SMASH コーパス DLC：対戦ゲーム動画に対する掛け合い実況解説音声コーパス

齋藤 佑樹[1]　川松 亮太[1,2]　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：本稿では，我々が新たに構築した「SMASH コーパス DLC (DiaLogue Commentary)」を紹介する．このコーパスは既存の SMASH コーパスの拡張版であり，「大乱闘スマッシュブラザーズ SPECIAL」の対戦ゲーム動画に対し，プロ話者 1 名と非プロ話者 1 名が掛け合いながら対戦の様子を実況解説した音声を含む．本稿では，話者 2 名による掛け合い実況解説音声と，各話者が個別で収録した実況解説音声の言語的・音声的な違いを分析し，その振る舞いを計算機的に再現するために必要な要素について洞察を与える．SMASH コーパス DLC は，エンターテインメント領域における音声言語情報処理の新たな研究領域を開拓し，マルチモーダル音声言語理解・合成といった応用を通じて，人間を楽しませるための AI 技術開発の基盤となることを期待する．

1. はじめに

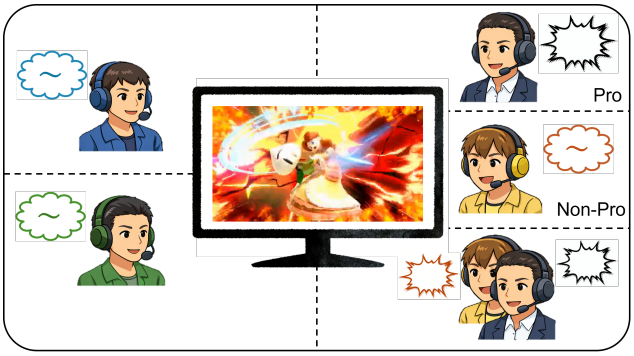
音声言語情報処理 (Spoken Language Processing; SLP) の研究を支える基盤として，多様な音声コーパスがこれまでに構築されてきた．特に，LibriVox[*1]をデータソースとする英語の多話者読み上げコーパス (LibriSpeech [1], Libri-Light [2], LibriHeavy [3]) とその多言語版 (MLS [4]) は，深層学習に基づく音声認識 [5] および汎用的な音声特徴量抽出を目的とした大規模事前学習モデル [6], [7] に関する技術の進展に大きく貢献している．音声認識技術を主眼に構築されたこれらのコーパスは必ずしも高品質でないが，VCTK [8] や JSUT&JVS [9] に代表されるスタジオ品質音声コーパスや，Miipher シリーズ [10], [11] や Sidon [12] などの音声復元技術により高品質化された音声認識向けコーパス (LibriTTS-R [13], FLEURS-R [14], MLS-Sidon[*2]) を用いることで，多様な話者・発話スタイルの音声を人工的に生成可能な音声合成技術 [15] も実現されている．

[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.
[*1] https://librivox.org/
[*2] https://huggingface.co/datasets/sarulab-speech/mls_sidon

このような背景をもとに，SLP の研究は，人間のように (1) 外界から得られる多様な情報を観測・統合し，(2) 環境や他者とのリアルタイムな相互作用を通じて，自らの振る舞いを適応的に制御可能な SLP システムの実現へと拡張されつつある．(1) は，音声・環境音といった聴覚モダリティだけでなく，動画像などの視覚モダリティからの情報を処理するための**マルチモーダル**機械学習 [16], [17], [18] を基盤とし，コミュニケーションに必要な情報の抽象化と取捨選択を実現する．(2) は，相手の発話が終了してから話し出すような half-duplex 対話ではなく，常に双方が音声でやり取りをし続ける **full-duplex 対話** [19], [20] により，あたかも実在する人間と実時間で音声コミュニケーションをしているかのような体験を提供する．これらに加える形で，自らが作り出した音声が他者に対してポジティブな影響（例：共感 [21]，好意 [22]，盛り上がり [23]）を与えるような制御機構も重要である．

