音声認識 API「Whisper」を用いた歴史的音源に対する音声認識に関する研究

Research on Speech Recognition for Historical Sound Sources using Whisper as Speech Recognition API

山崎　右京[1)]
指導教員　三輪　賢一郎

1）サレジオ工業高等専門学校　機械電子工学科　情報コミュニケーション研究室

キーワード： 音声認識, 文字起こし, SP レコード, Whisper

1．緒　言

SP レコード(Standard Playing Record)は、1887 年から 1950 年代後半まで用いられた。当時の歴史的価値ある音源の録音・再生を行う技術である。SP レコードには演説、講談などの「語りもの」や落語など、様々なジャンルが録音されている。これらの膨大な価値ある音源を文字として後世に遺すことが是非とも必要であると考える。しかしながら、当時用いられた原稿の存在は不明である場合が少なくない。そこで、そのための文字起こしを人が行うとなると、莫大なコスト、時間、労力がかかることは自明である。そのため、別の効率的な方法で文字起こしを省力化する必要がある。

本研究室では令和 3 年度から歴史的音源の文字起こしの研究を行っている。昨年度は、音声認識 API（Application Programming Interface）を用いた方法により、認識精度が飛躍的に向上したことを確認した[1]。しかしながら文字起こし後の文章を読み返すと、一部に読みにくさ感じる部分も散見され、作業の省力化に向けて認識精度のさらなる向上が望まれる。そこで本研究では、音声認識 API の最新形である「Whisper」を用いて、歴史的音源に対する最適な文字起こし手法について検証を行う。

2．方　法

本研究では、音声認識を行うにあたり、OpenAI 社の Whisper を採用した[2]。Whisper には tiny・base・small・medium・large の 5 つのモデルが存在する。これらのモデルは、tiny＜base＜small＜medium＜large の順に性能が向上する。また、認識に要する時間も上記の順番で長くなる。

今回、ハードウェアにはノートパソコンを用い、Whisper を使用する上でのプログラムは、Visual Studio Code と Google Colaboratory を開発環境として Python を用いて行った。本実験で使用した音源は、「国立国会図書館ディジタルコレクション「歴史的音源」」に所蔵されている「ソロモン海戦に就いて 1」（日本コロムビア、昭和 19 年頃、収録時間 3 分 4 秒）を国立国会図書館の許可のもとに使用した[3]。

実験結果の評価指標として、平仮名ベースのモーラ（拍）に着目し文字認識率の算出を行う。算出式は（1）式に示す。

文字認識率＝

$$\frac{\text{正解文文字数} - \text{誤挿入文字数} - \text{誤削除文字数} - \text{誤置換文字数}}{\text{正解文文字数}} \times 100 \quad \cdots (1)$$

3．結　果

認識結果のある一文を比較したものを表 1 に示す。また、文字認識率の結果を図 1 に示す。

結果から、Whisper を用いた場合、small 以上のモデルでは、先行研究で使用された GoogleSR、AmiVoice の認識率を上回る結果となった。また、large モデルでは文字認識率が 98%以上となり、非常に高い精度で文字認識が行えることを確認した。

表 1　認識結果の比較

種類	結果
Whisper (large)	猛撃を加え敵艦隊並びに輸送戦艦に大なる打撃を与え、もっかなお作戦続行中であること
GoogleSR	攻撃オープン関寛斎ならびに輸送戦団に大なる打撃を与え、もっかな大作戦続行中であること
AmiVoice	攻撃を加え敵艦隊並びに輸送戦団に大なる打撃を与え、目下なお 8006 校中であること
正解文	猛撃を加え敵艦隊並びに輸送戦団に大なる打撃を与え、目下なお作戦続行中であること

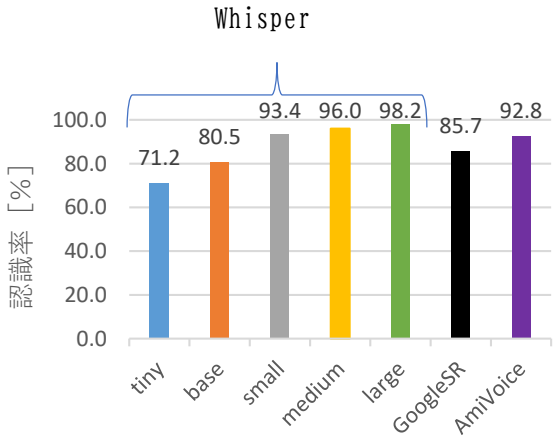

図 1　「ソロモン海戦に就いて 1」の文字認識率

4．結　言

本研究では、音声認識 API「Whisper」を用いた歴史的音源の文字起こしを行い、その可能性について検証した。検証の結果、「Whisper」による文字認識率（平仮名ベース）は 98%超にも達し、歴史的音源の文字起こしの省力化が十分期待できる結果となった。

今後は、落語など演説とは異なるジャンルの歴史的音源や、録音方式の異なるさらに古い年代の音源などを用いて検証を進める予定である。

謝　辞

本研究は、国立国会図書館のご厚意により、「国立国会図書館ディジタルコレクション歴史的音源」に所蔵の音源を用いております。

参考文献

[1] 小野﨑圭吾,三輪賢一郎,“API を用いた歴史的音源に対する音声認識に関する研究,” 大学コンソーシアム八王子　第 14 回学生発表会, A212,Dec.2022.
[2] OpenAI Whisper (https://platform.openai.com/docs/guides/speech-to-text)
[3] 国立国会図書館ディジタルコレクション「歴史的音源」 Web サイト (https://rekion.dl.ndl.go.jp/ja/)