TEPPAY: ゲームのプレイ動画を入力とする実況 AI Tuber システムの提案

栗原健太郎 [1,2] 吉野哲平 [1] 高市暁広 [1] 岩田伸治 [1]
長澤春希 [1,2] 佐藤志貴 [1] 岩崎祐貴 [1]
[1] サイバーエージェント [2] 株式会社 AI Shift
{kurihara_kentaro,yoshino_teppei,
takaichi_akihiro,iwata_shinji,nagasawa_haruki,
sato_shiki,iwazaki_yuki}@cyberagent.co.jp

概要

YouTuber や VTuber などの動画配信への関心が高まっている。しかし、動画配信を支援するソフトウェアの複雑さや、実況能力への不安から動画配信の参入障壁は高い。本研究では、大規模言語モデル等を活用して気軽な実況配信を実現するシステム: TEPPAY を提案する。TEPPAY はシーン検出や発話生成などを行う 7 つのモジュールで構成されている。TEPPAY を構成するモジュールの性能調査の結果、配信に必要な最低限の性能は担保しているものの、視聴者の興味を惹く魅力的な実況動画を作成する上でいくつか課題があることを確認した。

1 はじめに

YouTube[1] などの動画配信サービスを用いて、任意のコンテンツに関する解説や実況を配信する人（YouTuber）の認知が広がっている。特に、自身ではなく 3D モデルのキャラクターなどの仮想的なキャラクターによる解説・実況を行う VTuber（Virtual YouTuber）は、日本発祥のコンテンツ文化として広く普及しつつある [1]。これらの普及の背景には、配信コンテンツや配信者そのものへの人気のみならず、誰もが配信活動を行うことができるという点に関心が集まっていることが挙げられる。

配信者は、OBS Studio[2] などのソフトウェアを用いることで容易に動画配信を行うことができる。また、VTuber における 2D・3D モデルの作成や、音声合成・変換による任意のスタイルの音声作成を支援するソフトウェアもこれらの活動の普及を加速させている。しかし、前述した目的ごとに複数のソフトウェアを同時に扱う必要があるため、準備にかかる労力や、事前知識のない人にとっての参入障壁は依然として高い。また、インターネット上でのコミュニティとのつながり獲得を目的として、自身のゲームプレイを配信しようと考えたことがある人の中には、実況の能力への不安ゆえに参入できない人もいる。一方で、ChatGPT[3] などの大規模言語モデルを用いたサービスの台頭によって、配信へのコメントに対する応答の生成など、より柔軟性の求められる発話の実現可能性も高まりつつある [2]。

これらのソフトウェアやサービスの登場、および参入障壁に関する課題を背景として、本論文では気軽な配信を実現するシステムとして「動画像を入力とし、VTuber による実況を生成するシステム: TEPPAY (The Enhanced Product Playing AI YouTuber)」を提案する。ゲームのプレイ自体は人が行い、アバターの準備や実況中の表情変化、あるいはゲームの画面に基づいた実況の発話などを TEPPAY によって自動化することを目的としている。TEPPAY は、動画像のシーン検出やシーンからの状態抽出などの複数のモジュールから構成される。本研究では、配信する動画コンテンツとして人気のあるカテゴリの 1 つである「ゲーム配信」のうち、特に「対戦型格闘ゲーム」のプレイ動画を題材として、TEPPAY の各モジュールの役割や課題について整理し、現状のシステムの性能調査、および改良に向けた議論を行う。TEPPAY の各モジュールの性能について実験した結果、必要最低限の性能は担保されているものの、視聴者の興味を惹く魅力的な実況の提供を実現する上での課題が存在することを確認した。

