スポーツ実況のための日本語固有名詞音声認識

Japanese proper noun speech recognition for Live Sports Situations

野々村 美紗
Misa Nonomura
法政大学情報科学部ディジタルメディア学科 21K1126
misa.nonomura.4c@stu.hosei.ac.jp

abstract

This research proposes a Japanese proper noun speech recognition system for sports live audio. Currently, TV subtitles are assigned manually and partially using speech recognition.There is a need for subtitling using full speech recognition.In this study, the accuracy of proper nouns in sports live audio was improved.There are several effective ways to improve recognition performance of proper nouns, but a method using a Large language Model is focused on.First, a list containing several results from existing speech recognition systems is output.The list, natural language instructions and prior information are then prompted to a Large language Model to generate correct subtitles. These results showed improvements of approximately 2〜6% for CER and 5〜20% for PCER for the two large language models. We would like to apply the proposed method to various situations in the future.

1 はじめに

音声情報処理の分野では、音声認識の研究が進められており、日本語の音声認識の性能もかなり高くまで上がっている。音声認識は、情報にアクセスする利便性を高めるために有用である。しかし、日本語の音声認識は英語等に比べ認識性能が低い。その理由として挙げられるのは、常用する文字 (ひらがな 46 字、カタカナ 46 字、常用漢字 2136 字) が多く、同音異義語が多数、単語の区切りが曖昧であること、固有名詞の多さなどである。今回はこれらの課題の中でも、固有名詞に着目したい。固有名詞は、通常学習データ中に登場する回数が少ないために、正確に認識されないことが多い。以下の表 1 は今回収集したデータからランダムに 100 発話抽出し、既存の音声認識システムを用いた認識結果で単語誤り率を求めたものである。固有名詞の誤り率は高くなっている。しかし、固有名詞は文章の中で重要な意味を占めていることが多い。例えば、「次はアゼルバイジャンのトカエフのレースだ」という文で、固有名詞である「アゼルバイジャン」や「トカエフ」が正しく認識されなければ、文章の意図は全く伝わらない。そのため、固有名詞を正しく認識することは重要である。

表 1 単語誤り率

CER(文全体)	CER(固有名詞部分)
21.24%	32.54%

また、2024 年夏はパリオリンピックが開催されスポーツ実況を目にする機会が多かった。本稿ではスポーツ実況音声に焦点を当てたい。スポーツ実況では、技の合間に解説が挟まれ、選手名、技の名前、ルールなどスポーツに応じた様々な固有名詞が登場するが、これらの単語は一般的ではなく、そのスポーツに詳しくなければ聞き取るのが難しい。そのため、字幕を見ながらの番組視聴の需要がある。しかし、スポーツ実況はそれぞれの試合の曲面に応じて発する言葉が変わるため、原稿が用意されているようなニュースやドラマなどとは異なり、あらかじめ字幕を用意することが難しい。そのため、字幕を放送と同時に付与していかなければならない。

現在、実況音声字幕は大まかに 2 つの方式で付与されている。[1]1 つ目は、一般的なキーボードや特殊なキーボードを用いて人手で字幕を作成・付与する方式である。オペレーターが実況音声を聞きながら、リアルタイムで字幕を入力する。この方法は、即時性や正確性があるが、長時間の作業による負担や、作業効率の低下や人手不足などの課題がある。2 つ目の方式は、部分的に音声認識技術を活用した字幕付与である。この方式では、音声認識技術を適用する範囲を限定し、認識精度を向上させる工夫が行われている。たとえば、比較的認識精度が高いアナウンサーや解説者が発話している部分だけを音声認識で処理したり、専用の字幕制作担当者が番組音声を聞きながら復唱した音声を認識する形で字幕を生成したりしている。また、音声認識の誤りを手動で修正したり、音声認識したテキストをそのまま字幕にしたりしているものもある。しかし、すべての音声に音声認識を用いることは未だ難しい。以上の 2 つの方法で、日本では一般的に使用されない中国人選手の名前など、難しい漢字も正確に字幕化することできている。しかし、実況音声では上記の課題に加え、字幕に間違いが発生する場面が存在する。以下の表 2) はある番組の 10 分間の誤字脱字をまとめたものである。約 7.5% ほどの間違いがある。そのため、完全な音声認識による字幕の実現が求められる。

