6C－10

新聞・ニュース文をタスクとした大語彙連続音声認識システムの評価

赤松裕隆　花井建豪　甲斐充彦　峯松信明　中川聖一
豊橋技術科学大学 情報工学系

1　はじめに

本研究では大規模音声・テキストコーパスを利用した大語彙連続音声認識において効果的な音響モデルと言語モデルの構築を検討してきた。

まず、世界的に標準となっている triphone による音響モデルと音節単位による音響モデルを比較検討した。次に、新聞記事の読み上げ文の音声データ、および朗読音声と対照的な自然発話である NHK ニュース音声データに対して、大語彙連続音声認識実験で評価し、比較検討した。

2　認識システムの構成

本音声認識システムは 2 パス方式になっている。1 パス目で言語モデルに bigram を用いて認識を行ない、スコアの高い上位 N 個 (N-best) の候補を出力する。そして 2 パス目で言語モデルに trigram を用いて N-best の候補のリスコアリングを行なう。

連続音声認識のアルゴリズムには、Viterbi アルゴリズムに基づく One Pass サーチ法 を用いた。これは、各フレーム毎に各単語境界と仮定し、言語モデルによる確率の対数値をマッチング終了後の音響累積尤度に加えることを繰り返すことによって最尤の単語列候補を計算する [1]。

3　音響モデル

triphone による音響モデル (HTK により作成) と音節単位による音響モデルを比較検討するために、表 1の音響モデルを表 2に示す分析条件で構築し、連続音節認識実験を行なった [2]。毎日新聞読み上げ文に対する結果を表 3に示す。triphon モデルでは、音節に対応する音韻間の接続の制約を課した (なお、CMN とはケプストラム平均正規化を意味する)。

正解率は triphon のフレーム単位と音節のセグメント単位のものとではほとんど変わらなかったが、triphone モデルは挿入誤りが多いため音節モデルの方が正解精度はフレーム単位の音節モデル、セグメント単位の音節モデルの両方で triphon モデルよりもかなり良い結果となった。

次に朗読音声と自然発話 (spontaneous speech) の認識の困難さを調べるために、毎日新聞と NHK ニュースのタスクで連続音節認識実験を行なった。その結果を表 4に示す。この結果より NHK ニュース音声タスクが音響的に難しいタスクであることが分かる。今後は大量の自然発話を用いた音響モデルの学習など、NHK ニュース音声タスクに適した音響モデルの作成が必要であることが分かる。

表 1: 音響モデル

triphon モデル
5 状態 3 出力分布
連続出力分布型 HMM
16 混合ガウス分布, 対角共分散行列, モデル数 7921
音節モデル
5 状態 4 出力分布
離散継続時間分布付き連続出力分布型 HMM
4 混合ガウス分布, 全共分散行列, モデル数 114

表 2: 音声の分析条件

サンプリング周波数	12kHz
窓関数	21.33ms ハミング窓
フレーム周期	8ms
分析	14 次元 LPC 分析
学習データ	
	ASJ ATR503 文 A～J セットの 6 名の男性話者と 216 単語の音声データから初期モデルを作成
	ASJ ATR503 文 A～J セットの 30 名の男性話者と JNAS 新聞記事文 125 名の男性話者を MAP 推定で追加学習 (総発話数 17221 文)
特徴パラメータ	
	LPC メルケプストラム (フレーム単位:10 次元, セグメント単位:10 次元 × 4 フレームの特徴量を KL 展開で 20 次元に圧縮)
	+ Δ ケプストラム (10 次元)
	+ ΔΔ ケプストラム (10 次元)
	+ Δ パワー + ΔΔ パワー

表 3: 音響モデルの比較 (音節認識率 %)
(話者 9 名, CMN なし, 言語モデルなし)

音響モデル	Cor.	Acc.	Sub.	Ins.	Del.	Seg.
triphon(フレーム単位)	79.9	64.2	17.7	15.7	2.4	82.0
音節 (フレーム単位)	75.7	67.3	21.4	8.3	3.0	88.7
音節 (セグメント単位)	78.3	70.1	19.1	8.2	2.7	89.1

Cor.(正解率) = 100 − Sub.(置換率) − Del.(脱落率)
Acc.(正解精度) = Cor.(正解率) − Ins.(挿入率)

表 4: 連続音節認識実験結果 (%)
(話者 9 名, CMN あり, 言語モデルなし)

コーパス	Cor.	Acc.	Sub.	Ins.	Del.
毎日新聞読み上げ文	79.1	71.8	18.1	7.3	2.8
NHK(アナウンサ)	64.5	52.8	28.2	11.7	7.3
NHK(レポータ)	57.9	41.8	33.4	16.1	8.7
NHK(全体)	63.8	51.6	28.8	12.1	7.5

Evaluation of a large-vocabulary continuous speech recognition system for newspaper and broadcast news
Hirotaka Akamatsu, Kengo Hanai, Atsuhiko Kai, Nobuaki Minematsu and Seiichi Nakagawa
Toyohashi University of Technology

4 N-gram 言語モデル

本研究では、単語 bigram と単語 trigram を大語彙音声認識システムの言語モデルとして採用する。N-gram 言語モデルは CMU SLM Toolkit を用いて作成している。

