多ジャンルのスポーツ音声実況における音声特徴量の時間的構造の調査

松下 嶺佑[1,a)]　高道 慎之介[1,2,b)]　齋藤 佑樹[2]　ニュービッグ グラム[3]　須藤 克仁[4]　高村 大也[5]
石垣 達也[5]

概要：本研究では，スポーツの試合に付与された実況音声を対象とし，実況音声の特徴を時間的構造の観点から定量的に分析する．実況は，試合映像の価値や臨場感を高め，視聴者の理解や没入感を促進する重要な役割を担う一方で，その表現様式や実況者に求められるスキルは競技ごとに異なる．しかし，競技間で実況音声の時間的構造や音声特徴量がどのように異なるのかについては，十分に明らかにされていない．そこで本研究では，スポーツの実況音声に対して基本周波数（F_0）や発話速度といった音声特徴量を抽出し，それらの時間的変化に着目した分析を行う．これにより，スポーツごとの実況の傾向や音声的特徴の違いを明らかにすることを目的とする．本研究の成果は，自動実況生成システムの自動評価に資する知見を提供することが期待される．

1. はじめに

本研究で扱う実況とは，プレイヤー以外の人間が，試合を盛り上げることを目的として行う音声表現を指す．実況の役割としては，試合映像の価値を高めること，臨場感を提供すること，および競技知識が十分でない視聴者でも試合内容を理解できるようにすることが挙げられる [1], [2]．また，スポーツ実況に限っては，競技の魅力を細大もらさず視聴者に伝えることや，主役である選手を最大限引き立てることも重要な役割とされている [3], [4]．特に，本研究で対象とするような試合を盛り上げることを主眼とした実況は，試合情報を客観的に解説する実況と比較して，視聴者に盛り上がりの感情を喚起しやすく，試合に引き込まれる体験を生みやすいことが報告されている [3]．

このように，実況は試合観戦体験に大きな影響を与える一方で，適切に実況するには，実況対象となる競技やコンテンツに関する十分な知識と，高度な実況スキルが必要とされる [1], [5]．さらに，良いとされる実況の在り方はスポーツによって異なることが指摘されている [4], [6], [7]．

一方で，スポーツやボードゲーム等を対象とした自動実況生成の研究が進展しており，実際に人間の実況を模倣した音声やテキストを生成するシステムが提案されている [1], [8], [9]．これらの技術は，実況者不足の解消やコンテンツ制作の効率化といった観点から高い期待を集めているが，生成された実況の実況らしさをどのように評価すべきかについては，いまだ統一的な指標が確立されていない．多くの場合，評価は主観的評価や経験則に基づいて行われており，人間実況との違いやスポーツジャンルごとの特性を音声的観点から定量的に比較・分析する枠組みは十分に整備されていない．このことは，自動実況生成技術の客観的な性能評価や改良を進める上での課題となっている．

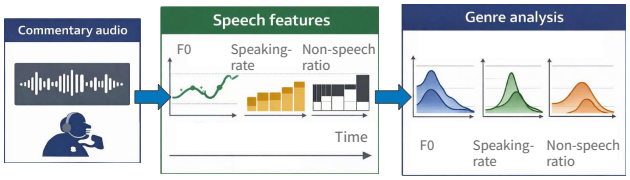

図 1: 本研究の概要．実況音声の特徴を定量的に分析する．

そこで本研究では，複数スポーツの実況を対象とし，実況音声の特徴を定量的に分析することで，実況らしさを明らかにする．本研究の示すスポーツは，リアルスポーツも e スポーツも対象とする．具体的には，図 1 に示されたように，試合に付与された実況音声の基本周波数 F_0 や発話速度，非発話率に着目し，スポーツごとの実況の傾向を分析する．また，本研究を進めるにあたり，各スポーツについての実況音声データはあるが，スポーツ間を比較できるデータセットがないため，その作成も行う．本研究の成果は，プロの実況を模倣した AI 実況生成のための設計指針

1 慶應義塾大学
2 東京大学
3 カーネギーメロン大学
4 奈良女子大学
5 国立研究開発法人産業技術総合研究所
a) ryosuke.jp66@keio.jp
b) shinnosuke_takamichi@keio.jp

となることや，試合観戦体験の最適化に資することが期待される．

2. 関連研究

2.1 スポーツごとに良いとされる実況

良い実況音声には，スポーツを問わず共通する発話特徴が存在することが指摘されている [4], [10], [11]．具体的には，必要以上に感情を誇張せず，重要な局面では要点を明確に示す張りのある声で締めること，選手の動作や試合状況を的確かつ迅速に描写すること，さらに試合全体の流れを把握した上で，解説者に対して適切なタイミングで質問を投げかけることなどが挙げられる [4]．

一方で，望ましい実況のあり方は競技によって異なることも報告されている．例えば野球中継では，ランナーが存在する局面において即時的なプレー描写が強く求められる一方で，それ以外の場面では配球の意図や選手の細かな仕草など，間を活かした情報提示が重視される．また，ラグビー中継では詳細なルール説明よりも，プレーの迫力や競技固有の魅力を直感的に伝える実況が重要視される [4]．さらに，スポーツの実況は，従来のスポーツ実況と比較して専門用語に対する解説の比率が高いことが報告されている [10], [11]．

