

ニュース音声認識のための時期依存言語モデル

小林 彰 夫[†] 今 井 亨[†]
安 藤 彰 男[†] 中 林 克 己[†]

本論文では、ニュース番組の音声認識の性能を向上させるため、最新ニュースの内容に依存した言語モデル（時期依存言語モデル）を提案する。まず、ニュース音声認識のための言語モデルについて、学習期間、学習時期を調べた。その結果、学習データは長期間であるほど、また、評価時期に近いほど有効であることが分かった。そこで、長期間の大量のニュース原稿に対して、最新の少量のニュース原稿を MAP 推定に基づいて重み付け混合し、時期依存言語モデルを作成した。長期間ニュース原稿に対する最新ニュース原稿の重みは、EM アルゴリズムにより推定される。提案手法の特徴は、n-gram を最新ニュースに適応化するだけでなく、語彙を自動的に更新することである。認識実験の結果、時期依存言語モデルはテストセット 297 文中、最新ニュースに関連する 140 文について、単語正解精度を 2.0% 向上させ、未知語の延べ数を 29.4% 削減した。一方、関連のないニュースについて単語正解精度は低下しなかった。

Time Dependent Language Model for Broadcast News Transcription

AKIO KOBAYASHI,[†] TORU IMAI,[†] AKIO ANDO[†]
and KATSUMI NAKABAYASHI[†]

In this paper we propose a new linguistic technique to make an adapted language model (Time Dependent Language Model, TDLM) which gives an improvement in broadcast news transcription. Examining a good condition of a language model for broadcast news transcription, we found out that long-term training scripts and the latest news scripts were effective. In our proposed method, the TDLM is trained from long-term news scripts and short-term latest news scripts based on MAP estimation. The mixture weight for short-term news scripts are derived from the EM algorithm. The method features not only an adaptation of n-grams for the latest news scripts but updating of TDLM's vocabulary automatically. We had a transcribing experiment for 297 sentences in NHK's broadcast news. The TDLM achieved 2.0% improvement in word accuracy over the baseline (no adapted) language model for 140 sentences which are associated with the latest news content and reduced 29.4% of unknown words in these topic related sentences. On the other hand, it didn't lower the word accuracy for the rest sentences.

1. はじめに

筆者らは現在、ニュース番組に対する字幕放送の実現を目的として、音声認識を用いたニュース番組用字幕作成システムの研究を進めている。ニュースの場合、アナウンサーの発声速度が速いため、キーボード入力では同音異義語の選択などが間に合わない。そこでリアルタイム音声認識などの自動手法を併用する必要がある。

ニュースを対象にした連続音声認識の研究は、欧米

では Hub4 タスクを中心に活発に行われており^{1),2)}、日本においても同様の報告がされている^{3),4)}。これら一連の研究が進化した背景には、新聞・放送に関連した大量のテキストコーパスが整備され、精度の高い言語モデルの構築が可能になったことがあげられる。しかし、ニュースというタスクに着目した場合、その特徴を利用した言語モデルの学習法は、これまであまり提案されていない。

ニュースには、関連する話題が数日間にわたって報道されるという特徴がある。時間的に継続した話題では、特定の単語や言い回し（単語の組合せ）が共通して出現することが多い。また、ニュースの内容は時間とともに変化するため、人名や企業名のような固有名

[†] NHK 放送技術研究所
NHK Science and Technical Research Laboratories

詞は、使われなくなったり頻繁に使われるようになったりする。

本論文ではまず、ニュースの特徴という観点から言語モデルを検討し、ニュース音声認識のための言語モデルを構築するのに有効なニュース原稿の性質を調べる。

次に、最新のニュース原稿と長期間のニュース原稿から新しい語彙と適応言語モデルを作成する手法を提案する。言語モデルの適応化の研究として、MAP推定に基づくタスク適応^{5)~7)}があるが、これらは語彙を更新することは試みていない。本論文では、最新ニュースに依存した n-gram・語彙が得られるように、n-gram と語彙の両方を更新する適応手法について提案する。この手法によって作成された言語モデルを時期依存言語モデルと呼び、最新の少量ニュース原稿の期間(原稿量)や、異なる放送日におけるテストセットの評価を通じて、時期依存言語モデルの有効性について検討を行う。

2. ニュース用言語モデル

2.1 言語モデルの学習期間

統計的言語モデルには大量の学習データが必要である。そこで、ニュース原稿を用いて、学習期間の異なる言語モデルによるテストセットパープレキシティの変化を調べた。

学習データはNHKニュース原稿データベース(1991年4月から1996年6月3日までの約5年間、78万文、38M単語)を利用し、形態素解析ツール茶筌⁸⁾を用いて形態素に分割した。茶筌による形態素解析の精度について、ニュース原稿データベース中の987文(82,884文字)を調べたところ、再現率95.7%、適合率94.4%であった。単語分割精度の指標であるF-尺度⁹⁾は0.951となった。単語分割の際のエラーの多くは固有名詞(人名や企業名、その略称など)であった。なお、本論文では茶筌により分割された形態素を単語と呼ぶことにする。