前述の能力を有するマルチモーダル full-duplex 音声対話技術は，複数名による視聴覚モダリティの共有を前提とする，様々な場面における応用も期待できる．例えば，観光地を巡るユーザを現地で案内するような場面では，周囲の環境音・風景から得られる情報をユーザと対話システムの間で共有し，ユーザからの質問にリアルタイムで応答しつつ，ユーザの驚きや感動に寄り添った発話スタイルでの対話が求められる．また，サッカーなどのスポーツ中継やゲーム配信における複数名でのライブ実況解説では，試合の中で各話者がそれぞれの役割を理解・分担しつつ，正確

© 1959 Information Processing Society of Japan



(a) SMASH [24]:
“Solo” Commentary x 2

(b) **DLC** (this study):
“Solo” commentary x 2
& **“Dialogue”** Commentary x 1

図 1: SMASH コーパスと DLC の比較．切り抜かれているスマブラ SP のシーンは SMASH コーパス [24] に含まれている対戦動画データに由来する．

な情報伝達のみならず，映像コンテンツの視聴者を楽しませるような対話が望ましい．これらの応用を視野に入れると，(1) 視聴覚モダリティから得られる情報を対話参加者間で共有・理解し，進行中のイベントに即応した発話生成，(2) 複数話者が同時的・相互的に発話しながら，役割分担や掛け合いを行う対話的振る舞いの適切な制御は必要不可欠な要素といえる．

本研究では，アドリブ性の高い視聴覚イベントが頻発するマルチモーダル full-duplex 音声対話の一例として「対戦ゲーム動画の掛け合い実況解説」に着目し，そのための新たな音声コーパスである **SMASH コーパス DLC** (**D**ia**L**ogue **C**ommentary) を新たに構築する．SMASH コーパス DLC は，図 1 に示すように，既存の SMASH コーパス [24] の拡張版であり，対戦ゲーム「大乱闘スマッシュブラザーズ SPECIAL（スマブラ SP）[*3]」の対戦動画に対し，プロ話者 1 名と非プロ話者 1 名が掛け合いながら実況を行った音声を収録している．実況解説の熟練度に差がある 2 名による同時的・相互的な発話を含む点に特徴があり，さらに各話者が同一動画に対して個別に収録した実況解説音声も併せて収録している．これにより，単独での実況解説音声と掛け合いが生じる実況解説音声の違いを，言語的特徴，音声的特徴，対話的特徴などの観点から体系的に分析することが可能となる．本稿では，SMASH コーパス DLC の設計方針と収録条件を述べるとともに，基礎的な分析を通じて，掛け合い実況解説の振る舞いを計算機的に再現するために必要な要素について議論する．

2. SMASH コーパス DLC の構築

2.1 スマブラ SP 対戦動画データの準備

SMASH コーパス [24] に含まれる 69 件の対戦動画データ（約 4 時間）を使用した．各動画データは制限時間を 2 分 30 秒とした，スマブラ SP の時間制乱闘[*4]の様子を収録したものである．

2.2 実況解説音声の収録

2.1 節で述べた対戦動画データを用いて後付けの実況解説音声を収録した．収録にあたり，スマブラ SP の大会での実況解説経験を有する男性プロ話者 1 名 (**Pro**) と，スマブラ SP のプレイ経験を有するが，実況解説経験は有さない男性非プロ話者 1 名 (**Non-Pro**) を雇用した．収録は東京都内のスタジオを利用し，2025 年 1 月に 4 時間 ×5 日という日程で実施した．収録セッションは動画データ単位で進行し，話者は録音ブース内に配置された椅子に座りつつ，机の上に置かれたディスプレイを通じて再生される対戦動画の内容を，事前に台本等は用意せずに実況解説した．対戦動画はゲームの背景音・効果音に加え，SMASH コーパス [24] の収録に参加したプレイヤー (FF〜MF2) の音声を含むが，話者はこれらをヘッドフォンで聴取しながら実況解説したため，収録音声にはこれらの音は含まれなかった．

本研究では**話者 1 名でのソロ (Solo) 実況解説**と**話者 2 名での掛け合い (Dialogue) 実況解説**という 2 パターンの音声を収録した．

