the training infrastructure and computational costs are not publicly available. However, GAN-based VC was trained only on a batch size of 10, whereas Auto-VC [17] and Style-VC [18] were trained with batch sizes of 32 and 24, respectively. For Diff-VC [14], our approach reduces the RTF from 0.15 to 0.11, with a 26.7% faster inference speed and decreases the training budget by approximately 60% (from 500 epochs to only ∼0.2 M steps) on a single 8 GB GPU. In the hierarchical settings, DiffHier-VC + Ours reduces the GPU memory usage from 80 GB VRAM to 12 GB VRAM (with ∼85% reduction), reduces the training time from 5 days to 2 days (with ∼60% reduction), and lower computational complexity from 624 to 12.1 FLOPS (with ∼98% reduction), while reducing RTF 0.32 to 0.27 (with ∼15.6% improvement) by utilizing selective SSM in diffusion-based VC to achieve high-fidelity synthesis with reduced computational cost, in order to make system suitable for real-world deployment.

### B. Subjective Results

This study uses NISQA-MOS [21] and UT-MOS [22] in order to assess perceptual naturalness and speech intelligibility through subjective evaluation. As presented in Table III, the proposed approach outperforms the baselines among the SOTA baselines. SOTA-VC GAN-based baseline, StarGAN-VC2 [19] achieves a high NISQA-MOS of $4.29 \pm 0.38$, with substantial perceptual similarity with reference speech utterances. Style-VC [18] achieves the highest UT-MOS of 3.43, indicating that the generated utterances from Style-VC are more natural than those from SOTA VC baseline methods, such as Auto-VC [17] and StarGAN-VC2 [19]. In the diffusion-based VC baselines, Diff-VC [20] achieves the highest NISQA-MOS of $4.61\pm0.04$, by surpassing all baseline models. It shows the intrinsic strength of diffusion-based

TABLE III
COMPREHENSIVE SUBJECTIVE (REPORTED WITH 95% CI), OBJECTIVE, AND QUANTITATIVE EVALUATION ANALYSIS ON CSTR VCTK [15] DATASET.

| Model | Subjective | | Objective | | | | | | Quantitative | |
|---|---|---|---|---|---|---|---|---|---|---|
| | NISQA | UTMOS | PESQ | MCD | MSD | RMSE-$F_0$ | FD | SS | WER | CER |
| | (↑) | (↑) | (↑) | (↓) | (↓) | (↓) | (↓) | (↑) | (↓) | (↓) |
| **SOTA VC Baselines (Pre-trained Models)** | | | | | | | | | | |
| Auto-VC [17] | $3.52 \pm 0.12$ | $2.56 \pm 0.26$ | 1.20 | 3.85 | 144.01 | 55.10 | 14212.87 | 0.96 | 99.22 | 71.08 |
| Style-VC [18] | $4.18 \pm 0.09$ | $\mathbf{3.43 \pm 0.25}$ | 1.17 | **3.51** | **119.68** | **45.95** | **2247.82** | 0.99 | 15.92 | 8.12 |
| StarGAN-VC2 [19] | $\mathbf{4.29 \pm 0.38}$ | $3.24 \pm 0.10$ | **1.30** | 3.96 | 149.28 | 68.86 | 12596.17 | 0.97 | **6.26** | **2.74** |
| **Diffusion-based VC Baselines** | | | | | | | | | | |
| Diff-VC [20] | $\mathbf{4.61 \pm 0.04}$ | $\mathbf{2.96 \pm 0.12}$ | **1.26** | 3.88 | 140.23 | 49.97 | 3553.17 | 0.97 | 96.34 | 73.16 |
| Diff-VC + Ours | $3.95 \pm 0.11$ | $2.82 \pm 0.35$ | 1.11 | **4.12** | **127.72** | **46.85** | **2429.09** | **0.98** | 23.60 | 13.42 |
| DiffHier-VC [14] | $3.88 \pm 0.14$ | $2.76 \pm 0.29$ | 1.14 | 4.12 | 127.72 | 46.85 | 2429.09 | 0.98 | 21.35 | 11.35 |
| DiffHier-VC + Ours | $\mathbf{3.90 \pm 0.11}$ | $\mathbf{2.92 \pm 0.24}$ | **1.15** | **4.33** | **133.12** | **42.44** | **2356.96** | **0.99** | **18.15** | **8.73** |

models for generating perceptually smooth speech. Compared to baseline Diff-VC, the integration proposed method results in a competitive NISQA-MOS (3.95) by reducing perceptual artifacts observed in the Diff-VC [20]. While the baseline achieves slightly higher NISQA-MOS and UT-MOS, the Diff-VC + Ours approach shows higher standard deviations across utterances. This increased variance suggests that the model produces a wider range of prosodic variations in the VC speech. However, when the proposed method is integrated, DiffHier-VC + Ours consistently improves both subjective metrics, achieving $3.90 \pm 0.11$ NISQA-MOS and $2.92 \pm 0.24$ UT-MOS. This improvement indicates that the proposed method enhances perceptual stability while preserving naturalness, particularly in hierarchical diffusion architectures. Furthermore, we observed that Diff-VC [20] frequently produces choppy, robotic artifacts in generated VC utterances, resulting in diminished temporal coherence between the VC speech and the associated linguistic content.

