音声言語情報処理　9－10
(1995. 12. 14)

新聞記事を用いた大語彙連続音声認識の検討

大附克年＊　森岳至＊＊　松岡達雄＊＊＊　古井貞熙＊＊＊＊　白井克彦＊

＊早稲田大学理工学部電気工学科
〒169 東京都新宿区大久保3-4-1
＊＊東京工業大学情報理工学研究科
〒152 東京都目黒区大岡山2-12-1
＊＊＊NTTヒューマンインターフェース研究所
〒180 東京都武蔵野市緑町3-9-11

あらまし　不特定話者大語彙連続音声認識の研究のための新聞記事読み上げによる大語彙音声データベースの構築を行った．約5年分の新聞記事680万文中から出現した60万語のうち上位7,000語，30,000語，150,000語の語彙数を設定して読み上げ文を抽出し，54名話者による5,400発声を収録した．また，文脈依存音素HMMと単語バイグラム文法を用いた大語彙音声認識システムを構築し，収録した音声データベースの7,000語彙のセットによって評価を行った．

キーワード　連続音声認識，大語彙，音声データベース

Study of Large-Vocabulary Continuous Speech Recognition Using Read-Speech Corpus

Katsutoshi Ohtsuki*, Takeshi Mori**, Tatsuo Matsuoka***,
Sadaoki Furui** ***, and Katsuhiko Shirai*

*Department of Electrical Engineering, Waseda University
3-4-1 Okubo, Shinjiku-ku, Tokyo, 169
**Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2-12-1 Ookayama Meguro-ku, Tokyo 152
***NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180

Abstract　We recorded Japanese read-speech corpus with text obtained from newspapers for studies of large vocabulary speaker-independent continuous speech recognition. A word-frequency list (WFL) was formed from 6.8M sentences from about 5 years of newspaper articles and this yielded a list of 600K words. We selected three sets of recorded texts according to vocabulary sizes: top 7K, 30K, and 150K words of the WFL. We recorded 5,400 utterances by 54 speakers. We also developed a large-vocabulary continuous speech recognition system using context-dependent phoneme HMMs and a bigram grammar and we evaluated the system using this corpus.

Keywords　Continuous Speech Recognition, Large Vocabulary, Speech Corpus

1. はじめに

近年の音声情報処理，言語処理および計算機処理技術の著しい進歩により，音声認識の分野において，これまでに比べてより自然な発話やより大きな語彙，より複雑な雑音環境などに対しても研究の目標が向けられるようになってきた．特に，大語彙の連続音声認識に関しては，ARPA の Wall Street Journal (WSJ) コーパス [1] を用いて，各研究機関において研究が進められ，着実に成果があげられている [5]．さらに最近では，アメリカ英語だけでなくイギリス英語やフランス語，ドイツ語に対しても，WSJ タスクで培われたノウハウを生かして，新聞読み上げコーパスを対象とした大語彙の連続音声認識の研究がなされている [6,7]．しかしながら，これまで大語彙連続音声認識のために構築されてきたデータベース [1-4] に日本語によるものはなく，日本語による大語彙連続音声認識の研究のためには，まず，日本語による大語彙の連続音声データベースの整備を進めることが不可欠である．

そこで，我々は5年分の新聞記事を用いて，日本語の連続音声認識の研究のための音声データベースの設計について検討した．まず，学習用の新聞記事テキストデータ 680 万文に出現した 60 万語のうち出現頻度が上位 7,000 語，30,000 語，150,000 語の語彙数を設定してテキストを選択し，54 名の話者による各 100 発話，計 5,400 発話の収録を行った．このデータベースを用いて，連続音声認識システム構築のための音響モデル，言語モデルについて検討を進めており，多数話者の発声データにより 2,000 種類以上の文脈依存音素 HMM の学習を行い，新聞記事テキストによりNグラム言語モデルの推定を行っている．

本稿では，数千語以上の語彙からなる連続音声データベースの設計と，文脈依存音素HMMと単語バイグラム文法を用いた連続音声認識システムの構築とその評価について述べる．