表 2 字幕の 10 分間の誤字脱字

置換誤り	挿入誤り	脱落誤り	総文字数
6	8	187	2678
0.22%	0.30%	6.98%	100%

第 2 節で詳しく述べるが、固有名詞などの稀な単語に対するアプローチは、音声認識システムのファインチューンや大規模

Supervisor: Prof. Katunobu Itou

言語モデルを用いた手法など様々な種類があるが、有効な手法は決定的ではない。

本研究では、音声認識システムから出力される ASR 仮説を使用した大規模言語モデルのプロンプトチューニングによって固有名詞の認識性能改善を行う。

2 関連研究

2.1 Whisper

OpenAI 社の Alec Radford らが開発した汎用音声認識モデル [2] である。音声認識だけでなく翻訳等も可能で、多言語多タスクに対応している。従来の音声認識モデルよりもノイズを含む音声の認識性能が高いことが特徴である。Web から収集した 68 万時間分の多言語音声データを教師付きデータで学習している end-to-end モデルで、日本語の word error rate は最大で 4.9% (FLEURS) という高い結果を出している。しかし、日本語の固有名詞に関しては誤認識してしまうことも多くある。前述した表 1 は Whisper-large で認識をした際の性能を示している。特に固有名詞の認識性能は向上の余地がある。

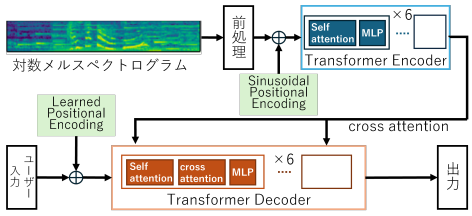

図 1 Whisper approach

2.2 従来の稀な単語に対応した音声認識

Jungwon Chang らは韓国語の特有語彙に焦点を当てファインチューニングを行った。[3] 大量のデータ (969 時間分) と少量のデータ (865 文) の韓国語データセットを比較してファインチューニングを行った。少量のデータでも語彙誤り率が相対的に 32% 改善した。しかし、日本語固有名詞が大量に含まれているデータセットは存在せず、作成にはコストがかかる。

2.3 大規模言語モデル (LLM) を活用した音声認識

また、Ziyang Ma らは大規模言語モデルを用いて英語音声認識の単語誤り率の改善を図った。その結果 3.81% の改善を見せている。[4] この研究では、音声埋め込みを用いて、プロンプトチューニングを行っており、事前に音声埋め込みでモデルをトレーニングしている。しかし、このトレーニングにもデータセットが不可欠である。また、研究の過程で音声認識タスクにおける大規模言語モデルのパフォーマンスの比較を行っていた。その結果、事前学習済みモデルよりも教師ありファインチューニングモデルが音声認識タスクには有効であるとわかった。公開されているものの中で LLaMA-2-Chat などのモデルと比較した結果、カルフォルニア大学等が開発する VICUNA[5] というモデルが音声認識タスクには最も優れていることが分かった。

2.3.1 プロンプトチューニング

Ada Defne Tur らは、ASR n-best の再スコアリングのための新しいゼロショット法を導入した。[6] 相対的に 単語誤り率 (WER) は 5% から 25% の範囲で有意な改善が見られた。

3 提案手法

本稿では、スポーツ実況音声で固有名詞の認識性能が高い音声認識の開発を目指す。音声認識システム Whisper から出力した疑似的な n-best リストを大規模言語モデルに与え、事前情報などとプロンプトチューニングを行っていくことで正しい字幕を生成する。(図 2)