### C. Objective Results

Objective metrics are used to assess the speech quality with respect to spectral, prosodic, and distributional similarity between the generated VC utterances and the reference speech utterances. Lower MCD and MSD values indicate better spectral reconstruction. Among SOTA baselines, Style-VC [18] achieves the lowest MCD and MSD of 3.51 and 119.68, respectively, indicating strong *spectral alignment* between the generated and reference VC utterances. However, diffusion-based systems augmented with the proposed method show competitive performance with limited training, with DiffHier-VC + Ours achieving improved MCD (4.33) and MSD (133.12) compared to its baseline counterpart, demonstrating that the proposed method preserves spectral consistency even under stochastic diffusion sampling. RMSE-$F_0$ measures prosodic accuracy; lower values indicate better pitch ($F_0$) modeling. The proposed method significantly reduces $F_0$ error in both diffusion architectures. In particular, Diff-VC [20] and DiffHier-VC + Ours achieve the lowest RMSE-$F_0$ of 42.44, outperforming the baseline (46.85) by 4.41dB in pitch ($F_0$) trajectory modeling and temporal stability. FD and SS measures capture speaker identity preservation during the VC. The proposed method consistently reduces FD across diffusion models (e.g., 2356.96 for DiffHier-VC + Ours) while maintaining SS at 0.99, indicating that the global distribution of the VC utterances matches that of the reference utterances.

### D. Quantitative Results

ASR-based metrics quantify intelligibility and linguistic content preservation. Among all the models, StarGAN-VC2 [18] achieves the lowest WER (6.26) and CER (2.74), while maintaining strong linguistic consistency. However, diffusion-based baselines suffer from severe intelligibility degradation. Diff-VC [20] achieves the WER of 96.34. In simple terms, the ASR model cannot accurately transcribe. Notably, DiffHier-VC + Ours reduces WER from 21.35 to 18.15 and CER from 11.35 to 8.73, ensuring that the proposed approach significantly reduces diffusion-induced linguistic content degradation while maintaining the *phonetic structure* of the speech in relation to the reference speech utterances.

### E. Observations: Same and cross-Gender conversion

In the same and cross-gender VC scenarios, Table IV encapsulates the performance of diffusion-based models on the VCTK dataset. The baseline Diff-VC and DiffHier-VC models demonstrated relatively stable performance across same-gender VC conditions, such as M2M and F2F. However, cross-gender, e.g., M2F and F2M, reveals the noticeable degradation due to significant acoustic and physiological mismatches, including differences in pitch ($F_0$) range, formant structures, and spectral characteristics. With the proposed approach ("+ Ours"), NISQA-MOS [21] shows consistent improvement in several challenging cross-gender cases, indicating strong modeling capability for large-domain speech prosody transformation and effective generalization across diverse prosodic variations. However, these gains lead to increased variance, as reflected in larger standard deviations. This behavior arises from increased model capacity, which improves the expressiveness but slows convergence and increases instability under limited training. The resulting trade-off between quality and stability is due to the undertrained diffusion models. Even with the proposed architecture, it shows strong potential for cross-gender VC by providing without large-scale training or larger computational budgets.

TABLE IV
VOICE CONVERSION PERFORMANCE FOR DIFFUSION-BASED MODELS ON SAME AND CROSS-GENDER CONVERSION ON CSTR VCTK [15] AND LIBRITTS [16] DATASET

