自然言語処理 125−13
音声言語情報処理 21−11
(1998. 5. 29)

新聞・ニュース文の大語彙連続音声認識

赤松裕隆† 甲斐充彦†† 中川聖一†

†豊橋技術科学大学 情報工学系
†† 豊橋技術科学大学 情報処理センター
E-mail: {akamatsu, kai, nakagawa}@slp.tutics.tut.ac.jp

あらまし

本研究では大規模なテキストと音声のコーパスを用い、音声認識のための精度の良い N-gram 言語モデルの構築を検討した。N-gram 言語モデルはタスクに依存するため、タスクに関する大量のデータベースを用いて構築される必要がある。そこで新聞記事テキストデータベースを用いて、同一ジャンルの過去の記事を用いるタスク適応化の方法とその有効性を調べた。また新聞記事には、使用頻度の高い (特殊) 表現や固定的な複数形態素から成る定型表現が多いため、それらを自動的に抽出し、1 形態素として捉えた上で N-gram 言語モデルを構築する方法を検討した。以上の言語モデルを朗読音声 (read speech) に適用し、評価した。

更にもう一つの利用可能な大規模なテキストコーパスとして、NHK ニュース原稿コーパスを用いた言語モデルを自然発話 (spontaneous speech) に適用し、比較評価した。

キーワード 音声認識, 大語彙連続音声認識, 言語モデル, N-gram, タスク適応化, 定型表現

Large-vocabulary speech recognition experiments using newspaper and broadcast news corpora

Hirotaka Akamatsu† Atsuhiko Kai†† Seiichi Nakagawa†

† Department of Information and Computer Sciences, †† Computer Center,
Toyohashi University of Technology

Abstract

In this paper, we describe a method that constructs language models using a task-adaptation strategy and idiomatic expressions of news articles. To build an effective n-gram based language model, it should be noted that the training data must be prepared as much as possible. However, for a given task/topic, it is very difficult to prepare a sufficient size of data. First, we investigated the effect of a task adaptation method of N-gram language model using a limited amount of target articles. Second, we investigated the effect of the use of idiomatic expressions as morpheme units, since some specific expressions and idiomatic expressions are frequently observed in news articles. Some experiments using news article read speech database were conducted to investigate the effectiveness of these methods for constructing N-gram language models. Experimental results using a broadcast news speech (spontaneous speech) and text corpora is also presented.

key words Speech Recognition, Large Vocabulary Continuous Speech Recognition, Language Model, N-gram, Task Adaptation, Idiomatic Expression

1 はじめに

近年の著しい計算機速度の向上、及び、音声処理技術/自然言語処理技術の向上により、音声ディクテーションシステムやパソコンで動作する連続音声認識のフリーソフトウェアの公開など、音声認識技術が実用的なアプリケーションとして社会に受け入れられる可能性がでてきた [1, 2]。我が国では、大量のテキストデータベースや音声データベースの未整備のため欧米と比べてディクテーションシステムの研究は遅れていたが、最近になって新聞テキストデータやその読み上げ文のデータが整備され [3], ようやく研究基盤が整った状況である。

このような背景を踏まえ、本研究では大規模音声・テキストコーパスを利用して音声認識において効果的な言語モデルの構築を検討してきた。まず言語モデルの構築に関して新聞記事テキストデータベースを用いて適応化について検討した [4]。

最近では付属語列を新たな認識単位とした場合、高頻度の付属語連鎖、関連率の高い複合名詞などを新しい認識単位とし、これらを語彙に追加する方法が試みられている [5, 6]。なお、連続する単語クラスを連結して一つの単語クラスとする方法や句を一つの単位とする方法は以前から試みられている [7, 8]。これと同じような効果を狙った方法として、N-gram の N を可変にする方法も試みられている [9]。

我々も同様に、新聞テキストやニュース文には、使用頻度の高い (特殊) 表現や、固定的な言い回しなどの表現 (以下、定型表現と呼ぶ) が非常に多く出現するため、定型表現を抽出し、これらの (複数形態素から成る) 定型表現を 1 形態素として捉え、N-gram 言語モデルを構築する方法を検討した [10]。