このように，実況における望ましい発話様式はスポーツごとに異なっており，実況音声の評価や分析においては競技を考慮する必要がある．本研究では，実況音声の特徴量にこれらの競技がどのように反映されるかを検証する．

2.2 マルチモーダルなデータセット

実況音声を対象とした研究においては，映像や数値情報と音声を統合的に扱うマルチモーダルデータセットの整備が重要な課題とされている．例えば，カーレースゲーム実況を対象とした研究では，映像情報および位置や速度などの数値データに加え，過去の発話情報を統合し，人間の実況に近い自然な発話生成を目的とした新たなタスクが提案されている [1], [12]．この研究では，発話の「いつ」「何を」話すかを段階的に決定する二段階生成タスクを導入し，レース展開に応じた発話タイミングと内容の対応関係を定量的に示している [1], [12]．

一方，サッカー中継を対象とした研究では，試合映像に実況音声，文字起こし，英語翻訳を付与し，発話ごとに開始時刻および終了時刻をアノテーションした大規模マルチモーダルデータセットが構築されている [13], [14], [15]．このデータセットにより，実況音声を含む統合的な分析が可能となり，イベント検出やハイライト生成など，多様な応用への有用性が示されている．

本研究においても，試合映像，実況音声，および発話内容を示したテキストを含むマルチモーダルなデータセットを用いて実験を行う．具体的には，各スポーツについての実況音声データはあるが，スポーツ間を比較できるデータセットがないため，それを作成する．

2.3 実況らしさを司る音声特徴量

実況の音声特徴量に関する研究では，興奮や緊張といった感情状態が音声韻律にどのように反映されるかが定量的に分析されている [16], [17], [18]．例えば，競馬実況を対象とした先行研究では，実況音声を序盤・中盤・終盤の三つのフェーズに分割し，各区間における F_0，音圧，話速など複数の韻律指標を比較することで，盛り上がりの音声的特徴を明らかにしている [18]．その結果，レース終盤に向かって F_0 や音圧が上昇し，声の張りが増加するなど，興奮状態に特有の韻律変化が確認されている．

また，スポーツを対象として盛り上がりの度合いを検証した研究も存在する．例えば，大乱闘スマッシュブラザーズ SPECIAL を対象とした研究 [17] では，対戦状況やゲーム内イベント情報から盛り上がり度を定義し，盛り上がりレベルを条件情報として入力することで，同一内容の発話であっても，より興奮した実況や落ち着いた解説を合成できる可能性が示されている．

しかし，これらの研究は単一スポーツの実況に焦点を当てたものが多く，競技の異なる他スポーツに対して同様の韻律的傾向が成り立つかについては十分に検討されていない．本研究では，競技を限定せず，複数のスポーツの実況を対象とし，盛り上がり表現の音声的特徴を横断的に分析することを試みる．

3. 提案手法

本研究で用いた手法は，データセットの構築と実況音声の分析の二つに大別される．データセットの構築では，スポーツ間を比較できるデータセットがないため，それを作成した．具体的には，試合動画の収集，試合開始および終了時刻の付与，ならびに実況音声に対する音声強調処理を行った．実況音声の分析では，実況音声の特徴量を定量的に分析するために，基本周波数 F_0 の抽出および発話速度の測定を行い，それらの時間的変化を定量的に評価した．

3.1 データセットの構築

3.1.1 試合動画の収集

実況音声が付与されたスポーツの試合動画は，YouTube 上に公開されている動画を対象に収集した．対象とする動画は，公式大会もしくはそれに準ずる試合を扱ったものであり，実況者によるリアルタイム実況音声が明瞭に含まれているものに限定した．また，試合進行が途切れずに記録されている動画のみを選定し，試合をダイジェスト形式で編集した動画は分析対象から除外した．動画のダウンロー

ドには yt-dlp[*1]を用いた．

3.1.2 試合開始，終了時刻の付与

ダウンロードした実況付き試合動画には，1 本の動画内に複数の試合が含まれる場合が多く存在した．そこで，各動画に含まれる試合ごとに開始時刻および終了時刻を付与した．この処理により，試合間に挿入される CM や雑談など，分析対象外の区間を除去することを目的とした．

付与方法としては，まず試合映像中の特定の動画フレームに注目した．これらのフレームには，スコア表示やラウンド開始表示など，試合の進行段階を示す特徴的な画面が含まれているため，本研究ではあらかじめ取得した試合開始・終了画面の画像を参照した．次に，それらのフレームを注目領域（ROI）として設定し，それらと視覚的に一致するフレームを検出することで，試合区間の開始時刻および終了時刻を決定した．また，一部の動画については，YouTube 上の概要欄に記載された試合開始・終了時刻のメタデータ情報を参照し，映像内容と整合することを確認した上で区間の特定に用いた．

しかし，これらの方法が適用可能な動画は限定的であったため，一部の動画については手作業により開始および終了時刻を付与した．

3.1.3 実況音声の音声強調

実況音声の分析において，ゲーム内 BGM や観客の歓声などの背景音が含まれると，実況者の発話に由来する音響特徴量を正確に測定することが困難となる．そこで，ダウンロードした実況音声に対して sidon [19] を用いた音声強調処理を行い，背景雑音やゲーム内 BGM の抑圧を行った．

