話し言葉特有の現象の統計的モデリングに関する研究

2013年 9月

博士(工学)

太田 健吾

豊橋技術科学大学

目次

図目次

表目次

第 1 章

序論

1.1 本研究の背景

近年，音声言語処理の対象は書き言葉から話し言葉へと移行しつつある．たとえば，音声認識の分野では，国会答弁の自動速記のために音声認識が導入されている [4][5][6]．また，日本放送協会では，テレビ番組の字幕生成に音声認識を利用している [7]．このように，話し言葉を対象とした音声認識は社会の様々な場面で求められており，今後もその重要性は高まっていくことが予想される．音声合成や音声対話の分野でも，より親しみやすい話し言葉の合成音声やインタラクションの実現が望まれている．これらの分野において話し言葉を高精度に処理するためには，話し言葉に現れる特有の現象をモデル化する必要がある．本研究では，そのような話し言葉特有の現象を統計的にモデル化する研究を行った．

従来の音声認識は，マンマシンインタフェースでの利用を想定して研究が行われてきた．マンマシンインタフェースの音声認識タスクでは，ユーザは，発話内容を事前に考えた上で，文法的で単純な文を，協力的に発声することを要求する．従って，モデル化が比較的容易であり，また，書き言葉のモデル化のための学習データも大規模に利用可能であった．そこで，大規模な言語データ（コーパス）から音声の音響的・言語的特徴をモデル化する手法が確立された．特に，音素や音節ごとの音響特徴量の分布をモデル化した音響モデルに加えて，単語の連鎖の統計量をモデル化した言語モデルが重要な役割を果たした．しかし，話し言葉の音声認識タスクでは，ユーザは，発話内容を考えながら発声を行うため，非文法的で複雑な文を，不明瞭に発声することが多い．そのため，既存の学習データによる従来のモデル化では実用的な認識精度を達成できなかった．特に，(1) フィラー（「えっと」，「あのー」などの場繋ぎ的に発声される単語）や言い直しが出現すること，(2) 意味的な区切りとは異なる位置に無音が挿入されること，(3) 話し言葉調の言い回しが出現すること，などの話し言葉特有の現象は，既存のデータには含まれないことが多

く，モデル化が困難であった．

この問題に対し，データを収集する方向と，既存のデータを活用する方向の二方向のアプローチが検討された．データを収集する方向として，我が国では，「日本語話し言葉コーパス (CSJ)」が構築された [8]．これは，学会講演や模擬講演を含む約 600 時間の音声を収録し，忠実な書き起こしと言語的なアノテーションを含んでいる．これにより，学会講演の認識において 80% 以上の高い精度を達成することができた．また，CSJ から学習したモデルに対して，話者や話題の適応を加えることで，講義音声の認識において 70% 程度の精度を達成することができた．なお，米国における英語の講義音声の認識でも，同等の精度が達成されている [9]．しかし，精度 70% の音声認識結果は，話題やキーワードは概ね理解できるが，会議録として利用したり，字幕として提示することは難しい．会議録として利用するには 85% 以上，字幕として提示するには 95% 以上の認識精度がそれぞれ必要と言われている [10]．

既存のデータを活用する方向として，対象とする音声とは異なるドメインの話し言葉スタイルの言語モデルと，対象とする音声と同一のドメインの書き言葉スタイルの言語モデルを組み合わせる手法が提案されている．たとえば，T.Hain ら [11] は，会議音声の認識のために，Switchboard コーパス [12] や Fisher コーパス [13]（電話での対話），HUB4 コーパス（放送ニュース）[14] などから構築された様々な言語モデルの N-gram 確率を線形補間して用いている．また，A.Park ら [15] は，講義音声の認識のために，講義テキストや Switchboard コーパスなどの複数のコーパスの N-gram 頻度を重み付き混合して言語モデルの構築を行っている．このように，複数の言語モデルやコーパスを組み合わせることで，話し言葉の N-gram，および認識対象と同一のドメインの N-gram の確率を推定することができる．

特に近年では，Web 上から収集したテキストを利用して，言語モデルのドメインやスタイルを補完する試みが多数行われている．たとえば，Zhu ら [16] は，出現頻度の低い 3-gram の確率を頑健に推定するために，当該 3-gram 自身をクエリとして Web 検索を行い，収集された Web テキストにおける 3-gram 頻度を利用している．また，Bulyko ら [17] は，話し言葉コーパス中で出現頻度の高い 3-gram をクエリとして Web 検索を行い，収集されたテキストをスタイルの補完に利用している．さらに，認識対象と同一ドメインの少量の話し言葉コーパスからキーワードを含む N-gram を抽出し，これをクエリとした Web 検索によってテキストを収集・利用することで，ドメインの補完も行っている．一方で，翠ら [18] は，対話システムの音声認識タスクにおいて，認識対象と同一ドメインの話し言葉コーパスを必要とせずに適切な Web テキストを収集する手法を提案している．具体的には，対話システムの知識ベースから抽出したキーワードを用いた Web 検索によって認識対象とドメインの一致した Web テキストを収集する．次に，別ドメインの音声対話コーパスと知識ベースの混合コーパスから作成した言語モデルを用いて，これら

の Web テキストの中から認識対象とスタイルの一致したテキストを抽出する．これにより，ドメイン・スタイル共に認識対象とマッチしたテキストを獲得している．

言語モデルの教師なし適応やスタイル変換に基づく手法も提案されている．南條ら [19] は，認識結果を適応データとして言語モデルを特定話者の発話スタイルに適応させる教師なし話者適応を提案している．また，秋田ら [20] は，まったく同一の内容を対象とした書き言葉と話し言葉のパラレルコーパスから，書き言葉を話し言葉に変換する統計的なモデルを学習し，書き言葉言語モデルを話し言葉言語モデルに変換する手法を提案している．渡邊ら [21] は，学会講演の音声認識において，秋田らの手法とルールベースの話し言葉変換手法 [22] を組み合わせることにより，書き言葉スタイルの予稿テキストから話し言葉スタイルの言語モデルを構築する手法を提案している．

1.2 本研究の目的

これらの背景を踏まえ，本研究では，話し言葉特有の現象を統計的にモデル化することにより，話し言葉を対象とした音声言語処理の高度化を行うことを目的とする．そのために，まず，話し言葉コーパスと書き言葉コーパスに対する分析により，フィラーや言い淀み，倒置といった話し言葉特有の現象や，音声認識における入力音声の認識処理単位と，音声認識用言語モデルの学習用コーパスにおける文単位との不一致といった問題について調査・分析した．そして，これらの分析に基づき，個々の問題を考慮した話し言葉特有の諸現象のモデル化について検討し，話し言葉の音声認識や音声対話に適用した．具体的には，それぞれ以下の目的に取り組んだ．

(a) フィラーのモデル化による話し言葉音声認識の高精度化

まず，話し言葉特有の現象の中でも最も発生頻度の高い現象であるフィラーに注目した [23]．フィラーを考慮した音声認識に関する検討としては，たとえば，フィラーを後続の単語の予測に影響を与えない透過単語として扱う手法が様々なタスク・ドメインで試みられている [24][25][26]．また，稲垣ら [27] は，韻律的特徴を利用したフィラー検出器を音声認識器に組み込むことで，話し言葉音声に対する認識率の改善を図っている．しかしこれらは基本的に，認識対象のドメインを含み，かつ，フィラーにも対応した話し言葉言語モデルが利用できることを前提としたものである．そして，前述の通り，このような言語モデルをいかに獲得するかの方が実用的にはより重要な問題である．従って，実際に利用が可能な既存のコーパスから，フィラーに対応した話し言葉言語モデルを構築する手法を検討する必要がある．

そこで本研究では，フィラーのタグ付けを含まないコーパスから，こうした言語モデルを構築する手法について検討し，話し言葉音声認識の高精度化に取り組んだ．具体的に

は，対象とする音声とは異なるドメインの正確な話し言葉コーパスからフィラー予測モデルを学習し，この予測モデルに基づいて，対象とする音声と同一のドメインの，フィラーのタグ付けなどがない話し言葉コーパスに対してフィラーの復元を行い，フィラーが復元されたコーパスから言語モデルを学習するという手法を提案した．本手法を国会会議録を用いた実験によって評価した結果，従来手法と比べ単語認識精度を 6.6 %改善することができた．すなわち，フィラーをモデル化することにより，既存のコーパスからフィラーを考慮した言語モデルを構築し，話し言葉の音声認識を高精度化することができた．

(b) フィラーのモデル化による話し言葉音声対話の高度化

次に，音声対話におけるフィラーの役割にも注目した．たとえば，Watanabe[28] らは，フィラーやポーズの時間長が，後続するフレーズの複雑さを予測する手がかりになることを示している．また，Somiya[29] らは，講義音声中のフィラーが受講者の感じる聞き易さや理解のし易さに影響を与えることを示している．これらの先行研究にも見られるように，フィラーやポーズは，聞き手の感じる聞き易さや理解のし易さに影響を与える．こうしたフィラーやポーズの効果は，音声対話システムを実装する際にも考慮すべきである．

音声対話におけるフィラーやポーズの影響に関する検討として，伊藤 [30] らは，音声対話システムにおける応答文間のフィラーについて分析を行なっている．伊藤らの分析によると，システムが次文の生成に時間を要する場合に，システムが動作していることをユーザに示すためのサインとして，応答文間でフィラーを発声させることが有効である．また，Shiwa[31] らは，人間とロボットの対話において同様の結論を示している．ユーザインタフェースの分野では，”2 秒ルール”という知見がよく知られている．すなわち，システムは，ユーザの入力を受けてから応答を返すまでに 2 秒以上を要するべきではない [32][33]．しかし，これらの先行研究は，応答文間のフィラーやポーズを対象としており，応答文内に出現するフィラーやポーズの影響を考慮していない．人間の音声対話においては，聞き手の理解度や聞き易さを考慮し，文内にフィラーやポーズを発生させることが少なくない．従って，より自然で高度な音声対話システムを実現するためには，文間だけでなく，文内のフィラーやポーズの影響も考慮する必要がある．

そこで本研究では，特に文内のフィラーやポーズに注目し，音声対話システムの応答文内にフィラーを挿入することで，応答音声の自然性や聞き易さを向上させることについて検討した．特に，文頭や文節境界にランダムにフィラーを挿入するのではなく，文脈等を考慮してより適切な位置に挿入することで，自然性や聞き易さ, 理解度等の改善を図った．すなわち，フィラーをモデル化することにより，話し言葉による音声対話システムの高度化を図った．

(c) ポーズのモデル化による話し言葉音声認識の高精度化

続いて，ポーズに基づく話し言葉の認識処理単位の問題に注目した．音声認識時における入力音声の認識処理単位や，言語モデルの学習に用いるコーパスの文単位は，音声認識上必ずしも最適な区切りではない．音声認識においては，入力音声に対応する文字列を探索する際のエラーが少なく，かつ，実用的な時間で探索が終了するような認識処理単位が理想的であり，具体的には，統語的・意味的なまとまりである文単位や，話者の意図の上で 1 つのまとまりであるような発話単位が適格であると考えられる．しかし，このような文単位・発話単位は，入力音声から観測することが不可能な単位であり，その予測も容易ではない [34][35][36]．そこで，現在の音声認識では，パワーや零交差数などの情報に基づいて区切られた単位を認識処理単位とし，それぞれの単位が独立であると仮定して認識を行うのが一般的である．しかし，実際にはこのようにして得られた認識処理単位は，独立であるとは限らない．こうした状況で，南條ら [37] は，文区切りのショートポーズの閾値と認識率の関係について調査しており，日本語話し言葉コーパスを用いた認識実験において，無音区間をすべてショートポーズとして扱った場合，500msec 未満の無音区間をショートポーズとして扱った場合，1000msec 未満の無音区間をショートポーズとして扱った場合をそれぞれ比較し，1000msec 未満の無音区間をショートポーズとして扱った場合に最も認識率が良くなるという結果を得ている．また，寺尾ら [38] や上西ら [39] は，F0 の統計的モデルに基づいて韻律句境界の推定を行い，韻律句境界を跨ぐ場合と跨がない場合とで言語モデルを使い分ける手法を提案している．同様に，鄭ら [40] や細田ら [41] は，品詞情報に基づいて文節境界の推定を行い，文節境界を跨ぐ場合と跨がない場合とで言語モデルを使い分ける手法を提案している．しかし，これらの手法では，文節や韻律句の同定が必要である．従って，文節や韻律句などの言語情報ではなく，検出の容易なポーズ情報の観点から，認識処理単位の扱い方について検討する必要がある．

そこで本研究では，話し言葉におけるポーズに基づく認識処理単位の問題を考慮することにより，話し言葉音声認識の高精度化を行った．具体的には，音声認識時における入力音声の認識処理単位 (認識対象区間) や，言語モデルの学習に用いるコーパスの文単位の独立性についてパープレキシティの観点から検証を行い，直前の単位の単語情報を，後続する単位の先頭単語の認識に利用することが有効であることを日本語話し言葉コーパス，国会会議録，および毎日新聞を用いた実験によって確認した．

また，読み上げ音声を対象とする場合には，コーパス中の句読点をポーズに対応付けることができたが，話し言葉の音声を対象とする場合には，実際の音声中に発生するポーズは必ずしも句読点に対応しない．従って，話し言葉の音声認識を高精度に行うためには，句読点を単純にポーズに対応づけるのではなく，実際のポーズの出現パターンをモデル化する必要がある．

そこで本研究では，先のフィラー予測モデルと同様の方法を用いてコーパスにポーズ

情報の復元を行い，ポーズ情報が復元されたコーパスから言語モデルを学習するという手法を提案した．本手法を国会会議録を用いた実験によって評価した結果，従来手法と比べ単語認識精度を 4.5 %改善することができた．すなわち，ポーズをモデル化することにより，話し言葉の音声認識を高精度化することができた．なお，書き言葉コーパスにフィラーやポーズを挿入して話し言葉コーパスを作成する我々の提案手法は増村ら [42] によって引用され，その効果が実証されている．

(d) 整形された会議録を用いた音響モデリングによる話し言葉音声認識の高精度化

最後に，話し言葉の音声認識のモデルを学習するために利用可能なコーパスとして，不正確な書き起こしに注目した．話し言葉を対象とする音声認識システムには，対象音声とドメインが一致し，かつ，話し言葉特有の言い回しに対応した言語モデルや音響モデルが不可欠である．そのようなモデルを構築する最も単純な方法は，対象音声と同一ドメインの大規模な話し言葉コーパスからモデルを学習するという方法である．しかし，話し言葉を正確に書き起こす作業は極めて高いコストを必要とするため，あらゆるドメインに対して，そのようなコーパスが入手できると仮定することは現実的ではない．

これに対し，速記録や会議録は，正確な書き起こしより広く作成されており，比較的容易に入手が可能である．ただし，速記録や会議録では，可読性を高めるために，フィラーや言い淀み，言い直しなどの話し言葉特有の現象は削除され，話し言葉特有の言い回しは適切な書き言葉に置き換えられていることが一般的である．また，近年では，クラウドソーシングにより，非熟練者による書き起こしを安価に収集する取組みが広く行われている．たとえば，Williams らは，多数の非熟練作業者を利用して書き起こしを作成する方法を提案している [43]．この方法は，熟練作業者による書き起こしに比べて，作業コストの面では非常に安価ではあるものの，多数の書き起こしミスが含まれる欠点がある．従って，このような整形された書き起こしやミスを含む書き起こしから，正確に書き起こされた箇所を検出することができれば，話し言葉の音声認識のモデルを学習する上で有用である．

そこで本研究では，整形された書き起こしから正確に書き起こされた箇所を自動検出する手法を提案し，検出された箇所をモデルの学習に用いることで，話し言葉の音声認識の高精度化を行った．提案法は 2 段階からなる．第 1 に，整形された書き起こしと原音声のアラインメントを行い，第 2 に，アラインメントによって得られた素性に基づく SVM を用いて，整形箇所を検出する．国会会議録を用いた評価実験により，提案法は，整形された書き起こしと正確な書き起こしのパラレルコーパスをできるだけ少ないコストで用意するための目標性能を達成できることを示した．また，提案法によって取り出された非整形箇所は，話者適応の発音ラベルとして有効であることを示した．すなわち，既存の整形された書き起こしを利用することにより，話し言葉の音声認識を高精度化することがで

きた.

1.3 本論文の構成

本論文の構成は以下のようになっている.

2 章では，音声認識の基本原理と，音声認識で一般に用いられる音響モデルと言語モデル，および音声認識性能の評価方法について説明する.

3 章では，話し言葉コーパスや書き言葉コーパスを用いて，フィラーや言い淀み，倒置といった話し言葉特有の現象や，音声認識における認識処理単位 (認識対象区間) と，コーパスの文単位との不一致といった問題について調査・分析した結果を述べる.

4 章では，フィラーを考慮した言語モデルの構築法として，フィラー予測モデルに基づく手法を提案し，日本語話し言葉コーパスおよび国会会議録を対象とした実験によって従来手法との比較結果を述べる.

5 章では，音声対話システムの応答音声にフィラーを挿入することで，応答音声の自然性や聞き易さを向上させることについて検討した結果を述べる．特に，文頭や文節境界にランダムにフィラーを挿入するのではなく，文脈等を考慮してより適切な位置に挿入することで，自然性や聞き易さ, 理解度等の改善を図った結果を述べる.

6 章では，音声認識における認識処理単位やコーパスの文単位が必ずしも音声認識上最適な区切りではないことを考慮し，直前の単位の単語を，後続する単位の単語の認識に利用する手法を述べる．さらに，認識処理単位とコーパスの文単位の不一致を考慮し，コーパス中にポーズを挿入する手法を述べる.