上述した 2 つの言語モデルの有効性を検討するために、新聞記事の読み上げ文 (朗読音声) を対象に大語彙連続音声認識実験で評価した。また、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も評価し、比較検討した。

なお新聞の朗読音声の認識実験は文献 [11, 12]、ニュース文の認識実験は文献 [13, 14] に報告されている。

2 認識システムの構成

2.1 N-gram 言語モデル

一般に、大語彙音声認識システムでは、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 言語モデルが用いられる。 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ているので、(1) 式のように定義される。

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

本研究では、単語 bigram を言語モデルとして採用する。N-gram 言語モデルは CMU SLM Toolkit[15] を用いて作成している。

言語モデルの評価基準の一つであるパープレキシティは次式で定義される。

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

この CMU SLM toolkit では語彙に含まれないもの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。そのため語彙サイズのセットが小さい程 (カバー率が小さい程)、パープレキシティは小さくなるということになり好ましくない。そこで評価テキスト中に出現した未知語の種類 m と、未知語の出現回数 n_u を用いてパープレキシティを補正する [16]。

補正パープレキシティは

$$APP = (P(w_1 \cdots w_n)m^{-n_u})^{-\frac{1}{n}} \quad (3)$$

で与えられる。これは、複数の未知語はそれぞれ等確率に生じると仮定して補正したものである。

2.2 単語 bigram による大語彙連続音声認識

連続音声認識のアルゴリズムとして、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次の式を満たす最尤の単語列候補を計算する [17, 18]。

$$P(w^*|y_1^T) = \underset{\{w_1^N\}\{t_1^N\}}{\text{argmax}} \left\{ \sum_{n=1}^{N} \log(P_{\text{acoust}}(w_n|y_{t_{n-1}+1}^{t_n})) + weight \cdot \sum_{n=1}^{N} \{\log(P_{\text{lang}}(w_n|w_{n-1})) + \Delta \cdot \gamma(w_n)\} \right\} \quad (4)$$

ここで、$P_{\text{acoust}}(w_n|y_{t_{n-1}+1}y_{t_{n-1}+2} \cdots y_{t_n})$ は観測パターン系列 $y_{t_{n-1}+1}y_{t_{n-1}+2} \cdots y_{t_n}$ に対する単語 w_n の音響モデルの尤度、$P_{\text{lang}}(w_n|w_{n-1})$ は単語 w_{n-1} の次に単語 w_n が接続する言語確率、Δ は文長に対するパラメータ、$\gamma(w_n)$ は単語 w_n の音節数である。本稿の評価実験では、文長は考慮していない ($\Delta = 0$)。

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。なお今回の実験で示したtrigramの結果は複数のパラメータを設定した時に認識精度が最大となったものである。

2.3 探索方法の改良

認識精度を高めるために後続単語の予測単語数の制限を緩めることを考えた場合、コンテキストに依存して生成されるサブツリー辞書のコピーが増大するため、認識システムはメモリを大量に消費する。その上、一文当りの認識時間も長くなるという問題があった。そこで単語の履歴ごとにサブツリーを生成し、計算した尤度をそれぞれに保持していたのを、常に各処理時刻までの1-bestの結果をコンテキスト部分の候補の近似として用い、1つの語彙全体をカバーするツリー状の探索空間に対して尤度計算を行なう近似的な方法に変更した。

同一の予測単語幅だと探索方法変更後の方が認識率が落ちてしまうが、予備的な実験の結果、語彙が5000単語で予測単語数を最大の5000にした場合でも、従来予測単語幅を500にした場合と比べてメモリ使用量は1/2~1/10で、認識精度は第1パスで8.1 %、第2パスで7.9 %改善できた。探索方法の改良の効果については、4.1節で述べる。

3 N-gram言語モデルの構築

3.1 毎日新聞記事データベース

毎日新聞読み上げ音声コーパスは日本音響学会により作成された大語彙音声データベースである。このデータベースはパープレキシティにより、いくつかのランクに分けられ、話者ごとに用意した男女それぞれ155セットの音声からなる。

本研究ではこの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。このうち1991年1月から1994年9月までを学習データ(train)、1994年10月から1994年12月までをテストデータ(test1)とした†。また、音声認識実験では、これらのうち100文(test2)を用いた。N-gramの構築のための学習データとしてはRWCの形態素解析結果を使用した。学習データの形態素総数は約8600万個(約330万文)である。

