くずし字翻刻のための古文用音声認識システムの検討

古文用音声認識システム研究会
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

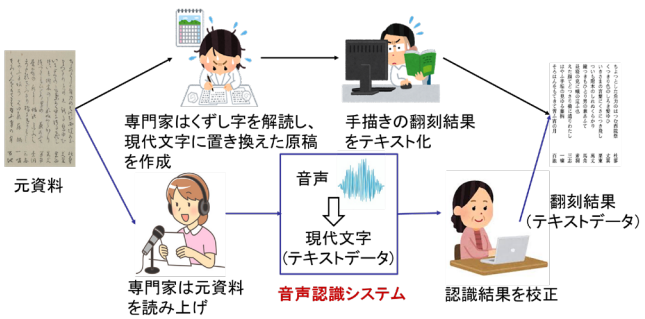
□ 現状

- 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、特に翻刻のテキストデータ化を効率化させることが現在の重要課題

□ くずし字翻刻の音声入力システム

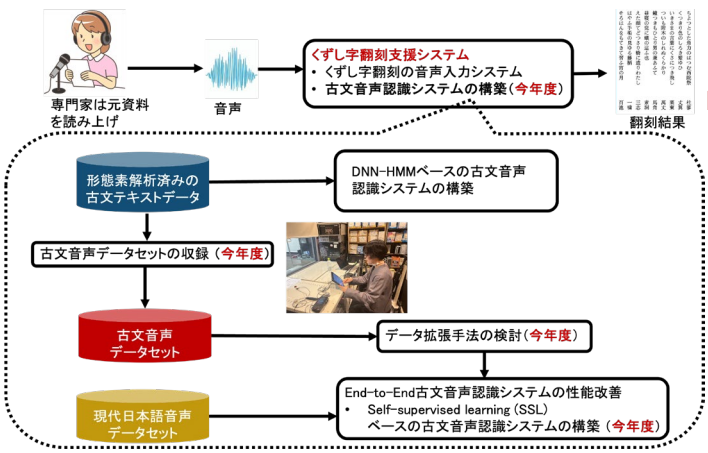
- 負荷を減らし、翻刻のテキストデータ化効率化させるため
  - 画像だけからではなく、**音声** からくずし字翻刻をサポートする方法を検討
  - 翻刻に**音声入力システムの導入**を検討



2. 古文用音声認識システムの構築

□ 現代日本語音声認識システムを直接利用することは困難

- 古文に特有の語彙や文法体系、イントネーションがあるため、認識性能が低下
- **古文用音声認識システム**を構築することは必要



□ 古文音声データの収録（現在音声認識し対象：江戸時代の古文）（**今年度**）

- 話者：古文を流暢に読める11人（女性7名、男性4名）（今年度）7人分のデータを追加収録した
- 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：合計で約7.5時間

□ 古文用音声認識システム

❖ **Problem**：学習データの不足
高精度なEnd-to-End音声認識システムを構築するには、少なくとも数百時間分のラベル付けされた音声データが必要

➢ 大量のラベル付き古文音声データを収集することは困難

➡ **認識性能を改善**するための検討（**今年度**）

➢ 自己教師あり学習モデル(SSL) モデルと現代日本語音声データを利用し、End-to-End古文音声認識モデルを構築

➢ データ拡張手法を検討
学習データセットを人工的に増やし、データの多様性を高める

3. 評価実験

□ 現時点で構築された古文音声認識システムの性能評価

□ テストデータ：
より多様なデータで評価する
長さ：**30分くらい**
（昨年度は**4分くらい**）

	昨年度提案した手法の結果	今年度の結果
文字誤り率	43	37.2

4. 今後の計画

□ **古文音声学習データ**の追加収録

□ **画像と音声情報**を併用したくずし字認識システムの構築

- 画像と音声情報を統合し、識別精度の向上を目指