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

Huu Tuong Tu[1, 2], Vu Huan[3], Ngo Dien Hy[1,3], Nguyen Tien Cuong[1], Nguyen Thi Thu Trang[2]
[1]VNPT AI, VNPT Group, Vietnam
[2]Hanoi University of Science and Technology, Vietnam
[3]Business AI Lab, National Economics Univesity, Vietnam

Introduction

Voice Conversion aims to transform a source speaker's voice to match a target speaker while **preserving the original linguistic content**. Recent VC methods typically rely on **disentangling** speech into separate speaker identity and content representations, which are then used to reconstruct the converted audio.

Approaches and Challenges

- Text-based method:
  - a content extractor trained with transcript labels. ([Liu, et al., INTERSPEECH 2018], [Hussain, et al., ICASSP 2023])
  - Phonetic Posteriorgram (PPG) features for content representation. ([Liu, et al., TASLP 2021])

→ **Misalignment** between speech and text in ASR, and **loss of speaker-independent information** (e.g., accent, emotion, etc.)

- Text-free method:
  - Adversarial Training ([Chou, et al., ICASSP 2018])
  - Normalization ([Chou, et al., INTERSPEECH 2019])
  - Vector quantization ([Wu, et al., ICASSP 2020])
  - Data augmentation, Bottleneck ([Li, et al., ICASSP 2023])

→ Difficult to completely remove speaker information from the source speech, leading to **speaker leakage** during inference.

- KNN Method ([Baas et al., INTERSPEECH 2023], [Shan et al., AAAI 2024])

→ Generated audio tends to be **oversmoothed**, reducing the naturalness and quality of the converted speech

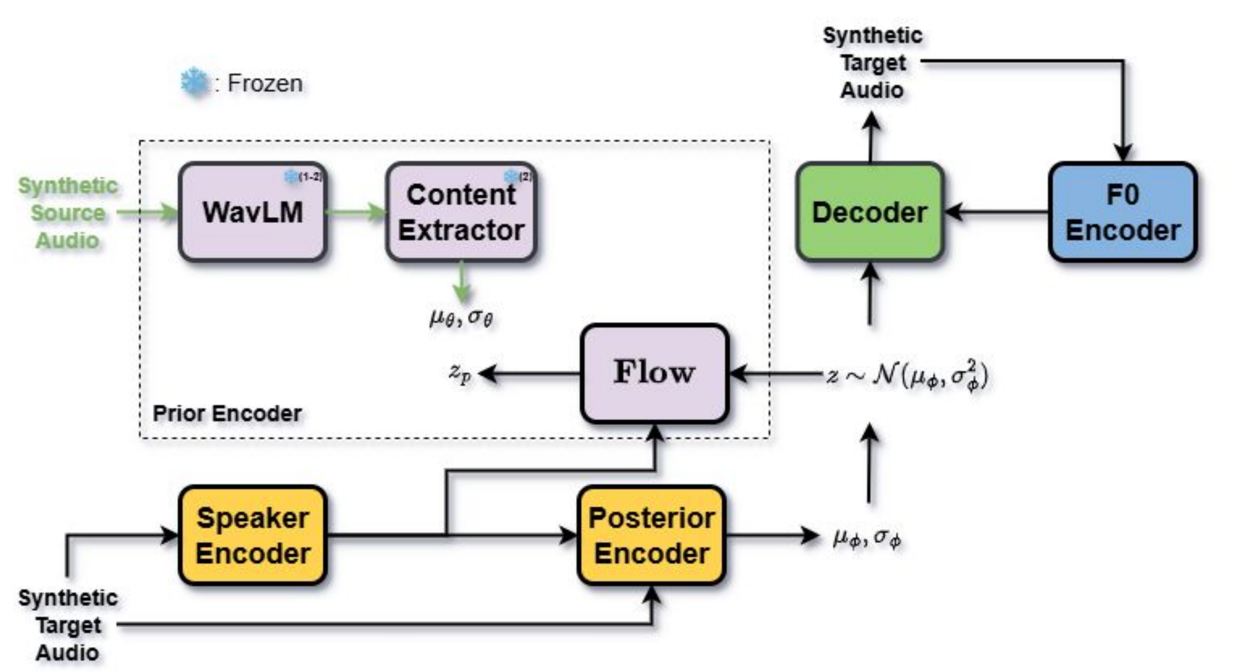
- Diffusion Method ([Choi et al., INTERSPEECH 2023], [Choi et al., AAAI 2024])

