自然言語処理 131－2
音声言語情報処理 26－2
（1999．5．28）

# 講演ディクテーションのための話題独立言語モデルと話題適応

加藤 一臣　李 晃伸　河原 達也

京都大学大学院 情報学研究科 知能情報学専攻
〒606-8501 京都市 左京区 吉田本町
e-mail: kazuomi@kuis.kyoto-u.ac.jp

**あらまし**　講演音声ディクテーションのための話し言葉のモデル化とその話題適応の方法を提案する。まず多数の話題からなる講演録を利用して、話題と出現単語の相互情報量に基づいて話題独立語の選択を行い、この語彙によって話題独立の言語モデルとした。このモデルを講演の予稿テキストから構築した言語モデルと重み付け混合することで話題適応を行い、当該講演の言語モデルを構築した。実際に男性話者1名の約10分間の口頭発表に対してディクテーションを行った結果、単語認識精度77.5%という結果が得られた。最後に、間投詞に対処したモデルを用いた結果、80.5%まで単語認識精度を向上できた。

**キーワード**　講演、話し言葉、ディクテーション、統計的言語モデル、話題適応

# Topic Independent Language Model and its Adaptation for Dictation of Lecture Speech

Kazuomi Kato　Akinobu Lee　Tatsuya Kawahara

Graduate School of Informatics
Kyoto University, Kyoto 606-8501, Japan
e-mail: kazuomi@kuis.kyoto-u.ac.jp

**Abstract**　We present a method to construct a language model for the dictation of lecture speech. Topic independent lexicon is selected based on mutual information between the topics and a word using transcriptions of various lectures. This model is adapted to a specific lecture to be transcribed. Specifically it is mixed with the language model which is built from the preprint paper of the lecture. We have evaluated the model by dictation of oral presentation of the paper. The word accuracy was 77.5%. And by dealing with filler words, the accuracy was improved to 80.5%.

**key words**　lecture speech, spoken language, dictation, statistical language model, topic adaptation

# 1 はじめに

近年何千語あるいは何万語といった大語彙で連続音声認識の研究が盛んに行われており、ディクテーションシステムとしても大きな成功を収めつつある[1]。その基盤は大量の読み上げデータ、テキストデータに基づく統計的手法である。

これまでのディクテーションの研究対象は、新聞記事の読み上げ音声やニュースの朗読音声、いわゆる「書き言葉」であった[2][3]。しかし、これからはより自然な音声である対話音声、すなわち「話し言葉」に対するディクテーションが求められるであろう。話し言葉に対するモデルを構築するためには、話し言葉の特徴を反映した大量のデータが必要となるが、現実には、そのようなデータは収集が困難であり、モデル構築の大きな障害の一つとなっている。

本研究では、話し言葉の特徴を持つ講演音声をディクテーションをするための言語モデルの構築を行う。ここで得られた知見は、話し言葉のモデル化にも有用であろう。また、このディクテーションの実現により、講演録を人手で書き起こすという労力とコストが削減できる。

講演には何らかの主題があり、この話題（主題）に関する部分が占める割合は非常に大きい。また講演調の言い回しや一般的な言葉は、講演の話題に依存しないと考えられる。このことから複数の講演から講演調にのみ依存した部分を抽出し、話題に独立した統計的言語モデルを構築した。その上で対象とする講演の予稿等を用いて話題に適応することで、ディクテーションを行うのに必要な話題依存の言語モデルとし、実際にディクテーションを試みた。

# 2 講演音声に対する言語モデル

## 2.1 講演音声のモデル化における問題点

音声認識の対象として「書き言葉」と「話し言葉」という二つの発話の種類がある。「書き言葉」は原稿の読み上げ音声や推敲してからの発声である。省略・倒置などは少なく冗長な部分が入る余地も少ない。「話し言葉」は自発的な発声で、会話のやりとりから成る言葉である。主語や助詞の省略などが多く、また間投詞（「え～」「あの～」など）が頻繁に出現し、言い直しもよく生じる。文末も、終助詞が付随する場合が多い。

本研究で扱う講演音声は、「書き言葉」と「話し言葉」の中間であると考えられる[4]。文章がある程度整っているという書き言葉の特徴が反映されていると共に、間投詞・文末表現・言い直しなどに講演音声の言語的特徴が現れている。以下、テレビの講演調の独話音声の観察に基づいてその特徴を列挙する。