テストセットはニュース音声データベース¹⁰⁾から、1996年6月4日放送のニュース83文(総単語数2,974)を選んだ。言語モデル作成にはCMU-Cambridge SLM Toolkit¹¹⁾を利用した。語彙サイズは学習データ中の出現頻度の高い20k単語とし、bigramを作成した。cut-off値は1とし、Good-Turingによるバックオフスムージングを行った。また、テストセットは補正パープレキシティ¹²⁾に従って評価した。

学習期間をテストセットの評価時点(96年6月4日)から過去にさかのぼって1年から5年までとし

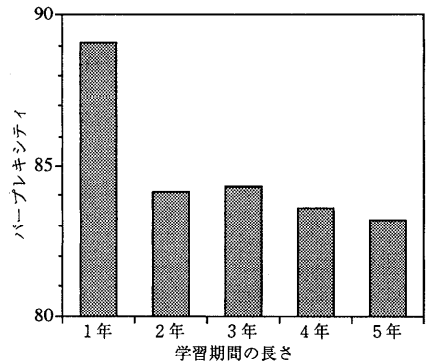


図1 学習期間に対するパープレキシティの変化
Fig.1 Perplexity results with various lengths of training period.

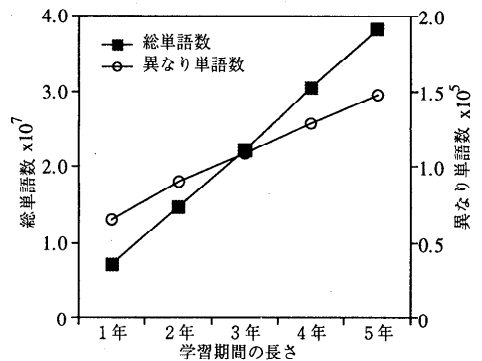


図2 総単語数・異なり単語数の変化
Fig.2 The number of total words and unique words with various lengths of training period.

て、各々の学習期間に対する言語モデルを作成し、テストセットパープレキシティを求めた(図1)。

その結果、学習期間が長くなるに従って、パープレキシティが低下し、5年分の原稿で学習した言語モデルのパープレキシティが最も小さくなった。1年分のニュース原稿は最近のニュースを相対的に多く含んでいるが、パープレキシティが他の言語モデルと比べて高くなっており、十分な精度のモデルが得られなかった。以下、5年分のニュース原稿で学習した言語モデルをベースライン言語モデルと呼ぶ。ベースライン言語モデルにおいて、学習に用いたニュース原稿に対するサイズ20kの語彙のカバー率は97.9%であった。

2.2 異なり単語数の変化

前節で作成した言語モデルの総単語数と異なり単語数の学習期間ごとの変化を調べた(図2)。その結果、異なり単語数は学習期間に対し単調に増加していることが分かった。ただし、この実験結果では、形態素解析エラーによって発生した未知語が異なり単語の中に

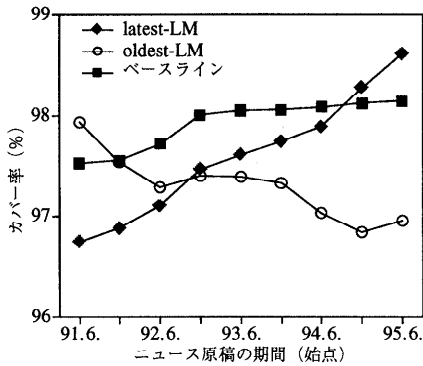


図3 カバー率

Fig. 3 Coverage results for language models with various training periods.

含まれている可能性があるが、形態素解析の適合率が94.4%であることと、5年分のニュース原稿に対する出現頻度上位20k語のカバー率が97.9%であることから、仮にエラーが含まれているとしても小量だと考えられる。異なり単語数の増加が示すのは、時間の経過に従って新しい単語が出現しているということであり、ニュースの内容が時間とともに変化していることを表している。

2.3 言語モデルの学習時期

新しい単語が時間の経過ともなって出現しているのであれば、言語モデルの学習を行った時期によって、異なる期間のニュース原稿に対するカバー率が変化すると考えられる。そこで、ニュース原稿データベースから、データベースの最も古い1年分(1991年6月~1992年5月まで)と最近の1年分(1995年6月~1996年6月3日まで)のニュース原稿データを選び、それぞれ言語モデルを作成した。前者をoldest-LM、後者をlatest-LMと呼ぶことにする。各モデルとも語彙サイズを学習データ中の単語の出現頻度に従って20kとした。

