音声言語情報処理　27−13
（1999．7．24）

「読み」情報を利用した大語彙連続音声認識

廣瀬良文[†]　伊藤克亘[††]　鹿野清宏[†]　中村哲[†]

[†] 奈良先端科学技術大学院大学 情報科学研究科
〒630-0101 奈良県生駒市高山町 8916-5
[††] 電子技術総合研究所
〒305-8568　茨城県つくば市梅園 1-1-4
E-mail: yosifu-h@is.aist-nara.ac.jp, kito@etl.go.jp,
{shikano,nakamura}@is.aist-nara.ac.jp

あらまし　日本語の表記法は、かな漢字混じりであるため、同一の語に対しても多様な表現が可能である。そのため、大語彙連続音声認識において語彙を制限した場合、言語の被覆率が低下してしまう。ディクテーションなどでは、より広い語彙を被覆することは非常に重要である。本研究では、形態素解析により得た「読み」情報を利用することにより、(1) 少数のエントリで従来と同程度の被覆率を実現し、認識時の探索空間を小さくする。(2)「読み」情報を利用することにより「読み」「漢字」混合モデルでより高い被覆率を実現し、入力音声中の未登録語を対処する実験をおこなったので報告する。

キーワード　ディクテーション, 言語モデル, 被覆率, 未登録語

Large Vocabulary Continuous Speech Recognition Using Reading-based Language Model

Yoshifumi Hirose[†]　Katsunobu Itou[††]　Kiyohiro Shikano[†]　Satoshi Nakamura[†]

[†] Graduate School of Information Science, Nara Institute of Science and Technology,
Takayama-cho, Ikoma,Nara 630-0101, Japan
[††] Electrotechnical Laboratory,
1-1-4,Umezono,Tsukuba,Ibaraki,305-8568 Japan
E-mail: yosifu-h@is.aist-nara.ac.jp,kito@etl.go.jp,{shikano,nakamura}@is.aist-nara.ac.jp

Abstract

As Japanese sentences are written using kana and kanji, some words are often written in various ways. In large vocabulary continuous speech recognition, where the vocabulary size is generally restricted, this word writing ambiguity causes the deterioration of the word coverage.

In this paper, we propose the following reading-based language models, (1) a reading-based language model with the same coverage by smaller vocabulary than morpheme-based language model, and (2) a kana and kanji mixed language model to cope with unknown words.

The effectiveness for these language models are confirmed by the LVCSR experiments.

key words　dictation, language model, word coverage, unknown word

1　はじめに

近年の急速な計算機技術の進歩により、人とコンピュータのインターフェースとして、音声を使用するシステムが多く現れてきた。なかでも、音声ワープロやディクテーションなどの大語彙連続音声認識は、音声認識技術を結集して行なわれ、非常に盛んに研究がなされている。

大語彙連続音声認識においては、一般に音響モデルと言語モデルを使用して認識が行なわれる。大語彙連続音声認識の語彙としては、無限の語彙を扱うことは不可能であり、通常は言語モデルの学習データ中に出現する高頻度語を語彙として使用することが多い。このように、制限語彙を使用する場合には、語彙に含まれない単語 (未登録語) は絶対に認識することはできない。 また、入力音声中に未登録語が存在した場合は、その単語のみが認識誤りを起こすのではなく、その前後にまで影響を及ぼすと考えられる。したがって、このような未登録語に対する処理は必要不可欠で、これまでにも研究がなされている [1, 2, 3]。

大語彙連続音声認識でよく用いられる言語モデルとして N-gram 言語モデルが挙げられるが、日本語においてこの N-gram モデルを構築するには、大量のテキストデータが必要となり、語彙はこのテキストデータに出現したものを使用することになる。しかし、日本語は、同一の語を記述する場合においても複数の表記をすることが可能であるため、さまざまな問題が生じる。

本稿では、N-gram 言語モデルを構築する際の日本語特有の問題点について、形態素解析結果の「読み」情報を利用することによる改善を試みたので報告する。

2　大語彙連続音声認識における統計的言語モデル

2.1　N-gram 言語モデル

一般に大語彙連続音声認識では、N-gram 言語モデルがよく使用される。これは、単語の生起を $N-1$ 重マルコフ過程で近似したモデルである。N-gram モデルでは、ある時点での単語の生起は、直前の $N-1$ 単語にのみ依存すると考えるので、次式のように定式化することができる。

