盛り上がり制御可能な対戦ゲーム実況解説音声合成モデルの検討

井浦 昂太[1,2]　齋藤 佑樹[1]　高道 慎之介[3,1]　ニュービッグ グラム[4]　須藤 克仁[5]　猿渡 洋[1]
高村 大也[2]　石垣 達也[2]

概要：近年，音声合成モデルは感情をはじめとして表現力豊かな音声を合成することが可能になっているが，多くのモデルではどのような表現を行うか事前に指定する必要がある．人間は，周りの状況に応じて発話スタイルを制御するアドリブ性を有している一方，このようなアドリブ性は現状の音声合成モデルにはない．そこで，本研究はシーンの変化が激しく発話スタイルのアドリブ性が強く要求される「対戦ゲーム実況解説」に焦点を当て，ゲームシーンに応じて発話の盛り上がりを制御できる音声合成モデルを提案する．提案手法は，実況解説音声コーパスの音声を，盛り上がりの有無に基づいて “high”（盛り上がりあり）と “low”（盛り上がりなし）に分類し，それぞれのラベルを与えて学習することで low/high の切り替えを可能とする音声合成モデルと，対戦ゲーム動画を入力として映像の盛り上がりを予測し，シーンに合わせた low/high の切り替えを可能とするモデルからなる．実験的評価では，音声合成モデルが盛り上がりの有無をラベルによって制御できているのかと，動画から制御を行うモデルも含めたシステム全体が盛り上がりのある実況解説を生成できるのか主観評価によって検証した．その結果，音声合成モデルは盛り上がりをラベルによって制御できること，システム全体で生成された映像はベースラインと比較して盛り上がりや面白さを向上できることを確認した．

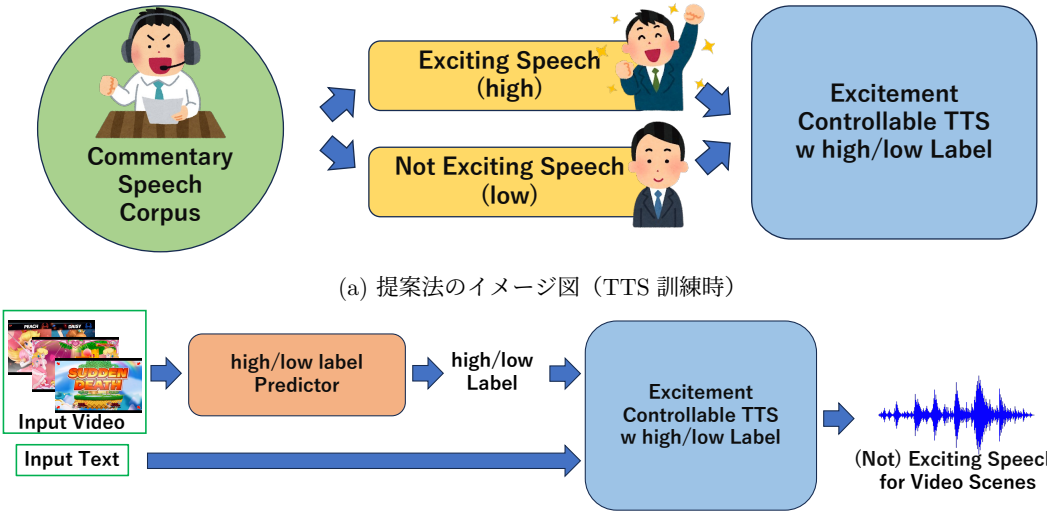
1. はじめに

Deep Neural Network (DNN) を用いた音声合成をはじめとする近年の音声合成技術の発展により，自然性だけでなく感情をはじめとする表現力豊かな text-to-speech (TTS) が研究されている [1], [2]．しかしながら，表現力豊かな TTS の多くは，音声を合成する際に感情ラベルなどの情報を与える必要があり，プロンプトなどを用いて制御性が向上しているものの，多くの場合で人手で指定するものが多い [3], [4]．人間による実際の音声は，視覚的情報や聴覚的情報などの周りの状況に影響を受け，状況の変化に応じてリアルタイムに発話スタイルを変化させるという特徴をもつが，現状の TTS ではその再現までできていない．

本研究では，話者の視覚情報に応じて発話スタイルを変化させるようなアドリブ性を持つ TTS の構築を目指す．音声のアドリブ性が要求される場面は多く存在するが，本研究は (1) 先行研究 [5] によりデータセットが整備されている，(2) 対戦ゲームは状況の変化が激しく，特にアドリブ性を要求される，(3) 実況解説者は視聴者を盛り上げるため，発話スタイルを大きく変化させる，という 3 点の理由から「対戦ゲーム実況解説」の分野に注目する．対戦ゲーム実況解説の感情表現として，音声を聞いたときに聴衆が感じる「盛り上がり」に注目し，ゲームシーンに合わせて聴衆を盛り上げるような話し方を制御できる TTS システムを提案する．提案する手法の概要を図 1 に示す．提案するシステムは大きく 2 つのモジュールに分けられる．1 つ目は，盛り上がりの有無に基づいて “high”（盛り上がりあり）と “low”（盛り上がりなし）に分類し，それぞれのラベルを与えて学習することで盛り上がりの制御を可能とする TTS モデルである．2 つ目は，実況解説が行われる対戦ゲーム動画を入力とし，入力されたシーンが盛り上がっているかどうかを識別するモデルである．