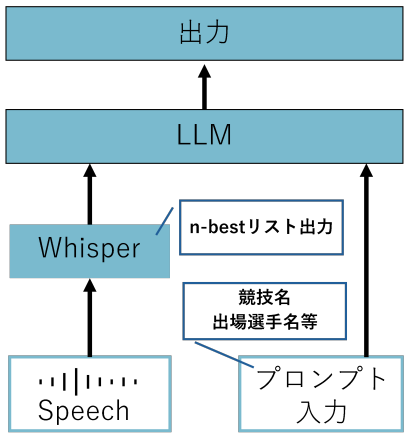

図 2 提案手法のモデル図

4 検討事項

4.1 音声認識

音声認識システムは様々なものが開発されているが、日本語固有名詞の認識を行うにあたり、日本語の認識性能がいいものを使用する。また、固有名詞の認識性能を上げるために n-best リストを用いた認識結果の修正を行う。n-best をプロンプトとして用いることで、未知語部分の認識結果に幅を持たせることができる。未知語を文脈に基づいて適切に補完、固有名詞を正確に認識、文法的に妥当な表現を優先的に選択することが期待できる。この手法は固有名詞が多く、ノイズの多い実況音声に適当である。

4.2 大規模言語モデル (LLM)

日本語固有名詞のエラー訂正に適している LLM は議論されていないが、英語の音声認識エラー訂正に有効な LLM は先行研究から VICUNA、GPT-4 などがあげられている。そのため、音声認識エラー訂正に有効な LLM が日本語固有名詞のエラー訂正に有効であるのかを検証する必要がある。また、LLM と同様に固有名詞のエラー訂正に有効なプロンプトも議論されていない。そのため、音声認識エラー訂正に有効なプロンプトを修正して、固有名詞のエラー訂正に有効なプロンプトを検証する必要がある。

4.3 手法

4.3.1 音声認識

音声認識には whisper の中で最も性能の良いモデル whisper-large モデルを用いる。ここでは、出力として複数の仮説が含まれているリスト疑似的な n-best リストを取得する。本来 n-best リストとは音声認識の複数の上位の仮説をその仮説がどれ

ほど尤もらしいか示すスコアとペアにしたリストである。しかし、今回はデコーダーを複数回実行することで、最も可能性の高い単一の結果ではなくより多様な (間違っている可能性もある) 結果を出力する。これにより、疑似的な n-best リストを得ている。今回は $n = 10$ でリストを出力し、そこから認識結果のテキストが重複していた場合はリストから削除する。スコアは同一のテキストで最も結果が良いものを用いる。この作業の結果 $n = 10$ の約半数のリストを得ることができる。本研究では以上の手順で得られたリストを n-best リストとして使用する。音声認識する際の言語は日本語に指定した。

4.4 大規模言語モデル (LLM)

大規模言語モデルは合計 4 つの LLM を使用する。まず、Ziyang Ma らの研究で音声認識タスクに優れているとされていた VICUNA-13b[5] と Ada Defne Tur らの研究で最も良い性能を記録した GPT-4[7] の最新モデルである GPT-4o[8] を使用する。

また、これらのモデルは日本語に対応しているが、日本語に特化しているものではないので、比較用に日本語に特化した LLM である Llama-3.1-ELYZA-JP-70B[9] を使用する。加えて、GPT-4o とのパラメーター数の比較のために GPT-3.5-turbo も用いる。各モデルのパラメータ数は以下の表 3 の通りである。

表 3　LLM のパラメーター数

	パラメーター数
VICUNA-13b	13B
Llama-3.1-ELYZA-JP-70B	70B
GPT-3.5-turbo	1.75T(GPT-4 推定) 以下 (推定)
GPT-4o	1.75T(GPT-4 推定) 以上 (推定)

4.5 プロンプトチューニング

プロンプトは自然言語の指示と n-best リスト、番組情報で構成される。番組情報は試合が始まる前に知りうることができる情報をプロンプトとして与える。今回は競技名、出場予定した選手の情報を与えた。追加で与える自然言語の情報は番組の内容と字幕情報に基づいてニュースサイトを参考に番組ごとにラベルを手動で付与した。プロンプトは Pranay Dighe ら [10] の研究で使われていた音声認識結果のエラー訂正のための英語のプロンプトをもとに手動で設計を行う。以下の図 3 のように n-best の概念の説明し、n-best リストを提示する。その後追加の情報を使用する意図を説明し、競技名と選手名を与えた。最後に出力形式を指定している。ここでは zero shot 法でプロンプトチューニングを行っている。