- **Solo 実況解説：**SMASH コーパス [24] の収録と同様に，Pro もしくは Non-Pro が単独で対戦動画の内容を実況解説した．音声は話者の前方に配置されたコンデンサー型マイクロホン (NEUMANN U87 AI) により収録した．
- **Dialogue 実況解説：**Pro と Non-Pro が共通の対戦動画を視聴しつつ，共同で動画の内容を実況解説した．この収録にあたり，話者には「基本的には Pro が先導して実況解説を行い，Non-Pro はその内容を補足し，重要イベントの発生に際し適宜介入する」ように指示した．Pro と Non-Pro は椅子に隣り合って座り，各話者の前方に 1 つずつ，75 cm 間隔で配置された NEUMANN U87 AI に向かって発話した[*5]．

2.3 書き起こし・タグのアノテーション

2.2 節で収録した音声に対し，以下のアノテーションを人手で付与した．

- **書き起こし：**ミリ秒単位の発話開始・終了時刻を記録し，当該区間での発話内容をかな漢字混じり文で書き起こした．この書き起こし結果には，“(F)”，“(D)”（それぞれ，フィラー，感情表出系感動詞と，言い直し，言い淀み等による語断片），そして “<息>”，“<笑>”（それぞれ，言語音と独立に話者の息，笑いが生

[*3] https://www.smashbros.com/

[*4] 制限時間内に相手キャラクターをより多く撃墜したプレイヤーもしくはチームが勝利となる，スマブラ SP における基本的な対戦形式である．

[*5] 即ち，掛け合い実況解説における Pro/Non-Pro 音声のチャネルは明確に分けられていない．

© 1959 Information Processing Society of Japan

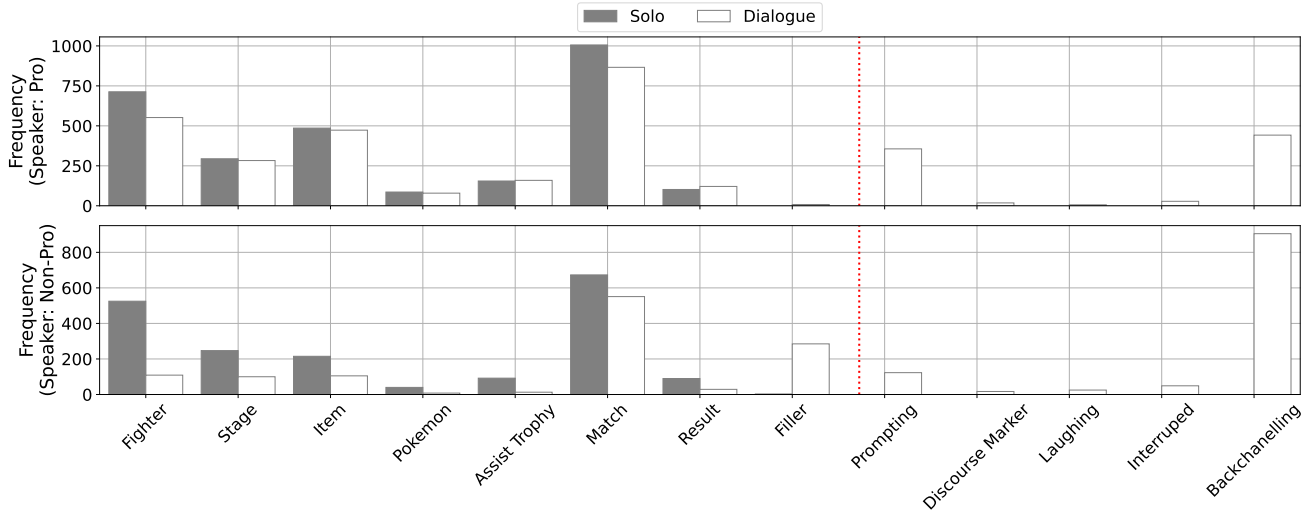


図 2: 発話トピックタグの出現頻度．赤の点線より右側のタグは掛け合い実況解説のみに出現するタグである．

表 1: 話者ごとの発話数と総発話時間

	Pro		Non-Pro	
	# utterance	Dur [h]	# utterance	Dur [h]
Solo	2,871	2.02	2,111	2.31
Dialogue	3,469	2.25	2,389	0.74