実験的評価では，まず TTS の盛り上がりの再現ができているかを評価し，実際に盛り上がりの有無をラベルにより制御できることを確認した．最後に連結したシステム全体で合成した音声が視聴者の盛り上がりをはじめとする視聴体験を向上できるか評価し，ベースラインモデルに勝る

[1] 東京大学，Hongo，Bunkyo-ku，Tokyo，113-8656，Japan.
[2] 国立研究開発法人産業技術総合研究所，Aomi，Koto-ku，Tokyo，135-0064，Japan.
[3] 慶應義塾大学，Hiyoshi，Kohoku-ku，Yokohama，223-8522 Japan
[4] カーネギーメロン大学，Forbes Avenue，Pittsburgh，PA，15213，USA.
[5] 奈良女子大学，Kitauoya Nishimachi，Nara，Nara 630-8506，Japan.



(a) 提案法のイメージ図（TTS 訓練時）

(b) 提案法のイメージ図（推論時）

図 1: 動画シーンに合わせた盛り上がり制御可能な実況解説 TTS システム

盛り上がりの評価を得ることを確認した．

2. 関連研究

2.1 実況解説システムに関する研究

実況生成に関する研究として，スポーツ実況の分野では発話生成の研究が多く行われている [6], [7]．また，リアルタイムで発話生成から音声合成までを行う実況解説システムも提案されている [8], [9]．しかし，これらの研究はテキスト生成に重きを置いており，シーンに合わせた発話スタイルや，音声の表現力による盛り上がりの向上などは考慮されていない．

感情音声合成の先行研究として，話者の意図ではなく聴衆の知覚に基づくラベルを用いることで，正確な感情制御を行えるとする研究がある [10]．そこで本研究では，視聴者が感じる盛り上がりをもとにラベルを作成し，盛り上がりのある音声表現によって視聴者を盛り上げ楽しませることができるシステムの構築を目指す．

2.2 SMASH コーパス

SMASH コーパス [5] は，大乱闘スマッシュブラザーズ SPECIAL (SSBU) という有名な対戦アクションゲームに対する実況解説を収録したコーパスである．約 2 分 30 秒の試合が 69 個収録されており，2 名の話者 (MC1, MC2) がそれぞれ 2.5 時間と 1 時間実況解説を行っている．試合の中で pitch，energy にピークがあり，またキャラクター名や固有名詞などが用いられている特徴がある [11]．本研究ではこのコーパスを用いて実況解説音声を合成することを試みる．

3. 盛り上がり制御可能な TTS

ここでは，提案する盛り上がり制御可能な TTS について説明する．まず，盛り上がりと音声特徴量との調査と結果について述べ，続けて盛り上がり制御を行うためのラベル作成方法について説明する．その後，提案する TTS モデルの概要について説明する．

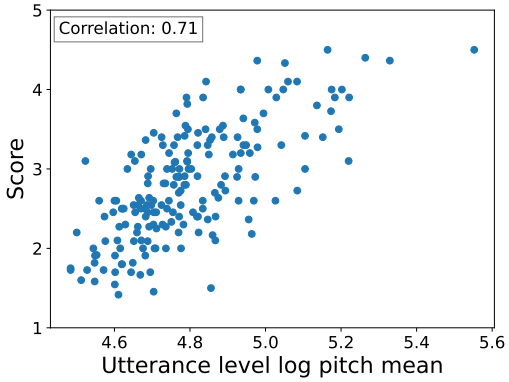
3.1 実況解説音声の盛り上がりスコアの収集

盛り上がりの制御を可能とする TTS を実現するため，最初に盛り上がりと音声特徴量との関連を調査する．SMASH コーパス [5] の音声に対し，盛り上がりを評価する主観評価実験を行い，音声特徴量との関連を分析する．SHASH コーパスの中から，6 試合分となる合計約 18 分，191 発話の音声を選択し，評価対象の音声とした．評価者は 100 名で，1 人当たり 20 個の音声を評価した．評価は 5 段階（1：全く盛り上がっていない ～ 5：非常に盛り上がっている）で，クラウドソーシングサービスであるランサーズ[*1]を用いて行った．

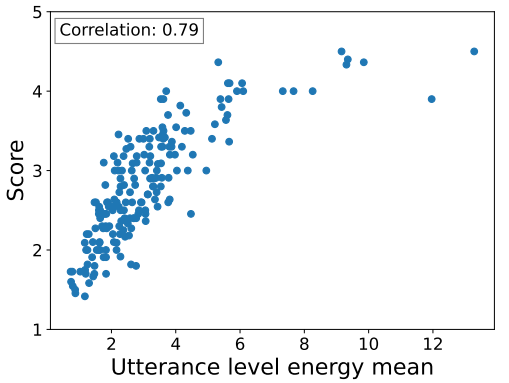
3.2 盛り上がりスコアの分析

集めた評価結果について，opensmile[*2]と呼ばれる音声分析ツールを用いて分析を行った．特徴量セットとして emobase プリセットを利用し，各音声から特徴量を抽出した．抽出された特徴量を用いて，各音声に対する盛り上がり評価スコアの平均値をランダムフォレストによる予測を行う形で学習し，予測に特に用いられた特徴量を確認した．ランダムフォレストのハイパーパラメータは，max depth: 5，min samples split: 3，n estimators: 20 とした．特に予測に重要視された上位 5 つの特徴量は順に，pcm loudness sma linregerrQ，pcm loudness sma de linregerrQ，pcm loudness sma stddev，IspFreq sma de[7] linregrerrQ，pcm loudness sma linregerrA となった．上位 5 つの内 4 つは loudness という声の大きさを表すものに関係した特徴量であり，盛り上がり評価には声の大きさが大きく影響していることがわかる，また残り 1 つの IspFreq sma de[7] linregrerrQ は，音声信号スペクトルの特定の周波数帯（この場合は 7 番目）の時間的な変化をモデル化

