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（東大院・情報理工）

1 はじめに

スポーツ実況解説を始めとして，実況解説者は対象とする映像に音声を追加することで，視聴者の認識や興奮度に影響を与える [1]．実況解説音声を再現できる音声合成手法の実現は，視聴者の理解やコンテンツへの参加を促すような，エンターテインメント分野への応用が期待される．実況解説音声の先行研究として，Kumano ら [2] はスポーツ中継の解説音声を自動で生成する手法を提案しているが，状況の補足説明にとどまっており，より視聴者を盛り上げるような実況解説音声の研究は十分になされていない．

本稿では，実況解説音声の中でも，近年エンターテインメント分野として注目されている，対戦ゲームプレイに対する実況解説者の発話内容と音響的特徴の分析を行い，実況解説音声の発話スタイルを再現する音声合成手法の検討を行った．

2 実況解説音声の分析

実況解説音声として，SMASH コーパス [3] を用いた．SMASH コーパスは，大乱闘スマッシュブラザーズ SPECIAL（スマブラ SP）の対戦動画に対する実況解説音声が収録されており，本研究が対象とする音声に合致している．SMASH コーパスの音声の韻律特徴量（pitch と energy）の分析結果の一例を Figure 1 に示す．この図は，ある一試合の発話ごとの pitch と energy を音素単位で平均して箱ひげ図にしたものである．横軸の数値は何番目の発話かを表しており，数値が大きくなるほど時間が経過したことを表す．横軸のアルファベットはトピックタグを表している．このトピックタグは SMASH コーパスに付与されており，各発話が意味する内容を大まかに分類している．タグは全 8 種類であり，F: Fighter, S: Stage, I: Item, P: Pokemon, A: Assist Trophy, M: Match, R: Result, C: Chat である．Figure 1 を見ると，試合全体の発話の中にピークが存在することが読み取れ，試合の中での場面に合わせて，実況解説者が発話スタイルを変化させていると考えられる．ピークに関わる発話テキストを Table 1 に示す．7M と 15M では，「撃墜」，「スマッシュボール」といった，試合内容に大きな影響を与える事象を直接表す単語が用いられており，スタイル変化にも関連していると考えられる．また，33M では「サドンデス」という，試合の中で大きく盛り上がるものを示した単語が使われており，33M 以降のピークに関連していると考えられる．

分析の結果，実況解説音声の発話スタイル変化を再現するには，発話内容の言語的情報を履歴も含めて扱うことが効果的であると予想される．また，実況解説者は試合単位で連続的に話しているため，韻律情報の発話履歴も考慮することが効果的であると考えられる．

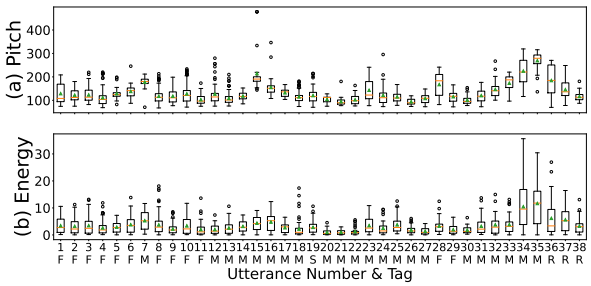

Fig. 1 pitch と energy の一試合中の変化の例．横軸の数値が大きくなるほど時間が経過している．pitch, energy 共にピークがあり，試合状況に応じて発話スタイルが変化している．

Table 1 発話内容の例．ID は Figure 1 の横軸と対応している．

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**

3 実況解説音声合成モデル構造の検討

本研究では，実況解説音声合成モデルのベースラインとして，発話履歴を考慮した音声合成モデルを採用する．

3.1 言語情報からの文脈埋め込みベクトル推定

Guo ら [4] は，音声対話履歴の言語情報を考慮した End-to-End 音声合成の枠組みを提案している．この手法では，対話履歴の各発話テキストを BERT [5] を用いて文埋め込みベクトルに変換し，それを Bi-directional GRU (BGRU) を用いて圧縮した固定次元ベクトルを用いて音声合成モデルに条件づけする．