「まー」「えー」などの間投詞は言葉をつないでいく際に多用されるが、現れる頻度や種類が話者によって大きな偏りがあるので、厳密なモデル化は非常に難しい。文末では、終助詞が「ですね」の形でよく現れ、文の合間にも挿入される。

新聞記事などの読み上げではほとんど現れないが、講演音声に固有の特徴としては、「～ございます」「～して頂きまして」などの「です・ます」調などの丁寧表現や、「～と思います」「～というわけです」などの話者の気持ちを表す言葉や説明の言い回し、「～でしょうか」といった問いかけなどが挙げられる。

さらに、講演には題目が掲げられることからもわかるように、その講演特有の話題というものが存在する。したがって、講演音声には講演調の言い回しだけではなく、講演話題に関する特殊な単語キーワードが必ず存在し、しかもそれが繰り返し用いられる。様々な種類の内容のテキストを混ぜ合わせたものから単純に言語モデルを構築しても対応できない。

## 2.2 講演音声に適した言語モデルの構築法

講演録などを収集しそこから学習を行うことで、講演調の特徴を反映したモデルが構築できると考えられるが、講演の話題に依存した語彙の変化を考慮する必要がある。収集した講演データには、各講演のそれぞれの話題に依存する語が多数含まれる。これらの不必要な話題依存語は削除し、ディクテーションを行う講演の話題に関する単語を、言語モデルの語彙に含めなければならない。

さらに講演録を学習コーパスとする場合、書き起こし時の後処理・編集によって、間投詞などの話し言葉の特徴がコーパスから欠落することがある。実際の講演音声は、このような間投詞が必ず発生するので、誤認識が起こる可能性は高く、避けて通れない問題となる。

このような点を考慮して本研究では、図1に示す方法で講演音声ディクテーションのための言語モデルを構築する。まず講演調の単語・言い回しは講演の話題に依存しないとして、各話題と出現単語の相

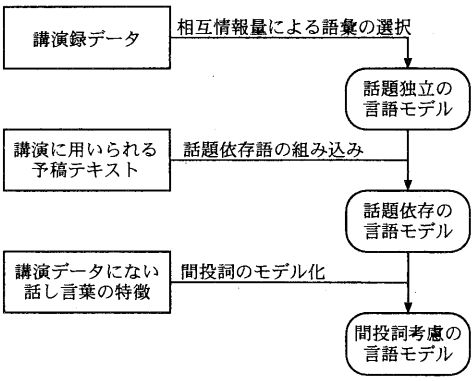


図 1: 講演音声に対する言語モデル構築の手順

表 1: 学習に用いる講演データのタイトルと語彙数

| タイトル | 単語総数 | 語彙数 |
|---|---|---|
| 「活力ある農業農村」 | 10900 | 3167 |
| 「有人宇宙活動の社会的意義」 | 11060 | 3822 |
| 「始末の美学」 | 5009 | 2204 |
| 「Web コンテンツと知的財産」 | 5231 | 2160 |
| ⋮ | ⋮ | ⋮ |
| 合計（39 講演） | 666843 | 28870 |

互情報量を利用して、話題独立の講演調言語モデルを構築する。次にこのモデルを、対象講演の予稿テキストを利用して対象話題に適応させる。また、実際の独話音声に対する間投詞の分析を行い、その特徴を反映するようにモデルを調整する。

# 3 話題独立の講演調言語モデル

## 3.1 学習に用いた講演データ

学習に用いた講演データは、主に World Wide Web 上で公開されている講演録を利用した。表 1 に使用データのタイトルとその規模を示す。ここで、単語総数とはテキストデータに出現したのべ単語数であり、語彙数は出現単語の種類数である。全部で 39 種類の講演データからなり、単語総数は約 66 万 6 千語、出現語彙数は 2 万 8 千語であった。これらを学習テキストとして、単語辞書と言語モデルを構築する。

## 3.2 認識に用いる単語の単位

音声認識で用いる単語の単位には様々なものが考えられるが [5]、本研究では形態素を単語の単位とする。形態素解析には JUMAN ver3.5 [6] を用いた。実際に認識の際に用いる単語には、形態素の表記だけではなく、その読みと品詞・活用情報カテゴリを持たせ、単語の正確な発音と品詞情報まで考えた単語連鎖の生起確率を用いる。