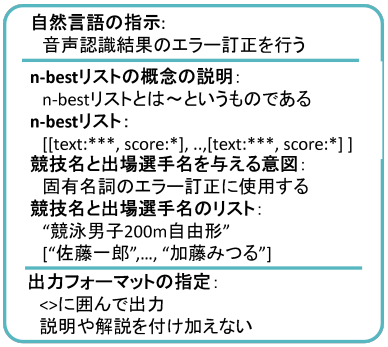

図 3　プロンプトの概略

5 評価

5.1 データセット

まず、ワンセグの TV 録画でスポーツ実況の音声、映像、字幕を収集する。今回用いるスポーツ実況音声は、2024 年夏に行われているパリオリンピック、パラリンピックの実況中継のものである。ワンセグで字幕を含むテレビ映像を録画した。以下の表 4 のようなデータを取得している。

表 4　収集した映像データ

番組数	競技数	時間	発話数
72	21	約 107 時間	約 90000 発話

その後、音声認識を行うために、BonTsDemux というソフトを用いて録画した映像から音声、字幕を切り離し、別のファイルへ保存する。そして、音声ファイルを一度 whisper において音声認識し、その際に付与されるタイムスタンプをもとに音声を分割する。

5.1.1 評価セット

上記で得られたデータの中から評価セットを作成した。まず上記のデータの中から 400 発話をランダムで抽出する。その中から固有名詞を含む発話をすべて手動で抽出する。今回は、シチュエーションに合った固有名詞を認識するために選手名とスポーツ特有の語 (技名、反則名、ポジション名など) を固有名詞とした。抽出された固有名詞を含む発話の中からランダムに 100 発話抽出した。

次にその音声に正解の字幕情報を付与する。生放送番組で字幕が付与されている際には 10 秒以上の遅れを伴うことも多い [11]。また、第 1 章にも記述したように字幕には誤りがある部分もある。そのため、評価セットに対して音声に対応した字幕を手動で付与する。まず、評価セットの発話を音声認識システムを用いてテキスト化する。その後、その認識結果の尤もらしい部分と一致する字幕を検索していく。字幕が脱落していたり、誤っていたりした場合は音声をもとに字幕を補完した。音声から発話内容がわからなかった場合はその発話を評価セットから削除し、ランダムに選んだ別の音声を評価セットに加えた。

以上の手順で作成したものを評価セットとする。

5.2 評価手法

評価指標として 2 つの指標を用いる。一般的な音声認識性能を測る指標として、character error rate(CER) と固有名詞の認識結果を測る指標として文章中の固有名詞部分のみの CER(PCER) を用いる。

5.2.1 CER

CER とは、元の文と認識結果の文が文字単位でどれほど一致しているかを表している指標である。文字の置換誤りだけでなく、脱落誤り、挿入誤りなどをカウントしている。分母を正解の文の文字数として、認識結果中の置換誤り、脱落誤り、挿入誤りしていた文字数を分子として算出する。この指標は、認識

結果が元の文章とどれだけあっていないかを示す指標であり、低ければ低いほど、正しく認識ができていることになる。この際に、句読点、人名の間に用いられる・や=、不要な空白は削除する。全角数字は半角数字に変換する。これらの文字は入れるべき位置が明確ではなく、認識結果には影響しないが、CER を下げる原因となるので削除する。英語など他言語の音声認識の際には CER でなく word error rate(WER) が一般的に用いられているが、日本語は英語などと異なり単語の単位が曖昧であるため、WER での評価が難しい。

$$CER = \frac{\text{置換誤り、脱落誤り、挿入誤りしていた文字数}}{\text{正解の文の文字数}}$$

5.2.2 文章中の固有名詞部分のみの CER(PCER)

固有名詞を必ず含む発話を抽出し、その発話の固有名詞部分ののみを使用した CER を計算する。この指標も低ければ低いほど正しく認識できている。

5.3 挿入誤りをカウントしない CER

今回の実験では、ファミリーネームをフルネームに修正してしまったがために、挿入誤りが増え、CER や PCER が増加してしてしまった結果が数多くみられた (例: ピオベサナがルービャナ・ピオベサナに)。しかし、ファミリーネームでもフルネームでも字幕の意味には影響を与えない。そのため、伝統的な CER だけでなく、挿入誤りをカウントしない CER を指標の一つとして加える。

