テレビ音声コーパスに対するテキスト補完と読み付与

手塚 絢子　滝口 雅人　松崎 拓也
東京理科大学 理学部第一部 応用数学科
{1422072,1422065}@ed.tus.ac.jp　matuzaki@rs.tus.ac.jp

概要

本研究では，テレビ音声に付随する不完全な字幕に対して，音声情報を用いたテキストの補完と読み付与を同時に行う手法を提案する．具体的には，カタカナ列を直接出力するよう追加学習した Whisper と，形態素解析結果のラティスを組み合わせ，漢字かな混じりの音声認識結果，カタカナによる音声認識結果，字幕の三者をアライメントすることで，読み付きテキストを生成する．LaboroTVSpeech を用いた実験では，人手で作成した正解カタカナテキストとの文字単位編集距離により評価を行い，字幕と音声の乖離が大きい条件において，提案手法が Whisper による漢字かな混じりの書き起こしや字幕に読み推定を適用した場合と比較して読み付与および補完の精度を向上させることを確認した．

1　はじめに

テレビ放送音声は，ニュース，バラエティ，ドラマなど多様な話者・話題・音響条件を含むため，音声処理研究において極めて有用なデータ源である．一方で，テレビ放送に付随する字幕は，視聴者にとっての可読性を優先して作成されるため，実際の発話内容を忠実に反映していない場合も多い．具体的には，助詞や言いよどみの省略，言い換え，さらには発話そのものが字幕に現れない箇所も存在する．従って，このような字幕を音声と対にしただけのデータを，音声合成や読み推定モデルの学習に直接用いることは困難である．安藤と藤原 [1] は，字幕が必ずしも正確ではないという問題を抱えつつも，準教師ありデコーディングを用いたアライメントにより，大規模な日本語音声認識コーパスが構築できることを示した．しかし，この研究は主に音声認識モデルの学習を目的としており，字幕テキストの欠落部分を補完したり，各語に対して正確な読みを付与したりすることまでは対象としていない．

本研究では，テレビ音声に付随する不完全な字幕に対して，音声情報を手がかりとして字幕テキストの補完を行うとともに，各語の読みをカタカナ表記で付与することを目的とする．具体的には，カタカナ列を直接出力するよう追加学習した Whisper [2] と，形態素解析結果のラティスを組み合わせ，字幕・カタカナによる音声認識結果・漢字かな混じりの音声認識結果の 3 つの間のアライメントを用いて補完・読み付与を行う手法を提案する．実験では，LaboroTVSpeech コーパス [1] を用いて評価用データを作成し，人手で作成したカタカナ正解テキストとの編集距離により性能評価を行った．その結果，字幕のみや Whisper による漢字かな混じりの書き起こし結果単独と比較して，本手法により読み付与およびテキスト補完の精度が向上することを確認した．

2　関連研究

この節では音声認識を用いたテキストへの読み付与に関する先行研究について簡単にまとめる．

Ohnaka ら [3] は，事前学習済み BERT を用いて書き起こしから抽出したテキスト特徴量と，音声基盤モデルを用いて抽出した音声特徴量を併用した上で，推論時に書記素と矛盾する音素ラベル仮説を削除することで読みを得る手法を提案し，音声特徴量のみを用いた時よりも読み付与精度が向上することを示した．

佐藤ら [4] は，青空文庫のテキストと，「サピエ」に収録された音声デイジーを組み合わせ，Whisper から複数の認識候補を取得することで，大規模な振り仮名注釈付き音声コーパスを作成した．

これらの先行研究に対し，本研究では，既存の字幕が不完全であるテレビ音声を対象とし，音声に基づいて字幕テキストの補完と読み付与を同時に行う．佐藤らの研究が主として文学作品を対象とし，朗読音声を用いているのに対し，本研究はニュースを始め様々なテレビ番組における発話に含まれる口語表現やフィラーを含む音声に対して，精度の高い読み付与を実現することを目的としている．