2つの言語モデルとベースラインモデルを用い、ニュース原稿データベースの異なる期間のニュース原稿に対するカバー率を求めた。様々な時期のカバー率を求めるために、ニュース原稿データベース(1991年6月から1996年6月3日まで)を半年の重複を与えながら1年分ずつ分割し、計9つのテキストを用意した。各期間のニュース原稿に対するカバー率を図3に示す。oldest-LM、latest-LMの言語モデルはともに、学習した時期から時間的に離れるに従いカバー率が低下した。oldest-LMのカバー率の変化は、語彙中の単語が新しいニュースでは使われなくなっていることを示し、latest-LMのカバー率の変化は古いニュー

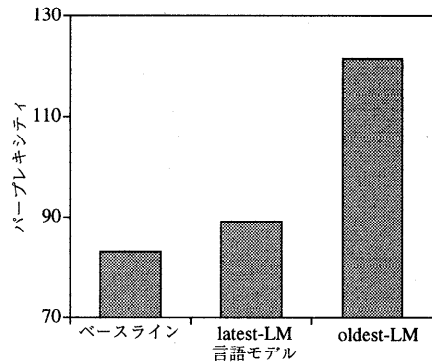


図4 学習時期によるパープレキシティの比較

Fig. 4 Perplexity results for language models with various training periods.

スにはない新しい単語が使われるようになったことを示すと考えられる。一方、5年分のデータで学習したベースラインモデルでは、各ニュース原稿に対するカバー率の変化は少なく、ほぼ一様にカバーしていることが分かる。

次に、ベースライン、oldest-LM、latest-LMの3つの言語モデルを用い、2.1節のテストセット83文に対するパープレキシティを求めた。その結果を図4に示す。カバー率に対する結果から予想されるように、テストセットの時期に近いlatest-LMとベースライン言語モデルによるパープレキシティはoldest-LMよりも低い値となった。

以上の結果から、パープレキシティを低下させるには、学習期間については長期間の方が良く、学習時期については評価する時点に近いニュース原稿を多く用いる方が良いといえる。

3. 時期依存言語モデル

3.1 言語モデルの適応化

この章では、長期間のニュース原稿と、最新のニュース原稿を用いて、最新ニュースの内容を反映するような時期依存言語モデルの作成方法について検討する。

少量の原稿と大量の原稿から、言語モデルの確率分布を推定する方法として、MAP(事後確率最大化)推定に基づく手法^{6),7)}がある。伊藤ら⁶⁾は、少量の原稿を重み付けして大量の原稿と混合し、混合した原稿から、語彙制限を行って言語モデルを得る手法をとっているが、適切な重みの値を取得するのにヒューリスティックな試行を重ねる必要がある。また、赤松ら⁷⁾は、少量の原稿および大量の原稿それぞれから言語モデルを作成し、両方の線形補間によって適応化した言語モデルを得る手法をとっている。この方法では事前

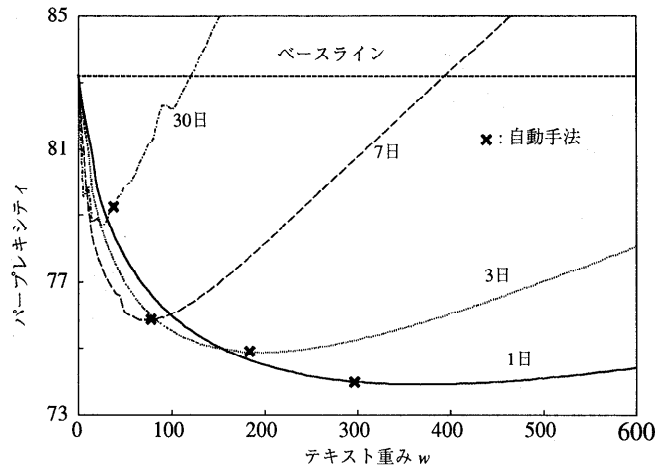


図5 時期依存言語モデルとテキスト重み w

Fig. 5 Perplexity results for TDLMs with various adaptation data and text weight w .

表1 最新ニュース原稿

Table 1 Short-term latest news scripts.

最新ニュース原稿	文章数	総単語数
1日分	457	21.3 k
3日分	986	45.9 k
7日分	3074	149.9 k
30日分	11384	555.8 k

に語彙を決定しなければならず、少量の原稿と大量の原稿からバランス良く単語を選びだして語彙を決定することが困難となる。

本論文の提案手法は上記の問題点を解消し、自動的に最新ニュース原稿の内容を考慮した言語モデルを作成する。また、提案手法により作成された言語モデルを時期依存言語モデル (Time Dependent Language Model, TDLM) と呼ぶことにする。