1） https://youtube.com
2） https://obsproject.com
3） https://openai.com/chatgpt

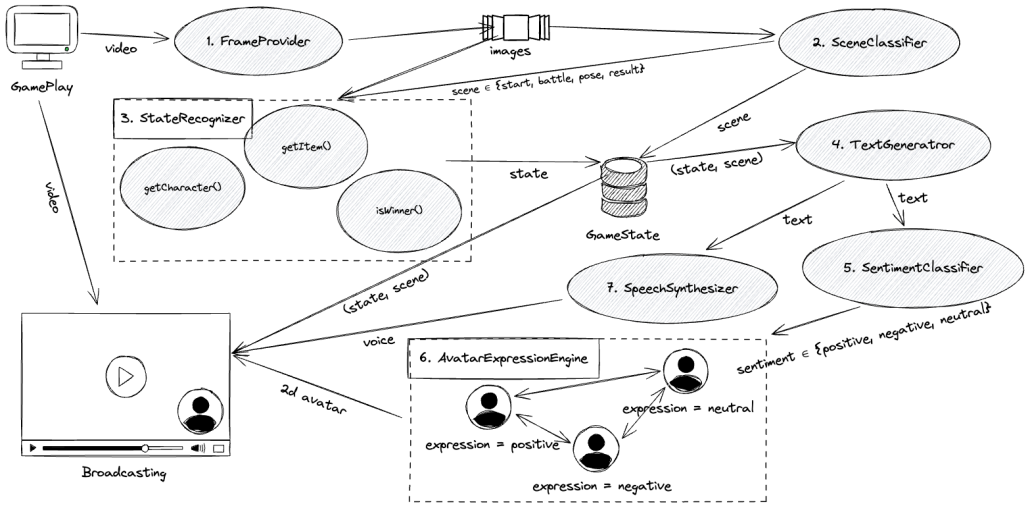

図 1　TEPPAY の構成

2　関連研究

2.1　VTuber

VTuber は仮想アバターを通じて、リアルタイム性を持つ動画配信等を行うコンテンツである。配信活動としては歌・チャット・ゲームの 3 つがよく見られる [3]。プレイしているゲームなど自分との共通部分があることや [4]、アバターの外見やキャラクター設定に視聴者は興味を惹かれている [3, 4]。

VTuber が広く注目を集めている一方、VTuber として配信を始める障壁は高い。具体的には、ゲームのプレイと実況の同時進行や、ゲームの内容に即しつつ視聴者の興味を惹く話をすることへの技量的な不安が障壁となると考えられる。そのため、ゲームのプレイのみに集中できる配信システムが求められている。例えば、アバターアニメーションの自動表出 [5] や、配信中に反応するコメントの自動選択 [6] がその構成要素として挙げられる。

2.2　ゲーム実況生成

リアルタイムで発生・放送される事象やコンテンツに対する実況テキストや実況音声の自動生成は、スポーツ実況を中心に試みられるようになった [7, 8, 9]。一方、昨今の e スポーツやゲーム実況配信に対する関心の高まりを背景として、ゲーム映像等を入力としたゲーム実況テキスト・音声の自動生成も盛んに取り組まれるようになった [10, 11, 12]。特に近年は、大規模言語モデルの利用によって高品質な実況が生成可能となってきている [12]。

本研究では、大規模言語モデルを用いた高品質なゲーム実況の自動生成と、2.1 節で例示した配信支援技術を組み合わせることで、ゲームのプレイのみを入力として VTuber によるゲーム実況を生成するシステムを構築する。

3　TEPPAY

TEPPAY のシステム構成を図 1 に示す。TEPPAY は動画を入力として、VTuber による実況動画を生成するシステムで、以下の 7 つのモジュールで構成する。本研究では、対戦型格闘ゲーム（対戦ゲーム）[4] 実況において本システムを使用することを想定して、各モジュールの概要と技術要素、および現状の課題について述べる。

1. Frame Provider　ゲームのプレイ動画からの実況発話生成は、入力情報の大きさゆえに難しい。そこで、より軽量な入力情報に変換するためのモジュールとして Frame Provider を構築する。Frame Provider は動画から定数秒ごとにフレームを抽出して後続処理にフレームを受け渡す。