3　手法

この節では，まず音素ないしカタカナ列として音声認識結果を出力するための Whisper の追加訓練方法について説明し，次に，音声に基づく書き起こしテキストの補完および読み付与の方法を述べる．

3.1　音素/カタカナ出力 Whisper の訓練

Whisper [2] は，約 68 万時間に及ぶ多言語音声データで訓練された音声認識システムである．Whisper はエンコーダー・デコーダー型の Transformer で実装されており，音響特徴量をエンコーダーへの入力とし，各言語の正書法による書き起こしがデコーダーの出力である．従って，日本語の音声に対する出力は漢字かな混じりテキストであるため，発話における漢字の読みは分からない．そこで，音素あるいはカタカナで出力するように Whisper をファインチューニングした．

具体的には，音声とそれに対応する読みを音素あるいはカタカナ 1 文字ごとの分かち書きにしたものを訓練データとして，フルファインチューニングを行った．その際，デコーダーの入出力埋め込みを日本語の音素あるいはカタカナのみのものとし，高速化を図った．各トークンに対応する元々ある埋め込みはそのまま用いて，無いものはランダムに初期化した．

3.2　音声に基づく字幕の補完と読み付与

テレビ番組音声に付随する字幕テキストに対して，音声情報を用いた補完および読み付与を行い，最終的に音声内容と整合性の高いカタカナ表記テキストを構築する手法を説明する．本手法では，OpenAI が公開している Whisper（以下，漢かな Whisper）による音声認識結果，字幕テキスト，および 3.1 節で述べたカタカナ出力を行う Whisper（以下，カナ Whisper）の出力を統合し，複数の読み候補を考慮しながらテキストを構築する．

まず，入力音声に対して漢かな Whisper を用いて漢字かな混じりの書き起こしテキストを得る．同時に，カナ Whisper を用いて，同一音声からカタカナ列を出力する．

次に，漢かな Whisper による書き起こし結果およびテレビ字幕テキストのそれぞれに対して MeCab [5] を適用し，形態素解析に基づく読み候補のラティスを構築する。本研究では，MeCab の N-best 出力を用い，最大 10 通りの読み候補列を生成した結果をラティスにまとめた．単一の最尤読みのみを用いる場合，表記曖昧性や音声との不一致に起因する誤りが固定されてしまう．そのためラティスを用いることで複数の読み候補を保持し，音声情報に基づいた柔軟な選択を可能にした．ラティス中の各読み候補（カタカナ列）に対して，Kaji の動的計画アルゴリズム [6] を用い，カナ Whisper の出力との編集距離が最も小さい候補を選択する．この処理を，漢かな Whisper による書き起こしおよび字幕の双方に対して行うことで，それぞれから音声と整合性の高いカタカナ列を得る．

その後，以下の三つの系列間でアライメントを行う：(1) 漢かな Whisper による書き起こし由来のカタカナ列，(2) 字幕由来のカタカナ列，(3) カナ Whisper による音声認識結果である．アライメントはカタカナ文字列に対して編集距離に基づいて行った．

最後に，このアライメント結果をもとにテキストの補完および統合を行う．字幕に存在しないが音声には現れる要素や，逆に字幕には含まれているが音声では発話されていない要素を区別し，音声とテキストの不一致に対処する統合処理を行った．具体的には，音声との整合性を最優先とする方針に基づき，カナ Whisper の出力が存在する場合はそれを優先的に採用した．カナ Whisper の出力に欠落がある場合は，漢かな Whisper と字幕で一致する文字のみを採用し，不確実な挿入や重複を抑制した．