3.2 ヒューリスティックな適応化

まず、長期間ニュース原稿と最新の少量ニュース原稿を足しあわせて作成した言語モデルについて調べた。このときの足しあわせ回数をテキスト重み w と呼ぶことにする。テキスト重み w は、原稿中の単語または単語対の出現頻度に対する重みとなる。語彙は足しあわされたニュース原稿での単語の頻度に基づいて制限した (サイズ 20 k)。長期間ニュース原稿は、ベースラインモデル作成時と同様に5年分とした。評価データは2.1節のテストセット83文とした。重み付けに用いた最新ニュース原稿は、テストセットの前1日～30日分の4種類を用意した (表1)。

ベースラインモデルと比較した結果を図5に示す。すべての最新ニュース原稿でパープレキシティは減

少したが、テキスト重み w が大きすぎると逆に増加した。また、最新ニュース原稿の期間が短いほど、パープレキシティはより小さい値を与えた。パープレキシティの最小値は、1日分の原稿による重み付けで350回程度としたときに、73.9となった。このとき、ベースラインモデルのパープレキシティ (=83.2) に対する削減率は11.2%であった。

このように、試行を重ねてヒューリスティックに最適なテキスト重みを探索することは非効率的であり、パープレキシティ最小値を与えるテキスト重みの値の判定も困難である。

3.3 時期依存言語モデル

次に、テキスト重み w を自動的に取得し、語彙を更新する時期依存言語モデルについて述べる。

時期依存言語モデルの学習の流れを図6に示す。以下では、時期依存言語モデルとして bigram を用いた場合について説明を進める。

始めに、長期間のニュース原稿の語彙を V_0 、最新ニュース原稿の語彙を V_1 として、 $V^{(0)} = V_0 \cup V_1$ となるように初期語彙 $V^{(0)}$ を決める。ただし、 V_0, V_1 と語彙サイズの上限を与える (以下の実験では語彙サイズを各ニュース原稿の単語の出現頻度に従って20 kとした)。単語 $u, v \in V_0 \cup V_1$ について、時期依存言語モデルの bigram $P(v|u)$ を、長期間のニュース原稿と最新ニュース原稿から作られた bigram $P_0(v|u), P_1(v|u)$ の線形補間で表現する。

$$P(v|u) = \lambda P_0(v|u) + (1 - \lambda) P_1(v|u)$$

各言語モデルの確率に対する重み (以下、確率重み) λ は EM アルゴリズムの再推定式から以下のように求

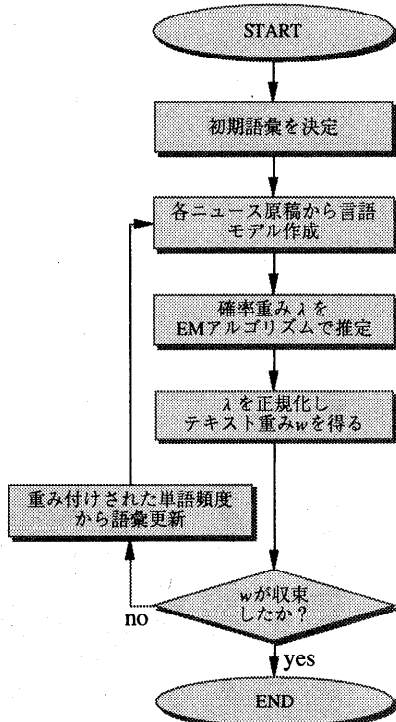


図 6 時期依存言語モデルの学習
Fig. 6 Training algorithm for the TDLM.

める。

$$\hat{\lambda} = \frac{1}{N} \sum_{n=1}^N \frac{\lambda P_0(y_n|y_{n-1})}{\lambda P_0(y_n|y_{n-1}) + (1-\lambda)P_1(y_n|y_{n-1})}$$

ただし、 y_n は EM アルゴリズムで確率重み λ の推定を行うための重み推定用テキスト $S = y_1, y_2, \dots, y_N$ の単語である。重み推定用テキストは長期間原稿、最新原稿とは別に用意する。

長期間ニュース原稿に対する最新ニュース原稿のテキスト重み w は上で推定した確率重み λ を正規化して求める⁵⁾。

$$w = \frac{(1-\lambda)m_0}{\lambda m_1}$$

ただし、 m_0, m_1 はそれぞれニュース原稿の大きさ(総単語数)である。

テキスト重み w で最新ニュース原稿を足しあわせ、あらたに重み付けされた単語の頻度からサイズ 20k の語彙 $V^{(1)}$ を決め直す。新しい語彙 $V^{(1)}$ で bigram P_0, P_1 を再度計算し、テキスト重み w を再決定する。語彙 $V^{(j)}$ の更新とテキスト重み計算の操作を、テキスト重み w が一定の値に収束するまで繰り返し、時期依存言語モデルを得る。

テキスト重み w で語彙を更新することにより、最

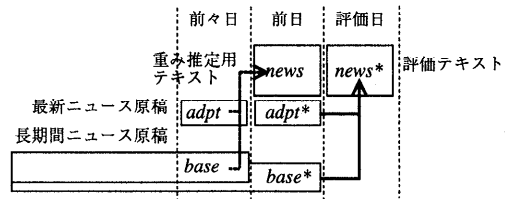


図 7 未知データの評価方法
Fig. 7 The way of evaluating open data $news^*$ with a pseudo-weight derived from $base$ and $adpt$.

