音声認識 API 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 Speech Recognition API

小野﨑　圭吾
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

キーワード：SP レコード,文字起こし,音声認識, API

１． 緒言

過去に使用された音声を録音する技術として SP レコードというものがある。SP レコード(Standard Playing Record)は、1887 年から 1950 年代後半まで使用され、その時代の歴史的価値のある様々な音源の録音に使用された。SP レコードのジャンルに演説、講談などの「語りもの」があり SP レコード自体は現存している。しかし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。一部の落語などは当時の速記本が現存しているケースもあるが、必ずしも実際の録音と文言が完全一致するわけではないことが分かっている[1]。それら膨大な文化遺産を文字の形で後世に遺すためには、何らかの方法で文字起こしを実施する必要がある。

昨年度、本研究室ではオープンソースの音声認識エンジンである Julius を用いた検証を実施した[2]。しかしながら、その認識精度はごく低水準にとどまった。近年 Google home、Alexa などのスマートスピーカに使用されるクラウド音声認識エンジンが API (Application Programming Interface) として利用できるようになった。そこで本研究では音声認識エンジンに代えて音声認識 API を用い、同研究と同じ音源を使用した音声認識を実行し、歴史的音源に対する最適な文字起こしについて検証する。

２．方法

本研究では、音声認識には音声認識 API である Google Speech-Recognition 及びアドバンスト・メディア社の AmiVoice を用いた。環境設定として、Google Speech-Recognition（以下、GoogleSR と略す）ではソースエディタとして Visual Studio Code を用い、プログラム言語の Python で記述した。AmiVoice では認識結果が JSON 形式で出力され、そのままでは認識率の計算に使用できないため、テキスト形式に変換することとした。

本実験で用いた音源は、「国立国会図書館ディジタルコレクション「歴史的音源」」[3]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）であり、国立国会図書館の許可のもとに使用した。音源には SP レコード特有のノイズが含まれるが、音源のまま使用した場合と、ノイズ除去した場合との２通りについて認識を行った。ノイズ除去には NCH Software 社の音声編集ソフトウェア「WavePad」を用いた。認識結果の精度の評価指標としては、今回は仮名ベースでの認識パフォーマンスを考えることとし、モーラ（拍）に着目した文字認識率を採用した。下記に文字認識率の算出式を示す。

$$\text{文字認識率} = \frac{\text{正解文文字数} - \text{誤挿入文字数} - \text{誤削除文字数} - \text{誤置換文字数}}{\text{正解文文字数}}$$

3．結果

本実験では 2 種類の音声認識 API を用いて同じ音源に対して音声認識を実行した。誤認識数より認識率を算出し、昨年度の結果[2]と比べたものを図 1 に示す。また図 2 と図 3 に API ごとの誤認識文字数を示す。図 1 の結果より、各 API の認識率は GoogleSR が約 85%、AmiVoice が約 93%と、Julius を用いた場合よりも格段に認識率が向上したことがわかる。また、SP レコード特有のノイズが認識精度に影響していると考えられたため、音源に対してノイズ除去を実施した場合でも実験を行ったが、今回使用した API の音声認識においてはどちらも認識率の低下が確認された。この要因として、恐らく音声認識 API のシステムの中にノイズを除去する工程が含まれており、結果として 2 重のノイズ除去を行ってしまっていたのではないかと考えられる。また GoogleSR と AmiVoice の誤認識数を比べると GoogleSR のほうが文字単位の誤認識が多い傾向が見られた。これは GoogleSR の Google が保有するビックデータによる予測機能の影響ではないかと考えられる。

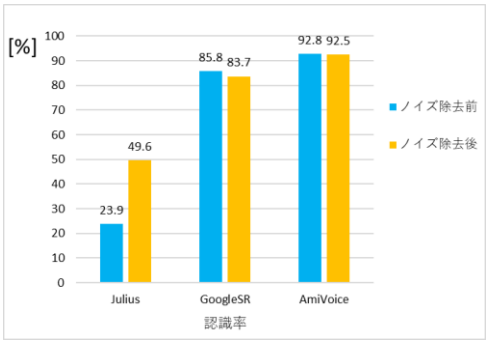

図 1　各音声認識システムの認識率

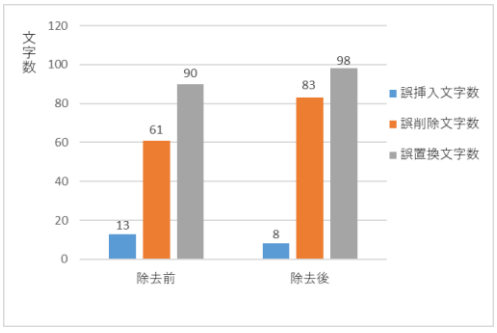

図 2　GoogleSR 誤認識文字数

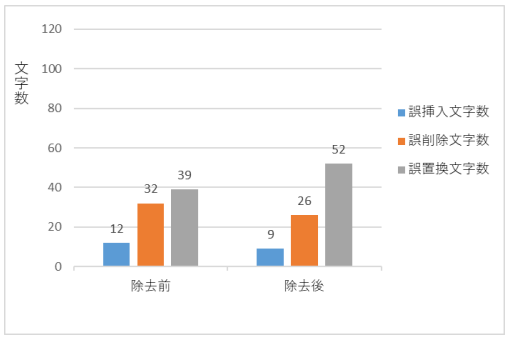

図 3　AmiVoice 誤認識文字数

4．結言

本研究では音声認識 API を用いたシステムを構築し、歴史的音源に当たる SP レコードに遺された「語りもの」の文字起こしの可能性について検討した。先行研究で検討された Julius よりも、音声認識 API である Google Speech-Recognition 及び Amivoice を用いた場合のほうが認識率を大幅に向上できることがわかった。また、ノイズ除去の効果は確認できなかった。

今後はジャンルの違う音源に対して同様の実験を行い、システムの汎用性を検証する。また今回の結果について漢字ベースでの認識率を算出し、より使用環境に即した認識率を検証する。

5．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」[3]に所蔵の音源(同図書館ウェブサイトで公開中のもの)を用いております。

6．参考文献

[1] 金澤裕之，“現代に繋がる近代初期の口語的資料における言語実態：速記本と SP レコードによる東西の落語を対象として，”国立国語研究所論集，no. 10，pp. 58–84，2016 年 1 月

[2] 小泉朝陽，三輪賢一郎，”音声認識技術を用いた歴史的音源のテキスト起こしに関する研究，”大学コンソーシアム八王子　第 13 回学生発表会，Q114，Dec. 2021.

[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/)