2. 連続音声データベース

2.1 データベースの設計

2.1.1 テキストデータベース

データベースは，不特定話者による新聞記事読み上げ音声により構成される．これまで新聞記事を用いた読み上げ音声のコーパスは，英語 (Wall Street Journal) [1,3] やフランス語 (Le Monde) [2]，ドイツ語 (Frankfurter Rundschau) [4] で構築されてきた．本研究では，1990 年から 1994 年までの日本経済新聞の記事 5 年分 [12]，約 750 万文を用いる．5 年分の記事のうち，1990 年 1 月から 1994 年 9 月までの 57ヵ月分 (95%) を単語頻度リスト構築および言語モデル学習用のデータとして用い，1994 年 10 月から 12 月までの 3ヵ月分 (5%) を評価用のデータとして用いる．

2.1.2 形態素解析

日本語の文章は，英語やフランス語などのように単語間に空白を持たないため，単語の境界が明確ではない．テキストデータベースの単語出現頻度の計算や言語モデルの学習を行うためには，連続した文字列を単語ごとに区切ることが必要である．また，認識を駆動する単位である単語の定義にも様々なものが考えられる．今回，我々は，語彙の単位として形態素を採用し，学習用のデータに対して形態素解析を行った．用いた形態素解析ツールの新聞記事に対する解析正解率は約 95% である．

形態素解析を行う前に，読み易さや統計的言語モデルの学習を考慮して，いくつかの前処理をテキストに対して施した．以下に，その例をいくつか示す．

テキスト前処理の例：

(1) ○，●，◎，？，！などは，文書における強調記号であり発音できないので，削除する．

ex.) ◎ＥＣが経済通貨同盟で政府間会議．

(2)「」，『』，“”，‘’ などは，括弧の中身が文を構成しているので，括弧のみ削除する．

ex.) ノイマン型コンピューターの情報は「０」か「１」かで表現される．

(3) ()，【】，〔〕，《》，〈〉 などは，語句の説明や見出しなどに使われており，削除しても文の構造は保たれるので，括弧・中身とも削除する．

ex.) やわらかな日差しがそそぐアトリウム（吹き抜け空間）．

前処理を施したテキストの形態素解析の結果より，1 文に含まれる形態素数の分布を正規分布とみなし，1 文当たりの形態素数が平均値±2 σ (95.45%：1～53 形態素)に含まれる文のみを単語頻度リストの構築および言語モデルの学習に用いた．その結果，学習用データの規模は，約 680 万文，約 1 億 8000 万形態素となった．

以下，「単語」は形態素を表すものとする．

2.1.3 読み上げ文選択

読み上げテキストの語彙サイズ設定のために，まず単語頻度リストを作成した．これは，学習データ中に出現した単語を頻度順に並べたものである．その結果，約 60 万語からなる単語リストが得られた．

続いて,長すぎる文は読み上げが困難であることから，1文当たりの単語数が前述の分布の平均値±σ (68.3%：3～39単語) に含まれない文は除外した．また，テキストの誤植や形態素解析のエラーは，低い頻度で現われることが考えられるため，単語頻度リストの上位150,000語（カバー率99.6%）に含まれない単語を含む文も除外した．

単語頻度リストに基づいて，いくつかの語彙サイズの設定を行う．今回は，7,000語と30,000語の2つのサイズを設定した．表1に，各語彙サイズの全単語に対するカバー率とWSJコーパス [1] との比較を示す．

表1: 語彙サイズとカバー率

日経		WSJ	
Size	Coverage[%]	Size	Coverage[%]
7K	90.3	5K	91.7
30K	97.5	20K	97.8
150K	99.6	64K	99.6
623K	100.0	173K	100.0