単語の読み付与は重要であるが、これまでは認識に用いる単語辞書上で表記に対して何重にも発音を登録していた。例えば「2」には「(ニ・ツー)」の二つの発音が登録されており、「通用」という音声に対して、「2・秒」と認識する可能性があった。本研究では、単語の読みまで考慮に入れることで単語の発音を一意に定め、このような誤りが起こらないように言語モデルを構築した。ただし「私（ワタシ・ワタクシ）」のように、連鎖する前後の単語が変化しないが、一意に読みが定まらないものに関しては、二重に発音を登録するようにしている。発音のない句読点については、文の区切りで出現することから、ショートポーズに対応させる。

また、講演中に頻出すると思われる数詞は、多種多様な表記法を統一するために漢数字表記を採用し、語彙数の増加を抑制するために位取りで分離し、一つの単語とする。例えば、「1234 万 5000」は「千」「二百」「三十」「四」「万」「五千」のように分割する。

## 3.3 話題独立の語彙の構成

本研究ではどのような話題でも普遍的に出現する語彙を選択するために、講演の話題と出現単語の相互情報量を用いて話題独立の語彙を構成する [7]。話題独立語彙の構成の手順を図 2 に示す。

相互情報量とは、情報理論的な意味での関連率を表し、話題の集合を $T$, 単語を $w$ とすると次のような式で表される。

$$I(T;w) = \sum_t P(t)\log_2\frac{1}{P(t)} - \sum_t P(t/w)\log_2\frac{1}{P(t/w)}$$

$$P(t/w) : \frac{\text{話題 } t \text{ における単語 } w \text{ の出現回数}}{\text{全講演における単語 } w \text{ の出現回数}}$$

この相互情報量が大きい単語は、ある特定講演に突出して出現している単語であり、話題依存語と判定できる。逆に相互情報量が小さい単語は、どの講

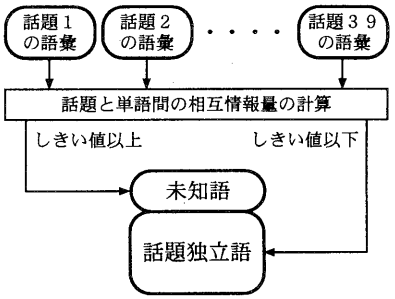


図 2: 話題独立の語彙の構成

演にも偏りなく出現している話題独立の単語である。このことから適当な相互情報量のしきい値を設定して、話題独立の語彙を構成することができる。

相互情報量の計算に先立ち、表 1 に示した講演データの各話題の出現単語数によって単語頻度の正規化を行っている。これによって、全ての話題は同等に扱われる。

講演データから、本研究で扱う相互情報量に基づいて構成した話題独立語彙と、単語の出現頻度によって構成した語彙を、同じ 5000 単語の語彙サイズで構成し、どちらか一方にのみ存在する単語を調べた。その結果を図 2 に示す。二つの語彙間で共通の単語は、5000 語中 3866 単語で、入れ替わったと考えられる単語数は、1134 単語である。

表 2: 話題独立語彙と出現頻度での語彙の内容比較

| 話題独立語彙 | 出現頻度語彙 |
|---|---|
| あたかも　あっち | ASEAN　アトリエ |
| このように　だったり | エンドユーザ　高橋 |
| 気づいて　嫌だ　出来た | 委譲　現物　人権 |
| 集め　情けない　頂いて | 稟議　利回り　揚力 |

図 2 に示すように、本研究で提案した手法で構成した語彙では、アルファベット単語や固有名詞が頻度が大きくても削除され、話し言葉の特徴を持つような副詞、動詞や文の終わり・つなぎの用言部分などが代わりに含まれている。

| | |
|---|---|
| 1 | 違う点もいくつかあります。<br>第一にはですね、さっきにも申し上げましたけども、活動に関してです。 |
| 2 | 例えばみなさんよくご存知のような、 |
| 3 | 気持よく聞いて頂けるように<br>気をつかっております。 |

図 3: 講演調データとして使用した文例

### 3.4 話題独立の言語モデルの予備評価

前節で述べた語彙を用いて、表 1 に示した講演データから話題独立の N-gram 言語モデルを構築する。言語モデルの構築には、CMU_TK ver.2 を用いて tri-gram を構築した。

