(14)[一般発表] 離散音声トークン生成によるテキスト音声合成のための音声主観評価値予測に基づく decoding 戦略

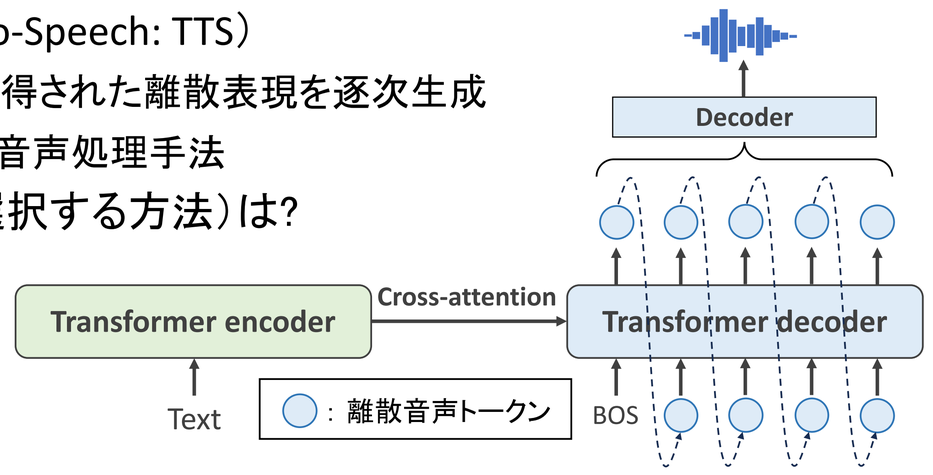
◎山内 一輝, 中田 亘, 齋藤 佑樹, 猿渡 洋（東京大学）

音声サンプルはこちら！

概要：離散音声トークン生成に基づく TTS モデルおける decoding 戦略の探求

- **背景**：離散音声トークンを中間表現として用いる自己回帰型テキスト音声合成（Text-to-Speech: TTS）
  - 従来のメルスペクトログラムに代わり，**Neural Audio Codec (NAC)** [1] などのデータ駆動で獲得された離散表現を逐次生成
  - 自然言語処理（特にテキスト生成）分野で研究されてきた離散トークン処理手法に類似した音声処理手法
- **問い**：最適な decoding 戦略（= 計算された出力確率から実際に出力するトークンを選択する方法）は?
  - 例：greedy search は同じトークン列の繰り返し生成問題を引き起こす
- **提案**：音声に対する**主観評価値の自動予測を活用**して最適な出力トークンを選択

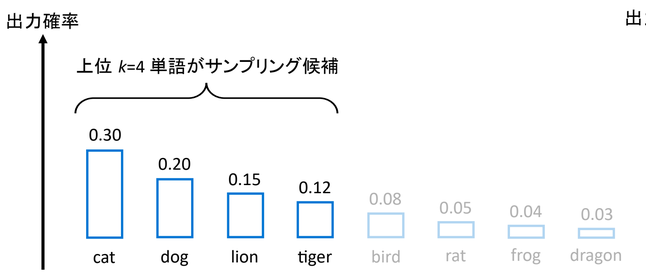
TTS モデルの追加学習なしで合成音声の自然性を向上！



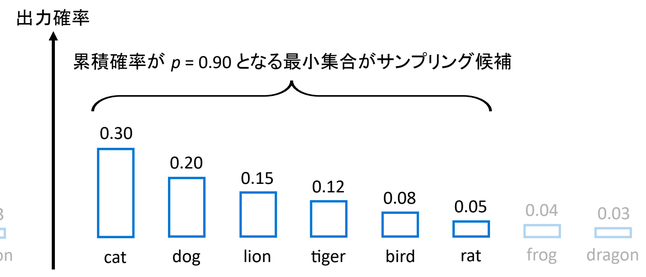
関連研究：テキスト生成における decoding 戦略

top-k / top-p サンプリング

- **サンプリング**により出力を多様化し，繰り返し生成問題を**軽減**
  - 出力されうるトークンの候補を絞り，不適切なトークンの生成を**軽減**



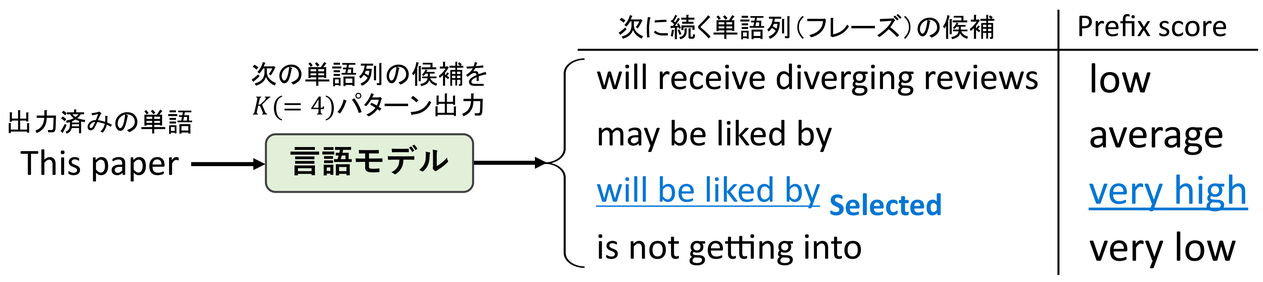
$k = 4$ とした top-k サンプリング[2]の概要



$p = 0.9$ とした top-p サンプリング[3]の概要

Controlled decoding (block-wise best-of-K)