語彙サイズおよび文中に含まれる未知語の数から5種類のサブセットを学習用,評価用それぞれのデータに対して合計10種類定義した．以下に各サブセットの定義を示す．

7K： 単語頻度リストの上位7,000語のみにより構成される文

7K+： 単語頻度リストの上位7,000語および1ないし2語の未知語により構成される文

30K： 単語頻度リストの上位30,000語のみにより構成される文

30K+： 単語頻度リストの上位30,000語および1ないし2語の未知語により構成される文

30K++： 単語頻度リストの上位30,000語および3語以上の未知語により構成される文

上述のように上位150,000語に含まれない単語は除外されているので，ここで未知語とは，7K+の場合は，上位150,000語に含まれかつ上位7,000語に含まれない語，30K+，30K++の場合は，上位150,000語に含まれかつ上位30,000語に含まれない語となる．つまりすべての読み上げ文は150,000語の語彙で閉じていることになる．

各サブセットに含まれる文の割合を表2に示す．表中の最下段 (30K, 30K+, 30K++) は，語彙サイズ150Kで表現できる文の割合を表している．

以上のように得られた各サブセットから，読み上げが困難な地名，人名，法人名を含む文や非文などを除外したものを読み上げテキストとして採用した．なお，読み上げの際には，句読点は読まないものとした．

表2: 各サブセットが表現できる文の割合[%]

サブセット	train(100days)	test(92days)
7K	10.7	11.1
7K, 7K+	40.8	41.8
30K	42.0	42.9
30K, 30K+	62.0	62.2
30K, 30K+, 30K++	64.2	63.9

2.2 音声収録

音声収録は，比較的静かな実験室環境で行った．すべての発声は，2本のマイク (head-mounted Senheiser HMD-410/desk-mounted Crown PCC-160) を用いて，DAT (SONY 77ES) に2チャンネル録音した．

各発声者は，前述の10のサブセットから各10文ずつ抽出した計100文を発声した．100文の収録時間は一人当り1時間程度であった．発声者は，大学生，大学院生を中心に集められ，男性51名，女性3名について収録を終えている．各サブセットの1文当たりの平均単語数および平均継続時間を表3に示す．

表3: 1文当たりの平均単語数と平均継続時間

サブセット	単語数	継続時間[sec]
7K	20.9	6.3
7K+	23.3	7.0
30K	23.4	7.1
30K+	25.7	7.9
30K++	27.4	8.6

3. 連続音声認識システム

3.1 音響モデル

大語彙の連続音声認識において，単語 (whole-word) を単位としてモデルを構築することは，学習データの量，学習効率の観点から現実的ではない．一般的に大語彙を扱う場合，より小さい単位 (sub-word) の音響モデルを連結して単語モデルとして扱うことが多い．今回我々は音素を単位とする音響モデルを構築した．

3.1.1 文脈独立音素HMM

多数話者による大量の発声データを用いて文脈(音素環境)に独立な音素HMMを学習し，評価を行った．まず，視察ラベルによりセグメンテーションされたデータ（ATR B）を用いて初期モデルを学習し，その

後，大量のデータ（ATR B，ASJ連続音声，ASJ対話）を用いて連結学習を行った．音素カテゴリは無音を含む42種類（表4），音素HMMは5状態3ループ4混合（無音および促音の閉鎖部は3状態1ループ）とした．学習および評価に用いた音声試料を表5に，分析条件を表6に示す．なお，HMMの学習および評価にはHTK [13] を用いた．連結学習において繰り返し学習4回で尤度がほぼ飽和し，そのモデルを評価および文脈依存HMMの初期モデルに用いた．

表4: 音素カテゴリ

母音	a, e, i, o, u aa, ee, ei, ii, oo, ou, uu
子音	b, d, g, p, t, k ch, j, sh, ts, f, h, s, z N, m, n, r, w, y by, gy, hy, ky, my, ny, py, ry
促音閉鎖部	Q
無音	#

3.1.2 文脈依存音素HMM

単語中の発声変動に対処するために，学習データ中における音素連鎖の出現頻度により，数種類の文脈（音素環境）依存HMMのセットを用意し，それぞれについて音素認識による評価を行った．文脈依存HMMは，先行音素のみに依存するbiphoneと先行および後続音素に依存するtriphoneの2種類を用意した．学習は前節で述べた文脈独立HMMを初期モデルとして連結学習（繰り返し回数：4回）を行った．