2. Scene Classifier　対戦ゲームの実況において、その発話内容や発話のトーンなどはゲームのシーンによって大きく異なることが想定される。例えば、戦闘シーンでは戦闘状況に応じてポジティブ、あるいはネガティブな発話が増え、キャラクター選択などの変化の少ないシーンではニュートラルな発話が主となると考えられる。Scene Classifier は、キャラクター選択画面・戦闘画面などの画面種類を識別し、識別結果を発話生成や音声合成などに用いる

4）　プレイヤーとコンピュータあるいはプレイヤー同士が操作するキャラクターが戦う、アクションゲームの一種。

ことを想定する。Frame Provider から提供されるフレーム全てに対して識別を実行するため、軽量かつ高速な処理を実現する必要がある。そこで本研究では、識別時において、判定したいシーンの見本画像を 1 件ずつ取得した上で、見本画像と Frame Provider から提供された画像をハッシュ化アルゴリズムで比較してシーンを識別した。

3. State Recognizer 対戦ゲームにおいては、プレイヤーを楽しませる要素の 1 つとして勝敗結果などの対戦成績（戦績）を残している場合がある。過去の戦績を参照した実況をすることで、より発話のバリエーションの増加が期待されるため「プレイヤー名」「戦績」を取得するモジュールとして State Recognizer を用意する。「プレイヤー名」や「戦績」はゲーム画面に対して OCR を実施することで取得する。本研究では、複数の OCR サービスを予備実験による定性評価によって比較・検討し、Paddle OCR[13] を採用した。

4. Text Generator ゲーム実況中では、ゲームと関連しない雑談もあるものの、基本的には種々のゲームシーンにある程度関連した発話が多い。Text Generator では、Scene Classifier や State Recognizer で獲得した Scene や State 情報を元に、実況発話を生成する。本研究では実況発話の生成に GPT-4o[5] を用いた。

5. Sentiment Classifier VTuber による実況において、アバターに生成内容を発話させる上では、アバターの表情やトーンも視聴者の体験価値を向上させる重要な要素となる。発話内容と整合性の取れた表情や音声のトーンを実現するために、発話から事前に定義した感情ラベルのいずれかで分類する。本研究では positive、negative、neutral の 3 ラベルで構成した。分類には GPT-4o[6] を用いた。

6. Avatar Expression Engine VTuber による実況において、アバターの表情と、ゲームのプレイ状況・実況内容との整合性を保つ必要がある。そこで、感情ラベルを入力としてラベルと一貫したアバターやキャラクター画像を取得するためのモジュールとして、Avatar Expression Engine を設ける。本研究における Avatar Expression Engine は、Sentiment Classifier によって獲得した感情ラベルと一貫した表情のキャラクターの画像を獲得するモジュールを想定する。

7. Speech Synthesizer 声のトーンが視聴者の体験価値に与える影響は大きい。感情ラベルを受け取り、VTuber の音声を合成するモジュールとして Audio Generator を設ける。本研究では、音声合成ソフトウェア VOICEVOX[7] を使用する。同一キャラクターの音声のうち、入力された感情ラベルに対応した感情パラメータを音声合成に用いることを想定する。

4 各種モジュールの性能評価

4.1 State Recognizer

State Recognizer では、Frame Provider から提供される画像を入力として、Text Generator による実況発話生成の元となるプレイヤーの情報を提供する。

実験設定 State Recognizer 内で採用している Paddle OCR の認識性能を評価した。題材とした格闘ゲーム内の「対戦相手決定シーン」と「対戦結果シーン」の 2 種類の画像を入力した場合の state の構成要素を正しく取得できるかを確認した。題材の格闘ゲーム内の「キャラクター名」とキャラクターを操作する「プレイヤー名」の 2 種類を取得した。キャラクター名の認識結果は、ゲーム内の登場キャラクターのいずれかになる。そのため、認識結果と最も編集距離の近い登場キャラクター名を出力とした。各シーンの画像は 70 件ずつ自動で収集した。OCR 結果と正解に対する一致数・編集距離を評価指標に用いた。