ここで構築した言語モデルが、話題独立の講演調言語モデルになっていることを確認するために、予備実験として他のモデルとの比較実験を行った。

音響モデル・認識エンジンは、IPA の日本語ディクテーション基本ソフトウェア 97 年度版 [1] を使用した。音響モデルは、新聞記事読み上げ音声による 2000 状態 16 混合連続分布の triphone モデルであり、認識エンジンには、本研究室で開発された 2 パス探索を行う JULIUS [8] を用いた。

テレビの講演調の独話音声の中からその話題に依存しないと思われる文を 50 文選び、男性話者 1 名によりその文を再現してもらうことで、話題独立の講演調音声とした。その際、採用した実際の音声中の言い直し・間投詞は削除し、滑らかな発話音声によって実験データとした。使用した文は、四種のテーマの独話から得た。この文の中から 3 つを例として表 3 に示す。

比較に用いた言語モデルは以下の通りである。語彙サイズはどの言語モデルについても 5000 語である。

(a) 講演録から相互情報量による言語モデル
(tri-gram の cut-off 1)

(b) 講演録から出現頻度による言語モデル
(tri-gram の cut-off なし)

(c) 新聞記事から構築した言語モデル
IPA の日本語ディクテーション基本ソフトウェア 97 年度版 [1] に含まれる言語モデル

実験を行うにあたり、テストセットを忠実に書き起こしたテキストに対する各言語モデルの単語カバ

表 3: 話題独立文に対するカバレージと単語認識精度

| 言語モデル | 講演録 | | 新聞記事 |
|---|---|---|---|
| | 相互情報量 | 出現頻度 | |
| カバレージ | 94.6% | 93.5% | 84.3% |
| 認識精度 (bigram) | 70.4% | 69.1% | 34.6% |
| (trigram) | 75.1% | 72.0% | 37.3% |

レージを求めた。新聞記事言語モデルの単語は、本実験で用いた単語と比べて読み付与がなく品詞情報も異なるのでカバレージは単語の表記のみに対するものを求めた。また、形態素解析基準が異なる点も考慮した。その結果、カバレージは表 3 に示す通りになった。

単語認識精度を表 3 に示す。正解文は、漢字かな混じり表記で与えた。漢字とかなの使い分けによって、考えられる正解文を複数用意する (例:「皆さん」と「みなさん」)。使用している形態素基準が異なることを考慮して、新聞記事言語モデルでの正解文は人手で出来る限り対応を取った。

各言語モデルについて、語彙サイズを 5000 語にそろえて実験を行った結果、講演録データを学習テキストとして用いた言語モデルの方が、カバレージ・認識率共に大きく上回り、講演調のモデル化が行えていることを示した。また、同じ講演録から構築した言語モデルでも、本研究で提案した相互情報量に基づく語彙構成の手法が、単語出現頻度で語彙を構成する方法と比べて、話題独立の講演調音声に対して有効であることが示された。今後はこれを話題独立の言語モデルとする。

# 4 講演調言語モデルの話題適応

## 4.1 話題適応の手法

講演には、多くの場合発表者があらかじめ用意する原稿や予稿資料などが存在する。本研究ではこの予稿テキストを利用して、話題独立の言語モデルを対象講演の話題に適応させる。

大規模なテキストデータから構築した言語モデルを、特定タスクに適応させるという研究は、いくつか行われてきた [9][10]。そこでタスク適応の手法として用いられているのは、大量のデータから構築した N-gram と、目的タスクのデータで作成

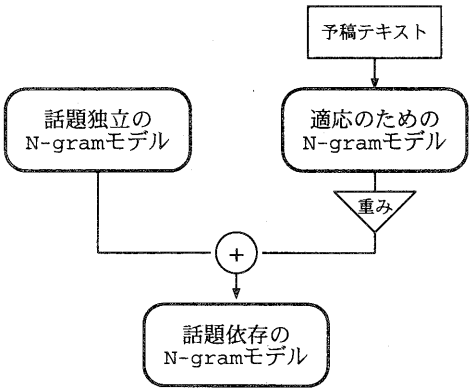


図 4: 話題適応の手順

した N-gram とを重み付けにより混合する方法や、MAP(Maximum A-posteriori Probability) 推定による方法である。

本研究では、まず予稿テキストから tri-gram を構築し適切な重み付けを行い、話題独立の言語モデルと混合する手法 [9] を用いる。話題適応後の語彙は、話題独立の語彙と予稿テキストから得られた語彙を併合して構成する。この話題適応の方法を図 4 に示す。単語総数 $m$ のコーパスから構築した話題独立モデルで単語 $x$ が出現する確率を $f(x)$、単語総数 $m'$ の予稿テキストで同じ単語 $x$ が出現する確率を $f'(x)$ とする。混合に用いる重みを $w$ として話題適応した後の単語 $x$ が生起する確率 $P(x)$ は式 (1) で表される。

$$P(x) = \left(\frac{m}{m+wm'}\right) f(x) + \left(\frac{wm'}{m+wm'}\right) f'(x) \quad (1)$$