3.1.4 実況音声の発話文字起こし

発話区間の決定には，実況音声を入力とし，音声認識モデルである Whisper [20] base を用いて自動文字起こしを行った．Whisper は各発話に対して開始時刻および終了時刻を含むタイムスタンプ情報を出力するため，本研究ではこの情報を用いて，実況音声中の各発話区間を推定した．

3.2 分析

以降の分析では， 3.1.3 節で準備した強調済み実況音声を使い，分析を行う．また，発話区間は 3.1.4 節で抽出した実況発話書き起こしを参考に決定する．

3.2.1 F_0 の抽出

まず，基本周波数 F_0 の抽出を行った．その際に，倍ピッチおよびハーフピッチを防ぐために，F_0 の探索範囲を設定し，各発話区間において平均 F_0 を算出した．

さらに，試合進行に伴う F_0 の変化を定量的に評価するため，式 (1) を以下に示す:

$$F_0(t) \approx at + b \qquad (0 \le t \le 1). \tag{1}$$

ここで，t は，試合時間を正規化したものを指す．また，係数 a, b は:

$$(a, b) = \arg\min_{a,b} \sum_{i=1}^{N} (F_0(t_i) - (at_i + b))^2 . \tag{2}$$

最小二乗法により推定した．ここで，N は当該区間において線形近似に用いた $F_0(t_i)$ のサンプル数を表す．

3.2.2 発話速度の抽出

本研究における発話速度は，以下の式により定義される:

$$\text{発話速度} = \frac{\text{発話区間内のモーラ数}}{\text{発話区間の時間長}}. \tag{3}$$

この時，各発話区間は，連続した発話が行われている区間として定義し，音声中に一定時間以上の非発話状態が検出されるまでを 1 つの発話区間とした．このようにして得られた各発話区間に対し，区間内の発話長および文字数を用いて発話速度を算出した．

得られた発話速度についても，試合進行に伴う変化を評価するため，式 (1) 同様に，最小二乗法により一次関数で近似した．

3.2.3 非発話率の測定

本研究における発話速度は，以下の式により定義される:

$$\text{非発話率} = \frac{\text{非発話区間の時間}}{\text{試合時間}}. \tag{4}$$

試合時間全体の非発話率だけでなく，試合時間を 3 分割し，序盤，中盤，終盤に分け，それぞれの非発話率について測定した．そして，非発話時間の時間長の分布を考察するために，非発話時間の分布のヒストグラムを作成した．その後，ヒストグラムを以下の関数でフィッティングする [21].

$$g(x) = \begin{cases} \dfrac{A}{x} \exp\left(-b\,(\log x - \log t)^2\right) & , \quad x > 0 \end{cases} \tag{5}$$

t は, ヒストグラムの中央値に当たる時の非発話時間である．この関数は，$x = t$ を中心とした対数ガウス関数である．$a \ge 0$ はスケールパラメータ，$b > 0$ は精度パラメータである．

4. 実験的評価

4.1 実験条件

4.1.1 対象データ

本研究では，YouTube 上で公開されているスポーツの試合映像を対象とした．対象スポーツは計 15 種であり，具体的なスポーツ名は表 1 に示す．

また，試合開始終了時刻のアノテーション方法も同様に表 1 に示す．また，アノテーション結果は GitHub[*2]にオープンソースとして公開する．試合開始，終了時刻のアノテーションにより，1 試合ずつに切り出した各スポーツの試合の総数も同表に示す．また，Whisper [20] による文字起こし結果は同様に GitHub にて公開している．

[*1] https://github.com/yt-dlp/yt-dlp

[*2] https://github.com/takamichi-lab/matsushita-M1-anygenre-commentary

表 1: データセットに用いたスポーツ，試合開始終了時刻のアノテーション方法，試合数

Sports	Start-end annotation	#matches
apex	メタ情報	62
baseball	必要なし	21
basketball	手動	13
dota2	テンプレートマッチング	11
hearthstone	手動	7
judo	手動	16
keiba	必要なし	28
league of legends	手動	17
legends of runeterra	テンプレートマッチング	10
mahjong	手動	44
poker	手動	12
shadowverse	メタ情報	59
street fighter	手動	18
tekken	手動	88
valorant	テンプレートマッチング	389
total		795

4.1.2 F_0 抽出

F_0 抽出には librosa の pyin[*3] を用い，探索範囲を $f_{\min} = 60$ Hz，$f_{\max} = 450$ Hz と設定した．有声フレームからなる連続区間を有声セグメントとして抽出した．各有声セグメントに対して，セグメント内の平均値を平均 F_0 として算出し，セグメントの開始時刻および終了時刻は，それぞれ当該セグメント内で最初および最後に得られた有声フレームの時刻として定義した．

4.1.3 発話区間

発話区間は WebRTC VAD[*4] を用いて検出を行った．この時，発話区間として採用する最短長を 0.30 sec とした．0.30 sec 未満の非発話区間は発話区間の区切りとして扱わず，前後の発話を結合した．さらに，Whisper の文字起こしとの誤該当を防ぐべく，発話区間と発話文字起こしの対応付けには，両者の重なり時間が 0.70 sec 以上であることを条件とした．

