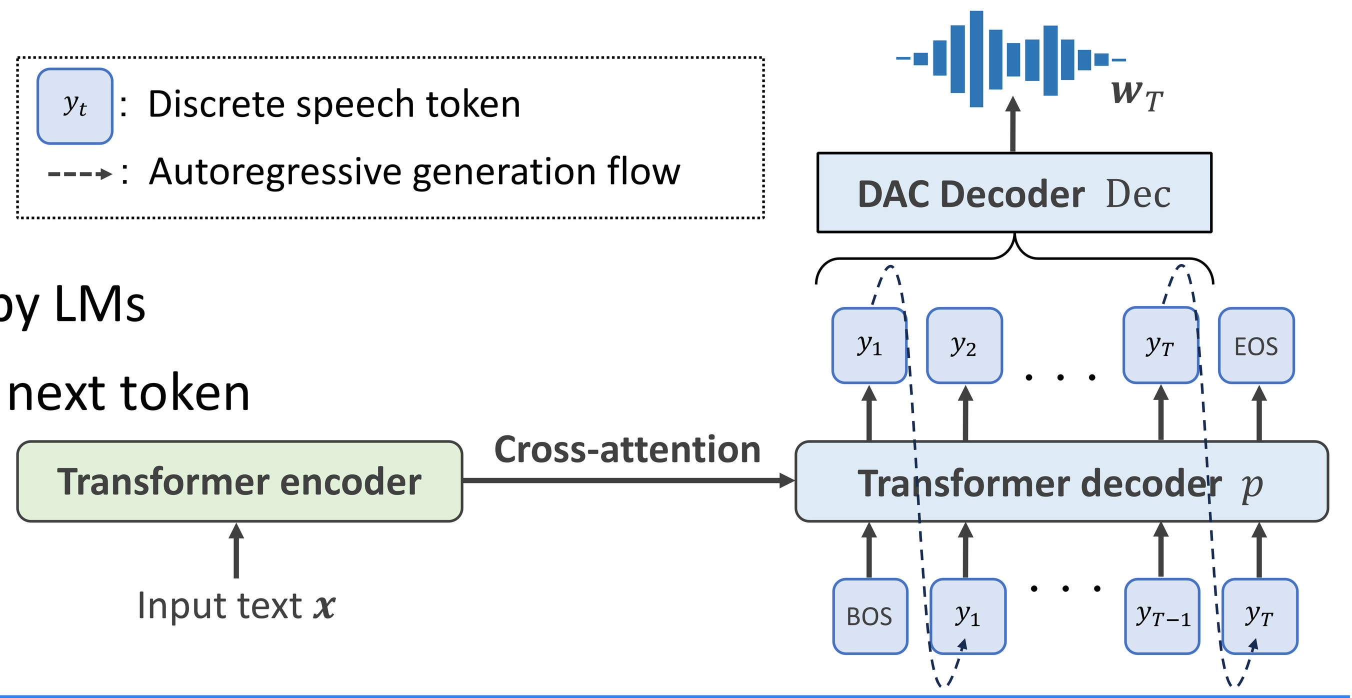
Decoding Strategy with Perceptual Rating Prediction for Language Model-Based Text-to-Speech Synthesis

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

Overview: Exploring decoding strategies for language model (LM)-based text-to-speech (TTS)

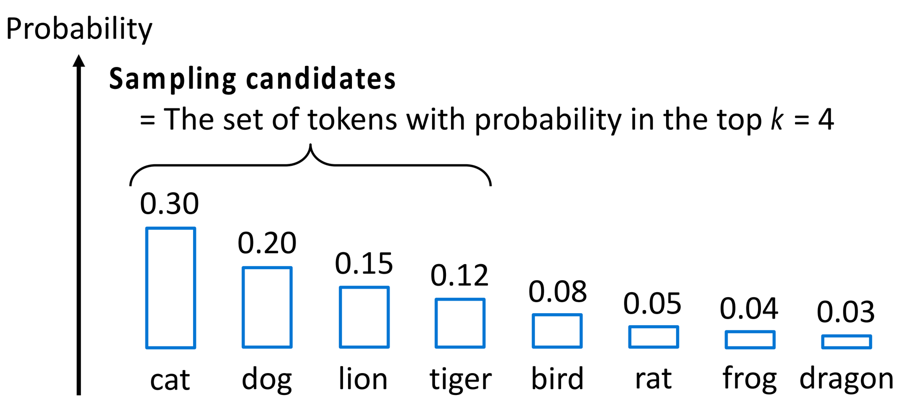
- **Background**: LM-based TTS model has recently attracted much attention
  - LM autoregressively generates discrete speech tokens such neural audio codec [1]
- **Question**: Which is the **optimal decoding strategy** for LM-based TTS?
  - Decoding: Process of selecting output tokens based on the probability distribution computed by LMs
  - ex. **Greedy decoding**: ***Deterministically*** selecting the token with the highest probability as the next token
    - Lead to **repetitive generation**, causing the output to get stuck in loops of repeating the same tokens
- **Proposal**: **BOK-PRP**, a novel sampling-based strategy for LM-based TTS
  - Incorporate **best-of-K (BOK) selection** based on **perceptual rating prediction (PRP)**