特に，カナ Whisper の出力には現れる一方，漢かな Whisper による書き起こしおよび字幕テキストのいずれにも対応する文字が存在しない部分は，話し言葉特有のフィラーである可能性が高い．そこで本研究では，そのような連続部分を一時的にバッファに保持し，事前に構築したフィラー辞書と照合することで，フィラーであると判断された場合のみ最終結果に反映する処理を行った．フィラー辞書は，複数の辞書資源から抽出したフィラー語を基に構築し，長音記号や促音を除去した正規化形による照合も許容することで，表記揺れに対して頑健な判定を可能とした．この処理により，音声には存在するが意味内容を持たない要素を無制限に挿入してしまうことを防ぎつつ，「エー」「アノー」などの実際の発話に基づくフィラー表現については音声忠実性を保った形で出力することができる．以上の処理を通じて，音声内容を忠実に反映したカタカナ表記の完成テキストを構築する．

4　実験設定

4.1　Whisper の追加訓練データと設定

ベースモデルとして huggingface の whisper-medium を使用した．追加訓練のためのデータとして，日本語話し言葉コーパス（CSJ）を用いた．正解テキストデータとして，コーパスの転記テキストの発音形をフィラーやいい間違いなども含めて用いた．音素列を正解とする場合は転記テキストの発音形を Open JTalk [1)] で音素列に変換した．コーパス全体のうち約 441 時間分を用いて，各音声は無音区間で約 10 秒ごとに区切った．これを訓練・検証・テストデータに 8:1:1 に分割して用いた．訓練はミニバッチサイズを 16 として 20,000 ステップ行った．

4.2　字幕の補完・読み付与実験の設定

字幕テキスト補完および読み付与の評価には，テレビ番組音声と字幕テキストを対にした LaboroTVSpeech コーパス（バージョン 1）[1] を用いた．このコーパスは実際のテレビ放送音声を対象としており，字幕テキストには省略や表記揺れが含まれるため，本研究の目的に適したデータセットである．まず，コーパスに含まれる音声ファイルのうち約 1 万に対して漢かな Whisper を用いて音声認識結果を得た．次に， 3.2 項で述べた方法で漢かな Whisper の出力および字幕をカタカナ変換した結果の間の編集距離に基づいて評価対象を抽出した．具体的には，編集距離が 1，2，4，8 となる音声ファイルをそれぞれ 100 ファイルずつ選択し，合計 400 ファイルを評価データとした。これにより，字幕と音声の乖離度が異なる複数の条件において，本手法の有効性を検証できるようにした．評価用の正解テキストは，各音声ファイルを聴取した上で，人手によりカタカナ表記で作成した．これを音声内容を最も忠実に反映した参照テキスト（正解テキスト）とする．評価指標としては編集距離を用いた．編集距離はカタカナ文字列に対して文字単位で算出し，音声ファイルごとに計算した値を用いる．ベースラインとして漢かな Whisper の出力に対し Open JTalk を用いて読み推定を行った結果と正解テキストとの編集距離を算出し，これと提案手法による最終出力結果と正解テキストとの編集距離を比較することで，本手法による改善効果を評価した．

1）　https://open-jtalk.sp.nitech.ac.jp/

表 1　文字誤り率 (CER) の比較

出力形式	fine-tuning	CER (%)
漢字かな混じり	なし	24.93
音素	なし	14.20
音素	あり (音素)	4.51
カタカナ	なし	17.50
カタカナ	あり (カタカナ)	2.70

5　実験結果

5.1　音素/カタカナ Whisper の認識精度

表 1 に，音素あるいはカタカナを出力として追加訓練した Whisper の CSJ コーパスに対する認識精度を示す．評価指標としては文字誤り率（CER）を用いた．比較対象として，追加訓練前の漢かな Whisper の出力，及びそれに対して Open JTalk を用いて読み推定を行い，音素・カタカナに変換した場合の精度を示す．最も精度が高かったのはカタカナでファインチューニングしたもので，漢かな Whisper の出力に対して Open JTalk で読み推定を行った場合に比べて約 15 ポイント CER が改善した．大きな要因として考えられるのは，漢かな Whisper は「あのー」や「えっとー」などのフィラーを出力しない傾向にあることである．この評価ではフィラーを含め実際の読みを正解としたため，読み推定誤りによるエラーにこの効果が重畳し，大きな差となったと考えられる．また，漢かな Whisper の CER が他のモデルに比べて高くなっているが，その主な原因は同音異字の出力が多いことだと考えられる．