実験結果/考察 OCR の認識成功率を表 1 に、OCR 出力と正解テキストとの編集距離を図 2 に示す。全体的に編集距離が短く、Paddle OCR の認識性能は高かった。「キャラクター名」は、実装されているキャラクターのパターン数が限られ認識誤りが発生しなかったが（表 1）、「プレイヤー名」は任意の名前を設定できるため、「キャラクター名」と比較して認識が難しいと考えられる。一方いずれのシーンにおいても 8 割以上の事例で編集距離が 0、誤りがある事例でも編集距離は 3 以下が多いことから、プレイヤーを正しく識別できるといえる。同一 OCR モデルを使う限り同じ画像入力の推論結果は安定することから、以前に対戦したプレイヤーを異なるプレイヤーとして判定してしまうことも少なく、特定のプレイヤーに対する勝敗履歴の悪影響も起こりにくい（図 1）。

5） 実況発話の生成に用いたプロンプトは付録 A.1 に示す。
6） 発話の感情分類に用いたプロンプトを付録 A.2 に示す。
7） https://voicevox.hiroshiba.jp

表 1　対戦相手決定シーン、対戦結果のシーンのキャラクター名およびプレイヤー名の認識成功率

	対戦相手決定	対戦結果
キャラクター名	1.000	1.000
プレイヤー名	0.829	0.878

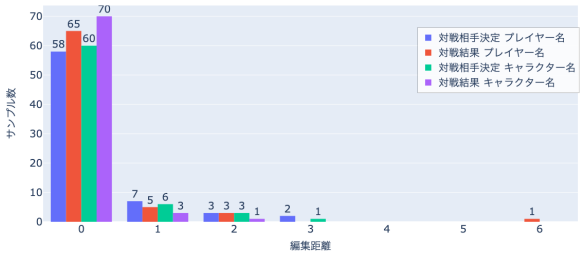

図 2　対戦相手決定時・対戦結果時のプレイヤー名およびキャラクター名の OCR 結果と正解との編集距離

4.2　Sentiment Classifier

実験設定　Sentiment Classifier 内で適用している GPT-4o の分類性能を評価した。人手で書き起こしたゲーム実況中の発話 100 件に対して人手付与した感情ラベルと、GPT-4o の分類結果を比較した。ラベルは positive、negative、neutral の 3 つで実施し、3 人の本論文著者による多数決でラベルを決定した。ただし 3 人とも異なるラベルを付与している場合は、neutral とした。GPT-4o による分類結果と人手付与した感情ラベルとの一致度で評価した。

実験結果/考察　人手評価によるラベルの分布の結果を A.3 に示す。GPT-4o 出力と人手付与したラベルの一致度合いは 72%となった。これは、感情の推定が難しい実況発話がある程度存在することを意味している。ラベルが一致した事例と不一致の事例それぞれにおける一致度として Fleiss' Kappa を算出したところ、それぞれ 0.549、0.203 と、不一致事例におけるアノテーションの一致度合いの低さが示された。本結果は、GPT-4o が推定を誤っている事例が、人にとっても感情の解釈が分かれやすい事例であると言える。GPT-4o の誤り事例を付録 A.4 に示す。GPT-4o が感情分類を誤る事例には、題材としたゲーム内固有の表現が含まれる発話、自身が優勢の場面における相手を煽る表現などがあることが確認できた。このことから、ゲーム実況におけるドメイン知識や高度な文脈を反映した分類を実施する必要性があると考えられる。

4.3　Text Generator

実験設定　モジュール内で用いる言語モデルの性能を評価した。3 種類の仮想的な state における実況発話を生成した。過去の戦績、および実況生成時の勝敗に応じてパターン分けした。[8] 各 state で 10 件ずつ、合計 60 件の実況発話を生成した。Text Generator で採用している GPT-4o の生成結果と、state の内容との間の一貫性の判断を人手により評価した。アノテータ 3 人によって、state の内容に基づいた生成結果となっていることを示す entailment、生成結果と state の内容の矛盾を示す contradiction、矛盾はしていないが state の内容と関係のない生成結果であることを示す neutral の 3 ラベルによる分類を実施し、多数決により分類結果を決定した。