7 章では，整形された書き起こしから整形箇所を自動検出する手法を述べる．本手法は，2 つのステップからなる．第 1 に，整形された書き起こしとその原音声とでアラインメントを行い，第 2 に，アラインメントによって得られた素性に基づく Support Vector Machine (SVM) を用いて，整形箇所を検出する．また，整形箇所以外を非整形箇所としてそのラベルを用いて音響モデルを適応した結果について述べる.

8 章では，本研究のまとめと，今後の課題について述べる.

第 2 章

音声認識概論

2.1 はじめに

本章では，音声認識の概要について述べる．現在の音声認識は，大規模な音声データに基づいて音声の物理的性質をモデル化する音響モデルと，大規模な言語データに基づいて単語の繋がり易さをモデル化する言語モデルが重要な役割を果たしている．まず，2.2 節では，これらのモデルを用いた音声認識の基本原理について述べる．次に，2.3 節では，現在主流となっている隠れマルコフモデルに基づく音響モデルについて説明する．また，2.4 節では，統計的言語モデルについて概説する．さらに，2.5 節では，音声認識の性能の評価尺度について述べる．

2.2 音声認識の基本原理

音声認識は，与えられた音声 X に対して，事後確率 $P(W|X)$ が最大となる単語列 $\hat{W}$ をみつける問題として，以下のように定式化される．

$$\hat{W} = argmax_W P(W|X) \tag{2.1}$$

ただし，この確率モデル $P(W|X)$ を直接推定するのは困難であるため，ベイズ則により，音響モデルの確率 $P(X|W)$ と言語モデルの確率 $P(W)$ に分解する．

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \tag{2.2}$$

ここで，右辺の分母 $P(X)$ は，W の決定には影響しないため無視することができる．従って，式 (2.1) は以下のように書き換えられる．

$$\hat{W} = argmax_W P(X|W)P(W) \tag{2.3}$$

一般には，さらに対数をとり，若干の修正を行った下式が音声認識の尤度として用いられる．

$$\hat{W} = argmax_W logP(X|W) + \alpha logP(W) + \beta|W| \tag{2.4}$$

音響モデルの確率 $P(X|W)$ は，単語列 W を構成する音素や音節に対する音声パターンの分布モデルとの照合により得られる確率の累積として求めることができる．言語モデルの確率 $P(W)$ は，単語列 W がどの程度尤もらしいかを表す尺度である．α は言語モデル重みと呼ばれるパラメータであり，音響モデルのスコアと言語モデルのスコアのレンジの違いを調整する役割を持つ．β は挿入ペナルティと呼ばれ，単語列の長さ $|W|$ に応じてスコアを調整する．これにより，β が正の値の場合には，認識結果に単語数が多いほど有利になり，β が負の値の場合には，単語数が少ないほど有利になる．

2.3 音響モデル

音響モデルは，音声の特徴量パターンの分布の統計モデルである．時系列パターンとしての音声をモデル化するために，確率的非決定性オートマトンである隠れマルコフモデル (Hidden Markov Model;HMM) が一般に用いられる．音響モデルは，音素や音節といった単位で構成されるのが一般的である．ただし，同一の音素や音節でも，直前あるいは直後の文脈の影響を受けて変形するため，十分な学習データが獲得できる場合には，直前後の文脈を考慮した文脈依存モデルが用いられる．

2.3.1 音声特徴パラメータ

音声の特徴量としては，パラメトリックな手法として，線形予測分析 (Linear Prediction)[44] や知覚線形予測分析 (Perceptual Linear Predictive analysis;PLP)[45]，ノンパラメトリックな手法として，ケプストラム係数やメルケプストラム係数 (Mel-Frequency Cepstrum Coefficients;MFCC) が用いられている．また，MFCC のような振幅スペクトル情報に加えて，位相情報を併用する手法も提案されている [46]．このほか，多層パーセプトロン (Multi Layer Perceptron;MLP) を用いた手法である Tandem[47] や TRAPs[48]，音声の構造的表象を用いた手法 [49]，発声者の発音動作の特徴量（Articulatory Feature;AF）を用いた手法 [50] なども提案されている．本節では，現在最も主流となっているメルケプストラム係数について述べる．

音声をディジタル処理するためには，まず，音声のアナログ信号を，サンプリング定理に従ってサンプリングし，A/D 変換器を用いてディジタル信号に変換する必要がある．たとえば音声認識では，サンプリング周波数は 16kHz，量子化ビット数は 16bit が一般に

用いられる．

続いて，前処理として，高域強調処理および窓掛け処理を行う．音声信号の成分は低域部分に偏っているため，1 次差分フィルタを用いて高域部分を強調することが一般的である．これをプリエンファシスと呼ぶ．プリエンファシスでは，時刻 t のサンプル値を s_t，高域強調後のサンプル値を s'_t，高域強調係数を α として，

$$s'_t = s_t - \alpha s_{t-1} \tag{2.5}$$

とする．高域強調係数としては，一般に $\alpha = 0.97$ が用いられる．

高域強調された音声は，一定区間 (フレーム) ずつに切り出されて処理される．これは，音声は全体として非定常であるが，短い区間では定常であるという仮定の基いた処理である．この時，切り出されたフレームに窓関数を乗ずる．これにより，切り出し区間の始点および終点において急峻な変化が起こらないようにする．窓関数は，一般に図 2.1 のように無数の峰を持ち，中央の峰をメインローブ，他の峰をサイドローブと呼ぶ．

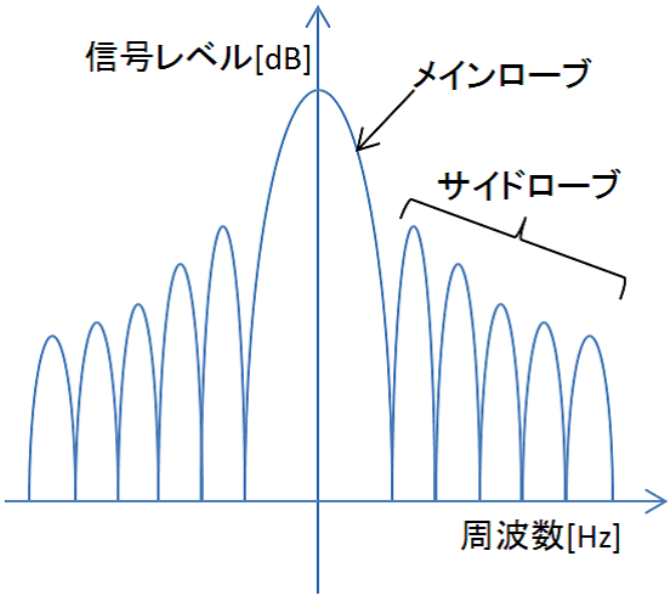


図 2.1　窓関数の例

窓関数としては，次のような特性を持つものが望ましいとされる [51].

- 周波数分解能が高い (メインローブが狭い).
- 他の周波数成分から生ずるスペクトルの漏れが少ない (サイドローブの減衰が大きい).

これらのことから，次式のハミング窓が一般に用いられる．

$$w_n = 0.54 - 0.46cos(\frac{2\pi n}{N-1}) \ (n = 0, 1, 2, ..., N-1) \tag{2.6}$$

フレームの長さとしては 25msec，フレームを移動させる周期 (フレームシフト) としては 10msec が一般に用いられる．窓関数を乗じたことによる情報損失を補うためには，フレームシフトはフレーム長の半分程度とし，分析区間をオーバーラップさせることが有効である．

前処理によって N 点の音声波形を切り出した後，フーリエ変換を行う．得られた N 点の振幅スペクトルに対して，周波数軸に等間隔に配置した L 個の帯域フィルタ (三角窓) を用いて，フィルタバンク分析を行う．これにより，窓の幅に対応する周波数帯域の信号のパワーを，単一スペクトルチャンネルの振幅スペクトル $|S'(k)|$ の重み付け和で求める．

$$m(l) = \sum_{k=lo}^{hi} W(k;l)|S'(k)| \ (l = 1, ..., L) \tag{2.7}$$

$$W(k;l) = \begin{cases} \frac{k-k_lo(l)}{k_c(l)-k_lo(l)} & \{k_lo(l) \leq k \leq k_c(l)\} \\ \frac{k_hi(l)-k}{k_hi(l)-k_c(l)} & \{k_c(l) \leq k \leq k_hi(l)\} \end{cases}, \tag{2.8}$$

ただし，$k_lo(l)$，$k_c(l)$，$k_hi(l)$ は，それぞれ l 番目のフィルタの下限，中心，上限のスペクトルチャネル番号であり，隣り合うフィルタ間で，

$$k_c(l) = k_hi(l-1) = k_lo(l+1) \tag{2.9}$$

なる関係がある．さらに，$k_c(l)$ は，メル周波数軸上で等間隔に配置される．メル周波数は，

$$Mel(f) = 2569log_{10}(1 + \frac{f}{700}) \tag{2.10}$$

により計算される．ここで，f の単位は $[Hz]$ にとる．最終的に，フィルタバンク分析によって得られた L 個の帯域におけるパワーを離散コサイン変換することによって，メルケプストラム係数が得られる．

$$c_{mfcc}(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} log\ m(l) cos(l - \frac{1}{2}) \frac{i\pi}{L} \tag{2.11}$$

ここで，フィルタの数を $L = 24$ とし，スペクトル包絡の概形情報として，低次の 12 次元のベクトルの MFCC を用いることが一般的である．

以上に述べた MFCC は，フレーム内の音声区間を定常とみなしたうえで得られる静的な特徴である．しかし，音声の音響的な特徴は，前後の文脈に影響を受けて変化することが知られており，音素から音素への渡りや，音節から音節への渡りの部分では，スペクトル特徴が時間とともに連続的に変化する．また，/r/ や /w/，/y/ といった半母音は，スペクトルの動きそのものに，音素の音響的な特徴が表現されている．これらのことから，

フレーム分析によって得られた静的な特徴に加えて，時間変化を考慮した動的な特徴を用いることで，音声認識の精度が改善することが知られている．

動的特徴量としては，たとえば，次式のような差分による近似が用いられる．

$$\triangle c(n;l) = c(n;l+K) - c(n;l-K) \tag{2.12}$$

また，次式の回帰係数も用いられる．

$$\triangle c(n;l) = \frac{\sum_{k=-K}^{K} k \cdot c(n;l+k)}{\sum_{k=-K}^{K} k^2} \tag{2.13}$$

これらの △ 特徴量に加え，△△ と呼ばれる，特徴量の 2 次の微分係数が用いられることも多い．

2.3.2　隠れマルコフモデルによる定式化

隠れマルコフモデルに基づく音響モデルにおいては，単語列を構成する音素や音節の系列 y に対し，音響特徴量の系列 x の出力確率 $P(x|y)$ を次式で定義する．

$$P(x|y) = \sum_{z} P(x,z|y) \tag{2.14}$$

$$= \sum_{z \in Z(y)} P(x|z)P(z) \tag{2.15}$$

$Z(y)$ は，モデルの構造により与えられる y が取りうる状態系列の集合を表す．

ここで，$P(x|z)$ は，フレーム間の音響特徴量の独立性の仮定により，次式のように近似できる．

$$P(x|z) \approx \prod_{t} P(x_t|z_t) \tag{2.16}$$

ある状態 z から音響特徴量 x(一般にはベクトル表現) が出力される確率 $P(x|z)$ は，音声の変動を考慮して，一般に混合正規分布 (Gaussian Mixture Model) を用いて次式のようにモデル化される．

$$P(x|z) = \sum_{m} w_{z,m} N(x;\mu_{z,m},\sigma_{z,m}) \tag{2.17}$$

$$N(x;\mu_{z,m},\sigma_{z,m}) = \frac{1}{\sqrt{2\pi|\sigma|^{-1}}} exp(-\frac{1}{2}(x-\mu_{z,m})^{\top}\sigma^{-1}(x-\mu_{z,m})) \tag{2.18}$$

$w_{z,m}$ は状態 z における正規分布 m の重みであり，この正規分布 m の平均が $\mu_{z,m}$，共分散行列が $\sigma_{z,m}$ でそれぞれ表される．

一方で，$P(z)$ は，マルコフ性の仮定により，次式のように近似される．

$$P(z) \approx \prod_{t=2}^{T} P(z_t|z_{t-1}) \tag{2.19}$$

ある状態 z' から状態 z へ遷移する確率 $P(z|z')$ は，離散分布を用いて次式のように定義される．

$$P(z|z') = \alpha_{z',z} \tag{2.20}$$

ここで，$\alpha_{z',z}$ は $\sum_z \alpha_{z',z} = 1$ を満たす．

2.3.3 パラメータ推定法 [1]

音響モデルのパラメータ推定の方法としては，最尤基準に基づく方法，相互情報量最大化基準に基づく方法，音素誤り率最小化基準に基づく方法などが提案されている．本節では，最尤基準に基づく方法について述べる．

最尤基準 (Maximum Likelihood Estimation;MLE) に基づく方法では, 学習データに対する対数尤度の和を最大化するようにパラメータを推定する．学習データを $D = \{x^i, y^i\}$ とするとき，最尤基準による目的関数は次式のようになる．

$$l(\lambda_{HMM}; D) \triangleq \sum_{d \in D} logP(x^d|y^d, \lambda_{HMM}) \tag{2.21}$$

$\lambda_{HMM} = \{a_{z',z}, w_{z,m}, \mu_{z,m}, \Sigma_{z,m}\}$ は，推定対象である HMM パラメータの集合である．ここで，式 (2.21) を変形すると，以下のようになる．最後の行は，Jensen の不等式による．

$$\sum_{d \in D} logP(x^d|y^d, \lambda_{HMM}) = \sum_{d \in D} log \sum_{z \in Z(y^d)} logP(x^d, z|y^d, \lambda_{HMM}) \tag{2.22}$$

$$= \sum_{d \in D} log \sum_{z \in Z(y^d)} P(z|x^d, y^d, \lambda'_{HMM}) \frac{P(x^d, z|y^d, \lambda_{HMM})}{P(z|x^d, y^d, \lambda'_{HMM})} \tag{2.23}$$

$$\geq \sum_{d \in D} \sum_{z \in Z(y^d)} P(z|x^d, y^d, \lambda'_{HMM}) log \frac{P(x^d, z|y^d, \lambda_{HMM})}{P(z|x^d, y^d, \lambda'_{HMM})} \tag{2.24}$$

ここで，クロスエントロピーの性質より，以下の関係が常に成り立つ．

$$\sum_{z \in Z(y^d)} P(z|x^d, y^d, \lambda'_{HMM}) logP(z|x^d, y^d, \lambda_{HMM}) \tag{2.25}$$

$$\leq \sum_{z \in Z(y^d)} P(z|x^d, y^d, \lambda'_{HMM}) log P(z|x^d, y^d, \lambda'_{HMM}) \tag{2.26}$$

以下の Q 関数を λ_{HMM} に関して最大化すればよい．これは，Baum-Welch のアルゴリズムによって効率的に実現できる．なお，この Q 関数の最大化は，式 (2.21) の下限を最大化することに相当する．

$$Q(\lambda_{HMM}, \lambda'_{HMM}; D) = \sum_{d \in D} \sum_{z} P(z|x^d, y^d, \lambda'_{HMM}) log P(x^d, z|y^d, \lambda_{HMM}) \tag{2.27}$$

2.3.4　音響モデルの学習

音響モデルの学習には，大規模な音声データが用いられる．このとき，学習データと認識対象の間で，入力環境 (帯域，周囲雑音) や話者の性質 (性別，年齢) が合致していることが非常に重要である．ただし，あらゆる環境や話者に対して学習データを獲得することは非現実的であるため，ベースラインのモデルから環境適応や話者適応を行うことが考えられる．モデルを適応する方法としては，次節で述べる最大事後確率推定法 (MAP 推定)[52] がある．

(a) 最大事後確率推定法 (MAP)

最大事後確率推定法 (MAP 推定)[53][54] は，Bayesian Succesive Estimation [55][56] とも呼ばれ，逐次的な教師あり学習手法である．そのため，1 つの学習サンプルが与えられるたびに，事後確率が最大となるようなパラメータ θ を推定する．例として，$X_1, \cdots, X_N$ の N 個の学習サンプルが与えられた場合の事後確率を次式に示す．

$$P(\theta|X_1, \cdots, X_N) = \frac{P(X_N|X_1, \cdots, X_{N-1}, \theta) P(\theta|X_1, \cdots, X_{N-1})}{\int P(X_N|X_1, \cdots, X_{N-1}, \theta) P(\theta|X_1, \cdots, X_{N-1}) d\theta} \tag{2.28}$$

以下の節では，MAP 推定を用いた多次元正規分布の平均ベクトルと共分散行列を学習する方法について説明する．

(b) 平均ベクトルの学習 [57]

MAP 推定によって多次元正規分布の平均ベクトルを学習する方法について述べる．推定対象のパラメータは平均ベクトル μ であることから，式 (2.28) において，

$$\theta = \mu \tag{2.29}$$

とする．

次に，1 個の学習サンプル X_1 は，$N(\mu, \Sigma)$ の正規分布に従うと仮定する．ここで，Σ はベースラインモデルの共分散行列であり，ここでは既知とする．

$$P(X_1|\mu) \simeq N(\mu, \Sigma) \tag{2.30}$$

また，推定対象の平均ベクトル μ について事前分布を仮定する．μ は，最も確からしい平均ベクトル μ_0，不確かさを表す共分散行列 K_0 の正規分布に従うと仮定する．

$$P(\mu) \simeq N(\mu_0, K_0) \tag{2.31}$$

以上の定義を，式 (2.28) に適用すると，

$$\begin{aligned} P(\mu|X_1) &= \frac{P(X_1|\mu)P(\mu)}{\int P(X_1|\mu)P(\mu)d\mu} \simeq N(\mu_1, K_1) \\ &= Cexp(-\frac{1}{2}(X_1-\mu)^T\Sigma^{-1}(X_1-\mu) - \frac{1}{2}(\mu-\mu_0)^T K_0^{-1}(\mu-\mu_0)) \end{aligned} \tag{2.32}$$