3.1.1 面種別の言語モデル

新聞記事は面種によって分類されているため、面種別に学習することで、認識タスクに対するよい言語モデルを得ることが可能であると考えられる。毎日新聞記事データベースの場合、表1のような面種で記事が分類されている。

これらの面種別で学習した言語モデルのパープレキシティを表2にまとめた。この表に示したパープレキシティは、面種毎に評価し、それらの平均をとったものである。

ここでPPは、未知後を一形態素としたパープレキシティで、APPは未知後の種類数を考慮した補正パープレキシティである。なお、音声認識実験に用いた評価用100文には未知語が1つ存在した。

この結果から、bigramについては全ての面種のデータを用いて言語モデルを構築するより、面種毎に言語モデルを構築し、テストデータの面種に応じて言語モデルを選択したほうが良いと考えられる。一方、trigramでは全面種で作成した言語モデルの方がよく、学習データが不足していることをことを示している。

全面種で学習した場合にバイグラムで、学習データの補正パープレキシティより、評価データの補正パープレキシティの方が小さくなっている。これは補正パープレキシティがデータのサイズに依存する特性を持つために現れたものである。全面種のトレーニングデータとテストデータの未知語について調べてみたところ、評価データと比べ、学習データの方が未知語が出現頻度で約20倍、種類数で約4倍ほど現れていた。

表1: 毎日新聞の面種

一面	二面	三面	解説	社説	国際	経済	特集
総合	家庭	文化	読書	科学	芸能	スポーツ	社会

表2: 面種別の評価結果(語彙サイズ5000)

言語モデル		面種別で学習			全面種で学習		
評価データ		train	test1	test2	train	test1	test2
bigram	PP	91.8	104.0	161.1	104.2	108.8	189.0
	APP	395.5	407.8	161.1	436.3	426.8	189.0
trigram	PP	67.9	87.2	145.6	69.4	81.1	139.2
	APP	292.6	342.0	145.6	290.6	318.1	139.2

train トレーニングデータ(1991.1~1994.9)
test1 テストデータ(1994.10~1994.12, 約23万文)
test2 音声認識評価文(100文)

3.1.2 適応化法

新聞記事では数日間に渡って関連のある記事が載っていることがある。そこである時期の記事に対してより良い言語モデルを与えるために、過去の数

†言語モデルには句読点(30種)を含む。但し、パープレキシティの算出時には句読点を無視した。

日間の記事で言語モデルを適応化する方法が考えられる。

ここでは、N-gram 言語モデルの適応化には MAP 推定 (最大事後確率推定) [4] を用いる。適応化サンプルを与えた後の推定値は次式で与えられ、推定前の条件確率と現在与えたサンプルの線形補間の形になっている [4]。

$$prob = \alpha \cdot prob_0 + (1-\alpha) \cdot prob_1 \qquad (5)$$

α　適応化の重み
$prob$　MAP 推定後の条件確率 (N-gram 確率)
$prob_0$ 標準言語モデルでの条件確率
$prob_1$ 適応化サンプルでの条件確率

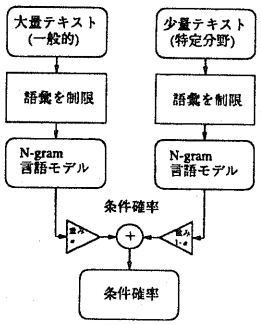

図 1: MAP 推定のブロック図

今回の実験では標準言語モデルと適応化サンプルによる言語モデルの 2 つを構築しておき、バックオフスムージングに基づく 2 つの条件確率を用いて MAP 推定を行なっている。この過程のブロック図を図 1に示す。標準言語モデル (全面種で学習) と適応化サンプルによるモデル (評価文の直前の n 日間のデータを使用) の学習では、新聞記事の全面種に対応する学習サンプルで出現頻度の高い 5000 個の形態素に限定し、使用した。適応化サンプルの語彙は全面種で学習した語彙と同じものを使用した。

表 3に MAP 推定による言語モデルの適応化の結果を示す。評価文の数日間分のデータを用いて適応化することによりパープレキシティは bigram で約 13%,trigram で約 17%減少した。