また、$c(s)$,$c'(s)$ をそれぞれ話題独立、予稿テキストでの単語列 $s$ の出現頻度とすると、単語列 $x$ から単語 $y$ への生起確率 $P(y|x)$ は、式 (2) で表される。

$$P(y|x) = \left(\frac{c(x)}{c(x)+wc'(x)}\right)\frac{c(xy)}{c(x)} + \left(\frac{wc'(x)}{c(x)+wc'(x)}\right)\frac{c'(xy)}{c'(x)} \quad (2)$$

## 4.2 評価に用いたデータ

評価の対象としては、本研究室で行われた卒業論文の発表を用いた。比較的文体がしっかり構築されているが、話題に依存した専門用語が多く出現し、

また言い直しや間投詞の存在など、話し言葉の特徴も有している。OHP シートを見ながらの発表音声であり、録音時間は約 10 分間、発表者は男性話者 1 名である。録音環境は、本研究室である。この音声を DAT に 48kHz で録音し、16kHz にダウンサンプリングした上で使用した。認識の際には、ポーズによって切出して適切な長さで用いている。

適応に用いる予稿テキストには、卒業論文を利用した。この中で説明のための図や表、また複雑な記号と式で構成されている部分は削除し、簡単な記号については発音されると予測される単位で単語を構成した。この卒業論文に形態素解析を施した結果、単語総数は 6707 語で語彙数は 972 語であった。話題独立の語彙サイズが 12000 語の場合、適応テキスト語彙に対するカバレージは 80.3%であり、カバーできない話題依存語は合計 191 語ある。この話題依存語の例を図 5 に示す。これらの単語を話題適応後に語彙に含まれるようにし、発表中に出現する話題依存語をモデル化する。

| |
|---|
| Ngram　bigram　wi　wj |
| エントロピー　パープレキシティ　ユーザ |
| 音響　音素　解析　形態素　語彙 |
| 尺度　生成　同音　品詞　分かち書き |

図 5: 話題依存語の例

## 4.3 適応に最適なパラメータの調査

本研究では、重みづけ混合という適応手法を用いているので、重みの値を変化させることで、話題適応後のモデルの特徴を変化させることができ、また話題独立の語彙を、どの程度の大きさにするかによっても、認識できる語彙が変化する。

そこで、あらかじめどの程度の混合重みや語彙が最適な値であるのかを、言語的な評価方法で調べた。まず発表音声を忠実に書き起こしたものを用意する。その際、言い直しや間投詞に関しては除外している。話題独立の語彙サイズは、5000,8000,12000 語の三通りについて考えた。12000 語の語彙は二つ以上の話題で出現した単語で構成される。それぞれについて話題適応後の語彙を用いて、発表音声に対する単語カバレージを求めた。これを表 4 に示す。この結果から、話題独立の語彙サイズが 5000 語でも非常に高いカバレージとなっており、語彙サイズを 12000 語にまで増加させてもそれほど変化がないことがわかる。

表 4: 発表音声に対する話題適応後のカバレージ

| 話題適応後の語彙数 | 話題独立の語彙数 | 適応語彙数 | カバレージ |
|---|---|---|---|
| 5K 語 | 5000 語 | 296 語 | 98.7% |
| 8K 語 | 8000 語 | 229 語 | 98.8% |
| 12K 語 | 12000 語 | 191 語 | 98.9% |

表 5: 種々の混合重みと語彙における単語パープレキシティ

| | w=2 | w=3 | w=4 | w=5 | w=10 | w=20 |
|---|---|---|---|---|---|---|
| 5K 語 | 65.7 | 62.5 | 70.6 | 69.5 | 70.2 | 74.4 |
| 8K 語 | 65.3 | 61.9 | 68.0 | 68.6 | 69.4 | 73.4 |
| 12K 語 | 65.4 | 61.9 | 67.8 | 68.4 | 69.2 | 73.1 |

