StreamVC: Real-Time Low-Latency Voice Conversion

Yang Yang, Yury Kartynnik, Yunpeng Li, Jiuqiang Tang, Xing Li, George Sung, Matthias Grundmann
{yanghm, kartynnik, yunpeng, jqtang, simplyxing, gsung, grundman}@google.com

Introduction

Research Goal:

- Lightweight streamable voice conversion solution.
- Real-time low-latency inference on mobile devices.
- Comparable quality with existing SOTA.

Key contributions:

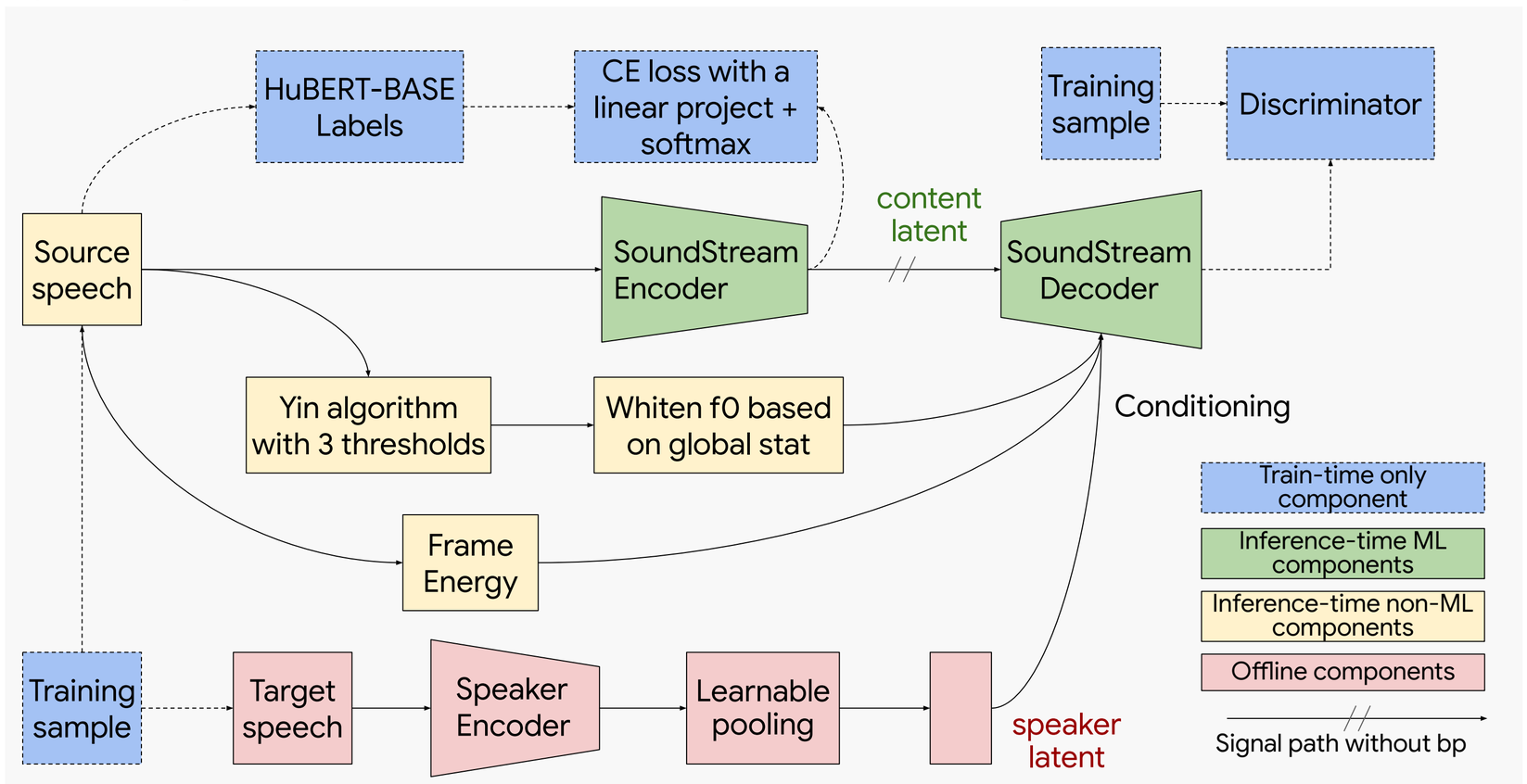
- Demonstrate the feasibility of learning soft speech units using a **causal** conv-net.
- Introduce the injection of whitened f0 contour as a way to improve pitch consistency.
- Achieve real-time inference on a mobile device.