5.4 実験

評価セットに対して、VICUNA-13b、Llama-3.1-ELYZA-JP-70B、GPT-3.5-turbo、GPT-4o の 4 つの LLM に対して提案手法を実施する。その結果に対して CER、PCER を算出し、評価する。

6 結果

結果は以下の表 5 の通りである。VICUNA では CER が 53.79%(挿入誤りなしでは 3.24%) 増加、PCER は 18.12%(挿入誤りなしでは 0.34%) 増加してしまっている。また、GPT-3.5-turbo では、CER が 15.34%(挿入誤りなしでは 1.28%) 増加、PCER は 12.85% 増加 (挿入誤りなしは 7.64% 減少) してしまっている。一方、Llama-3.1-ELYZA-JP-70B では CER が 2.21%(挿入誤りなしでは 4.77%) 減少、PCE が 5.41%(挿入誤りなしでは 16.89%) 減少した。GPT-4o でも CER が 4.15%(挿入誤りなしでは 6.35%) 減少、PCER は 12.94%(挿入誤りなしでは 20.94%) 減少することができた。特に、baseline では CER と PCER の差が、10% 以上あったが、GPT-4o では最大で 0.05% ほどに縮めることができている。

表 5 実験結果 (S: 置換誤り、D: 脱落誤り、I: 挿入誤り)

	CER(%)		PCER(%)	
	S+D+I	S+D	S+D+I	S+D
baseline(whisper-large)	23.29	18.34	37.13	35.72
VICUNA-13b	77.08	21.58	50.25	36.06
Llama-3.1-ELYZA-JP-70B	21.08	13.57	26.72	18.83
GPT-3.5-turbo	38.63	19.62	49.98	28.08
GPT-4o	19.14	11.99	19.19	14.78

7 考察

Llama-3.1-ELYZA-JP-70B や GPT-4o では PCER、CER が改善したことが確認された。このことから提案手法は有効であり、日本語固有名詞タスクにおいて十分に性能を発揮していることがわかる。また、baseline では約 1% ほどしかない挿入誤りありと挿入誤りなしの PCER の差が GPT-4o や Llama-3.1-ELYZA-JP-70B では、4%、8% と幅が大きくなっていることから、フルネームでは認識できているものが一定数あることがわかる。また、GPT-4o で最も低い CER と PCER をとったため、GPT-4o が最も日本語固有名詞音声認識タスクに有効であることが示された。GPT-4o では CER と PCER に大きな差は見られず、固有名詞の認識性能を文章全体と同程度まで下げることができたことがわかる。

7.1 GPT-4o

GPT-4o では以下の表 6 ような間違いが散見された。n-best での音声認識結果が漢字変換されていたり、カタカナになっていて固有名詞に見えずらい場合に固有名詞が修正されず、認識間違いが起こっているようである。音声認識結果が漢字に変換されてしまっていることで、発話が持つ読みがその漢字に依存していまっている。そのため、この誤りを直すには正しい読みをの情報を LLM に渡すことが有効である。その手段として n-best リストの n を増やす手段も有効であるように思われる。しかし、n を増やしすぎてしまうと、表 7 のように、発話からより離れた認識結果「姉は座に行く」を n-best リストに含めることになってしまう。これもまた、認識誤りの一因となる。さらに、Whisper ではトレーニングなしに日本語音声の読みを出力することはできないため、LLM に音声情報を渡すには、音声埋め込みなどを用いることが有効であると考える。

7.2 Llama-3.1-ELYZA-JP-70B

Llama-3.1-ELYZA-JP-70B では GPT-4o に次ぐ性能を発揮している。しかし、GPT-4o と比べ、PCER の値が高くなっている。このことからも分かる通り、Llama-3.1-ELYZA-JP-70B では修正できた固有名詞が GPT-4o よりも少なく、別の固有名詞、または音声認識の出力のままに修正していまっていることが多かった。このモデルは日本語能力が高く、日本語の誤りを修正する能力は高かったものの、固有名詞を修正するというタスクをこなす能力が低かったことが考えられる。