推定された平均ベクトル $\hat{\mu_1}$ と不確かさ $\hat{K_1}$ は次式のようになる．

$$\begin{aligned} \hat{\mu_1} &= K_0(K_0+\Sigma)^{-1}X_1 + \Sigma(K_0+\Sigma)^{-1}\mu_0 \\ \hat{K_1} &= K_0(K_0+\Sigma)^{-1}\Sigma \end{aligned} \tag{2.33}$$

K_0 は推定前に μ がどの程度の不確かさを持つかを表す共分散行列である．ここでは，適応化パラメータ α を導入し，実験的に求める．

$$K_0 = \alpha^{-1}\Sigma \tag{2.34}$$

ここで，α を 0 に近づければ K_0 は大きくなり，μ の不確かさが大きいことを仮定することになる．逆に，α を非常に大きな値にすれば K_0 は小さくなり，μ の不確かさが小さいことを仮定することになる．式 (2.33) より，

$$\hat{\mu_1} = \frac{\alpha\mu_0 + X_1}{\alpha+1} \tag{2.35}$$

N 個の学習サンプルを繰り返し与えた後の推定値は次式のようになる．

$$\hat{\mu_N} = \frac{(\alpha+N-1)\mu_{N-1} + X_N}{\alpha+N} = \frac{\alpha\mu_0 + \sum_{i=1}^{N} X_i}{\alpha+N} \tag{2.36}$$

これは，常に推定前の平均ベクトルと現在与えたサンプルの間で，サンプル数で重み付けされた線形補間に対応している．

α はすべての音節カテゴリの各状態で同一の値を用い，実験的に求める．また混合分布の場合は，α を各要素分布ごとに同一の値を用いる．

(c) 平均ベクトルと共分散行列の同時学習 [57]

平均ベクトルと共分散行列の同時学習では，推定対象のパラメータが 2 組あるため，仮定する事前分布と事後確率は同時分布となる．1 個および N 個の学習サンプルによって推定された共分散行列の推定値は次式のようになる．

$$\hat{\Sigma_1} = \frac{X_i X_i^T - (\alpha+1)\mu_1\mu_1^T + \beta K_0 + \alpha\mu_0\mu_0^T}{\beta+1} \tag{2.37}$$

$$\begin{aligned}\hat{\Sigma_1} &= \frac{1}{\beta+N}(\sum_{i=1}^{N} X_i X_i^T - (\alpha+N)\mu_N\mu_N^T + \beta K_0 + \alpha\mu_0\mu_0^T) \\ &= \frac{1}{\beta+N}(X_N X_N^T - (\alpha+N)\mu_N\mu_N^T \\ &\quad +(\beta+N-1)\Sigma_{N-1} + (\alpha+N-1)\mu_{N-1}\mu_{N-1}^T)\end{aligned} \tag{2.38}$$

平均ベクトルの推定値は式 (2.36) と同様である．

2.4　言語モデル

計算機で自然言語を扱うにあたっては，様々なモデル化の方法が考えられてきた．近年では, 言語モデルを, 与えられた単語列 $w_1w_2\cdots w_n$ に対し, その出現確率 $P(w_1w_2\cdots w_n)$ を与える確率モデルとして考え，大量のサンプルデータを用いた統計的な手法によって確率の推定を行うのが主流となっている（統計的言語モデル）．

こうした統計的言語モデルにも様々なものがあるが，音声認識の分野では，確率付き文脈自由文法や N-gram モデルなどが主に用いられてきた．

特に，地名や人名の認識といった単純なタスクに対しては確率付き文脈自由文法がよく用いられるのに対し，本研究で扱うような大語彙を対象とした話し言葉の認識に対しては，N-gram モデルを用いるのが一般的となっている．

2.4.1　N-gram 言語モデル

N-gram モデルでは，与えられた単語列 $w_1w_2\cdots w_n$ に対する確率 $P(w_1w_2\cdots w_n)$ を，次式のようにして推定する．

$$P(w_1w_2\cdots w_n) = \prod_{i=1}^{n} P(w_i|w_{i-N+1}\cdots w_{i-1}) \tag{2.39}$$

つまり，i 番目の単語 w_i の生起確率が，直前の $N-1$ 単語にのみ依存し，それ以外の単語からはまったく影響されないものと考えて確率推定を行う．ここで，特に N=1 のモ

デルをユニグラムモデル，N=2 のモデルをバイグラムモデル，N=3 のモデルをトライグラムモデルという．

N グラム確率 $P(w_i|w_{i-N+1}\cdots w_{i-1})$ の推定は，大量のテキストデータ（コーパス）からの最尤推定によって行うのが一般的である．たとえば，トライグラム確率は次式で推定することができる．ここで，$w_{i-2}w_{i-1}$ を w_{i-2}^{i-1} と表すものとする．また，$N(w_{i-2}^{i-1})$ は単語列 w_{i-2}^{i-1} が学習コーパスに出現した頻度であるとする．

$$P(w_i|w_{i-2}^{i-1}) = \frac{N(w_{i-2}^{i})}{N(w_{i-2}^{i-1})} \tag{2.40}$$

しかし，このような単純な推定では，学習コーパスにたまたま出現しなかった単語列に対する出現確率が 0 になってしまう．そこで，これを回避するために，次節で述べるスムージングが適用される．

2.4.2 N-gram 言語モデルのスムージング [2]

N-gram モデルのスムージング手法としては，バックオフ・スムージングが一般的である．これは，学習コーパスに出現しなかった N-gram 確率を，(N-1)-gram 確率から推定する方法である．たとえば，学習コーパスに出現しなかったトライグラムの確率は，次式のようにバイグラム確率を用いて推定する．

$$P(w_i|w_{i-2}^{i-1}) = \begin{cases} \lambda(w_{i-2}^{i-1})f(w_i|_{i-2}^{i-1}) & \text{if } N(w_{i-2}^{i}) > 0 \\ (1-\lambda_0(w_{i-2}^{i-1}))\alpha P(w_i|w_{i-1}) & \text{else if } N(w_{i-2}^{i-1}) > 0 \\ P(w_i|w_{i-1}) & \text{otherwise} \end{cases}, \tag{2.41}$$

ここで，$f(w_i|_{i-2}^{i-1})$ は最尤推定によって求めたトライグラム確率である．

$$f(w_i|_{i-2}^{i-1}) = \frac{N(w_{i-2}^{i})}{N(w_{i-2}^{i-1})} \tag{2.42}$$

λ はディスカウントと呼ばれる係数であり，学習コーパスに出現した N-gram の確率を割り引いて，学習コーパスに出現しなかった N-gram の確率へと割り当てる．また，λ_0 は次式のようにして求められる．

$$\lambda_0(w_{i-2}^{i-1}) = \sum_{w} \lambda(w_{i-2}^{i-1})f(w|w_{i-2}^{i-1}) \tag{2.43}$$

α は，確率の総和を 1 にするための正規化係数であり，次式で求められる．

$$\alpha = \left(1 - \sum_{N(w_i^{i-2})>0} P(w_i|w_{i-1})\right)^{-1} \tag{2.44}$$

ディスカウント量の決め方についても様々な方法が考案されているが，近年では次式の Witten-Bell 法 [58] がよく用いられる．

$$P(w_i|w_{i-2}^{i-1}) = \begin{cases} \frac{N(w_i^{i-2})}{N(w_{i-1}^{i-2})+R(w_{i-1}^{i-2})} & \text{if } N(w_{i-2}^{i}) > 0 \\ \frac{R(w_{i-1}^{i-2})}{N(w_{i-1}^{i-2})+R(w_{i-1}^{i-2})} \alpha P(w_i|w_{i-1}) & \text{else if } N(w_{i-2}^{i-1}) > 0 \\ P(w_i|w_{i-1}) & \text{otherwise} \end{cases} , \tag{2.45}$$

ここで，$N(w_i^{i-2})$ は単語列 w_i^{i-2} が学習コーパスに出現した回数であり，$R(w_{i-1}^{i-2})$ は単語列 w_{i-1}^{i-2} の後に出現した単語の種類数である．

2.4.3　言語モデルの評価尺度

言語モデルの評価には，評価用テキストに対する単語パープレキシティがよく用いられる．単語パープレキシティは，次式のように，単語の出現確率の相乗平均の逆数として定義される．

$$PP = (P(w_1 w_2 \cdots w_n))^{-\frac{1}{n}} \tag{2.46}$$

このときの $w_1 w_2 \cdots w_n$ を評価用テキストの単語列として算出したものが，テストセットパープレキシティである．もし，w_i が未知語であるような場合，確率計算がスキップされることが多い．しかし，未知語を示すシンボル UNK を用いて，$P(UNK|w_{i-2}^{i-1})$ として計算する場合がある．

また，テストコーパスに出現した未知語の割合を考慮するパープレキシティとして，前述の後者により求めた PP に対して，補正テストセットパープレキシティという尺度が提案されている [59]．補正テストセットパープレキシティ PP^* は，テストコーパス中に出現した未知語の延べ頻度を o，異なり数を m，総単語数を n とすると，未知語の出現確率を $\frac{1}{m} \cdot P(UNK|w_{i-2}^{i-1})$ として，次式によって定義される．

$$\log_2 PP^* = -\frac{1}{n}\left(\log_2 P(w_1, w_2, \cdots, w_n) - o \log_2 n\right) \tag{2.47}$$

$$= \log_2 PP + \frac{o}{n} \log_2 m \tag{2.48}$$

2.5 音声認識の評価法

2.5.1 単語正解率と単語認識精度

音声認識の性能評価には，単語正解率 (Correct) と単語認識精度 (Accuracy) が一般に用いられる．単語正解率と単語認識精度は，正解単語数（H），脱落誤りの単語数（D），置換誤りの単語数（S），挿入誤りの単語数（I）を用いて，次式のように定義される．

$$Correct = \frac{H}{H + D + S} \tag{2.49}$$

$$Accuracy = \frac{H - I}{H + D + S} \tag{2.50}$$

ここで，$H + D + S$ は入力 (発話) 単語数である．

2.5.2 符号検定 [3]

多くのパターン認識と同様に，音声認識の研究でも，統計検定を利用してアルゴリズムの有効性を検証することが望ましい [3]．特に，音声認識では，現実的に利用可能な音声データベースの規模を考慮すると，数千サンプル以上のテストデータを用意することは困難である．従って，たとえば二項分布の平均の差による検定を行うには，データ数が不足してしまうことが多く，認識率が 100% に近くかつ認識率の差が小さいような場合には，有意な差となることは少ない．

そこで，符号検定を用いることにより，テストデータ数を大きく削減することができる．この場合，認識データベースを共有するだけでなく，各テストデータに対する認識結果を比較する必要がある．すなわち，何番目のデータで誤認識を起こしたかというデータを共有する必要がある．

符号検定を説明するにあたり，まず，認識方法 A と B を同じパターンサンプル集合に適用して認識実験を行った結果，次のような結果が得られたとする．

- 方法 A,B ともに正解 $\cdots n_1$ 個
- 方法 A で正解，方法 B で誤認識 $\cdots n_2$ 個
- 方法 A で誤認識，方法 B で正解 $\cdots n_3$ 個
- 方法 A,B ともに誤認識 $\cdots n_4$ 個

ここで，方法 A と B で認識結果が異なった n_2 および n_3 について考える．$N = n_2 + n_3$ とし，N が決まったという条件の下での n_2，n_3 の起こる確率をそれぞれ p_1，p_2 とする．

もし，方法 A と B の間に性能差がなければ，n_2 と n_3 はほぼ同じ数になるはずである．すなわち，仮説 $H_0 : p_1 \equiv p_2 (= 1/2)$ が成り立つはずである．このとき，実際の観測値 $n_2 (n_3$ でもよい$)$ は二項分布 $B(N, 1/2)$ に従う．そこで，

$$Z = \frac{n_2 - N/2}{\sqrt{N/4}} \tag{2.51}$$

とおくと Z は基準正規分布 $N(0, 1)$ に従う．実際には，連続型分布でないための補正として，

$$Z = \frac{n_2 - N/2 \pm 1/2}{\sqrt{N/4}} \tag{2.52}$$

とおく．正規分布表より，$P(|Z| \geq 1.96) = 0.05$ であるから，確率 5% で $|Z| \geq 1.96$ が起きれば，仮説 H_0 を否定し，起こらなければ危険率 5% で仮説を認めることにする．すなわち，事象

$$\frac{n_2 - N/2 \pm 1/2}{\sqrt{N/4}} \geq 1.96 \tag{2.53}$$

が成り立つかどうかを判定し，成り立てば方法 A と B の認識結果に有意差あり，成り立たなければ両者の認識性能に有意差なしとする．

2.6　まとめ

本章では，音声認識の基本原理と，音響モデル，言語モデル，および音声認識の性能評価の尺度について述べた．現在の音声認識は統計的手法に基づいており，大規模な音声データから学習した音響モデルと，大規模な言語データから学習した言語モデルを用いて，与えられた音響特徴量 X に対して最適な単語列 W を推定する問題として定式化される．

音響モデルは，隠れマルコフモデルと混合正規分布を用いて，音素や音節といった単位に対して音響特徴量の分布をモデル化する．音素や音節は前後の文脈によって性質が変化することから，直前や直後の文脈を考慮した文脈依存モデルが用いられる．また，学習データと認識対象の間で入力環境や話者の性質が合致していることがきわめて重要であり，モデルの適応を行なって両者のミスマッチに対処することが考えられる．

言語モデルは，N-gram モデルによって単語の繋がり易さを統計的にモデル化する．学習データの不足によるデータスパースネスの問題に対処するために，低次の N-gram 確率を用いて高次の N-gram 確率を推定するバックオフ平滑化が行われる．言語モデルの性能は，評価用テキストに対するテストセットパープレキシティで評価することができる．

音声認識の性能評価には，正解テキストに対する，音声認識結果の脱落誤り，置換誤り，挿入誤りの頻度を考慮した単語正解率と単語認識精度が用いられる．また，提案手法の有

効性を既存手法との比較によって検証する場合には，統計検定の一つである符号検定が有効である．

第 3 章

話し言葉特有の現象に関する分析

3.1　はじめに

本章では，話し言葉特有の様々な現象について分析を行う．まず，3.2 節では，講演音声や講義音声などを対象として，フィラーや言い淀みといった話し言葉特有の非流暢現象 (disfluency) の分析を行う．次に，3.3 節では，話し言葉に出現するポーズについて分析を行い，書き言葉に出現する句読点では，話し言葉のポーズの位置情報として不十分であることを示す．また，3.4 節では，会議録や速記録などの整形が加えられた書き起こしにおける，話し言葉特有の現象の扱いについて分析する．

3.2　フィラー，言い淀み，倒置に関する分析

話し言葉では，フィラーや言い淀み，倒置といった話し言葉特有の現象が存在し，これらが音声認識において大きな障害となる．そこで，こうした諸現象の性質を明らかにするために，話し言葉コーパスの分析を行った．分析を行うコーパスとしては，日本語話し言葉コーパス [8] と，本研究室で構築を進めている講義音声データベース [60] を用いた．

3.2.1　話し言葉特有の現象

話し言葉特有現象にはフィラーや言い淀み，言い直し，倒置のほか，助詞落ち，繰り返しなど様々なものが挙げられるが，今回は特に以下の 4 つを分析の対象とした．

1. フィラーおよび感情表出系感動詞
   例）こういうことを えー 行うわけですが ...
2. 言い直し等による語断片
   例）この前の だいが 大学の学部の会議で ...

3. 助詞，助動詞，接辞の言い直し
   例）評価値 が の数値が ...
4. 倒置
   例）私は耐えられないんです これは

3.2.2 日本語話し言葉コーパス

日本語話し言葉コーパス (Corpus of Spontaneous Japanese:以下 CSJ) は，日本語の話し言葉を対象として豊富な研究用付加情報を付与した非常に大規模なデータベースである [61][8]．CSJ には学会講演を始めとした 5 種類の発話が含まれているが，今回は学会講演，模擬講演，対話の 3 種類の発話を分析対象とする．それらの発話に関する情報を表 3.1 に示す．表 3.1 から分かるように，CSJ の全 3302 講演の内，学会講演や模擬講演が大部分を占めており，その各講演時間は 10〜15 分程度である．

また，CSJ では，フィラーや言い直し等の話し言葉特有の現象に対し，一定の基準に基づいてタグの付与がなされている [62]．今回分析の対象とする現象のうち，1.〜3. の 3 つについては，それぞれ以下のタグが対応する．

1. フィラーおよび感情表出系感動詞：タグ F
   例）こういうことを (F えー) 行うわけですが ...
2. 言い直し等による語断片：タグ D
   例）この前の (D だいが) 大学の学部の会議で ...
3. 助詞，助動詞，接辞の言い直し：タグ D2
   例）評価値 (D2 が) の数値が ...