新のニュースに現れやすい単語をテキスト重みに応じて語彙に反映させられる点が本手法の特長である。

3.4 未知データの評価

提案手法では、長期間原稿、最新原稿および EM アルゴリズムで確率重みを推定する重み推定用テキストの 3 種類が必要である。これら一連のテキストの関係を時系列で表したのが図 7 である。音声認識しようとする評価テキスト $news^*$ に対して最適な時期依存言語モデルは、評価前日までの長期間原稿 $base^*$ と評価日前日の最新原稿 $adpt^*$ から、 $news^*$ に対して確率重み λ を推定した場合と考えられるが、 $news^*$ は通常未知なので、適切な確率重みを求めることができない。そこで、 $news^*, adpt^*, base^*$ から得られる確率重みの値は、これら 3 種類のテキストをそれぞれ 1 日ずつ前にシフトしたときの、評価日の前日の重み推定用テキスト $news$ 、評価日前々日の最新原稿 $adpt$ および長期間原稿 $base$ から得られる確率重みと同じ値であると仮定する。 $news$ に対する確率重み λ の値を求めるために、 $base$ と $adpt$ からそれぞれ言語モデルを作成し、両言語モデルから、2.1 節に基づき EM アルゴリズムを用いて確率重みを推定する。ただし、確率重み λ からテキスト重み w へは $base^*, adpt^*$ の総単語数で正規化し、 $news^*$ に対するテキスト重みとする。このテキスト重みと、ニュース原稿 $base^*, adpt^*$ から、 $news^*$ に対する時期依存言語モデルを作成する。

4. 実験と考察

4.1 時期依存言語モデルと最新ニュース原稿

前述の提案手法に従って、最新ニュース原稿の期間を変えて語彙サイズ 20k の時期依存言語モデルを作成した。提案手法によって得られたテキスト重み w と時期依存言語モデルのパープレキシティの関係を、図 5 (×印) に示す。このときのテキスト重み w の値を表 2 に記す。30 日分を除くと、ヒューリスティックな手法のパープレキシティの最小値に近い値をとった。また、1 日分、3 日分、7 日分の最新ニュース原稿による時期依存言語モデルでは、最新原稿中のすべ

表2 各最新原稿に対するテキスト重み w Table 2 Text weight w for each latest script.

最新ニュース原稿	テキスト重み w
1 日分	297
3 日分	184
7 日分	78
30 日分	38

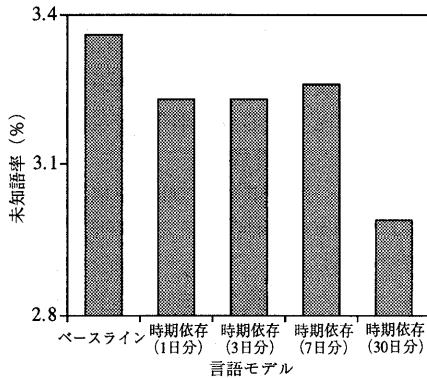


図8 各時期依存言語モデルによる未知語率の変化

Fig. 8 Out of vocabulary rates for TDLMs with various adaptation data.

での異なり単語が語彙に登録された。

最新ニュース原稿が30日分の場合、原稿のサイズが大きく、テキスト重みのわずかな増減で語彙の中の多数の単語が入れ替わる。このため、提案手法によるテキスト重みの値がヒューリスティックな手法による値に対してずれたと考えられる。

時期依存言語モデルによる、テストセットの未知語率を図8に示す。その結果、30日分の最新原稿を使った時期依存言語モデルの場合に未知語率が最小となった。テストセット中の未知語をカバーするには、最新原稿の期間を長くする必要があることが分かるが、同時に不要なニュース原稿の文章数も増加するため、パープレキシティの削減効果が期待できない。

評価直前のテキストを利用した言語モデルの適応化については、文献7), 13)と同様にパープレキシティ削減効果が確認された。しかし、文献7), 13)では、大量テキストと小量テキストから作成された言語モデルの語彙はそれぞれ独立しており、小量テキストから新たに得られた単語を、大量テキストの言語モデルに反映させることなく適応化を行っている。本手法の場合、線形補間の確率重みを求めた後、単語の頻度を再計算して語彙を更新する。このため、すべてのテキストに対して、出現が期待される単語を n -gram の計算に利用できる。

表3 最新ニュース原稿・テストセット

Table 3 Series of latest news scripts and evaluation news scripts.