[*1] https://www.lancers.jp/
[*2] https://github.com/audeering/opensmile



(a) pitch と盛り上がり評価のプロット．



(b) energy と盛り上がり評価のプロット．

図 2: SMASH コーパスの pitch, energy と盛り上がりのプロット．

し，変化がどれだけ不規則かを表す値であり，周波数が急激に変化することで大きくなる．よって，発話中で声の高さ (pitch) が急激に変化していると予想されるため，盛り上がりには pitch も影響を与えていると考えられる．

3.3 音声韻律特徴量と盛り上がりスコアの相関分析

SMASH コーパスの各発話の声の高さ (pitch)，声の大きさ (energy) を計算し，この 2 つを発話単位で平均した値と，各音声の平均盛り上がりスコアをプロットした結果を図 2 に示す．Pitch の計算は WORLD [12] (D4C edition [13]) を用いて基本周波数を抽出する形で行なった．音声のサンプリングレート (sr) は 22050 Hz，frame period は 256 × 1000/sr とした．Energy の計算は，まず無音区間を除去するため，librosa[*3]ライブラリの split 関数を用い，音圧レベルが 40 dB 以下の区間を除外した．その後，フレームサイズ 1024，シフト幅 256，窓長 1024 の短時間フーリエ変換を行い，振幅スペクトルを計算した．振幅スペクトルの各時間フレームごとにノルムを計算することで，音声の大きさを相対的に表す energy を計算した．図 2 から分かるとおり，どちらも盛り上がりスコアと強い相関があり（それぞれ相関係数が 0.71, 0.79），音声の盛り上がりを再現するには，この 2 つを重点的に再現する必要があると確認できる．

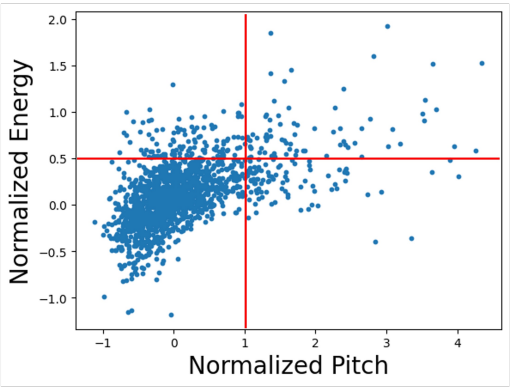


図 3: SMASH コーパスの各発話の pitch, energy を normalize し，その平均値をプロットしたもの．赤い線は low/high ラベル作成のために設定した pitch, energy の閾値を示す．

3.4 音声韻律特徴量に基づく盛り上がりラベル作成

盛り上がりラベルの作成は，pitch，energy それぞれに一定の閾値を設定し，そのどちらも超えた場合を high，そうでない場合を low として作成する．ラベル作成の閾値と pitch，energy をプロットした様子を図 3 に示す．図 3 の pitch，energy は平均 0 分散 1 に変換している．Pitch の閾値は 1.0, energy の閾値は 0.50 と定め，そのどちらも超えた音声を high とする．この閾値により，SMASH コーパスの話者 MC1 の全 1,611 発話のうち，73 発話が high，1,538 発話が low に分類された．

3.5 ラベルを用いて盛り上がりを制御する TTS モデル

ベースとして使用するモデルは，VITS [14], JETS [15] の 2 種類を用いる．それぞれのモデルでラベルを挿入する位置として，VITS は，PromptStyle [4] での style embedding を挿入した箇所を参考にし，JETS は ESPnet[*4]で実装されている speaker embedding を参考にする．VITS，JETS のどちらの場合も，盛り上がりラベル low/high を ID とし，学習可能な埋め込み表現に変換した後，各モデルの潜在表現に足し合わせることで条件付けを行う．また，VITS，JETS 共に実装は ESPnet を参考にする．

4. 動画からの盛り上がりラベル予測

ここでは，提案 TTS モデルにおいて動画シーンから盛り上がりを制御するために，音声から作成した盛り上がりラベルを動画から予測するモデルについて説明する．まず，対戦動画に対する主観的な盛り上がりを判断するため，動画シーンに対する盛り上がりアノテーションを行う．その後，動画から盛り上がりを判断する方法として，(1) フレーム画像から計算される動画指標を用いたシーンの盛り上がりを判断する手法と，(2) 動画を入力とし，その動画の盛り上がりスコアを予測する DNN ベースの手法の 2 つを提案する．

