05

Real-time voice conversion with high quality and low latency

Changing your voice and speaking style in real-time

Abstract

Conventional voice conversion technology waits for the end of speech before converting it, making it difficult to convert speech while it is being spoken. We present a technique for real-time conversion with high fidelity and low latency, using a deep generative model that accounts for inter-speaker differences in intonation and voice timbre. By ensuring robustness to inter-speaker variation, it is possible to learn speech representations that are less speaker-dependent, resulting in high-quality real-time conversion that does not wait for the end of speech. Additionally, by improving the waveform synthesizer, we have achieved even higher speed and lower power consumption simultaneously. The system can be applied to various forms of voice communication, both face-to-face and remote. For example, it can be used in live streaming and call centers to conceal the speaker's identity. In the future, we aim to incorporate text-to-speech and speech recognition technologies to enable smoother communication.

Communication augmentation system

Using voice conversion (VC) technique to help overcome barriers that prevent us from smooth communication

Ex) Conversion to an easy-to-hear voice

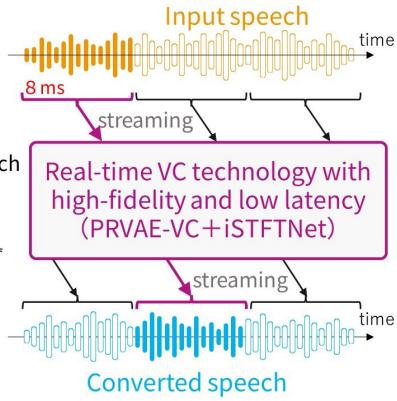
Real-time VC

Necessary condition

Converts speech segments every 8 milliseconds into segments of converted speech within 8 milliseconds

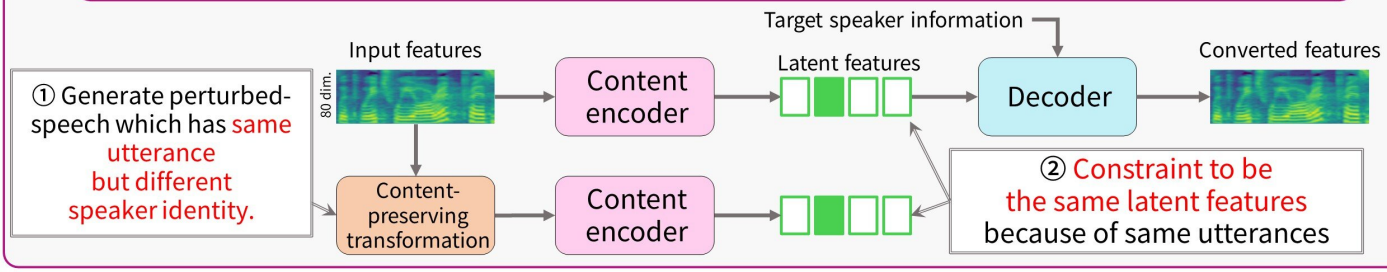
Sufficient condition

high-fidelity and low latency

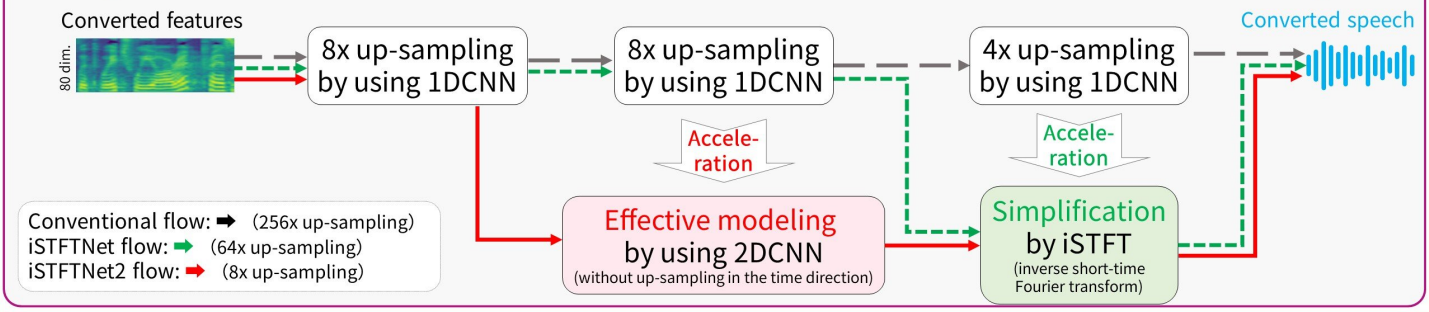


Core techniques

PRVAE-VC: High-fidelity conversion achieved by speech representation with low speaker dependence



iSTFTNet: Hybrid processing of DNN up-sampling and iSTFT for low-latency waveform synthesis.



References

[1] K. Tanaka, H. Kameoka, T. Kaneko, “PRVAE-VC: Non-Parallel Many- to-Many Voice Conversion with Perturbation-Resistant Variational Autoencoder,” in *Proc. the 12th Speech Synthesis Workshop (SSW)*, 2023.

[2] T. Kaneko, H. Kameoka, K. Tanaka, S. Seki, “iSTFTNet2: Faster and More Lightweight iSTFT-Based Neural Vocoder Using 1D-2D CNN,” in *Proc. INTERSPEECH*, 2023.

Contact

Kou Tanaka, Computational Modeling Research Group, Media Information Laboratory