表 3: MAP 推定による適応結果
(適応化の重み 0.8, 語彙サイズ 5000,
音声認識評価文 test2 のパープレキシティ)

適応化サンプル		なし	5 日	15 日	30 日
bigram	PP	189.6	167.3	163.6	162.5
trigram	PP	139.6	115.9	116.4	115.9

3.1.3 定型表現

新聞テキストやニュース文には、定型的な表現が多く含まれる。そこで、高頻度の定型表現を 1 形態素として捉えた上で言語モデルを構築すれば、より精度の良いモデルが出来ると考えられる。

今回、使用頻度の高い定型表現を抽出するアルゴリズムとして、池原らの提案した方法 [19] を用いた。この方法では、最長一致の文字列抽出 (ある文字列が抽出されたとき、その文字列に含まれる部分文字列は統計量を求める際にはこの部分文字列を定型表現とはカウントしない) を条件とし、任意の長さ以上、任意の使用頻度以上の表現を、もれなく自動的に抽出する。文献 [19] では文字列単位で抽出していたが、これを形態素単位で適用した。抽出例を表 4に示す [10]。

表 4: 定型表現抽出例

連語数 2 定型表現 (頻度)	連語数 3 定型表現 (頻度)
て/いる (318691)	し/て/いる (106121)
は/ない (56333)	に/よる/と (24093)
東京/都 (23452)	に/なっ/て (19718)
大統領/は (14647)	話し/て/いる (6130)
国民/の (9909)	記者/会見/し (4297)

以下に定型表現を用いた言語モデル構築のための手順を示す。

Step.1 定型表現抽出

RWC の毎日新聞形態素解析結果に対して、前述の方法で連結数 2 の定型表現を抽出する。こうして得られた連結数 2 の定型表現のうち句読点を含んでおらず出現頻度が高い上位 2000 個をそれぞれ一つの単語として元の 5000 語彙に追加する。

Step.2 定型表現の連結

Step.1 の定型表現を用い、トレーニングデータ内の定型表現を図 2のように 1 つの単語にまとめる。

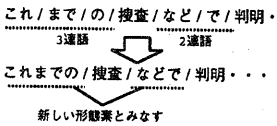

図 2: 形態素の連結例

Step.3 言語モデルの構築

CMU SLM Toolkit を用いてトレーニングデータから、語彙サイズ 7000(ベース 5000, 定型表現

2000) の辞書を作成し、bigram,trigram 言語モデルを構築する。

定型表現での評価結果を表5に示す (パープレキシティは元の形態素当りの値)。定型表現は既登録形態素の連結で定義されるため、定型表現を追加してもカバー率は変化しない。この結果より、期待に反して定型表現を用いることでパープレキシティは全体的に大きくなっている。句読点を含めた場合はパープレキシティが小さくなったのとは対照的であった。

表 5: 定型表現のパープレキシティ
(語彙サイズ 5000+2000)

データセット		train		test1		test2	
定型表現		なし	2 連結	なし	2 連結	なし	2 連結
bigram	PP	104.2	118.4	108.8	124.2	189.0	238.7
	APP	436.3	586.4	426.8	593.2	189.0	238.7
trigram	PP	69.4	84.1	81.1	98.8	139.2	202.3
	APP	290.6	416.5	318.1	472.1	139.2	202.3

train : トレーニングデータ (1991.1~1994.9)
test1 : テストデータ (1994.10~1994.12, 約 23 万文)
test2 : 音声認識評価文 (100 文)

3.2 NHK ニュース原稿コーパス

もう一つの大量の言語データベースを用いることができるコーパスとして、NHK ニュース原稿データベースを使用した。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。NHK のニュース音声は 1996 年 6 月 1 日~10 日の 10 日間のニュース音声からなる。そこで、言語モデルの学習には 1991 年 4 月から 1996 年 5 月までのニュース原稿 (train) を用い、言語モデルの評価に 1996 年 6 月のニュース原稿 (test1) を用いた。形態素解析には京大で作成された JUMAN を使用した。学習データの総形態素数は約 2000 万個 (47 万文) である。なお、句読点は削除した。また、3.1.3 節と同様に 2000 個の定型表現も抽出した。言語モデルのパープレキシティを表 6に示す。定型表現の登録によりパープレキシティは小さくなっている。

