O_O-VC: Synthetic Data-Driven One-to-One Alignment for Any-to-Any Voice Conversion

Anonymous ACL submission

Abstract

Traditional voice conversion (VC) methods typically attempt to separate speaker identity and linguistic information into distinct representations, which are then combined to reconstruct the audio. However, effectively disentangling these factors remains challenging, often leading to information loss during training. In this paper, we propose a new approach that leverages synthetic speech data generated by a high-quality, pretrained multispeaker text-to-speech (TTS) model. Specifically, synthetic data pairs that share the same linguistic content but differ in speaker identity are used as input-output pairs to train the voice conversion model. This enables the model to learn a direct mapping between source and target voices, effectively capturing speaker-specific characteristics while preserving linguistic content. Additionally, we introduce a flexible training strategy for any-to-any voice conversion that generalizes well to unseen speakers and new languages, enhancing adaptability and performance in zero-shot scenarios. Our experiments show that our proposed method achieves a 16.35% relative reduction in word error rate and a 5.91% improvement in speaker cosine similarity, outperforming several state-of-the-art methods. Voice conversion samples can be accessed at: https://voiceconversion-emnlp-2025.github.io/

1 Introduction

Voice conversion specifically aims to transform a source speaker's voice to match a target speaker while preserving the original linguistic content. This is typically done by disentangling speech into content and speaker identity representations, which are combined during training to reconstruct the audio. At inference time, the source content is paired with a target speaker embedding to generate the converted speech.

Several methods have been proposed for VC, with supervised training being a common approach. Content encoders are trained with text labels to extract linguistic features, and speaker encoders use speaker labels to capture identity-specific traits (Huang, 2023; Liu et al., 2021). Alternatively, phonetic posteriorgrams (PPGs) can be used directly as content representations (Sun et al., 2016; Tian et al., 2018). However, both approaches often struggle to capture speaker-independent prosody and accent information. Moreover, training content encoders as ASR models can introduce alignment errors or recognition errors, which can negatively impact conversion quality (Hussain et al., 2023).

On the other hand, some methods avoid using text labels by leveraging self-supervised learning (SSL) to extract high-level phonetic representations (Polyak et al., 2021; Lin et al., 2021; Huang et al., 2022b,c). These approaches aim to remove speaker identity from source audio while preserving speaker-independent features such as accent and content. To achieve this separation, techniques such as vector quantization (Wu and Lee, 2020), instance normalization (Chen et al., 2021b), heuristic transformation (Neekhara et al., 2024), bottleneck, and data augmentation (Li et al., 2023) are commonly applied. However, despite these efforts, such methods still struggle to completely eliminate speaker information from the source speech. This often leads to speaker leakage, where the converted audio retains unintended characteristics of the source speaker, resulting in mismatches between the synthesized voice and the intended target speaker (Baas et al., 2023).

Previous studies have primarily focused on feature disentanglement methods and audio reconstruction in voice conversion systems. However, feature disentanglement remains a challenging task and training models to reconstruct the audio may not be well suited for the voice conversion objective, which inherently involves transforming

speech from one speaker to another. To address these limitations, we propose a novel training strategy that leverages synthetic data generated by a high-quality multi-speaker text-to-speech (TTS) system to directly establish input-output mappings for voice conversion, bypassing the need for traditional reconstruction-based approaches. Despite the high fidelity of synthetic data, such TTS systems are typically constrained to a fixed set of speakers, limiting their applicability to any-to-any voice conversion, particularly for unseen speakers. To mitigate this issue, we introduce a training framework that promotes generalization to unseen speakers without relying on additional text or speaker labels, thereby enhancing the system's adaptability and performance in zero-shot voice conversion scenarios. In summary, we make the following contributions:

- Synthetic data for Voice Conversion Training: We propose the use of synthetic data generated by a high-quality multi-speaker TTS system to train voice conversion models. This approach eliminates the need for audio reconstruction and feature disentanglement, enabling direct learning of input-output mappings.
- Improved Generalization: We introduce a training strategy that allows the model to generalize to unseen speakers or unseen languages, making it well suited for zero-shot voice conversion.
- We validate the effectiveness of our approach through extensive experiments, showing significant improvements over traditional reconstruction-based methods, especially in challenging zero-shot settings.

2 Literature Review

The goal of voice conversion (VC) is to transform the voice of a source speaker into that of a target speaker while preserving the original linguistic content. Achieving this requires an effective decomposition of speech signals into distinct components such as linguistic content, speaker timbre, and prosodic characteristics. Early VC systems were typically trained as speech-to-speech models on parallel datasets, where multiple speakers uttered the same sentences, defining the task as a sequence-to-sequence autoregressive problem (Sun et al., 2015; Chen et al., 2014). Recent VC approaches have focused on reconstructing speech using disentangled representations of linguistic content and speaker identity.

