2-Q-30 対戦ゲーム動画の実況解説音声の分析と合成の検討

☆**井浦 昂太**，齋藤 佑樹，猿渡 洋（東京大学）

概要：対戦ゲーム実況解説音声の特徴の分析＆合成

- 対戦ゲーム実況解説音声
  - 対戦ゲーム動画に対して音声を追加する
  - 視聴者の認識や興奮度に影響を与える[1]
  - スポーツ解説音声合成[2] → 状況説明にとどまる
- 実況解説特有の韻律を再現する音声合成
  - 視聴者を盛り上げるような音声の合成
  - エンターテインメント分野への応用

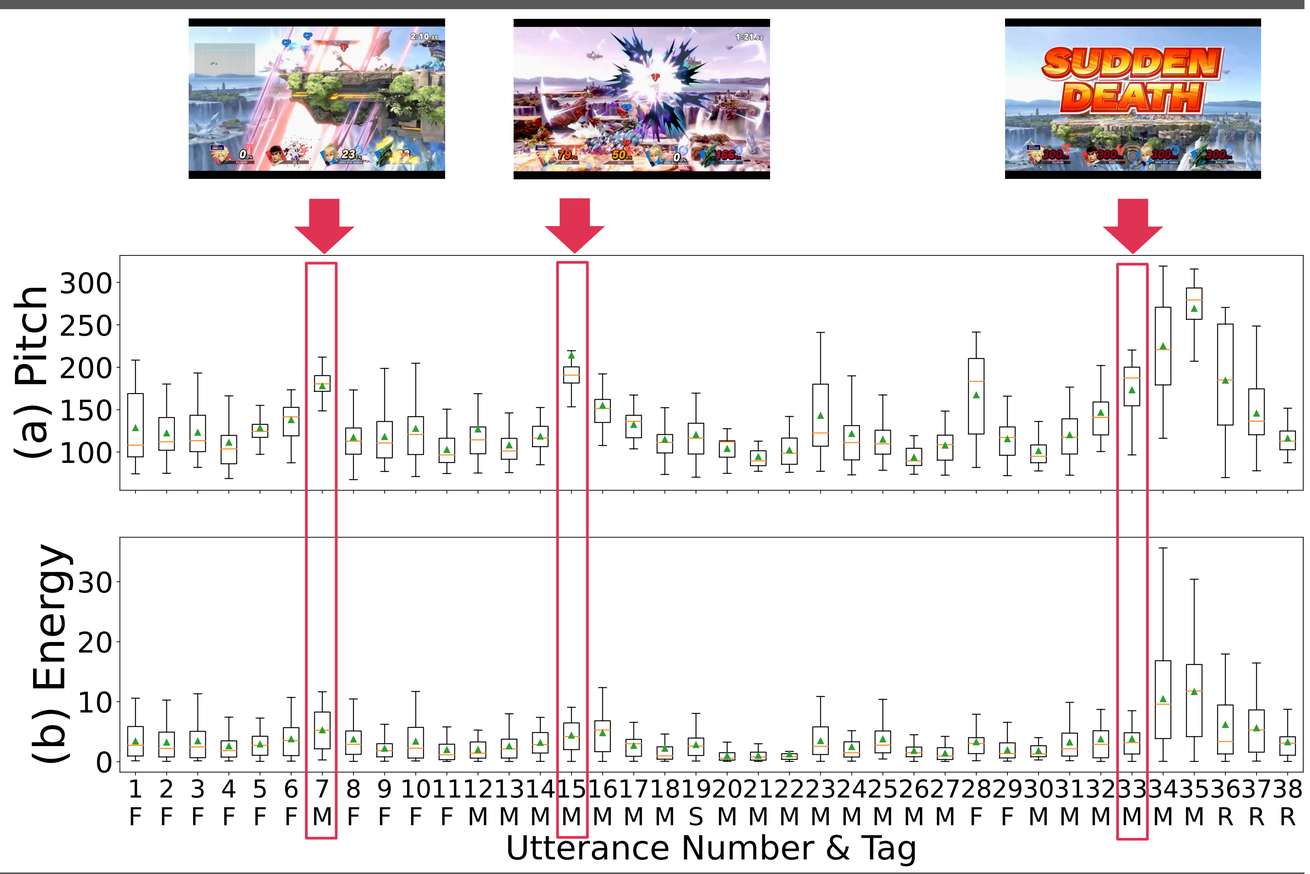
対戦ゲーム実況解説音声の分析

- SMASH corpus[3]
  - 大乱闘スマッシュブラザーズ SPECIAL の実況解説音声
  - コーパス音声の pitch, energy を分析