4.2 F_0 に関する分析

F_0 が試合時間の進行に伴ってどのように変化するかを調査するため，各スポーツに，横軸を F_0 の傾き，縦軸を回帰直線からの RMSE とした散布図を作成した．横軸の絶対値が大きくなるほど，当該 スポーツは序盤から終盤にかけて F_0 が上昇／下降することを表す．縦軸の値が大きいほど，F_0 の変化が時刻に対して線形でないことを表す．図 2 は各スポーツにおける平均的な傾向を示しており，図 3 は各試合の傾向を 2 次元のカーネル密度推定（2D-KDE）で示している．2D-KDE にはガウスカーネルを用い，各試合の傾向分布を滑らかに推定した．

[*3] https://librosa.org/doc/main/generated/librosa.pyin.html
[*4] https://github.com/wiseman/py-webrtcvad

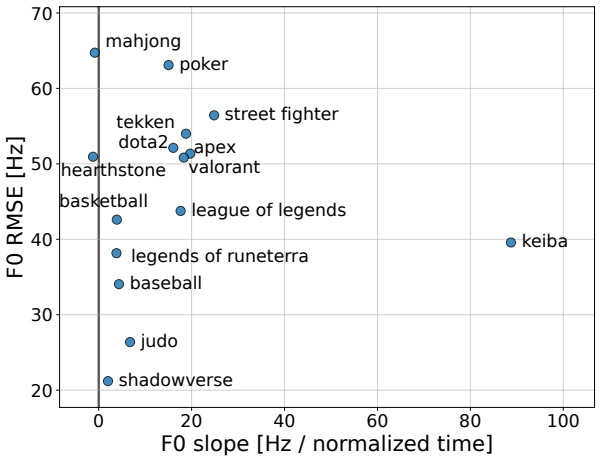

図 2: 実況音声のスポーツ別の F_0 傾きと RMSE の散布図．

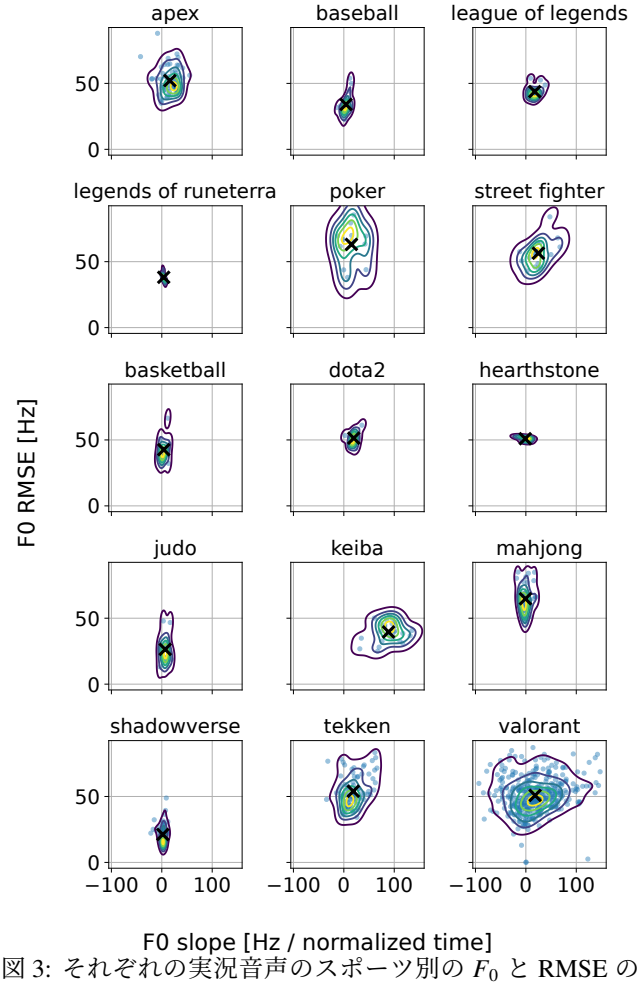

図 3: それぞれの実況音声のスポーツ別の F_0 と RMSE の 2D-KDE．バツは全試合の平均値，点は各試合の値を指す．

図 2 に示されるように，全体的な傾向として，F_0 の傾きが正となるスポーツが多く見られた．これは，試合の序盤から終盤にかけて，実況音声における盛り上がり度合いが増加している可能性を示唆している．

特に競馬は他のスポーツと顕著に異なり，F_0 の高い傾きと低い RMSE を示している．すなわち，競馬は序盤から終盤にかけて F_0 が典型的に上昇することを表す．これは，

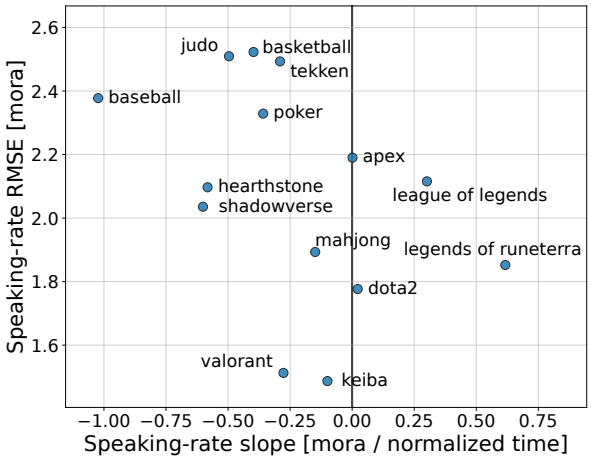

図 4: 実況音声のスポーツ別の平均発話速度と RMSE の散布図.

