第155回 音声言語情報処理研究会 (IPSJ SIG-SLP)

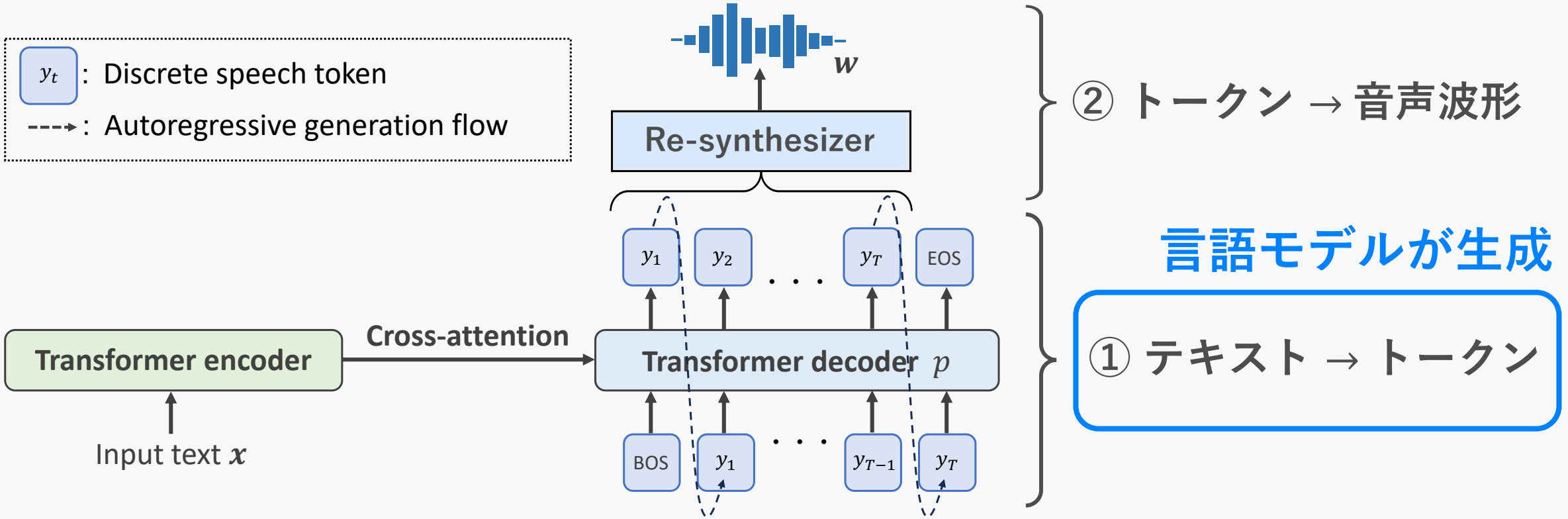
離散音声トークン生成に基づく感情合成音声のための多目的知覚評価値を活用したdecoding戦略