表 6: ニュース文の言語モデルの評価
(語彙サイズ 5005+2000)

データセット		train		test1		test2	
定型表現		なし	2 連結	なし	2 連結	なし	2 連結
bigram	PP	58	48	60	52	100	99
	APP	246	143	247	135	100	99
trigram	PP	25	24	32	32	58	63
	APP	77	73	84	84	58	63

train : トレーニングデータ (1991 年 4 月~1996 年 5 月)
test1 : テストデータ (1994 年 6 月, 約 2 万文)
test2 : 音声認識評価文 (70 文)

4 音声データベースを用いた評価実験

4.1 音声データベースと実験条件

認識実験の基本的な実験条件を表 7に示す。言語モデルは語彙サイズが 5000 である。

表 7: 認識実験の基本的な実験条件

音響モデル	
	5 状態 4 出力分布 (2 混合ガウス分布, 全共分散行列)
	離散継続時間分布付き連続出力分布型 HMM
音節カテゴリ数	113 音節
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 (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

認識文は表 8に示すように 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) を使用した。

表 8: 評価用音声データ (毎日新聞)

男性話者 10 名の初めの各 10 文
NM006 NM014 NM017 NM021 NM026
NM089 NM102 NM109 NM115 NM122

初めに、探索方法の近似に伴う単語認識率の変化を調べた結果を表 9に示す。この表の bigram は認識システムの第 1 パスでの第一候補の認識率、200best は認識システムの第 1 パスでの上位 200 候補で認識率の最も高い候補文を選んだ場合の認識率、trigram は第 1 パスで得られた 200best の候補を trigram を用いてリスコアリングした場合の認識率、Time は一文当りの認識時間を示している。ちなみに、この実験では単語辞書に単一の読みしか登録していなかったため (読みの間違いがあったため)、他の実験と比べて認識率が全体的に約 3 %ほど低くなっている。

この認識実験の結果を見ると、同じ予測単語幅だ

と探索方法の変更後のほうが変更前と比べて認識精度が落ちていることが分かる。しかし探索方法変更前にはシステムが動作する限界点であった予測単語幅1000の結果よりも、探索方法変更後の予測単語幅5000の方が認識率が良い。また1文当りの認識時間も探索方法変更後のほうが短い。そこで、この後の認識実験では、探索方法変更後の認識システムを予測単語幅5000に設定して使用した。

表 9: 探索方法の改良による単語認識率 (%)
(単一の読みを登録)

言語重み	探索方法変更前		探索方法変更後	
	予測単語500	予測単語1000	予測単語500	予測単語5000
	Cor. Acc.	Cor. Acc.	Cor. Acc.	Cor. Acc.
bigram	78.7 71.9	81.7 76.9	78.1 71.8	83.4 80.0
trigram	79.7 75.2	83.6 79.5	78.8 73.5	87.5 83.1
200best	86.4 82.5	90.4 87.4	85.8 81.4	93.1 90.7
Time	117 (sec)	211 (sec)	41 (sec)	54 (sec)

Cor. = 100 - Sub. - Del. (%)
Acc. = 100 - Sub. - Del. - Ins. (%)

4.2 全面種と面種別で学習したときの認識実験

全ての面種の記事を学習データとして言語モデルを構築した場合と、面種別に学習データとして言語モデルを構築した場合の認識結果を表10に示す(表2参照)。

この結果より、面種別で学習した言語モデルを使用するよりも全面種で学習した言語モデルを使用するほうが認識率が高くなることが分かる。これはパープレキシティによる学習モデルの評価の結果と異なっている。この原因を調べてみたところ、全面種で学習した言語モデルだと認識文に対するバイグラムのヒット率は97%で、面種別で学習した言語モデルだと認識文に対するバイグラムのヒット率は89%だった。ヒット率が悪いとパープレキシティは大きくなるはずだが、逆の結果になっている。スムージングの方法に問題がある可能性がある。

trigramによるリスコアリングを行なった結果は1パス目のバイグラムの結果と比べて0.6 %ほどの向上しか見られなかった。この傾向は他の認識実験でも見られる傾向である。200best での結果でも、単語正解率で95 %、単語正解精度で93 %である。trigramによるパープレキシティの減少が比較的少ないのと、200bestの精度が不十分なため、trigramによる十分な認識率の向上が見られなかったと考えられる。