競馬が終盤において順位変動を生じさせやすいことを鑑みれば，妥当な結果と言える．

また，RMSE が大きいスポーツも存在した．これは，序盤や終盤といった時間的区分に依らず，実況の盛り上がりが変動していることを示している．具体的には，麻雀や poker といった頭脳戦型のゲームにおいてこの傾向が見られた．これらのスポーツでは，試合の進行段階に関わらず，勝負を左右する局面が突発的に訪れるため，盛り上がりが時間変化に必ずしも依存しなかったと考えられる．

図 3 に示されるように，legends of runeterra や hearthstone，shadowverse では，分布が偏っている．これは，この 3 つのゲームが対戦カードゲームというジャンルであり，実況中の発話内容が戦況の分析や次の手の予測など，論理的で落ち着いた説明に偏りやすいためであると考えられる．

4.3 発話速度に関する分析

発話速度が試合時間の進行に伴ってどのように変化するかを調査するため，各スポーツに，横軸を発話速度の傾き，縦軸を 発話速度の RMSE とした散布図を作成した．横軸の絶対値が大きくなるほど，当該 スポーツは序盤から終盤にかけて発話速度が上昇／下降することを表す．縦軸の値が大きいほど，発話速度の変化が時刻に対して線形でないことを表す．図 4 は各スポーツにおける平均的な傾向を示しており，図 5 は各実況音声の傾向を 2 次元のカーネル密度推定（2D-KDE）で示している．

図 4 に示されるように，全体的な傾向として，発話速度の傾きは絶対値として小さい値を取っており，実況中を通して発話速度の時間的変動が比較的小さいことが分かる．これは，多くの実況者が試合進行に応じて急激に話速を変化させるのではなく，一定のテンポを保ちながら実況を行っていることを示唆している．

また，多くのスポーツにおいて RMSE の値が比較的大

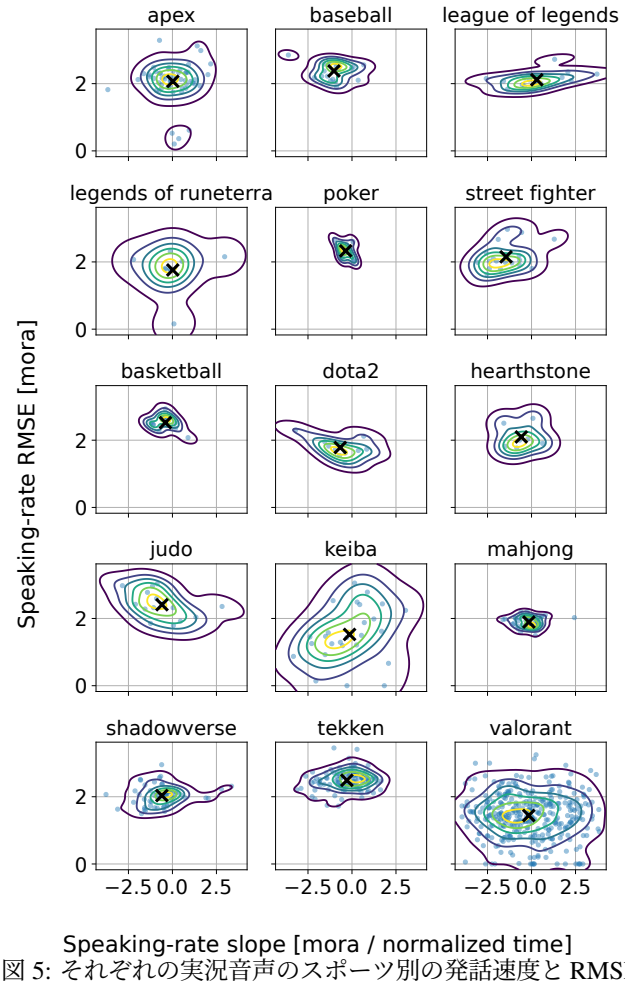

図 5: それぞれの実況音声のスポーツ別の発話速度と RMSE の 2D-KDE．バツは全試合の平均値，点は各試合の値を指す．

きい傾向が確認された．これは，発話速度の変動が試合時間の進行に一様に依存するものではなく，局所的なイベントや状況変化によって大きく変動していることを示唆している．

特に，柔道やバスケットボールでは RMSE の値が大きい．これは，柔道の技をかける場面や，バスケットボールのゴール場面といった重要なイベントが，試合時間の特定の位置に集中せず，試合中のどのタイミングでも起こり得るためである．その結果，イベントの発生に応じて実況者の発話速度が変化すると考えられる．

次に，図 5 に示されるように，競馬や valorant では，各試合の分布にばらつきがあることがわかる．これは，競馬や valorant では試合展開や局面の緊張度が各試合で大きく異なり，それに応じて実況者が発話速度を柔軟に変化させているためと考えられる．すなわち，これらのスポーツでは実況スタイルが試合内容に強く依存し，結果として試合間で発話速度の分布にばらつきが生じやすいことが示唆される．