4.1 動画のみを見た盛り上がりアノテーション

動画のどのようなシーンが盛り上がっていると判断され

[*3] https://librosa.org/doc/latest/index.html

[*4] https://github.com/espnet/espnet

るのかを確かめるため，SMASH コーパスの映像のみを見せ，そのシーンが盛り上がっているかどうかを 5 段階で評価する実験を行うことで，盛り上がりに関するアノテーションを実施する．評価対象の動画は，SMASH コーパスに収録されている動画から用意した．音声との関連も調査するため，SMASH コーパスの発話音声のタイムスタンプに従い，動画も発話音声の分割と同様に分割した．ただし，実況解説音声は動画を見て，その見たシーンに対して実況解説を行うため，言及している内容と映像のシーンにずれがあると考えられる．そこで，分割する動画は音声のタイムスタンプから 2 秒早く切り取るように設定した．分割された動画は発話数と同じで合計 1,611 個である．ゲーム音や実況解説音声，プレイヤー音声を全て消し，映像のみを見てそのシーンが盛り上がるかどうかを評価した．評価は全てクラウドソーシングサービスであるランサーズで行った．評価者は合計 780 名であり，一人当たり 21 個の動画を見て，5 段階（1：全く盛り上がっていない 〜 5：非常に盛り上がっている）で評価した．一つの動画あたり，少なくとも 10 名の評価が集まった．

4.2 動画由来の盛り上がりラベル分析

集めた盛り上がりラベルがどのようなシーンで高い，あるいは低い評価を受けているか確認するための分析を行った．分割された各動画についた 5 段階の盛り上がりスコアの平均値を計算し，各動画の盛り上がりスコアとした．アノテーションスコア全体の平均値は 3.32，中央値は 3.40 であり，比較的盛り上がった評価を受けた動画が多いことがわかった．これは，対戦ゲームには様々なエフェクトがあるという特徴から，盛り上がっている評価を受けやすいものであると考えられる．また，評価が特に高かった動画（スコア 4.5 以上）は全て戦闘中の画面であり，「最後の切り札」と呼ばれる必殺技を打つシーンや，撃墜のシーン，試合が決着するシーンなど，試合内容に大きな影響を与えつつ，画面の変化としても大きいシーンが多かった．一方，特に評価が低かった動画（スコア 1.8 以下）は試合開始画面やリザルト画面などの対戦中ではない，画面がほとんど変化していないシーンが多かった．映像のみをみた場合の盛り上がりの判断は，試合内容への影響や，画面全体の動きの大きさが影響を与えていると考えられる．

4.3 動画指標を用いた盛り上がり予測

動画由来の盛り上がりラベルの分析から，盛り上がるシーンでは画面の変化が大きく，シーンとして激しいものであると予想される．そこで，動画を構成するフレーム画像から計算できる指標を 2 つ検討し，画面の変化や盛り上がっている場面の特徴を数値として表現することで，盛り上がっているかどうかの判断を行うことを試みる．

Frame Difference (FD)：画面の変化が大きいシーンでは，フレーム間のピクセル変化が大きいと考えられる．そこで，前のフレームからのピクセルの変化量をフレーム画像全体で計算し，その平均を取ることで画面の動きの激しさを数値化できると考える．動画フレームのうち，前から n 番目のフレームを I_n と表すと，I_n は縦 h，横 w，カラーチャネル c の 3 次元テンソルで表現できる．よって，n 番目のフレームの変化量を表す指標 D_n は，

$$D_n = \frac{1}{h \cdot w \cdot 3} \|I_n - I_{n-1}\|_1$$

で表される．ただし，$\|\cdot\|_1$ は要素ごとの絶対値の総和を意味する．D_n を用いることで，変化が激しいシーンを数値として表現できると考えられる．本研究では，この D_n を Frame Difference (FD) と呼ぶことにする．

Laplacian Variance (LV)：盛り上がっていると判断されるシーンでは，撃墜の爆風や最後の切り札での攻撃など，特徴的なエフェクトが多く用いられている．そこで，画像のシャープネスさを表し，画像がボケているかどうかの判断に用いられる Laplacian variance (LV) [16] を用い，盛り上がったシーンの判断を行うこととする．LV は次の手順で求められる．まず，モノクロ画像を $I(m,n)$ と置く．ここで：

- $I(m,n)$：画像のピクセル（位置 (m,n) における輝度値）
- $m = 1, 2, \ldots, M$：画像の行インデックス（高さ方向）
- $n = 1, 2, \ldots, N$：画像の列インデックス（幅方向）
- $M \times N$：画像のサイズ（ピクセル数）

である．LV で用いられる Laplacian 演算は，次式で表される：

$$L(m,n) = (I * K)(m,n)$$

$L(m,n)$ は Laplacian 演算後の値であり，$I * K$ は画像 I とカーネル K の畳み込み演算である．ここで，Laplacian カーネル K は：

$$K = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

である．$L(m,n)$ に対し，画像全体での分散を以下のように求める：

$$LV(I) = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(|L(m,n)| - \bar{L} \right)^2$$

ただし，$\bar{L}$ は Laplacian 絶対値の平均を表し：

$$\bar{L} = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} |L(m,n)|$$

である．フォーカスがあった画像ではエッジが多く，$LV(I)$ が大きくなり，フォーカスがあっていないボケた画像では，$LV(I)$ が小さくなる．ゲーム画像では，エフェクトが多いシーンでエッジが多くなると考えらるため，盛り上がりシーンの検出が可能になると考えられる．

4.4 DNN ベースの盛り上がり予測

画像フレームから計算される動画指標は，簡単・高速に計算できる一方，複雑な対戦ゲーム映像の特徴を細かく分

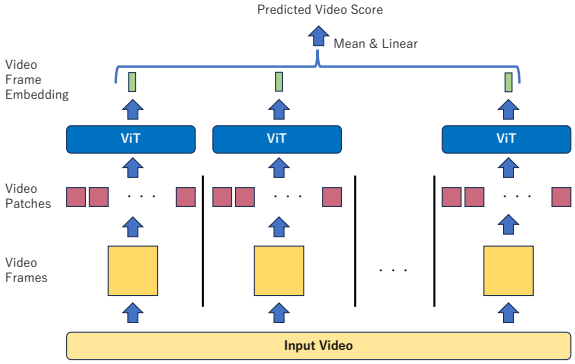


図 4: ViT で盛り上がりスコアを予測する手法の概要.