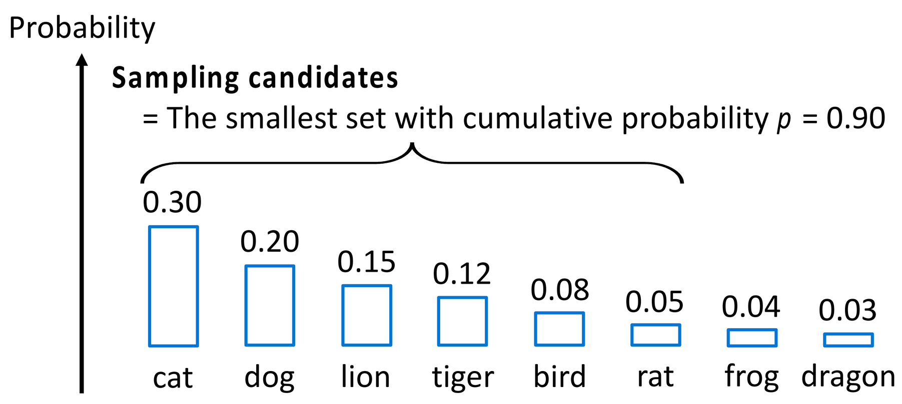
Conventional decoding strategies

Top-k sampling [2] / Top-p sampling [3]

- ***Stochastically*** select tokens based on the distribution of tokens
  - Introduce diversity and effectively address repetitive generation



top-k sampling (k = 4)



top-p sampling (p = 0.90)

Challenges of sampling-based decoding strategies

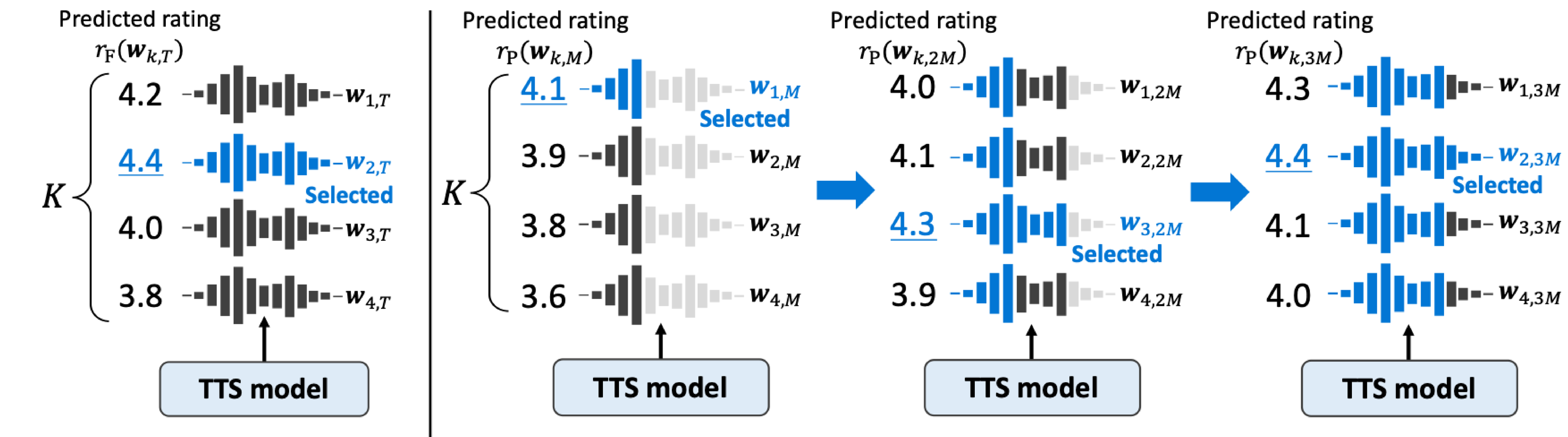
- **Challenge**: Sampling randomness **destabilizes generation**
  - Sampling randomness can lead to undesirable output, such as artifact
  - To alleviate this, top-k / top-p sampling **limit candidate tokens**
  - However, narrowing down candidates reduces output diversity and can lead to repetitive generation issues