山内一輝，中田亘，齋藤佑樹，猿渡洋

東京大学 情報理工学系研究科

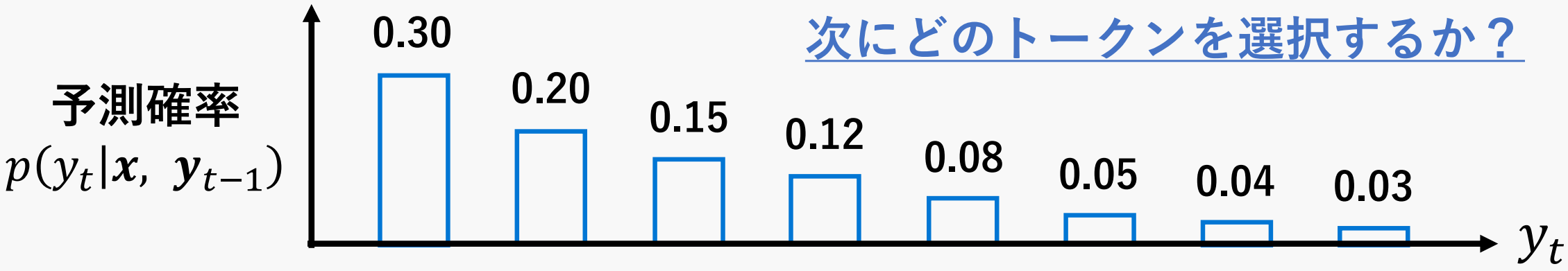
言語モデル（LM）に基づくテキスト音声合成（TTS）

- **言語モデルにより離散音声トークンを自己回帰的に生成**
  - 離散音声トークン: 音声特徴量を量子化することで得られる離散潜在表現



Decoding 戦略

- Decoding: 言語モデルによってモデル化された各トークンの生起確率分布に基づき，次に出力するトークンを選択するプロセス



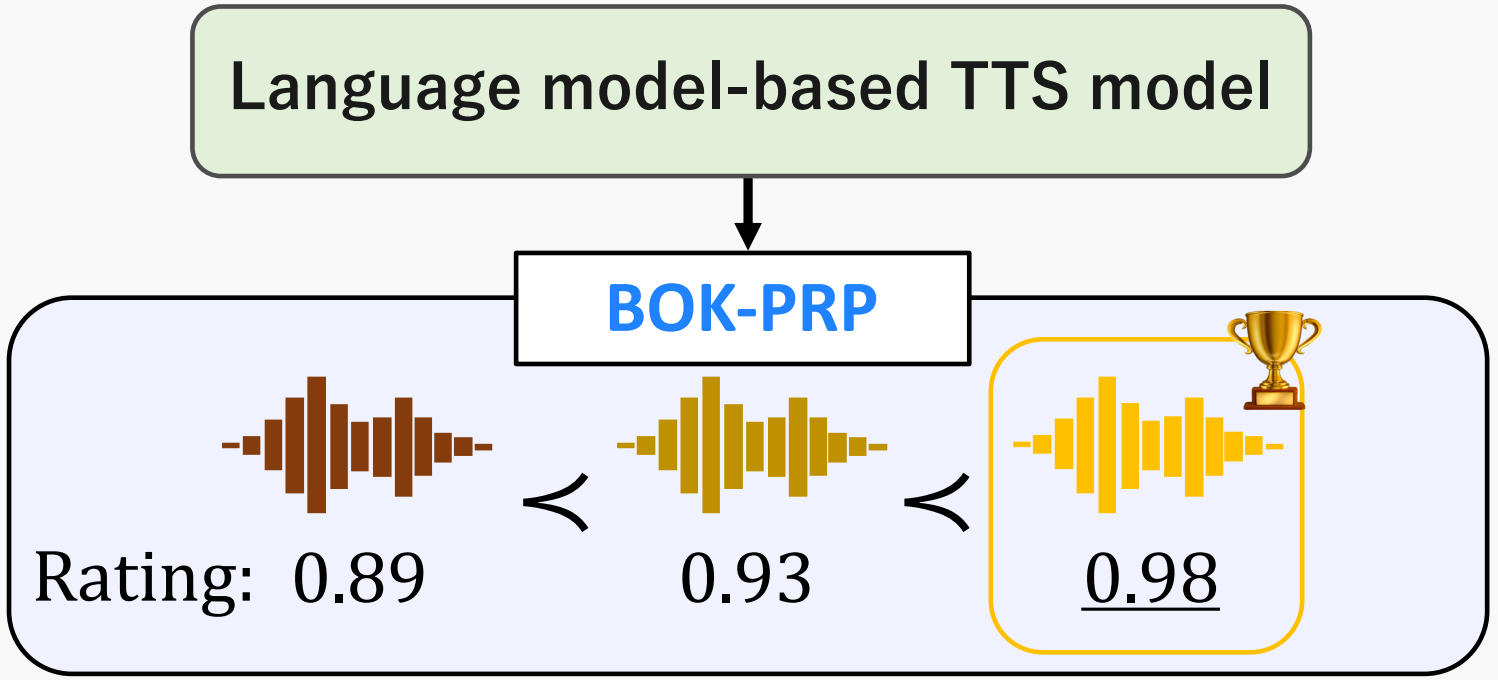
- Greedy decoding
  - 予測確率が最も高いトークンを次の出力トークンとして選択
  - 出力が同じトークンの繰り返しに陥る問題が生じる

従来のサンプリングを伴う decoding 戦略

- **Decoding にサンプリングを導入**
  - 代表例: top-k [A. Fan+2018] / top- p サンプリング[A. Holtzman+2020]
  - トークンを予測確率分布に基づいて確率的に選択
    - 繰り返し生成問題を効果的に対処
- **利点: サンプリングベースの戦略により韻律の多様性が向上**
  - 言語モデルに基づく感情音声合成モデルに適用することで，合成音声の感情表現が豊かになる[Z. Ju+2024]
- **課題: サンプリングのランダム性により生成が不安定化**
  - 雑音など不適切な出力が生成される可能性があり，合成音声の自然性が不安定になる

BOK-PRP: Best-of-K (BOK) selection based on perceptual rating prediction (PRP) [K. Yamauchi+2024]

- サンプリングによって K 通りの音声を生成し，その中から最も“Rating” が高いサンプルを選択
  - 出力の多様性を保ちながら，不適切な出力を効果的にフィルタリング



自然性 MOS 予測に基づく BOK-PRP

- **自然性予測器 $r_{\mathrm{nat}}(\boldsymbol{w})$:**

$$r_{\mathrm{nat}}(\boldsymbol{w}) = \frac{\mathrm{UTMOS}(\boldsymbol{w}) - 1}{4}$$

  - UTMOS [T. Saeki+22]: 広く用いられている自然性に関する 5 段階の Mean Opinion Score (MOS) 予測モデル
  - $r_{\mathrm{nat}}(\boldsymbol{w})$ の出力値は **0 (非常に不自然)** ～ **1 (非常に自然)** までの実数値

先行研究の課題

- **課題 1 : 実験条件が限定的**
  - 実験は単一話者による読み上げ音声合成のみを対象としている
  - サンプリングベースの戦略がより効果を発揮する，感情表現を伴う音声合成などの高度なタスクは検討されていない
- **課題 2 : 知覚評価の観点が限定的**
  - 知覚評価値として自然性のみを検討している
  - 自然性以外(例：感情の表出度)の観点に基づく評価値を複合的に活用する手法は検討はされていない

多目的知覚評価に基づく BOK-PRP

LM-based TTS model

自然性を評価 → $r_{\mathrm{nat}}(w_k)$	0.63	0.70	0.72	0.74
感情強度を評価 → $r_{\mathrm{emo}}(w_k)$	0.75	0.71	0.67	0.61

$$r = \nu_{\mathrm{nat}} \cdot r_{\mathrm{nat}} + \nu_{\mathrm{emo}} \cdot r_{\mathrm{emo}}$$
$$\nu_{\mathrm{nat}} = 0.6,\ \nu_{\mathrm{emo}} = 0.8$$