3.2 韻律情報を考慮した文脈埋め込みベクトル推定

Nishimura ら [6] [7] はテキストを用いた言語情報に加えて，韻律情報の履歴を用いた対話音声合成の枠組みを提案している．畳み込み層と BGRU からなる prosody encoder を用いて対話履歴の各発話のメルスペクトログラムから対話履歴の韻律埋め込みベクトルを抽出し，BGRU を用いて圧縮した固定次元ベクトルを得る．言語情報は Guo ら [4] と同様の手法で固定次元ベクトルを抽出し，韻律・言語それぞれのベクトルを音響モデルに足し合わせることで条件づけする．また，Nishimura らは 2 段階の学習 [6] を提案している．1 段階目の学習で，現在発話の正解音声メルスペクトログラムを用いた条件付を行い，音響モデルと prosody encoder を学習する．2 段階目の学習で，音響モデルと prosody encoder の重みを固定し，現在発話の音声から得られた韻律埋め込みベクトルを，発話履歴のテキストと音声から得られた文脈埋め込みベクトルで予測するように学習することで，より効率的な学習を可能にしている．本研究ではこの手法も採用した．

* Analysis and synthesis of commentary audio for competitive game videos by IURA, Kota, SAITO, Yuki, and SARUWATARI, Hiroshi (The University of Tokyo).

4　実験的評価

4.1　実験条件

データセットとして，SMASH コーパス [3] を用いた．実況解説者 2 名の内，収録データが多い MC1 の音声を用いた．本研究では発話履歴を考慮したモデルを用いるため，データセットの分割は試合単位で行った．学習データ，検証データ，テストデータはそれぞれ 40 試合，6 試合，6 試合とした．1 試合約 3 分であり，1 試合あたり平均して 30 発話に分割されている．また，音響モデル (FastSpeech2 [9]) の事前学習に JSUT コーパス [8] を用いた．FastSpeech2 の実装は日本語音声向けのオープンソース[1]を参考に構築した．メルスペクトログラムから音声波形を生成するニューラルボコーダとして，HiFi-GAN [10] を利用した．HiFi-GAN のモデルとして，公開されている事前学習済みの UNIVERSAL_V1 モデル[2]を利用した．

音響モデルへの条件付けの概要を説明する．現在発話のみを用いるモデルとして，トピックタグ ID から埋め込みベクトルを生成して条件付けするモデル (**tag**)，発話テキストから BERT を用いて抽出したテキスト埋め込みベクトルで条件付けするモデル (**bert**)，**tag** と **bert** のどちらも用いて条件付けするモデル (**tag+bert**) を作成した．発話履歴を用いるモデルとして，Guo ら [4] の実装に従うモデル (**TH**) と Nishimura ら [6] [7] の実装に従うモデル (**PTH**) を作成し，これらのモデルのテキスト埋め込みベクトルに 1 次元のトピックタグ ID を結合したモデル (**TH+tag**，**PTH+tag**) を作成した．また，Nishimura らのモデルでは 2 段階学習を用いたもの (**PTH MT**，**PTH+tag MT**) も作成した．いずれのモデルも，固定長のベクトルを FastSpeech2 の Encoder の出力に足し合わせることで条件づけしている．BERT は日本語用に fine-tune されたモデル[3]を用いた．