3 種類の語彙で混合重みを変化させて得たモデルを用いて、発表音声の書き起こしに対するパープレキシティの値によって評価を行い、混合重みの最適値を求める。様々な混合重み・語彙によるパープレキシティの値を表 5 に示す。この結果から、どの語彙サイズでも混合重み w=3 の時にパープレキシティが最小値となった。語彙サイズの変化によるパープレキシティの値には目立った差は見られなかった。

## 4.4 ディクテーション実験

このモデルを使用して発表音声に対するディクテーションを試みた。

ここでは新聞記事読み上げ音声による 2000 状態 16 混合連続分布の triphone モデルを音響モデルとして使用した。認識エンジンは JULIUS [8] を用いた。

様々な適応テキストの混合重み・話題独立の語彙の組み合わせに対する単語認識精度の結果を表 6 に示す。表中 tri-gram の結果が最終結果を示している。認識率を計算するに当たって、正解文は漢字・かな文で与えた。間投詞が含まれる正解文は、「エ 二百 七十 ア 二百 五十 七 単語 …」のように間投詞の部分を片仮名で設定した。したがって、偶然間投詞の部分に平仮名を当てるような認識結果が得られても、このモデルにおいて間投詞をモデル化していな

表 6: 種々の混合重みと語彙に対する単語認識精度 (%)

| | w=2 | w=3 | w=5 | w=10 | w=30 | w=50 | w=70 | w=100 |
|---|---|---|---|---|---|---|---|---|
| 5K 語 (bi-gram) | 70.2 | 72.0 | 71.5 | 73.1 | 74.9 | 74.6 | 74.5 | 74.6 |
| 5K 語 (tri-gram) | 76.1 | 77.5 | 75.1 | 76.2 | 77.3 | 76.2 | 75.9 | 76.1 |
| 8K 語 (bi-gram) | 69.9 | 71.9 | 72.2 | 73.1 | 74.1 | 74.5 | 74.7 | 74.2 |
| 8K 語 (tri-gram) | 75.7 | 77.5 | 75.7 | 76.0 | 77.1 | 77.1 | 77.2 | 76.5 |
| 12K 語 (bi-gram) | 69.8 | 71.1 | 72.1 | 73.2 | 73.4 | 74.2 | 74.1 | 74.3 |
| 12K 語 (tri-gram) | 74.9 | 76.2 | 75.7 | 76.1 | 76.0 | 75.8 | 76.0 | 76.5 |

いので、全て誤りと判断される。

認識実験の結果、話題独立の語彙が 12000 語のものは、5000 語や 8000 語と比べて全体的に認識率が低下している。5000 語と 8000 語の認識率はそれほど差が見られない。パープレキシティが最も低かった混合重みが 3 の時に認識率が一度極大値を示し、5000 語と 8000 語の両方で 77.5%となる。しかし、混合重みが 30 以上の大きな値の時に再び認識率が上昇し、5000 語で最高 77.3%、8000 語で 77.2%を示す。また、混合重みの増加にしたがって bi-gram による認識率が向上する一方、tri-gram による認識率向上はそれほど見られない。このことは、混合重みが大きければ発表の話題依存語の生起・連鎖の確率が大きくなり、キーワードを確実に認識することで認識率を向上させていると考えられる。その効果は bi-gram でも十分達成されるが、キーワード以外の部分の認識の向上が見込めないので、tri-gram を用いてもそれほどの効果が得られない。

# 5 間投詞を考慮したモデル

## 5.1 モデル化

講演録では書き起こし及び編集の過程で間投詞・言い直しが削除・修正されており、考慮すべきこれらの現象が実際のモデルに反映されていない。

これらの統計的な性質を得るためには、忠実に書き起こしたデータが必要であるが、入手することは非常に困難である。言い直しに関しては、突発的な現象であり統計的性質が推定できないと考えられるので、本研究では扱わない。一方間投詞については、テレビの講演調の独話音声（話題 10 種類、話者 17 名）に対する観察をしたところ、前後にショートポーズが多い、文頭によく現れるなどの特徴があり、出現位置を N-gram の範囲で比較的容易に扱えると予想できる。そこで、間投詞にも適当な N-gram 確率値を割り振って語彙に含め、通常の単語と同様にモデル化する。