表5: 音声試料

初期モデル学習用データ	ATR Bセット (5名話者，2515発話)
連結学習用データ	53名話者，13270発話 ATR Bセット (5名話者，2515発話) ASJ連続音声DB (30名話者，4518発話) ASJ対話DB (18名話者，6327発話)
評価用データ	ATR Bセット(closed) (5名話者，250発話) ATR Cセット(open1) (5名話者，250発話) 日本経済新聞(open2) (5名話者，500発話)

表6:分析条件

A/D変換	12kHz, 16bit量子化
フレーム長	32ms（ハミング窓）
分析周期	8ms
音響特徴量	LPCケプストラム[1-16] Δケプストラム[1-16]

文脈独立HMM (CI) および各文脈依存HMMのセットのモデル数と音素認識正解率 (%Correct) と正解精度 (Accuracy) を図1に示す．正解精度は，正解率と挿入誤り率との差をとったものである．図中の

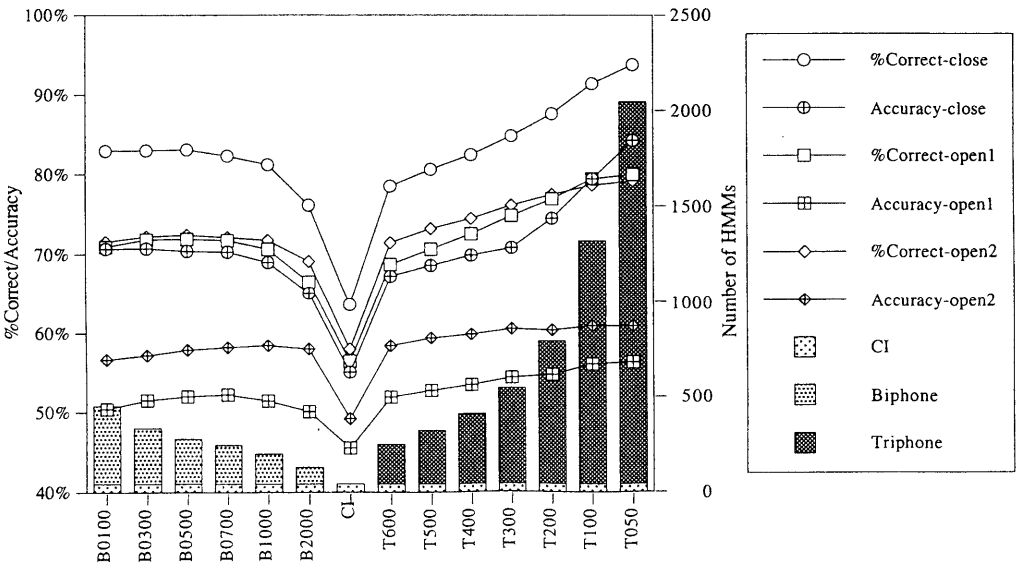

図1: 音素認識実験結果

B1000，T500などの数値は，文脈依存HMMの種類(B：biphone, T：triphone) とその出現頻度を表している．図1をみると，biphoneの場合はモデル数が200を超えた辺りから性能が上がらなくなるのに対して，triphoneではモデル数が2,000に達するまで性能が上がり続けていることがわかる．さらにbiphoneモデルとtriphoneモデルを同時に用いて最大尤度を示したものを認識結果とみなした場合の音素認識率および正解精度を表7に示す．B700とT300とを併せて用いた場合（モデル数：748）に正解精度が最も高くなっている．

表7: 文脈依存HMMの評価(open2)

	B1000	B700	B500
T500	76.1(60.9)	76.2(61.0)	76.5(60.9)
T400	76.6(60.9)	76.8(61.2)	77.1(61.1)
T300	77.3(61.2)	77.5(61.6)	77.8(61.5)
T200	77.9(60.6)	78.1(60.9)	78.4(60.9)

