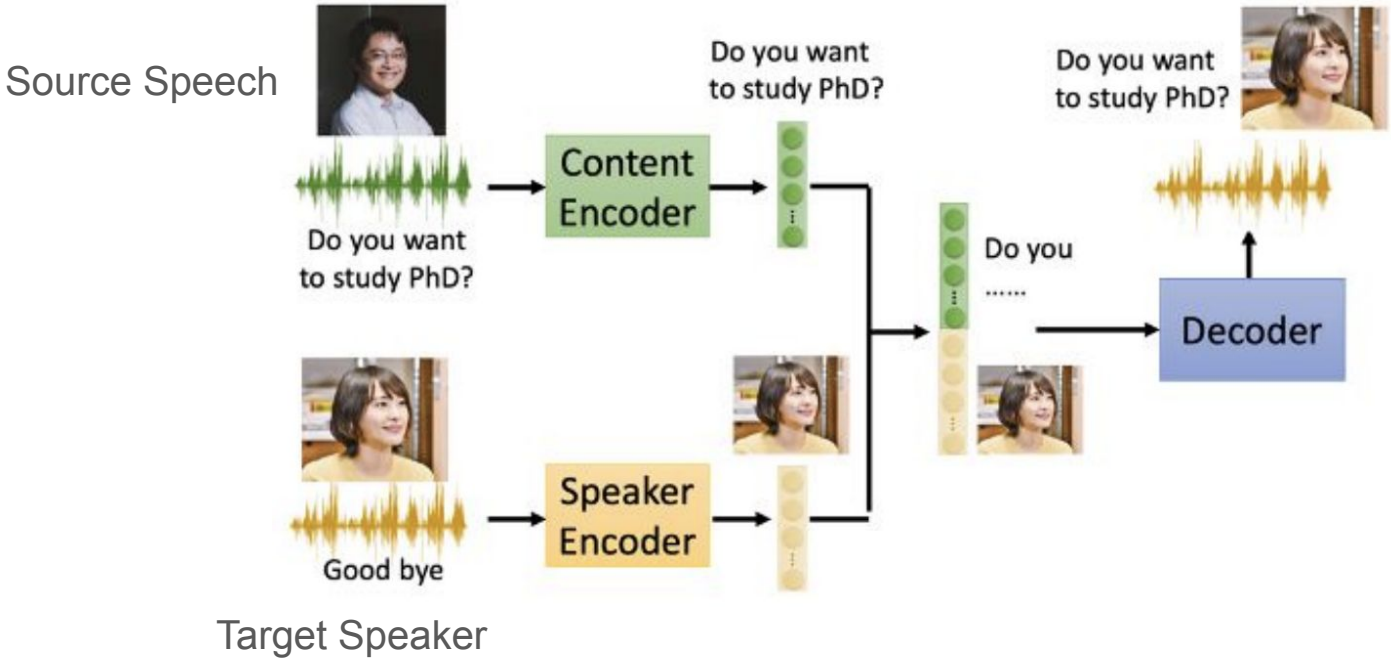
Streaming Voice Conversion Through Chunk-wise Training and Lookahead Loss

ECE 477 Course Project
Baotong Tian

Introduction

Voice Conversion: altering the style of a speech signal while preserving its linguistic content