表 10: 全面種と面種別の言語モデルによる単語認識率 (%)

	LM	Cor.	Acc.	Sub.	Ins.	Del.
bigram	全面種	86.4	83.3	11.1	3.1	2.5
	面種別	85.3	81.8	11.7	3.5	3.0
trigram	全面種	87.6	83.9	9.9	3.7	2.5
	面種別	84.4	81.7	11.8	2.7	3.8
200best	全面種	95.1	92.8	3.5	2.4	1.3
	面種別	95.2	92.8	3.5	2.5	1.2

4.3 タスク適応化の認識実験

適応化の重みを0.2, 0.4, 0.6, 0.8と変化させて認識実験を行なった。その結果は適応化の重みを小さくするほど単語認識精度が落ちる結果となったため、以後0.8を用いた。

タスク適応化の認識実験の結果を表11,12にまとめた。この結果より、適応化による認識率の向上があまり見られず、30日分の適応化サンプルを使用したときで、バイグラムでの単語正解精度が1 %しか上がらなかった。

表 11: 適応化による単語認識率 (%)

	サンプル	Cor.	Acc.	Sub.	Ins.	Del.
bigram	なし	86.5	82.6	11.4	3.9	2.1
	5日	85.5	81.2	11.9	4.3	2.6
	15日	86.6	83.4	10.7	3.2	2.7
	30日	87.1	83.7	10.4	3.3	2.6
trigram	なし	86.6	83.9	9.9	2.8	3.4
	5日	88.0	83.9	10.2	4.1	1.8
	15日	86.6	84.0	9.8	2.7	3.5
	30日	86.6	83.9	9.9	2.8	3.4
200best	なし	94.8	91.8	3.9	3.0	1.2
	5日	94.8	91.5	3.8	3.3	1.3
	15日	94.7	91.7	3.8	3.0	1.4
	30日	94.8	91.8	3.8	3.0	1.3

4.4 定型表現を用いた時の認識実験

パープレキシティによる評価の結果、定型表現を用いるとパープレキシティは全体的に大きくなったため、認識率の向上は期待できない。しかし、見ための語彙サイズが増加したためにパープレキシティが増加した可能性がある。そこで第1パスで使用するバイグラムの構築に定型表現を使用して、このことを調べた。ベースとなる語彙5000個に定型表現2000個を追加して認識実験を行なった。この2000個の定型表現はベースとなる語彙5000に存在する単語で構成された定型表現の中で出現頻度の高い上位2000個をとったものである。

定型表現を用いて構築した言語モデルを使用した

表 12: bigram の適応化による音声認識結果 (%)

適応化サンプル		なし		30 日	
面種	評価文の数	PP	Acc.	PP	Acc.
一面	6	119.5	81.9	109.7	84.7
二面	7	178.3	87.1	134.0	87.1
三面	5	177.5	91.7	160.3	91.7
解説	7	123.3	78.5	119.4	77.2
社説	7	183.3	86.7	168.2	85.0
国際	5	218.2	78.6	198.7	82.1
経済	9	208.4	84.6	170.6	83.3
特集	3	131.5	86.5	113.2	89.2
総合	20	241.7	76.2	223.6	77.7
家庭	3	284.8	75.0	202.1	82.1
文化	3	237.5	78.3	175.8	82.6
科学	1	128.6	100.0	110.1	100.0
芸能	3	144.3	88.0	74.2	92.0
スポーツ	12	189.6	86.0	152.1	86.9
社会	9	228.7	87.5	197.8	88.8
Total	100	189.0	82.6	162.5	83.7

ときの認識結果を表 13に示す。この結果から、定型表現を用いても認識率が向上せず、悪くなっていることが分かる。

表 13: 定型表現の使用による単語認識率 (%)