じている場合）というタグも内包されている．

- **発話トピック：**SMASH コーパス [24] と同様に，発話トピックのタグとして (1) **Fighter**（使われているファイターや，そのファイターに関連する作品など），(2) **Stage**（使われているステージや，そのステージに関連する作品など），(3) **Item**（出現しているアイテムや，そのアイテムに関連する作品など），(4) **Pokémon**（出現しているポケモン），(5) **Assist Trophy**（出現しているアシストフィギュア），(6) **Match**（対戦状況など，上記のタグ以外で対戦内容に関連するもの），(7) **Result**（対戦結果に関するもの）を用意した[*6]．また，「おっと」や「あー」といったフィラーのみから構成される発話区間のために (8) **Filler** を用意した．Dialogue 実況解説に対しては，これらに追加する形で，(9) **Prompting**（発話の促し），(10) **Discourse Marker**（談話標識），(11) **Laughing**（談笑），(12) **Interrupted**（発話の中断），(13) **Backchanelling**（相槌）を用意した．

Dialogue 実況解説の書き起こし・各種タグのアノテーションは話者ごとに実施した．

3. SMASH コーパス DLC の分析

3.1 コーパススペック

表 1 に SMASH コーパス DLC のスペックを示す．話者ごとの総発話数は Pro が 6,340（約 4 時間），Non-Pro が 4,500（約 3 時間）であった．総発話数および総発話時間には，実況形態 (Solo/Dialogue) の違いにより顕著な差が確認できる．具体的には，Pro/Non-Pro ともに発話数は Dialogue 実況解説で増加したが，Dialogue 実況解説の総発話時間は Pro で増加，Non-Pro で減少となった．即ち，本コーパスに含まれる話者は，実況形態の違いに応じて発話の密度と長さを適応的に変えていることが示唆された．

3.2 発話トピックタグの分析

図 2 に発話トピックタグの出現頻度を示す．実況形態 (Solo/Dialogue) の違いは，Non-Pro の発話において特に顕著に観測される．具体的には，Non-Pro の Dialogue 実況解説の中で，Filler の頻度は大幅に増加している一方で，SMASH コーパスに由来する Fighter–Result の頻度はすべて減少していることが確認できる．Solo 実況解説では，対戦の中で「現在何が起きているか」を話者が単独で逐次的に説明する必要があるが故に，Fighter, Stage, Item, Match, Result といった，スマブラ SP の客観的事実・試合の進行に直接関わるトピック（以降，「戦況説明タグ群」と表記）は，Solo 実況解説において出現頻度が全体的に高い．一方，Dialogue 実況解説では複数話者で対戦の様子を実況解説するため，Solo 実況解説と同様の戦況説明に加え，Prompting や Backchanneling を通じた互いの発話タイミング制御（ターンテイキング）が要求される．これに伴い，戦況の説明を主な役割とする Pro は戦況説明タグ群の発話の頻度を大きく変化させないのに対し，Non-Pro は戦況説明の頻度を減らし，Filler, Prompting, Backchanelling を通じて重要イベントの発生に言及しつつ，Pro の実況解説を補助するような役割を担っている．以上より，本研究で取り扱う「掛け合い実況解説」は，SMASH コーパス [24]