Inspired by:

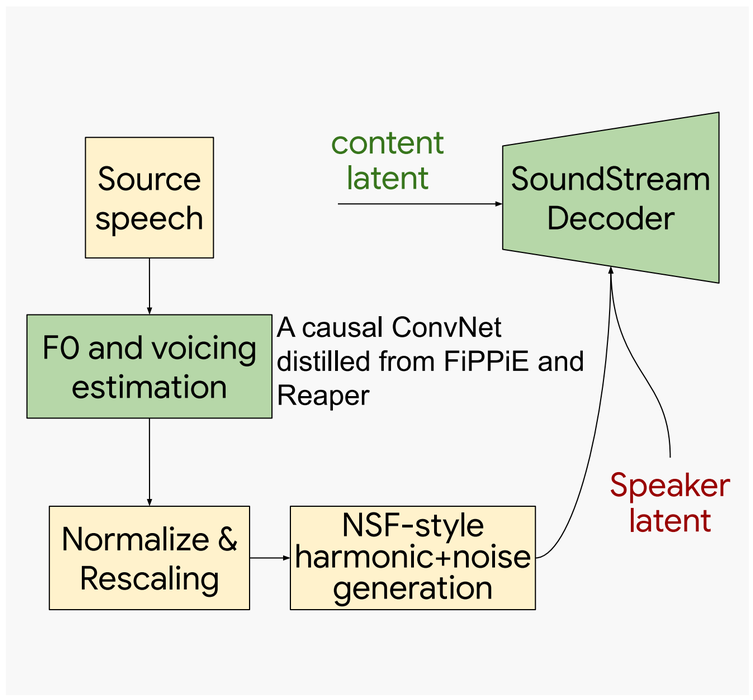
- Content disentanglement using **SoftVC.**
- High-quality causal speech synthesis using **SoundStream**.
- F0-rescaling and **NSF**-style harmonic-plus-noise conditioning used in **RVC.**

Method

System diagram:



Updates in StreamVC V2 with NSF-style f0 conditioning:



Results

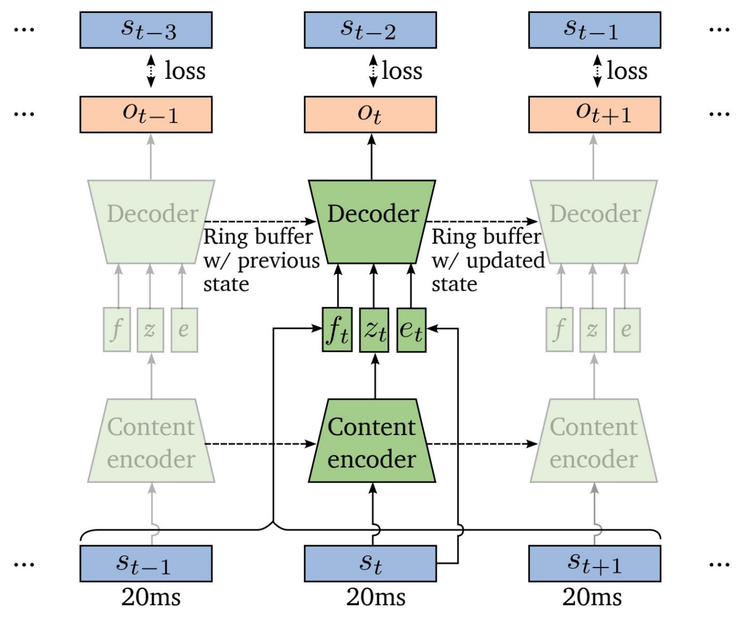
	Metrics	Naturalness (DNSMOS)			Intelligibility		Speaker similarity	f_0 consistency
		SIG	BAK	OVRL	WER	CER	Resemblyzer score	f_0 PCC
Baselines (VCTK)	VQMIVC [3]	3.36	3.29	3.86	57.88%	35.50%	65.72%	0.680
	BNE-PPG-VC [6]	3.54	3.42	3.89	11.43%	5.24%	80.17%	0.723
	Diff-VCTK [22]	3.57	**3.60**	**4.11**	11.64%	4.38%	**84.23%**	0.219
	QuickVC [9]	**3.61**	3.59	4.08	6.30%	3.03%	78.51%	0.758
Ours	StreamVC (LibriTTS)	3.57	3.53	4.07	**6.22%**	**2.17%**	77.81%	**0.842**
	+ fine-tuning on VCTK	3.56	3.48	4.02	6.54%	2.25%	80.34%	0.830
Ablations	$-$ f_0 whitening	3.48	3.40	4.00	6.12%	2.04%	75.59%	0.704
	$-$ f_0	3.56	3.55	4.06	6.97%	2.56%	77.46%	0.461
Oracles	Source	3.57	3.56	3.99	5.41%	2.43%	53.86%	1.000
	Target	-	-	-	-	-	82.42%	-

Evaluation setup

- StreamVC is trained on libriTTS and evaluated on VCTK.

Results highlights

- StreamVC delivers the best intelligibile score (WEB and CER) and f0 consistency score among baselines.
- Speaker similarity score can be improved by fine-tuning on VCTK.
- Effectiveness of injecting whitened f0 contour is verified in the ablation study.



Streaming inference

- Causal convnet converted to streamable tflite using google-research/kws_streaming.
- Inference is done once every 20ms.

Latency

- 3-frame/60ms lookahead.
- ~10ms inference latency on Pixel 7.
- A total of ~70ms end-to-end latency.

Conclusion

In this work, we propose a light-weight (~20M param.) causal voice conversion solution that can run in real-time with low latency on a commercially available mobile device. The key design elements are: (1) using a causal encoder to learn soft speech units; (2) injecting whitened f0 to improve pitch stability without leaking source speaker info.

In our later V2 version, we found that f0 rescaling followed by a NSF-style harmonic-plus-noise conditioning (as is done in RVC) results in better quality.

Live demo available upon request