2.1 Text-Based Method

A common strategy is to leverage pretrained automatic speech recognition (ASR) models to extract phonetic features such as PPGs, which provide a speaker-independent representation of the input speech (Sun et al., 2016; Liu et al., 2021; Tian et al., 2018). Specifically, speaker information is obtained using a pretrained speaker verification (SV) model. The speaker embeddings are combined with the content features during decoding, allowing the system to generate speech in the target speaker's voice. Some approaches leverage hidden text representations from pretrained multispeaker text-to-speech (TTS) models, using them either as semantic features or as target representations for learning a mapping from audio to text (Park et al., 2020; Zhang et al., 2021).

Despite their advantages, text-based methods suffer from several limitations. PPGs and other textual representations often fail to capture fine-grained attributes such as accent, prosody, and speaker-independent speaking style. As a result, these systems often produce speech that lacks expressiveness and sounds overly neutral (Hussain et al., 2023). Although ASR-based disentanglement methods have shown progress in separating speaker and content information, their reliance on textual supervision and limited prosodic modeling remain significant challenges for achieving natural and expressive voice conversion across diverse speakers and languages.

2.2 Text-Free Method

To address the limitations of text-based methods, text-free approaches have emerged, leveraging self-supervised learning models to extract content representations without requiring transcriptions (Polyak et al., 2021; Lin et al., 2021; Huang et al., 2022b,c). Although self-supervised learning (SSL) features capture high-level information related to linguistic content, they often retain residual speaker characteristics. To address this, methods such as bottleneck layers (Li et al., 2023), vector quantization (Wu and Lee, 2020), and instance normalization (Chen et al., 2021b) have been proposed to compress SSL features and extract speaker-independent content representations. However, effective disentanglement heavily depends on the choice of bottleneck configuration: if the bottleneck dimension

is too large, speaker information may be retained; if too small, important content information can be lost. A similar trade-off exists in vector quantization: large codebooks may retain speaker traits, while overly small codebooks may lead to excessive loss of content information. Moreover, these compression techniques often degrade the quality of the generated audio, and full disentanglement of speaker and content information remains an open challenge.

2.3 KNN Method

Recent work has introduced k-nearest neighbor (kNN)-based voice conversion methods (Baas et al., 2023), offering a simpler alternative to traditional feature disentanglement approaches. These methods operate directly on frame-level self-supervised representations extracted from both source and target speech, which encode both phonetic and speaker-specific information. Voice conversion is performed by replacing each frame of the source with its nearest neighbor from the target set, followed by vocoder-based synthesis.

However, in one-shot scenarios, the limited size of the target set restricts the pool of candidate neighbors, often resulting in higher word error rates. To address this, the Phoneme Hallucinator (Shan et al., 2024) was proposed, leveraging a permutation network to synthesize additional target representations and expand the neighbor set, thereby improving intelligibility. Nonetheless, averaging features in kNN-based retrieval can lead to oversmoothing, reducing speaker distinctiveness and clarity in the synthesized speech.

2.4 Diffusion Method

Diffusion models have shown exceptional performance in generative tasks across a variety of domains, including images, videos, and audio. In speech processing, diffusion models have been successfully applied to tasks such as audio generation (Kong et al., 2021; Chen et al., 2021a) and text-to-speech (TTS) synthesis (Popov et al., 2021; Huang et al., 2022a). Furthermore, diffusion models have been investigated for VC tasks with the aim of enhancing the conversion process. In particular, diffusion-based VC models (Popov et al., 2022) have demonstrated high performance in zero-shot speaker adaptation through iterative sampling processes. While recent works such as Diff-HierVC (Choi et al., 2023) and DDDM-VC (Choi et al., 2024) have further improved zero-shot VC performance through source-filter disentanglement and disentangled denoising processes, the audio quality of diffusion models is still limited.

3 Methodology

In text-free VC systems, content and speaker identity are often not fully disentangled. As a result, speaker information can leak into the content representation, which undermines the system's ability to perform clean speaker conversion. This leakage reduces the system's generalization to unseen voices or speaking styles and often results in converted speech that retains characteristics of the source speaker.

In text-based VC, ASR-derived content representations are highly sensitive to transcription errors, mispronunciations, and noisy labels, which can compromise their reliability and degrade the quality of converted speech (Sun et al., 2016; Liu et al., 2021; Tian et al., 2018). Some approaches attempt to leverage knowledge transfer from hidden representations of text encoders in multispeaker TTS models (Park et al., 2020; Zhang et al., 2021). However, mapping audio representations directly to these text-based features is a difficult task. In addition, this process typically requires an explicit alignment mechanism between speech and text, which introduces further complexity.