$$P(w_n|w_1\cdots w_{n-1}) \quad = \quad P(w_n|w_{n-N+1}\cdots w_{n-1})$$

N=2 の時を **bigram**、N=3 の時を **trigram** と言う。

2.2　日本語特有の問題点

大語彙連続音声認識において N-gram 言語モデルを作成するときの日本語特有の問題点として、次のようなものが挙げられる。

(1) 文章がわかち書きされていないため、形態素解析システムなどを用いて文章を N-gram の単位に分割する必要がある。

(2) 日本語の表記は、かな漢字混じりであるため、同じ語でも、複数の表記が可能である。そのため、高頻度語により語彙を制限した時の単語カバー率が低下する。

(1) の問題点は、形態素解析システムを使用することにより、解決することが可能である。本研究においては、形態素解析システムとして、Chasen[6] を使用した。また、形態素解析用の辞書は、日本語ディクテーション用にエントリと読み情報を整備したものを使用した [4]。

一方、(2) の問題点により単語カバー率が低下すると、制限語彙中に含まれない単語 (未知語) が増加する。未知語が存在する文章を認識しようとすると、未知語だけではなく、その前後の単語にも悪影響を与える。そこで、本稿では形態素解析結果の「読み」情報を利用することによる改善策を検討した。

「読み」情報を用いることにより、同一の読みのものは表記が統一され、結果的には高頻度語による制限語彙を使用した場合と比較して単語カバー率を改善することが可能であると考えられる。

そこで、次に示す 2 つのアプローチで、前述した日本語特有の問題点を解決しようと試みた。

(1) 「読み」情報のみを利用することにより、少数の語彙エントリで、従来の「漢字」表記によるものと同等の単語カバー率を得ることができる。認識辞書は小さくなり、認識時の探索空間を狭くできる。

表 1: 「読み」と「漢字」によるカバー率

	カバー率	
エントリ数	「漢字」表記	「読み」表記
5k	88.16 %	91.90 %
20k	96.35 %	98.54 %

(2) 学習データ中に出現する高頻度語以外の語を「読み」情報のみに変換する。この高頻度語以外は、「読み」情報をもとに N-gram 確率を学習することにより、単語カバー率を向上させる。

3 「読み」表記によるエントリ数の削減

3.1 言語モデルの作成

毎日新聞 91 年 1 月～ 94 年 9 月までの 45 カ月分の学習データを用いて、「読み」表記による言語モデル、「漢字」表記による言語モデルを作成した。言語モデルの作成には、CMU-SLM-toolkit[5] を使用した。形態素解析には Chasen[6] を使用した。形態素解析結果は、"形態素", " 読み","原形","品詞情報"である。

「漢字」表記の N-gram の単位は"形態素"+"読み"+"原形"+"品詞情報"とし、「読み」表記では "読み"のみを使用した。 表 1 に学習データに対するカバー率を示す。

表 1 から、エントリ数が 5k では 3.74%、20k では 2.19% カバー率が改善できていることがわかる。この結果から「読み」情報のみを用いることにより、「漢字」表記よりも、少ないエントリで言語モデルを構築することが可能となる。

本稿では、従来の「漢字」表記のエントリで表現できる語彙について、「読み」情報のみの言語モデルを作成した。この時、言語モデルのエントリ数は 5k の語彙では、「漢字」のシステムで 5,000、「読み」のシステムでは 4,207 となり、同様に 20k の語彙では「漢字」で 20,000、「読み」で 14,954 となった。

3.2 実験

3.2.1 言語モデルの評価

言語モデルの評価データを毎日新聞 94 年 10 月～ 12 月としたときの、言語モデルのテストセットパープレキシティを表 3に示す。この表から、「読み」表記で作成した言語モデルの方が、カバー率が 5k で、1.9%、20k で 0.94% 改善できている。

「読み」表記で作成した言語モデルのパープレキシティは、従来法の「漢字」表記で作成したものと比較して、bigram では差があるものの、trigram では余り差が見られないことがわかる。

3.2.2 実験方法

本稿の認識実験においては、京都大学で開発された大語彙連続音声認識エンジン JULIUS[7] を使用した。JULIUS は段階的探索法を採用しており、第 1 パスでは bigram を用いたフレーム同期のビームサーチを行ない、第 2 パスでは trigram を用いた A^* サーチによるスタックデコーディングを行なっている。このとき A^* サーチのヒューリスティクスとして、第 1 パスの前向きスコアを用いて探索を行なっている。

この認識実験では「読み」情報のみを利用しているので、最終的には「漢字」表記に変換する必要が生じる。そこで、「漢字」表記への変換法として次の 2 つの方法を実装し、評価実験を行なった。かな漢字変換は「漢字」表記の言語モデルの"読み"に従って変換を行なう。