多目的知覚評価 → $r(w_k)$	0.978	0.988	0.968	0.932

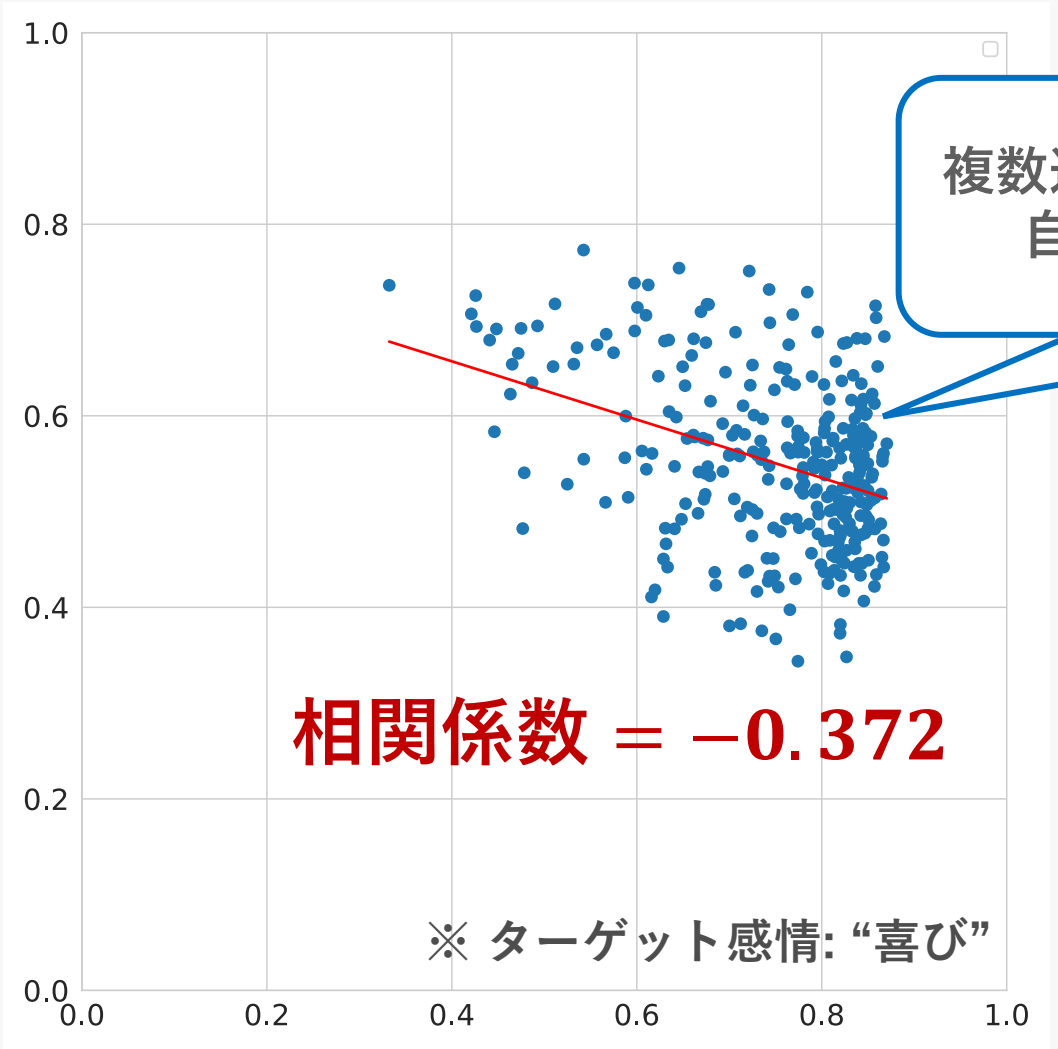
感情の表現力を評価するための感情強度予測器

- **感情強度予測器 $r_{\mathrm{emo}}(\boldsymbol{w})$:**

$$r_{\mathrm{emo}}(\boldsymbol{w}) = \begin{cases} (\mathrm{arousal}(\boldsymbol{w}) + (1 - \mathrm{valence}(\boldsymbol{w})))/2 & \text{if ターゲット感情が”怒り”} \\ (\mathrm{arousal}(\boldsymbol{w}) + \mathrm{valence}(\boldsymbol{w}))/2 & \text{if ターゲット感情が”喜び”} \\ ((1 - \mathrm{arousal}(\boldsymbol{w})) + (1 - \mathrm{valence}(\boldsymbol{w})))/2 & \text{if ターゲット感情が”悲しみ”} \end{cases}$$

  - 事前学習済みの次元感情認識モデル [L. Goncalves+24] モデルを活用
    - 覚醒度 (arousal)，支配性 (dominance)，感情価 (valence) を 0 (低い) ～ 1 (高い) の範囲の実数値で予測
  - ターゲット感情: 怒り (angry), 喜び (happy), 悲しみ (sad)
  - $r_{\mathrm{emo}}(\boldsymbol{w})$ の出力値は **0 (非常に弱い)** ～ **1 (非常に強い)** までの実数値

自然性と感情強度にトレードオフが存在



複数通りサンプリングされた合成音声の
自然性と感情強度の評価値の分布

多目的知覚評価値

- **多目的知覚評価値予測器 $r(\boldsymbol{w})$:**

$$r(\boldsymbol{w}) = \nu_{\mathrm{nat}} \cdot r_{\mathrm{nat}}(\boldsymbol{w}) + \nu_{\mathrm{emo}} \cdot r_{\mathrm{emo}}(\boldsymbol{w}),$$

where $\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}} \in \mathbb{R},\ \nu_{\mathrm{nat}}^2 + \nu_{\mathrm{emo}}^2 = 1$

