百認一取：歌読み上げのための音声認識システムの評価

木村 翔真[1]　濱武 右京[1]　黄 思韵[2]
柴田 美桜[2]　筒口 拳[1]　岡本 学[1]

概要：本研究は，コンピュータとかるた取りを行えるシステムを実現する百認一取プロジェクトの一環である．本稿では，Julius を利用した百人一首用の音声認識システムで，マイクから入力された歌読み上げ音声を認識・評価した結果を報告する．本システムの認識精度を向上させることを目的とし，重複している句を含む辞書と削除した辞書を作成して比較実験を行った．さらに，リアルな場での対戦を想定しアナウンサー音声と一般話者の音声で比較実験を行った．2 つの実験により，辞書の単語数を減らすことによって認識精度が向上することと，一般話者の音声でもある程度正しく認識できることを確認した．

キーワード：音声認識システム，言語モデル，システム評価，百人一首，ゲーム AI

Hyakunin Isshu Project: Evaluation of a speech recognition system for song reading

SHOMA KIMURA[†1]　UKYO HAMATAKE[†1]　SIYUN HUANG[†2]
MIO SHIBATA[†2]　KEN TSUTSUGUCHI[†1]　MANABU OKAMOTO[†1]

Abstract: We are working on the Hyakunin Isshu Project, which aims to create a system that can play Karuta with a computer. In this paper, we report the results of recognizing and evaluating song reading speech input from a microphone using the Hyakunin Isshu system that uses the speech recognition engine "Julius." With the aim of improving the accuracy of our system, we created dictionaries that included duplicate phrases, and which deleted them, and conducted a comparative experiment. Furthermore, we conducted a comparative experiment with the announcer's speech and general speakers. Through two experiments, we confirmed that accuracy improves by reducing the number of words in the dictionary, and that even the general speakers can be recognized almost correctly.

Keywords: Speech recognition system, Language model, System evaluation, “One Hundred Poets, One Poem Each”, Game AI

1. はじめに

我々は，百認一取プロジェクトの一環として，音声認識や画像認識を用いてリアルな場で人間とコンピュータが対戦可能な百人一首システムの開発をめざしている[1][2][3][4].「リアルな場」では，取り札は現実世界に配置し，人は読み上げられた音声に対し取り札を手で取る．一方，システムは配置された札の位置をカメラで撮影した画像から認識し，読み上げられた音声の札に投光するなどして勝敗を競う（図 1）.

本システムにおいて，我々はオープンソースの音声認識ソフトウェアである Julius[5][6]を用いて，マイク入力される歌の読み上げ音声を認識し，その歌番号を出力するシステムの作成，および評価を行った[4]．図 2 に音声認識システムの概略図示す.

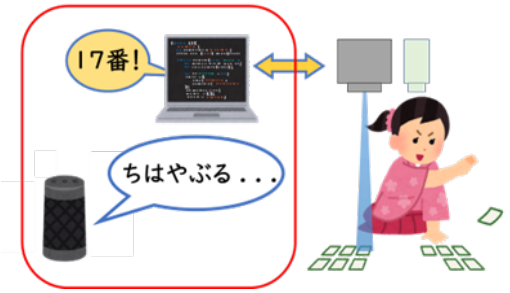

図 1　百認一取プロジェクトの全体図

図 2　音声認識システムの概略図

1 崇城大学情報学部
Faculty of Computer & Information Sciences, Sojo University
2 崇城大学大学院工学研究科
Graduate-School-of-Engineering, Sojo University