(1) 「読み」(カナ) 表記の言語モデルで、第 2 パスまで認識を行なう。カナ認識結果を第 3 パスで、「漢字」表記の trigram での文の生起確率が最大になるような「漢字」系列への変換を行なう。

(2) 第 1 パスは、「読み」表記の bigram を用いて認識を行なう。第 2 パスで、漢字変換を行ないながら、「漢字」表記の trigram を用いて A^* スタックデコーディングを行なう。

3.2.3 実験結果

日本音響学会新聞記事読み上げコーパスのうち男性話者 23 名による、語彙規模 5k 語および 20k 語の読み上げ音声各 100 文を評価セットとして認識実験を行なった。音響モデルとしては、日本音

表 2: 実験条件

sampling 周波数	$16kHz$	
フレーム長	$25ms$ (ハミング窓)	
フレーム周期	$10ms$	
プリエンファシス	0.97	
特徴パラメータ	メルケプストラム 12 次元 + Δ ケプストラム 12 次元 + Δ パワー (計 25 次元)	
音響モデル	混合数 状態数 triphone	16 2,000
言語モデル	学習テキスト: カットオフ	毎日新聞 91.1 ~ 94.9 bigram 1 trigram 2

響学会研究用連続音声データベースの男性話者全部と新聞記事読み上げ音声コーパスのうち男性話者 100 名分で学習した 状態数 2,000 、混合数 16 の triphone HMM を使用した。実験条件を表 2 に示す。

表 4に各手法による認識率を示す。認識率は出力と同じ表記の正解ファイルと自動比較した結果である [8]。単語正解率 (Corr.) および単語正解精度 (Acc.) は次の式で計算している。

$$Corr. = \frac{(\text{総数} - \text{挿入誤り} - \text{脱落誤り})}{\text{総数}}$$

$$Acc. = \frac{(\text{総数} - \text{挿入誤り} - \text{脱落誤り} - \text{挿入誤り})}{\text{総数}}$$

5k 評価セットに対して、方法 (1) では、第 2 パスでの単語正解率は、従来法と比較して、1.2% 良くなっていることがわかる。これはエントリ数が従来法と比較して削減できているため、第 2 パスでの探索空間を狭くすることができ最適な候補を出力できた結果であると考えられる。 一方、方法 (2) では、従来法と比較すると 1.12% 低下した。

20k 評価セットに対しては、方法 (1) では、正解率は 0.57% 低下しているもののほぼ同等の精度を得ることができている。しかし、方法 (2) では正解率は 1.64% 低下している。

この理由として考えられるのは、これは認識エンジン内の第 2 パスで A* サーチのヒューリスティックとして「読み」のモデルを使用しているので、「漢字」のモデルとの整合性がうまくとれていないからであると考えられる。また、第 2 パスで生成される仮説数を各方法で比較してみると、5k セットの場合、従来法 22,490 、方法 (1) 22,794、方法 (2) 97,248、20k セットの場合、従来法 48,360、方法 (1) 42,028、方法 (2) 222,804 と仮説数が方法 (2) で大きくなっている。このことから方法 (2) の第 2 パスで生成される仮説数が「漢字」変換の際に膨大になり、A* サーチで最適解が見つけられていないからであると考えることができる。

表 3: テストセットパープレキシティ

	5k		20k	
	「漢字」	「読み」	「漢字」	「読み」
entry	5,000	4,207	20,000	14,954
bigram	53.42	60.35	75.20	82.35
trigram	36.26	38.60	50.52	51.96
OOV	11.88%	9.98%	3.96%	3.02%

表 4: 単語認識率 (%)

		5k set	20k set
	pass	Corr./Acc.	Corr./Acc.
従来法	1st. 漢字	78.78/77.02	75.96/74.89
	2nd. 漢字	92.79/91.51	91.48/89.84
方法 (1)	1st. 読み	78.61/76.68	77.54/75.84
	2nd. 読み	93.99/92.47	91.61/89.97
	3rd. 漢字	93.03/91.51	90.91/89.21
方法 (2)	1st. 読み	78.61/76.68	77.54/75.84
	2nd. 漢字	91.67/90.55	89.72/87.26

以下にその一例を示す。方法 (2)' は第 2 パスで扱える仮説数を 10 倍に増やして認識を行なった結果である。この場合、第 2 パスで扱える仮説数を大きくとることにより、正しい候補を出力することができている。