Introduction

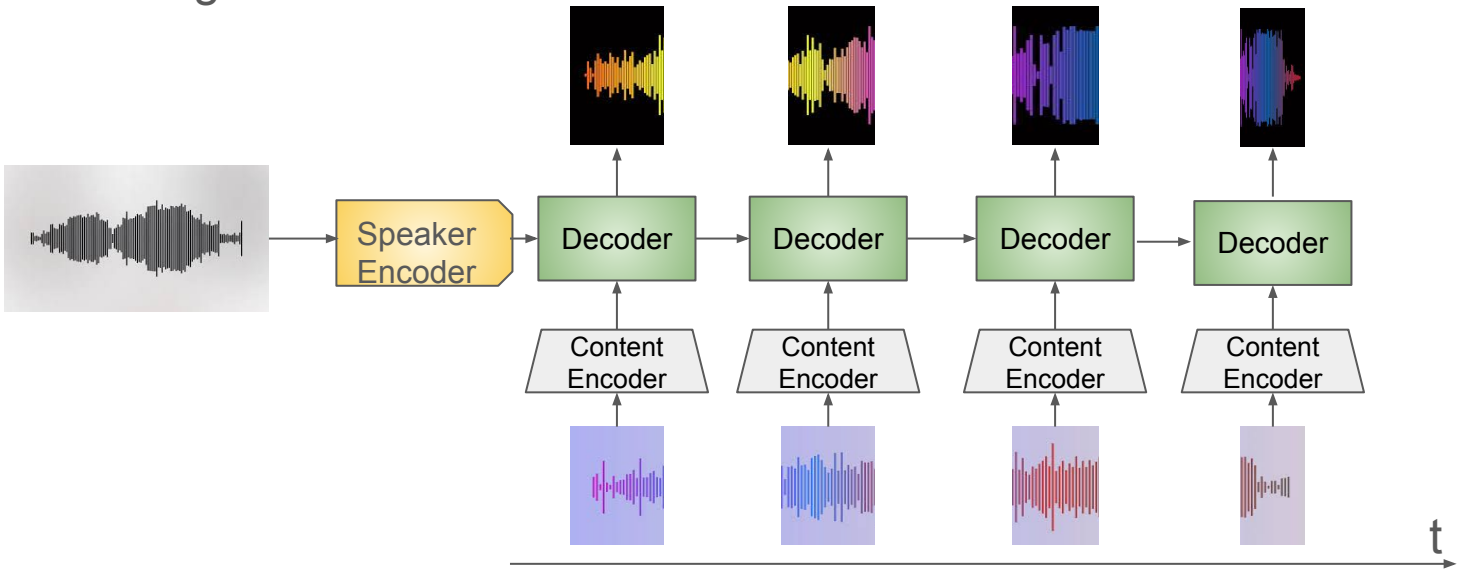
What Voice Conversion has achieved: high quality, naturalness

What current Voice Conversion systems lacked: feature disentanglement (speaker information leakage), controllability (e.g. prosody & speed), fast inference in real-time streaming scenarios (calls and video conferencing, voice anonymization)

Problem Formation

Any-to-any streaming voice conversion system

During inference, the speaker embedding is extracted from the whole utterance, the content is obtained chunk-wise. the model takes a series of chunks of input speech and generate converted chunks



Non-Streaming VS Streaming

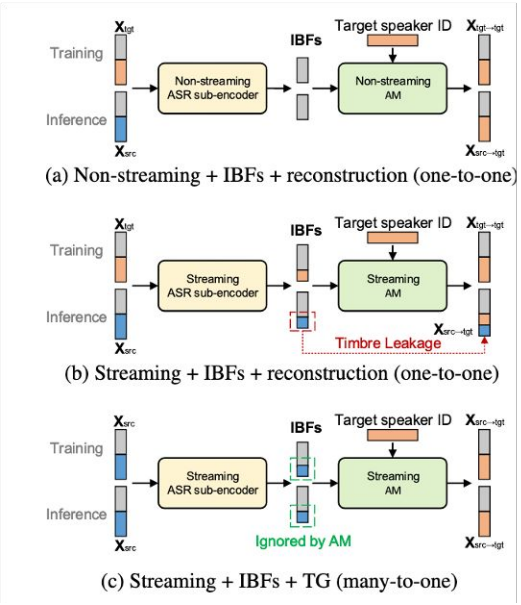
Directly applying non-streaming system to streaming scenarios:

- Noticeable artifacts
- Less Natural
- Lack of coherency among output chunks

Missing future information

Previous Works

Intermediate Bottleneck Features (IBF) instead of Phonetic Posteriorgrams (PPG) to get more low-level information & Non-streaming teacher guidance