一方で，麻雀や poker では，各試合の分布にばらつきが少ない．麻雀や poker では，相手の手牌や心理を推測する

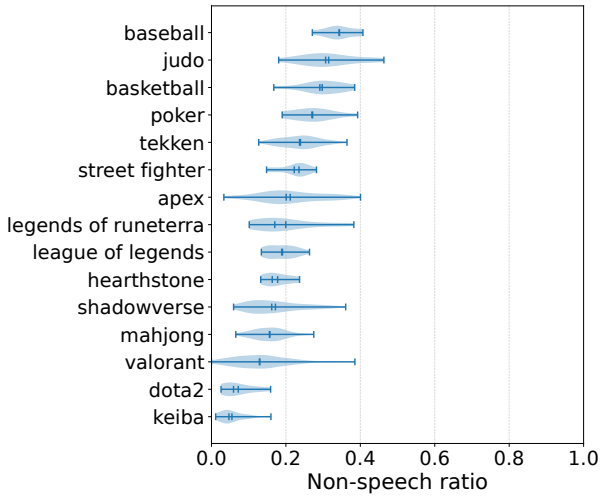

図 6: 試合時間全体の非発話率のバイオリンプロット．

局面が多く，実況者が状況を整理しながら説明する時間が確保されるため，発話速度が各試合で大きく変動しにくいと考えられる．

4.4 非発話率の測定結果

4.4.1 試合時間帯別の非発話率

非発話率が試合時間の進行に伴ってどのように変化するかを調査するため，各スポーツに非発話率のバイオリンプロットを作成した．図 6 は各スポーツにおける平均的な非発話率の分布を示しており，図 7 は，試合時間を 0 から 1 に正規化した後，序盤，中盤，終盤の 3 区間に分割して測定した非発話率の分布を示している．

図 6 に示されるように，競馬，dota2，valorant では，他のスポーツと比較して非発話率が低い傾向が確認された．これらのスポーツに共通する特徴として，試合中に状況変化が頻繁に発生する点が挙げられる．特に競馬では，レース中に順位変動が連続して発生し，実況者がレース展開を継続的に描写する必要があるため，非発話区間が短くなる傾向が生じたと考えられる．

一方で，野球，柔道，バスケットボール，poker といったスポーツでは，非発話率が比較的高い傾向が見られた．これらの競技に共通する点として，プレー中に戦略的な駆け引きや心理的要素が強く関与することが挙げられる．例えば野球では，投手と打者の配球を巡る駆け引きや，走者のスタートタイミングを巡る緊張状態が発生する．このような場面では，実況者が状況を注視し，発話を控える時間が生じるため，非発話率が高くなった可能性が示唆される．

図 7 に示されるように，全体としては，試合時間の進行に伴う非発話率の大きな変化は確認されなかった．しかし，apex においては，終盤にかけて非発話率が低下する傾向が見られた．これは，試合終盤において対戦が激化し，状況変化が頻発するため，実況者が逐次的に状況を説明する必要が生じたことに起因すると考えられる．

表 2: スポーツ別の対数ガウス関数のパラメータ．Peak Position は Peak Height の時の非発話時間を示す．

Sports	Peak Position	Peak Height	Precision (b)
apex	0.447	0.072	0.753
baseball	0.459	0.057	0.569
basketball	0.510	0.065	0.794
dota2	0.440	0.101	1.173
hearthstone	0.354	0.087	0.675
judo	0.398	0.072	0.646
keiba	0.349	0.120	1.069
league of legends	0.465	0.081	0.959
legends of runeterra	0.397	0.070	0.600
mahjong	0.494	0.101	1.497
poker	0.397	0.091	0.900
shadowverse	0.452	0.088	1.027
street fighter	0.496	0.087	1.164
tekken	0.505	0.083	1.131
valorant	0.468	0.085	1.010

4.4.2 各スポーツの非発話時間の分布

また，非発話区間の時間長を調べるために，ヒストグラムを作成し，ヒストグラムの概形を回帰曲線でフィッティングした．また，フィッティングした関数のピークの位置、ピークの高さ、精度を表 2 に示す．

図 8 から示される通り，非発話長分布を見ると，いずれのスポーツにおいても短い非発話が高頻度で出現し，非発話長が長くなるにつれて確率が単調に減少する分布形状を示している．表 2 からもわかるように，非発話時間の長さのピークは，およそ 0.4 – 0.5 sec に収まっている．これは，実況音声において短いポーズが頻繁に挿入される一方で，長時間の沈黙はまれであり，スポーツを問わず実況者が試合状況を継続的に言語化していることを反映していると考えられる．

次に，ピークの高さは分布の集中度を示す指標であり，非発話率が特定の値に強く偏るスポーツほど高くなる．特に競馬や dota2 はピーク高さが 0.10 以上と大きく，実況スタイルが比較的一様であることが示唆される．これらは，試合中に状況変化が頻繁に発生し，その都度，状況説明の実況すると考えられるので，非発話時間が短いことが多くなったと考えられる．

一方で，精度は，分布の裾の広がりを示す指標であり，この数値が高いほど分布の裾の広がりは狭くなる．これは各スポーツの差異が見られた．具体的には，麻雀では最も高く 1.50 付近となった．麻雀では，プレイヤーの動作や，試合の流れに決まったものがあり，そのため，麻雀実況では各局面の沈黙と解説が一定のリズムで繰り返されることが示唆される．反対に，野球や legends of runeterra では，0.57 – 0.60 付近と小さい値となり，非発話率が特定の値に集中せず広い範囲に分布していることを示している．これは，実況者の発話量が場面によって大きく変動しやすく，