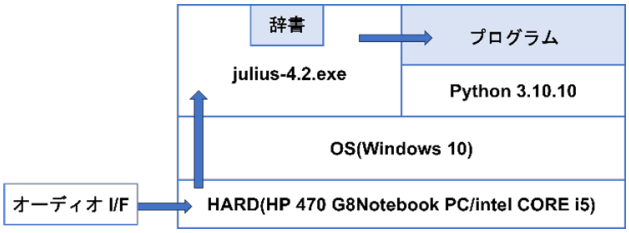

図 3　音声認識システムの構成

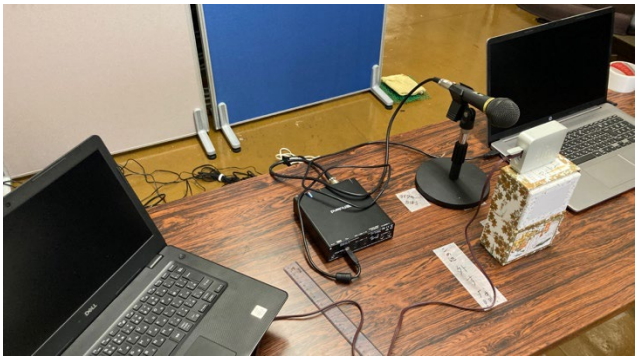

図 4　実験装置の全体像

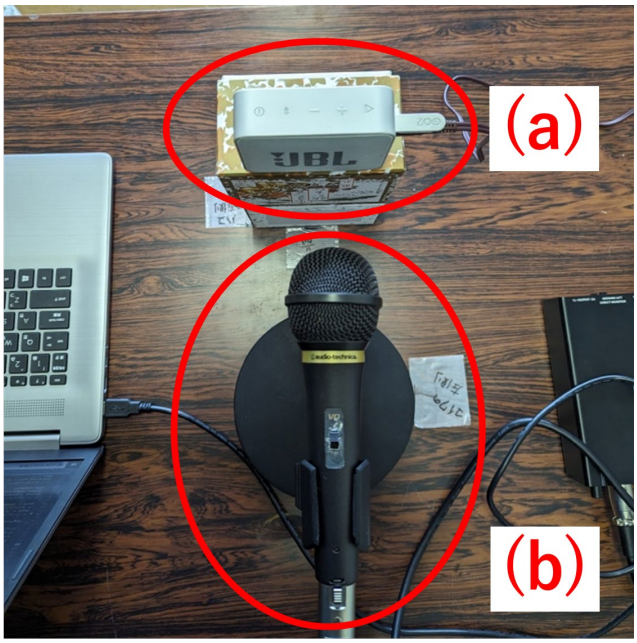

図 5　実験装置（スピーカとマイク）

アナウンサーが百人一首を読み上げた NHK クリエイティブ・ライブラリーの音声[7]（以下，NHK 音声と称する）を用い評価を行った結果，誤認識のパターンとして第 1 句と同じ言葉が別の歌の第 3 句にあり間違えたケース，音声が想定した音素で読み上げてないケースがあった．それ以外で認識自体を間違ったというケースもあり，発話がはっきりしていない音声に対する性能は不明である．

本報告では，辞書の単語数を減らすことで重複する句による誤認識を無くし正答率を向上させる検討を報告する．

また，発話がはっきりしていない音声でも対応できるかを確認するため，一般話者が読み上げた音声（以下，一般話者の音声と称する）を用いた実験についても述べる．

第 2 章で評価環境について説明し，第 3 章で本研究の実装について述べる．第 4 章で実験およびその結果について記述し，第 5 章で考察，第 6 章でまとめを述べる．

2. 評価環境

先行実験[4]では，録音された音声をマニュアル操作で再生し，音声認識システムの出力を記録していた．そのため，実験のたびに音声をカットする部分が変わり，再現性が確保できない，記録が正しいかを後から確認することができないなどの課題があった．

これらの課題を解決し，また周囲のノイズや残響の影響を最小限に抑えるために，評価実験環境を構築した．

これにより，実験の信頼性や再現性が向上し，より安定した結果が得られることが期待される．

2.1 システムの構成

本研究で用いる音声認識システムのうち，音声認識の機能は Julius に百人一首に合わせた辞書を作成し組み込むことで実装した．モジュールモードで実行した Julius から認識結果を受け取り，認識結果の表示と札を取得する処理に渡すためのプログラムを Python で作成した（図 3）．PC（HP 470 G8 Notebook PC・Windows10 Pro・64bit）にはマイク（audio-technica・AT-VD4）が接続されたオーディオ I/F（Roland・Rubx22）を接続する．

