論文 / 著書情報
Article / Book Information

論題(和文)	ニュース音声を対象とした大語彙連続音声認識
Title(English)	
著者(和文)	田口雄一, 大附克年, 松岡達雄, 古井貞熙, 白井克彦
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 1997年春季講演論文集, Vol. , No. 2-6-11, pp. 65-66
Citation(English)	A Large Vocaburaly Continuous Speech Recognition Using Broadcast News Speech., Vol. , No. 2-6-11, pp. 65-66
発行日 / Pub. date	1997, 3

2—6—11　ニュース音声を対象とした大語彙連続音声認識*

◎田口雄一[1] 大附克年[2] 松岡達雄[2] 古井貞熙[23] 白井克彦[1]
([1]早大 [2]NTT ヒューマンインタフェース研究所 [3]東工大)

1　はじめに

大語彙連続音声認識の研究では，これまで新聞記事をはじめとするテキスト読み上げ音声を対象とした研究が盛んに行なわれてきた．日本語では1995年，日本経済新聞を題材とする読み上げ音声データベースが構築され[1]，7000語彙での実験において90%に近い認識結果が報告されている[2]．

一方で，ニュース音声のように，読み上げ音声とは異なるより自然な発声に近い音声を扱った大語彙連続音声認識についての研究はほとんど行なわれていなかったが，1995年ARPAではニュース音声を対象としたプロジェクトが発足し，その音声認識においては読み上げ音声に比べて非常に困難な題材であることが報告された[3]．しかし，これらを大語彙音声認識に適用することは，字幕の書き起こし，放送素材へのインデックスの付与または音声による検索など，応用が広範囲である．本研究ではNHKニュース番組の原稿によるテキストデータベースから言語モデルを構築し，ニュース音声[4]を対象とする大語彙連続音声認識に適用した．さらに日経新聞記事による言語モデルと比較することにより，ニュース音声認識における新聞記事の有効性について検討した．

2　ニュース原稿テキストデータベース

2.1　テキストデータベース

1992年8月から1996年8月までのニュース原稿のテキストデータベースから言語モデルを学習するため，テキストに対して幾つかの前処理を施した．

原稿は日付，見出し，本文などからなるが，本研究では本文のみを対象とした．また，日本語の文章は英語のように単語間に空白が存在せず，そのままでは単語境界が明確ではない．しかし連続単語認識を試みるためにはテキストを単語に区切る必要がある．本研究では語彙の単位として形態素を採用し，ニュース原稿テキストに対して形態素解析を行なった．なお，用いた形態素解析ツールは[1]で使用されたものと同一である．

2.2　テキスト前処理

ニュース原稿テキストには音声認識とは無関係な記号などが含まれるが，言語モデルにはこれらは含まれないことが望ましい．すべてのテキストについて以下のような前処理を施した．

- ◎○●◇◆□■△▲▽▼※「」『』・などを削除
  これらの記号は主に注意を引きつけるためのものであり，発声されないことから削除する．
- ( ) [ ] 《 》は中身ごと削除
  これらは主に文の説明，特に単語の読みがなに使われる例が多く，発声されないため，削除することに支障はない．

3　言語モデル

連続音声認識に用いる言語モデルをニュース原稿DB，日経新聞DBから個別に学習した．その語彙をニュース原稿DBに出現する単語の頻度の高いものから選択し，語彙サイズはWSJのカバー率を参考に20k語に決定した．その内訳を表1に，また学習テキストデータ量を表2に記す．

表1: 語彙サイズと単語カバー率(%)

ニュース		日経		WSJ(米国)	
size	cov.	size	cov.	size	cov.
5k	91.5	7k	90.3	5k	91.7
20k	98.0	30k	97.5	20k	97.8
63k	99.7	150k	99.6	64k	99.6
114k	100.0	623k	100.0	173k	100.0

表2: 学習テキストデータ量

	ニュース	日経
総文数	477k	6M
総単語数	24M	180M

ニュース原稿DBは114k語で閉じている．そのデータ量は日経DBに比べて非常に少量であり，20k語の正確な言語モデル推定に十分である確証はない．

こうして構築した言語モデルの規模を表3に示す．データ量が少ないため，ニュース原稿言語モデルのbigramの総種類数は日経新聞言語モデルに比べてはるかに少なくなっている．

表3: 言語モデルの規模

	モデル	総種類数	平均頻度
ニュース	unigram	20k	1160
	bigram	0.9M	24
日経	unigram	20k	8747
	bigram	3.6M	44

4　音響モデル

4.1　Tree-based clustering に基づく音素HMM

今回の実験で採用した音響モデルは，Tree-based clustering に基づいて状態共有化を行なった音素HMMである．学習に用いたデータを表4に示す．構築されたモデルの総状態数は2106となった．なお，混合数はすべて4とした．

表4: 音声データ

音響モデル学習用	ASJ 音素バランス文 連続音声DB ASJ 案内タスク連続音声DB ATR B (計53名, 13270発話)
音素認識評価用	NHK ニュース (5名, 49文) 日経新聞 (10名, 100文)

4.2　音素認識実験

音響モデルの性能を調査し，ニュース音声および新聞記事読み上げ音声に対する音響モデルの性能の違いを調べるために音素認識実験を行なった．日本語の音節の制約に基づく音素ネットワークを用いて連続音素認識を行ない，正解率(%Correct)および正解精度(Accuracy)を算出した(表5)．この結果，ニュース音声は新聞記事読み上げ音声