言語モデルの評価にはパープレキシティと補正パープレキシティを用いた。それぞれ、式 (1), 式 (2) に示す。ここで、n は評価テキストの総単語数、n_u は未知語の出現回数、m は未知語の種類数である。

$$PP = P(w_1 \cdots w_n)^{-\frac{1}{n}} \quad (1)$$

$$APP = (P(w_1 \cdots w_n)m^{-n_u})^{-\frac{1}{n}} \quad (2)$$

パープレキシティでは未知語は全て一つの未知語のカテゴリにまとめられ、語彙に含まれる形態素と等価に未知語のカテゴリは扱われる。補正パープレキシティでは複数の未知語は別個の単語として扱い、それぞれ等確率に生じると仮定したものである。

本研究では毎日新聞読み上げ音声コーパスの評価用の言語モデルの構築のために、4 年間分の毎日新聞記事データベース (1991 ～ 1994) を使用した。形態素解析には RWC の形態素解析結果を使用した。また、NHK ニュース音声コーパスの評価のために NHK ニュース原稿データベースを用いた言語モデルを構築した。このデータベースは 1991 年 4 月から 1996 年 6 月までの 5 年間のニュース原稿で構成されている。形態素解析には京大で作成された JUMAN を使用した。表 5に実験データを、表 6に 5000 語彙に対するこれらの言語モデルの評価結果を示す。

表 5: 実験データ

毎日	
train	学習セット (1991.1~1994.9, 約 330 万文)
test1	評価セット (1994.10~1994.12, 約 23 万文)
test2	音声認識評価文 (100 文)
NHK	
train	学習セット (1991.4~1996.5, 約 47 万文)
test1	評価セット (1994.6, 約 2 万文)
test2	音声認識評価文 (70 文)

表 6: 言語モデルの評価結果

		毎日新聞			NHK		
データセット		train	test1	test2	train	test1	test2
bigram	PP	67	71	94	58	60	100
	APP	280	275	94	246	247	100
trigram	PP	39	47	61	25	32	58
	APP	164	184	61	77	84	58

5 評価実験

認識文として、毎日新聞読み上げ音声コーパスの 10 名の男性話者の初めの各 10 文の総計 100 文 (すべて異なる文) と、NHK ニュース音声の語彙で閉じているノイズの比較的少ない男性話者 (不特定話者) の文の中から、5 形態素から 30 形態素で形成される 70 文の評価用データを使用した。

使用した音響モデルはセグメント単位入力の音節 HMM で、状態毎の継続時間分布の継続時間長を新聞タスクでは 1/1.1 に、ニュースタスクでは 1/1.7 に縮めた。認識実験の結果を表 7,8に示す。

表 7の結果を見ると、trigram のリスコアリングにより認識率が bigram と比べ上昇していることが分かる。しかし、200best との差が大きく、改善の余地が残っていると考えられる。

ニュース音声の結果 (表 8) は毎日新聞の結果 (表 7) と比べて非常に悪い結果となっている。これは、新聞の音声データは朗読音声であるのに対して、ニュース文の音声データは自然発話であることから、音声認識が難しいことによる。また、全体的に紙の擦れる音や機械音などのノイズが含まれていること、間投詞 (16 個) や言い淀み (2 個) が含まれていることが挙げられる。

表 7: 新聞記事読み上げ文での単語認識率 (%)
(話者 10 名, CMN あり, 語彙サイズ 5000)

	Cor.	Acc.	Sub.	Ins.	Del.
bigram	85.8	82.8	11.0	3.0	3.2
200best	96.1	93.1	2.9	3.0	1.0
trigram	88.5	85.9	8.7	2.6	2.8

表 8: NHK ニュース音声タスクでの単語認識率 (%)
(CMN あり, 語彙サイズ 5005)

話者	言語モデル	Cor.	Acc.	Sub.	Ins.	Del.
アナウンサー (59 文)	bigram	73.0	66.5	22.3	6.5	4.8
	200best	82.6	77.2	13.3	5.3	4.1
	trigram	76.7	71.5	18.3	5.1	5.0
レポーター (11 文)	bigram	59.2	53.1	31.5	6.2	9.2
	200best	73.1	68.5	20.8	4.6	6.2
	trigram	63.1	57.7	27.7	5.4	9.2
Total (70 文)	bigram	71.4	65.0	23.3	6.4	5.3
	200best	81.5	76.3	14.2	5.3	4.3
	trigram	75.2	70.0	19.3	5.2	5.5

6 まとめ

新聞記事読み上げコーパスを用いて読み上げ音声に対する評価と、NHK ニュース音声タスクで自然発話に対する評価を行なった。

この実験で、我々が従来から用いている音節単位の音響モデルの有効性、および trigram による言語スコアのリスコアリングの有効性が明らかとなった。

また NHK ニュース音声タスクでの認識実験によって自然発話を用いた音響モデルの適応化の必要性が明らかとなった。

参考文献

[1] 赤松裕隆, 甲斐充彦, 中川聖一: 新聞・ニュース文の大語彙連続音声認識, 情報処理学会研究報告 98-NL-125-13,98-SLP-21-11,pp.115-122(1998-5)

[2] 花井建豪, 山本一公, 峯松信明, 中川聖一: セグメント単位入力 HMM のコンテキスト依存, 混合分布化による連続音声認識, 信学技報 SP98-29, pp.45-52(1998-6)