4ZK-02

台本なしテレビ番組の音声データを用いた自動台本生成

藤原 有希† 横山 泰知† 大政 裕哉† 近藤 里咲‡ 梶原 智之‡ 二宮 崇‡ 後藤 功雄‡
愛媛大学工学部工学科† 愛媛大学大学院理工学研究科‡

1 はじめに

テレビ番組における台本は、番組の構成や撮影の順序を記述する目的のほか、過去の番組情報を参照する目的でも使用される。出演者のアドリブによる発話が中心となるテレビ番組では、台本がないケースが多く存在し、編集作業に対する負担の増大や番組データの検索性・再利用性の低さが問題視されている。

近年では、番組制作を支援する目的[1]のほかに、番組データからの効果的なテキスト分析[2]やテレビ番組を話題とした発話文生成[3]などに番組の情報が用いられている。こうしたメディア分野における技術発展のためには、台本を代表とした言語資源が必要とされている。しかし、番組映像からの台本作成には、出演者の発話内容の抽出や話者の特定、場面転換の言語化など高コストなアノテーションを要するため、台本自動生成手法の確立が望まれている。

本研究では、台本自動生成を実現するために、台本生成において主要な要素となる「発話内容の文字起こし」に取り組む。具体的には、複数の音声認識モデルを用いて音声データをテキストに変換し、モデルサイズが音声認識性能に与える影響を調査する。実際のテレビ番組の音声データを用いた評価実験の結果、最もサイズの大きいモデルが最良の結果を示した。

2 音声認識モデルを用いた文字起こし手法

本研究では、OpenAI 社が開発した音声認識モデルである Whisper [4]を用いて、番組の音声データをテキストに変換する。Whisper は Transformer をベースとした音声認識モデルである。多言語かつマルチタスクの大規模な音声データセットで学習されており、日本語においても高い認識精度を示している[4]。Whisper では、表 1 に示すように、複数のモデルサイズが公開されている [1]。

表 1 利用可能なモデルサイズの例

モデルサイズ	パラメータ数
small	224M
medium	769M
large	1,150M

3 実験設定

3.1 データセット

テレビ番組の映像データとして、南海放送で実際に放送された 5 本の番組データを用いた。各番組データは 1 本あたり約 1 時間の映像であり、番組音声には発話以外にも BGM や効果音が含まれる。MP4 形式の動画データから ffmpeg[2] を用いて音声データに変換した。これらの音声を、日本語母語話者の大学生 3 名が人手で書き起こし、最終的に 69,104 文字分の正解データを作成した。

3.2 音声認識モデル

本研究では、音声認識モデルとして、Whisper の small と medium および large-v2 を採用した。なお、デコード時には GreedyDecoder[3] を用い、logprob_threshold を −1.0、no_speech_threshold を 0.6 として、ノイズテキストの生成を抑制した。

3.3 評価指標

モデルの評価には文字誤り率（CER：Character Error Rate）を用いた。CER は、式(1)に示すように正解文の文字数に対する予測文の文字単位の編集操作（挿入・置換・削除）の割合で計算される。この CER の値が低いほど正解文との差が少ないため、モデルの認識精度は高いといえる。

$$\mathrm{CER} = \frac{\text{挿入数}+\text{置換数}+\text{削除数}}{\text{正解文の文字数}} \quad (1)$$

サイズの異なる音声認識モデルを用いて 5 番組分の音声データからそれぞれ発話内容を生成し、3.1 節で作成した正解データとの間で CER を算出した。5 つの番組の CER の平均値 $\mathrm{CER}_{\mathrm{AVG}}$ をそのモデルの評価値とした。また、文字起こしに要した処理時間を測定し、元の音声データの長さとの比率を算出した。測定には NVIDIA TITAN RTX (24GB) のシングル GPU を使用した.