放送日	最新原稿		テストセット	
	文章数	総単語数	文章数	総単語数
6/5	440	22.6 k	183	7.9 k
6/10	233	10.7 k	190	8.0 k
6/15	620	30.3 k	117	4.8 k
6/20	458	21.4 k	199	7.6 k
6/25	456	20.7 k	250	9.2 k
6/30	299	14.2 k	97	4.5 k
7/11	389	18.2 k	297	10.4 k

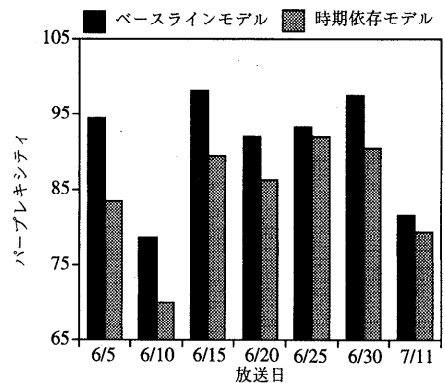


図9 パープレキシティの削減効果

Fig. 9 Perplexity results for the baseline model and the TDLM over a series of test sets.

4.2 異なる放送日での評価

異なる放送日のニュースをテストセットとして選び、時期依存言語モデルによるパープレキシティの削減効果を調べた。放送日は1996年6月5日、10日、15日、20日、25日、30日、7月11日の7日とした(表3)。

ベースラインモデルはそれぞれ、放送日の前日までの約5年分の長期間原稿から作成した。

最新ニュース原稿は放送日の前1日分とし、長期間ニュース原稿と最新ニュース原稿(表3)から時期依存言語モデルを作成した。ベースラインを含め、各言語モデルとも2.1節で述べた条件に従い、語彙サイズ20kとした。

各モデルで評価したテストセットパープレキシティを図9に示す。すべてのテストセットの中で、パープレキシティ削減率が最大となったのは6月5日放送のニュースを評価したときで、ベースラインモデルに比べパープレキシティは11.6%削減した。反対に、6月25日放送のニュースではパープレキシティ削減率が最小となった(0.9%)。

なお、最新原稿が3日分の場合についても上記の一連のテストセットについて実験を行ったが¹⁴⁾、同程度

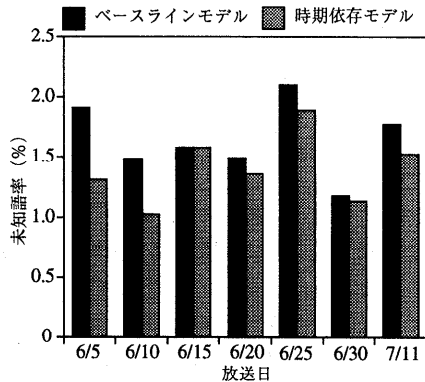


図 10 未知語の削減効果

Fig. 10 Out of vocabulary rates over a series of test sets with the baseline model and the TDLM.

のパープレキシティ削減効果であった。

ベースラインモデル・時期依存言語モデルいずれの場合も放送日によってパープレキシティの値にばらつきがあった。原因として、テストセット中には気象情報などのニュースがあるが、学習に用いたニュース原稿中には存在しないことなどがあげられる。

パープレキシティ削減率も放送日によって大きく異なった。最新原稿中にない新しいニュースや関連の低いニュースの量によって削減率が変わることが考えられる。

ベースラインモデルと時期依存言語モデルそれぞれを用い、各テストセットの未知語率について調べた(図 10)。未知語の削減効果は、パープレキシティ同様、放送日によって異なった。効果が最も大きかったのは、6月5日の時期依存言語モデルで、テストセットに対する未知語率を1.9%から1.3%に引き下げた(テストセットの未知語の延べ数を31.2%削減)。逆に6月15日の時期依存言語モデルは、未知語をまったく削減できなかった。原因は、テストセット中に最新原稿にない新しいニュースが存在したこと、ベースラインでは登録されていた単語が語彙の更新により登録されなかったことである。

4.3 音声認識実験

ニュース音声データベースから、放送日を1996年7月11日としてテストセット(297文、総単語数10,430)を選び、時期依存言語モデルを用いた音声認識実験を行った。実験では時期依存モデルとして bigram および trigram を作成した。ただし、trigram の語彙は時期依存 bigram を作成したときの語彙と同一とした。trigram 学習時の cut-off 値は2とし、bigram 同様、Good-Turing のスムージングを行った。

表 4 テストセット(放送日7/11)のパープレキシティ
Table 4 Perplexity results (Jul. 11).

	ベースライン	時期依存モデル
bigram (第1パス)	81.6	79.5
trigram (第2パス)	41.2	39.7

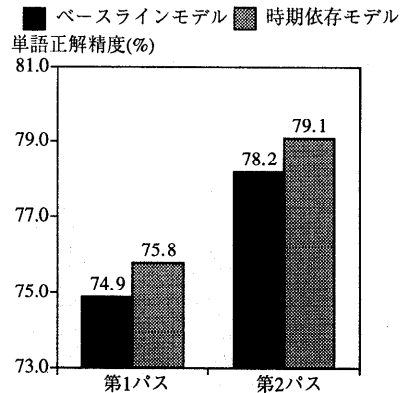


図 11 テストセット(放送日7/11)の単語正解精度
Fig. 11 Word accuracy (Jul. 11).

