SMASH コーパス：
ゲーム動画の後付け実況解説音声収録に基づく自発発話音声コーパス *

◎齋藤 佑樹，高道 慎之介，猿渡 洋（東大院・情報理工）

1　はじめに

自発発話音声の研究は，より実用的・発展的な音声言語情報処理技術の実現に不可欠である．近年の機械学習技術の進展に伴い，自発発話音声の処理と理解に関する種々の研究（認識 [1]，生成 [2]，要約 [3] など）が行われてきた．これらの研究をさらなる進展には，多様な自発発話音声を収録したコーパスが必要である．また，コーパス自体だけでなく，自発発話音声を収録するための方法論の確立も重要である．

これまでに構築された，日本語で最大規模の自発発話音声コーパスは，日本語話し言葉コーパス (Corpus of Spontaneous Japanese: CSJ) [4] であり，1,417 名の日本語母語話者による合計約 660 時間の自発発話音声を含む．ここで，多様な発話様式の音声を収録するために，学会講演，模擬講演，課題志向対話などの様々な収録環境が用意された．CSJ は音声・自然言語研究の進展に大きく貢献しているが，このような大規模なコーパス構築方針に他の研究者が追従するのは困難である．より容易な方針の一つは，話者にとって親しみやすく，自身の考えや感情を躊躇なく発話可能な収録環境を用意することで実現できると考えられる．感情評定値付きオンラインゲーム音声チャットコーパス [5] はこの方針に基づいて構築されたコーパスの一例であり，オンラインゲームをプレイ中の音声チャットとして自然な感情音声を収録している．

本稿では，ゲームプレイの実況解説の収録に基づく自発発話音声コーパスの構築法を提案する．この手法は，Fig. 1 に示すような，実況解説者の (1) 時々刻々と変化するゲームプレイの様子を正確に述べる能力，そして (2) 適切な語彙・トピックを選択し，時には感情的な発話で聴衆を楽しませる能力に着目したものである．コーパスの構築に必要なのは，ゲームとそのプレイヤー，そして当該ゲームを熟知している実況解説者の用意のみであり，これまでよりも容易な自発発話音声の収録が可能となる．また，ゲームプレイシーンと実況解説音声のペアは，visually-grounded な機械学習 [6] に基づく先進的なマルチモーダル音声情報処理技術の実現にもつながる．

我々はこのコーパス構築法の初期検討として，世界的に有名な対戦ゲームである大乱闘スマッシュブラザーズ SPECIAL（スマブラ SP）のゲームプレイ実況解説音声を含む “SMASH コーパス” を構築した．スマブラ SP は，突如出現するアイテムや，画面上に存在するキャラクターの予期しない動きなど，高いアドリブ性を有するゲームであり，多様な自発発話音声の収録に適していると考えた．本稿では，このコーパスの具体的な構築法と，種々のアノテーション結果について述べる．

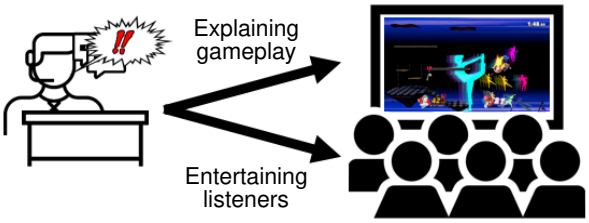


Fig. 1　ゲーム実況解説の概念図

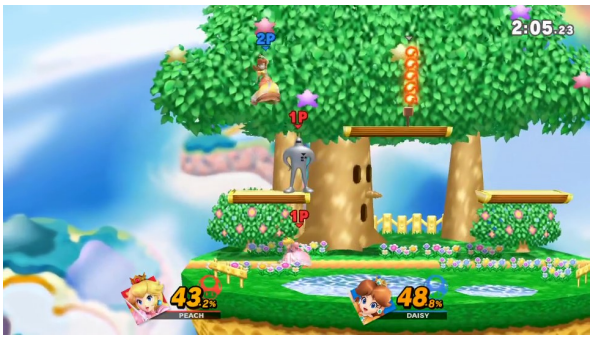


Fig. 2　スマブラ SP のゲームプレイシーンの例．画面上には，2 体の**ファイター**と 1 体のゲストキャラクター（**アシストフィギュア**）が出現している．また，ステージの上には 1 つの**アイテム**が生成されている．