As an alternative solution, synthetic data offers several advantages for voice conversion. When both source and target audio are generated from the same linguistic content, it provides a clean and direct supervisory signal. This shared content allows precise frame-level alignment between source and target audio, enabling more stable and fine-grained learning of the conversion function. Unlike traditional approaches that rely on symbolic representations (e.g., phonemes or characters), synthetic data eliminates the need for such intermediates and avoids issues like label noise commonly found in real-data training. Furthermore, with controlled or predefined durations, speaker-independent features are inherently aligned across domains, removing the need for forced alignment algorithms used in previous work (Park et al., 2020; Zhang et al., 2021). Unlike methods that map audio to hidden text representations from multispeaker TTS models, our approach directly maps source audio to target audio, simplifying the learning process. This enables one-to-one frame alignment, allowing the model to focus more effectively on speaker trans-

formation while preserving linguistic content. Finally, synthetic data enables the creation of diverse speaker pairs with uniform content, supporting the learning of generalizable speaker conversion mappings. Motivated by these advantages, this work is the first to propose using synthetic data as a training paradigm for voice conversion models.

3.1 Synthetic Data Strategy

Building upon these advantages, we propose a synthetic data strategy to improve the disentanglement of speaker and content representations in voice conversion. Instead of relying on real-world utterances, we generate high-quality synthetic pairs with identical linguistic content but varying speaker identities. This provides ideal supervision for isolating speaker-independent content.

We use a multi-speaker TTS system that produces natural, intelligible speech across speakers from a shared linguistic latent space. Our selection criteria for the TTS system are: (1) the generated speech must be of high fidelity, exhibiting natural prosody and clarity, and (2) the model must synthesize source and target utterances from the same linguistic latent space, ensuring consistent phonetic and prosodic alignment across speakers. These criteria ensure that the synthetic speech pairs are perfectly aligned in linguistic structure while differing only in speaker identity.

We adopt VITS (Kim et al., 2021) as the backbone of our TTS system because it combines variational inference, flows, and adversarial learning to generate high-quality speech with precise duration control and the ability to sample two audios with different speakers conditioned on a shared latent linguistic representation. This enables the creation of large-scale, controllable training data that improves disentanglement, reduces speaker leakage, and enhances voice conversion robustness.

Given a text input c_{text}, source speaker embedding s_{src}, target speaker embedding s_{tgt}, and noise vector w, we generate a pair of utterances by first encoding the linguistic content:

$$h_{\text{text}} = \text{TextEncoder}(c_{\text{text}}) \tag{1}$$

We then sample a global speaker token:

$$g = \text{random}(s_{\text{src}}, s_{\text{tgt}}) \tag{2}$$

and predict the duration based on speaker condition using the duration predictor (DP):

$$\text{dur}_{\text{text}} = \text{DP}(h_{\text{text}}, g, w) \tag{3}$$

Next, we project the encoded content into a linguistic latent distribution:

$$\mu_p, \sigma_p = \text{Projector}(h_{\text{text}}) \tag{4}$$

We then sample latent linguistic features with duration expansion, where the length regulator (LR) is defined as:

$$\mu_p, \sigma_p = \text{LR}(\mu_p, \sigma_p, \text{dur}_{\text{text}}) \tag{5}$$

$$z_p = \mu_p + \sigma_p \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{6}$$

To generate speaker-specific representations, we apply the inverse flow conditioned on each speaker:

$$z_{\text{src}} = \text{Flow}^{-1}(z_p, s_{\text{src}}) \tag{7}$$

$$z_{\text{tgt}} = \text{Flow}^{-1}(z_p, s_{\text{tgt}}) \tag{8}$$

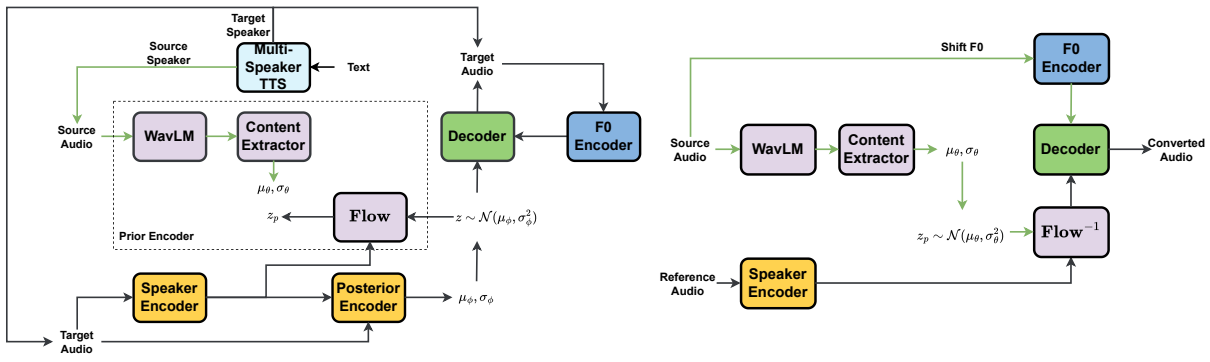
Finally, we decode both representations into waveform audio:

$$a_{\text{src}} = \text{Decoder}(z_{\text{src}}), \quad a_{\text{tgt}} = \text{Decoder}(z_{\text{tgt}}) \tag{9}$$

This process results in a pair of utterances with identical content, prosody, and speaker-independent information, but differing in speaker identity. Such pairs provide a clean and consistent training signal for learning voice conversion.