	定型表現	Cor.	Acc.	Sub.	Ins.	Del.
bigram	なし	86.4	83.3	11.1	3.1	2.5
	2 連結	86.3	83.1	10.9	3.2	2.8
trigram	なし	87.6	83.9	9.9	3.7	2.5
	2 連結	83.5	79.8	12.8	3.7	3.6
200best	なし	95.1	92.8	3.5	2.4	1.3
	2 連結	95.1	92.8	3.4	2.4	1.4

4.5 NHK ニュース音声タスクでの認識実験

5005 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、4 形態素以下の短い文と 31 形態素以上の長い文を省き、最終的に 70 文の評価用データ (test2) を使用した。このうちアナウンサーの発話が 59 文、レポーターの発話が 11 文である。

使用した音響モデルは毎日新聞のタスクと同じものを使用した。しかし音響モデルの学習に使用した音声データと比べ、ニュース音声の発声スピードが全体的に速いことから、HMM の状態毎の継続時間分布の継続時間長を 1/1.7 に縮めた。

標準言語モデルと定型表現を考慮した言語モデルによる認識結果を表 14,15に示す。ニュース音声にはアナウンサーの他にレポーターの音声が含まれている。アナウンサーは発話の訓練を受けているが、レポーターは訓練を受けていない。またレポーターは現場で立ちながら話しているため、アナウンサーと比べ音声認識が困難になっていると考えられる。そこでアナウンサーとレポーターを分けて表に示している。

この結果は毎日新聞の結果と比べて非常に悪い結果となっている。その原因として考えられるのは、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 (spontaneous speech) であることである。次に、毎日新聞よりニュース音声のほうが認識文の平均単語数が多いことが挙げられる。更に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。また、定型表現を用いた場合、レポーターに対しては良くなったが、全体としては悪くなった。

表 14: NHK ニュース音声タスクでの単語認識率 (%)

言語モデル	話者	Cor.	Acc.	Sub.	Ins.	Del.
bigram	アナウンサー	68.1	62.9	24.2	5.2	7.7
	レポーター	50.0	46.2	37.7	3.8	12.3
	Total	66.1	61.0	25.7	5.1	8.2
trigram	アナウンサー	69.1	64.1	22.8	5.0	8.1
	レポーター	55.4	52.3	27.7	3.1	16.9
	Total	67.6	62.8	23.3	4.8	9.1
200best	アナウンサー	77.0	73.3	17.0	3.6	6.1
	レポーター	68.5	64.6	19.2	3.8	12.3
	Total	76.0	72.4	17.2	3.6	6.8

表 15: NHK ニュース音声タスクでの単語認識率 (%)
(定型表現, 語彙サイズ 5005+2000)

言語モデル	話者	Cor.	Acc.	Sub.	Ins.	Del.
bigram	アナウンサー	64.3	59.1	23.9	5.3	11.4
	レポーター	56.2	52.3	32.3	3.8	11.5
	Total	63.4	58.3	24.9	5.1	11.4
trigram	アナウンサー	64.9	58.8	25.7	6.1	9.4
	レポーター	57.7	53.1	31.5	4.6	10.8
	Total	64.1	58.1	26.4	5.9	9.6
200best	アナウンサー	73.0	69.3	17.5	3.7	9.2
	レポーター	73.8	69.2	17.7	4.6	8.5
	Total	73.1	69.3	17.5	3.8	9.1

音響的に困難なタスクであるか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 16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。

表 16: 連続音節認識実験結果 (%)
(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	78.6	69.8	18.6	8.8	2.9
NHK ニュース音声	57.8	48.6	32.9	9.2	9.3

5 まとめ

新聞記事読み上げコーパスを用いて、言語モデルのタスク適応化、定型表現を用いた言語モデルの構築の評価を行なった。また、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。

まず言語モデルについては、bigram の有効性は明らかとなったが、適応化や trigram などでは、パープレキシティは小さくなったものの認識率の大幅な向上には繋がらなかった。

定型表現の登録に関しては認識率の向上には繋がらなかったが、認識誤りを起こしやすい箇所を定型表現を用いて訂正に利用していく予定である。

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学習の必要性が現れた。また、間投詞などを考慮に入れた言語モデルや認識アルゴリズムの構築なども考えていかなければならない [20]。