析することはできない．そこで，DNN ベースの手法である Vision Transformer (ViT) [17] を用い，動画盛り上がりアノテーションで集めたスコアを予測する学習を行うことで，シーンの盛り上がり予測を行う方法も検討する．

ViT を用いて盛り上がりを予測する手法の概要を図 4 に示す．各フレーム画像を patch に分割した後，1 つ目の手法のように patch 全体を ViT に入力するのではなく，各フレームごとに ViT に入力し，フレーム画像の潜在表現を得る．その後，各フレームの潜在表現を平均して動画潜在表現とみなし，線形層を通してスコアを予測する．

5. 実験

本研究で提案した手法が，盛り上がる実況解説を行えるのか確認するための主観評価実験を行った．まず，TTS 部分について，提案するラベル付き学習による盛り上がり制御が可能かを確認するため，音声のみを対象とした実験を行い，音声の自然性と盛り上がりについて評価を行った．続いて，動画による制御を合わせた提案システム全体が，視聴者の盛り上がりや楽しさを向上できるか確認するため，数発話で構成される実況解説付き動画を対象とした実験を行い，盛り上がりや面白さを感じられるかを評価した．

5.1 実験条件

実験に用いるコーパスとして，SMASH コーパス [5] を用いた．実況解説者として MC1，MC2 の 2 名が存在するが，本研究は単一話者 TTS を想定するため，より収録時間の長い，MC1 による合計約 2.5 時間のデータを用いた．合計 52 試合，1,611 発話のデータを train，validation，test データとしてそれぞれ 40，6，6 試合 (1,234，186，191 発話) に分割した．

TTS モデルについて，提案法である low/high ラベルを用いた学習の有効性を調べるため，low/high ラベルを用いず学習させたモデルをベースラインとして用意した．基本設計は ESPnet に則っている．VITS は，JSUT [18] による事前学習済みモデル[*5]を SMASH コーパスで fine-tune する形で学習した．バッチサイズは 20 であり，100 エポック学習した中でスコアが良かった上位 10 個のパラメータを平均したものを使用した．JETS は ESPnet に JSUT による学習済みモデルが存在しなかったため，まず JSUT による学習を行った．バッチサイズは 20 とし，400 エポック学習した中でスコアが良かった上位 5 個のパラメータを平均したものを使用した．その後 SMASH コーパスを用いた fine-tune を行った．バッチサイズは 20 とし，300 エポック学習した中でスコアが良かった上位 5 個のパラメータを平均したものを使用した．どのモデルでも，optimizer として AdamW [19] を用い，学習率は 10^{-4} とし，$\beta_1 = 0.8$，$\beta_2 = 0.99$ とした．VITS，JETS のその他基本的なパラメータは，ESPnet の実装に準じた．

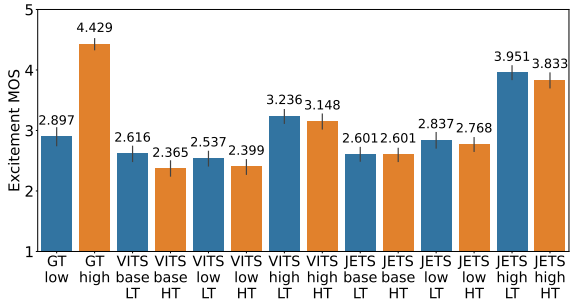
盛り上がり予測を行うモデルとして，動画指標から予測するモデル (VF) と，ViT を用いた予測モデルの 2 つを用意した．VF は，入力する動画の各フレームで FD と LV を計算し，計算された FD，LV を入力動画全体で平均し，この平均値がどちらかの指標で閾値を超えた場合にその動画を high と判断するよう設計した．ただし，試合単位で FD と LV を求め，各試合ごとに平均を 0，分散を 1 とするよう変換を行い，試合の違いに影響されず統一の閾値で盛り上がりの判断ができるようにした．盛り上がりの有無を決める閾値は，FD，LV どちらも 0.5 とした．ViT による盛り上がりアノテーション予測は，torchvison ライブラリ (version 0.17.2) の ViT_B_16 を用いた．また重みの初期値は事前学習済みモデルとして IMAGENET1K_V1 を使用した．画像サイズは 224，各フレームの patch size は 16，hidden size は 256，encoder の layer 数は 12，head 数は 12 とした．学習データとして SMASH コーパスの動画を使用し，train，validation，test データの分割は TTS 学習での分割と対応させた．Optimizer として AdamW を用い，学習率は 10^{-6}，$\beta_1 = 0.9$，$\beta_2 = 0.999$ とし，100 エポック学習した．