[*6] SMASH コーパスでは **Chat**（雑談）も含むが，本研究ではこのタグはアノテーション作業中に使用されなかったため除外する．

© 1959 Information Processing Society of Japan

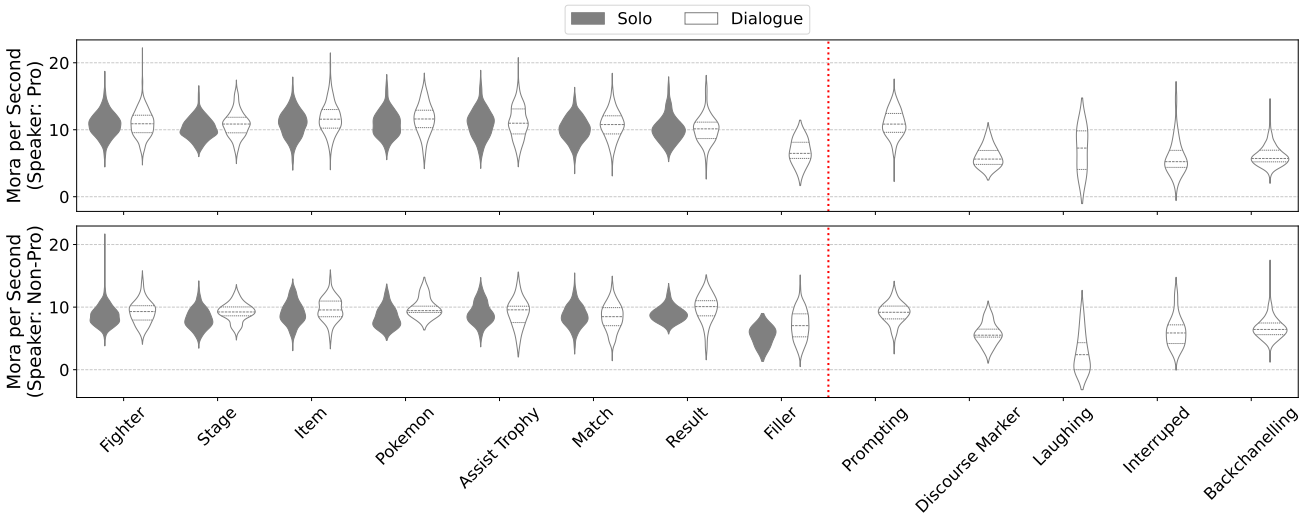


図 3: 話速（モーラ毎秒）のバイオリンプロット

表 2: 1 試合あたりに発生するフィラー (F)，非流暢発声 (D)，呼吸音 (B) の回数の平均と標準偏差

(a) Pro

	F	D	B
Solo	0.23±0.52	1.27±1.20	2.25±2.24
Dialogue	0.74±0.91	2.41±1.91	2.41±2.91

(b) Non-Pro

	F	D	B
Solo	1.75±1.80	2.70±1.63	13.47±5.09
Dialogue	7.20±2.56	1.07±1.03	0.86±1.41

でも考慮されていた (1) リアルタイムに変化する戦況の即興的な正確な説明と (2) 聴衆を楽しませるような発話計画に加え，(3) 話者の役割をの違いを明確にしつつ，適切な対話管理を要する対話タスクだといえる．

3.3 言い淀み・非言語的発声の分析

表 2 は, SMASH コーパス DLC に含まれる Pro/Non-Pro の 1 試合中の実況解説に，フィラー (F)，非流暢発声 (D)，呼吸音 (B) が何度生じるかを，Solo/Dialogue の条件ごとに比較した結果である．

Pro の結果に着目すると，Solo 実況解説から Dialogue 実況解説へ移行した際，フィラー，非流暢発声，呼吸音のいずれについても平均回数が増加していることが分かる．特に非流暢発声の増加が顕著であり，対話形式では即時的な戦況解説に加え，他の話者とのターンテイキング・応答が求められるため，発話計画に伴う認知的負荷が高まることが一因であると考えられる．ただし，その増加幅は比較的小さく，本コーパスに含まれる Pro は実況形態の変化に対しても流暢な発話をある程度維持できていることが示唆される．

一方，Non-Pro の結果では，実況形態による影響がより顕著に現れている．特に Solo 実況解説における呼吸音の発生回数が非常に多く，Pro と比較して顕著な差が見られる．これは，発話訓練を受けていない話者にとって，長時間の単独実況における発話ペースや呼吸・発話の協調を適切に制御することが困難であったためであると考えられる．また，興味深い点として，Non-Pro は Dialogue 実況解説に移行することで，呼吸音および言い淀みの発生回数が大幅に減少する一方，フィラーの発生回数が著しく増加している．これは，対話における役割の変更により発話量や時間的負荷が分担され，生理的負担が軽減された結果，呼吸音が減少した可能性を示している．一方で，発話開始時や応答時にフィラーを多用することで，発話権の保持や間の調整を行っていると解釈でき，掛け合い対話実況に特有の談話調整スキルが十分に獲得されていない可能性も示唆されている．