2　スマブラ SP の予備知識

スマブラ SP は，任天堂から Nintendo Switch 向けに発売された対戦ゲームであり，2019 年 10 月までに世界各国で累計 15 百万本以上の売上を記録している．

2.1　ゲームプレイの基本ルール

Figure 2 にスマブラ SP のゲームプレイシーンの一例を示す．基本的なルールは非常に単純であり，プレイヤーが使用するキャラクター（**ファイター**）の攻撃などで相手により多くのダメージを与え，場外に飛ばす（KO する）ことでスコアを稼ぐ．Figure 2 の画面下部に示されている “43.2%” などの値は，当該ファイターが受けたダメージの総量を表し，より多くのダメージを受けたファイターはより遠くに飛ばされやすくなる．基本的には，より多くのスコアを稼いだプレイヤー（もしくは non-player character: NPC）がゲームの勝者となる．また，**ステージ**上にランダムに出現する**アイテム**を活用することで，プレイヤーは戦況をより有利に運ぶことができる．

ゲーム開始時に，プレイヤーは対戦に用いるステージとファイターを選択する．インターネットを通じて購入できるダウンロードコンテンツを除き，スマブラ SP では 103 個のステージと 74 体のファイターを選択できる．各ファイターは攻撃手段や移動能力など

*SMASH corpus: spontaneous speech corpus of audio commentaries on gameplay by SAITO, Yuki, TAKAMICHI, Shinnosuke, and SARUWATARI, Hiroshi (The University of Tokyo).

の多くの点で異なるが，基本的な操作方法はほぼ共通である．ファイターだけではなく，ステージにも対戦を盛り上げるための様々な要因（例えば，ファイターの行動を阻害するステージギミック）が存在する．

以降の節では，スマブラ SP におけるいくつかの重要な概念について簡潔に述べる．

2.2 アイテムの利用

スマブラ SP には，対戦を有利に進めるための 83 種類のアイテムが存在する．本節では，特に重要な 2 種類のアイテムについてさらに説明する．

2.2.1 モンスターボールとマスターボール

モンスターボールやマスターボールを投げると，スマブラ SP で特集されている 55 種類の**ポケモン** (Pokémon) の中からランダムに選ばれた 1 体がステージ上に召喚され，ファイターを一定時間サポートする．これら 2 つのアイテムの機能は基本的に同じだが，マスターボールは特に強力な伝説ポケモンのみを召喚するという点で異なる．

2.2.2 アシストフィギュア

アシストフィギュア (Assist Trophy) を取得すると，様々なゲームから特集された 59 体のゲストキャラクターの中からランダムに 1 体がステージ上に出現する．ほぼ全てのゲストキャラクターがファイターをサポートする効果を持つが，中には，画面を見えなくするといった妨害効果を与えるものも存在する．ゲストキャラクターにダメージを与えて KO することで，プレイヤーはスコアを稼ぐこともできる．

3 SMASH コーパスのデータ収集

3.1 スマブラ SP ゲームプレイの収録

スマブラ SP ゲームプレイの収録にあたり，本稿では最も基本的な対戦ルールである時間制乱闘を用いた．ここで，ゲームの勝者は制限時間 2 分 30 秒の間に最も多くのスコアを稼いだプレイヤー（もしくは NPC）で決まり，制限時間終了時にトップタイが生じた場合はサドンデスマッチにより唯一の勝者が選ばれる．その他のオプション（例えば，アイテムの出現頻度など）は，デフォルトの設定を用いた．ステージやファイターの選択に制限は設けなかったが，ダウンロードコンテンツは選択肢から除外した．

ゲームのプレイヤーとして，スマブラ SP をプレイした経験がある 4 ペアを集めた．ここで，様々なゲームプレイのシチュエーションを用意するために，男女全てのペアとして男性 2 人 (MM)，女性 2 人 (FF)，男女 2 人 ×2 ペア (MF1, MF2) を雇用した．各ペアはオフラインでスマブラ SP を約 1 時間プレイし，前半の 30 分で 1 対 1 のシングルスマッチ（プレイヤー同士の対戦）を，後半の 30 分で 2 対 2 のダブルスマッチ（プレイヤー 2 人と NPC2 体でのチーム戦）を収録した．ダブルスマッチでの NPC の強さ（9 段階から選択）はプレイヤーの好みに設定させた．対戦のシーンだけでなく，ステージやキャラクターの選択画面も収録した．収録の解像度とフレームレートは，それぞれ 1920×1080 と 30 fps に設定した．Table 1 に収録されたゲームプレイの内訳を示す．