5.2　字幕の補完・読み付与の結果

図 1 に，漢かな Whisper の出力に Open JTalk による読み推定を行なったカタカナ書き起こし（Baseline）および提案手法による最終出力（Proposed）について，正解テキストとの編集距離分布を示す．4.2 項で説明した distance=1 の条件では，Baseline が正解テキストに近い分布を示しており，提案手法による改善は確認されなかった．これは，字幕と音声の乖離が小さいケースでは，処理の介入によって編集距離が増加するためと考えられる．一方，distance=2 および distance=4 の条件では，提案手法により編集距離が 0 および 1 のものの割合が増加し，Baseline と比較して正解テキストに近い出力が得られている．この傾向は distance=8 において顕著であり，Baseline では編集距離が大きく分散しているのに対し，提案手法では編集距離の小さい範囲に分布が集中する傾

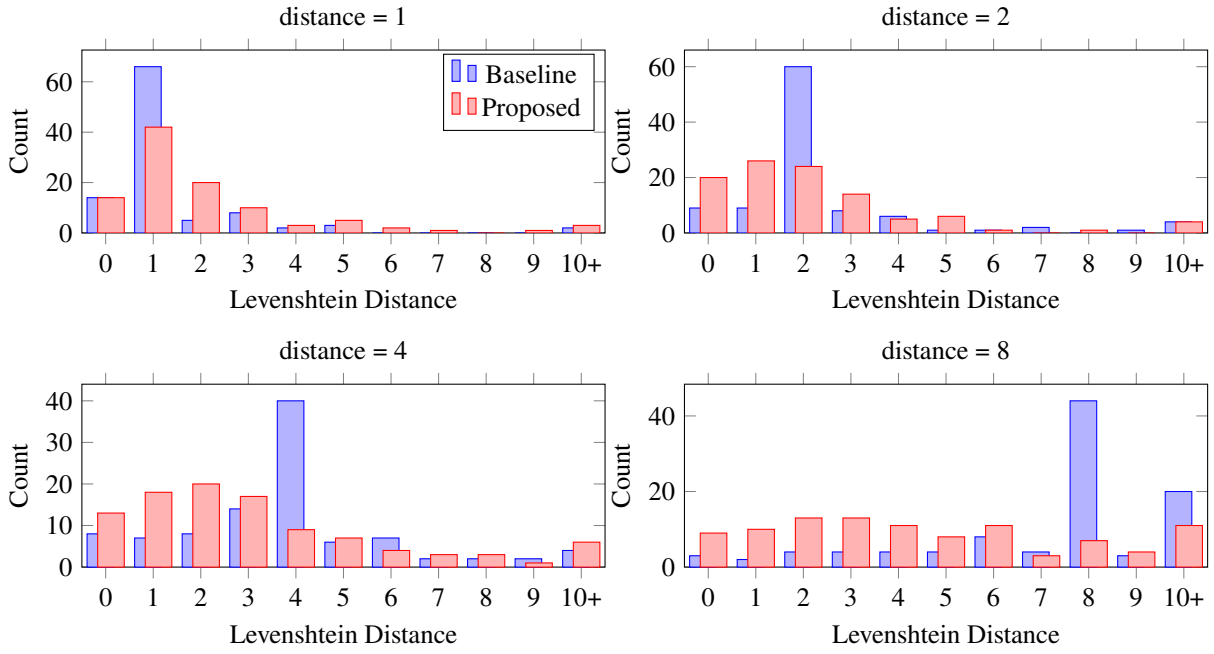

図 1　正解カナ列と読み推定・テキスト補完結果との編集距離の分布

向が見られた．以上の結果から，本手法は Whisper の書き起こしと字幕との乖離が大きい条件において有効であり，テレビ字幕に対する補完および読み付与の精度向上に寄与することが示された．