謝辞

今回の認識実験を行なうにあたり、大語彙連続音声認識システムの改良をして頂いた廣瀬良文氏、定型表現を用いた言語モデルを作成して頂いた西崎博光氏、連続音節認識実験を行なって頂いた花井健豪氏に感謝致します。

参考文献

[1] 西村雅史, 伊藤伸泰, 山崎一孝, 荻野紫穂: 単語を認識単位とした日本語の大語彙連続音声認識, 情報処理学会研究報告 97-SLP-20-3,pp.17-24(1998-2)

[2] 甲斐充彦, 伊藤敏彦, 山本一公, 中川聖一: 自然な発話を対象としたパソコン／ワークステーション用連続音声認識ソフトウェア, 日本音響学会秋季講演論文集 2-Q-30(1997)

[3] 伊藤克亘 他: 語彙日本語連続音声認識研究基盤の整備—学習・評価テキストコーパスの作成—, 情報処理学会研究報告 97-SLP-18-2,pp.7-12(1997-10)

[4] 赤松裕隆, 中川聖一: 新聞記事のトライグラムによるモデル化と適応化, 言語処理学会, 第 3 回年次大会 D5-2,pp.533-536(1997)

[5] 小林紀彦, 中野裕一郎, 和田陽介, 小林哲則: 統計的言語モデルにおける高頻度形態素連鎖の辞書登録に関する一考察, 情報処理学会, 音声言語情報処理 SLP-20-5, pp.33-38(1998-2)

[6] 小黒玲, 高木一幸, 橋本顕示, 尾関和彦: ニュース音声認識のための言語モデルの比較, 日本音響学会春季講演論文集 1-6-22, pp.47-48(1998)

[7] E.P.Giachin: " Phrase bigrams for continuous speech recognition ", Proc.ICASSP, pp.225-228(1995)

[8] 政瀧浩和, 松永昭一, 匂坂芳典: 連続音声認識のための可変長連鎖統計言語モデル, 信学技報 SP95-73, pp.1-6(1995-11)

[9] S.C.Marlin, J.Liermann, H.Ney: " Adaptive topic dependent language modelling using word-based varigrams", Proc. EuroSpeech, pp.1447-1450(1997)

[10] 西崎博光, 中川聖一: 音声認識のための定型表現を用いた言語モデルの検討, 言語処理学会, 第 4 回年次大会 C4-3, pp.520-523(1998)

[11] 大附克年, 森岳至, 松岡達雄, 古井貞照, 白井克彦: 新聞記事を用いた大語彙連続音声認識の検討, 信学技報 NLC95-55,SP95-90,pp.63-68(1995-12)

[12] 河原達也 他: 日本語ディクテーション基本ソフトウェア (97 年度版) の性能評価, 情報処理学会研究報告 98-SLP (1998-5)

[13] 大附克年, 松岡達雄, 松永昭一, 古井貞照: ニュース音声を対象とした大語彙連続音声認識と話題抽出, 信学技報 SP97-27,pp.67-74(1997-6)

[14] 小林彰夫 他: ニュース音声認識システムの検討, 日本音響学会秋季講演論文集 3-1-9, pp.103-104(1997)

[15] R.Rosenfeld: "The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation", Proc. ARPA Spoken Language Systems Technology Workshop, pp.47-50(1995)

[16] J.Ueberla:"Analysing a simple language model - some general conclusion for language models for speech recognition", Computer Speech and Language, vol.8, No.2, pp.153-176(1994-4)

[17] 周旻 , 堤真理子 , 中川聖一: 確率モデルにおける大語彙連続音声認識の評価, 情報処理学会研究報告 96-SLP-11-6 , pp.31-36 (1996-5)

[18] 甲斐充彦, 廣瀬良文, 中川聖一: N-gram 言語モデルと効率的探索法を用いた大語彙連続音声認識システムの検討, 信学技報, SP97-99,pp.31-38(1998-1)

[19] 池原悟, 白井諭, 河岡司: 大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法, 情報処理学会論文誌, Vol.36, No.11,pp.2584-2596(1995)

[20] 廣瀬良文, 甲斐充彦, 中川聖一: バイグラム言語モデルに基づく対話音声認識における冗長語・未知語処理, 信学技報, SP98-4,pp.25-32(1998-4)