- 人間の嗜好度の予測値（**prefix score**）に基づいてトークンを選択
  - 尤度が高いフレーズと好ましいフレーズは**異なる**場合がある

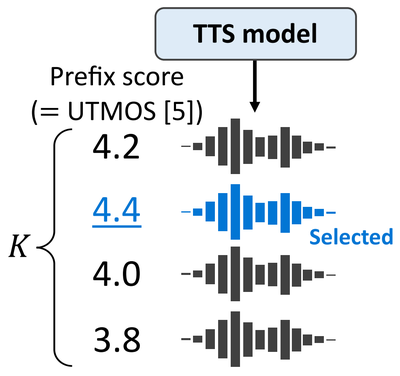


Controlled decoding [4] の概要

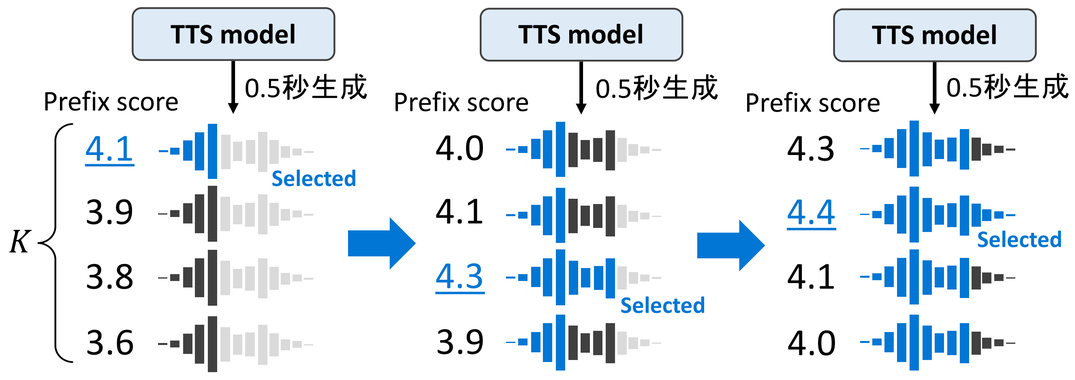
提案手法：音声主観評価値の自動予測を活用した decoding 戦略

TTS のための sequence-level best-of-K / block-wise best-of-K

- top-k top-p サンプリングにより K 通り生成し，**prefix score が最も高い候補**を選択



sequence-level best-of-K の概要



block-wise best-of-K (controlled decoding) の概要

主観評価値の予測に基づく prefix scorer

- 自然性 MOS 予測モデル **UTMOS** [5]を活用
  - 5 段階の自然性 MOS の予測値を prefix score とする
  - 学習時，**入力音声をランダムな時刻以降切り捨てる**

→ 途中まで合成された音声の主観評価値予測が可能に

block-wise best-of-K のための prefix scorer の学習方法の概要

実験的評価：提案手法は合成音声の自然性向上に有効か?

実験条件

- TTS モデル：Transformer TTS [6]
- 離散音声トークン：
  - Descript Audio Codec (DAC) [7]
- データセット：LJSpeech [8]
  - 単一女性英語話者による読み上げ音声
  - 学習/検証/評価：12,600/250/250 発話
- 比較手法：
  - greedy search
  - naive sampling
  - top-k top-p sampling
  - sequence-level best-of-K（**提案手法**）
  - block-wise best-of-K（**提案手法**）

主観評価実験の結果

- 自然性に関する 5 段階 MOS 評価の結果：
  - サンプリングに基づく手法 > greedy search
  - **提案手法** > 従来のサンプリングに基づく手法

提案手法は合成音声の自然性向上に有効！

Method	MOS (↑)	UTMOS (↑)
greedy search	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-level best-of-K	3.71 ± 0.07	4.46
block-wise best-of-K	3.73 ± 0.07	4.43
ground-truth	3.92 ± 0.07	4.43

※受聴者はネイティブ英語話者200人，1人の評価回数は24

サンプル数Kに関する ablation study

- block-wise best-of-K の K と MOS の関係：
  - K を 2 から 8 まで上げると MOS は向上
  - K を 8 から 32 まで上げると MOS は低下
  - 一方，K を大きくするほど UTMOS は 向上

→ UTMOS は主観評価値と必ずしも一貫しない

K を大きくしすぎると prefix scorer に過適合

サンプル数 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

今後の展望

- 自然性向上に有効な**逐次的な** decoding 戦略の提案
  - 実験では逐次的 decoding の自然性向上に対する有効性は示されず
  - 逐次的 decoding は長時間音声のストリーミング合成などに適応可能
- 自然性以外の観点の主観評価値に基づく decoding 戦略の提案
  - Prefix score は自然性に限らず様々な観点の主観評価値に拡張可能
  - **韻律の自然性**や**感情の適合度**などに関する自動評価手法の構築

参考文献

[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] S. Mudgal et al., in *Proc. NeurIPS Workshop*, 2023.
[5] T. Saeki et al., *in Proc. INTERSPEECH*, 2022. [6] N. Li et al., in *Proc. AAAI*, 2019. [7] R. Kumar et al., in *Proc. NIPS*, 2023. [8] K. Ito et al., https: //keithito.com/LJ-Speech-Dataset/, 2017.