| Model | Conversion Type | NISQA-MOS (↑) | UTMOS (↑) |
|---|---|---|---|
| Diff-VC [20] | M→M | **4.23 ± 0.23** | **2.95 ± 0.25** |
| | F→F | **3.72 ± 0.38** | **2.96 ± 0.43** |
| | M→F | 3.59 ± 0.77 | **2.84 ± 0.34** |
| | F→M | 4.07 ± 0.53 | **3.05 ± 0.30** |
| Diff-VC + Ours | M→M | 3.93 ± 1.34 | 2.76 ± 1.09 |
| | F→F | 3.47 ± 0.75 | 2.60 ± 0.96 |
| | M→F | **4.09 ± 0.23** | 2.81 ± 0.59 |
| | F→M | **4.08 ± 0.67** | 2.87 ± 0.49 |
| DiffHier-VC [14] | M→M | 3.93 ± 1.34 | 2.76 ± 1.09 |
| | F→F | 3.47 ± 0.75 | 2.56 ± 0.88 |
| | M→F | 4.09 ± 0.13 | 2.82 ± 0.55 |
| | F→M | 3.47 ± 0.67 | **2.89 ± 0.53** |
| DiffHier-VC + Ours | M→M | **4.23 ± 0.23** | **2.91 ± 0.49** |
| | F→F | **3.72 ± 0.38** | **2.62 ± 0.40** |
| | M→F | **3.59 ± 0.77** | 2.37 ± 0.52 |
| | F→M | **4.07 ± 0.53** | 2.56 ± 0.81 |

## VI. SUMMARY AND CONCLUSIONS

In this paper, we introduced SSMamba-VC, a diffusion-based VC framework that replaces the *quadratic* complexity of the heavy attention mechanism with *linear* complexity by using a SS-based Mamba model in both the models, thereby stabilizing the reverse of the Gaussian diffusion process for long-range prosody preservation. By integrating selective scan into both the content encoder and the diffusion decoder, the proposed method effectively minimizes reliance on large-scale training while preserving phonetic structure and speaker identity. Subjective, objective, and quantitative evaluations demonstrate that with limited training, diffusion-based VC improves intelligibility, speaker similarity, and linguistic content modeling over SOTA diffusion-based baselines. In future work, we plan to focus on scaling the encoder-decoder architecture for zero-shot scenarios and multilingual voice conversion to improve robustness to non-parallel data. Still, SSM models can preserve long-range prosody in VC, however, they are insufficient to control prosodic attributes, such as pitch range, speaking rate, and emotion. Designing the controllable prosody mechanism using a linear-time SSM-based VC framework remains an open research problem.

## ACKNOWLEDGMENT

The authors sincerely thank the *Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India,* through project titled 'BHASHINI' under Grant 11(1)2022-HCC(TDIL).

## REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS, pp. 2672–2680*, 2014, Montreal, Canada.

[2] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *EUSIPCO, pp. 2100–2104*, 2018, Rome, Italy.

[3] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Conference on Neural Information Processing Systems (NeurIPS), pp. 2234–2242*, 2016, Barcelona, Spain.

[4] R. Ferro, N. Obin, and A. Roebel, "CycleGAN voice conversion of spectral envelopes using adversarial weights," in *EUSIPCO, pp. 406–410*, 2021, Amsterdam, Netherlands.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS, pp. 6840–6851*, 2020, Vancouver, Canada.

[6] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv:2011.13456*, 2020 {Last Accessed Date: January $30^{th}$, 2026}.

[7] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning (ICML), pp. 8599–8608*, 2021, Virtual Conference.

[8] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *NeurIPS, pp. 11895–11907*, 2019, Vancouver, Canada.

[9] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv:2202.00512*, 2022 {Last Accessed Date: January $30^{th}$, 2026}.

[10] N. Li, C. Wang, Y. Gu, and Z. Li, "Mitigating timbre leakage with universal semantic mapping residual block for voice conversion," *arXiv:2504.08524*, 2025 {Last Accessed Date: January $30^{th}$, 2026}.

[11] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint*, 2021 {Last Accessed Date: January $30^{th}$, 2026}.

[12] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," *NeurIPS, Vol. 35, pp. 22982–22994*, 2022, New Orleans, USA.

[13] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Conference on Language Modeling*, 2024.

[14] 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," in *INTERSPEECH, pp. 2283–2287*, 2023, Dublin, Ireland.

[15] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[16] 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:1904.02882*, 2019 {Last Accessed Date: January $30^{th}$, 2026}.

[17] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning (ICML) , pp. 5210–5219*, 2019, Long Beach, CA, USA.

[18] I.-S. Hwang, S.-H. Lee, and S.-W. Lee, "StyleVC: Non-parallel voice conversion with adversarial style generalization," in *International Conference on Pattern Recognition (ICPR), pp. 23–30*, 2022, Montreal, Canada.

[19] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *INTERSPEECH, pp. 1349–1353*, 2021, Brno, Czech Republic.

[20] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," vol. arXiv:2109.13821, 2021 {Last Accessed Date: January $30^{th}$, 2026}.

[21] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *INTERSPEECH, pp. 2127–2131*, 2021, Brno, Czech Republic.

[22] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," in *INTERSPEECH, pp. 4521–4525*, 2022, Incheon, South Korea.