5.3　字幕の補完・読み付与の誤り例

本項では，提案手法によって編集距離が低減しなかった，あるいは増加した事例について，誤り例に基づいて分析を行う．特に，distance=1 の条件において性能向上が見られなかった要因を明らかにする．表 2 に，distance=1 の条件で確認された誤り例（Baseline 出力は正しく，提案手法では誤ったもの）を示す．これらの多くは，表記揺れや音韻的に近い表現の差異，助詞や機能語レベルの置換，およびフィラーや発話境界に起因するものである．まず，表記揺れや音韻的差異に関する誤りとして，「テイウカ」と「テユーカ」，「オ」と「ヲ」，長音の有無といった違いが確認された．これらは音声的にはほぼ同一であり，読みとしての自然さや意味理解に影響を与えないが，文字単位の編集距離では誤りとして扱われる．そして，フィラーや発話の開始・終了部分に起因する誤りも確認された．文頭・文末の短い発話や間投詞（「えー」「ナー」など）は音声的に不明瞭となりやすく，字幕や正解テキストとの対応関係が不安定となる場合がある．このような誤りは，テレビ音声データに特有の課題であると考えられる．一方で，これらの誤りは表記揺れや音声境界に起因するものであり，実運用上の読み付与品質に与える影響は限定的である．この結果は，字幕と音声の乖離が大きい条件において提案手法が有効であるという 5.2 節の結果と整合的であり，本手法の適用条件を明確に示すものとなっている．

表 2　字幕補完・読み付与における誤り例（distance=1）

誤りタイプ	Whisper 出力（Baseline）	提案手法出力（Proposed）
表記揺れ	テイウカ	テユーカ
助詞の置換	イキヲハイテ	イキオハイテ
フィラーの差異	ナー	エー
語頭の欠落	ワアルケドネ	アルケドネ
外来語表記	ウィズコロナ	ウイズコロナ
長音の有無	ダローホー	ダローホ

6　おわりに

本研究では，テレビ音声データに付随する不完全な字幕に対し，音声認識結果を用いた字幕テキストの補完および読み付与の手法を提案した．Whisper を音素・カタカナ出力に拡張し，ラティスを用いて音声に基づく読み候補を生成・選択することで，字幕と音声の乖離が大きい条件において，正解テキストとの編集距離を低減できることを示した．一方で，Whisper の書き起こし精度がすでに高い条件では，表記揺れや音声境界に起因する不要な補完が生じる場合があることも明らかになった．今後の課題として，補完処理を適用すべき条件の判定や，音韻類似度を考慮した評価指標の導入が挙げられる．

参考文献

[1] 安藤慎太郎, 藤原弘将. 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. 情報処理学会研究報告, 2020.

[2] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In **Proceedings of the 40th International Conference on Machine Learning**, ICML'23. JMLR.org, 2023.

[3] Hien Ohnaka, Yuma Shirahata, Byeongseon Park, and Ryuichi Yamamoto. Grapheme-coherent phonemic and prosodic annotation of speech by implicit and explicit grapheme conditioning. In **Interspeech 2025**. ISCA, 2025.

[4] 佐藤文一, 吉永直樹, 豊田正史, 喜連川優. 音声認識を用いた青空文庫振り仮名注釈付き音声コーパスの構築の試み. 言語処理学会第 30 回年次大会講演論文集, 2024.

[5] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[6] Nobuhiro Kaji. Lattice path edit distance: A Romanization-aware edit distance for extracting misspelling-correction pairs from Japanese search query logs. In Mingxuan Wang and Imed Zitouni, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track**, pp. 233–242, Singapore, December 2023. Association for Computational Linguistics.