End-to-End Neural Speech Coding for Real-Time Communications

Paper ID #5220

Xue Jiang[1], Xiulian Peng[2], Chengyu Zheng[1], Huaying Xue[2], Yuan Zhang[1] ,Yan Lu[2]

[1]Communication University of China, [2]Microsoft Research Asia

Introduction

Background

- Existing neural speech codecs mostly still target at coding efficiency with little attention on low latency and error resilience for real-time communications (RTC).

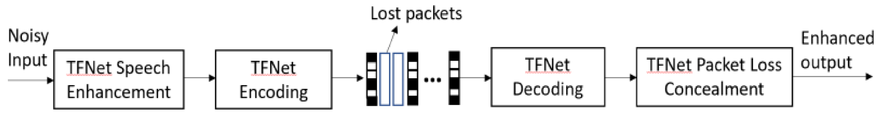
Highlights

- TFNet: a low-latency (20ms), error-resilient neural speech codec for RTC
- One-for-all: joint optimization with speech enhancement and packet loss concealment

Joint optimization

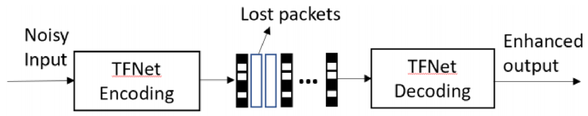
Cascaded network

- Pre-processing enhancer – codec – post-processing PLC
- Two-stage training



All-in-one network

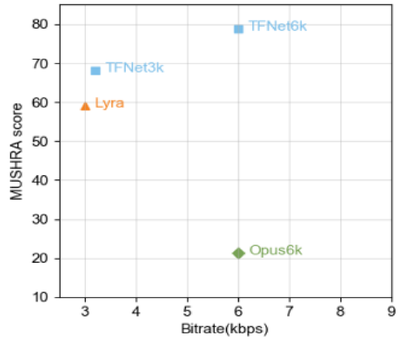
- A single TFNet codec model for all tasks



Experiments

Comparison with other codecs

- TFNet vs. Lyra @3kbps
- TFNet vs. Opus @6kbps



Comparison of different temporal filtering modules

	PESQ	STOI
TCM	2.447	0.869
GRU	2.360	0.863
Interleave-TCM-G-GRU	**2.501**	**0.870**

- Interleaved structure efficiently captures both short-term and long-term temporal correlations

Evaluation on joint optimizations

- Comparison of coding efficiency under background noises

	PESQ	STOI	DNSMOS
Baseline	1.740	0.811	3.29
Cascaded	1.777	**0.822**	**3.51**
All-in-one	**1.794**	0.821	3.48

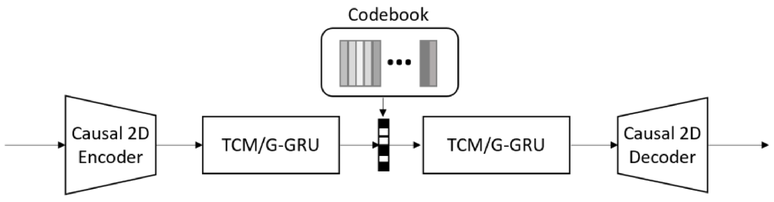
- Comparison of coding efficiency under background noises and packet losses

	PESQ	STOI	DNSMOS
Baseline	1.413	0.747	3.04
Cascaded	**1.545**	**0.778**	**3.38**
All-in-one	1.510	0.770	3.33

- All-in-one is on par with cascaded
- TFNet is efficient for both separation(speech enhancement) and restoration(packet loss concealment and coding)

Proposed TFNet Codec

Encoder-temporal filtering-decoder paradigm

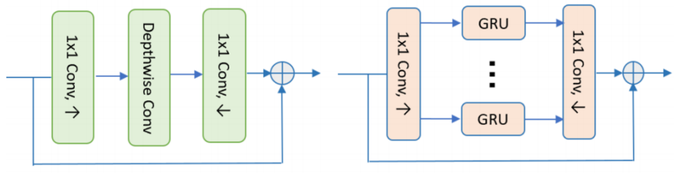


- Causal encoder and decoder
  - Local frequency and temporal correlations
- Causal temporal filtering
  - Long-term temporal dependencies
- Vector quantization
  - Group-wise quantization

Loss function

- $\mathcal{L} = \mathcal{L}_{recon} + \alpha\mathcal{L}_{VQ}$,
- $\mathcal{L}_{recon}$ is the MSE loss on the power-law compressed STFT spectrum. $\mathcal{L}_{VQ}$ is the commitment loss.
- STFT consistency

Temporal filtering



(a) TCM (b) G-GRU

- TCM
  - Explore short-term and middle-term temporal evolutions
- G-GRU
  - Capture long-term dependencies with frequency-aware temporal filtering
- Interleaved structure
  - Combine TCM and G-GRU in an interleaved way for joint short-term and long-term temporal correlation exploitation at different depths