ID	テキスト
7M	あーリュウが**撃墜**されてしまった
15M	クラウドが**スマッシュボール**取りましたそして
33M	これは引き分けいや引き分けですね**サドンデス**

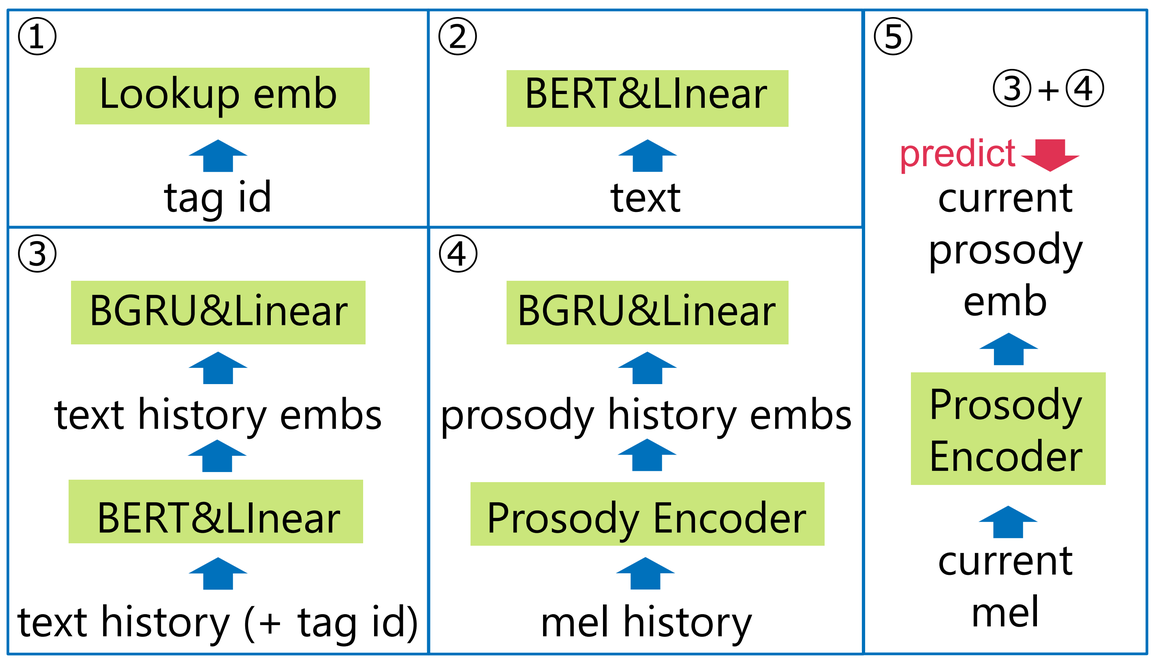
- 音声の pitch, energy
  - 撃墜や必殺技のシーンでピークになりやすい
  - 特定の単語が使われている場合が多い
  - 過去の発話内容がピークに関連している (33M)