3.4 音声的な違いの分析

発話ごとの話速と声の平均的な高さを分析するために，1 秒間に発話されるモーラ数（モーラ毎秒）と基本周波数の平均値（平均 F0）の統計量を算出した．F0 の抽出には WORLD ボコーダ [25] を用いた．

話速：図 3 に，発話トピックタグで集計されたモーラ毎秒のバイオリンプロットを示す．全体的に，Pro は Non-Pro と比較してモーラ毎秒の値が大きく，より早口で実況解説する傾向が確認できる．また，Pro では Solo/Dialogue 実況解説の間で話速に大きな差は見られない．一方で，Non-Pro では Dialogue 実況解説において話速の分布形状が大きく変化する傾向が観察された．これは，実況解説での Non-Pro の役割の変化に起因すると考えられる．

声の高さ：図 4 に，発話トピックタグで集計された F0 平

© 1959 Information Processing Society of Japan

表 3: 実況者ごとの発話量に関する統計量．各値は平均値 ± 標準偏差を示す．

話者	実況形態	発話時間の累積	発話時間の平均	沈黙時間の累積	沈黙時間の平均	発話数	発話長
Pro	Solo	105.35 ± 8.30	2.56 ± 0.33	64.60 ± 7.90	1.62 ± 0.27	41.6 ± 4.6	22.3 ± 2.6
Non-Pro	Solo	120.70 ± 10.55	4.08 ± 0.87	49.45 ± 10.87	1.71 ± 0.39	30.6 ± 5.1	32.8 ± 6.6
Pro	Dialogue	117.51 ± 7.82	2.36 ± 0.28	53.65 ± 7.06	1.11 ± 0.16	50.3 ± 4.8	21.7 ± 2.2
Non-Pro	Dialogue	38.60 ± 8.27	1.15 ± 0.25	120.28 ± 8.78	3.95 ± 1.83	34.6 ± 8.17	8.8 ± 2.0

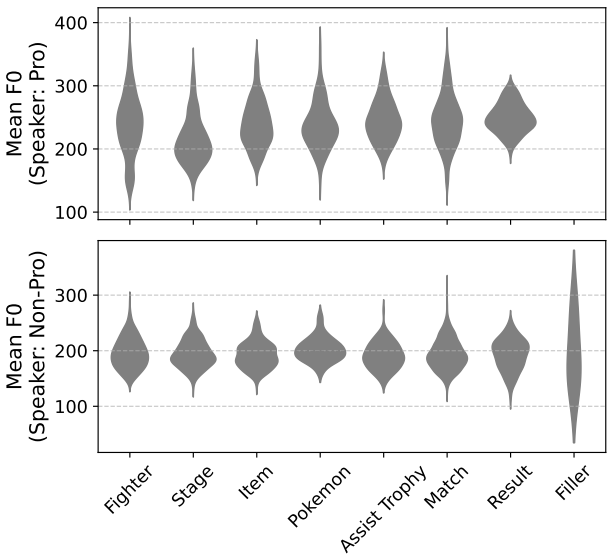


図 4: 平均 F0 のバイオリンプロット

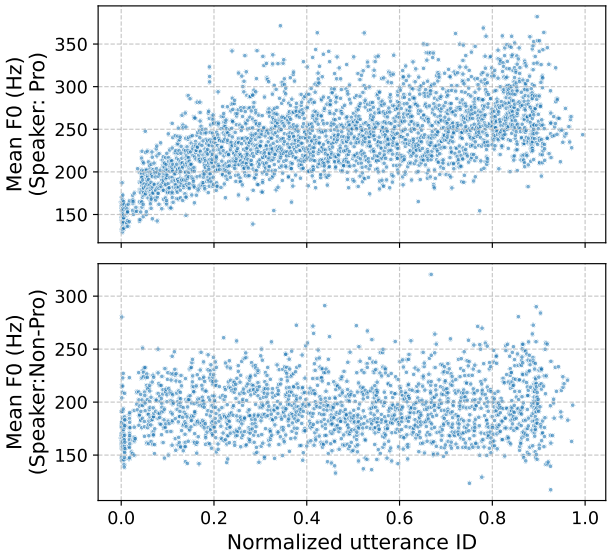


図 5: 平均 F0 の時間的変化．横軸は試合中の正規化された発話 ID であり，0 と 1 はそれぞれ試合内での最終発話に対する ID を意味する．即ち，横軸の値が増加するにつれて試合は進行し，その時点で発話された声の高さが平均的にどの程度なのかを示している．