ここでは独話音声の観察の結果に基づいて 15 種類の間投詞を導入した。[ ] を省略可能部分として、「ア [ー], アノ [ー], イ [ー], エ [ー], エ [ー] ト, オ [ー], ト, マ [ー]」である。

前節までは認識の際に、開始記号をモデルに含めていなかったが、間投詞がポーズの直後や文頭に出現しやすいという制約を表現するために開始記号<s>も導入する。まず uni-gram の確率について言語モデル中の未知語カテゴリの確率値から、開始記号<s>と間投詞に対する確率を一定値ずつ配分する。次に bi-gram の確率値は、本来「読点→未知語」のように未知語の遷移確率があるので、元の未知語確率に占めるそれぞれの配分確率値にしたがって「<s>→間投詞」「句読点→間投詞」「間投詞→句読点」の遷移確率値を定める。tri-gram については変更していない。

話題適応させたモデルに対して前述の処理を行い、実際にディクテーションする際に間投詞の部分が認識されているか、また他の部分に悪影響を与えていないかを調べた。

## 5.2 実験結果と考察

次に、話題独立の語彙を 8000 語に固定し、間投詞の特徴を上記のように言語モデルに反映させた。混合重みと開始記号・間投詞に対する確率配分の割合を変化させ、ディクテーションを行った結果を表 7 に示す。間投詞を考慮していない場合と比較して、認識

表 7: 確率配分・混合重みに対する単語認識精度

| 間投詞 | <s> | w=3 | w=30 | w=70 |
|---|---|---|---|---|
| 0.010 | 0.005 | 78.8% | 79.5% | 80.5% |
| 0.010 | 0.001 | 78.6% | 79.4% | 79.8% |
| 0.005 | 0.001 | 78.3% | 79.1% | 79.5% |

話題独立の語彙数：8000 語

率は重み係数が 70 の時に最大 3.3%向上し、80.5%を達成した。また、混合重みが大きな値になるほど認識率が向上した。これはキーワード以外の部分であった間投詞が認識されるようになった結果である。このモデル化による他の部分の認識への悪影響も見られなかったので、本研究で用いた間投詞のモデル化は有効であった。

## 6 まとめ

本研究では、講演音声をディクテーションするために必要な言語モデルを構築するために、話題独立の言語モデルを構築し、その言語モデルを話題適応させる手法を提案した。

実際に研究発表の音声に対して、話題適応の重みと語彙サイズを変化させて認識実験を行った。パープレキシティが最も低かった重み係数 3 の時に認識率 77.5%と最も高くなり、一定の成果を収めた。さらに、間投詞の出現の特徴に基づいてモデル化を行った。間投詞を考慮することにより認識率が 3.3%向上し、最終的な認識率は 80.5%を得た。以上より提案手法の有効性を示すことができた。今後はデータを増やして評価を進めていく予定である。

## 参考文献


[1] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 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-21-10, 1998.

[2] 緒方淳, 西田昌史, 有木康雄. 自動抽出されたアナウンサー発話に対するニュースディクテーションと記事分類. 情報処理学会研究報告, 98-SLP-21-5, 1998.

[3] 赤松裕隆, 廣瀬良文, 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-SLP-21-11, 1998.

[4] 峯松信明, 片岡嘉孝, 中川聖一. 講演調の話し言葉に対する言語的解析. 情報処理学会研究報告, 95-SLP-8-7, 1995.

[5] 西村雅史, 伊東伸康, 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, 98-SLP-20-3, 1998.

[6] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.5, 3 1998.

[7] T.Kawahara and S.Doshita. Topic independent language model for key-phrase detection and verification. In *Proc. IEEE Int'l Conf. Acoust. Speech & Signal Process.*, pp. 685–688, 1999.

[8] 李晃伸, 河原達也, 堂下修司. 単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識. 電子情報通信学会論文誌, Vol. J82-DII, No. 1, pp. 1–9, 1999.

[9] 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. 電子情報通信学会技術研究報告, SP96-81, 1996.

[10] 政瀧浩和, 匂坂芳典, 久木和也, 河原達也. 最大事後確率推定による N-gram 言語モデルのタスク適応. 電子情報通信学会論文誌, Vol. J81-DII, No. 11, pp. 2519–2525, 1998.