Improving X-Codec-2.0 for Multi-Lingual Speech: 25 Hz Latent Rate and 24 kHz Sampling

Husein Zolkepli*

Abstract

X-Codec-2.0 has shown strong performance in neural audio compression and multilingual speech modeling, operating at a 50 Hz latent rate and a 16 kHz sampling rate using frozen HuBERT features. While effective, this configuration limits efficiency and audio fidelity. In this work, we explore a simple yet effective modification: introducing additional pooling and increasing the decoder hop size. This reduces the latent rate from 50 Hz to 25 Hz and simultaneously raises the output sampling rate from 16 kHz to 24 kHz, improving temporal efficiency and perceptual quality without altering the core architecture. Evaluated on the multilingual Common Voice 17 test set, the proposed configuration achieves a 0.5 MOS improvement over the original X-Codec-2.0 baseline. The source code, checkpoints and generation comparison released at xcodec2-25tps.github.io.

1 Introduction

.

2 Method

.

3 Experiments

.

4 Results

.

5 Acknowledgement

Special thanks to my wife for her patience and for not getting too upset about the electricity bill caused by running two RTX 3090 Ti GPUs around the clock for two months.

References

*husein.zol05@gmail.com

Technical Report.