履歴を考慮した音声合成モデルが効果的？



実況解説音声合成モデル構造の検討

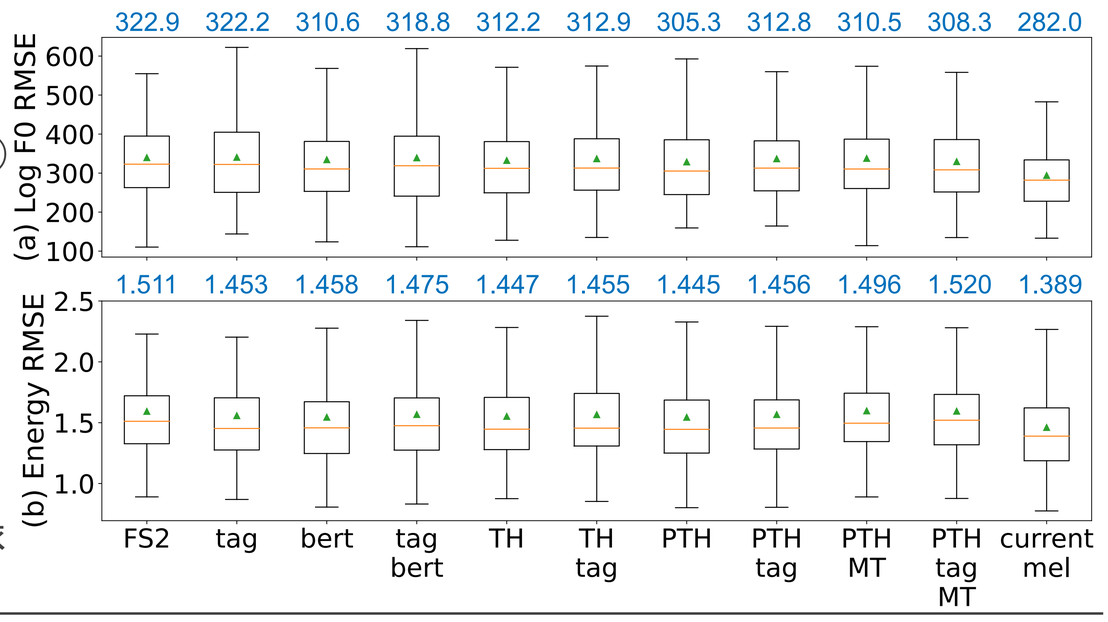
- 発話履歴を考慮したモデル (ベースは FastSpeech2:FS2)
  1. 言語情報からの文脈埋め込みベクトル推定[4] (③)
     - 履歴の各発話を BERT[5]を用いて埋め込みベクトル化
     - BGRU を用いて固定次元ベクトルに圧縮し条件付け
  2. 韻律情報を考慮した文脈埋め込みベクトル推定[6] (③+④)
     - 履歴の各メルスペクトログラムを prosody encoder (畳み込み層 + BGRU) により prosody emb (pe) を得る
     - [4]と同様に BGRU を用いて固定次元化
     - 二段階学習 (1. 現在発話の pe を用いて条件付け 2. 音響モデルと prosody encoder の重みを固定し，現在発話の pe を発話履歴ベクトルから推定) も存在[7] (⑤)
  - それぞれ tag id を加えたものも用意
- 比較として，現在発話の情報のみを用いるモデルも用意
  - tag id (①) , bert emb (②)



※ 埋め込みベクトルは FS2 の Encoder の出力に加える

客観評価実験＋結果

- train: 40試合，val: 6試合，test: 6試合 (1試合約3分)
- **FS2**: 条件付けなし，**tag**:①，**bert**:②，**tag bert**:①+②
  TH:③，**TH tag**:③w/tag，**PTH**:③+④，**PTH tag**:③w/tag+④
  PTH MT:⑤，**PTH tag MT**:⑤w/tag, **current mel**:⑤第1段階
- log f0, energy についてRMSEを計算 (右図青字は中央値)
  - 条件付けなし FS2 から有意な改善はなし
  - 正解音声を用いた pe では有意に改善
  - 話者はプロではない → スタイル制御がうまくない可能性
- 今後の課題：正解音声の pe の予測，プロ実況解説者の音声収録



参考文献

[1] J. Bryant et al., Journal of Communication, vol. 32, no. 1, 1982. [2] Kumano et al., Symp BMSB, 2019. [3] Saito et al., Proc. LREC, 2020.
[4] H. Guo et al., Proc. SLT, 2019. [5] J. Devlin et al., Proc. NAACL-HLT, 2019. [6] Nishimura et al., Proc. INTERSPEECH, 2022. [7] 西邑 他, 情報処理学会研究報告, 2022