実験で利用した音声認識デコーダ³⁾は2-パスで構成されており、第1パスは bigram を使った Viterbi ビームサーチを行って N-best 候補を求め、第2パスでは trigram によるリスコアリングで認識結果を得る。今回の実験では第1パスで 400 best を得ることにした。また、音響パラメータは 39 次元(12 次元 MFCC + 対数パワーおよびそれぞれの Δ , $\Delta\Delta$)とし、男女別の特定話者を対象とした混合ガウス分布型 HMM (triphone) を用いた。

時期依存言語モデルを作成したときの最新原稿のテキスト重み w は bigram で 144, trigram で 110 となった。テストセットパープレキシティを表 4 に示す。パープレキシティ削減率はベースラインモデルに対し bigram で 2.5%, trigram で 3.6% となった。ベースラインモデル、時期依存言語モデルによる単語正解精度を図 11 に示す。時期依存言語モデルによる単語正解精度はベースラインモデルに比べ 0.9% 向上した。

テストセットの未知語率はベースラインモデルの 1.8% から時期依存言語モデルの 1.6% に減少した。時期依存言語モデルにより新しく語彙に登録された単語は、延べ 26 単語だった。このうち 25 単語が名詞(固有名詞は 11 単語)、1 単語は動詞であった。認識実験の結果、26 単語中の名詞 14 単語(53.8%) が正しく認識された。固有名詞に限ると、11 単語中 6 単語(54.5%) が正しく認識された。

次に、テストセット中の最新ニュース原稿に関連したニュースと、関連のないニュースのそれぞれについて

表 5 話題の関連性によるパープレキシティ
Table 5 Perplexity results (topic related).

	テストセット 1		テストセット 2	
	ベースライン	時期依存	ベースライン	時期依存
bigram (第 1 パス)	76.8	71.7	85.7	86.7
trigram (第 2 パス)	38.6	34.9	43.4	44.1

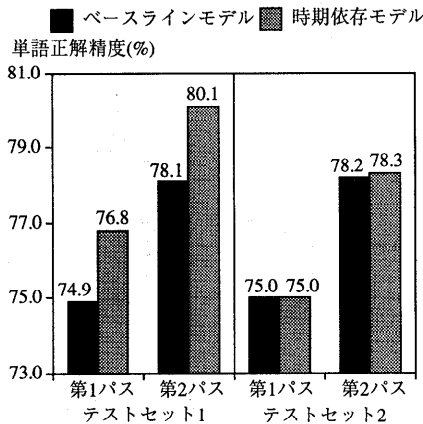


図 12 話題の関連性による単語正解精度
Fig. 12 Word accuracy (topic related).

て、ベースラインモデルと比較したときの認識率の変化を調べた。テストセットを最新ニュース原稿中の話題と関連のあるもの(テストセット 1, 140 文)とないもの(テストセット 2, 157 文)との 2 つに人手で分けて評価した。それぞれのテストセットについて第 1 パス、第 2 パスでのパープレキシティを表 5 に示す。時期依存モデルのテストセット 1 に対するパープレキシティ削減効果は bigram で 6.6%, trigram で 9.6%であった。一方、テストセット 2 では bigram, trigram とも、ベースラインモデルに比べパープレキシティはわずかに大きくなった。テストセット 1 について、未知語の延べ数は 29.4%削減したが、テストセット 2 では変わらなかった。

認識結果を図 12 に示す。最新ニュース原稿と関連のあるテストセット 1 では単語正解精度が 2.0%向上し、関連のないテストセット 2 の場合は単語正解精度が下がらなかった。

以上から、時期依存言語モデルは、最新ニュース原稿に関連した話題について、パープレキシティ・未知語とも削減し認識率を向上させること、一方関連のない話題の認識率を下げないことが分かった。

5. ま と め

ニュースの特徴を利用した、言語モデルの適応化手

法について述べた。提案手法は、最新のニュース原稿の内容を反映するように、語彙と n-gram を自動的に求めることを特徴としている。実験の結果、パープレキシティや未知語の減少については、ばらつきはあるものの、ある程度の効果が得られた。また、最新の 1 日分のニュース原稿を用いた時期依存言語モデルによって、単語正解精度はベースラインモデルに比べ 0.9%向上し 79.1%となった。また、最新のニュース原稿と関連した話題に限った場合、単語正解精度は 2.0%向上し、80.1%となった。一方、関連のないニュースに対しては単語正解精度を下げないことが分かった。