5.2 TTS モデルの盛り上がりを評価する音声評価実験

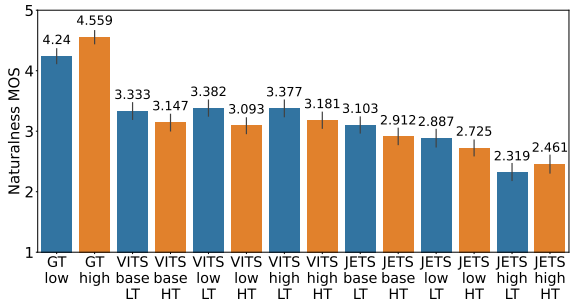
提案するラベルによって盛り上がりを制御するモデルが盛り上がりのある音声を合成できるかを確認するため，音声のみを対象とする主観評価実験をおこなった．評価は全てクラウドソーシングサービスであるランサーズで行った．評価者は合計 200 名で実施された．評価者は同一テキストの音声をランダムな順番で聞き，その音声が盛り上がっているかどうかを 5 段階（1：全く盛り上がっていない ～ 5：非常に盛り上がっている），音声の自然性（人間が話すような自然な音声に聞こえるかどうか）を 5 段階（1：非常に悪い ～ 5：非常に良い）で評価した．手法間の比較のため，有意水準 0.05 の多重検定（Steel-Dwass 法）を行った．

まず，図 5a に盛り上がりに関する主観評価の結果を示す．正解音声が low のテキスト (LT) と high のテキスト (HT) でグラフを分けている．JETS，VITS のどちらのモデルでも，ラベルによる盛り上がりの制御ができているとわかる．特に JETS の方がより顕著に盛り上がりの制御が

[*5] https://zenodo.org/records/5521360



(a) 盛り上がりに関する主観評価の結果.



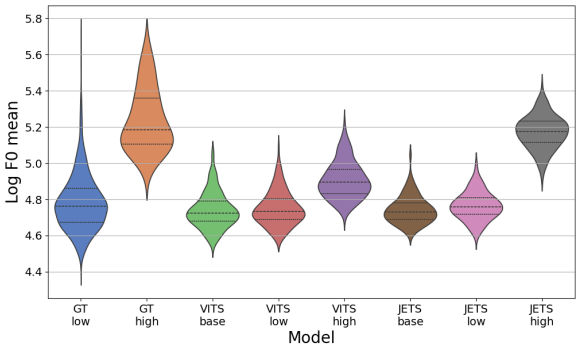
(b) 自然性に関する主観評価の結果.

図 5: TTS に関する主観評価結果．LT，HT はそれぞれ自然音声のテキストが low/high のどちらなのかを表し，LT が low，HT が high である．

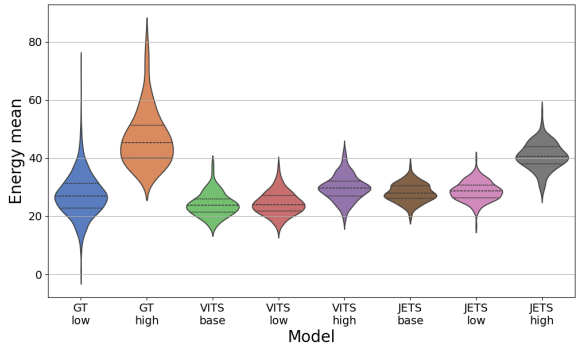
できている．また，ベースラインモデルは VITS，JETS のどちらでも low ラベルと同程度の結果になっており，提案手法であるラベルを与えた学習が盛り上がりのある音声を合成するために有効な手段であるとわかる．

次に，自然性の評価結果を図 5b に示す．VITS，JETS どちらのモデルも，自然音声と比較すると自然性は有意に低かった．これは，学習データが約 2 時間と少なく，SMASH コーパスは SSBU の実況解説音声という特徴上，話し言葉が多く使われており，キャラクター名や技名などの固有名詞が多いため，合成音声のアクセントが不自然になり自然性が低くなってしまったと考えられる．また，提案手法であるラベル付き学習は，VITS は low，high 共にベースラインと同程度の自然性である一方，JETS は high ラベルを与えたものが特にベースラインより自然性が低い結果になった．これは，JETS の high ラベルによる合成は，pitch と energy を大きくできる一方で，高音部分が掠れたように聞こえることがあり，自然性が低く判断されてしまったと考えられる．

Pitch，energy 由来のラベルによる制御性の確認として，自然音声と合成音声の pitch，energy の可視化を行った．その結果を図 6 に示す．VITS，JETS のどちらでも high ラベルのほうが値が高くなっており，特に JETS のほうが差が顕著であり，自然音声により近い結果だった．これは主観評価の盛り上がりの評価にも一致しており，ラベルによる制御が有効に働いていると分かる．



(a) 各音声の pitch の可視化.



(b) 各音声の energy の可視化.

図 6: 各音声の pitch，energy を可視化した結果.

5.3 実況解説システム全体の実験と評価

続いて，動画から盛り上がりラベルの制御を行い，シーンに合わせた実況解説を行うシステムが実際に盛り上がりや面白さを向上できるのかを確認する評価実験をおこなった．評価対象は，(1) 自然音声 (GT)，(2) low/high ラベルを用いず学習したベースライン (baseline)，(3), (4) 該当シーンの発話を全て high あるいは low にしたもの (all high, all low)，(5) 自然音声の low/high ラベルに従う制御 (GT label)，(6) ViT の正解に当たる盛り上がりアノテーションの値による制御 (annotation)，(7), (8) DNN ベースの盛り上がり予測で閾値を 2 種類用意 (ViT 3.5, ViT 3.8)，(9) 動画指標による予測 (VF) の 9 通りとした．自然音声以外は TTS モデルの違いとして VITS，JETS の区別があるため，合計 17 手法を評価した．