3.2 Model Overview

After generating synthetic data consisting of source and target speech pairs for supervised training, these utterances are directly utilized as input-output pairs for the voice conversion model. As the backbone architecture, we adopt a VITS-base model. Following the design of FreeVC (Li et al., 2023), our model structure retains its core components. However, we note that while the source and target speech pairs share the same underlying linguistic content, the target speech is conditioned on a different speaker identity, which primarily manifests in variations in pitch. To avoid mismatch between input and output during training, we incorporate the fundamental frequency ($F0$) of the target speech as an additional conditioning feature when decoding the final audio. The general model pipeline is illustrated in Figure 1.



(a) Training (b) Inference

Figure 1: Voice conversion with synthetic data.

3.2.1 Training Procedure

In the training phase, the source and target audios are processed through different stages:

Source Audio Processing: The source audio is passed through WavLM (Chen et al., 2022) and a Content Extractor to obtain a distribution of content features $\mathcal{N}(\mu_\phi, \sigma_\phi^2)$.

Target Audio Processing: The target audio is passed through a Speaker Encoder and a Posterior Encoder to extract the posterior latent distribution $\mathcal{N}(\mu_\phi, \sigma_\phi^2)$. Then, a latent variable z is sampled from this distribution $z \sim \mathcal{N}(\mu_\phi, \sigma_\phi^2)$. The sampled latent vector z is passed through a flow-based module to obtain z_p, transforming the posterior distribution to match the prior distribution. A Kullback–Leibler (KL) divergence loss is calculated to minimize the discrepancy between the posterior and prior distributions.

Fundamental Frequency ($F0$) Adjustment: To address the mismatch in $F0$ between the source and target audio, we extract the $F0$ of the target audio and pass it through an $F0$ encoder to obtain pitch-related features. The decoder then takes the transformed latent representation along with the $F0$ features to generate the target audio. In this work, we extract $F0$ using Parselmouth[1].

3.2.2 Fine-Tuning Adaptation

Training with synthetic data often results in poor generalization to unseen speakers. Multi-speaker TTS models are typically trained with fixed speaker embeddings, which can lead to reduced speaker similarity for out-of-domain speakers. Moreover, slight discrepancies in the linguistic content between source and target utterances may persist, despite both being generated from the same text.

[1] https://github.com/YannickJadoul/Parselmouth

To mitigate these challenges, we adopt a two-phase training strategy:

- **Phase 1:** Train the model with synthetic data, where the WavLM and Content Extractor components are responsible for learning independent speaker representations. The WavLM Large is frozen during training.

- **Phase 2:** Fine-tune the model using real large multispeaker speech recordings corpus with a reconstruction-based objective, the input and output of the model is the same audio. During this phase, we freeze WavLM and Content Extractor to preserve speaker-independent representations learned in phase 1. This fine-tuning phase helps the model adapt to new speakers and improves the fidelity of linguistic content. Additionally, the model becomes more versatile and can easily adapt to different domains, such as various languages or accents.

3.2.3 Inference

During inference, the source audio is processed by WavLM and the Content Extractor to obtain the content distribution $\mathcal{N}(\mu_\phi, \sigma_\phi^2)$. A latent sample z_p is drawn and combined with the speaker embedding from the reference audio via the inverse flow model, producing a feature that captures both content and speaker information.

To address the mismatch in $F0$ between the source and reference audio, we shift $F0$ source to $F0$ target with same median level, the following

steps are performed:

$$F0_{\text{src}} = \text{Get_F0}(\text{Source Audio}) \quad (10)$$

$$F0_{\text{ref}} = \text{Get_F0}(\text{Reference Audio}) \quad (11)$$

$$\log F0_{\text{shifted}} = \log(F0_{\text{src}}) - \text{med}(\log(F0_{\text{src}})) + \text{med}(\log(F0_{\text{ref}})) \quad (12)$$

$$F0_{\text{shifted}} = \exp(\log F0_{\text{shifted}}) \quad (13)$$

The shifted $F0$ is then used as a conditioning feature along with the fused linguistic and speaker representation to generate the final audio output.

3.2.4 Objective Function

Following the methodology proposed in (Li et al., 2023), we formulate the objective function by combining losses from conditional variational autoencoders (CVAE) and generative adversarial networks (GAN) (Mao et al., 2017; Larsen et al., 2016). CVAE-related losses include the KL divergence loss L_{kl}, which measures the discrepancy between the prior and posterior distributions of the flow-based model, and a phase-dependent reconstruction/conversion loss, either L_{rec} in phase 2 or L_{cv} in phase 1, defined as the L1 distance between the predicted and target mel-spectrograms. GAN-related losses include the adversarial loss for the discriminator $L_{\text{adv}}(D)$, the adversarial loss for the generator $L_{\text{adv}}(G)$, and the feature matching loss $L_{\text{fm}}(G)$. We further incorporate two distillation losses into the total objective. The final loss function is defined as:

$$L(D) = L_{\text{adv}}(D) \quad (14)$$

$$L(G) = L_{\text{rec/cv}} + L_{\text{kl}} + L_{\text{adv}}(G) + L_{\text{fm}}(G) \quad (15)$$