正解 :	... 大会 に 比べ 二 個 マイナス 。
方法 (2) :	... 大会 に 比べ 二 五 マイナス 。
方法 (2)':	... 大会 に 比べ 二 個 マイナス 。

3.3 まとめ

「読み」情報のみを用いることにより、大語彙連続音声認識におけるエントリ数の削減、およびカバー率の改善を試みた。認識エンジンの第 2 パス内での実装では、A* サーチの問題から従来法ほどの認識精度を得るにはいたらなかった。また「読み」による認識結果を後処理的にかな漢字変換することにより、若干の認識精度の改善が見られた。

以上のことから、かな漢字変換システムにおいて、方法 (1) の「読み」情報のみを用いた大語彙連続音声認識を入力部として導入すること等が可能であると考えられる。

表 5: 各言語モデルの学習データに対する被覆率

モデル	被覆率
「漢字」	96.4 %
「読み」	98.1 %
「読み」「漢字」混合	97.3 %

4 「読み」「漢字」混合モデルによる大語彙連続音声認識

4.1 言語モデルの作成

毎日新聞 91 年 1 月～ 94 年 9 月までの 45ヶ月分の学習データを使用して、以下の手順で言語モデルを学習した。

1. 学習データ中に出現する「漢字」表記による 20k 高頻度語以外の全ての単語を「読み」表記 (カタカナ) に変換する。
2. 新しく「読み」「漢字」混合語彙を以下のように作成する。
   (a) 「漢字」表記での 20k 高頻度語
   (b) 「読み」表記での 20k 高頻度語において、(a) に含まれない語

   (a) と (b) をマージして「読み」「漢字」混合語彙とした。語彙サイズは、24002 単語であった。
3. 1. で作成したテキストを学習データとして、2. の「読み」「漢字」混合語彙を用いて、N-gram 言語モデルを作成した。

「読み」「漢字」混合モデルでは、従来の「漢字」表記の言語確率だけでなく、「読み」表記での言語確率も同時に学習することが可能である。

各言語モデルの学習データに対する単語カバー率を表 5に示す。新たに作成した「読み」「漢字」混合モデルは「読み」表記モデルと比較した場合、0.8% 及ばないものの、従来の「漢字」表記に対しては、0.9% 単語カバー率が改善されていることがわかる。このことから、本稿で作成した言語モデルを使用することにより、入力音声に含まれる高頻度語以外の語が入力された場合にも、「読み」表記で出力することができる可能性を示すことができた。

4.2 実験

4.2.1 言語モデルの評価

評価用データとしては、日本音響学会新聞記事読み上げコーパス 男性話者全体の発話テキストを学習モデルと同じ形態素システムで処理したときに、高頻度語以外の語を 3 つ以上含む文を抽出した [2]。発話データのテキストは学習データには含まれていないので、言語モデルに関してはオープンな条件での実験である。107 話者による 187 文が該当した。

評価用データには、「漢字」表記での 20k 高頻度語に含まれない 574 個の未登録語が存在する。しかし、実際にはそのすべてが未登録というわけではなく、登録語を複数組み合わせることにより構成することが可能となる単語も存在する。1 単語で構成できるもの (形態素が同じ) は 98 単語、2 単語で構成できるものは 92 単語、3 単語で構成できるものは 21 単語存在した。また、「読み」「漢字」混合語彙でカバーできた未登録語は 113 単語 (16 単語は登録語から合成可能)、「読み」表記語彙では 282 単語 (同 57 単語) であった。

表 6: テストセットパープレキシティ

	「漢字」	「読み」「漢字」
bigram	113.9	127.1
trigram	86.0	96.3
OOV	10.8%	8.67%

表 6に各言語モデルの評価用データ (187 文) に対するテストセットパープレキシティを示す。この表からテストセットパープレキシティが、「読み」「漢字」混合モデルを使用することにより、若干大きくなってることがわかる。しかしながら、未知語率が 2.1% 減少していることから、実際にはほとんど影響しないと考えることができる。

4.2.2 実験結果

前述の評価用データに対して、認識エンジン JULIUS を使用して認識実験を行なった。実験条件は表 2に示す通りである。言語モデルの重みは固定で行なった。

表 7に各モデルの単語認識率を示す。正解文に出現する未登録語に対しては、正解率が高くなるように人手により、アライメントを取り直した。例えば、未登録語"米子高島屋"に対して 認識結果"ヨナゴ" "高島屋"を正解にした。この表より、漢字 20k の語彙で認識を行なった時よりも、「読み」「漢字」混合の語彙により認識を行なった場合の方が、単語正解精度で 4.92% 改善していることがわ