タグ F は，フィラーや感情表出系感動詞に付与されるタグである．ただし，CSJ では，フィラーの語彙を限定しており，これに含まれないものに関してはたとえ場つなぎ的な機能を有する表現であっても，タグ F は付与されない．また，フィラーの中でも特に，「あの/その」はフィラーか連体詞かの判断が困難である場合が少なくないが，その場合でもタグ F は原則付与される．感情表出系感動詞は，驚いた時や落胆した時などに発する感動詞であり，こちらは語彙は限定されず，タグ F の付与対象となる．

また，タグ D は，言い直し等によって生じた語断片に付与されるタグである．具体的には，何かを言い掛けた途中で別の表現で言い替えたとき，およびその他の発声上の問題によって語の断片が生じた場合に本タグが付与される．言い直しがされた場合でも，語の断片が生じなかったときは本タグの対象外となる．

タグ D2 は，助詞・助動詞・接辞の言い直し部分に付与されるタグである．タグ D と混

同し易いが，本タグは言い直しにおける言い掛け部・訂正部が，共に助詞・助動詞・接頭辞・接尾辞あるいは数字から構成される場合にのみ付与される．本タグは，言い掛け部が語断片でなくとも付与される．ただし，言い掛け部が機能語の断片であった場合に付与されるタグは，本タグ (D2) ではなく，タグ D である．

ここで，倒置に対しては，タグ付与はされていないが，その代わりに，コアと呼ばれる一部のデータ（全体の 6.6 %程度）に限って，節単位認定および係り受け構造付与の際に，必要に応じて倒置に対するラベリングが行われている [63][64]．CSJ では，統語的・意味的な妥当性を備えた処理単位を抽出するために，書き起こしテキストを節単位 *1 に分割する作業を行っている．この作業は，まず節境界検出プログラム CBAP-csj によって自動的に行われる [63]．しかし，CBAP-csj は局所的な形態素列のみを参照して節境界を検出するため，倒置を始めとした話し言葉特有の現象が発生した場合に対応できない．そこで，CSJ では CBAP-csj による自動検出の結果に対し，さらに人手による修正を加えており，この修正作業において，倒置に対するコメント付与を行っている．また，CSJ では，係り受け構造を付与する際に，倒置が生じている箇所については右から左への係り受けとしてラベリングしている．なお，CSJ における係り受け構造付与は人手で行われている．

以上のタグおよびラベルを利用して，対応する各現象の分析を行った．

3.2.3　講義音声データベース

講義音声データベース（Corpus of Japanese classroom Lecture speech Contents:以下 CJLC）には，本学をはじめとした複数の大学で実際に開講されている全 89 講義の録音音声が収録されている [60]．CSJ と同一の基準に基づき，フィラーや言い直し等にタグが付与されており，同様の分析が可能である．また，これらの一部（5 講義）についてのみ，倒置と判断される箇所に独自の倒置タグが付与されている．この倒置タグの付与にあたっては，[63][64] を参考にした．

CJLC に関する情報を表 3.1 に示す．講義の録音時間は平均で約 43.5 分であり，CSJ の講演と比べてかなり長い．また，倒置タグを付与した 5 講義に関する情報を表 3.2 に示す．

3.2.4　フィラー，言い淀み，倒置の出現頻度に関する分析結果

コーパスの分析結果を以下の図表に示す．タグ F,D,D2 の出現頻度の平均に関する調査結果が表 3.3，出現頻度の分布に関する調査結果が図 3.1-3.3，倒置に関する講義音声の調査結果が表 3.4 である．ただし，倒置に関しては，CSJ はコアに含まれるコーパスのみ

*1 3.3 節参照.

表 3.1 CSJ と CJLC のデータ量

コーパス	種別	講義数（講演数）	収録時間 [時間]	平均時間長 [秒/講義]	平均単語数 [単語数/講義]
CSJ	学会講演	987	275	1,003	3,358
	模擬講演	1,715	330.6	694	2,122
	対話	58	12.3	765	2,613
CJLC	講義	89	63	2,610	6,636

表 3.2 倒置タグの付与された講義データ

話者	語数（形態素）	時間（sec）	タグ数		
			F	D	D2
話者 A（中川）	12451	4065	525	75	4
話者 B（秋葉）	12330	4195	1557	81	4
話者 C（北岡）	13611	3949	687	56	3
話者 D（新田）	9953	4168	779	53	10
話者 E（小宮）	16103	5349	630	102	10

を，CJLC は倒置タグの付与された 5 講義のみをそれぞれ調査の対象とした．

なお，ここで表 3.4 の CSJ コア平均とは，CSJ のコアに含まれる学会講演および模擬講演について平均を取ったものであり，そのような講演全体で見た平均と，それらのうちで倒置が少なくとも 1 回以上出現した講演で見た平均とを示している．なお，CSJ に関しては，コアに含まれる全 215 講演のうち，倒置が 1 回以上発生したのは 71 講演であり，倒置の発生回数は全部で 215 回であった（1 講演当り 1 回）．

タグ F の出現頻度で比較すると，対話のみが，比較的フィラーが多く現れているのに対して，それ以外の 3 種の音声では，あまり差が見られなかった．タグ D およびタグ D2 の出現頻度で比較すると，講義のみが比較的少なめであり，それ以外の 3 種の音声には，あまり差が見られなかった．また，ほとんどは学会講演および模擬講演からなる CSJ コアと講義とを比較すると，講義の方が倒置の発生回数は小さかった．

また，言い直しや言い淀み，倒置は，全体的に発生頻度が非常に小さい．一方で，フィラーは発生頻度が非常に高いため，優先的に対処するべきであるといえる．

なお，このように話し言葉特有の現象に注目した分析として，中川ら [65] が模擬対話を対象としたフィラー，言い直し，倒置，助詞落ち等の分析を行っており，本実験と同様にフィラーの出現率が極めて高い（言い直しや助詞落ちの出現頻度が高々 0.1 回/文であるのに対し，フィラーは 1.1 回/文）という結果を得ている．

表 3.3　CSJ と CJLC の平均タグ出現頻度

コーパス	種別	平均タグ出現頻度 [頻度/講義]（[頻度/秒]）		
		F	D	D2
CSJ	学会講演	229.2(0.229)	44.5(0.0448)	3.4(0.00343)
	模擬講演	118.8(0.169)	26.0(0.0370)	1.4(0.00201)
	対話	322.2(0.420)	43.9(0.0588)	1.4(0.00195)
CJLC	講義	410.3(0.172)	49.9(0.0139)	3.9(0.00194)

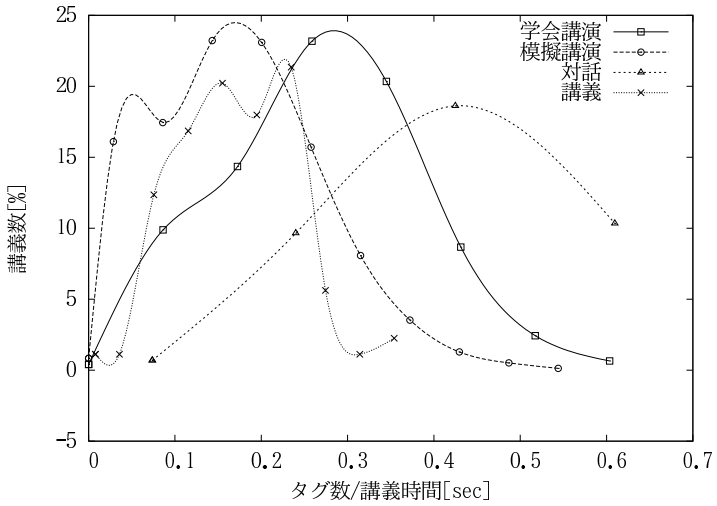


図 3.1　タグ F（フィラー）の出現頻度

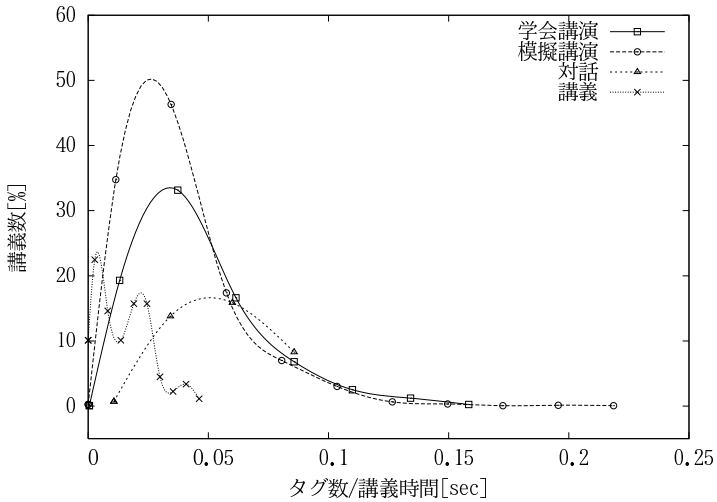


図 3.2　タグ D（言い淀み）の出現頻度

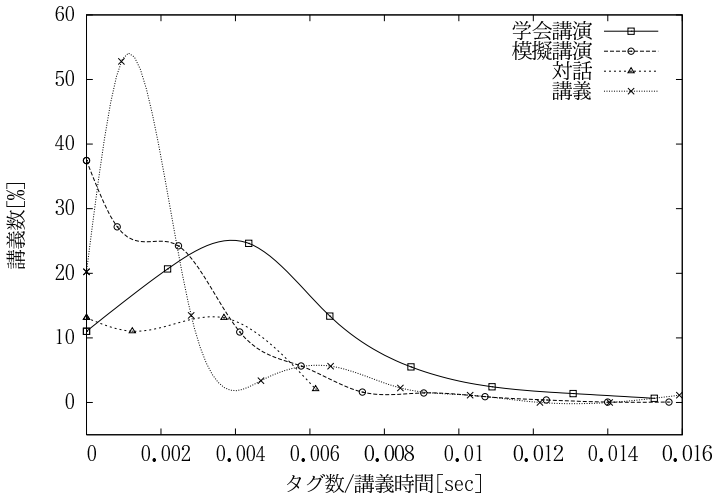


図 3.3　タグ D2（助詞・助動詞・接辞の言い直し）の出現頻度

表 3.4　倒置の出現頻度

(講義者)	話者 A (中川)	話者 B (秋葉)	話者 C (北岡)	話者 D (新田)	話者 E (小宮)	講義 平均	CSJ コア平均 全体	CSJ コア平均 倒置を含む
倒置発生回数	33	2	10	7	2	10.8	1.07	3.03
発生回数/講義時間 (min)	0.049	0.029	0.15	0.10	0.022	0.07	0.088	0.27
発生回数/総語数	0.0027	0.00016	0.00073	0.00070	0.00012	0.00088	0.000475	0.00144

3.2.5　フィラーの種類に関する分析結果

CSJ の学会講演および模擬講演について，出現する全てのフィラーの語彙を列挙し，各語彙の頻度を調査した．その結果，合計で 151 種類もの語彙が列挙された．しかし，これらの多くは，長音・促音の有無や語尾音節の繰り返しなどの発音上の揺れによる派生形である．そこで，これらの語彙のうち，互いに類似した語彙同士をグループとして統合した．具体的には，各語彙の表層形に以下の処理を施し，等しくなったもの同士を統合した．

- 長音（ー）あるいは促音（っ）の除去
  例）あのー　⇒　あの
  　　えっと　⇒　えと

- 語尾で２つ以上連続するモーラの除去
  例）ありゃりゃりゃ　⇒　ありゃ
  　　おれれれ　　⇒　あれ

- この他，語中の撥音など，若干の違いがある語彙は，その頻度が小さい（10 に満たない）場合のみ統合
  例）あんのー（頻度 6）　⇒　あの
  　　えーっとですね（頻度 1）　⇒　えと

統合の結果，151 種類の語彙が 58 のグループにまとめられた．統合後の語彙グループとその頻度分布を表 3.5 に示す．なお，頻度が 10 に満たないような低頻度なグループはここでは省略している．表 3.5 から分かるように，「え，あの，まー，えとー，その，あ」の上位 6 グループの語彙のみで全体の 9 割がカバーされている．これは中川ら [65] の模擬対話に対する分析結果とも一致している（上位 6 語彙が全体の 95.8% をカバー）．

3.2.6　フィラーの出現とコンテキストの関係に関する分析結果

続いて，直前のコンテキストの情報が，フィラー語彙の予測のために有効であるかどうかを調査した．具体的には，モーラ，母音，品詞のフィラー直前における分布とユニグラム分布とを比較した．結果を表 3.6-3.11 に示す．なお，各表中において，

- “head”は講演の開始点を表す．
- 品詞やモーラが空欄となっていることがある．
  これは主に，以下のような場合である．
  1. “？”タグ（聞き取りに自信なし）が付与されている場合
     ⇒　品詞もモーラも空欄になる．
  2. 言いよどみが発生した場合
     ⇒　品詞は“言いよどみ”となり，モーラが空欄となる．
  3. ボーカルフライ等で母音が特定できない場合
     ⇒　品詞は空欄となり，モーラが“φ”となる．

である．

まず，表 3.6 および表 3.7 より，フィラーの直前では“デ”，“ワ”，“テ”，“ノ”，“ガ”といったモーラの出現率が高い．これらはユニグラムの分布では上位になかったものである．なお，第 21 位以下を含む表は付録 A に示した．

また，表 3.8，表 3.9 より，フィラーの直前では“え”や“う”の出現率が高くなるなど，母音コンテキストにおいても一定の特徴が見られる．

さらに，表 3.10，表 3.11 より，フィラーの直前においては助詞や感動詞，接続詞が比

表 3.5　類似語彙統合後のフィラー語彙リスト

No.	フィラー語彙	出現数	[%]	累積 [%]
1	え	169105	38.05	38.05%
2	あの	79563	17.9	55.95%
3	まー	78132	17.58	73.53%
4	えとー	27686	6.23	79.76%
5	その	22968	5.17	85.0%
6	あ	22834	5.14	90.07%
7	んー	15634	3.52	93.59%
8	お	12245	2.75	96.34%
9	う	7788	1.75	98.09%
10	い	4345	0.98	99.07%
11	とー	2748	0.62	99.69%
12	んとー	610	0.14	99.83%
13	あれ	149	0.03	99.86%
14	うん	148	0.03	99.89%
15	はー	116	0.03	99.92%
16	あとー	86	0.02	99.94%
17	あら	62	0.01	99.95%
18	わー	52	0.01	99.96%
19	うんと	39	0.01	99.97%
20	うわー	38	0.01	99.98%
21	へー	31	0.01	99.99%
22	おっと	21	0	99.99%
23	ふーん	20	0	99.99%
24	ほー	11	0	99.99%

（頻度が 10 に満たないものは省略）

較的出現し易い.

以上のように，フィラーの直前におけるモーラ，母音，品詞といったコンテキストには一定の特徴が見られることから，フィラーは発話中の任意の部分に現れるのではなく，その出現位置には偏りがあるといえる．よって，こうしたコンテキストが，フィラーの出現の予測にあたって有用であると考えられる．

表 3.6　フィラーの直前に出現したモーラ (上位 20 種)

No.	モーラ	出現数	[%]	累積 [%]
1	デ	42669	9.60	9.60
2	ワ	36912	8.30	17.90
3	テ	32910	7.40	25.31
4	ノ	30841	6.94	32.25
5	ガ	27170	6.11	38.36
6	ス	26466	5.95	44.31
7	ニ	24005	5.40	49.71
8	モ	22068	4.96	54.68
9	ト	21857	4.92	59.60
10	ネ	14476	3.26	62.85
11	タ	13654	3.07	65.93
12	エ	13410	3.02	68.94
13	オ	13342	3.00	71.94
14	ラ	11664	2.62	74.57
15	ン	11348	2.55	77.12
16	カ	9418	2.12	79.24
17	マ	8892	2.00	81.24
18	イ	7834	1.76	83.00
19	ナ	6214	1.40	84.40
20	リ	5927	1.33	85.73

表 3.7 モーラのユニグラム統計 (上位 20 種)

No.	モーラ	出現数	[%]	累積 [%]
1	ン	745928	5.72	5.72
2	イ	645148	4.94	10.66
3	ト	576029	4.41	15.07
4	ノ	559506	4.29	19.36
5	カ	470397	3.60	22.96
6	テ	434972	3.33	26.30
7	デ	420469	3.22	29.52
8	シ	394829	3.03	32.54
9	タ	393722	3.02	35.56
10	ス	383961	2.94	38.50
11	マ	372173	2.85	41.35
12	ッ	349257	2.68	44.03
13	ナ	339944	2.60	46.64
14	オ	339269	2.60	49.24
15	コ	336643	2.58	51.81
16	エ	315840	2.42	54.23
17	ニ	312064	2.39	56.63
18	ワ	286922	2.20	58.82
19	ア	279158	2.14	60.96
20	ク	276394	2.12	63.08

表 3.8 フィラーの直前に出現した母音

No.	母音	出現数	[%]	累積 [%]
1	あ	124693	28.05	28.05
2	え	107498	24.18	52.24
3	お	100825	22.68	74.92
4	う	49747	11.19	86.11
5	い	47361	10.66	96.77
6	ン	11348	2.55	99.32
7	head	1140	0.26	99.58
8	ッ	800	0.18	99.76
9		649	0.15	99.90
10	φ	389	0.09	99.99
11	×	35	0.01	100.00

(空欄は，”？”タグや言いよどみにより母音が特定できなかった場合)

表 3.9　母音のユニグラム統計

No.	母音	出現数	[%]	累積 [%]
1	あ	3161505	24.22	24.22
2	お	2997836	22.97	47.19
3	い	2160749	16.56	63.75
4	え	1928952	14.78	78.53
5	う	1663388	12.75	91.28
6	ン	745928	5.72	96.99
7	ッ	349257	2.68	99.67
8	×	28672	0.22	99.89
9	φ	9176	0.07	99.96
10		3965	0.03	99.99
11	××	1406	0.01	100.00

（空欄は，”？”タグや言いよどみにより母音が特定できなかった場合）

表 3.10　フィラーの直前に出現した品詞

No.	品詞	出現数	[%]	累積 [%]
1	助詞	232944	52.41	52.41
2	助動詞	66082	14.87	67.27
3	感動詞	44441	10.00	77.27
4	名詞	20868	4.69	81.97
5	接続詞	19388	4.36	86.33
6	副詞	18918	4.26	90.59
7	言いよどみ	15812	3.56	94.14
8	動詞	12226	2.75	96.89
9	接尾辞	3789	0.85	97.75
10	連体詞	3426	0.77	98.52
11	形容詞	2836	0.64	99.16
12	代名詞	1204	0.27	99.43
13	head	1140	0.26	99.68
14		626	0.14	99.82
15	形状詞	293	0.07	99.89
16	記号	290	0.07	99.95
17	接頭辞	202	0.05	100.00

（空欄は，”？”タグやボーカルフライにより品詞が特定できなかった場合）

表 3.11 品詞のユニグラム統計

No.	モーラ	出現数	[%]	累積 [%]
1	助詞	2221543	30.69	30.69
2	名詞	1627149	22.48	53.17
3	動詞	953400	13.17	66.34
4	助動詞	858112	11.85	78.19
5	感動詞	450044	6.22	84.41
6	副詞	218310	3.02	87.43
7	接尾辞	205016	2.83	90.26
8	代名詞	149601	2.07	92.32
9	形状詞	120186	1.66	93.98
10	形容詞	101205	1.40	95.38
11	言いよどみ	91996	1.27	96.65
12	連体詞	91432	1.26	97.92
13	接続詞	81650	1.13	99.04
14	接頭辞	41725	0.58	99.62
15	記号	23828	0.33	99.95
16		3605	0.05	100.00

(空欄は，”？”タグやボーカルフライにより品詞が特定できなかった場合)

3.2.7 フィラーと節境界の関係

CSJ では，文節よりも長い統語的・意味的な単位として，節単位が定義されている．節境界は絶対境界，強境界，弱境界の 3 つのレベルに区分され，それぞれ以下のように定義されている [66].