また，実験用音声を再生するために別の PC に再生用のプログラムを python で作成し用いた．評価用の音声を PC の音声出力に接続したスピーカ（JBL・Go 2）で自動的に再生し，再生した音声のログを記録する．音声認識システムも，認識結果のログを出力するよう改良し，再生音声と認識結果を照らし合わせ正答率を求められるようにした．評価中，人手による操作を不要とし，実験の再現性を高めた．実験装置の設置状況を図 4 に示す．また，図 5 の(a)はスピーカで，図 5 の(b)はマイクである．

本システムにおいて，外部のノイズや干渉を極力 排除し，システムによる正確な音声データを取得するために，先行実験と下記に記す条件を変更した：

(1) マイクの位置，スピーカの高さ
(2) 実験音声の音量
(3) 評価を行う部屋

スピーカから再生する音声の音量について，先行実験よりも音量を上げた．ただし，本実験では音声を再生するコンピュータも変更したため，音量以外に音声出力の特性影響も考慮する必要がある．そこで，我々は先行実験で利用した音声も再度本実験で評価した．

3. 実装

3.1 辞書ファイル

百人一首に合わせ，文法規則を定義する grammar ファイルと単語集合を定義する voca ファイルを作成した[8].

grammar ファイルは, 第 1 句を FIRST, 第 2 句を SECOND のように定義した.

一方で, voca ファイルは grammar ファイルの構成に従い，句ごとに区切って作成し，百人一首の歴史的仮名遣いの読みに合わせて音素を変更した．音素は文献[9]を参考にした.

先行実験では第 1 句と第 3 句が同じ句を誤認識するという課題があった．また，百人一首では読み上げが始まってから早い段階で正しく認識することが重要なため，百人一首用の辞書を第 1 句と第 2 句のみに限定した.

比較する Julius の辞書には識別ができるよう，辞書の名前に“1”，“2”のように番号を付与し，第 1 句から第 5 句で構成された Dictionary1 と，それを改良し第 1 句と第 2 句で構成された Dictionary2 という名称とした.

それぞれの辞書の voca ファイルと grammar ファイルの違いについては以下の通りとした.

voca ファイル：

Dictionary1 は歌番号のみを出力するように登録した（表 1）．Dictionary2 は第 1 句と第 2 句のどちらが出力されたのかを区別できるよう，“-1”，“-2”を歌番号の後に付けて出力するように登録した（表 3）.

grammar ファイル：

Dictionary1 は, 第 1 句から第 5 句まで（表 2）, Dictionary2 は第 1 句から第 2 句の部分集合をすべて定義している（表 4）.

これにより，実際に読み上げた音声の一部のみ入力されても，認識が可能になると考えられる.

3.2 評価用音声

リアルな場でシステムを利用することを想定し, NHK 音声の明瞭で，発話がはっきりしている音声以外に，発話がはっきりしていない一般話者の音声（あるいは，一般的な人々が歌っているような音声）を 4 人分収録した（表 5）．これら一般話者の音声を，それぞれ A 音声，B 音声，C 音声，D 音声と呼ぶことにする.

チャネル数が 1（モノラル），サンプリング周波数が 32000Hz で収録した．その後，一般話者の音声の音量は平均の音量レベルを A 特性で NHK 音声に合わせた.

評価用音声の作成手順を次のページに示す.

表 1　Dictionary1 の grammar ファイル

```
S       :  NS_B FIRST SECOND THIRD FORTH FIFTH NS_E
S       :  NS_B FIRST SECOND THIRD FORTH NS_E
S       :  NS_B FIRST SECOND THIRD NS_E

S       :  NS_B FORTH NS_E
S       :  NS_B FIFTH NS_E
```

表 2　Dictionary1 の voca ファイル

```
%FIRST
1/      a k i n o t a n o
2/      h a r u s u g i t e

%FIFTH
1/      ts u y u n i n u r e ts u ts u

100/    m u k a sh i n a r i k e r i
% NS_B
<s>             silB
% NS_E
</s>            silE
```