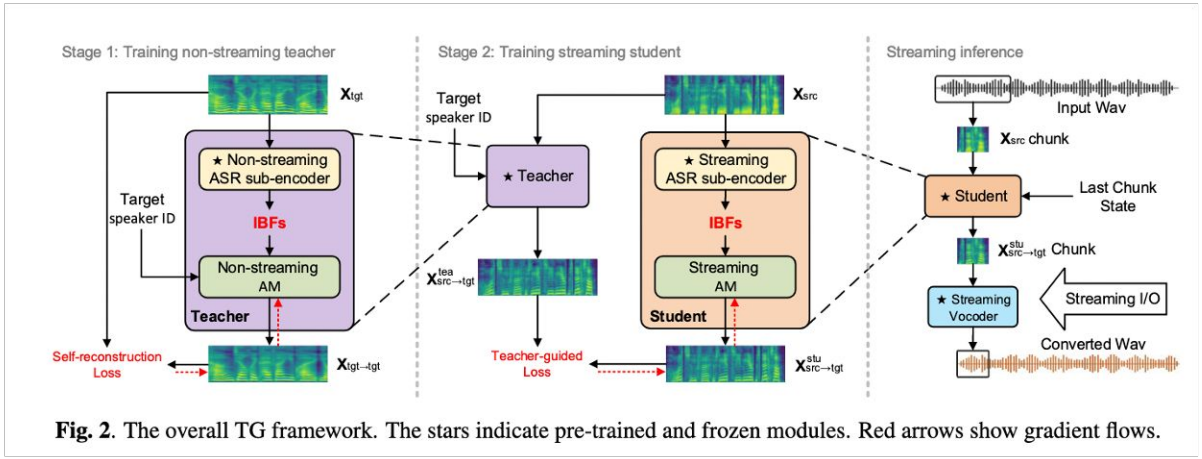


Fig. 2. The overall TG framework. The stars indicate pre-trained and frozen modules. Red arrows show gradient flows.

Chen, Yuanzhe, et al. "Streaming voice conversion via intermediate bottleneck features and non-streaming teacher guidance." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

Previous Works

Hybrid Predictive Coding (HPC) to capture common feature structure & Teacher Guidance (Dual Mode)

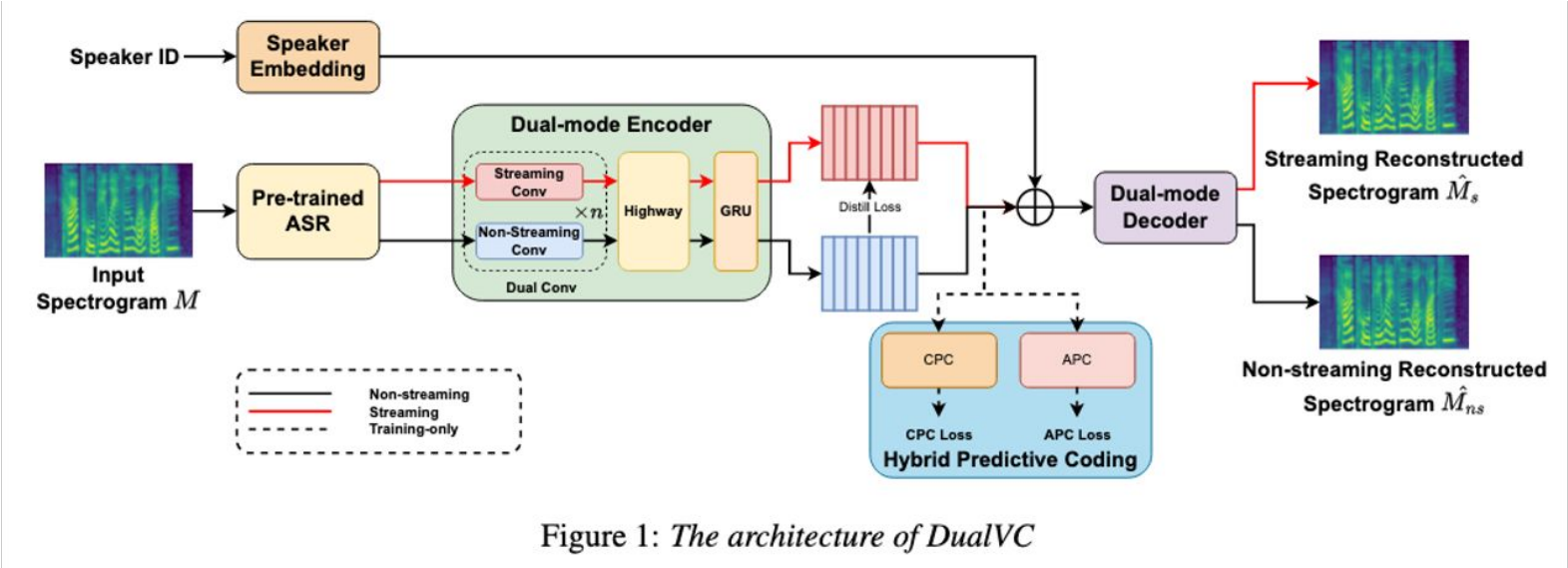
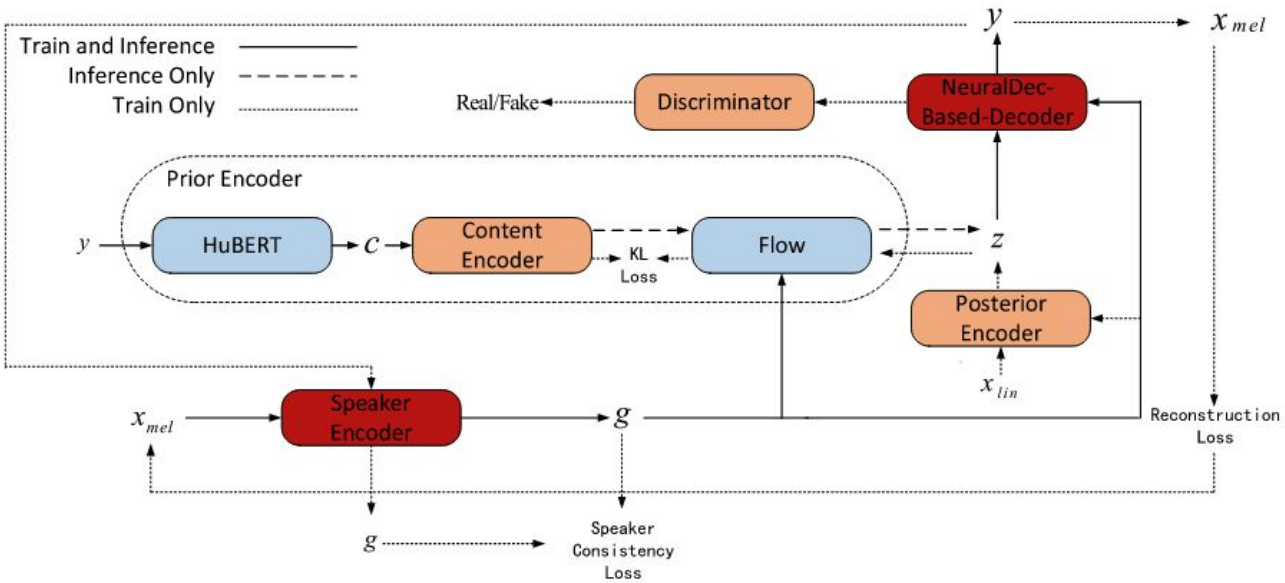