- 絶対境界：形式上明示的な文末表現に相当する節境界.
- 強境界：いわゆる文末ではないが，発話の大きな切れ目として考えられる節境界.
- 弱境界：節境界ではあるが，通常は発話の切れ目になることはないと考えられる節境界.

本節では，CSJ のコアに含まれる学会講演および模擬講演を対象として，フィラーが節境界 (弱境界，強境界，絶対境界) に出現する割合を調査した．この結果，すべてのフィラーの内，節境界に出現したフィラーの割合はそれぞれ表 3.12 のようになった．表 3.12 のように，フィラーの 41.1% がいずれかの節境界の位置に出現している．このことから，フィラーは，ある程度の意味的なまとまりを区切る位置に出現しやすいことがわかる.

表 3.12　節境界に出現したフィラーの割合

節境界の種類	出現したフィラーの頻度 (%)
弱境界	21.7
強境界	9.4
絶対境界	10.0
合計	41.1

3.3　ポーズに関する分析

話し言葉の音声を対象とする場合には，ポーズに基づいて得られた認識処理単位 (認識対象区間) と言語的なまとまりとは必ずしも一致しない．例として，日本語話し言葉コーパス (CSJ)[8] における転記基本単位と節単位，および国会会議録 *2 と毎日新聞における文単位の長さの分布を図 3.4 に示す． CSJ の転記基本単位は，200msec 以上の無音区間に基づいて区切られており，一般的な音声認識器における認識処理単位として用いられている．また，3.2.7 節でも述べたように，CSJ では，統語的・意味的な単位として，節単位が定義されている．ここでは，発話の大きな切れ目に相当する絶対境界および強境界によって区切られた単位を節単位とする．なお，節境界認定は，話者ではなくコーパス作成者によって行われているため，節単位は，話者の発話意図をそのまま表現した単位ではない．国会会議録は，国会審議の速記録からフィラーや言い直し，繰り返しなどの話し言葉的な現象を整形したコーパスであり，整形作業者の主観に基づいて句読点が挿入されて

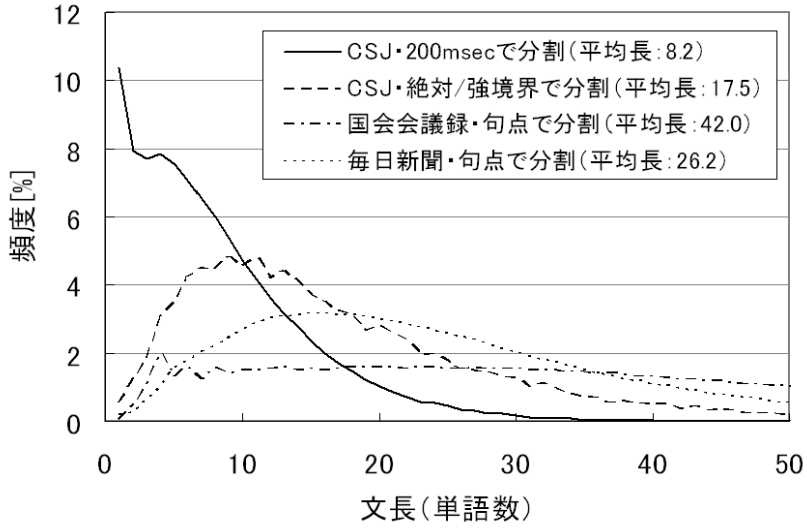


図 3.4　認識処理単位と文単位の長さの比較

*2 http://kokkai.ndl.go.jp/

いる．よって，国会会議録における文単位も，話者の発話意図に基づく単位ではない．また，毎日新聞の文単位は，筆者の言語的直感に基づいて挿入された句点に基づいている．図 3.4 より，CSJ の転記基本単位は，国会会議録および毎日新聞の文単位に比べて，非常に短い単位が多い．このように，ポーズに基づいて得られた処理単位と，コーパス作成者または筆者によって設定された文単位との間には明白な不整合が存在する．一方で，CSJ の節単位の長さの分布と，国会会議録および毎日新聞の文単位の分布は良く似ている．そのため，入力音声を自動的に節単位に分割することによって，句点に基づいた文単位の言語モデルで適切に扱う手法が提案されている．しかし，実際には，話し言葉に対して節単位を自動検出することは容易ではない [34][35][36].

ポーズに基づく処理単位と句点に基づく文単位が不整合な理由は，言語的なまとまり以外の要因に基づくポーズが数多く出現するためである．節境界の定義より，言語的まとまりに起因するポーズは，節境界と対応する位置に出現するはずである．例として，CSJ のコア・コーパスに出現するポーズ（200msec 以上）の分布を図 3.5 に示す．図 3.5 より，言語的まとまり以外の要因によるポーズが過半数を占めており，句点のみでは，ポーズの位置情報として不十分であることが分かる．また，国会会議録から取り出した 88 文 [*3] について，ポーズ（200msec 以上）と句読点の位置を比較した．なお，読点は，国会会議録作成者の言語的直感に基づいて挿入されている．比較例を，図 3.6 に示す．ポーズは 452 個所，読点は 357 個所，句点は 88 箇所に出現したが，ポーズの内，42.3% が読点と一致し，15.9% が句点と一致していた．また，読点の 53.5% がポーズと一致し，句点の

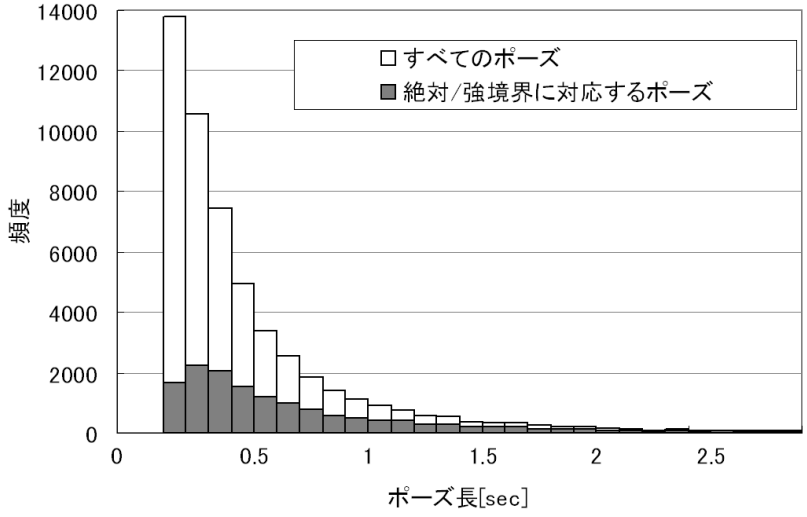


図 3.5　CSJ に出現するポーズの長さの分布

[*3] ただし，国会会議録への記録にあたって大幅な整形処理が行われている文は，音声データとの直接の比較が困難であるため対象から除外した.

81.8% がポーズと一致していた．よって，読点のみでも，句読点の組み合わせでも，ポーズの位置情報として不十分である．以上の分析より，話し言葉の書き起こしまたは書き言葉コーパスに含まれる句読点は，ポーズの位置情報として不十分といえる．

確かにＮＨＫは、借入金に対して、これを返していかなければならないと <p> いうことが、ございます。一方では、<p> やはり視聴者の方々に <p> しっかりと、番組をつくって届けるというこの役目と両方持っている中で、<p> 大変厳しい財政状況でありますから、<p> やはり基本的には番組の方で <p> しっかりと <p> ＮＨＫに対する期待を <p> 果たしていくということがまずポイントかと考えております。<p>

図 3.6　国会会議録における句読点とポーズの例（<p> は 200msec 以上のポーズ）

3.4　会議録・速記録における話し言葉特有の現象の扱いに関する分析

会議録や速記録などの整形された書き起こしでは，フィラーや言い直しが除去されているほか，話し言葉調の言い回しが書き言葉調に整形されるなど，様々な整形処理がなされている．こうした整形処理の頻度について，実際の国会会議録を分析した結果を表 3.13 に示す．表 3.13 のように，話し言葉では文節の「ですね」や文末の「けども」といった冗長な表現が出現するが，これらはすべて整形されている．例を図 3.7 に示す．図 3.7 に見られるように，国会会議録では，可読性を重視するために，フィラー（例．”えー”, ”いー”）や言い直し（例．”け”），および冗長な言い回し（例．”ですね”, ”と”）はすべて削除されており，また，口語体の言い回し（例．”てる”）は文語体（例．”ている”）に置き換えられている．さらに，助詞の不足が補われている（例．”を”）ほか，一部の読点は速記者の判断によって追加あるいは削除されている．

3.5　まとめ

本章では，話し言葉特有の様々な現象について分析した．フィラーや言い直しといった非流暢現象について，講演音声や講義音声などを対象に分析を行った結果，特にフィラーの出現頻度が高いことが明らかになった．フィラーの出現位置は一様ではなく一定の偏りがある．また，フィラーの語彙は，出現頻度の高い上位 6 種類でフィラー全体の 9 割をカバーできる．

話し言葉に出現するポーズは言語的な区切りとは異なっており，従って，書き言葉に出現する句読点では，話し言葉のポーズの位置情報として不十分である．

ところ が ですね， えー，この 資料，見 て み ます と 神奈川 県 の 場合 は，け，結果 と し て 財政 的 に いー 豊か に なっ てる と．

(i) 正確な書き起こし

ところ が，この 資料 を 見 て み ます と，神奈川 県 の 場合 は，結果 と し て 財政 的 に 豊か に なっ て いる．

(ii) 整形された書き起こし

図 3.7 正確な書き起こしと整形された書き起こしの例

表 3.13 国会会議録における整形処理

整形処理	頻度 [%]	例
表現の置換	5.3	「じゃあ逆の視点から…」 →「では逆の視点から…」
文節の「ですね」の削除	31.3	「支出予算のですね、削減は…」 →「支出予算の削減は…」
言い直し・言い淀みによる語断片の削除	17.6	「そちらの資本のぶぶ、部分に…」 →「そちらの資本の部分に…」
言い直しの削除	13.0	「二百四十四の、違う、二百四十七の…」 →「二百四十七の…」
余分な助詞の削除	3.8	「できるだけ削減をして…」 →「できるだけ削減して…」
省略された助詞の挿入	10.7	「これ１ページ目ですが…」 →「これは１ページ目ですが…」
倒置の修正	6.1	「借金一方で三百億を…」 →「一方で借金は三百億を…」
その他、冗長な部分の削除	6.9	「一方でやはり視聴者の方々に…」 →「一方で視聴者の方々に…」
文末の整形	5.3	「いろいろ言い訳をしておられましたけども」 →「いろいろ言い訳をしておられました」

会議録や速記録といった整形された書き起こしでは，話し言葉特有のフィラーや言い直しといった非流暢現象のほか，冗長な言い回しも整形される．

第 4 章

話し言葉音声認識のためのフィラーの統計的モデリング

4.1 はじめに

本章では，フィラーを含まないコーパスから，フィラー予測モデルに基づいてフィラー付きの話し言葉言語モデルを構築する方法を提案する．本手法では，音声認識対象とは異なるドメインのコーパスからフィラー予測モデルを学習し，認識対象のドメインのフィラーを含まないコーパスに対してフィラーの挿入を行い，フィラーに対応した言語モデルを構築する．日本語話し言葉コーパスを対象とした実験の結果，本提案手法は，フィラーを含む正確な話し言葉コーパスから作成した 3-gram モデルにきわめて近い言語モデルを再現できた．また，国会会議録を対象として，この手法によって作成した言語モデルと従来手法とを比較したところ，より高い音声認識性能を達成することができた．

4.2 フィラー予測モデルの定式化

例として，文 (1) のようなフィラーを含まない文から，フィラーに対応した言語モデルを作成する方法を考える．

(1) この画面を見ると ···

この場合，2 つの方法が考えられる．第 1 の方法は，秋田ら [20] のように，文 (1) からフィラーを含まない言語モデルを学習しておき，その言語モデルをフィラーに対応した言語モデルに変換するという方法である．第 2 の方法は，文 (1) 中の適切な個所にフィラーを挿入して，文 (2) のようなフィラーを含む文を作成し，その文からフィラーに対応した言語モデルを学習するという方法である．

(2) この画面をえー見ると ···

しかし，第 1 の方法には，幾つかの欠点がある．まず，この方法では，対象とする言語モデルに対応した変換規則または変換モデルが必要となり，別種の言語モデルを利用するためには，変換規則または変換モデルを作成し直す必要がある．たとえば 3-gram 言語モデルに対して作成した変換規則または変換モデルを，確率文脈自由文法などの別種の言語モデルに対してそのまま適用することはできない．

加えて，言語モデルよりも長い文脈情報を言語モデルの変換に利用しにくいという問題点も挙げられる．たとえば言語モデルが 3-gram の場合，変換モデルが利用できる文脈情報は直前 2 形態素および現在の形態素のみであり，直後の形態素の情報やモーラの情報を利用することは困難である．たとえば秋田ら [20] の方法では，3-gram 言語モデルを対象としていることから，形態素の 3 つ組および品詞の 3 つ組に対する変換パターンを用いている．

また，第 1 の方法では，言語モデルの変換にあたり，変換後の言語モデルの確率を推定する必要がある．従って，近年音声言語処理の分野でよく用いられているような種々の機械学習手法が適用しにくい．たとえば，秋田ら [20] の方法では，形態素や品詞の N-gram モデルを用いて変換確率の推定を行っているが，これらの代わりに決定木やニューラルネットワークなど，出力値と確率値との対応付けが困難な手法を適用することは不可能と考えられる．

一方で，第 2 の方法では，フィラーの挿入個所と種類を予測するモデルが必要になるが，そのようなモデルさえ得られれば，言語モデルの変更には容易に対応可能である．加えて，言語モデルより長い文脈情報を容易に利用することができる．また，フィラーの挿入箇所と種類の予測さえできれば良いので，第 1 の方法のように確率値を取り扱うことが必ずしも必要とはならないことから，様々な機械学習手法を適用しやすいと言える．

以上の理由から我々は，第 2 の方法によるフィラーを含む言語モデルの構築方法を提案する．フィラーの挿入にあたっては直後の形態素やモーラなど，言語モデルよりも長い文脈情報を利用し，また，挿入箇所を決定するために Conditional Random Field(CRF)[67] を適用する．

本研究ではこれ以降，文 (1) を文 (2) のように書き換えるためにフィラーの挿入個所と種類を予測するモデルを**フィラー予測モデル**と呼ぶ．3 節で述べたように，フィラーには多様な派生形が存在する．従って，フィラーの挿入箇所と種類を同時にモデル化すると，データスパースネスが生じる恐れがある．そこで，我々は，フィラーの挿入箇所と種類は独立に推定できるという仮定をおく．すなわち，フィラーを挿入する箇所を推定する**フィラー挿入モデル**と，推定された箇所に挿入するべき適当なフィラーを選択する**フィラー選択モデル**，という 2 つのモデルの組み合わせとしてフィラー予測モデルを定式化する．

4.2.1　フィラー挿入モデル

フィラー挿入モデルとは，ある形態素列が与えられた時に，その形態素列中においてフィラーを挿入すべき箇所を推定するモデルである．本研究では，このモデルを，形態素列を対象とし，個々の形態素に対して，その形態素の直後にフィラーを挿入するべきかどうかというラベルを付与するという，系列ラベリング問題として定式化する．例えば，文 (1) を文 (2) に変換する場合には，最初に文 (1) を形態素列に分解し，図 4.1 のように個々の形態素に対して，直後にフィラーを挿入すべきである場合にはラベル F を付与し，フィラーを挿入すべきではない場合にはラベル O を付与する．なお，この定式化では，フィラーが 2 つ以上連続して出現するような状況を表現することができない．しかし，日本語話し言葉コーパスを調査した結果から，フィラーが連続して出現する確率は約 6% 程度と極めて低いことが分かっている (模擬対話データでは約 12% 程度 [65])．従って，今回はそのような状況は扱わない．

形態素列	この　画面　を　見る　と　・・・ (文頭) 連体詞　名詞　助詞　動詞　助詞 (文頭)　コノ　ガメン　ヲ　ミル　ト
ラベル列	O　O　O　F　O　O　・・・

図 4.1　フィラー挿入モデルの学習用ラベル

本研究では, このような問題を解くフィラー挿入モデルを, 条件付き確率場 (Conditional Random Field．以下，CRF）[67] を用いて作成する．CRF は，隠れマルコフモデルなどのモデルと比べて柔軟な素性設計が可能であり，また，比較的少量の学習データでも高い性能を示すことが知られている識別モデルである．