→ Diffusion models require significant **computation time**

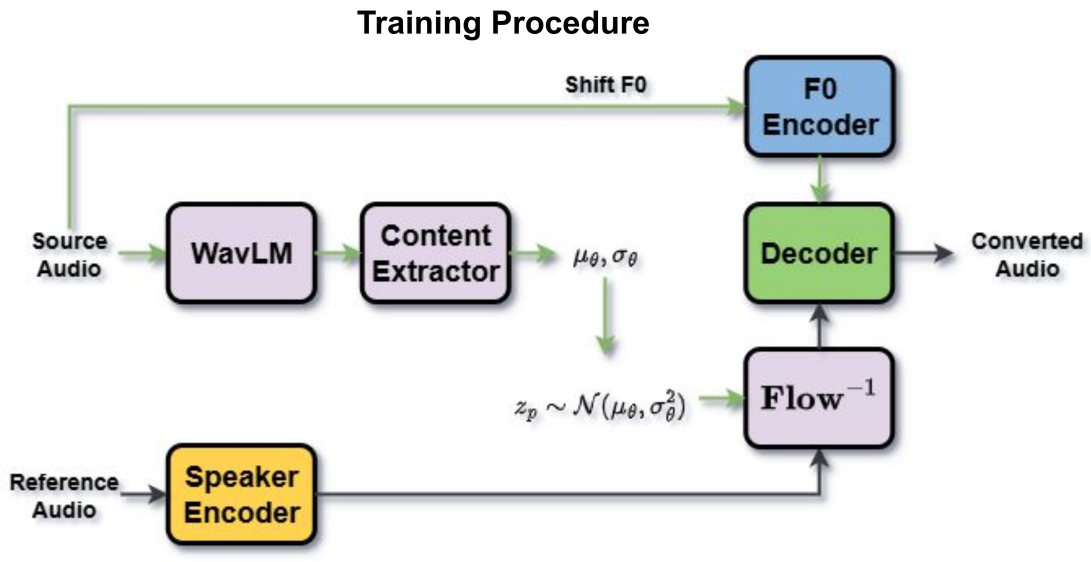
Proposed Method

A novel approach that **bypasses the need for feature disentanglement and audio reconstruction**:

- **Leveraging Synthetic Data:** We use a high-quality, pretrained multi-speaker **Text-to-Speech (TTS)** model to generate synthetic data.
- **Direct Mapping:** We train the VC model using synthetic data pairs that share the **same linguistic content but differ in speaker identity** as input-output of model.
- **Any-to-Any and Zero-Shot VC:** We introduce a flexible training strategy that promotes **generalization** to unseen speakers and new languages.



Training Procedure



Inference Procedure

The difference lies in speaker identity, which causes an **F0 mismatch** → We incorporate an **F0 encoder** to address this issue.

