音声認識API「Chirp」による歴史的音源の文字起こしに関する研究
～落語音源での評価試験～

A Study on Text Generation from Historical Sound Sources using Speech Recognition API “Chirp”: Evaluation by RAKUGO Audio

西野 嘉祥
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

歴史的音源を文字情報として後世に伝えるために、文字起こしを行う必要がある。しかしながら、膨大な量の音源を人力で行うことは時間、労力、コスト面の問題があり、現実的ではない。そこで本研究では、Google の音声認識API「Chirp」を用いて、歴史的音源の文字起こし作業の効率化を検討する。

キーワード：歴史的音源,音声認識,文字起こし,Chirp

１．緒　言

国立国会図書館では、1900 年初頭から 1950 年頃までに国内で製造された音楽・演説等、約 5 万件の歴史的音源をデジタル化した上で保管している[1]。中でも、語り物と呼ばれる音源については、検索や分析の容易化、そして教育・研究に役立てるために、できるだけ文字起こしをしておくのが望ましい。しかし膨大な音源の文字起こしをすべて人の手で行うとすると、時間や労力、コスト面の問題も含め、言語の変化への対応など様々な問題が存在する。その解決策として近年、音声認識ソフトを用いた作業の効率化が検討されている。

本研究室では令和 3 年度から引き続き、歴史的音源に対する文字起こしの研究を行っている。昨年度は、音声認識 API の最新型「Whisper」を用いて昭和初期の落語の音源の認識を実行したところ、漢字ベースの文字の認識率は 56.9%にとどまった[2]。この結果から、「Whisper」は会話文に対して不向きな API なのではないかという結論に至った。そこで本研究では、会話文の認識に定評のある Google の音声認識 API である「Chirp」を使用することで、歴史的音源に対する文字認識精度の向上を図る。

２．方　法

本研究で使用する Chirp は、Google Cloud 上の Cloud speech-to-text V2 における一つのモデルとして使用することができる[3]。

評価に用いた音源は、昨年度と同じく「居酒屋（一）」である。「居酒屋（一）」は、昭和 5 年頃に録音された三代目三遊亭金馬による作品で、収録時間は 2 分 58 秒、文字数は 1062 文字である。この音源は、日本コロムビアから発売され、SNR（信号対雑音比）は約 17dB で、当時の録音技術の限界を反映している。

評価指標としては、文字認識率を使用する。下記の式を用い、仮名ベースの場合と漢字ベースの場合とそれぞれ文字認識率を算出し、昨年度の「Whisper」による結果との比較を行った。

$$\frac{正解文文字数-誤置換文字数-誤削除文字数-誤挿入文字数}{正解文文字数}\times 100[\%]$$

上式で文字認識率を正確に算出するために、次の 3 つの要素の文字数を手作業でカウントした。

①誤置換文字・・・認識された文字が間違って別の文字に置き換わる
②誤削除文字・・・本来認識される文字が欠落する
③誤挿入文字・・・余分な文字が誤って挿入される

カウントの方法としては、正解文と認識された文とを比較し、誤っている部分を数取器で記録し

た。表 1 に、そのカウント方法の一部を示す。

表 1　認識結果の比較方法

正解文	おはようへーいみやしたおまえなんだだいじんぐうさまのしたがえ×ておりますおみやのしたおまえお
認識された文の一部	おはよう**え**〇いみやした〇**ごう**なんだ〇**おおばやし**さんのしたが**あ**いておりますおみやのし**もごう**〇お

誤置換文字　誤削除文字〇　誤挿入文字×

3．結　果

認識誤り文字数の内訳を表２に示す。Chrip と Whisper を比較すると、全体的に Chirp の方が認識誤り文字数が多いことがわかるが、特に誤削除文字数の割合が多いことがわかる。

表２　認識誤り文字数の内訳

種別	Chirp		Whisper
算出法	漢字ベース	仮名ベース	
誤置換文字数	285	476	301
誤削除文字数	228	287	69
誤挿入文字数	57	52	38

文字認識率の結果を図 1 に示す。Chirp における漢字ベースの文字認識率は 41%、仮名ベースの文字認識率は 29%となり、先行研究で使用された Whisper の 56.9%と 61.6%を大きく下回る結果となった。

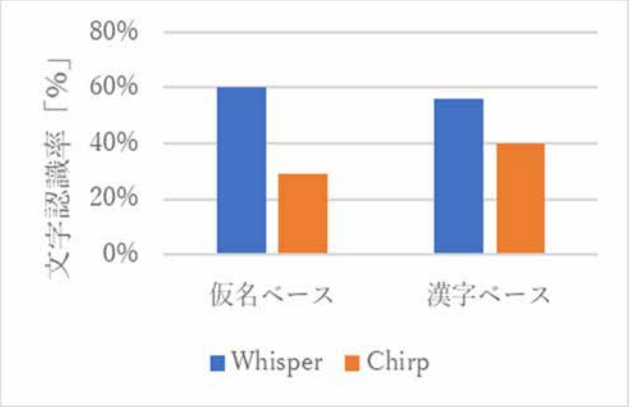

図 1　文字認識率の比較

4．結　言

本研究では音声認識 API「Chirp」を用いて、歴史的音源に対する文字起こしを行った。検証の結果、Chirp は Whisper よりも文字認識精度が低いということが確認された。

今後は、会話文に強いとされている Chirp の文字認識精度が低くなってしまっている原因を究明し、対策を検討する。

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。

文　献

[1] 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/）

[2] 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，令和五年度サレジオ工業高等専門学校機械電子工学科卒業論文

[3] 音声認識モデル　Chirp https://console.cloud.google.com/speech/transcriptions/list?project=robust-tracker-428102-d2