Figure 1: *The architecture of DualVC*

Ning, Ziqian, et al. "Dualvc: Dual-mode voice conversion using intra-model knowledge distillation and hybrid predictive coding." *arXiv preprint arXiv:2305.12425* (2023).

Framework

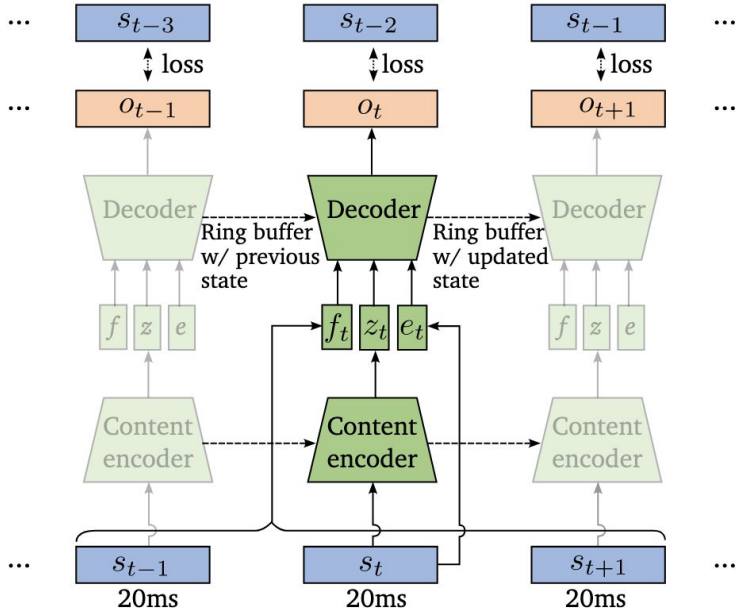
VITS-based Model + Light-weight Decoder → Fast Inference (but not for streaming scenarios)



Cao, Danyang, Zeyi Zhang, and Jinyuan Zhang. "NeuralVC: Any-to-Any Voice Conversion Using Neural Networks Decoder For Real-Time Voice Conversion." *IEEE Signal Processing Letters* (2024).