Automatic Script Generation from Audio Dara for Unscripted Television Programs
†Yuki Fujiwara (fujiwara@ai.cs.ehime-u.ac.jp)
†Taichi Yokoyama (yokoyama@ai.cs.ehime-u.ac.jp)
†Yuya Omasa (omasa@ai.cs.ehime-u.ac.jp)
‡Risa Kondo (kondo@ai.cs.ehime-u.ac.jp)
‡Tomoyuki Kajiwara (kajiwra@cs.ehime-u.ac.jp)
‡Takashi Ninomiya (ninomiya.takashi.mk@ehime-u.ac.jp)
‡Isao Goto (goto.isao.fn@ehime-u.ac.jp)
Ehime University

[1] https://github.com/openai/whisper
[2] https://ffmpeg.org
[3] https://github.com/openai/whisper/whisper/decoding.py

表 2 各モデルサイズの評価結果

モデルサイズ（パラメータ数 [M]）	相対処理速度	CER_{AVG}	挿入文字数	置換文字数	削除文字数
small (224)	**0.087**	0.313	3,326	13,032	5,155
medium (769)	0.151	0.226	**2,092**	9,304	4,087
large-v2 (1,150)	0.209	**0.135**	3,017	**4,044**	**3,017**

4 実験結果

表 2 の左側に各モデルサイズの元の音声データに対する相対処理速度と CER_{AVG} を示す。モデルサイズが大きい large-v2 の場合に最も良い CER_{AVG} が得られた。妥当な CER の値として、会議録作成システムでは0.15以下が適正とされているため[5]、large-v2 は十分な性能であるといえる。

また、相対処理速度ではモデルサイズの小さい small が最も高速に動作し、large-v2 に比べて約 2.4 倍の処理速度を実現した。音声認識の性能と処理速度はトレードオフの関係にあるため、用途に応じたモデル選択が望まれる。

表 2 の右側に 5 つの番組における各編集操作（挿入・置換・削除）の総数を示す。置換および削除はモデルサイズが大きいほど減少する傾向にある。一方、挿入については、small から medium にかけては減少しているものの、medium と large-v2 の間では減少がみられなかった。

さらに、各モデルサイズにおける編集操作の割合に着目すると、認識誤りのうち置換操作に起因する件数が 43.7～60.7%と多くを占めることが確認できた。

5 分析

最後に、Whisper の実際の誤った出力例を示し、その原因について考察する。Whisper の出力に含まれる誤りの中から、置換操作の例を図 1 に示す。

図 1 の①に示す表記の誤りによる置換は、音声は一致しているが漢字や片仮名の表記が異なるものであり、②に示す音声認識の誤りによる置換は、モデルの予測と音声が一致しておらず誤認識した結果である。前者については、音声認識に加え文脈に応じた適切な文字変換が求められるという、日本語での音声認識タスク特有の性質に起因しており、後者は音声認識そのものの性能に起因している。

したがって、Whisper の出力に含まれる置換誤りは、モデルの音声認識精度そのものを改善するほか、後編集として修正を施すことで効率的に認識精度を向上させられることが示唆される。

① 同音異義語の誤り
予測：特技は初動
正解：特技は書道

予測：おいしい一軸の甘さ
正解：おいしいイチジクの甘さ

② 音声認識の誤り
予測：そういうからずっと
正解：創業からずっと

予測：はがきは明日決心有効です
正解：はがきは明日消印有効です

図 1 置換による誤りの例

6 おわりに

本研究では、Whisper を用いてテレビ番組の音声データから発話内容を自動で文字起こしし、モデルサイズごとの性能評価および結果の分析を行った。評価実験の結果、モデルサイズの大きいモデルが最高性能を達成した。より正確な発話内容の文字起こしを実現するために、同音異義語や音声認識の誤りを、LLM（大規模言語モデル）を用いて自動検出・訂正することで精度向上を目指している。今後、この手法によって音声認識精度の改善に取り組みたい。

謝辞

南海放送株式会社からデータを提供いただきました。

参考文献

[1] 三島剛，萩原愛子，伊藤均，小森智康，佐藤庄衛．番組制作を支援する音声認識を用いた字起こしシステムの開発．映像情報メディア学会誌．Vol. 75，No. 1，p. 118−124，2021.
[2] 織田一輝, 佐々木稔. テレビ番組データを対象とした人名抽出と番組ジャンル推定. FIT2020 (第 19 回情報科学技術フォーラム), 2020.
[3] 金子豊, 星裕太, 村﨑康博, 上原道宏. テレビ番組に含まれるキーワードに基づく発話文生成手法. 情報処理学会研究報告, 2018.
[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. International Conference on Machine Learning, PMLR, 2023, pp. 28492– 28518.
[5] 河原達也. 議会の会議録作成のための音声認識－衆議院のシステムの概要. 情報処理学会研究報告 SLP－93－5, 2012.