4 Experiment Setup

4.1 Datasets

For phase 1 of training, we use synthetic speech generated by a publicly available pretrained VITS model[2]. Specifically, we adopt the model released in the official repository. The amount of synthetic data corresponds to the VCTK training set used in the original VITS implementation. For each sample, the target audio is synthesized using the ground-truth text and speaker ID, while the source audio is generated by sampling a different random speaker ID. In phase 2, we fine-tune the model on the LibriSpeech dataset (Panayotov et al., 2015), using the train-clean-360 and train-clean-100 subsets, totaling approximately 460 hours of speech from 1,172 speakers. Evaluation is conducted on the test-clean subset under any-to-any voice conversion scenarios.

4.2 Model Configuration and Training Details

We follow the implementation and hyperparameter setup of FreeVC (Li et al., 2023). Training occurs in two phases: up to 450k steps on synthetic data, followed by 150k steps of fine-tuning on real speech. All experiments are conducted on four NVIDIA Tesla A100 GPUs.

We compare our method with several recent state-of-the-art voice conversion models, including FreeVC (Li et al., 2023), DDDM-VC (Choi et al., 2024), Diff-HierVC (Choi et al., 2023), FaCodec (NaturalSpeech 3) (Ju et al., 2024), and KNN-VC (Baas et al., 2023). For all baselines, we use official publicly released pretrained models. For KNN-VC, we use an 8-minute real speech segment as the reference pool for nearest-neighbor retrieval.

4.3 Evaluation Metrics

Objective Evaluation: We evaluate system performance using four objective metrics: Character Error Rate (CER), Word Error Rate (WER), Speaker Encoder Cosine Similarity (SECS), and Objective Naturalness. CER and WER assess intelligibility of both source and converted speech, using the HuBERT model[3] (Hsu et al., 2021). SECS measures speaker similarity using the cosine similarity between embeddings extracted by Resemblyzer[4]. Naturalness is assessed using NISQA (Mittag et al., 2021), which estimates perceptual speech quality without reference audio. We compute these metrics on 1,000 randomly sampled audio pairs from LibriSpeech test-clean.

Subjective Evaluation: For human evaluation, we use Mean Opinion Score (MOS) and Speaker Similarity Mean Opinion Score (SMOS). MOS rates naturalness, while SMOS rates speaker similarity, both on a 1–5 scale. We randomly select 30 audio pairs from the objective set, each evaluated by three different annotators, resulting in a total of 540 labeled audio samples. A total of 12 volunteer listeners participate in the evaluation. Final scores are calculated by averaging the ratings across annotators for each pair to ensure reliability. In voice

[2] https://github.com/jaywalnut310/vits

[3] https://huggingface.co/facebook/hubert-large-ls960-ft

[4] https://github.com/resemble-ai/Resemblyzer

Model	Objective Evaluation				Subjective Evaluation		
	SECS↑	WER↓	CER↓	NISQA↑	MOS↑	SMOS↑	B-MOS↑
FreeVC	75.66	2.37	0.78	**4.60**	**3.60 ± 0.26**	3.01 ± 0.28	3.31
KNN-VC	78.33	2.16	0.62	3.92	3.17 ± 0.23	2.89 ± 0.21	3.03
Diff-Hier	81.42	3.82	1.51	3.80	2.87 ± 0.28	3.42 ± 0.25	3.15
DDDM-VC	81.86	6.84	2.92	3.91	2.89 ± 0.28	**3.61 ± 0.23**	3.25
Facodec	81.54	2.08	0.64	3.90	2.49 ± 0.29	2.66 ± 0.27	2.58
O_O-VC (Ours)	**86.70**	**1.74**	**0.53**	4.04	3.42 ± 0.24	3.48 ± 0.23	**3.45**

Table 1: Any-to-any voice conversion results. **Blue** indicates best performance, Underline indicates second best. Subjective evaluation results showing MOS and SMOS scores, along with 95% confidence intervals.

conversion systems, both MOS and SMOS help evaluate model quality. To provide an overall comparison, we introduce a new metric called balance-MOS (B-MOS), defined as the average of these two scores.

5 Results and Analysis

5.1 Zero-Shot Voice Conversion

We evaluate our model in a zero-shot setting, where the target speaker is unseen during training. The results in Table 1 demonstrate that our model achieves the best performance in terms of content consistency, with the lowest WER and CER. Furthermore, our model archives the second highest MOS, only slightly behind FreeVC. This can be attributed to the fact that FreeVC is trained on a high-quality speech dataset, whereas our model is fine-tuned using LibriSpeech, which is of comparatively lower quality. Despite FreeVC's strong MOS, it performs notably worse in terms of speaker similarity and content intelligibility compared to our model. Although DDDM-VC achieves the highest Similarity Mean Opinion Score (SMOS), its speech quality is comparatively poor. Overall, our model achieves the best intelligibility while maintaining a strong balance between naturalness (MOS) and speaker similarity (SMOS), outperforming recent systems in a zero-shot scenario.