Lookahead Loss

Adopted From StreamVC



Yang, Yang, et al. "StreamVC: Real-Time Low-Latency Voice Conversion." *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

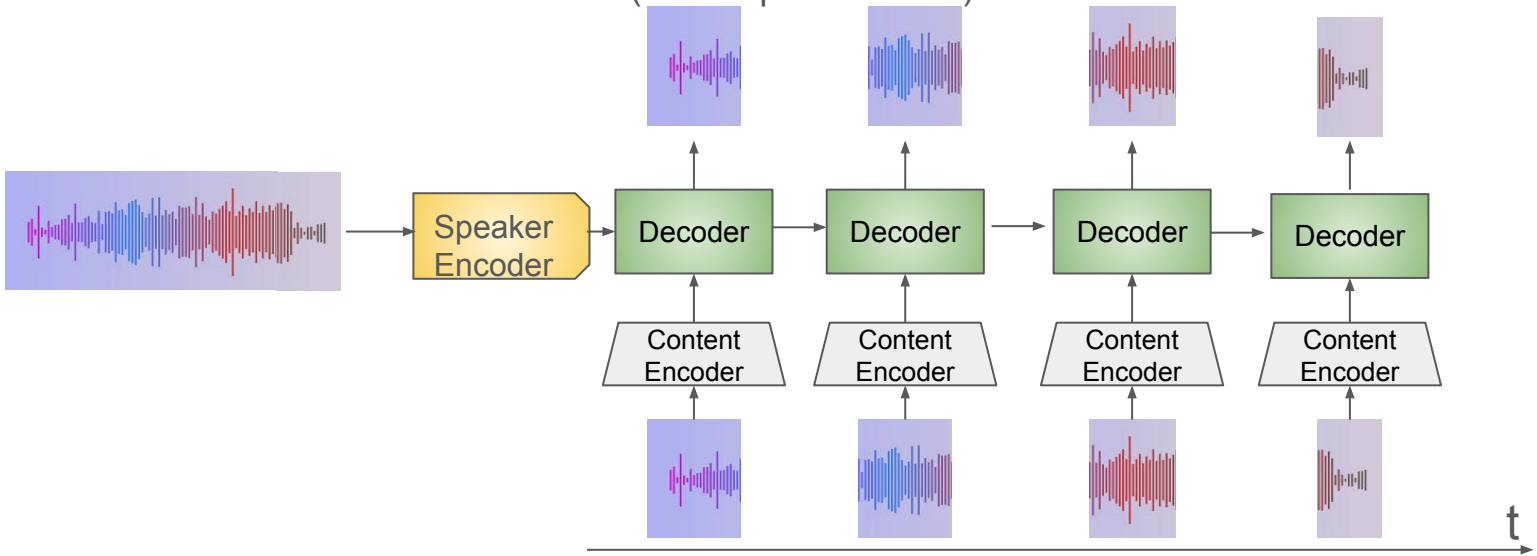
Training Settings

Chunk size = 3200 samples (200ms under 16k sample rate), 25% ratio of overlap between chunks

Non-causal Module (convolution) → Causal Module

VCTK Dataset (110 English speakers)

Reconstruction instead of conversion (lack of parallel data)



Results

Table 1. Non-streaming inference for different training settings.

Training Settings	Inference Time	Similarity	WER
Whole	0.1384	0.7839	0.0336
Chunk	0.1277	0.7502	0.133
Chunk+Lookahead	0.1342	0.7404	0.4344

Table 2. Streaming inference for different training settings.

Training Settings	RTF	Similarity	WER
Whole	0.4943	0.7371	0.6298
Chunk	0.3584	0.7644	0.2686
Chunk+Lookahead	0.3629	0.7253	0.6074

Table 3. Streaming performance with different inference settings.

Chunk Size, Overlap Ratio, Buffer	RTF	Similarity	WER
3200, 0, True	0.3644	0.7529	0.3719
3200, 0, False	0.178	0.7851	0.4269
3200, 25%, True	0.3584	0.7644	0.2686
3200, 25%, False	0.3492	0.7842	0.3553
3200, 12.5%, True	0.3282	0.7632	0.3072
320, 25%, True	2.706	0.5079	1.009

Results

Samples

A→B

A

Whole+Whole

Chunk+Whole

Ahead+Whole

B

Whole+Chunk

Chunk+Chunk

Ahead+Chunk

C

C→A

C→B

Conclusion & Future Work

- Whole utterance training - degrade performance in streaming scenarios → chunk-wise training

- Processing extremely short chunks (e.g., 20ms) is hard to maintain intelligibility and temporal consistency
- One model for variable-length chunk input

Thanks！