均値のバイオリンプロットを示す．ここで，Dialogue 実況解説では発話中の話者オーバーラップが頻発するため，本分析からは除外した．Pro 発話の F0 平均値は分布の幅が広く，トピックごとに分布形状にも違いが見られる．一方で，Non-Pro は分布の幅が比較的狭く，Filler を除くとトピック間で分布形状に大きな違いは見られない．このことから，Pro は発話トピックに応じて声の高さを柔軟に使い分け，発話訓練を受けていない Non-Pro は，声の高さを意図的に制御することが困難であった可能性が示唆される．

また，図 5 に 1 試合内における声の高さの時間的変化を可視化した散布図を示す．この散布図では，横軸を正規化された発話 ID，縦軸を平均 F0 とした．Pro では試合終了が近づくにつれ，実況解説音声の平均 F0 が上昇する傾向が見られた一方，Non-Pro では同様の傾向は確認されなかった．これは Pro が試合終盤に向けて聴衆を盛り上げるような韻律制御をしている可能性を示唆する．同様の傾向は，競馬の実況解説においても確認されている [26]．

3.5 音声対話的な振る舞いの分析

SMASH コーパス DLC に収録された掛け合い実況解説音声に注目し，音声対話的な振る舞いの特徴を定量的に分析する．

オーバーラップ発話時間の分析：Dialogue 実況解説の full-duplex 性を定量的に評価するため，Nguyen ら [27] および Ohashi [28] らの手法に基づき，1 分あたりのオーバーラップ時間（秒）を算出した．分析の結果，Dialogue 実況解説における Pro/Non-Pro の平均的な発話オーバーラップ率は 5.27 秒/分 であった．この値は，J-Moshi モデルの学習データのオーバーラップ率（約 8.0 秒/分）と比較すると小さいことがわかる．

発話・沈黙パターンの分析：表 3 に，話者ごとの発話量に関する統計量を示す．ここで，発話時間の累積とは，話者が 1 試合において発話した時間の合計を各試合について算出し，それらを平均した値である．発話時間の平均は，全試合における全発話を対象として，1 発話あたりの発話時間を平均したものである．沈黙時間は，当該話者が発話していない区間を指し，掛け合い実況においては他方の話者が発話している時間も沈黙時間に含めた．沈黙時間の累積および平均も，同様の手続きにより算出した．発話数は，1 試合における話者の発話回数を各試合についてカウントし，それらを平均した値である．さらに，発話長については，全試合における全発話を対象として，1 発話あたりの文字数の平均を算出した．

まず，ソロ実況に注目すると，Non-Pro は Pro と比較し

© 1959 Information Processing Society of Japan

て，発話時間の累積が長く，一方で発話数は少ない．これに伴い，1 発話あたりの平均発話時間も Non-Pro の方が大きい．一方で，Dialogue 実況解説では，話者間で顕著な差が観察された．Pro は発話時間の累積および発話数が大きく，平均発話時間も Solo 実況解説の場合と同程度であるのに対し，Non-Pro は発話時間の累積が大幅に減少し，その代わりに沈黙時間の累積が増加している．また，Non-Pro の沈黙時間の平均が 3.95 秒と，他条件と比較して最も長い値を示した．さらに，発話長も大きく減少する傾向が確認された．これは 2.2 節で述べたように，Dialogue 実況解説において Pro が先導して実況解説を行い，Non-Pro はその内容を補足するという指示が与えられていたため，掛け合い実況時には Non-Pro の発話量が減少したと考えられる．

4. おわりに

本研究では，対戦ゲーム「大乱闘スマッシュブラザーズ SPECIAL」のプレイ動画を対象として，プロの実況解説者 1 名と非プロの実況解説者 1 名が掛け合う音声を含む「SMASH コーパス DLC (DiaLogue Commentary)」を新たに構築した．今後はこのコーパスを実況音声合成 [23] などのタスクに使えるかどうか検証を進める．本コーパスは，アカデミック機関での研究，非商用目的の研究，個人での利用に用途を限定して使用可能とする予定である．