3.2 言語モデル

比較的小さな語彙の連続単語認識では，主として音響的特徴を利用して認識が行われるが，大語彙の連続音声（文）認識を行うためには，文法的，意味的な言語情報を利用することが必要である．言語モデルには大きく分けて，文脈自由文法 (Context Free Grammar : CFG) などの構文規則によるものと，単語Nグラムなどの統計的モデルによるものとがある．今回のように数千語以上の大語彙で，ある程度の学習データが得られるタスクの場合には，統計的な言語モデルの方が導入が容易でかつ効果が大きいと考えられる．

3.2.1 統計的言語モデル

統計的言語モデル$P(W)$は単語列Wの生起確率を表し，式 (2) のように定義される．

$$W = w_1 w_2 ... w_k \tag{1}$$

$$P(W) = \prod_{i=1}^{k} P\left(w_i | w_1 w_2 ... w_{i-1}\right) \tag{2}$$

しかし，ここですべての単語のすべての単語列の長さに対して$P\left(w_i | w_1 ... w_{i-1}\right)$を推定するのに十分な統計量を得ることは不可能なので，式 (3) のような近似が用いられる．

$$P\left(w_i | w_1 w_2 ... w_{i-1}\right) \approx P\left(w_i | w_{i-N+1} ... w_{i-1}\right) \tag{3}$$

今回我々は，2.1.2で述べた680万文を用いて，N=2（バイグラム）の場合とN=3（トライグラム）の場合とについて言語モデルの推定を行った．

3.2.2 パープレキシティ

連続音声認識の複雑度を表す尺度として，単語系列についてのエントロピーをHとするときの2^Hで定義されるパープレキシティがある．また，言語モデルを導入したときのパープレキシティの減少から言語モデルの性能をはかることができる．表8に今回のタスクおよびWSJタスクのテストセットパープレキシティ [8] を示す．ここで，言語モデルの推定における統計量の不足を補うために，言語モデルに対してback-off平滑化 [9] を行った．

表8: パープレキシティ

日経†			WSJ			
Vocab	LM	Perp	Vocab	LM	Perp	
					VP	NVP
7K	BG	74	5K	BG	80	118
	TG	45		TG	44	68
30K	BG	89	20K	BG	158	236
	TG	65		TG	101	155

BG: bigram; TG: trigram
VP:verbalized punctuation; NVP: non-VP
†各語彙サイズについて個々に言語モデルを推定した

4. 連続音声認識実験

今回収録したデータのうち，語彙サイズ7,000，未知語無し (7K) のデータ5人分100文章（学習用，評価用各50文章）を用いて連続音声認識実験を行った．音声データは，接話型マイク (head-mounted Senheiser HMD-410) で収録されたものを用いた．音響モデルは文脈独立HMM (CI) および単語内文脈依存HMM (CD) を用い，言語モデルは単語バイグラムモデルを用いた．文脈依存HMMによる実験では，音素認識実験で最も性能の良かったモデル数748個のセットを用いた．実験結果を表9に示す．単語誤り率 (%WordError) および単語正解率 (%WordCorrect) の計算は次式のように行った [10]．Sは置換誤り単語数，Dは脱落誤り単語数，Iは挿入誤り単語数，Nは認識対象テキストの総単語数である．

$$\%\text{WordError} = \frac{S+D+I}{N} \times 100 \tag{4}$$

$$\%\text{WordCorrect} = \left[1 - \frac{S+D}{N}\right] \times 100 \tag{5}$$

単語バイグラムを用いることにより，単語誤り率が半分以下に減少している．また，音響モデルに文脈依存HMMを用いることで，単語正解率が文法無し (NG) の場合で約6%，単語バイグラム (BG) の場合で約3%改善されている．

表9: 実験結果

AM	LM	training set		test set	
		%Correct	%Error	%Correct	%Error
CI	NG	28.0	73.3	26.6	73.8
CD	NG	34.2	67.1	33.0	67.7
CI	BG	74.6	31.7	74.3	30.9
CD	BG	77.4	31.3	77.8	29.6

NG: No Grammar

5. まとめ

本稿では,大語彙連続音声認識の研究のための新聞記事読み上げによる音声データベースの設計,連続音声認識システムの構築および評価について報告した.