CRF では，形態素列 X に対するラベル列 Y の条件付き確率 $P(Y|X)$ を，次式のように表す．

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{i}^{n} \sum_{a} \lambda_a f_a(X_i, Y_i) \right) \tag{4.1}$$

ここで，f_a は素性関数，λ_a は素性関数に対する重み，$Z(X)$ は正規化項である．なお，CRF の学習時には，パラメータの事前分布として Gaussian Prior を用いて事後確率を最大化することにより，パラメータを正則化した．

4.2.2 フィラー選択モデル

フィラー選択モデルは，適当な形態素列とフィラーの挿入箇所が指定された時に，挿入すべき適当なフィラーを選択するモデルである．本研究では，単純に，周囲の形態素やモーラなどの文脈 h に対してフィラー f が生起する条件付き確率 $P_s(f|h)$ を，フィラー選択モデルとして用いる．条件付き確率 $P_s(f|h)$ は，Witten-Bell スムージングを適用して [58]，次式のように推定する．

$$P_s(f|h) = \begin{cases} \frac{c(h,f)}{c(h)+r(h,f)} & \text{if } c(h,f) > 0 \\ \frac{r(h,f)}{c(h)+r(h,f)} \cdot P_s(f|h') & \text{otherwise} \end{cases} \tag{4.2}$$

ただし，$c(h,f)$ はフィラーを含む正確な話し言葉コーパスにおいて文脈 h とフィラー f が同時に生起する頻度，$c(h)$ は文脈 h の生起する頻度，$r(h,f)$ は文脈 h の直後に現れるフィラーの種類の数である．文脈 h' は，文脈 h から条件を 1 つ取り除いた文脈である (バックオフ)．

4.2.3 比較手法

本研究では，提案手法に対し，以下の手法を比較する．

- 形態素トライグラムに基づくフィラー挿入モデル
- 品詞トライグラムに基づくフィラー挿入モデル
- フィラーのユニグラム確率に基づくフィラー挿入モデル

形態素トライグラムに基づくフィラー挿入モデルでは，直前 2 形態素を考慮したトライグラムモデルに基づいてフィラーの挿入位置を決定する．なお，トライグラムモデルの平滑化のために，Witten-Bell バックオフを用いる．品詞トライグラムに基づくフィラー挿入モデルでは，直前 2 形態素の情報の内，ドメインへの依存性の低い品詞情報だけを考慮したトライグラムモデルに基づいてフィラーの挿入位置を決定する．形態素トライグラムと同様に，Witten-Bell バックオフを用いてモデルの平滑化を行う．フィラーのユニグラム確率に基づくフィラー挿入モデルでは，文脈を考慮せず，単純に，ユニグラム確率に基づいてフィラーの挿入位置を決定する．

4.3　フィラー予測モデルを用いたフィラーつき言語モデルの構築

本研究では，フィラーを含まない不正確な話し言葉コーパスから，フィラーに対応した話し言葉言語モデルを作成する手順として，我々は，以下のような手順を提案する．

1. フィラーを含む正確な話し言葉コーパス（以後，**学習コーパス**と呼ぶ）から，フィラー予測モデルを構築．この部分は，更に以下の 2 段階に分けられる．
   (a) フィラー挿入モデルの構築.
   (b) フィラー選択モデルの構築.
2. フィラーを含まない不正確な話し言葉コーパス（以後，**開発コーパス**と呼ぶ）に対してフィラー予測モデルを適用し，フィラーを付与したコーパスを作成.
3. フィラーを付与したコーパスから，言語モデル（トライグラム）を構築.

ここで，手順 2 においては，上記のようなコーパスからのモデル化と，出現確率からのモデル化の 2 通りの方法が考えられるが，本節では，コーパスからのモデル化について述べる．なお，出現確率からのモデル化については，6 章で述べる．

4.3.1　フィラー予測モデルの学習

最初に，学習コーパスからフィラー挿入モデルを構築する．学習コーパスに対して，個々の形態素の直後がフィラーであるか否かを表すラベルを付与した上で，フィラーを取り除く．例えば，文 (2) を学習コーパス中の文とすると，図 4.1 のような学習データが得られる．この学習データに基づいて，形態素列 X に対するラベル列 Y の条件付き確率 $P(Y|X)$ を CRF を用いて求める．CRF の学習用プログラムとしては CRF++[*1] を用いた．素性としては，形態素の表層形や品詞，読みなどを用いる．具体的には，学習データとして与えられる形態素列中の i 番目の形態素 x_i に対するラベル y_i を決定する際には，周囲の 5 つの形態素（表層形と品詞の組）x_{i-2},x_{i-1},x_i,x_{i+1},x_{i+2} の組み合わせに加え，x_i の読みに対応するモーラ列 m_i の内の終端 2 モーラを素性として用いる．たとえば，図 4.2 のようなデータの場合，図中の網掛け部が y_7 に対する素性となる．

次に，学習コーパスからフィラー選択モデルを構築する．本研究では，単純に，周囲の形態素やモーラなどの文脈 h を条件として，フィラー f が生起する条件付き確率 $P_s(f|h)$ を，フィラー選択モデルとして用いる．この条件付き確率は，学習コーパスから式 (4.2)

[*1] http://chasen.org/~taku/software/CRF++/

i	形態素 (x)		モーラ	ラベル
	表層形	品詞	(m)	(y)
1	それ	代名詞	**ソ, レ**	O
2	で	助詞	**デ**	F
3	ハワイ	名詞	ハ, **ワ, イ**	O
4	と	助詞	**ト**	O
5	いう	動詞	**イ, ウ**	O
6	の	助詞	**ノ**	O
7	は	助詞	**ハ**	F
8	火山	名詞	カ, **ザ, ン**	O
9	の	助詞	**ノ**	O
10	噴火	名詞	フ, **ン, カ**	O
11	で	助詞	**デ**	O
12	だんだん	副詞	ダ, ン, **ダ, ン**	O
13	でき	動詞	**デ, キ**	O
14	てっ	助動詞	**テ, ッ**	O
15	た	助動詞	**タ**	O
16	島	名詞	**シ, マ**	O

図 4.2 学習データの例

に基づいて求められる．なお，3 節で述べた通り，フィラーは，長音・促音の有無や語尾音節の繰り返しなどといった発音上の揺れによる派生形が生じやすい．出現頻度が非常に小さい派生形について信頼できる条件付き確率 $P(f|h)$ を推定することは困難であるため，3 節と同様にして，151 種類のフィラーを 58 種類にまとめた上でフィラー選択モデルを学習した．

4.3.2 フィラー予測モデルを用いたコーパスの変換

次に，ここまでの手順によって得られたフィラー予測モデルを用いて，開発コーパスにフィラーを挿入する．具体的には，開発コーパス中のそれぞれの形態素 $x_i (i = 1, 2, \ldots)$ に対して，以下の処理を行う．ただし，フィラーが確率的な振る舞いをすることを考慮して，以下のように一様でランダムな確率変数 Q_i, Q'_i を導入する．

1. 形態素列 X 中のそれぞれの形態素 x_i の直後にフィラーが挿入される確率 $P(y_i = \mathtt{F}|X)$ を次式により求める．

$$P(y_i = \mathtt{F}|X) = \sum_{\{Y|y_i=\mathtt{F}\}} P(Y|X). \tag{4.3}$$

一様でランダムな確率変数 Q_i（ただし，$0 \leq Q_i \leq 1$）が，$Q_i \leq P(y_i = \mathtt{F}|X)$ を満

たすとき，形態素 x_i の直後にフィラーを挿入するため，次のステップに進む．そうでなければ，次の形態素に進む．

2. あるフィラー $f_k(k = 1, 2, \ldots, |F|)$ が次式を満たすとき，そのフィラー f_k を形態素 x_i の直後に挿入する．

$$\sum_{j=1}^{k-1} P_s(f_j|h_i) \leq Q'_i < \sum_{j=1}^{k} P_s(f_j|h_i) \tag{4.4}$$

ただし，Q'_i は一様でランダムな確率変数（$0 \leq Q'_i \leq 1$），h_i は形態素 x_i 周辺の文脈である．

Q_i, Q'_i の導入により，まったく同一のコーパスを用いた場合でも，上述の手順によって作成されたコーパス中のフィラーの位置や種類は一定とはならない．よって，次節以降では，10 回の試行の結果を平均した結果を実験結果として示す．このようにして得られたフィラーを付与したコーパスから，言語モデルとして形態素 3-gram モデルを構築することは，非常に容易である．なお，実際の実験においては，頻度順に上位 20,000 語の語彙のみを用い，残りの低頻度語は未知語と見なして処理した．

4.4　日本語話し言葉コーパスを対象とする評価実験

本節では，CSJ を学習コーパスおよび開発コーパスとして用いた実験結果について述べる．評価用のテストコーパスとして CSJ の学会講演を用いる場合には，CSJ からフィラーを取り除いたコーパスを開発コーパスとして用いると，会議録や議事録を開発コーパスとして用いる場合よりも理想的な結果が得られると考えられる．

4.4.1　実験条件

CSJ の内，模擬講演の一部（1665 講演）から作成した 20,000 語からなる辞書を用いて，模擬講演と学会講演それぞれの 50 講演の未知語率を求めると，表 4.1 のように大きく異なる結果が得られる．よって，学会講演と模擬講演は，たがいにドメインの異なるコーパスと考えることができる．

表 4.1　学会講演と模擬講演の比較
（辞書は模擬講演から作成）

テストコーパス	未知語率
模擬講演	0.86 %
学会講演	2.51 %

そこで，本節の実験では，CSJ の模擬講演を学習コーパスに用い，学会講演を開発コーパスとテストコーパスの 2 つに分割して用いた．それぞれのコーパスの諸元を表 4.2 に示す．ただし，開発コーパスとして用いる学会講演については，実験前にフィラーを削除しておき，フィラーを含まない不正確な話し言葉コーパスを模擬した．

表 4.2　実験データ諸元

	学習 コーパス	開発 コーパス	テスト コーパス
ドメイン	模擬講演	学会講演	学会講演
講演数	1715	937	50
収録時間 (hour)	329.9	258.4	16.0
総文数	498k	363k	22k
総単語数	3,606k	3,109k	170k
語彙サイズ	41k	29k	8k
フィラー発生頻度	175k	174k	11k
フィラー発生率	4.8%	5.6%	6.7%

作成した言語モデルの評価には，テストコーパスに対するテストセットパープレキシティ PP と補正テストセットパープレキシティ PP^* を用いた．

また，テストセットパープレキシティ PP をフィラー部分のみについて計算した PP_F と，フィラー以外の部分について計算した PP_O も補助的な尺度として用いた．PP_F は，テストセット w_1^n 中でフィラーが n_F 回出現し，それらの集合を F とした場合，次式によって計算される．

$$H_F = -\frac{1}{n_F} log \prod_{w_i \in F} P(w_i|w_{i-2}w_{i-1}) \tag{4.5}$$

$$PP_F = 2^{H_F} \tag{4.6}$$

同様に，PP_O は，テストセット w_1^n 中でフィラー以外の単語が n_O 回出現し，それらの集合を O とした場合，次式によって計算される．

$$H_O = -\frac{1}{n_O} log \prod_{w_i \in O} P(w_i|w_{i-2}w_{i-1}) \tag{4.7}$$

$$PP_O = 2^{H_O} \tag{4.8}$$

4.4.2　フィラー挿入モデルの評価

最初に，フィラー挿入モデルのみの性能評価を行うため，フィラーの種類の違いを区別せず，全てのフィラーを同一視した実験を行った．結果を表 4.3 に示す．表 4.3 より，形態素トライグラムや，品詞トライグラム，単純なフィラーのユニグラム確率などに基づくフィラー挿入モデルと比べ，CRF に基づくフィラー挿入モデルが，すべての評価尺度において最も優れた値を達成していることが分かる．また，この値は，開発コーパスからフィラーを取り除かずに作成した場合の値（目標値）に非常に近い．よって，フィラー挿入モデルとして CRF を用いた提案手法は，実際の話し言葉に極めて近い言語モデルを再現できると言える．これらの傾向は特に PP_F において顕著であることから，各モデル間の性能差は，主にフィラーへの対応の差によるものであると言える．また，ドメインに依存しやすい名詞や動詞・形容詞の表層形を素性として用いない場合でも，性能はほとんど低下していない．これらの素性は，今回のように学習コーパスとテストコーパスでドメインが異なるようなタスクでは重要性は低いことから，フィラーの予測にほとんど寄与せず，予測性能にも影響を与えないと考えられる．CRF はこうした素性の重要性を自動学習していることから，このような素性を利用した場合でも利用しなかった場合でも，予測性能はほとんど変化しない．

表 4.3　フィラー挿入モデルの性能比較

フィラー挿入モデル	素性					フィラー頻度	PP	PP^*	PP_F	PP_O
	直前 2 形態素，直後 2 形態素および現在の形態素				挿入箇所直前の 2 モーラ					
	表層形の文字列									
	名詞	動詞/形容詞	その他	品詞						
CRF	○	○	○	○	○	152614	60.5	68.3	13.7	67.7
	×	○	○	○	○	151269	60.7	68.5	14.0	67.8
	×	×	○	○	○	153722	60.9	68.7	14.0	68.0
形態素トライグラム						134234	62.9	70.7	17.1	69.3
品詞トライグラム						155463	63.5	71.7	16.3	70.4
ユニグラム						148452	67.6	76.3	29.3	72.0
フィラーを除去していない正確な開発コーパスから作成した言語モデル						175253	59.5	67.1	10.9	67.6

次に，直前 2 形態素および直後 2 形態素の基本形の文字列・品詞と挿入個所直前の 2 モーラを素性とする CRF をフィラー挿入モデルとして用いた場合について，学習コーパスの分量とテストセットパープレキシティの関係を図 4.3 に示す．図 4.3 より，フィラーの出現位置の傾向を十分に学習するためには，200 講演（約 42 万語）程度以上の学習コーパスが必要である．なお，以上の結果はそれぞれ 10 回の試行の結果を平均したものであるが，10 回の試行における標準偏差は平均値に対して 0.05〜0.15% 程度であり，非常に小さかった．

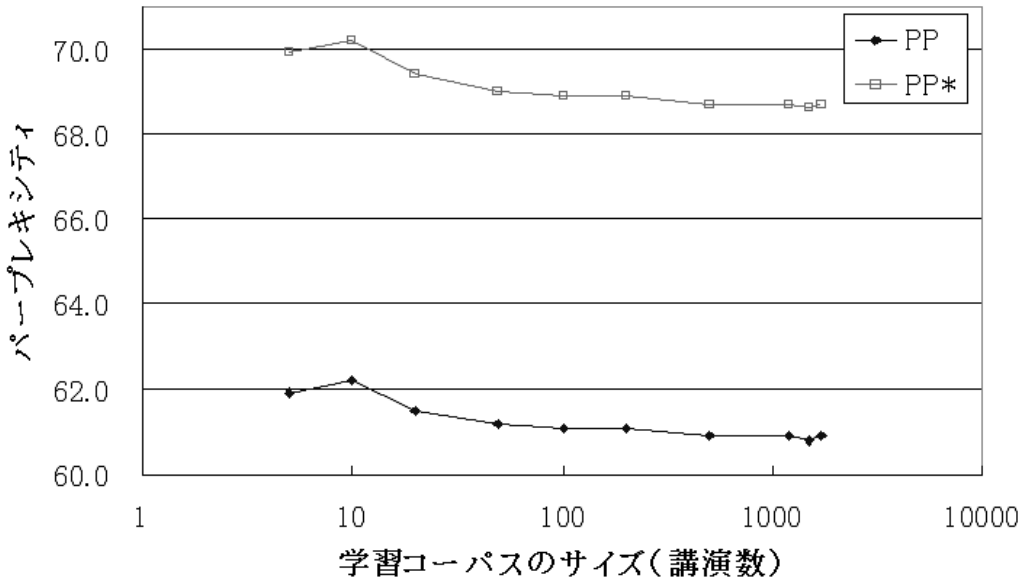


図 4.3　フィラー挿入モデルの学習曲線

4.4.3　フィラー選択モデルの評価

フィラー選択モデルは，指定された個所の周囲の文脈 h に応じて適切なフィラー f を選択するモデルである (4.2.2 節)．したがって，文脈 h に何を用いるかによって，様々なモデル化が可能である．ここでは，文脈 h として何が適当かを検討する．

用いる文脈を変化させたフィラー選択モデル間の比較の尺度として，フィラー部分のみに注目したパープレキシティ FP を用いる．

$$H(F) = -\sum_{i \in I} P(f_i) \log_2 P_s(f_i|h_i) \tag{4.9}$$

ただし $I \equiv \{w_j \in F$ である添字 $j\}$

$$FP = 2^{H(F)} \tag{4.10}$$

ここで，$P(f_i)$ はコーパスにおけるフィラー f_i の出現確率である．式 (4.9) より，$H(F)$ は，エントロピー $H(L)$ から，フィラーの予測に関係する部分のみを抜き出した評価尺度となっており，FP は，後続可能なフィラーの数を表す．

フィラーを決定する要因となっているのではないかと予想される幾つかの要因について，比較した結果を表 4.4 に示す [*2]．素朴には，フィラーは次単語の発話に詰まった場合に発生しやすい，発話中の口唇の形状によってフィラーが定まる，などという予想がある

[*2] 表 4.4 の実験に限り，日本語話し言葉コーパスに収録されている学会講演と模擬講演全体を，テストコーパスとして用いた

[68]．しかし，表 4.4 によると，直前がポーズであるといった文脈 (2) や，直前の母音や音節のみの文脈 (3),(5) を用いても，十分には FP は小さくなっていない．よって，フィラー選択モデルのモデル化にあたっては，素朴な予想では不十分であり，直前の形態素などを文脈として用いる必要があると考えられる．