Table 1　収録されたスマブラ SP のゲームプレイ

		対戦数	
ペア	収録時間	シングルス	ダブルス
MM	60 分 32 秒	9	9
FF	59 分 40 秒	9	8
MF1	58 分 41 秒	9	8
MF2	58 分 18 秒	9	8

Fig. 3　スマブラ SP ゲーム実況解説収録のイメージ図．コメンテーターは，プレイ動画を視聴しながら，机上に配置されたデスクトップマイクに向けて発話した．

3.2 ゲーム実況解説音声の収録

ゲーム実況解説音声の収録のために，スマブラ SP に関する深い知識を持った男性のコメンテーター 2 名 (MC1, MC2) を雇用した．各コメンテーターは対戦だけでなく全てのゲームプレイシーンを視聴し，MC1 は 3 ペア (MM, FF, MF1)，MC2 は残りの 1 ペア (MF2) のプレイに対して実況解説を行った．実況解説音声の収録は，音響監督による管理に基づいて収録スタジオで行われ，各コメンテーターの音声収録は 1 日以内で終えられた．全ての音声は単一指向性のデスクトップコンデンサマイクで録音された．音声のサンプリングレートは 48 kHz とし，オーディオフォーマットとしては，16 bit/sample RIFF WAV を用いた．Figure 3 に実況解説収録のイメージ図を示す．

4 SMASH コーパスのアノテーション

4.1 データの前処理

ゲームプレイと実況解説音声は全てのシーンに関して収録したが，本稿では最も重要なパートである対戦部分のみをアノテーション対象とした．そこで，収録されたゲームプレイと実況解説音声を対戦毎に分割し，それぞれ (1) ファイター紹介パート (約 3 秒)，(2) 対戦パート (約 2 分 40 秒，サドンデスを含む場合あり)，(3) 対戦結果パート (約 10 秒) により構成されるようにした．

4.2 自動書き起こし

自発発話音声の収録では事前にテキストを用意しないため，発話内容の書き起こしは非常に困難である．そこで，近年開発されたクラウドベースの音声書き起こしサービスにより自動で生成された書き起こしを利用することで書き起こし作業の効率化を図った．本稿では，この音声書き起こしサービスとして Google

Table 2　トピックタグのリスト

タグ	定義
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
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
Item	出現しているアイテムや，そのアイテムの関連する作品など
Pokémon	出現しているポケモン
Assist Trophy	出現しているアシストフィギュア
Match	上記のタグ以外で，対戦内容に関連するもの
Result	対戦結果に関するもの
Chat	コメンテーターの過去の経験など，対戦内容と関係がない雑談

Cloud Speech-to-Text[1]を利用し，対戦ごとに分割された実況解説音声単位で自動書き起こしを生成した．

4.3　アノテーションの付与

自動書き起こしの結果に基づき，本稿では (1) 修正書き起こしと (2) トピックタグのアノテーションを行った．アノテータとして健聴者の日本語母語話者を雇用し，実況解説音声付きのゲームプレイ動画を視聴しながら，これら 2 つのアノテーションを同時に遂行させた．このアノテーション開始の前に，実況解説音声を発話や文の長さ，ブレスの挿入を考慮して複数のセグメントに分割した．このとき，分割前の実況解説音声における各セグメントの開始・終了時刻は記録しておき，アノテータが作業を行う際に提示した．

4.3.1　修正書き起こし

ゲームプレイシーンとの対応関係を考慮しながら，アノテータに自動書き起こしの結果を手動で修正させた．ここで，ポーズ位置・長さの決定は困難であるため，修正書き起こしに句読点は挿入しないように指示した．また，自動書き起こしから完全に欠落したフィラーの挿入も禁止した．書き起こし結果に自信がない固有名詞についてはカタカナで表記するように指示し，結果の集計時に本稿の第 1 著者が目視でその固有名詞を確認・必要に応じて手動で修正した．

4.3.2　トピックタグ付け

修正書き起こしと並行して，各アノテータに発話トピックのタグ付けを依頼した．本稿では，スマブラ SP のゲームプレイにおいて特に重要であると判断した Table 2 に示す 8 つのタグを定義した．

5　アノテーション結果

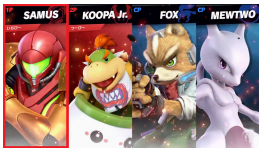
5.1　対戦シーンと対応付けられたアノテーション

Table 3 にアノテーション結果の一例を示す．この表より，いくつかの修正書き起こし結果にはそのトピックを特徴づけるキーワード（ゲームやキャラクターの名前など）が含まれていることが確認できる．

