論文 / 著書情報
Article / Book Information

論題(和文)	新聞記事読み上げタスクを用いた大語彙連続音声認識における音響モデルの検討
Title(English)	
著者(和文)	大附克年, 森岳至, 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 1996年春季講演論文集, Vol. , No. 3-8-8, pp. 161-162
Citation(English)	, Vol. , No. 3-8-8, pp. 161-162
発行日 / Pub. date	1996, 3

3—8—8

新聞記事読み上げタスクを用いた 大語彙連続音声認識における音響モデルの検討 *

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

1. はじめに

大語彙連続音声認識の研究を行うためには，音響モデルおよび言語モデルの学習のために大量の音声データとテキストデータが必要である．しかしながら，これまで日本語の大語彙連続音声認識の研究に用いることのできるようなデータベースは構築されていなかった [4]．我々は5年分の新聞記事から連続音声データベースの構築を行い，音素HMMと統計的言語モデルを用いた日本語の大語彙連続音声認識について検討している [1]．本稿では，大語彙連続音声データベースの設計について，また我々が構築している大語彙連続音声認識システムの音響モデルについて述べる．

2. データベースの設計

2.1 テキストデータベース

日本語の文章は単語間に空白をもたないため単語境界が不明確であり，また単語の定義にも様々なものがある．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．本研究では語彙の単位として形態素を採用し，テキスト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．テキストデータとして日本経済新聞の記事5年分 [2] を用いた．5年分の記事のうち，95%にあたる57ヵ月分を言語モデル学習用 (training set) とし，残りを評価用 (test set) とした．形態素解析を行う前に，統計的言語モデルの学習を考慮して，記号や括弧などを取り除く前処理をテキストに対して施した．

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文あたりの形態素数が平均値±2σ (95.5%: 1-53形態素) に含まれる文のみを学習用データとした．その結果学習用データの規模は，約680万文，約1億8千万形態素となった．

学習データ中に出現した単語を頻度順に並べた単語頻度リストに基づいて，いくつかの語彙サイズを設定した．今回，ARPAのWSJタスク [5] のカバー率を参考にして，7千語，3万語，15万語の3つの語彙サイズを設定した．表1に，各語彙サイズの全単語に対するカバー率とWSJタスクとの比較を示す．

2.2 連続音声データベース

前節で設定した語彙サイズおよび文中に含まれる未知語の数に関して5種類のサブセットを学習用，評価用それぞれに対して定義した（合計10種類）．各サブセットの定義を表2に示す．ここで未知語は，語彙7千語または3万語に含まれず15万語に含まれる単語である．

表1: 語彙サイズとカバー率

語彙サイズ	カバー率[%]	
	日経	WSJ
5K	88.0	91.7
7K	90.3	-
20K	96.2	97.7
30K	97.5	-
64K	98.9	99.6
150K	99.6	-
173K	99.7	100.0
623K	100.0	-

音声収録は比較的静かな実験室環境で行われ，各発声者は10種類のサブセットから10文ずつを無作為に抽出した合計100文を発声した．マイクは接話マイク (Senheiser HMD-410) とバウンダリマイク (Crown PCC-160) の2本を用いた．現在54名について収録を終えている．

3. 音響モデル

3.1 文脈独立音素HMM

音素HMMの学習は，まずラベル付きのデータを用いて初期モデルを学習し，その後連結学習を行った [3]．音素カテゴリは42種類（無音含む），音素HMMの構造は5状態3ループ4混合とした．音声試料を表3に示す．サンプリング周波数12kHz，フレーム長32ms（ハミング窓），分析周期8msとし，音響特徴量には16次元のLPCケプストラムおよびΔケプストラムを用いた．

3.2 文脈依存音素HMM

単語間および単語内における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，文脈（音素環境）に依存したHMMのセットを数種類用意し，それぞれについて評価を行った．先行音素のみに依存するdiphoneのセットを6種類，先行および後続音素に依存するtriphoneのセットを7種類用意した．

表2: 各サブセットの定義

サブセット	定義
7K	7千語語彙のみにより構成される文
7K+	7千語語彙に未知語を2語まで含む文
30K	3万語語彙のみにより構成される文
30K+	3万語語彙に未知語を2語まで含む文
30K++	3万語語彙に未知語を3語以上含む文

* A Study of Acoustic Modeling for Large-Vocabulary Continuous Speech Recognition Using A Read-Speech Corpus.
By Katsutoshi Ohtsuki[1], Takeshi Mori[2], Tatsuo Matsuoka[3], Sadaoki Furui[2,3] and Katsuhiko Shirai[1]
([1]Waseda University, [2]Tokyo Institute of Technology, [3]NTT Human Interface Laboratories)