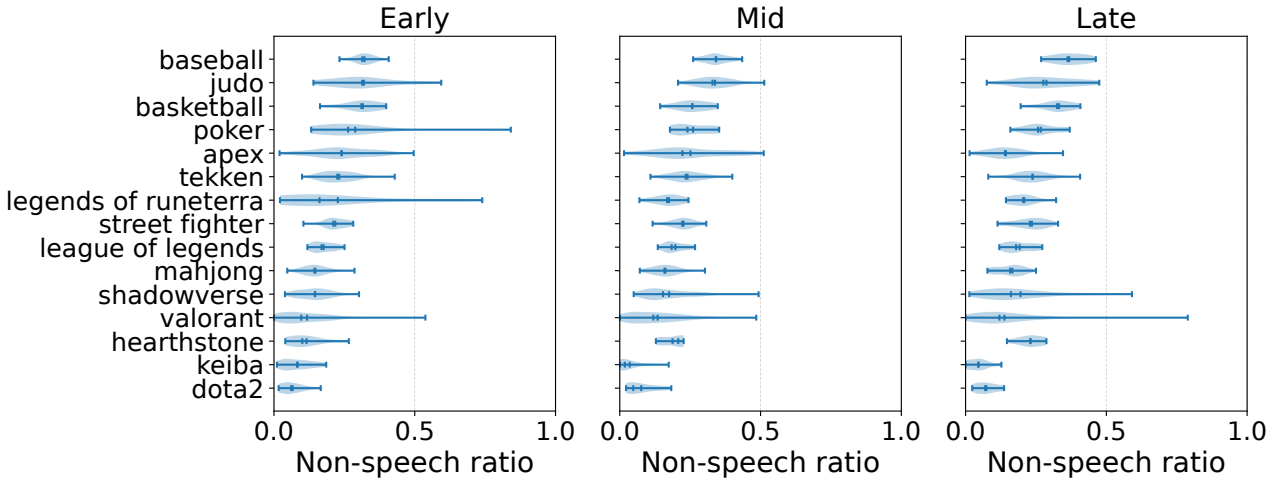

図 7: 試合時間を 3 分割にしたときの非発話率のバイオリンプロット．

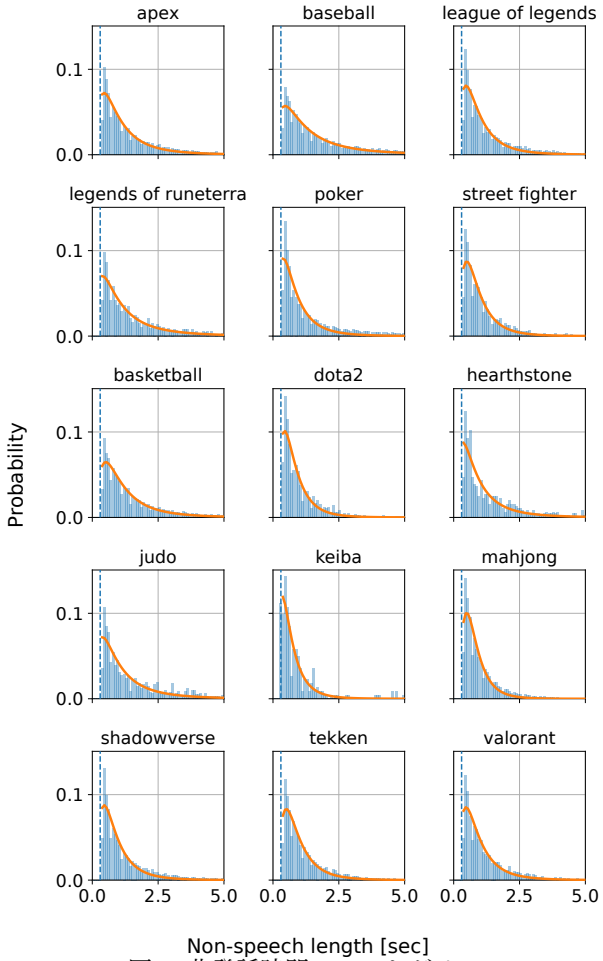

図 8: 非発話時間のヒストグラム．

沈黙と発話のリズムが一定になりにくいジャンルであるためと考えられる．

また，図 6 から示されるように，poker とバスケットボールは非発話率の平均値が近い．しかし，図 8，表 2 から示される通り，非発話時間の分布は異なる．具体的には，バスケットボールの方が，ピークとなる非発話時間が長く，分布の裾の広がりを示す精度の値も小さいので，より非発話時間の長さが長いことが示される．これは，バスケットボールでは，タイムアウトや，ボールがコートの外に出るといったプレイの切れ目に伴い，比較的長い非発話が生じやすいと考えられる．一方で，poker は，次の行動が確定するまでの思考的な間合いや心理的駆け引きによって非発話が発生するので，このような違いが生まれたと考えられる．

5. まとめ

本研究では，スポーツの実況を対象とし，実況音声の特徴を定量的に分析することで，盛り上がりの側面から実況とスポーツとの関係性を明らかにすることを目的とした．特に，実況音声における基本周波数 F_0 および発話速度，非発話率に着目し，試合進行に伴うそれらの時間的変化を分析した．

分析の結果，実況音声の F_0 や発話速度，非発話率の推移にはスポーツごとに異なる傾向が見られ，競技特性が実況音声の韻律的特徴に反映されている可能性が示唆された．これは，良いとされる実況の在り方がスポーツによって異なるという先行研究の知見を，音声的特徴に基づいて定量的に支持する結果であると考えられる．

謝辞：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る．また，JSPS 科研費 21H04900，JST 創発的研究支援事業 JPMJFR226V の支援を受けて実施した．

参考文献

[1] Ishigaki, T., Topic, G., Hamazono, Y., Noji, H., Kobayashi, I., Miyao, Y. and Takamura, H.: Generating Racing Game Commentary from Vision, Language, and Structured Data, *Proc. INLG*, pp. 103–113 (2021).