表 7: 単語正解率 (%)

モデル	Corr.	Acc.
「漢字」	73.05	65.73
「読み」「漢字」混合	76.71	70.65
「読み」	78.07	71.79

表 8: 認識できた未登録語の例

人名	地名	普通名詞
シマズ	サツマ	ブドーキューキン
テラヤマ	ヨナゴ	ダイレクトメール
シュージ	エトロフドー	フルホンヤ
ナガセ	リットー	ローソク
ニノミヤ		フェアプレー
ホリグチ		キョーシュージョ
ケンジロー		テンモンダイ

かる。また、「読み」「漢字」混合の語彙に含まれた評価用データ中の未登録語 113 単語のうち、82 単語 (72%) が正しく認識することができた。82 単語のうち、71 単語は、「読み」表記の 1 単語で認識でき、4 単語は「読み」表記と登録単語の 2 単語の組み合わせで認識ができ、残りの 5 単語は登録単語 2 単語として認識することができた。一方、本来登録語であるのに、「読み」表記で認識された単語は 2 単語であった。このことから、本手法による悪影響はほとんどないと考えることができる。

以下に、正しく認識できた例を示す。この場合、正解文中に含まれる未登録語は、"薩摩"と"島津"の 2 単語である。「漢字」表記で認識した場合は、両単語とも未登録語であるため、誤った認識を行なっている。特に"島津" に対しては、次の単語である"氏"にまで影響を与えている。一方、「読み」「漢字」混合の場合では、両未登録語ともに「読み」表記で認識することができている。

正解:	薩摩	の	島津	氏	が	攻撃
漢字:	三 妻	の	始末		が	攻撃
混合:	サツマ	の	シマズ	氏	が	攻撃

表 8に、評価用データに対して、「読み」表記で正しく認識できた未登録語の例を示す。正しく認識されるものは名詞が多く、特に人名・地名などが認識されやすい傾向にあった。

しかし、人名・地名など多様な表記が可能なものに対しては有効であるが、固有名詞など表記が一意に決定できるものは、「漢字」表記で登録する方が望ましいと考えられる。

4.3 まとめ

日本語の表記の特徴を考慮して、「読み」「漢字」混合の語彙を使用して言語モデルを作成し、単語カバー率の高い言語モデルを作成した。

未登録語を含む評価データに対して、未知語率、及びテストセットパープレキシティ、単語正解率による評価を行なった。未知語率で、2.1% 改善することができた。また、単語正解精度も 4.92% 改善することができた。

「読み」「漢字」混合の語彙による言語モデルを使用することにより、登録語以外の入力音声に対して、人名・地名など比較的他数の読みが存在するものの一部に対して、正しく認識することができた。

以上のことから、本手法が比較的容易に実装することが可能な未登録語を含む入力音声に対処する方法であると考えることができる。

謝辞

本研究は、情報処理振興事業協会「独創的情報技術育成事業」の一環として行なわれた。言語・音声資源として毎日新聞社による CD-毎日新聞データ集、音響学会データベース委員会による JNAS 新聞記事読み上げコーパスを利用した。関係者各位のご支援に感謝致します。

参考文献

[1] 伊藤克亘, 速水悟, 田中穂積. 連続音声認識における未知語の扱い, 信学技法,SP91-96,Dec.1991.

[2] 伊藤克亘, 田中穂積. 被覆率を重視した大語彙連続音声認識用統計的言語モデル. 音学講論, pp.65-66,Mar.1999.

[3] 甲斐, 廣瀬, 中川. 単語 N-gram 言語モデルを用いた音声認識システムにおける未知語・冗長語の処理. 情報処理学会論文誌 pp1383-1394,Vol.40 No.4 Apr.,1999

[4] 山本, 伊藤, 山田, 鹿野, 中村. ディクテーションにおける形態素辞書エントリと読みの整備の効果. 音学講論, pp53-54 Mar.1999.

[5] Ronald Rosenfeld. The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation. In Proc. ARPA SLS Workshop, pp.47-50,Jan.1995

[6] 松本, 北内, 山下, 今一, 今村. 日本語形態素解析システム『茶筅』 version1.5 使用説明書, 奈良先端大学松本研究室.

[7] 李, 河原, 堂下. 単語 N-gram と段階的探索に基づく大語彙連続音声認識エンジン JULIUS. 音学講論, pp.51-52, Mar.1998.

[8] 山本, 伊藤, 鹿野, 中村. ディクテーションにおける日本語の特性を考慮した単語正解率判定ツール. 音学講論, Mar.1999.