[1] https://cloud.google.com/speech-to-text/

Table 3　アノテーション結果の例．修正書き起こしの**太字**はトピックを特徴づけるキーワードを意味する

ID	修正書き起こし	タグ
(1)	もうね**メトロイド**のエンディングでね中身が女性だっていうことが分かるんですけれども	Fighter
(2)	さあサドンデスなりましたどっちが勝つのか	Match
(3)	じゃあもうこの二人はもうレベル 9 で行った方がいいんじゃないかなと思いますけれどもね	Chat
(4)	さあおお**イフリート**出ましたね	Stage



(1) ファイター紹介

(2) サドンデス

(3) 対戦結果


(4) ダブルスマッチ

Fig. 4　Table 3 のアノテーション結果に対応するゲームプレイシーン．(1) と (4) では，コメンテーターは図の赤枠で囲まれた部分である (1) サムス（メトロイドの主人公）と (4) イフリート（ステージの背景に出現しているキャラクター）について話した．

また，これらの結果は，Fig. 4 に示すゲームプレイシーンと対応付いていることも確認できる．

5.2　トピックの遷移

Figure 5 にシングルスマッチにおける実況解説音声のトピック遷移の例を示す．この例においては，Table 2 に示す全てのトピックタグが網羅されており，いくつかの急激なトピック遷移（すなわち，発話内容の割り込み）が発生していることも確認できる．例えば，この図中の赤で囲まれた部分では最後の切り札（ファイターによる強力な超必殺技）が発生しており，これによりトピックが “Chat” から “Match” に遷移している．全体的なトピック遷移のパターンとして，“Fighter” や “Match” がトピックの多くを占めている中で，“Item” や “Pokémon” などのランダムな要因が実況解説に対する割り込みを発生させる傾向にあることが確認された．

5.3　自動書き起こしの有効性

自動書き起こしを利用したアノテーションの有効性を検証するために，修正書き起こし結果をリファレンスとした word error rate (WER) を計算した．その結果として，全ての発話セグメントで平均した WER は 10.3%であった．すなわち，自動書き起こしの利用は自発発話音声の発話内容書き起こしの補助に有効であるということが示唆された．

5.4　トピックタグの分布

Figure 6 にトピックタグの分布を示す．ここでは，アノテーションの結果はプレイヤーペア毎にまとめ

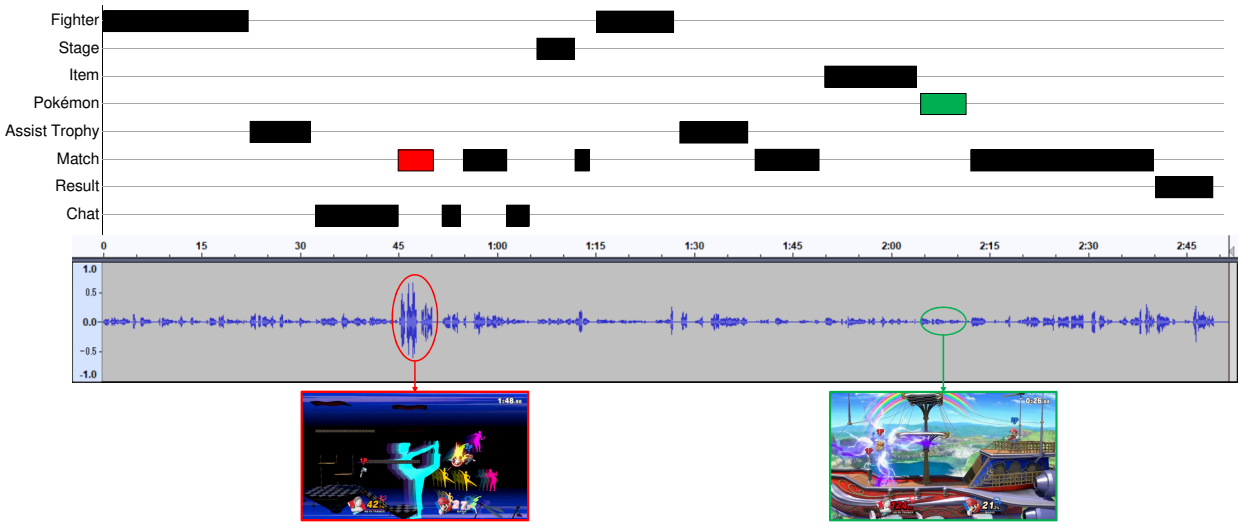


Fig. 5　実況解説におけるトピック遷移の例．図中の赤丸（最後の切り札が発動しているシーン）でトピックの割り込みが発生している．この図では，連続する同じトピックの発話セグメントは結合して表示している．