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] V. Panayotov et al., “LibriSpeech: An ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
[2] J. Kahn et al., “Libri-Light: A benchmark for ASR with limited or no supervision,” in *Proc. ICASSP*, 2020, pp. 7669–7673.
[3] W. Kang et al., “Libriheavy: A 50,000 hours ASR corpus with punctuation casing and context,” in *Proc. ICASSP*, 2024, pp. 10 991–10 995.
[4] V. Pratap et al., “MLS: A large-scale multilingual dataset for speech research,” in *Proc. INTERSPEECH*, 2020, pp. 2757–2761.
[5] R. Prabhavalkar et al., “End-to-end speech recognition: A survey,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 325–351, 2024.
[6] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. NeurIPS*, 2020, pp. 12 449–12 460.
[7] A. Babu et al., “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” in *Proc. INTERSPEECH*, 2022, pp. 2278–2282.
[8] J. Yamagishi et al., “CSTR VCTK Corpus: english multi-speaker corpus for CSTR voice cloning toolkit,” https://doi.org/10.7488/ds/2645, 2019.
[9] 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, Sep. 2020.
[10] Y. Koizumi et al., “Miipher: A robust speech restoration model integrating self-supervised speech and text representations,” in *Proc. WASPAA*, 2023.
[11] S. Karita et al., “Miipher-2: A universal speech restoration model for million-hour scale data restoration,” in *Proc. WASPAA*, 2025.
[12] W. Nakata et al., “Sidon: Fast and robust open-source multilingual speech restoration for large-scale dataset cleansing,” in *Proc. ICASSP*, 2026 (Accepted).
[13] Y. Koizumi et al., “LibriTTS-R: A restored multi-speaker text-to-speech corpus,” in *Proc. INTERSPEECH*, 2023, pp. 5496–5500.
[14] M. Ma et al., “FLEURS-R: A restored multilingual speech corpus for generation tasks,” in *Proc. Interspeech 2024*, 2024, pp. 1835–1839.
[15] S. Chen et al., “Neural codec language models are zero-shot text to speech synthesizers,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705–718, 2025.
[16] D. Michelsanti et al., “An overview of deep-learning-based audio-visual speech enhancement and separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.
[17] B. Shi et al., “Learning audio-visual speech representation by masked multimodal cluster prediction,” in *Proc. ICLR*, 2022.
[18] U. Cappellazzo et al., “Large language models are strong audio-visual speech recognition learners,” in *Proc. ICASSP*, 2025.
[19] A. Défossez et al., “Moshi: a speech-text foundation model for real-time dialogue,” *arXiv*, vol. arXiv:2410.00037, 2024.
[20] H. Lu et al., “SLIDE: Integrating speech language model with LLM for spontaneous spoken dialogue generation,” in *Proc. ICASSP*, 2025.
[21] Y. Saito et al., “STUDIES: Corpus of Japanese empathetic dialogue speech towards friendly voice agent,” in *Proc. Interspeech*, 2022, pp. 5155–5159.
[22] H. Suda et al., “Voice conversion for likability control via automated rating of speech synthesis corpora,” in *Proc. INTERSPEECH*, 2025, pp. 1363–1367.
[23] K. Iura et al., “Excitement-inducing commentary text-to-speech system for fighting game video scenes,” *IEEE Access*, vol. 13, pp. 216 748–216 758, 2025.
[24] Y. Saito et al., “SMASH corpus: A spontaneous speech corpus recording third-person audio commentaries on gameplay,” in *Proc. LREC*, 2020, pp. 6571–6577.
[25] M. Morise, “D4C, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, 2016.
[26] J. Trouvain, W. J. Barry, “The prosody of excitement in horse race commentaries,” in *Proc. ISCA Workshop Speech Emotion*, 2000, pp. 86–91.
[27] T. A. Nguyen et al., “Generative spoken dialogue language modeling,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.
[28] A. Ohashi et al., “Towards a japanese full-duplex spoken dialogue system,” in *Proc. Interspeech*, 2025, pp. 1783–1787.

© 1959 Information Processing Society of Japan