表3: 音声試料

初期モデル学習用	ATR B (5名, 2515発話)
連結学習用	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 (53名, 13270発話)
音素認識評価用	日本経済新聞 (5名, 500発話)
連続音声認識評価用	日本経済新聞 (10名, 200発話)

3.3 音素認識実験

音素HMMの性能を評価するために音素認識実験を行った．認識単位が音素である構文ネットワークを用いて連続音声（音素）認識を行い，認識結果と正解音素列とのDPマッチングを行うことにより正解率 (%Correct) および正解精度 (Accuracy) を算出した．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．音素認識結果を図1に示す．図中のD1000，T500などは文脈依存HMMの種類 (D: diphone, T: triphone) とその出現頻度に関するしきい値を表している．また文脈依存モデルを用いた実験においても常に文脈独立モデル (CI) を同時に用いている．D700とT300を併せて用いた場合（モデル数：748）に最も高い正解精度 (61.6%) が得られた．

3.4 パワー情報の利用

前述の32次元の音響特徴量に，正規化対数パワーおよびその変化率（Δパワー）の2次元の特徴量を加えたHMMの学習を行った．音素認識実験による評価を行ったところ，誤り率 (100% - 正解精度) が約10%改善された．

4. 大語彙連続音声認識

今回収録したデータのうち，接話マイクで収録された7Kセット10名分200発話 (training set, test setから各100テキスト) を用いて連続音声認識実験を行った．言語モデルは単語バイグラムを用いた [6]．文脈依存HMMは音素認識実験において最も性能の高かったセットを用い，文脈依存性は単語内でのみ考慮した．実験結果を表4に示す．文脈依存HMM (CD) を用いることでCIモデルのみを用いたときと比べて誤り率が，文法無し (NG) の場合で約6%，単語バイグラム (BG) の場合で約24%改善されている．さらに，パワー情報を用いることによってもCIモデルの場合で20%以上，CDモデルでは25%以上の誤り率改善がみられた．

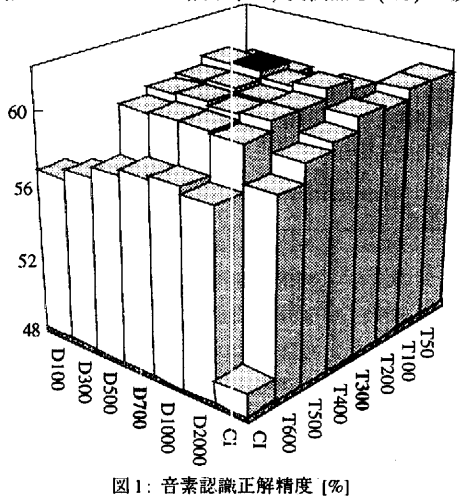

図1: 音素認識正解精度 [%]

表4: 連続音声認識実験結果

音響モデル		言語	training set		test set	
HMM	特徴量	モデル	%Correct	Accuracy	%Correct	Accuracy
CI	cepstrum Δcepstrum	NG	19.6	18.0	18.1	17.2
CD		NG	24.5	23.0	23.1	22.1
CI		BG	67.0	65.0	65.8	63.7
CD		BG	77.2	73.5	76.0	72.5
CI	+ power Δpower	BG	75.3	73.4	73.2	71.5
CD		BG	82.4	80.3	82.3	80.0

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，および連続音声認識システムにおける音響モデルの検討について報告した．

連続音声認識実験により音響モデルを評価し，文脈依存モデルおよびパワー情報を用いることにより単語誤り率が約45%改善され，その有効性が確認された．今回，音素認識において最も性能の高い文脈依存モデルのセットを用いて連続音声認識実験を行ったが，セットを変えた場合についても評価を行う必要がある．

さらに，単語間への文脈依存HMMの適用，助詞などの頻度の高い単語に対する単語依存HMMの導入，ΔΔケプストラムの利用，音韻規則の導入などが今後の課題としてあげられる．

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．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] 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.

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

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

[4] L. F. Lamel, and R. De Mori, "Speech Recognition of European Language," IEEE ASR Workshop, 1995-12.

[5] D. B. Paul, and J. M. Baker, "The Design for the Wall Street Journal-Based CSR Corpus, " Proc. ICSLP'92.

[6] 森，大附，松岡，古井，白井，" 新聞記事読み上げタスクを用いた大語彙連続音声認識における言語モデルの検討，" 音講論，1996-3.