  - 合成音声の**自然性と感情強度にトレードオフが存在**することに着目
  - 重み係数 ν_{nat} と ν_{emo} を手動で調整することで，decoding 時に**自然性と感情強度のどちらをより重視するかを制御可能**

実験条件

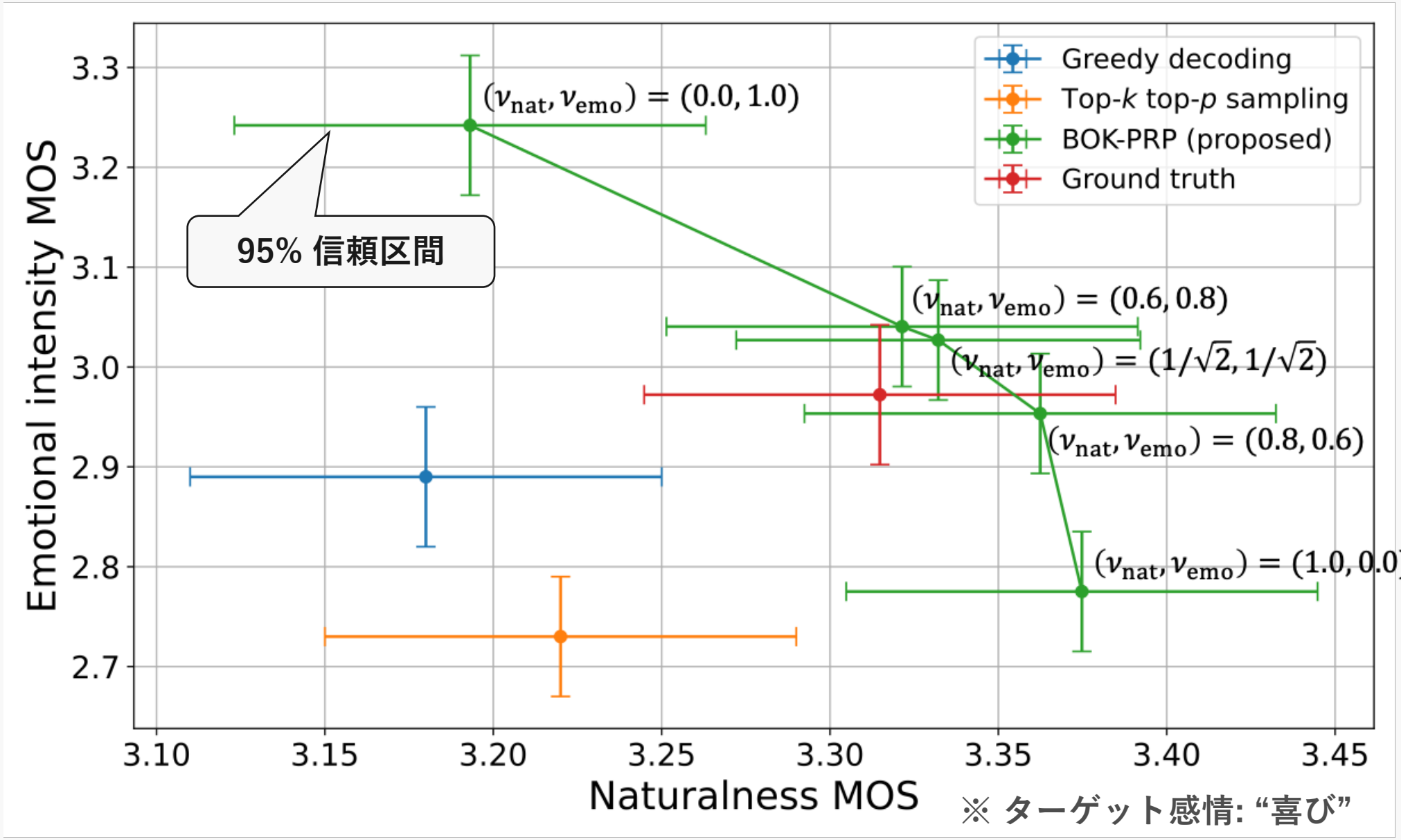
- **言語モデルに基づく感情 TTS モデル:**
  - Cosyvoice-300M-Instruct model [Z. Du+24] (事前学習済み)
- **比較評価のための参照音声:**
  - Emotional Speech Dataset (ESD)-English [K. Zhou+22] のテストセット
    - 10名の話者 × 3感情（怒り, 喜び, 悲しみ） × 各30発話（計900発話）
- **主観評価指標:** **※ 250 人の参加者が，それぞれ 24 サンプルを評価**
  - 自然性 MOS
    - 合成音声の自然性を **1 (非常に不自然)** ～ **5 (非常に自然)** の 5 段階で評価
  - 感情強度 MOS
    - 合成音声の感情強度を **1 (非常に弱い)** ～ **5 (非常に強い)** の 5 段階で評価