本論文では、話題の時間的な継続性に基づいた言語モデル適応化の実験結果を示した。この際、以下の問題点が明らかとなった。

- (1) テストセットの一部のニュースは、関連する話題が数日前に存在する場合がある。
- (2) あるニュースで使われる単語・単語対は、関連の希薄な話題の中に出現する場合がある。
- (3) 最新原稿の中には不要なニュースも含まれており、適応化を阻害する一因となる。

このような問題に対しては、事前に得られる情報を利用することが考えられる。たとえば、放送すべきニュースに関して、タイトルやキーワードなどの事前情報によって、類似した話題を時間的に近接した過去のニュース原稿から探し出し、適応データとして利用することがあげられる。

今後は、話題の時間的な継続性をさらにきめ細かく考慮できるような言語モデルについて検討する予定である。

参 考 文 献

- 1) Polymenakos, L., Olsen, P., Kanvesky, D., Gopinath, R., Gopalakrishnan, P. and Chen, S.: Transcription of Broadcast News - Some Recent Improvements to IBM's LVCSR System, *ICASSP-98*, pp.901-904 (1998).
- 2) Cook, G. and Robinson, T.: Transcribing Broadcast News with the 1997 ABBOT System, *ICASSP-98*, pp.917-920 (1998).
- 3) 今井 亨, 尾上一穂, 小林彰夫, 安藤彰男: ニュース音声認識用デコーダの開発, 音響学会平成

- 10 年度秋期研究発表会講演論文集, pp.105-106 (1998).
- 4) 大附克年, 松永昭一: 適応音響モデルを用いた放送ニュース音声認識, 音響学会平成 10 年度秋期研究発表会講演論文集, pp.49-50 (1998).
- 5) Federico, M.: Bayesian Estimation Methods for N-gram Language Model Adaptation, *ICASSP-96*, pp.240-243 (1996).
- 6) 伊藤彰則, 好田正紀: 対話音声認識のための事前タスク適応の検討, 電子情報通信学会研究会報告, SP96-81, pp.91-98 (1996).
- 7) 赤松裕隆, 甲斐允彦, 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-SLP-21-11, pp.97-104 (1998).
- 8) 松本裕治, 北内 啓, 山下達雄, 平野喜隆, 今一修, 今村友明: 日本語形態素解析システム『茶筌』version1.5 使用説明書 (1997).
- 9) 永田昌明: 単語頻度の再推定による自己組織化単語分割, 自然言語処理研究報告, 98-NL-121-2, pp.9-16 (1997).
- 10) 安藤彰男, 宮坂栄一: ニュース音声データベースの構築, 音響学会平成 9 年度春期研究発表会講演論文集, pp.157-158 (1997).
- 11) Clarkson, P. and Rosenfeld, R.: Statistical Language Modeling Using the CMU-Cambridge Toolkit, *EUROSPEECH-97* (1997).
- 12) Uebara, J.: 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).
- 13) 中川聖一, 赤松裕隆, 西崎博光: 音声認識用言語モデルのためのタスク適応化と定型表現の利用, 自然言語処理, Vol.6, No.2, pp.97-113 (1999).
- 14) 小林彰夫, 今井 亨, 安藤彰男: ニュース音声認識のための時期依存言語モデル, 音響学会平成 10 年度秋期研究発表会講演論文集, pp.71-72 (1998).

(平成 10 年 10 月 12 日受付)

(平成 11 年 2 月 8 日採録)



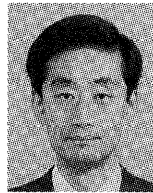
小林 彰夫

平成 3 年早稲田大学理工学部電気工学科卒業。同年 NHK 入局。岡山、広島放送局を経て、平成 8 年より同放送技術研究所勤務。音声認識の研究に従事。電子情報通信学会、日本音響学会、映像情報メディア学会各会員。



今井 亨

昭和 62 年早稲田大学理工学部電気工学科卒業。同年 NHK 入局。大阪放送局を経て、平成 2 年より同放送技術研究所勤務。平成 8 年米国 BBN 滞在研究員。現在、NHK 放送技術研究所において音声認識、話題抽出の研究に従事。平成 10 年日本音響学会粟屋潔学術奨励賞受賞。電子情報通信学会、日本音響学会各会員。



安藤 彰男 (正会員)

昭和 53 年九州芸術工科大学芸術工学部音響設計学科卒業。昭和 55 年同大学大学院修士課程修了。同年 NHK 入局。昭和 58 年より同放送技術研究所勤務。音声認識の研究に従事。昭和 62 年日本音響学会第 4 回粟屋潔学術奨励賞受賞。電子情報通信学会、日本音響学会、言語処理学会、映像情報メディア学会、IEEE 各会員。



中林 克己

昭和 42 年東京大学工学部電気工学科卒業。同年 NHK 入局。昭和 53 年日本音響学会佐藤論文賞受賞。電子情報通信学会、映像情報メディア学会、日本音響学会、Audio Engineering Society 各会員。