5.2 Ablation Study

We conduct an ablation study by modifying or removing key modules to evaluate their individual contributions, summarized in Table 2. We observe that removing the use of synthetic data, the F0 encoder, or phase 2 fine-tuning each leads to a noticeable drop in intelligibility, highlighting the importance of all three components. Eliminating phase 2 fine-tuning also causes a significant reduction in speaker similarity, likely due to the limited speaker diversity in the phase 1 dataset. However, since the phase 1 data is of higher quality, the phase 2 adaptation may slightly reduce speech quality.

Model	SECS↑	WER↓	CER↓	NISQA↑
O_O-VC (Ours)	86.70	**1.74**	**0.53**	4.04
w/o F0 Encoder	**87.00**	2.07	0.61	3.85
w/o Finetuning	70.78	2.18	0.66	4.59
FreeVC	75.66	2.37	0.78	**4.60**

Table 2: Ablation study results.

We quantitatively evaluate how effectively the prior encoder removes speaker information by comparing our model to FreeVC, which shares the same backbone architecture. Our goal is to demonstrate that training with synthetic data significantly improves the removal of speaker identity from source audio. To assess this, we use three clustering evaluation metrics: Adjusted Rand Index (ARI), Normalized Mutual Information (NMI) and Silhouette Score. The ARI measures the similarity between predicted clusters and true speaker labels, adjusted for chance. A lower ARI indicates that the clusters do not correspond well to speaker identities, suggesting better speaker information removal. NMI measures the amount of shared information between the predicted and true clusters; lower values indicate weaker correlation and thus stronger speaker anonymization. The Silhouette Score reflects how well each embedding fits within its cluster compared to the others. Lower scores imply that the model's embeddings are not tightly grouped by speaker, further indicating that speaker identity has been suppressed. The quantitative results are shown in Table 3. Our model, trained with synthetic data, consistently achieves lower scores across all metrics, demonstrating its improved ability to remove speaker-specific information compared to FreeVC.

To provide an intuitive understanding, we visu-

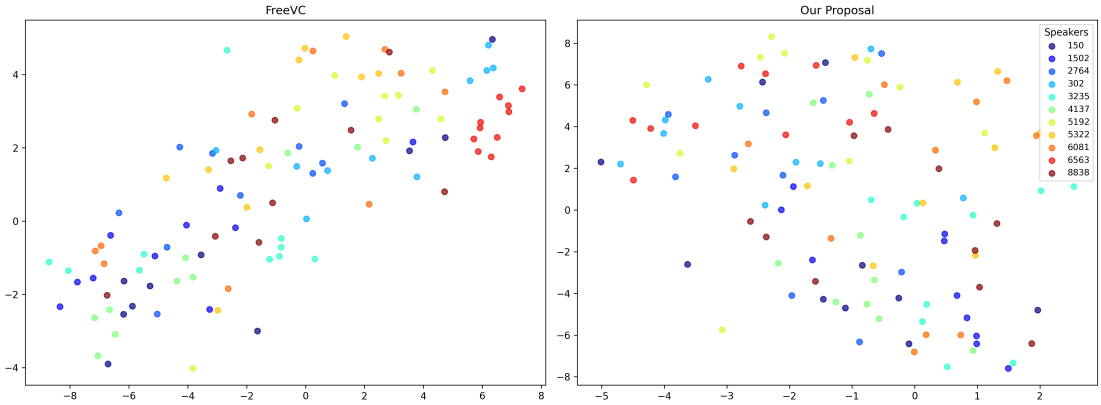


Figure 2: T-SNE visualization of speaker-independent features. More distributed points with no clusters indicate better speaker independence.

Model	ARI↓	NMI↓	Silhouette↓
O_O-VC (Ours)	**0.07**	**0.31**	**0.15**
FreeVC	0.13	0.41	0.17

Table 3: Evaluation of speaker information removal.

alize the speaker-independent features using t-SNE in Figure 2. Our model's embeddings are more evenly dispersed across speakers (different colors), indicating greater speaker independence. In contrast, FreeVC shows noticeable clustering, such as 6563 and 5192, which indicates that its features preserve a greater degree of speaker-specific information.

We provide further ablation studies on how synthetic alignment data and emotion information are preserved in the Appendices A.1 and A.2.

5.3 Adaptation to New Languages

We evaluate the adaptability of our approach to new languages by applying the model to speech data from previously unseen linguistic domains. In this experiment, we fine-tune the model in phase 2 using speech from three languages: Chinese (ZH), Italian (IT), and Vietnamese (VI). For Chinese, we use the AISHELL-3 training dataset (Shi et al., 2021); for Vietnamese, we use the same dataset as (Tu et al., 2025); and for Italian, we use the training subset from the Multilingual LibriSpeech (MLS) dataset (Pratap et al., 2020). We reserve a portion of each training set as a test set and randomly pair 400 utterances for evaluation. To measure content intelligibility, we use language-specific automatic speech recognition (ASR) tools: FunASR (Gao et al., 2023) with paraformer-zh (Gao et al., 2022) for Chinese, Chunkformer-large-vi (Le et al., 2025) for Vietnamese, and Whisper-large (Radford et al., 2023) for Italian. Figure 3 suggests that our phase 2 fine-tuning significantly improves the model's performance and allows it to adapt to any language using only audio, without requiring labeled data.