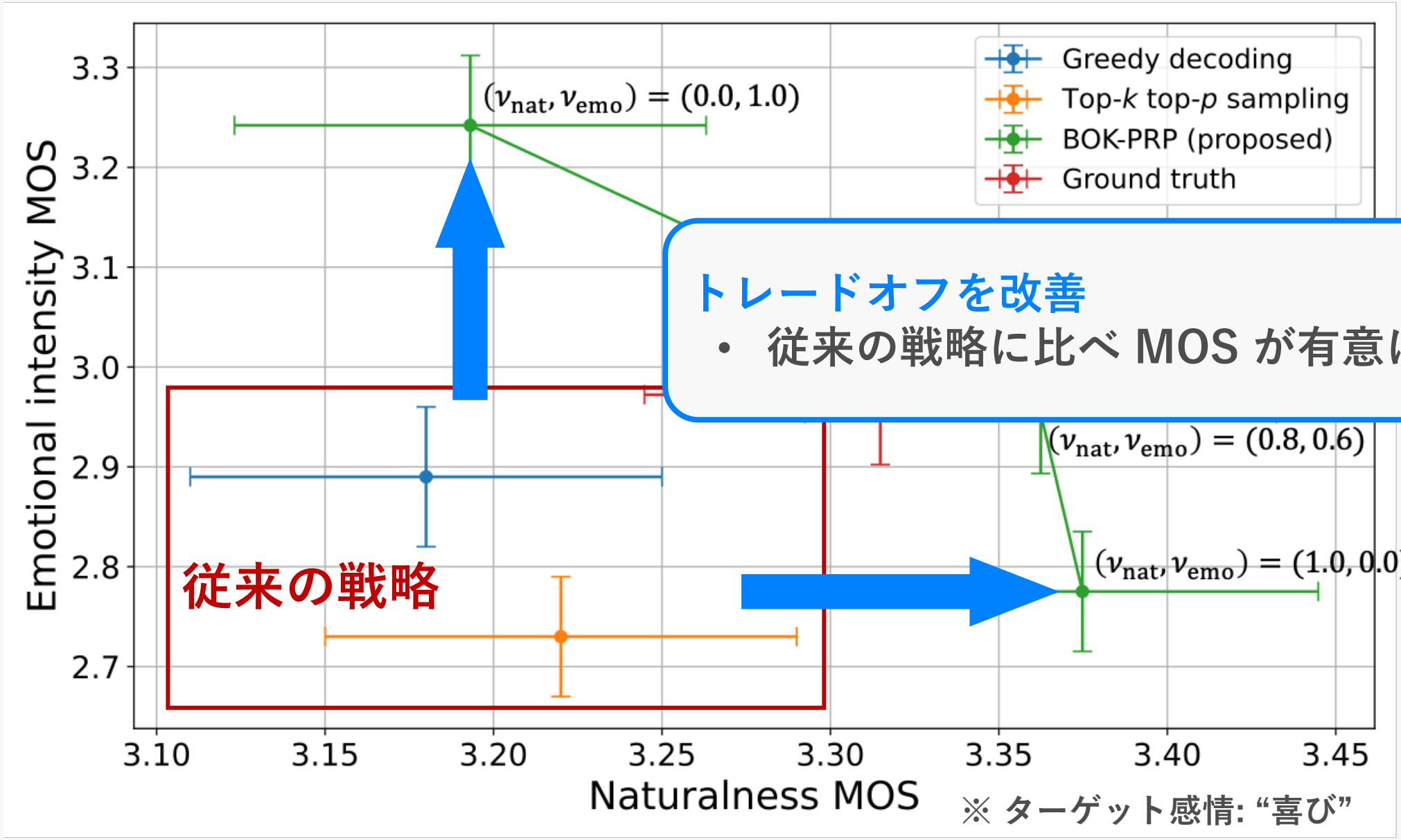
実験条件

- **比較手法:**
  - Greedy decoding
  - Top-k top-p sampling
  - 多目的知覚評価値に基づくBOK-PRP (Proposed)
- **BOK-PRP の設定**
  - サンプルサイズ K の値は 10 に設定
  - 多目的知覚評価値の重み係数は以下の 5 通りの組み合わせを比較:

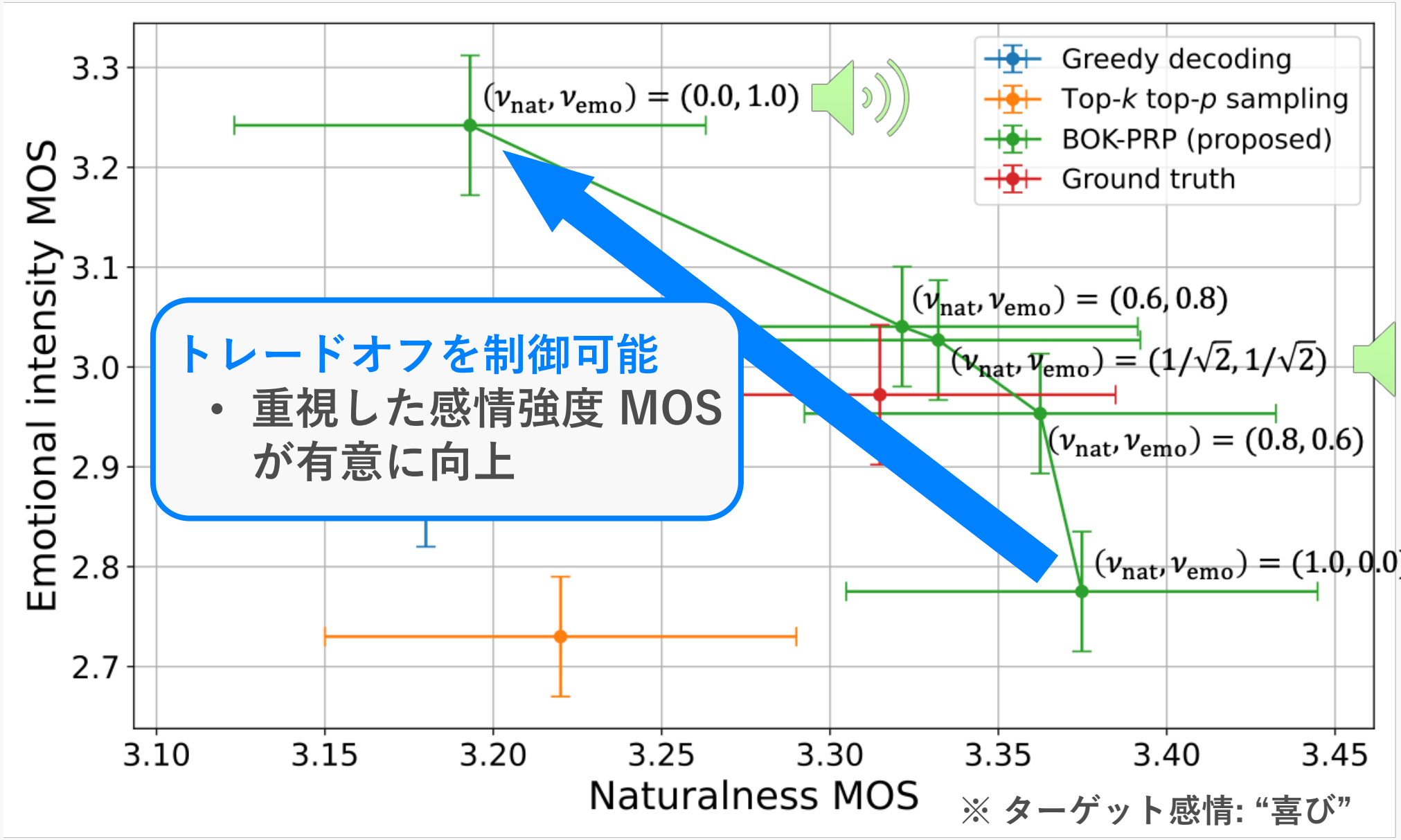
$$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (1.0, 0.0), (0.8, 0.6), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), (0.6, 0.8), (0.0, 1.0)$$