実験結果/考察　実験結果を表 2 に示す。entailment の割合が約 83%であること、および state ごとの entailment の件数にも偏りがないことから、state の種類を問わず、与えられた状況に忠実な生成をすることがある程度可能であると言える。生成事例を付録 A.5 に示す。連勝・連敗などの文脈の有無を問わず、勝利した際には「やったー！」と冒頭で述べる事例、敗北した際には「負けちゃった」と 1 文目で述べる事例など、類似した生成事例が散見された。本結果は、ゲーム実況における多様な発話を再現する上では不十分であるということを示している。

表 2　生成された発話と state との一貫性の人手評価結果

ラベル	All	state0	state1	state2
entailment	50 (0.833)	17	16	17
contradiction	8 (0.133)	2	3	3
neutral	2 (0.033)	1	1	0
Total	60	20	20	20

5　おわりに

本論文では気軽な配信を実現するシステム TEPPAY を提案した。TEPPAY を構成する各種モジュールの性能を評価したところ、実況生成に必要な性能をある程度担保しているものの、実況発話の感情分類や発話生成のバリエーションに課題があることを確認した。今後は、各種モジュールの性能向上および各種モジュールの繋ぎ込みによる TEPPAY の完成を目指す。

8）　本研究では、「state0: 初勝利 / 敗北」「state1: 連勝 / 連敗」「state2: 連勝後の敗北 / 連敗後の勝利」の 3 種類で検証する。

参考文献

[1] 矢野経済研究所. Vtuber 市場に関する調査（2023）. https://www.yano.co.jp/press-release/show/press_id/3304 (アクセス日:2025 年 1 月 8 日).

[2] Natale Amato, Berardina De Carolis, Francesco de Gioia, Mario Nicola Venezia, Giuseppe Palestra, and Corrado Loglisci. Can an ai-driven vtuber engage people? the kawaii case study (short paper). In Axel Soto and Eva Zangerle, editors, **Joint Proceedings of the ACM IUI 2024 Workshops co-located with the 29th Annual ACM Conference on Intelligent User Interfaces (IUI 2024), Greenville, South Carolina, USA, March 18, 2024**, Vol. 3660 of **CEUR Workshop Proceedings**. CEUR-WS.org, 2024.

[3] Zhicong Lu, Chenxinran Shen, Jiannan Li, Hong Shen, and Daniel Wigdor. More kawaii than a real-person live streamer: Understanding how the otaku community engages with and perceives virtual youtubers. In **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[4] 横田健治. バーチャル youtuber の提供価値の分析. 電子情報通信学会誌 = The journal of the Institute of Electronics, Information and Communication Engineers, Vol. 102, No. 7, pp. 654–659, 07 2019.

[5] Man To Tang, Victor Long Zhu, and Voicu Popescu. Alterecho: Loose avatar-streamer coupling for expressive vtubing. In **2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)**, pp. 128–137, 2021.

[6] Zhantao Lai and Kosuke Sato. Multi-criteria evaluation framework of selecting response-worthy chats in live streaming. In **Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 186–191, Kyoto, Japan, September 2024. Association for Computational Linguistics.

[7] Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Generating live sports updates from twitter by finding good reporters. In **Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)**, Vol. 1, pp. 527–534, 2013.

[8] Byeong Jo Kim and Yong Suk Choi. Automatic baseball commentary generation using deep learning. In **Proceedings of the 35th Annual ACM Symposium on Applied Computing**, pp. 1056–1065, 2020.

[9] 森雄一郎, 前川在, 小杉哲, 船越孝太郎, 高村大也, 奥村学. サッカー実況中継を付加的情報の提供という側面から見る. 言語処理学会 第 30 回年次大会, pp. 2040–2045, 2024.

