音声認識API「Chirp」を用いた歴史的音源の文字起こしに関する研究
～一般化調和解析によるノイズ除去処理の検証～

Research on Transcription for Historical Sound Sources using Speech Recognition API “Chirp”：Verification of Noise Reduction Effects by Generalized Harmonic Analysis

藤咲 勇汰
指導教員　三輪 賢一郎

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

歴史的に貴重な音源を後世に「文字」という形で遺すべく、音声認識技術を用いた文字起こしの可能性について検討を行っている。本稿では、大正期の音源に対して、音声認識API「Chirp」に一般化調和解析によるノイズ除去を併用することで、文字起こし作業の効率化・省力化が実用的に行えるかを検討する。

キーワード：音声認識API，文字起こし,SPレコード,ノイズ除去,Chirp

１．はじめに

昨今、SiriやAlexaなどのAIを用いた音声対話システムが普及しており、その性能も日々向上している。本研究室においても、それら音声認識API（Application Programming Interface）を用いることで、人の手で行われている“文字起こし”作業の省力化･効率化を図り、歴史的価値のある大正期の音源を文字という形で後世に遺すべく検討を行っている。

昨年度の先行研究では、ChatGPTで有名なOpenAI社が開発した音声認識API“Whisper”を用いて、大正期の歴史的音源に対する文字起こしの検証を行ったが、漢字ベースの文字認識率は70%にとどまり、実用化には程遠い結果となった[1]。そこで、今回は音声認識APIにGoogle社の“Chirp”を用い、大正期の歴史的音源にノイズ除去の前処理を行った上で、文字起こしが効率的に行えるかを検証する。

２．方法

本研究での評価実験の対象とした音源は、国立国会図書館の「国立国会図書館ディジタルコレクション歴史的音源」内に所蔵されている「政治の倫理化（一）」（日本コロムビア,大正13年頃,3分40秒）で、国立国会図書館の許可のもとに使用した[2]。音声認識APIはGoogle社のChirp[3]を用い、APIの動作はGoogleのSpeech-to-Textより音声文字変換の機能を用いて行った。また、1分未満の短い音声信号に適した“Chirp”と、長文に適した“Chirp2”の２種類のモデルが存在しており、それぞれ実施することとした。ノイズ除去手法は一般化調和解析を採用し、フリーソフトの「DHA Denoiser」（バージョン 2.0,64bit 版）[4]をノートパソコンにインストールした上で実施した。

評価指標としては、(1)式に示す文字認識率を使用し、漢字ベースでの文字認識率が90%を超えた場合に、システムとして実用的とみなすこととした。

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

３．結果

漢字ベースでの文字認識率を図1、Chirpによる認識結果の例を図2に示す。図1よりノイズ除去ありの音声は共に、ノイズ除去なしの音声よりも文字認識率が高くなっており、ノイズ除去処理に一定の効果が認められることが確認できる。しかし、文字認識率の最大値が66%であり、先行論文の

Wisper よりも低く、実用化には程遠いといえる。

また図 2 より、「懸案」、「欣快」、「施行」、「政界」など、カ行・サ行から始まる単語が同音異義語に変換されてしまうといった特徴も明らかとなった。

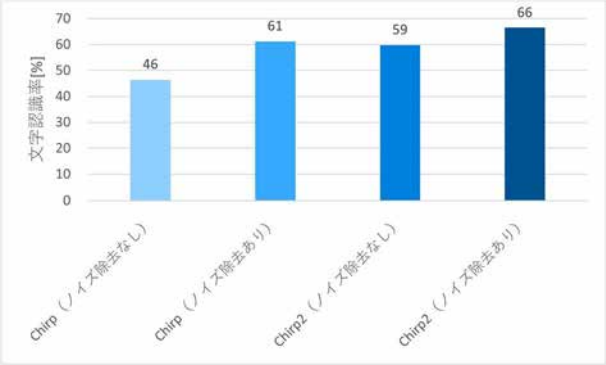

図 1 Chirp による文字認識率

今や我が正解多年の権妄であった普通選挙もようやく実現の趨向を下(たたむる)に↓
至りましたこの時にあたり新たに参政権 1（得て）憲法裁判の大（業）に↓
パ（関なので、脱字「か」は考えない）ワラントする諸君に向かって↓
予がべきの主張たる政治の人（理）化を提唱することが最も近海とするところであります↵
衆宣すでにご表示の通り我が正解近来の腐敗堕落は天人の共に一るところ↵
ことに牽制運用の収縮勢力を持って認ずる政党が我が瀬戸内閣という↵
極めて卑劣の文句を不面もなく交渉して民衆の診療を幻惑するがごときは↵
立憲（民主主義）の国民として切にも(猛は一文字の為、脱字「う」は未カウント)（省）を望む ↵
(未変換 3　誤変換 41　脱字 13　誤挿入 1 ↓
正解文 247-未変換 3-誤変換-41 脱字 13-御挿入 1=189 文字正解↓
→ 誤り率は 189/247×100=76.518%) ↵

図 2 認識結果の例

４．まとめ

今回は音声認識 API に Google 社の Chirp を用い、大正期の歴史的音源にノイズ除去の前処理を行った上で、文字起こしが効率的に行えるかを検証した。ノイズ除去を実行することで、文字認識率の向上がみられたものの、実用的と言えるレベルには届かなかった。

５．今後の予定

比較対象として、Chirp 本来の性能を調査するため、人工的な雑音を付加した合成音声による音源を用いた補足実験を実施する予定である。

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。

文献

[1] 山崎右京，“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，”サレジオ工業高等専門学校 令和 5 年度卒業論文
[2] 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/)
[3] Google Chirp (https://cloud.google.com/speech-to-text/v2/docs/chirp-model?hl=ja)
[4] GHA Denoiser (https://www.vector.co.jp/soft/winnt/art/se510574.html?srsltid=AfmBOor6maG8QGlFCGXqgNIsqQOMRvqW4nFGUfkNJCja6T21EJ3JmeMw)