音声データベースの設計においては,5年分の新聞記事から,単語の出現頻度を基準にして文章を抽出し,54名話者による5,400発話を収録した.現在,10名分1,000発話について,A/D変換,発話内容記述の確認が終了している.今後はより大規模な実験が行えるよう,残りのデータの整備を進めていく.

また,文脈依存音韻HMMと単語バイグラム文法とを用いた連続音声認識システムを構築し,7,000語彙のデータにより評価を行った.その結果,単語バイグラムを用いることにより単語誤り率が50%以上改善され,さらに文脈依存HMMを用いることでオープンテストで単語誤り率29.6%という結果が得られた.

今回,文脈依存HMMを単語内に導入することにより,単語誤り率が改善されているが,単語間には適用されないため,単語内と単語間の両方の文脈によって学習された文脈依存HMMの効果が十分現われているとはいえない.そこで今後は,文献[11]の手法により単語間にも文脈依存HMMを導入した実験を行っていく.また,今回用いなかったパワー,Δパワー,ΔΔケプストラムなどの音響特徴量の利用や音韻規則の導入についても検討する必要がある.言語モデルについては,バイグラム,トライグラムの推定において,単語の品詞クラスを用いることによりモデルの推定精度を向上することを検討している.

謝辞 形態素解析ツールを提供していただいたNTTヒューマンインターフェース研究所映像処理研究部の田中一男主幹研究員に感謝します.テキストデータの使用を許諾していただいた日本経済新聞社に感謝します.音声データベースの収録にあたり,読み上げ文発声に御協力いただいた皆様に感謝します.また,日頃御討論いただくNTTヒューマンインターフェース研究所古井特別研究室,早大白井研究室,東工大古井研究室の皆様に感謝します.

参考文献

[1] D. B. Paul, and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," Proc. ICSLP'92, pp. 899-902.

[2] J. Gauvain, L. F. Lamel, and M. Eskenazi, "Design Considerations and Text Selection for BREF, a large French read-speech corpus," Proc. ICSLP'90, pp. 1097-1100.

[3] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0 : A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," Proc. ICASSP'95, pp. 81-84.

[4] H. J. M. Steeneken and D. A. van Leeuwen, "Multi-Lingual Assessment of Speaker Independent Large Vocabulary Speech-Recognition Systems : The SQUALE-Project," Proc. EUROSPEECH'95, pp. 1271-1274.

[5] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. J. Young, "The 1994 HTK Large Vocabulary Speech Recognition System," Proc. ICASSP'95, pp. 73-76.

[6] D. Pye, P. C. Woodland, and S. J. Young, "Large Vocabulary Multilingual Speech Recognition using HTK," Proc. EUROSPEECH'95, pp. 181-184.

[7] L. Lamel, M. Adda-Decher, and J. L. Gauvain, "Issues in Large Vocabulary, Multilingual Speech Recognition," Proc. EUROSPEECH'95, pp.185-188.

[8] D. B. Paul, and B. F. Necioglu, "The Lincoln Large-Vocabulary Stack-Decoder HMM CSR," Proc. ICASSP'93, pp. 660-663.

[9] S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," Trans. ASSP-35, pp. 400-401, March 1987.

[10] F. Kubala, et al., "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-Word Resource Management Database," Proc. ICASSP'88, pp. 291-294.

[11] W. Chou, T. Matsuoka, B.-H. Juang, and C.-H. Lee, "An Algorithm of High Resolution and Efficient Multiple String Hypothesization for Continuous Speech Recognition Using Inter-Word Models," Proc. ICASSP'94-II, pp. 153-156.

[12] "日本経済新聞 CD-ROM 版 1990 年版～1994 年版," 日本経済新聞社, 1994-1995.

[13] Cambridge University Engineering Department Speech Group and Entropic Research Laboratory Inc., "HTK-Hidden Markov Model Toolkit V1.5," 1993.