表 3　Dictionary2 の grammar ファイル

```
S       :  NS_B FIRST SECOND NS_E
S       :  NS_B FIRST NS_E
S       :  NS_B SECOND NS_E
```

表 4　Dictionary2 の voca ファイル

```
%FIRST
1-1/    a k i n o t a n o
2-1/    h a r u s u g i t e

%SECOND
1-2/    k a r i h o n o i o n o

100-2/ f u r u k i n o k i b a n o

% NS_B
<s>             silB
% NS_E
</s>            silE
```

表 5　一般話者の音声情報

音声	種類
A 音声	男声
B 音声	男声
C 音声	女声
D 音声	女声

NHK 音声の場合：

(1) 一般話者の音声と同じチャンネル数，サンプリング周波数にする
(2) 音声ファイルに含まれる不要な空白カットし第 1 句のみの音声にする

一般話者の音声の場合：

(1) 100 首の中，第 1 句と第 3 句が重複していない歌からランダムで 25 首選ぶ
(2) 防音室にて上の句（第 1 句から第 3 句まで）を収録する
(3) Audacity[10]を用いて第 1 句のみの音声に編集する
(4) NHK 音声と音量を A 特性で揃える

NHK 音声はもともと第 5 句まで読み上げた音声である．しかし，本検討で行う実験では第 1 句のみを用いるため，第 2 句以降の音声を切り取った音声を用意した．また，一般話者の音声についても同じように第 1 句のみの音声を用意した．

なお，NHK 音声を利用するにあたり，100 首のうち 99 首の音声は，第 1 句のみを出力する音声に加工した．第 74 首は発話内で第 1 句と第 2 句の境目が存在せず，加工を行うと第 1 句の語尾が切れて認識に影響がでる可能性があるのを考慮し，境目がある第 2 句終了部分で切り取った．そのため，第 74 首については，第 1 句と第 2 句両方が正答した場合のみ，正しく認識ができたこととする．

4. 実験

第 3 章で，実装した音声認識システムと用意した評価用音声を用いて，以下の 2 つの実験を行った．

4.1 実験 1：辞書変更の効果の評価

実験 1 は，辞書ファイルへの登録単語数を減らすことにより，重複している句がある歌の誤認識を防ぐことができるかを評価する．

ここで，本実験での認識率の最大値（理論値）は 97%である．百人一首内には第 1 句が同じ歌が合計 6 首あり，それぞれ 2 首ずつ重複している．そのため，この 6 首については言葉自体の認識が出来ていても，半分の確率で間違えると思われる．そこで，6 首が正しく認識された場合，期待値としては，3 首正解するとみなす．残りの 94 首が正しく認識された場合，重複している 6 首を合わせて，正答率の期待値は 97%となる．

NHK 音声 100 首を用い，改良前の辞書 Dictionary1 と改良後の辞書 Dictionary2 の比較実験を行った．

その結果，Dictionary1 の認識率は 90%であった．一方，Dictionary2 の認識率は 94%であった（表 6）．実験 1 で誤認識したものについて，入力した音声が読み上げている句と認識結果として誤って出てきた歌番号（句）を表 7（表 8）に示す．

表 6　NHK 音声（100 首）を用いた比較実験の結果

音声	認識率（Dictionary1）	認識率（Dictionary2）
NHK 音声	90%	94%

表 7　NHK 音声（100 首）における入力音声が読み上げた句（正解）と認識結果として誤って出力された歌番号 (Dictionary1)

読み上げた句	誤って出力された歌番号
33-1：ひさかたの	76
34-1：たれをかも	1
40-1：しのぶれど	39
44-1：あふことの	20
45-1：あはれとも	66
50-1：きみがため	15
64-1：あさぼらけ	31
73-1：たかさごの	34
76-1：わたのはら	11
84-1：ながらえば	68

表 8　NHK 音声（100 首）における入力音声が読み上げた句（正解）と認識結果として誤って出力された句 (Dictionary2)