7.3 GPT-3.5-turbo

また、GPT-4o よりもパラメータ数が少ないとされる GPT-3.5-turbo では CER も PCER も baseline より増加してしまった。以下の表 8 は GPT-3.5-turbo での認識結果の一部であるが、与えられた情報を必要以上に使い、文章の補完も行っている。あまりタスクの意図を理解できていないことが考えられる。このことから、パラメーター数によって性能が変化することがわかる。特に、大規模なパラメータ数を持つモデルが、与えられたタスクの意図をより正確に理解し、結果として誤りを減少させる傾向があると考えられる。Llama-3.1-ELYZA-JP-70B のようには多言語でなく日本語に特化したモデルでは 70B ほど比較的大きくないのパラメーター数でも固有名詞性能を向上することができることが予想される。

表 6　GPT-4o での実験結果の一部

正解文	そしてフミレウスキ兄弟ポーランドのクシュツトフ
認識結果	そしてフミレウスキ兄弟ポーランドの北野不出と
正解文	最後ノーハンドイエス輪夢拍手
認識結果	最後ノーハンドイエスリム拍手

表 7　n-best リスト

	n = 10	n = 16
1	寝技に行くランディアも足技崩して	寝技に行くランディアも足技崩して
2	根技に行くランディアも足技崩して	根技に行くランディアも足技崩して
3	寝技に行くランディアも足技崩した	寝技に行くランディアも足技崩した
4	根技に行くランディアも足技崩した	根技に行くランディアも足技崩した
5	-	ランディアも足技崩して
6	-	ランディアも足技崩した
7	-	ランディアも足技を崩して
8	-	姉は座に行く

表 8　GPT-3.5-turbo での実験結果の一部

正解文	結果韓国キムソヨンコンヒヨンストレート勝ち
認識結果	結果パリオリンピックバドミントン自転車で キムソヨンがコンヒヨンをストレートで破りました
正解文	ローテーション演技は終わっていません ブラジルのスター登場レベッカアンドラーデ
認識結果	ローテーション演技は終わっていませんパリオリンピック体操女子団体決勝に 出場している選手レベッカアンドラーデです

表 9　VICUNA-13b での実験結果の一部

正解文	オーストラリアのジェマイマモンタグ オセアニア記録を持っている選手 26 歳です
認識結果	オーストラリアのジェマイマモンタオセアニア記録を持っている 選手 26 歳です, パリオリンピック競歩女子の優勝者です
正解文	そして比江島準備しまして
認識結果	そして比江島も準備しましてパリオリンピックバスケットボール男子予 「日本 vs フランス」で活躍した富樫勇樹選手が今大会でも高いパフォーマンスを 見せることが期待されています

7.4　VICUNA

VICUNA-13b では、CER も PCER も大幅に増加してしまった。以下の表 9 は VICUNA で見られた誤りの一部である。これらの誤りには、固有名詞の修正が適切に行われていないという問題が含まれるだけでなく、元の文章が意図しない形で大幅に補完されてしまっているという問題も見られる。これは、GPT-3.5-turbo と同様に、モデルが与えられたタスクを正確に理解できていないことが挙げられる。しかし、VICUNA-13b では、GPT-3.5-turbo と比較してもタスク理解度が低い例がみられる。具体的には、表 9 のような誤りのほかにも文章全体を競技名に置き換えてしまうような誤りや、与えられた 30 名ほどの選手名をすべて文章中に盛り込もうとする誤りがみられた。また、1 つ目の例のジェマイマ・モンタグという選手は実際にパリオリンピックで金メダルを獲得している。この情報はプロンプトでは与えておらず、LLM が付け足した情報である。このことから、VICUNA-13b には、自身が知っている情報を過剰に提供しようとする傾向があると考えられる。このような傾向がタスク遂行を妨げており、誤り率を増加させている可能性がある。