主観評価実験の結果


Greedy decoding
Top-k top-p sampling
BOK-PRP (proposed)
Ground truth
95% 信頼区間
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (0.0, 1.0)$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (0.6, 0.8)$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (1/\sqrt{2}, 1/\sqrt{2})$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (0.8, 0.6)$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (1.0, 0.0)$
Emotional intensity MOS
3.3
3.2
3.1
3.0
2.9
2.8
2.7
3.10
3.15
3.20
3.25
3.30
3.35
3.40
3.45
Naturalness MOS
※ ターゲット感情: “喜び”


Greedy decoding
Top-k top-p sampling
BOK-PRP (proposed)
Ground truth
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (0.0, 1.0)$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (0.8, 0.6)$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (1.0, 0.0)$
従来の戦略
トレードオフを改善
• 従来の戦略に比べ MOS が有意に向上
Emotional intensity MOS
Naturalness MOS
3.3
3.2
3.1
3.0
2.9
2.8
2.7
3.10
3.15
3.20
3.25
3.30
3.35
3.40
3.45
※ ターゲット感情: “喜び”

主観評価実験の結果

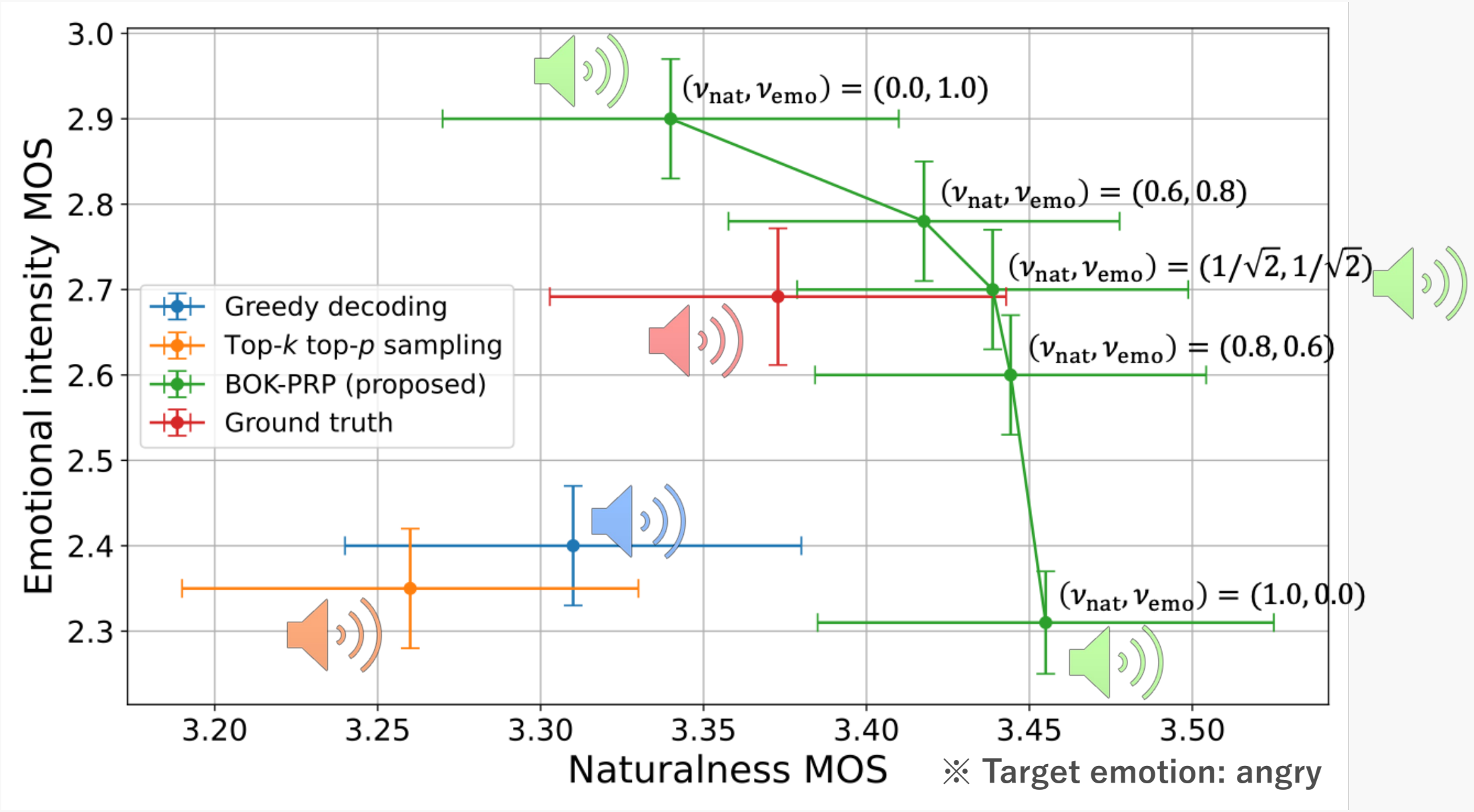

Greedy decoding
Top-k top-p sampling
BOK-PRP (proposed)
Ground truth
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (0.0, 1.0)$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (0.6, 0.8)$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (1/\sqrt{2}, 1/\sqrt{2})$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (0.8, 0.6)$
$(\nu_{\mathrm{nat}}, \nu_{\mathrm{emo}}) = (1.0, 0.0)$
トレードオフを制御可能
• 重視した感情強度 MOS が有意に向上
Emotional intensity MOS
3.3
3.2
3.1
3.0
2.9
2.8
2.7
3.10
3.15
3.20
3.25
3.30
3.35
3.40
3.45
Naturalness MOS
※ ターゲット感情: “喜び”

BOK-PRP に多目的知覚評価を新たに導入

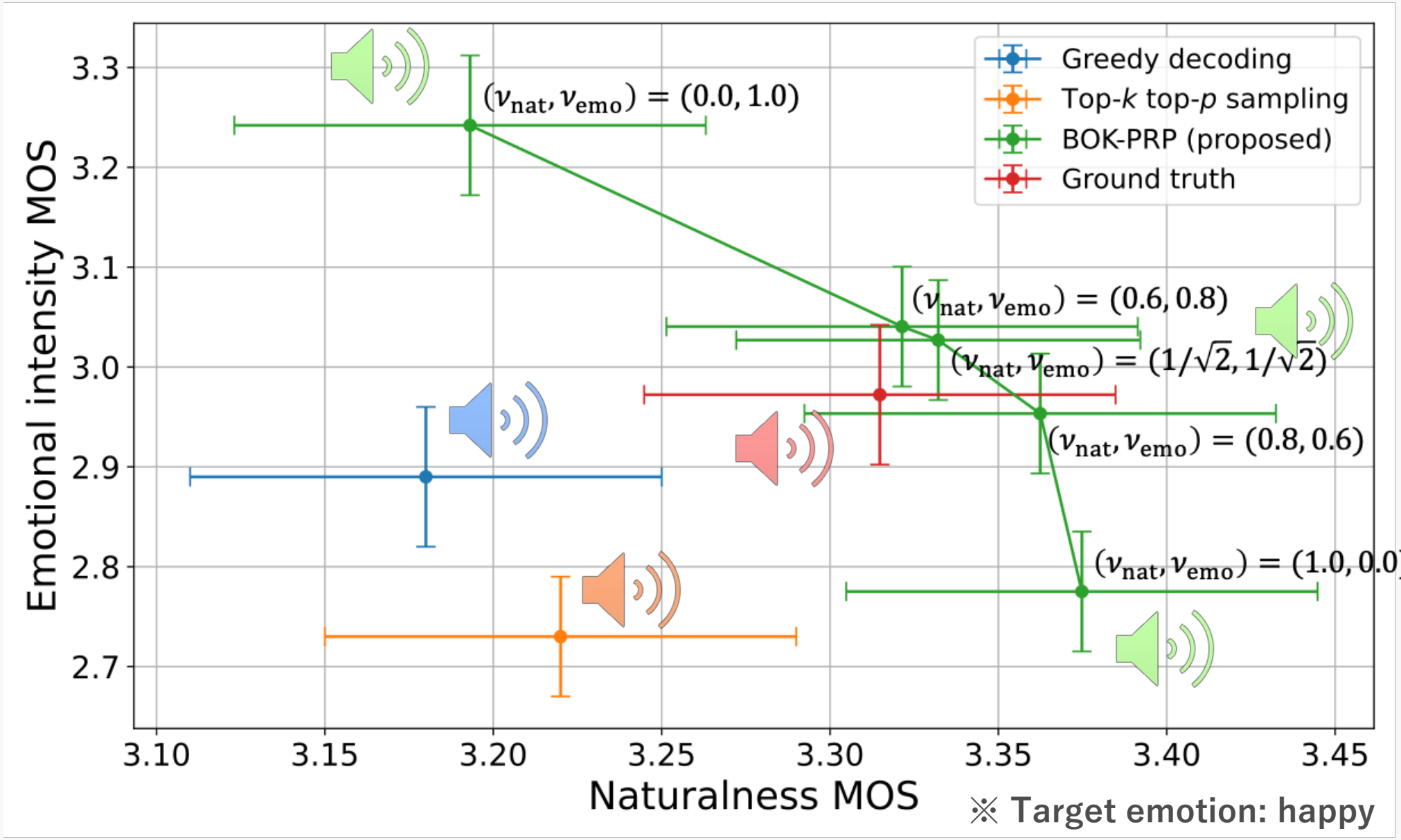
- **本研究の目的:**
  - 合成音声の自然性を保ちながら，多様で豊かな感情表現を生成可能な decoding 戦略の開発
- **提案手法:**
  - 感情音声合成のための**多目的知覚評価値**に基づく BOK-PRP
  - 合成音声の**自然性と感情強度のトレードオフを制御・改善**
- **今後の展望:**
  - BOK-PRP を韻律的な自然性や音声対話における流暢性など，より複雑で多様な知覚評価に拡張