読み上げた句	誤って出力された句
34-1：たれをかも	55-1：たきのおとは
44-1：おうことの	22-2：あきのくさきの
45-1：あはれとも	66-2：あはれとおもへ
50-1：きみがため	15-1：きみがため
64-1：あさぼらけ	31-1：あさぼらけ
76-1：わたのはら	11-1：わたのはら

今回の実験における認識率の理論値は 97%であるため，理論値に近い値となった．

4.2 実験 2：複数の話者の音声での評価

実験 2 は，発話がはっきりしない一般話者の音声でも正確に認識できるかの評価を行う．改良後の辞書 Dictionary2 を用いて NHK 音声と一般話者の音声の比較実験を行った．一般話者の音声は選定した 25 首のみを収録したため，実験 1 の結果から一般話者の音声と同じ 25 首の結果を抽出し，比較を行う．

実験 2 の結果，A 音声の認識率は 96%，B 音声の認識率は 92%，それ以外の音声の認識率では 100%であった（表 9）．また，入力した音声が読み上げている句と認識結果として出てきた句を表 10 に示す（表 9 と表 10 は，次のページに表している）．

表 9　音声の種類の違いによる認識率

音声	認識率（Dictionary2）
NHK 音声	100%
A 音声	96%
B 音声	92%
C 音声	100%
D 音声	100%

表 10　入力音声が読み上げた句（正解）と認識結果として誤って出力された句（Dictionary2）

音声	読み上げた句	誤って出力された句
A 音声	いまはただ	いまはたおなじ
B 音声	あらざらむ	あらざらむ/ ふゆぞさみしき
B 音声	かぜそよぐ	かぜそよぐ/ ふゆぞさみしき

5. 考察

実験 1 の結果，辞書に登録する単語数を減らすことにより認識率を上げることができた．

この理由としては，第 1 句と第 3 句の重複による誤認識がなくなったことが考えられる．辞書の単語数を減らした効果が出たといえる．具体的には第 1 句と第 3 句が同じ句である，第 40 首，第 73 首，第 84 首を誤認識することがなくなったと考えられる．

一方，誤認識したものに関しては，音素的に似た他の句と誤認識した可能性があった．ここでは，第 34 首の第 1 句と第 55 首の第 1 句，第 45 首の第 1 句と第 66 首の第 2 句といった誤認識結果であった．

また，先行実験と同様，NHK 音声が想定した音素で読み上げておらず，入ってきた音声を認識した際，辞書に登録されたものから似た句を結果として出力している可能性がある．

第 44 首の第 1 句は「あふことの」である．本研究では，辞書を作成する際に，参考した文献[9]では，歴史的仮名遣いである「あふ」は「おう」と読むとなっていたため，それを踏まえて辞書を作成した．しかしながら，NHK 音声は「あふ」を「あう」と発音している．辞書と読み方が違うため，辞書に登録された似た句の歌番号を出力していると思われる．

この対策としては，読み上げる人によって読み方が変わる場合，それぞれを辞書に登録することで誤認識を防ぐことができると考える．

その他，辞書については，より早く正確に認識結果を出力することができるような構成に変更する必要があると考える．

実験 2 の結果から，正解数は NHK 音声と 1，2 首の差であり，一般話者の音声でも，ある程度認識できることがわかった．

A 音声では，表 10 にあるように，途中まで似ている歌の番号が出力されるケースがみられた．

B 音声では，表 10 にあるように，認識結果が 2 つ繋がって出力され誤認識するケースがみられた．この理由として，第 98 首は「かぜそよぐ」であるが，B 音声では語尾まではっきりと発音しており，最後が「う」となっていた．そのため，第 28 首の第 2 句である「ふゆぞさみしき」の認識結果が第 1 句の結果と連続して出力されてしまったと考えられる．

第 56 首は「あらざらむ」である．これも第 98 首を入力させた時と同様，第 28 首第 2 句の認識結果と連続して出力されている．しかし，語尾が「ん」であるため，語尾を伸ばすことによって生じた可能性は低いと考える．実験時の騒音などの影響があった可能性はあるが，現時点では不明である．