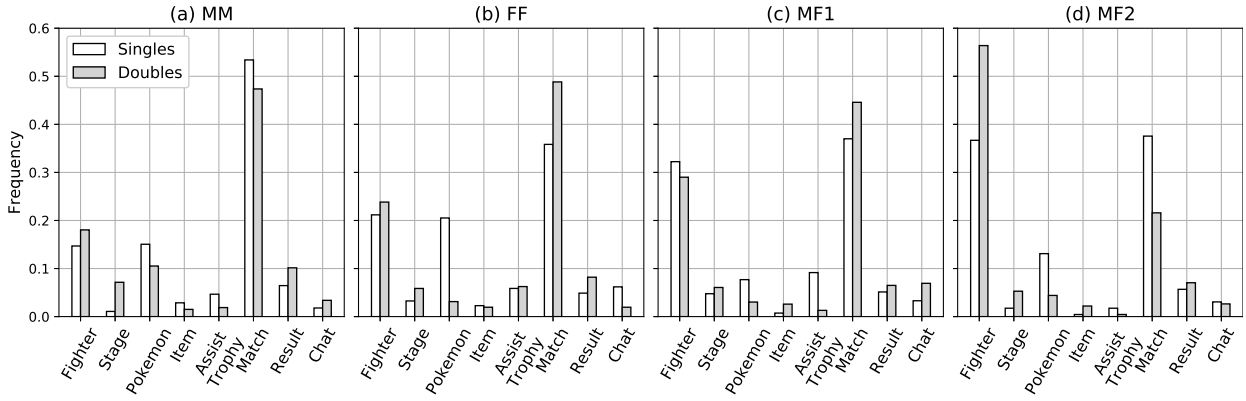


Fig. 6　トピックタグの分布．ここで，(d) MF2 のゲームプレイに対して実況解説を収録したコメンテーターは他と異なる．

ている．この図より，同じコメンテーター (MC1) によって収録された実況解説におけるトピックタグの分布 (Fig. 6(a)–(c)) は，プレイヤーペアの違いに依らず類似した形状となっている．一方で，別のコメンテーター (MC2) による実況解説のトピックタグの分布 (Fig. 6(d)) は他のものと大きく異なる．これらの結果は，(1) ゲームプレイの設定が固定であれば，同じコメンテーターはプレイヤーの違いに依らず類似したトピックの実況解説をすること，(2) コメンテーターの違いは，収録される実況解説のトピックの傾向に大きく影響を与えることが示唆された．

6　まとめと今後の展望

本稿では，ゲームプレイの実況解説に基づく自発発話音声の収録法と，それに基づく SMASH コーパスの構築について述べた．今後は，より複雑なルール設定などを用いた場合のゲームプレイ及び実況解説音声の収録，実況解説音声に表出する話者性の違いの分析，マルチモーダルなアノテーションの付与を行う．また，新たな音声言語情報処理技術として，動画の様子を音声で説明する際の発話文自動生成・自動読み上げの実現を検討する．

謝辞: 本研究の一部は，JSPS 科研費 18J22090，19H01116 の助成と，東京大学の丹治 尚子氏，田丸 浩気氏，佐伯 高明氏，三井 健太郎氏の協力を受け実施した．

参考文献

[1] S. Furui, *Proc. SSPR,* Tokyo, Japan, Apr. 2003.

[2] S. Werner *et al.*, *IEEE Trans. on SAP,* vol. 12, no. 4, pp. 436–445, July 2004.

[3] S. Furui *et al.*, *IEEE Trans. on SAP,* vol. 12, no. 4, pp. 401–408, July 2004.

[4] K. Maekawa *et al.*, *Proc. LREC,* pp. 947–952, Athens, Greece, May 2000.

[5] Y. Arimoto *et al.*, *Acoustical Science and Technology,* vol. 33, no. 6, pp. 359–369, June 2012.

[6] D. K. Roy, *Computer Speech and Language,* vol. 16, no. 3–4, pp. 353—385, July 2002.