* A Large Vocabulary Continuous Speech Recognition Using Broadcast News Speech.
By Yuichi Taguchi[1],Katsutoshi Ohtsuki[2],Tatsuo Matsuoka[2],Sadaoki Furui[23] and Katsuhiko Shirai[1]
([1]Waseda University,[2]NTT Human Interface Laboratories,[3]Tokyo Institute of Technology)

と同等の結果を得た．このことから音響レベルでの認識の難しさは変わらないことがわかった．

表 5: 音素認識結果

	Cor.	Acc.
ニュース	82.0	61.5
日経	80.7	64.7

5 大語彙連続音声認識

5.1 評価用音声データ

連続音声認識に用いる評価セットを選択した．スタジオ以外からの中継や音楽が重なった音声はすべて除外し，その話者がメインのアナウンサーであるか (anchor)，否か (others) に着目し，2 つのテストセットを構築した．また，比較に用いる日経新聞読み上げ音声には 30k セット†を採用した．

表 6にその概要を記す．

表 6: テストセット

	anchor	others	nikkei(30k)
話者	5 名	6 名	10 名
総発話数	100	125	100
総単語数	4184	2285	2168
未知語率	0.9%	3.7%	3.5%

5.2 実験結果

それぞれのテストセットについて連続音声認識実験を行なった．その結果とテストセット・パープレキシティを図 1 に重ねて記す．パープレキシティは bigram モデルから算出

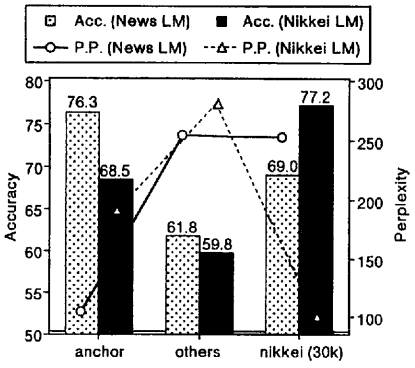

図 1: 連続単語認識正解精度とパープレキシティ

した値である．anchor に対するニュース原稿，日経新聞の各言語モデルによる結果を比較すると，前者において正解精度 76.3%が得られたのに対し，後者においては 68.5%であり，その差は顕著である．パープレキシティによる比較ではニュース原稿の 104 に対して日経言語モデルでは 190 となったことから，認識性能が言語モデルに依存していることがわかる．同様のことが nikkei テストセットについてもいえる．

さらに，anchor，others のセットを話者毎に評価する (図 2)．その認識結果の差は非常に大きく，認識性能は話者にも依存するという結果を得た．しかしパープレキシティとの相関係数が 0.652 と算出され，ある程度の相関関係が

†日経新聞データベースにおける出現頻度上位 30k 語からなるテストセット

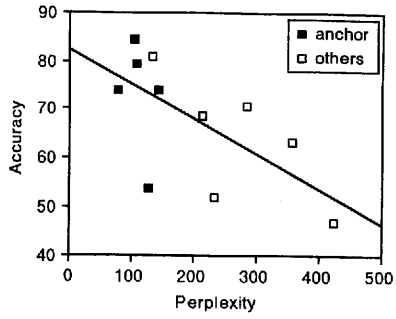

図 2: 話者毎の評価

得られたことから，その性能が低い話者についてはその発話内容自体が困難であった場合があるといえる．

6 まとめ

本稿では，NHK ニュースの原稿を用いた言語モデルの推定，およびその連続音声認識実験結果について報告した．ニュース原稿テキストから語彙を 20k 語に設定し，その語彙においてニュース原稿 DB，日経 DB から言語モデルを構築して音声認識を試みた．ニュース原稿から学習した言語モデルを用いた場合にニュース音声に対して単語正解精度 76.3%という結果を得た．

これらの言語モデルを使用しない音素認識において同等の結果が得られたにもかかわらず，連続音声認識においてはタスクに非常に依存した結果が得られた．それぞれのタスクについて言語的な構造が異なることがパープレキシティとの相関からも裏付けられるが，ニュース音声を認識しようとする立場から見れば，新聞記事のように大量に入手可能なテキストが学習に使えることが望ましく，タスク適応化技術についても検討が必要である．

今回用いたニュース音声と日経新聞記事には時期的におよそ 2 年間の隔たりがあり，認識対象に出現する単語連鎖 (「住専」,「阪神大震災」など) を学習できなかったことを考慮すれば，新たな記事を導入することによって言語モデルに適用できる可能性は残されているといえる．

単語 trigram による評価，話者適応の適用などが今後の課題として挙げられる．本稿では触れなかったが，中継や電話音声などのニュース音声特有の問題への対処が必要であると考えられる．

謝辞

NHK ニュース原稿と音声データを提供していただいた日本放送協会，日本経済新聞 CD-ROM 版 1990 年版～1994 年度版の使用を許諾していただいた日本経済新聞社，音響モデルの構築に尽力していただいた山形大の堀貴明氏に感謝致します．

参考文献

[1] 大附, 森, 松岡, 古井, 白井, "新聞記事を用いた大語彙連続音声認識の検討," 信学技報, SP95-90, 1995-12

[2] 吉田, 松岡, 大附, 古井, "単語 trigram を用いた大語彙連続音声認識," 信学技報, SP96-82, 1996-12

[3] Francis Kubala,Tasos Anastasakos,Hubert Jin, "Toward Automatic Recognition of Broadcast News," 1996 DARPA Speech Recognition Workshop

[4] 安藤, 宮坂, "ニュース音声データベースの構築," 音講論, 2-Q-9, 1997-3