また，このように連続して結果が出力された際，片方はノイズなどにより出力されたとすると，Julius の信頼度パラメーターは低くなると考えられるので，信頼度を取得し，高いほうの結果を画像認識の処理に渡すことで対処が可能になると考える．

しかし，各個人の話し方の癖などによって認識結果が左右されることが考えられる．

そのため，今後の実験においてはさらに人数を増やして実験を行い，今回の結果と比較を行いたい．

本研究の実験は，防音室などではなく通常の部屋を用いて実験を行ったため，スピーカとマイクの距離や音量が同じであっても，残響や騒音の問題により正答数が 1~2 首程度変動することも考えられる．

6. まとめ

今回の実験の目的として，辞書の変更によって認識結果を向上させること，またアナウンサー以外の音声を利用して認識を行うことができるかの 2 点について，システムの評価を行った．

実験 1 の結果として，辞書を改良したことによって認識率が 4%向上した．認識が改善された句については，それぞれ第 33 首，第 40 首，第 73 首，第 84 首であり，どれも第 1 句と第 3 句で同じ文字が存在する句であった．

実験 2 の結果として，一般話者の音声であっても認識率に差があまり生じないことが分かった．また，一般話者の中で，C 音声と D 音声の認識率については 100%の認識率を達成することができた．誤認識であった A 音声と B 音声においては，92%以上の正答率が得られた．よりリアルな場で高い認識精度が得られたため，現段階でも対戦することができるシステムであると考える．

一方，誤認識した句についても考察によって改善できる可能性があるため，認識率が 100%をめざしていきたいと考える．

我々が作成した音声認識システムの今後の課題としては,信頼度パラメーターや辞書の改良を行うことによって,より認識率を上げることが挙げられる．

また，さらに大人数で音声を取得し，比較実験を行うことによって各個人の話し方の癖などにも対応できる音声認識システムの構築をめざす．

さらに，画像認識領域と統合することを目的とし，システムの統合を含めて開発を行うことが挙げられる．

7. 謝辞

本研究での実験にあたり，NHK クリエイティブ・ライブラリーより音源をお借りした．また，音声の録音にご協力いただいた方々に感謝の意を表する．

参考文献

[1] 長嶺和紀，角田唯隼，仲茉夏，中内眞希，岡本学，筒口拳：百認一取(1):取札配置画像に対する歌番号付与システム，電子情報通信学会九州支部学生会講演会・講演論文集，D-26，2023．

[2] 吉川唯杜，佐藤礼一郎，安慶直哉，長嶺和紀，岡本学，筒口拳：百認一取(2):取札画像のサイズに対する文字読み取り精度の評価，電子情報通信学会九州支部学生会講演会・講演論文集，D-27，2023．

[3] 伊藤壮真，長嶺和紀，吉川唯杜，角田唯隼，佐藤礼一郎，岡本学，筒口拳：百認一取：色情報を用いた実画像からの取札領域抽出，火の国情報シンポジウム2024．

[4] 濱武右京，木村翔真，黄思韵，柴田美桜，筒口拳，岡本学：百認一取(3):歌読み上げのための音声認識システムの検討，電子情報通信学会九州支部学生会講演会・講演論文集，D-28，2023．

[5] 河原達也：大語彙連続音声認識エンジン Julius，〈https://julius.osdn.jp/index.php〉（参照 2023-07-18）．

[6] Akinobu Lee: Julius GitHub，〈https://github.com/julius-speech/julius〉（参照 2024-02-07）．

[7] 荒木雅弘：フリーソフトでつくる音声認識システム第 2 版，pp.166-216，森北出版（株），2017．

[8] NHK：NHK アーカイブス，〈https://www.nhk.or.jp/archives/creative/〉（参照 2024-02-08）．

[9] 本田:百人一首の読み方を小学生のために現代仮名遣いで一覧に，〈 https://honda-n2.com/ogurahyakuninisshu-hiragana-ichiran-shougakusei〉（参照 2024-02-08）．

[10] The Audacity Team: Audacity，〈https://www.audacityteam.org/〉（参照 2024-02-09）．