[10] Noah Renella and Markus Eger. Towards automated video game commentary using generative AI. In **CEUR Workshop Proceedings: AIIDE Workshop on Experimental Artificial Intelligence in Games**, pp. 341–350, 2023.

[11] Zihan Wang and Naoki Yoshinaga. From eSports data to game commentary: Datasets, models, and evaluation metrics. In **DEIM Forum 2021**, 2021.

[12] Zihan Wang and Naoki Yoshinaga. Commentary generation from data records of multiplayer strategy esports game. In Yang (Trista) Cao, Isabel Papadimitriou, Anaelia Ovalle, Marcos Zampieri, Francis Ferraro, and Swabha Swayamdipta, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)**, pp. 263–271, 2024.

[13] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. **arXiv preprint arXiv:2009.09941**, 2020.

A　付録

A.1　実況発話の生成に用いたプロンプト

あなたは {PERSONA} のゲーム実況者です
ゲームは {GAME} です

実況生成のフロー
以下のフローで試合終了時の実況を行ってください

1. 実況内容を決める
以下を踏まえてバリエーション豊かな実況を作成します
・試合の結果を喜んだり悔しがったりと感情をあらわにする
・キャラクター同士の相性 (有利や不利) を踏まえる
・前の試合までの連勝数や連敗数と今回の勝敗を踏まえる
・一般的なキャラの強さやマッチアップ勝率を踏まえる

2. 実況内容を生成する
実況者＝プレイヤー 1 です。プレイヤー 1 の目線で実況します
実況はテンション高く感情込めて行います
実況なので箇条書きなどは使わず、一言で喋りきれるよう生成してください

3. 実況内容を精査する
一言で喋れる内容に調整します
生成した内容が複数の文章になっている場合は 1 文のみを選んで他は削除してください
その上で長さを 20〜60 文字の範囲に収めてください

4. 最終結果を出力する
以下の json 形式で出力してください
{
”content”: ”3. で調整した最終の実況セリフのみ”
}

補足情報
player1: 実況者
player2: 対戦相手
character1: player1 が使うキャラ
character2: player2 が使うキャラ
character1_class: player1 が使うキャラの名前
character2_class: player2 が使うキャラの名前
tier: キャラの一般的に言われる強さ
win_rate_today: 今日の勝率 (%)
mu_win_rates: このマッチアップのプレイヤーの勝率 (%)
sequential_win: 前の試合までカウントした連勝数 (今回を含まない)
sequential_lose: 前の試合までカウントした連敗数 (今回を含まない)
general_win_rate: 一般的なこのマッチアップの勝率 (rank: （試合終了時のみ）player1 から見た試合結果。1 なら勝利、2 なら敗北

A.2　発話の感情分類に用いたプロンプト

感情分析タスクに取り組んでください。
与えられた文に対して、[”ネガティブ”, ”ポジティブ”, ”ニュートラル”] のなかから感情ラベルを一つ割り当ててください。
他のテキストは出力せず、ラベルのみ生成してください。

文:
感情ラベル:

A.3　人手付与した感情ラベルの内訳

100 件のうち、positive が 27 件、neutral が 37 件、negative が 36 件だった。1 件のみ全員が異なるラベルを付与したため、その事例は neutral とした。

A.4　GPT-4o による感情分類の誤り事例

ラグラグラグラグ
あんたでちょっと俺のスランプ直させてもらおうか
この {ATTACK_NAME} の差し合いまじで俺上手えから気をつけなはれよ

A.5　GPT-4o による実況発話の生成事例

やったー！同じ C ランクキャラ同士のバトルで勝利できました！
やったー！連敗ストップで {CHARACTER} の力が炸裂しました！
ああ、負けちゃった！でも次こそは絶対に勝つから応援してね！
また負けちゃったけど、次は絶対に勝つぞ〜！