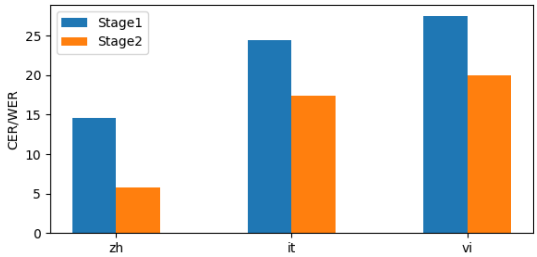


Figure 3: Performance of new language adaptation: CER for Chinese, WER for Vietnamese and Italian.

6 Conclusion

We presented a robust voice conversion framework based on synthetic data and a two-phase training strategy. Our method enhances speaker similarity, speech quality, and content consistency, particularly in zero-shot scenarios with unseen target speakers. Experiments and ablation studies confirm the effectiveness of our approach and demonstrate its ability to prevent speaker information leakage from the source audio. Additionally, we showed that the model generalizes well to unseen languages without requiring labeled data, making it highly suitable for low-resource settings.

Limitations

Although our model improves speaker similarity and content intelligibility, it still depends on access to a high-quality, labeled corpus to train the

TTS system. Furthermore, the effectiveness of synthetic data generation and its influence on the performance of voice conversion across different TTS systems remain insufficiently explored. Therefore, in future work, we plan to investigate alternative TTS models to gain a deeper understanding of their impact on overall system performance.

References

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice conversion with just nearest neighbors. In *Interspeech 2023*.

Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu, and Lirong Dai. 2014. Voice conversion using deep neural networks with layer-wise generative training. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22:1859–1872.

Nanxin Chen, Yu Zhang, Heiga Zen (Byungha Chun), Ron Weiss, Mohammad Norouzi, and William Chan. 2021a. Wavegrad: Estimating gradients for waveform generation. In *ICLR*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1–14.

Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee. 2021b. Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5954–5958.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2023. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. In *Interspeech 2023*, pages 2283–2287.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2024. 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*, volume 38, pages 17862–17870.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Zhifu Gao, Shiliang Zhang, Ian Mcloughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech 2022*, pages 2063–2067.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1.

Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022a. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. In *Proceedings of the IJCAI-22*.

Wang L. Yang J. et al. Huang, H. 2023. W2vc: Wavlm representation based one-shot voice conversion with gradient reversal distillation and ctc supervision. In *Journal AUDIO SPEECH MUSIC PROC*.

Wen-Chin Huang, Shu-Wen Yang, Tomoki Hayashi, Hung-Yi Lee, Shinji Watanabe, and Tomoki Toda. 2022b. S3prl-vc: Open-source voice conversion framework with self-supervised speech representations. In *Proc. ICASSP*.

Wen-Chin Huang, Shu-Wen Yang, Tomoki Hayashi, and Tomoki Toda. 2022c. A Comparative Study of Self-Supervised Speech Representation Based Voice Conversion. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1308–1318.

Shehzeen Hussain, Paarth Neekhara, Jocelyn Huang, Jason Li, and Boris Ginsburg. 2023. Ace-vc: Adaptive and controllable voice conversion using explicitly disentangled self-supervised speech representations. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Joonson Kim, Jungil Kong, and Jaehyeon Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5530–5540.

Ziyu Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, H. Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. *International conference on machine learning*, abs/1512.09300.

Khanh Le, Tuan Vu Ho, Dung Tran, and Duc Thanh Chau. 2025. Chunkformer: Masked chunking

conformer for long-form speech transcription. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jingyi Li, Weiping Tu, and Li Xiao. 2023. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yist Y. Lin, Chung-Ming Chien, Jheng-Hao Lin, Hung-yi Lee, and Lin-shan Lee. 2021. Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5939–5943.

Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng. 2021. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1717–1728.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. *Proc. ACL 2024 Findings*.

X. Mao, Q. Li, H. Xie, R. K. Lau, Z. Wang, and S. Smolley. 2017. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, Los Alamitos, CA, USA. IEEE Computer Society.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *Interspeech 2021*.

Paarth Neekhara, Shehzeen Hussain, Rafael Valle, Boris Ginsburg, Rishabh Ranjan, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. 2024. Selfvc: voice conversion with iterative refinement using self transformations. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Seung-won Park, Doo-young Kim, and Myun-chul Joe. 2020. Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data. In *Proc. Interspeech 2020*, pages 4696–4700.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Sergeevich Kudinov, and Jiansheng Wei. 2022. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. In *International Conference on Learning Representations*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23.