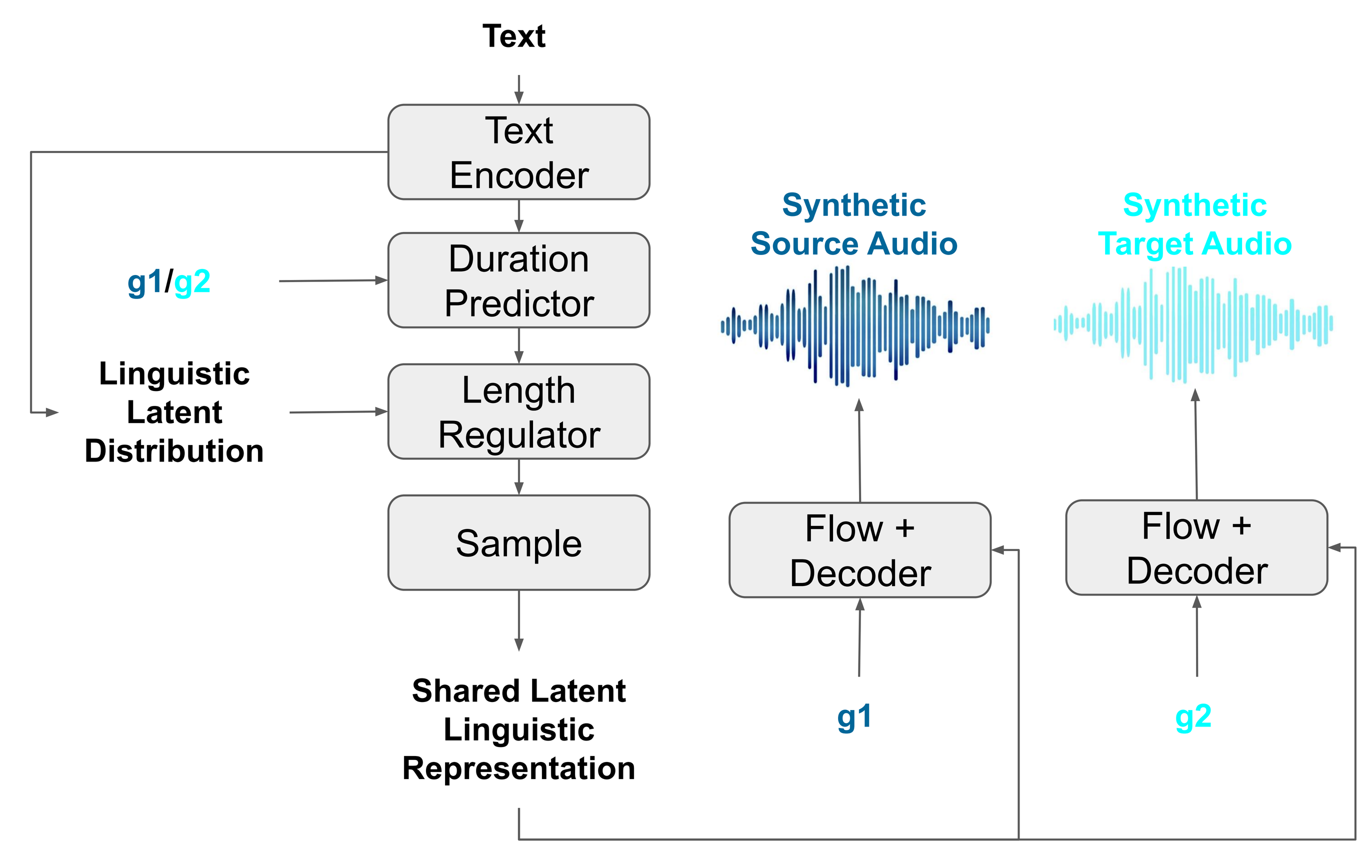
Limited generalization to unseen speakers and domains arises from the low speaker count and exclusive reliance on synthetic data → We adopt a Two-Phase Training Strategy:

1. Phase 1 (Synthetic Pretraining): Train with synthetic data to **enforce speaker-independent content representations** in the content extractor.
2. Phase 2 (Fine-tuning Adaptation): Fine-tune using a large-scale, real multi-speaker corpus with real audio only (reconstruction objective). **Freeze speaker-independent components** to **preserve content purity** while achieving **robust generalization** and **facilitating domain/data adaptation.**

Synthetic Data Generation

TTS System Selection Criteria:

1. **High Fidelity:** Produces natural, clear, and high-quality audio → Ensures the quality of the converted speech
2. **Shared Latent Space:** Synthesizes both source and target from same linguistic latent representation with identical duration → Ensures consistent phonetic alignment for one-to-one mapping



VITS Inference for Synthetic Data

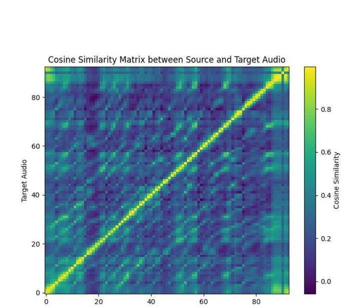
Experimental Setup

- Phase 1: Uses a pretrained VITS model on the VCTK dataset.
- Phase 2: Fine-tunes on the LibriSpeech train dataset for adaptation.
- Hyperparameters: Follow the configuration of the FreeVC model.
- Zero-shot evaluation: Performed on the LibriSpeech test set.
- Cross-lingual generalization: Tested on three languages to assess generalization ability.
- Ablation study: Conducted to analyze the impact of individual components.