評価用のシーンとして，合計 21 個用意した．シーンの選択基準として，正解データ由来の制御方法である GT label，annotation のどちらかで high が必ず含まれるシーンを切り出した．映像は音声の開始 1.5 秒前から切り抜き，終了時刻はどの合成音声でも時間内に収まるよう，自然音声の終了時刻の 1.5 秒後までを切り抜いた．各発話の開始時刻は既知とし，SMASH コーパスの自然音声の開始時刻を合成音声でも開始時刻とした．合成音声の長さにより，次の発話の開始時刻までに発話が終了しない場合は，強制的に次の発話を開始するように設計した．映像と音声のミックスは ffmpeg[*6] を用いた．ゲーム音は無音であり，実況解説

[*6] https://www.ffmpeg.org

のみが聞こえるようにした．作成した動画の長さは短いもので 10 秒，長いもので 33 秒だった．

評価は全てクラウドソーシングサービスであるランサーズで行った．評価者は合計 120 名で，評価者はランダムに選ばれた 1 つのシーンに対し，17 手法をランダムな順番で視聴しそれぞれを評価した．評価項目は次の 5 つとした：

- Q1. 実況解説音声も含め，この動画がどれだけ盛り上がっているか答えてください．
- Q2. 実況解説音声も含め，この動画がどれだけ面白かったか答えてください．
- Q3. 実況解説音声の自然性（人間らしく，自然に話しているかどうか）を評価してください．
- Q4. 実況解説音声が，動画をさらに盛り上げるような話し方ができていたか評価してください．
- Q5. 実況解説音声の盛り上がったような話し方が，映像の盛り上がるシーンと合っているか評価してください．

これらの評価項目を，一つの実況解説付き映像を見るたびに 5 段階で評価した．手法間の比較のため，有意水準 0.05 の多重検定（Steel-Dwass 法）を行った．

Q1 の評価結果を図 7a に示す．提案法の中で baseline と比較して，VITS 同士では all high，ViT 3.5 が有意に高く，JETS 同士では all high，ViT 3.5，VF が有意に高い結果だった．これは，提案法による実況解説システムが既存の TTS と比較してより盛り上がる実況解説を行えることを示している．また，JETS all high は GT よりも有意に高い結果を示しており，提案法が盛り上がりのある実況解説システムとして有効な手法であると言える．

Q2 の評価結果を図 7b に示す．提案法の中で，baseline と比較して VITS 同士では all high が，JETS 同士では all high，ViT 3.5，VF が有意に高い結果であった．提案法による実況解説音声が，既存の TTS よりも視聴者の楽しみを向上させることができたと言える．

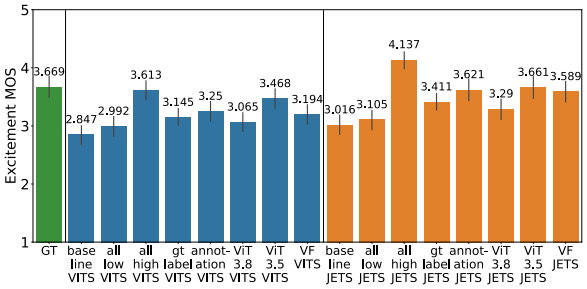
Q3 の評価結果を図 7c に示す．合成音声は GT と比較して低い結果であり，JETS の自然性が特に低いという結果だった．これは，音声のみの評価と一致している．

Q4 の評価結果を図 7d に示す．All high JETS が有意に高い結果であり，Q1 の結果と一致した．また，baseline との比較では，VITS は all high，ViT 3.5 が有意に高く，JETS は all high，ViT 3.5，VF が有意に高いという結果になった．

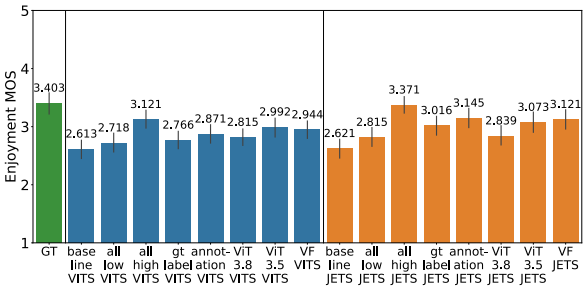
Q5 の評価結果を図 7e に示す．Baseline と比較して VITS は all high が有意に高く，JETS は all high，ViT 3.5，VF が有意に高い結果となり，提案法の中では all high JETS が最も高い評価であった．

6. 考察

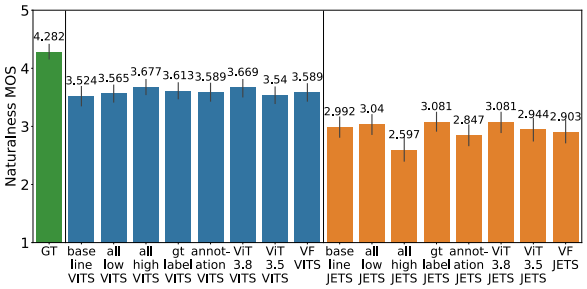
提案手法の中で，all high が最も良い結果であり，ま



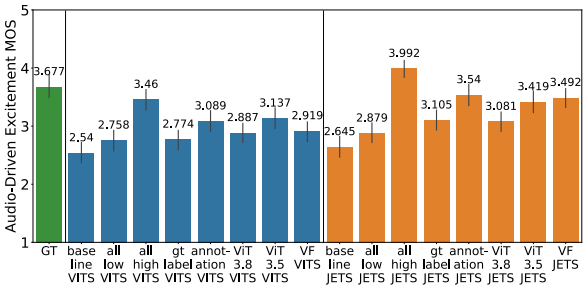
(a) Q1 の評価結果.