Siyuan Shan, Yang Li, Amartya Banerjee, and Junier Oliva. 2024. Phoneme hallucinator: One-shot voice conversion via set expansion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:14910–14918.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. Aishell-3: A multi-speaker mandarin tts corpus. In *Proceedings of Interspeech*, pages 2756–2760.

Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng. 2015. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4869–4873.

Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. 2016. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Xiaohai Tian, Junchao Wang, Haihua Xu, Eng Chng, and Haizhou Li. 2018. Average modeling approach to voice conversion with non-parallel data. In *Odyssey 2018 The Speaker and Language Recognition Workshop*, pages 227–232.

Huu Tuong Tu, Luong Thanh Long, Vu Huan, Nguyen Thi Phuong Thao, Nguyen Van Thang, Nguyen Tien Cuong, and Nguyen Thi Thu Trang. 2025. Voice conversion for low-resource languages via knowledge transfer and domain-adversarial training. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

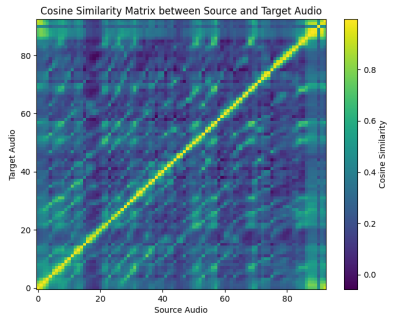
Da-Yi Wu and Hung-yi Lee. 2020. One-shot voice conversion by vector quantization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7734–7738.

Mingyang Zhang, Yi Zhou, Li Zhao, and Haizhou Li. 2021. Transfer learning from speech synthesis to voice conversion with non-parallel training data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1.

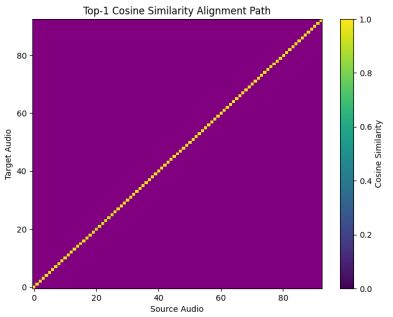
Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE.

A Appendix

A.1 Semantic Alignment of Synthetic Audio Pairs



(a) Cosine Pairwise Semantic Similarity



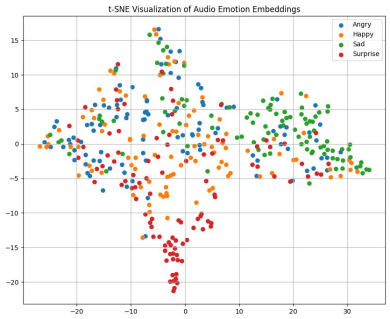
(b) Top-1 Cosine Similarity Alignment Path

Figure 4: Semantic Alignment of Source and Target Audio via Synthetic Data

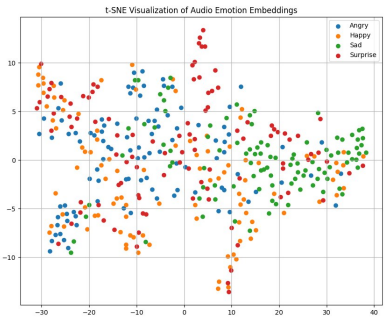
In this section, we evaluate the alignment between the source and target audio generated using a synthetic data strategy. To assess alignment quality, we extract semantic features with a pretrained HuBERT ASR model, as described in Section 4.3. We then compute the cosine similarity between all pairs of frames from the source and target audio, resulting in a pairwise similarity matrix. This matrix is visualized as a cosine similarity heatmap in Figure 4a.

The heatmap displays a clear diagonal of high similarity values, indicating strong frame-level alignment between the source and target audio. Furthermore, the top-1 cosine similarity alignment path, shown in Figure 4b, lies precisely along the diagonal, confirming perfect alignment. These results demonstrate that the synthetic data input-output pairs are ideal training examples for voice conversion, enabling the model to learn effective one-to-one mapping.

A.2 Emotional Information Preservation



(a) Our Proposal



(b) FreeVC

Figure 5: Emotion representation of converted audio using t-SNE

To evaluate how well emotional information is preserved, we compare our proposed model with the baseline FreeVC, which shares the same backbone. For this experiment, we use the ESD dataset (Zhou et al., 2021), which contains emotional speech. We randomly sample 10 audio clips from each of the 10 speakers across 4 emotions, resulting in a total of 400 source audio samples. For the target speakers, we randomly select a speaker from the

LibriSpeech test set. After performing voice conversion, we extract emotion embeddings from the converted audio using the emotion2vec_plus_large model (Ma et al., 2024) and visualize them using t-SNE. As shown in Figure 5, our proposed model produces more distinct emotion clusters, such as sad and surprise, while FreeVC exhibits little to no clustering, indicating that our model better preserves emotional characteristics.