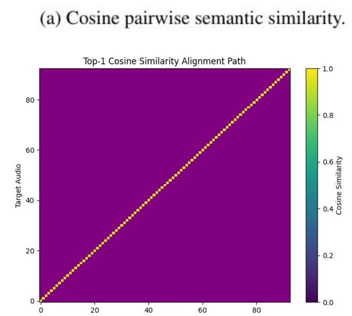
Results

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 the best performance, Underline indicates second best. Subjective evaluation results showing MOS and SMOS scores, along with 95% confidence intervals.



(a) Cosine pairwise semantic similarity.



(b) Top-1 cosine similarity alignment path.

Figure 5: Semantic alignment of source and target audio via synthetic data.

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.

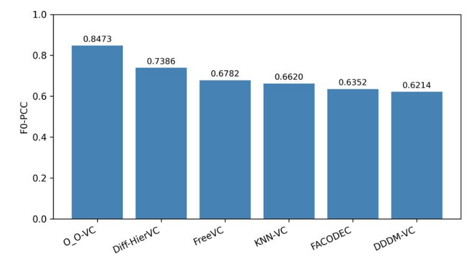


Figure 3: Comparison of systems on F0-PCC

Results

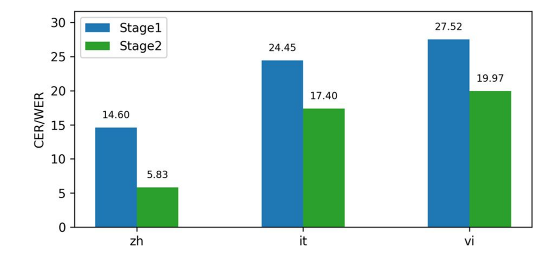


Figure 4: Performance of new language adaptation: CER for Chinese, WER for Vietnamese and Italian.

- Best in content intelligibility, balanced naturalness, and speaker similarity.
- F0 conditioning ensures superior pitch consistency.
- Synthetic data perform strong disentanglement between content information and speaker information.
- Perfect frame-level alignment in training data.
- Generalizes well to new domains.

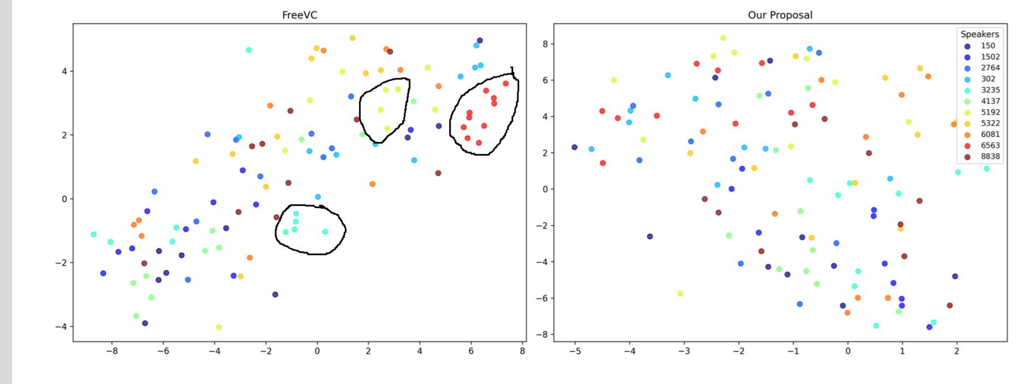


Figure 2: T-SNE visualization of speaker-independent features. More distributed points with no clusters indicate better speaker independence.

CONCLUSION.

We present a robust voice conversion framework using synthetic data and two-phase training. It improves speaker similarity, speech quality, and content consistency, especially in zero-shot settings. Experiments confirm its effectiveness in preventing speaker leakage and its strong generalization to unseen languages.

EMNLP 2025
Suzhou, China | 中国苏州