➡ **Filtering out undesirable outputs while maintaining diversity remains challenging!**

Proposed method: Best-Of-K selection based on Perceptual Rating Prediction (BOK-PRP)

Sequence-wise BOK-PRP / Block-wise BOK-PRP

- The sample with the highest rating is selected from the K samples from an LM-based TTS

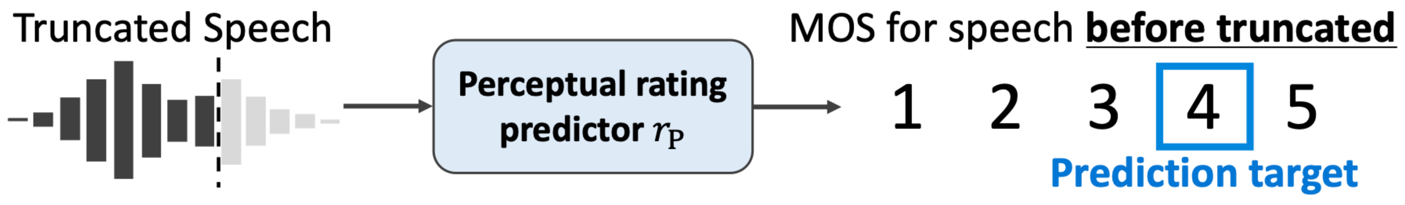


(a) Sequence-wise BOK-PRP (b) Block-wise BOK-PRP

Perceptual rating predictor

- Rating: Naturalness **M**ean **O**pinion **S**core (MOS)
  - Perceptual rating predictor: **UTMOS** [4]
  - Note: Rating predictor for block-wise takes as input a waveform synthesized from **partially** decoded tokens

→ Introduce a **truncation process** into the training process



Training process of the rating predictor for block-wise BOK-PRP

Experimental evaluation: Is BOK-PRP effective in improving the naturalness of synthetic speech?

Experimental conditions

- LM-based TTS model:
  - Transformer encoder-decoder TTS [5]
- Discrete speech tokenizer:
  - Descript Audio Codec (DAC) [6]
- Dataset: LJSpeech [7] (24 hours)
  - Speech from a single English speaker
- Compared methods:
  - Greedy decoding
  - Naive sampling
  - Top-k top-p sampling
  - Sequence-wise BOK-PRP (**proposed**)
  - Block-wise BOK-PRP (**proposed**)

Result of subjective evaluation

- Results of 5-point naturalness MOS:
  - Sampling-based strategies > Greedy decoding
  - Proposed strategy > Top-k top-p sampling

BOK-PRP improves subjective naturalness

Method	MOS (↑)	UTMOS (↑)
Greedy decoding	3.35 ± 0.09	4.27
Naive sampling	3.57 ± 0.08	4.31
Top-k top-p sampling	3.62 ± 0.08	4.36
Sequence-wise BOK-PRP	3.71 ± 0.07	4.46
Block-wise BOK-PRP	3.73 ± 0.07	4.43
Ground truth	3.92 ± 0.07	4.43

200 native English speakers each evaluated 24 samples

Ablation study on K

- Relationship between K and MOS:
  - Increasing K from 2 to 8 improves MOS
  - Increasing K from 8 to 32 degrades MOS

→ **Selecting tokens based on higher UTMOS does not always improve the actual MOS**

Excessively large K degrades naturalness

K	MOS (↑)	UTMOS (↑)
2	3.72 ± 0.08	4.40
4	3.74 ± 0.08	4.43
8	3.83 ± 0.07	4.43
16	3.79 ± 0.07	4.45
32	3.65 ± 0.08	4.46

Future direction

- Extend BOK-PRP to perceptual rating predictions from various perspectives, such as **emotional suitability**, beyond naturalness

References

[1] N. Zeghidour et al., *IEEE/ACM TASLP*, 2021. [2] A. Fan et al., in *Proc. ACL*, 2018. [3] A. Holtzman et al., in *Proc. ICLR*, 2020. [4] T. Saeki et al., *in Proc. INTERSPEECH*, 2022. [5] W. Nakata et al., arXiv:2403.13720, 2024. [6] R. Kumar et al., in *Proc. NIPS*, 2023. [7] K. Ito et al., https: //keithito.com/LJ-Speech-Dataset/, 2017.