[2] 小渡悟：e-sports 観戦における実況解説の有無による

影響，教育システム情報学会 2022 年度学生研究発表会，Vol. 1, pp. 257–258 (2023).

[3] Lee, M., Kim, D., Williams, A. S. and Pedersen, P. M.: Investigating the Role of Sports Commentary: An Analysis of Media-Consumption Behavior and Programmatic Quality and Satisfaction, *Journal of Sports Media*, Vol. 11, No. 1, pp. 145–167 (2016).

[4] 山本浩：スポーツアナウンサー――実況の真髄，岩波書店，東京 (2021).

[5] Taniguchi, Y., Feng, Y., Takamura, H. and Okumura, M.: Generating Live Soccer-Match Commentary from Play Data, *Proc. AAAI* (2019).

[6] Li, L., Uttarapong, J., Freeman, G. and Wohn, D. Y.: Spontaneous, Yet Studious: Esports Commentators' Live Performance and Self-Presentation Practices, *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, No. CSCW2, pp. 1–25 (2020).

[7] Kempe-Cook, L., Sher, S. T.-H. and Su, N. M.: Behind the Voices: The Practice and Challenges of Esports Casters, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Association for Computing Machinery, pp. 565:1–565:12 (online), DOI: 10.1145/3290605.3300795 (2019).

[8] Sadikov, A., Možina, M., Guid, M., Krivec, J. and Bratko, I.: Automated Chess Tutor, *Computers and Games* (van den Herik, H. J., Ciancarini, P. and Donkers, H. H. L. M. J., eds.), pp. 13–25 (2007).

[9] Sun, Z., Chen, J., Zhou, H., Zhou, D., Li, L. and Jiang, M.: GraspSnooker: Automatic Chinese Commentary Generation for Snooker Videos, *Proc. IJCAI*, pp. 6569–6571 (2019).

[10] Seo, Y.: Electronic sports: A new marketing landscape of the experience economy, *Journal of Marketing Management*, Vol. 29, No. 13-14, pp. 1542–1560 (online), DOI: 10.1080/0267257X.2013.822906 (2013).

[11] Stuart Reeves, B. B. and Laurier, E.: Experts at play: understanding skilled expertise, *Games and Culture*, Vol. 4, No. 3, p. 205–227 (online), DOI: 10.1177/1555412009339730 (2009).

[12] Marrese-Taylor, E., Hamazono, Y., Ishigaki, T., Topić, G., Miyao, Y., Kobayashi, I. and Takamura, H.: Open-domain Video Commentary Generation, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7326–7339 (2022).

[13] Rao, J., Wu, H., Liu, C., Wang, Y. and Xie, W.: MatchTime: Towards Automatic Soccer Game Commentary Generation, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (2024).

[14] Suglia, A., Lopes, J., Bastianelli, E., Vanzo, A., Agarwal, S., Nikandrou, M., Yu, L., Konstas, I. and Rieser, V.: Going for GOAL: A Resource for Grounded Football Commentaries, *CIKM 23* (2022).

[15] Gautam, S., Sarkhoosh, M. H., Held, J., Midoglu, C., Cioppa, A., Giancola, S., Thambawita, V., Riegler, M. A., Halvorsen, P. and Shah, M.: SoccerNet-Echoes: A Soccer Game Audio Commentary Dataset, *2024 International Symposium on Multimedia (ISM)*, IEEE, p. 71–78 (2024).

[16] Audibert, N., Aubergé, V. and Rilliard, A.: Prosodic correlates of acted vs. spontaneous discrimination of expressive speech: a pilot study, *Speech Prosody 2010*, p. paper 097 (2010).

[17] 井浦昂太，齋藤佑樹，高道慎之介，ニュービッググラム，須藤克仁，猿渡洋，高村大也，石垣達也：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，第 155 回音声言語情報処理研究発表会 (2025).

[18] Trouvain, J. and Barry, W. J.: The Prosody of Excitement in Horse Race Commentaries, *Proceedings of the ISCA Workshop on Speech and Emotion*, ISCA, pp. 86–91 (2000).

[19] Nakata, W., Saito, Y., Ueda, Y. and Saruwatari, H.: Sidon: Fast and Robust Open-Source Multilingual Speech Restoration for Large-scale Dataset Cleansing (2025).

[20] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I.: Robust speech recognition via large-scale weak supervision, *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, JMLR.org (2023).

[21] Matsushita, R., Sakai, R., Fukuda, K., Takamichi, S., Iura, K., Saito, Y., Neubig, G., Sudoh, K., Takamura, H. and Ishigaki, T.: Measuring Time Delay Tolerance in Third-Person Live Commentary for Super Smash Bros. Ultimate, *IEEE Conference on Games* (2025).