4.2　客観評価結果

客観評価の指標として，自然音声と合成音声の間の対数基本周波数 (log F0) 及び energy の Root Mean Squared Error (RMSE) を用いた．Figure 2 に客観評価の結果，表 2 に条件付けしていない FastSpeech2 と各モデルを比較した検定結果を示す．ただし，合成の際に音声の長さを揃えるため，duration のみ正解を与えて合成している．また，2 段階学習の 1 段階目は正解音声の韻律埋め込みベクトルで条件付けしたものであるため，2 段階学習の理想的な結果であると考えられる．そこで，2 段階学習の 1 段階目を用いて合成した音声 (**current mel**) を参考値として記載した．数値として最も評価の高かったものは **PTH** であったが，**current mel** を除く全てのモデルにおいて log F0 と energy のどちらにおいても，条件づけを行っていない FastSpeech2 との比較で有意差は見られなかった．原因として，SMASH コーパスで収録されている音声は，実況解説者としての仕事の経験がない人物の音声を収録しているため，実況解説に合った発話スタイルの制御がうまくできておらず，特徴の抽出が難しいことが考えられる．また，**current mel** で有意な改善が見られる一方，2 段階学習を行ったモデル (**PTH MT**，**PTH+tag MT**) は改善が見られないことから，2 段階学習が有効に働いていないとわかる．

[1] https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[2] https://github.com/jik876/hifi-gan
[3] https://huggingface.co/colorfulscoop/sbert-base-ja

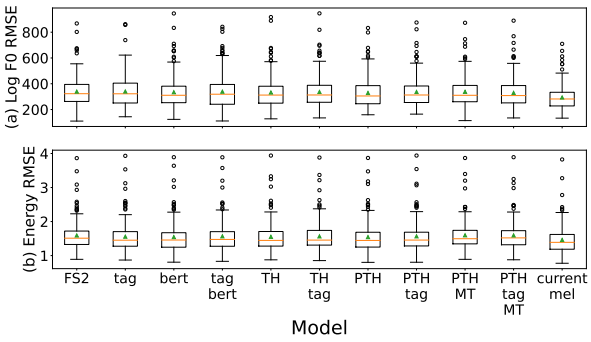

Fig. 2　log F0 と energy の客観評価結果．合成の際に，duration は正解を与えている．

Table 2　客観評価結果の検定結果．いずれも条件づけしていない FastSpeech2 と比較している．

FS2 vs	p 値 (log F0)	p 値 (energy)
tag	0.99	0.16
bert	0.43	0.077
tag+bert	0.49	0.28
TH	0.32	0.12
TH+tag	0.54	0.30
PTH	0.19	0.059
PTH+tag	0.50	0.30
PTH MT	0.75	0.79
PTH+tag MT	0.29	0.81
(current mel)	$< 10^{-4}$	$< 10^{-3}$

5　まとめと今後の展望

本稿では，対戦ゲーム実況解説音声の分析と，その音声の特徴的な発話スタイルを再現する音声合成の検討を行った．検討したいずれのモデルでも有意な改善は見られなかったが，現在発話の韻律埋め込みベクトルで条件付けした場合は有意な改善が見られた．今後は，この現在発話の韻律埋め込みベクトルを予測できる手法を検討する．また，大会などで実況解説経験のあるプロの話者の音声の収録を行い，よりスタイル制御が正確な音声を用いた合成を検討する．

謝辞: 本研究の一部は，JSPS 科研費 22K17945 の助成を受けたものです．

参考文献

[1] J. Bryant et al., Journal of Communication, Vol. 32, No. 1, pp. 109–119, March, 1982.
[2] T. Kumano et al., Symp. BMSB, pp. 1–4, Jun, 2019.
[3] Y. Saito et al., Proc. LREC, pp. 6571–6577, Marseille, France, May, 2020.
[4] H. Guo et al., Proc. SLT, pp. 403–409, Shenzhen, China, January 2021.
[5] J. Devlin, et al., Proc. NAACL-HLT, pp. 4171–4186, Minneapolis, U.S.A., June, 2019.
[6] 西邑 他, 情報処理学会研究報告, 2022-SLP-140, pp. 1–6, 2022.
[7] Y. Nishimura et al., Proc. INTERSPEECH, pp. 3373–3377, Incheon, Korea, September, 2022.
[8] S. Takamichi et al., Acoustical Science and Technology, Vol. 41, No. 5, pp. 761–768, Sep, 2020.
[9] Y. Ren et al., Proc. ICLR, Virtual Conference, May 2021.
[10] J. Kong et al., Proc. NeurIPS, pp. 17022–17033, Virtual Conference, December, 2020.