表 4.4　フィラー選択のための文脈の比較

No.	文脈	$H(F)$	FP
(1)	なし (ユニグラム)	2.67	6.36
(2)	直前がポーズである	2.60	6.07
(3)	直前の母音	2.55	5.86
(4)	直前単語の品詞	2.58	5.96
(5)	直前の音節	2.47	5.56
(6)	直前の音節 ＋ 直前単語の品詞	2.41	5.32
(7)	直前の形態素	2.32	4.99
(8)	直前の 2 音節	2.34	5.06
(9)	直前の 2 音節 ＋ 直前 2 単語の品詞	2.27	4.82
(10)	直前の 2 形態素	1.83	3.54

ここで，形態素，品詞＋モーラ (音節)，モーラ，母音という順に，利用する情報を落としていった場合のパープレキシティの変化に注目すると，図 4.4 のようになっていることが分かる．図 4.4 から，モーラの情報が特に有効であることがわかる．

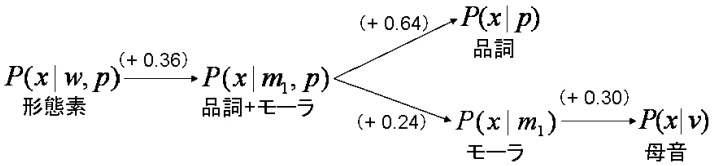


図 4.4　直前 1 つのコンテキストによる PP の比較

また，直前 2 つのコンテキストを利用した場合と，直前 1 つだけのコンテキストを利用した場合のパープレキシティを比較すると，図 4.5 のようになる．図 4.5 から，やはり直前 1 つだけよりも，直前 2 つのコンテキストを利用した方が有効であることが分かる．

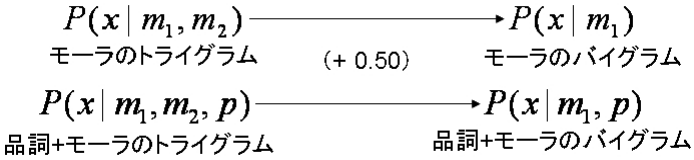


図 4.5　異なる長さのコンテキストによる PP の比較

4.4.4　フィラー予測モデルの評価

フィラー挿入モデルとフィラー選択モデルを統合した提案手法全体の評価を行うため，フィラー挿入モデルとして，CRF，形態素トライグラムや品詞トライグラムおよびユニグラムを用いた場合，および，フィラー選択モデルとして，形態素トライグラムやモーラトライグラム，品詞トライグラムおよびユニグラムを用いた場合を組み合わせた実験を行った．なお，各フィラー選択モデルのコンテキストとして，形態素トライグラムは直前２形態素を，品詞トライグラムは直前２形態素の品詞を，モーラトライグラムは直前１形態素の読みに対応するモーラ列の内の終端２モーラをそれぞれ用いた．バックオフ時には，トライグラムからはバイグラム，バイグラムからはユニグラムといったように，それぞれより短いコンテキストを用いた．結果を表 4.5 に示す．

表 4.5　フィラー予測モデルの性能比較

フィラー挿入モデル	フィラー選択モデル	フィラー頻度	PP	PP^*
CRF	形態素トライグラム	153722	70.6	79.6
	モーラトライグラム	153722	70.7	79.8
	品詞トライグラム	153722	70.5	79.6
	ユニグラム	153722	71.7	81.0
形態素トライグラム	形態素トライグラム	134234	72.7	81.8
	モーラトライグラム	134234	72.8	82.0
	品詞トライグラム	134234	72.6	81.7
	ユニグラム	134234	73.8	83.1
品詞トライグラム	品詞トライグラム	155463	73.2	82.7
ユニグラム	ユニグラム	148452	79.7	90.1
開発コーパスから言語モデル作成		175253	67.9	76.6

表 4.5 より，フィラー挿入モデルとフィラー選択モデルの両方のモデル化において，周囲のコンテキストを考慮している手法が，周囲のコンテキストを考慮していない手法（ユ

ニグラム）に比べて，優れた結果を達成していることが分かる．最も優れているのは，CRF に基づくフィラー挿入モデルと，形態素トライグラムや品詞トライグラムなどのコンテキストを考慮したフィラー選択モデルを組み合わせた場合であり，開発コーパスからフィラーを取り除かずに作成した場合の値（目標値）に非常に近い値が得られている．よって，周囲のコンテキストを考慮したフィラー挿入モデルとフィラー選択モデルを組み合わせたフィラー予測モデルによって，実際の話し言葉にかなり近い言語モデルを再現できるといえる．

なお，直前後のコンテキスト（直前 2 つの形態素，および直後 2 つの形態素）を利用した 5-gram をフィラー選択モデルとして用いることも検討したが，フィラー選択精度の改善は得られなかった（PP で 79.6，$PP*$ で 90.0）．直後のコンテキストの導入によるデータスパースネスが原因と考えられる．

4.5　国会会議録を対象とする評価実験

4.5.1　実験条件

4.3 節で述べた提案手法によって構築した言語モデルを，国会音声の認識実験で評価した．学習コーパスには前節と同様に CSJ の模擬講演を用い，開発コーパスには 1999 年から 2007 年にかけての衆議院で開かれた 1083 件の会議の会議録を用いた．また，テストコーパスとして，2007 年に衆議院で行われた会議から 4 件を選び，それぞれ 5 分ずつを抽出して，合計 20 分のデータを用意した．ここで，開発コーパスはテストコーパスにおいて発言している話者を含んでいない．各コーパスの諸元を表 4.6 に示す．

表 4.6 のコーパスを用いて，表 4.7 に示す 7 つの言語モデルを用意した．具体的には，まずベースラインとして，CSJ のデータベースに付属する **CSJ 付属モデル** [69]，国会会議録（開発コーパス）から単純に構築した**フィラーなし国会モデル**，CSJ の模擬講演から構築した**模擬講演モデル**，国会会議録と CSJ の模擬講演の混合コーパスから構築した**フィラーなし国会＋模擬講演モデル**を用意した．

これに対し，提案法のモデルとして，4.3 節の定義に基づくフィラー予測モデルを適用した国会会議録単独から学習した**フィラーつき国会（CRF）モデル**，模擬講演との混合コーパスから学習した**フィラーつき国会（CRF）＋模擬講演モデル**を用意した．さらに，より単純なフィラー予測モデルを適用した**フィラーつき国会（トライグラム）モデル**と**フィラーつき国会（ユニグラム）モデル**も比較のために用意した．

各言語モデルはいずれも形態素トライグラムモデルであり，平滑化のために Witten-Bell バックオフを適用した．なお，言語モデルの語彙は，フィラーが挿入された開発コーパスにおいて出現頻度の高かった上位 20,000 語を用いた．ただし，フィラーなし国会モ

デルではフィラー（21 語）を語彙から除いた．また，CSJ 付属モデルは他のモデルとは異なり，CSJ のコーパスにおいて 4 回以上出現した 25300 語を語彙として用いている．

音声認識用のデコーダには Julius ver. 4.0.1 を用い，音響モデルは，講演音声認識のための標準的なモデルとして CSJ に付属している，CSJ-APS,SPS を用いた [69]．このモデルは，混合連続分布 HMM（対角共分散）であり，HTK で作成されている．音素ごとに 3 状態 left-to-right HMM（飛び越し遷移なし）でモデル化を行い，音素環境依存モデル（状態共有 triphone モデル）を学習している．その際，決定木に基づく状態共有を行い，状態数 3,000 のモデル（16 混合）を学習している．各モデルには MLLR 適応のための回帰クラス情報が付加されている．このモデルの学習データは，合計 2,496 講演（486 時間）の学会講演および模擬講演である [69]．

音響分析条件は表 4.8 の通りに設定した．

表 4.6 実験データ諸元

	学習 コーパス	開発 コーパス	テスト コーパス
ドメイン	模擬講演	国会会議	国会会議
収録時間 (hour)	329.9	N/A	0.3
総単語数	3.6M	36M	3.6k
語彙サイズ	41k	55k	0.8k
フィラー発生頻度	175k	0	0.3k
フィラー発生率	4.8%	0.0%	8.3%

表 4.7 比較した言語モデル

<table>
<tr><th rowspan="2">言語モデル</th><th colspan="2">フィラー予測モデル</th><th rowspan="2">フィラー</th></tr>
<tr><th>挿入モデル</th><th>選択モデル</th></tr>
<tr><td>CSJ 付属</td><td colspan="2" rowspan="3">なし</td><td>含む</td></tr>
<tr><td>フィラーなし国会</td><td>含まない</td></tr>
<tr><td>フィラーなし国会
＋ 模擬講演</td><td>含む</td></tr>
<tr><td>フィラーつき国会
(CRF)</td><td>CRF</td><td>形態素
トライグラム</td><td>含む</td></tr>
<tr><td>フィラーつき国会
(トライグラム)</td><td colspan="2">形態素
トライグラム</td><td>含む</td></tr>
<tr><td>フィラーつき国会
(ユニグラム)</td><td colspan="2">ユニグラム</td><td>含む</td></tr>
<tr><td>フィラーつき国会
(CRF)+ 模擬講演</td><td>CRF</td><td>形態素
トライグラム</td><td>含む</td></tr>
</table>

単語辞書は，Mecab ver. 0.96（IPA 辞書 ver. 2.7.0）による形態素解析の結果から得られた単語の読みに基づいて作成した．ただし，CSJ 付属モデルと共通する語彙について

表 4.8　音響分析条件

サンプリング周波数	16kHz
プリエンファシス	0.97
分析窓	Hamming 窓
分析窓長	25ms
窓間隔	10ms
特徴パラメタ	MFCC（12 次）＋ Δ MFCC（12 次）＋ Δ パワー (計 25 次)
周波数分析	等メル間隔フィルタバンク
フィルタバンク	24 チャネル
CMS	発話単位

は，CSJ 付属モデルの発音エントリも追加した．さらに，フィラーについては，CSJ のコーパスにおいて特に出現率の高かった派生形に対応する発音エントリを追加した．これにより，発音エントリ数は 21,801 となった．なお，CSJ 付属モデルに関しては，CSJ 付属の単語辞書を使用した．これは CSJ のコーパスにおいて一定の閾値よりも高い出現率を持っていた発音エントリから構成されたものであり，発音エントリ数は 27,249 である．

言語モデルの評価尺度としては，認識実験における単語正解率と単語認識精度のほか，テストセットパープレキシティ PP と補正テストセットパープレキシティ PP^* を用いる．

4.5.2　フィラー予測モデルの評価

各言語モデルの評価結果を表 4.9 に示す．言語モデルをパープレキシティで評価した場合，まず，CSJ の模擬講演から構築した模擬講演モデルでは，テストコーパスとドメインが異なることから PP,$PP*$, 未知語率のすべてにおいて全モデル中で最も悪い結果となった．また，国会会議録から単純に構築したフィラーなし国会モデルでは，テストコーパスとドメインが一致することから PP は比較的良い結果が得られたが，フィラーがすべて未知語となることから，未知語率および $PP*$ は比較的悪い結果となった．これに対し，CSJ の模擬講演を混合したフィラーなし国会＋模擬講演モデルでは，テストコーパスとドメインが一致し，かつ，フィラーを含むモデルとなったことにより，未知語率および $PP*$ が大幅に改善した．しかし，このような従来法の混合モデルは，フィラーとドメイン内の単語にまたがるような N-gram を得ることができず，また，フィラーと同時にドメイン外の単語までもが語彙や N-gram に混入してしまうことから，性能の改善に限界が生じる．フィラーなし国会モデルと比べて PP が悪化したのも，このためであると考えられる．一方で，提案法によって構築されたフィラーつき国会モデルは，PP，$PP*$ の両方においてフィラーなし国会＋模擬講演モデルを上回った．フィラーつき国会モデルは，コーパス

に直接フィラーが挿入されていることから，ドメイン外の単語が混入することはなく，また，フィラーとドメイン内の単語にまたがるような N-gram も多数含んでいる．中でも，フィラー予測において最も長いコンテキストを考慮したフィラーつき国会（CRF）モデルは特に優れた性能を達成した．また，模擬講演を混合するとさらに性能が改善した．ここで，CSJ 付属モデルは品詞体系が他のモデルと異なるため，PP および PP^* による評価は行わなかった．

以上の結果のうち，フィラーつき国会モデルおよびフィラーつき国会＋模擬講演モデルの結果はそれぞれ 10 回の試行の結果を平均したものであるが，10 回の試行における標準偏差は平均値に対して 0.2% 程度であり，非常に小さかった．

なお，フィラーを考慮した音声認識手法として，フィラーを，後続の単語の予測に影響を与えない透過単語として扱うことが提案されている．この場合，フィラーは言語モデルの単語履歴には用いられない．しかし，このようにフィラーの有無を無視して単語予測を行っても，認識率はほとんど改善しないか，あるいはかえって悪化してしまう [24][25][26]．たとえば，西村ら [24] は，講義音声の認識において，フィラーを透過させた場合に認識率が改善することを示しているが，その改善率は非常に小さい（単語認識精度が 80.1% から 80.3% に改善）．一方で，Young ら [25] は，対話音声の認識においてフィラーを透過単語として扱った場合，パープレキシティ，単語認識精度共にかえって悪化することを示している．また，Stolke ら [26] も，フィラーを透過単語として扱った場合，特にフィラー周辺におけるパープレキシティが著しく悪化すると報告している．以上から，フィラーは話し言葉の音声認識においては重要な文脈情報であり，従って本手法のように，フィラーを積極的にモデル化する方が有効であると考えられる．

表 4.9　各言語モデルのパープレキシティと未知語率

言語モデル	語彙サイズ	フィラー頻度 (%)	PP	PP*	未知語率
CSJ 付属	25300	−	−	−	−
模擬講演	20000	4.8	114.0	226.3	12.75%
フィラーなし国会	19979	0	88.7	135.3	9.88%
フィラーなし国会＋模擬講演	20000	0.5	96.3	113.1	3.86%
フィラーつき国会 (ユニグラム)		5.5	86.2	101.2	
フィラーつき国会 (トライグラム)		4.6	86.1	101.1	
フィラーつき国会 (CRF)		3.9	83.2	97.7	
フィラーつき国会 (CRF) ＋ 模擬講演		4.1	78.6	92.3	

4.5.3　認識実験による評価

次に，これらの言語モデルを，テストデータに対する実際の認識性能で評価した．結果を表 4.10 に示す．なお，本節では，テストデータ全体に対する評価に加え，フィラーの周辺のみに限定した評価も行う．ここでフィラー周辺とは，フィラー直前の 2 単語，フィラー直後の 2 単語，およびフィラー自身を含む．

まず，フィラーなし国会モデルでは，テストデータ全体に対して比較的高い精度となったが，フィラーを含まないモデルであることから，フィラー周辺に対する精度は全モデル中で最も悪い結果となった．これに対し，CSJ 付属モデルでは，フィラーを含んだモデルであることから，フィラー周辺に対しては比較的高い精度となったが，ドメインの違いから，テストデータ全体に対する精度は全モデル中で最も悪い結果となった．これに対し，両者の混合にあたるフィラーなし国会＋模擬講演モデルは，テストデータ全体，フィラー周辺の両方に対して高い精度を達成した．しかし，前節と同様に，提案したフィラーつき国会モデルがこれをさらに上回る結果となった．特にフィラー周辺に対する精度においてベースラインとの差が顕著である．

これらの傾向は，フィラーの種類を区別した場合でも区別しなかった場合（フィラー間の混同は無視）でも変わらない．

以上の結果から，本提案手法は実際の話し言葉音声認識タスクにおいて，従来法よりも有効であることが示された．

表 4.10　各言語モデルの音声認識性能 (%)

言語モデル	語彙サイズ	未知語率	フィラーの種類を区別しない				フィラーの種類を区別する			
			全体		フィラー周辺		全体		フィラー周辺	
			Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.
CSJ 付属	25300	−	49.0	40.4	53.9	47.0	47.5	39.0	47.9	41.2
模擬講演	20000	12.75%	45.9	34.9	47.8	38.7	44.2	33.2	41.5	32.3
フィラーなし国会	19979	9.88%	54.0	49.6	35.3	31.3	54.0	49.6	35.3	31.3
フィラーなし国会＋模擬講演	20000	3.86%	57.7	51.6	48.1	41.0	57.1	51.1	45.7	39.0
フィラーつき国会 (ユニグラム)			59.2	52.5	59.7	52.1	57.4	50.7	52.3	45.0
フィラーつき国会 (トライグラム)			61.0	53.3	62.6	54.7	59.3	51.7	56.1	48.7
フィラーつき国会 (CRF)			61.3	55.0	62.9	55.3	59.7	53.4	56.5	49.2
フィラーつき国会 (CRF) ＋ 模擬講演			61.5	54.7	63.9	56.3	59.8	53.0	57.1	49.7

4.5.4　フィラー挿入モデルの挿入精度の評価

4.1 節で述べたように，国会会議録は不正確な話し言葉コーパスであり，文末の”ですね”等といった話し言葉調の表現は忠実に書き起こされている一方で，フィラー等の話し言葉特有の現象は省略されている．図 4.6 に，(a) 実際の国会会議録，(b) 人手でフィラーを書き起こした国会会議録，(c) 提案法によってフィラーを挿入した国会会議録の一例を

それぞれ示す．なお，図中の下線部がフィラーである．

図 4.6 の通り，フィラー予測モデルの適用により，フィラーの挿入を適切に行うことができる．

(a) 国会会議録：　その 中 で 今回 の ですね NHK 予算 の 審議 は ですね大臣 が 今 まで おっしゃっ て き た こと そして これから 大臣 が ですね･･･

(b) 人手でフィラーを書き起こした場合：　その 中 で 今回 の ですね<u>え</u> NHK 予算 の<u>お</u> 審議 は ですね<u>え</u> 大臣 が 今 まで おっしゃっ て き た こと そして これから<u>え</u> 大臣 が ですね･･･