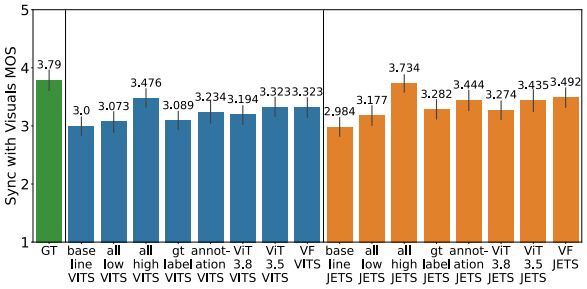
(b) Q2 の評価結果.



(c) Q3 の評価結果.



(d) Q4 の評価結果.



(e) Q5 の評価結果.

図 7: Q1 ～ Q5 の評価結果を棒グラフで表したもの．エラーバーは 95% 信頼区間を表す．

た ViT による予測で閾値が低い方が良い結果であったことから，high ラベルが多くなるほど評価が良くなると考えられ，実際 high ラベルの数が多い順に並べると，**all high** > **ViT 3.5** > **annotation** > **VF** > **GT label** > **ViT 3.8** > **all low** となり，この並びは，Q1，Q2 の評価の高い順と一致していることから，基本的に high ラベル数が多いほど，盛り上がり (Q1)，面白さ (Q2) の値が大きくなっていることが分かる．つまり，今回のタスクである SSBU に対する実況解説は，シーンの盛り上がりに関わらず high に分類される話し方を行うことが，視聴者の盛り上がりや面白さの向上につながり，逆に low に分類される話し方は盛り上がりを損なうためできるだけ減らすことが求められると考えられる．

この理由として，SMASH コーパスの音声は実際に SSBU の大会などで実況解説を行っているプロではないことが挙げられる．対戦ゲーム実況に限らず，実況解説はシーンに合わせて適切な内容を適切なスタイルで話す技術が求められるが，視聴者の盛り上がり・面白さを向上・維持するため，盛り上がりに欠ける話し方を避けていると予想される．この観点において SMASH コーパスの音声の low/high の分類で大部分を low が占めているのは，盛り上げる必要がある実況解説において盛り上がりに欠ける実況解説を行ってしまっており，スタイル制御が不十分だった可能性がある．この点において本研究の提案手法，特に all high は，非プロ実況解説者の発話の中で，盛り上がりのある発話ができている部分を抽出して学習し，合成時に実況解説として盛り上がりのある音声のみにすることで，より適切な実況解説音声を実現できていると考えられる．

7. まとめ

本研究では，対戦ゲーム実況解説音声合成システムを提案した．TTS が盛り上がりを制御できることを確認し，視聴者に盛り上がりや面白さを感じてもらえることを確認した．今後はより良い実況解説合成のため．プロ実況解説者の音声収録や自然性の改善方法の検討を実施する．

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．本研究には、内閣府が実施する「研究開発成果の社会実装への橋渡しプログラム（BRIDGE）/AI×ロボット・サービス分野の実践的グローバル研究」により得られた成果が含まれています．

参考文献

[1] Y. Wang et al., “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *Proc. International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[2] S.-Y. Um et al., “Emotional speech synthesis with rich and granularized control,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 7254–7258.

[3] Z. Guo et al., “Prompttts: Controllable text-to-speech with text descriptions,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[4] G. Liu et al., “Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions,” in *Proc. Interspeech*, 2023.

[5] Y. Saito et al., “SMASH corpus: A spontaneous speech corpus recording third-person audio commentaries on gameplay,” in *Proc. Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6571–6577.

[6] B. J. Kim, Y. S. Choi, “Automatic baseball commentary generation using deep learning,” in *Proc. ACM SAC*, 2020, p. 1056–1065.

[7] J. Rao et al., “Matchtime: Towards automatic soccer game commentary generation,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[8] T. Kumano et al., “Generation of automated sports commentary from live sports data.” in *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, H. Sun, Ed. IEEE, 2019, pp. 1–4.

[9] T. Ishigaki et al., “Audio commentary system for real-time racing game play,” in *Proc. Proceedings of the 16th International Natural Language Generation Conference: System Demonstrations*, C. M. Keet et al., Eds. Prague, Czechia: Association for Computational Linguistics, Sep. 2023, pp. 9–10.

[10] J. Lorenzo-Trueba et al., “Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis,” *Speech Communication*, vol. 99, pp. 135–143, 2018.

[11] 井浦昂太 et al., “対戦ゲーム動画の実況解説音声の分析と合成の検討,” in **日本音響学会秋季研究発表会**, 2023.

[12] M. Morise et al., “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[13] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, 2016.

[14] J. Kim et al., “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proc. International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[15] D. Lim et al., “JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech,” in *Proc. Interspeech*, 2022, pp. 21–25.

[16] J. Pech-Pacheco et al., “Diatom autofocusing in brightfield microscopy: A comparative study,” in *Proceedings 15th International Conference on Pattern Recognition*, vol. 3, 2000, pp. 314–317.

[17] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. International Conference on Learning Representations*, 2021.

[18] S. Takamichi et al., “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[19] I. Loshchilov, F. Hutter, “Decoupled weight decay regularization,” in *Proc. International Conference on Learning Representations*, 2019.