Appendix

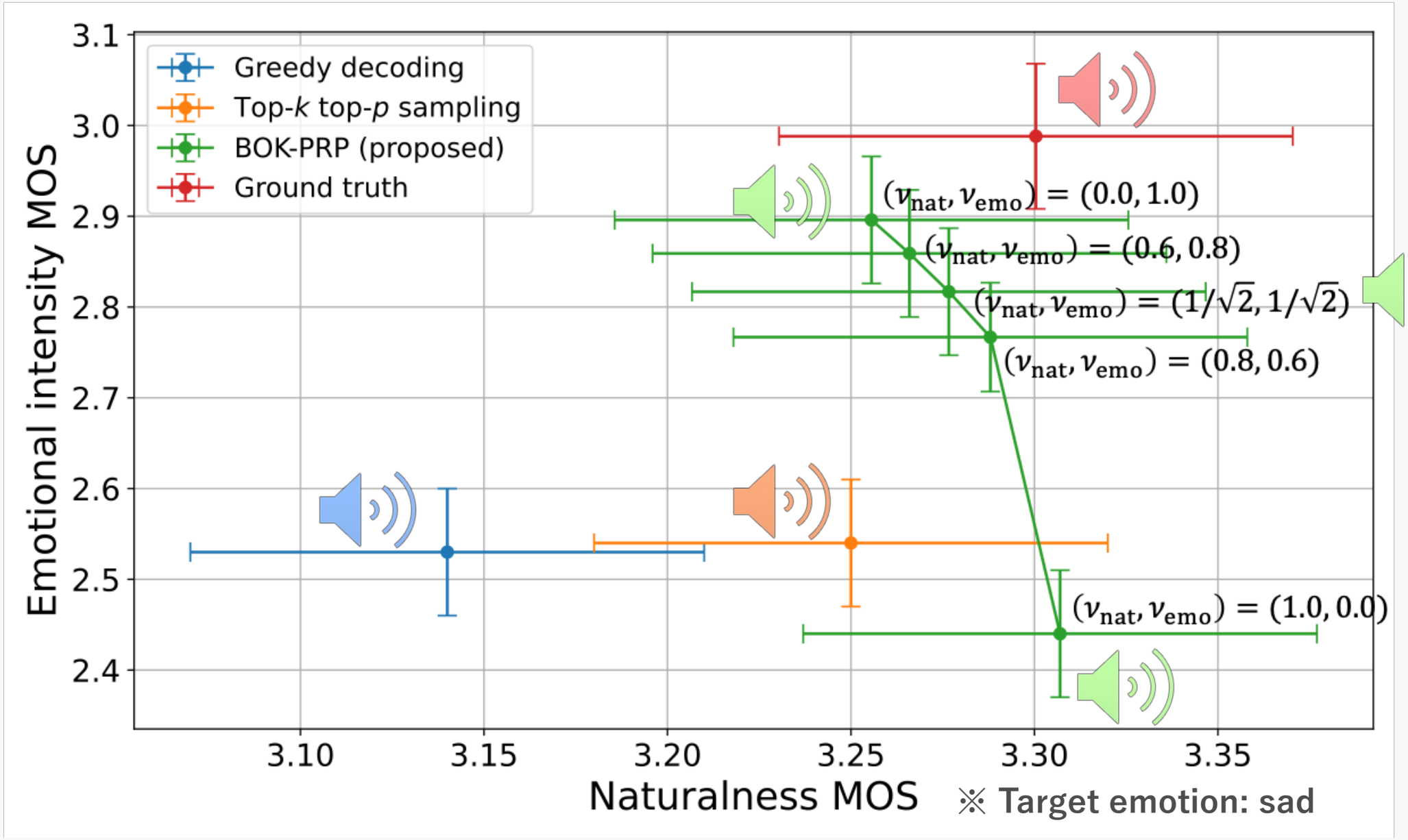
Results of subjective evaluations of BOK-PRP



Results of subjective evaluations of BOK-PRP



Results of subjective evaluations of BOK-PRP



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

Over-optimization:

Excessively large K **degrades naturalness**

Listening Evaluation Test

Please read the following instructions carefully.

Listen to the speech with anger and rate **the naturalness of the speech** (does it sound human-like and natural?) and **the intensity of anger** on a 5-point scale, respectively.

When you click [start] button, the audio starts to play.

Please perform the evaluation in a quiet environment and with headphones.

Do not use your browser's “back” or “reload” buttons during the test.

Select [1] to [5] (you can play it again with the [replay] button).

Question 1/24

Naturalness

1: Very Poor	2: Poor	3: Fair	4: Good	5: Very Good

Intensity of anger

1: Very Weak	2: Weak	3: Moderate	4: Strong	5: Very Strong

user:beethoven1166 start replay submit