くずし字翻刻のための古文用音声認識システムの検討

研究代表者
文学研究科　戸塚史織
情報理工学研究科　ZHANG Yutao

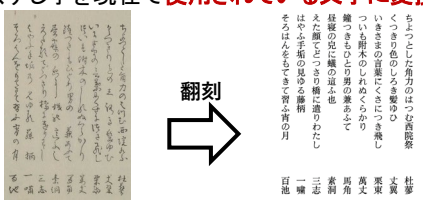
1. 研究背景

◻ くずし字

- 古典籍、古文書などの**前近代資料に使われてきた文字**
- 日本には数百万の古典籍・古文書があり、近年多くの機関がインターネットを通じてこれを公開し始めているが、くずし字で記されているため**現代の多くの人は読めない**

➢ **くずし字翻刻**

- くずし字を現在で**使用されている文字に変換する**



- **負荷が大きい**
- くずし字解読では近年**AIやOCR（光学的文字認識）**技術が注目されている

◻ 現状

- くずし字解読を得意とする人々の多くは**AIやOCR技術**を活用したくずし字解読技術を使用するよりも、使い慣れた**アナログ的な方法**の方がより効率的に翻刻できる
- 翻刻に慣れた人が**過去に自分の手元に手書きで残した翻刻のテキストデータ化**が求められる

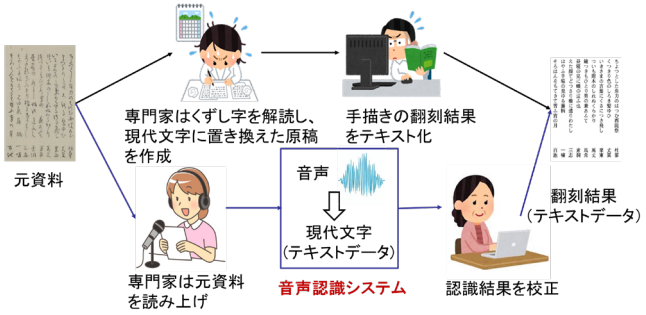
翻刻作業の負荷を減らし、特に翻刻の<u>テキストデータ化</u>を<u>効率化させる</u>ことが現在の重要課題

2. 古文用音声認識システム

◻ くずし字翻刻の音声入力システム

- 負荷を減らし、翻刻のテキストデータ化効率化させるため
  - 画像だけからではなく、**音声**からくずし字翻刻をサポートする方法を検討
  - 翻刻に**音声入力システムの導入**を検討

<u>高精度</u>な古文認識を実現できる音声認識システムが求められる



3. 古文用音声認識システムの構築

◻ 学習データの収集

- 古文テキスト学習データの収集
  - 既存データセットに含まれている作品
    - ➢ 江戸時代洒落本（15作品）　文字数：約17万
  - **独自の学習データ**
    - ➢ 役者評判記の形態素解析データ
      - ◆ 「役者白虎通 江戸の巻」（先頭5000語）（**昨年度**）
      - ◆ 「役者白虎通」の**より詳細な正誤判定を行った解析データ**（京都の巻先頭2000語・江戸の巻先頭2000語の計4000語解析修正データ）（**今年度**）
- 古文音声データセットの収録（**今年度**）

<u>気伝導マイク</u>と<u>皮膚密着型マイク</u>（NAM）を使用して同時に収録

- ➢ 話者：古文を流暢に読める8人の話者（女性6名、男性2名）
- ➢ 読み上げ作品：形態素情報が付与された洒落本作品の翻刻テキスト
- ➢ 収録場所：ARCスタジオの録音ブース
- ➢ 収録音声の長さ：各話者が40分間読み上げ、合計で約5時間のラベル付き音声データを収集

◻ 古文用音声認識システムの構築

- DNN-HMMベースの音声認識システム（**昨年度**）
- 商用音声認識APIを試す（**今年度**）
  - 古文の認識精度を向上させるため、カスタム言語モデル機能を利用（ドメイン固有の音声の文字起こしの精度を向上させるように設計される）
    - ➢ 古文学習データを用いて言語モデルをチューニング
- End-to-End 古文用音声認識モデルを構築（**今年度**）
  - ラベルあり古文音声データが限られているため、**自己教師あり学習モデル**wav2vec 2.0を活用

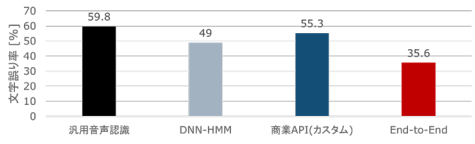
4. 評価実験

◻ 実験目的

- 現時点で構築された古文音声認識システムの性能評価

◻ 実験手順

- テストデータ
  - 洒落本の読み上げデータ　内容：「聖遊廓」の一部
- 認識テストに使用する音声認識システム
  - 汎用日本語音声認識システム
  - DNN-HMMベースの音声認識システム
  - 商用音声認識API（カスタム言語モデル使用）
  - **<u>現時点で構築されたEnd-to-End古文用音声認識モデル</u>**



➢ 昨年度より、古文音声認識システムの**認識性能を向上**させることができた