5pSC28

Speech Synthesis with Perceptual Rating-Guided Parallel Iterative Decoding

Kazuki Yamauchi, Yuki Saito, Hiroshi Saruwatari
The University of Tokyo, Japan

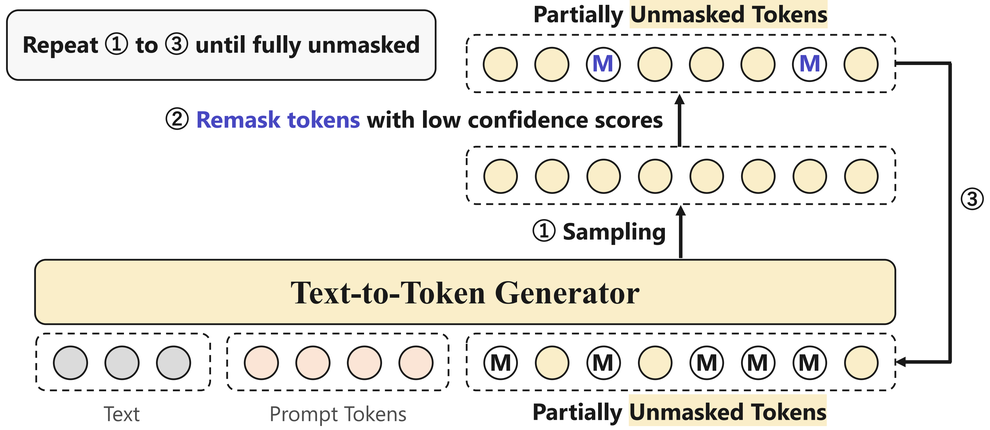
MOONSHOT RESEARCH & DEVELOPMENT PROGRAM
AVATAR SYMBIOTIC SOCIETY

Overview

- **Purpose**:
  - Explore ***inference-time optimization*** methods leveraging speech perceptual quality ratings for text-to-speech (TTS)
  - Focus on TTS model based on ***parallel iterative decoding***
- **Proposal**: **Perceptual Rating-Guided Parallel Iterative Decoding**
  - Introduce naturalness and speaker similarity **guidance** to parallel iterative decoding
  - **Improve zero-shot TTS performance**

Background

- **Masked Generative Codec Transformer (MaskGCT)** [Y. Wang+24]
  - Zero-shot TTS model based on **parallel iterative decoding**
  - TTS pipeline: **Text & Speech Prompt** → **Speech Tokens** → **Waveform**



- **Advantages**:
  - **High prosodic diversity** due to gradual sampling of tokens
  - **High controllability over duration** than autoregressive TTS model
- **Challenges**:
  - Selecting tokens to unmask based solely on the token's confidence score (= probability) **does not necessarily result in perceptually optimal speech quality**

Exploring *inference-time optimization* methods to optimize perceptual speech quality ratings during inference

Proposed Method

- **Proposed method**: **Perceptual Rating Guidance**
  - Multiple candidate tokens are sampled and evaluated, and the most perceptually promising candidate is selected
- **Explore three variants**:
  - **(1) Guided Decoding: Iterative selection at each decoding step**
  - **(2) Best-of-K (BOK): One-shot selection after the whole decoding**
  - **(3) Hybrid approach: Combining Best-of-K & Guided Decoding**
    - ➢ Generate K speech samples using Guided Decoding → Best-of-K selection

Overview of (1) Guided Decoding

- **Perceptual ratings**:
  - Naturalness: Predicted mean opinion score (MOS) by **UTMOS** [T. Saeki+22]
  - Speaker similarity (**SpkSim**): Cosine similarity between speaker embeddings of prompt and synthesized speech
    - ➢ Speaker embeddings are taken from a pre-trained ECAPA-TDNN [B. Desplanques+20]

Experiments

- **Experimental settings**:
  - **Backbone zero-shot TTS model**:
    - ➢ Pre-trained MaskGCT [Y. Wang+24]
  - **Dataset for evaluation**:
    - ➢ SeedTTS *test-en* dataset [P. Anastassiou+24]
      - Approximately 500 speakers × 2 samples from Common Voice Dataset [R. Ardila+19]
  - **Evaluation metrics**:
    - ➢ **Naturalness MOS (N-MOS)**
      - **1 (very unnatural)** to **5 (very natural)**
    - ➢ **Speaker similarity MOS (S-MOS)**
      - **1 (not at all similar)** to **5 (very similar)**
      - Similarity between prompt and synthesized speech
- **Results of subjective evaluation**:
  - Our method, especially based on UTMOS, significantly **improves naturalness and speaker similarity** compared to the original
  - Combining Best-of-K and Guided Decoding **improved the scores**

Method	N-MOS (↑)	S-MOS (↑)
Ground truth	4.00 ± 0.07	3.86 ± 0.09
MaskGCT (Original)	2.63 ± 0.08	2.42 ± 0.08
MaskGCT w/ BOK-UTMOS ($K = 16$)	2.89 ± 0.08	2.51 ± 0.08
MaskGCT w/ BOK-SpkSim ($K = 16$)	2.68 ± 0.08	2.42 ± 0.09
MaskGCT w/ Guide-UTMOS ($N = 16$)	2.82 ± 0.08	2.44 ± 0.09
MaskGCT w/ Guide-SpkSim ($N = 16$)	2.80 ± 0.08	2.43 ± 0.08
MaskGCT w/ BOK & Guide-UTMOS ($K = 4, N = 4$)	2.93 ± 0.08	2.56 ± 0.09
MaskGCT w/ BOK & Guide-SpkSim ($K = 4, N = 4$)	2.79 ± 0.08	2.51 ± 0.09

250 native English speakers each evaluated 24 samples.
*-UTMOS and *-SpkSim denote BOK or Guided Decoding based on UTMOS and SpkSim.

Conclusion & Future Work

- **Conclusion**:
  - Combining **Best-of-K** and **Guided Decoding** based on perceptual ratings **improved zero-shot TTS performance**
- **Future work**:
  - Extend perceptual ratings to various ratings, such as NISQA [G. Mittag+21], and **multi-objective ratings**

Acknowledgements: This work was supported by JST, Moonshot R&D Grant Number JPMJPS2011, JST, ACT-X, JPMJAX23CB, and JST, BOOST, JPMJBS2418.