8　プロンプト (競技名、選手名) の効果

本研究で付け加えたプロンプトの競技名、選手名の効果を確かめるために、n-best リストのみのプロンプトと n-best リスト、競技名、選手名を全て含むプロンプトにおいて、GPT-4o で比較実験を行った。選手名、競技名を与えないことで、LLM は文脈から場面を判断し、既知の単語の中から認識結果を修正することが求められる。結果は以下の表 10 の通りである。プロンプトが n-best のみの場合は、baseline よりは減少させる

表 10　プロンプト内容における比較実験結果 (S: 置換誤り、D: 脱落誤り、I: 挿入誤り)

	CER(%)		PCER(%)	
	S+D+I	S+D	S+D+I	S+D
baseline(Whisper-large)	23.29	18.34	37.13	35.72
GPT-4o(n-best)	22.14	16.58	29.64	27.03
GPT-4o(n-best+ 競技名 + 選手名)	19.14	11.99	19.19	14.78

ことができているが、競技名、選手名を含むプロンプトよりは CER、PCER が共に増加している。CER の増加は 5% ほどなのに対し、PCER は 10% ほどになっている。このことから、競技名、選手名等のプロンプトは固有名詞の認識性能向上に大きく寄与していることが分かる。競技名、選手名をプロンプトとして与えることで LLM が未知の固有名詞をカバーできることが示された。

9　おわりに

本稿では、ファインチューニングを行わない大規模言語モデルを用いた日本語固有名詞音声認識について述べた。本研究では、4 種類の大規模言語モデルを用い、GPT-4o と Llama-3.1-ELYZA-JP-70B を用いた提案手法の有効性を検証した。CER ではおよそ 2〜6%、PCER はおよそ 5〜20% の改善が見られた。今後の課題として、音声埋め込みの情報を用いることで、精度改善が見込まれると予想される。

また、ゲーム実況や将棋実況などスポーツ実況以外の分野のより多くのシチュエーションに提案手法を適応していくことが課題になる。特に、ゲーム実況では、武器の名前が省略される場面も多く、より多様な固有名詞が登場する。プロンプトとして与える事前情報について検討する必要がある。

以下の表 11 は、スプラトゥーン 3 のプレイ音声 1 発話について、提案手法を適用した結果である。プロンプトとして略称を含むブキの名前、スキルの名前を与えた。その結果 3 つある固有名詞のうち 2 つが正しく修正されている。このことから、提案手法はスポーツ実況以外の場面でも字幕修正を行えることが示唆される。

表 11　LLM による音声認識結果の修正

正解文	ラピ以外ステジャン。左高、スクスロ降りた
認識結果	ラピーガイスじゃん左高スプスローリータ
GPT-4o 修正後	ラピガイスじゃん左高スクスロ降りた

参考文献

[1] 小森智康. 生放送番組における自動字幕制作の最新動向, 2020. https://www.nhk.or.jp/strl/publica/rd/182/3.html [アクセス日： (2025/01/29)].

[2] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

[3] and . Exploring the feasibility of fine-tuning large-scale speech recognition models for domain-specific applications: A case study on whisper model and ksponspeech dataset. , Vol. 15, No. 3, pp. 83–88, 2023.

[4] Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*, 2024.

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[6] Ada Defne Tur, Adel Moumen, and Mirco Ravanelli. Progres: Prompted generative rescoring on asr n-best, 2024.

[7] OpenAI, Josh Achiam, et al. Gpt-4 technical report, 2024.

[8] OpenAI, : Aaron Hurst, et al. Gpt-4o system card, 2024.

[9] 株式会社 ELYZA. Elyza、llama 3.1 ベースの日本語モデルを開発, 2024. https://prtimes.jp/main/html/rd/p/000000052.000047565.html [アクセス日： (2025/01/29)].

[10] Pranay Dighe, Yi Su, Shangshang Zheng, Yunshu Liu, Vineet Garg, Xiaochuan Niu, and Ahmed Tewfik. Leveraging large language models for exploiting asr uncertainty. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12231–12235. IEEE, 2024.

[11] 安藤慎太郎, 藤原弘将ほか. テレビ録画とその字幕を利用した大規模日本語音声コーパスの構築. 研究報告音声言語情報処理 (SLP), Vol. 2020, No. 8, pp. 1–7, 2020.