(c) 提案法でフィラーを挿入した場合：　その 中 で 今回 の ですね<u>え</u> NHK 予算 の 審議 は ですね<u>まー</u> 大臣 が 今 まで おっしゃっ て<u>え</u> き た こと そして これから 大臣 が ですね･･･

図 4.6　国会会議録に対するフィラー挿入の例

表 4.11　フィラー挿入の精度と再現率

フィラー挿入モデル	フィラー選択モデル	精度	再現率	F 値
CRF	（フィラーの種類を区別しない）	0.26	0.21	0.23
形態素トライグラム		0.17	0.12	0.14
ユニグラム		0.06	0.05	0.05
CRF	形態素トライグラム	0.08	0.05	0.06
	ユニグラム	0.06	0.04	0.05
形態素トライグラム	形態素トライグラム	0.05	0.03	0.04
	ユニグラム	0.04	0.03	0.03
ユニグラム	ユニグラム	0.01	0.01	0.01

フィラー挿入の精度については，表 4.11 に示す．表 4.11 から分かるように，フィラー挿入モデルのみの評価，すなわち，フィラーの種類を区別せずにフィラーの挿入位置だけを評価した場合，今回提案した CRF に基づくフィラー挿入モデルの精度は 26% となり，フィラーの種類を区別した場合には，精度は 8% となる．この精度は一見非常に低く見えるが，フィラーは本来確率的な振る舞いをすることを考慮に入れる必要がある．例えば，

3-gram による単語の予測精度は約 17% である [70]．また，フィラー予測として周囲のコンテキストを考慮しないユニグラムを用いた場合の結果に比べると明らかに良い結果が得られていることから，フィラー挿入およびフィラー選択にあたっては，周囲のコンテキストを考慮する必要があることが分かる．

4.6　まとめ

本章では，フィラーを含む正確な話し言葉コーパスが十分に得られない状況のもとで，フィラーを考慮した言語モデルを構築するための手法として，フィラー予測モデルを用いる方法を提案した．提案手法は 2 段階からなり，最初に，正確な話し言葉コーパスからフィラー予測モデルを作成し，次に，このモデルから与えられる確率に基づいてフィラーを挿入したコーパスから言語モデルを構築した．日本語話し言葉コーパスを対象とした実験により，提案手法は，実際の正確な話し言葉コーパスから作成された言語モデルにかなり近い言語モデルを作成できることを示した．また，国会会議録を対象とした実験により，提案手法は，従来の手法よりも高い認識率を達成できることを示した．

第 5 章

音声対話応答文のためのフィラーの統計的モデリング

5.1 はじめに

人間同士の対話では，フィラーやポーズが重要な役割を果たすことが知られている．Watanabe[28] らは，フィラーやポーズの時間長が，後続するフレーズの複雑さを予測する手がかりになることを示している．Somiya[29] らは，講義音声中のフィラーが受講者の感じる聞き易さや理解のし易さに与える影響を調査している．この調査に基づいて，Naito[71] らは，決定木に基づくフィラーの使い方のアドバイスシステムを提案している．これらの先行研究にも見られるように，フィラーやポーズは，聞き手の感じる聞き易さや理解のし易さに影響を与える．

従って，こうしたフィラーやポーズの効果は，音声対話システムを実装する際にも考慮すべきである．たとえば，伊藤 [30] らは，音声対話システムにおける応答文間のフィラーについて分析を行なっている．伊藤らの分析によると，システムが次文の生成に時間を要する場合に，システムが動作していることをユーザに示すためのサインとして，応答文間でフィラーを発声させることが有効である．また，Shiwa[31] らは，人間とロボットの対話において同様の結論を示している．ユーザインタフェースの分野では，”2 秒ルール”という知見がよく知られている．すなわち，システムは，ユーザの入力を受けてから応答を返すまでに 2 秒以上を要するべきではない [32][33]．以上の先行研究からも，音声対話システムを設計する際に，フィラーやポーズの使い方を考慮することは重要であるといえる．

そこで，本研究では，音声対話システムの応答文におけるフィラーやポーズの影響を調査した．具体的には，対話コーパスの分析に基づいて，フィラーやポーズの挿入位置を人手ルール化し，応答文にフィラーやポーズを挿入した．観光案内タスクの被験者実験にお

いて，ユーザの理解度やユーザの感じる自然さと聞き易さの観点から，フィラーやポーズを挿入した応答文と，これらを含まない応答文を比較した．実験結果から，文中のフィラーが理解度や自然さを改善することが示された．なお，この評価は，実際の音声対話システムではなく，音声案内応答文の段階で行ったものである．

5.2　フィラーとポーズの挿入位置とポーズ長

フィラーやポーズは主に 2 つの要因により発生する．第一の要因は，息継ぎ等の生理学的な要因であり，これによって発生するフィラーやポーズは言語的な区切りとは無関係な位置に出現する．第二の要因は，発話のプランニングに基づくものであり，これによって発生するフィラーやポーズは情報のまとまりを区切る位置に出現する．峯松ら [72] は，音響的なポーズと，知覚的なポーズの関係を分析している．特に，知覚的なポーズの前後に出現するフィラーを分析した結果，”えー”，”えーと”，”で”等のフィラーが高確率で知覚的なポーズを発生させることが示されている．北原ら [73] は，音声の韻律やポーズが聴取者の理解度に与える影響を調査している．北原らの実験によると，軽作業の環境下のような，注意を集中しないで音声を聴取する状況において，抑揚およびポーズ情報が効果的に働き，音声の内容の理解を容易にする．このことから，ポーズは統語的な分割による文構造の明確化や，認知処理に対する時間的バッファの役割を果たすことから，注意の集中されない受聴環境下で重要な役割を果たすといえる．Watanabe[74] は，フィラーに関する仮説として，(1) フィラーは句境界や節境界，文境界，談話境界といった発話の切れ目に出現しやすく，切れ目が大きいほどフィラーの出現率が高くなるという”boundary hypothesis”と，(2) 後続の句や節が複雑であるほどフィラーの出現率が高くなるという”complexity hypothesis” の 2 つを考え，CSJ の学会講演と模擬講演の分析によってそれぞれの仮説が成り立つことを実証している．Watanabe の分析結果によると，切れ目の大きな節境界 (強境界) では，文境界と同等の割合でフィラーが出現する．また，Watanabe は，フィラーやポーズが聞き手の発話理解に与える影響についても調査している．Watanabe の実験では，物体の形状を説明する発話を被験者に聞かせ，説明に合致する物体を被験者に選択させている．この実験の結果，被験者の正解率は，先頭にフィラーのある発話では 99.2% ，先頭にポーズのある発話では 97.0% ，フィラーもポーズもない発話では 97.6% であった．また，被験者の反応速度は，先頭にフィラーのある発話とポーズのある発話が同等であり，フィラーもポーズもない発話が最も遅かった．このことから，フィラーは聞き手に発話理解の準備を促すといえる．

本研究では，文内の言語的な区切りに出現するフィラーやポーズに注目する．まず，フィラーの出現位置についてコーパスの分析を行い，その分析に基づいて，フィラーやポーズの挿入位置を決定する．

5.2.1 コーパスの分析

日本語話し言葉コーパス (CSJ) を対象として，フィラーの挿入位置に関する分析を行う．CSJ の対話データには，フィラーのタグ付けと共に，文節境界のラベリングが付与されている．そこで，対話音声におけるフィラーと文節境界の関係を調査する．

CSJ で定義される文節は，係り受け解析における基本単位としても利用され，節境界と同様に，文中における統語的・意味的な区切りの 1 つである．本節では，CSJ の対話データ 58 件を対象として，フィラーが文節境界に出現する割合を調査した．この結果，文節境界に出現したフィラーは，すべてのフィラーのうち 87.1% を占めていた．このことから，息継ぎ等の生理学的な要因による一部のフィラーを除き，多くのフィラーは意味的な区切りである文節境界に出現することがわかった．

また，3.2.7 節でも述べたように，CSJ のコアに含まれる学会講演および模擬講演では，フィラーの 41.1% が節境界に出現していた．このことからも，フィラーは，ある程度の意味的なまとまりを区切る位置に出現しやすいといえる．

5.2.2 コーパスの分析に基づく挿入位置のモデリング

コーパスの分析により，多くのフィラーは，文節境界や節境界といった統語的・意味的な区切り位置に出現しやすいことがわかった．特に，節境界のような，一定の情報のまとまりを区切る位置にフィラーを挿入することで，聞き手にとって発話内容が理解しやすくなると考えられる．

ここで，観光案内タスクにおける文の例を挙げる．実験に用いた文リストは，付録 B に示す．

> 札幌から美瑛へ行くには / 札幌駅からスーパーカムイに乗って旭川駅まで 1 時間 30 分 / 旭川駅から直行バスで美瑛駅まで 50 分です．

この例では，出発地，移動方法，到着地，移動時間の 4 つの要素が情報のまとまりを構成している．従って，このような情報のまとまりの境界部分，すなわち文中の”/”で示された位置はそれぞれ言語的な区切り位置にも対応する．これらの位置にフィラーを挿入することで，発話内容が理解しやすくなると考えられる．そこで，本研究では，このように定義される言語的な区切りの位置にフィラーやポーズを挿入する．なお，句読点とポーズ位置は必ずしも一致しないが [75]，本研究では，第一近似として一致するとした．

伊藤ら [30] の対話システムを用いた先行研究では，文間に発生する，”あ”, ”あの”, ”え”, ”えー”, ”えと”, ”えーと”, ”えっと”, ”じゃ”, ”その”, ”で”, ”ま”, ”まあ”, ”や”の 13

種類のフィラーを比較している．本研究でも，同様の種類のフィラーを対象とする．

5.3　被験者実験

5.3.1　実験方法

(a) タスク

本研究では，被験者に，画面に投影された 3D エージェントに対して，以下の対話を行なってもらう．

エージェント:　北海道について案内しますよ．
被験者:　札幌から○○にはどう行けばいいですか？
エージェント:　札幌から○○へ行くには…(案内文)．

ただし，内部的には，音声対話システムは動かさず，単純に，あらかじめ決められた質問文を被験者に発話してもらい，あらかじめ用意された案内文を再生する（Wizard of Oz）．エージェントの表示や案内文の再生には，NHK 放送技術研究所の開発した，TVML Player2 (Ver 2.3) を用いる．エージェントは発声に合わせてリップシンクを行う．

対話の開始時に，被験者には，出発地と目的地 (上記の例の”○○”) のみが示される．対話エージェントが出発地から目的地への移動経路を説明した後，被験者には，記憶している範囲で，移動経路を書き出してもらう．

また，案内文の再生と同時に，計算問題を画面に表示する．計算問題は，ランダムに生成された，1 桁の数字 2 つの加減算であり（例．$2 + 8, 4 - 1, 5 - 7$．答えが負の値になる場合もある），3 秒ごとに 1 題の間隔で，案内文の再生が終わるまで出題され続ける．被験者には，計算問題を解きながら，案内文を聴取してもらう．このような”2nd task”を導入することにより，被験者に知覚的な負荷を与え，案内文の理解 (1st task) を難しくする [76]．

被験者に与える回答用紙を図 5.2 に示す．また，被験者に提示するエージェントを図 5.1 に示す．

(b) 比較する音声応答文

本実験では，以下の 3 通りの音声応答文を比較する．

- フィラーつき音声応答文
- ポーズつき音声応答文
- フィラーのポーズ置換音声応答文 (ポーズも適宜入れて発声)

図 5.1 被験者に提示する対話エージェント

エージェントの案内音声は，名古屋大学演劇部の女性部員の録音音声を用いた．No.1〜14 の文について，「フィラーあり」の発話と，「フィラーのポーズ置換」の発話をそれぞれ独立に録音した．「ポーズあり」の発話は，「フィラーあり」の発話においてフィラーをポーズに置き換えることで生成した．各発話の話速は，時間領域調波構造伸縮 (TDHS: Time-Domain Harmonic Scaling) に基づく話速変換器である PICOLA[*1] を用いて，発話毎に 8 mora/sec に統一した．

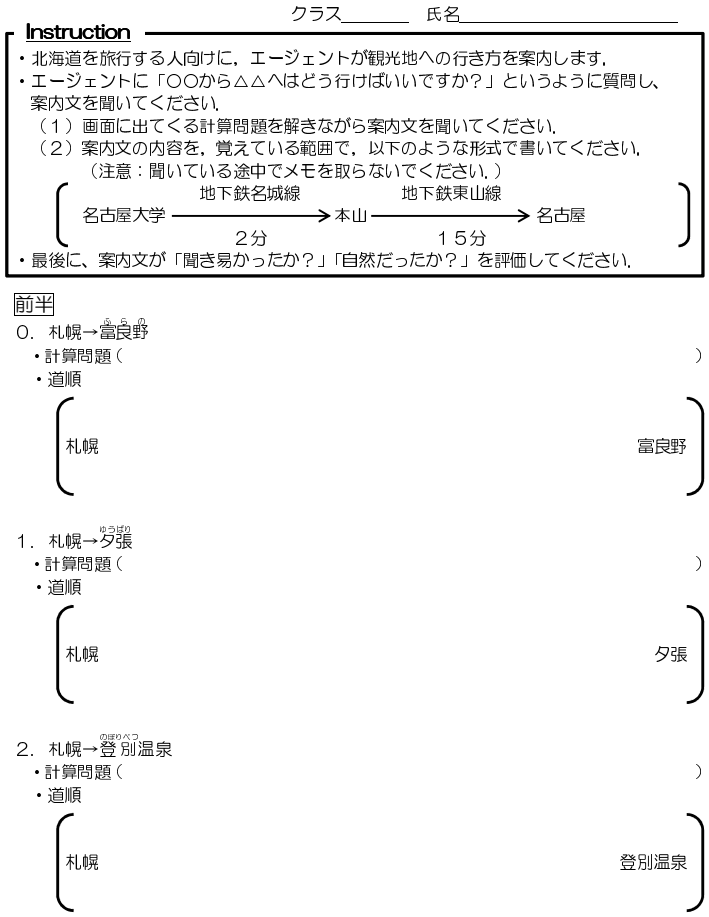
「フィラーのポーズ置換」と「フィラーあり」の案内文の例を以下に示す．また，使用した案内文の全文を付録に示す．

- 【フィラーのポーズ置換】 札幌から富良野 (ふらの) へ行くには，JR 特急で 2 時間です．
- 【フィラーあり】 札幌から富良野 (ふらの) へ行くには，えっと，JR 特急で 2 時間です．

被験者には，表 5.1 に示す No.1〜No.14 の案内文を聴取してもらう．これらの 14 文を，前半 (No.1〜7) と後半 (No.8〜14) に分け，この前半と後半で，(i) フィラーあり，(ii) ポーズあり，(iii) フィラーのポーズ置換，の条件を切り替える．すなわち，表 5.3 に示す 6 つの組み合わせを各被験者に割り当てる．

なお，「フィラーのポーズ置換」の発話にも，読み上げにおいて自然に発生するような短いポーズは含まれる．そこで，「フィラーのポーズ置換」の発話中のポーズの長さを調べた．各文中のポーズの長さを表 5.2 に示す．また，ポーズの長さの分布を図 5.3 に示す．

[*1] http://keizai.yokkaichi-u.ac.jp/~ikeda/research/picola.html


クラス_____ 氏名_____

Instruction

・北海道を旅行する人向けに，エージェントが観光地への行き方を案内します．
・エージェントに「○○から△△へはどう行けばいいですか？」というように質問し，案内文を聞いてください．
（１）画面に出てくる計算問題を解きながら案内文を聞いてください．
（２）案内文の内容を，覚えている範囲で，以下のような形式で書いてください．
（注意：聞いている途中でメモを取らないでください．）

名古屋大学 —地下鉄名城線 2分→ 本山 —地下鉄東山線 １５分→ 名古屋

・最後に、案内文が「聞き易かったか？」「自然だったか？」を評価してください．

前半

0．札幌→富良野（ふらの）
・計算問題（ ）
・道順
札幌 ⇒ 富良野

1．札幌→夕張（ゆうばり）
・計算問題（ ）
・道順
札幌 ⇒ 夕張

2．札幌→登別（のぼりべつ）温泉
・計算問題（ ）
・道順
札幌 ⇒ 登別温泉

図 5.2　被験者に与える回答用紙

図 5.3 より，「フィラーのポーズ置換」の発話でも，0.3〜0.5sec 程度のポーズが比較的多く発生していることがわかる．

(c) 評価方法

評価尺度としては，理解度・聞き易さ・自然さの 3 つを評価する．理解度は，聴取した案内文の道順をその場で書いてもらい，その道順における経由地名，移動手段，移動時間

表 5.1 案内文諸元

(a) 案内文ごとの諸元

No.	前半/後半	出発地	目的地	情報スロット数	モーラ数	単語数	理解度 (%, 平均値)	計算問題の正答率 (%)
1	前半	札幌	夕張	5	78	32	46.3	85.9
2			登別	5	96	40	46.7	83.5
3			小樽	5	92	39	32.2	88.2
4			網走	8	120	47	33.9	85.9
5			釧路	6	120	53	29.5	71.7
6			知床	8	128	53	30.7	77.7
7			宗谷	11	168	65	30.7	74.7
8	後半		旭山	5	96	39	58.3	82.3
9			美瑛	5	83	36	52.9	77.3
10			洞爺湖	5	93	39	42.5	82.7
11			函館	8	131	54	42.9	77.9
12			襟裳	8	133	57	28.1	83.0
13			五稜郭	11	164	69	31.4	80.2
14			松前	14	193	78	32.4	78.8

(b) 前半と後半の平均

前半/後半	情報スロット数	モーラ数	単語数	理解度 (%, 平均値)	計算問題の正答率 (%)
前半	6.86	114.6	47.0	35.7	81.1
後半	8.00	127.6	53.1	41.2	80.3

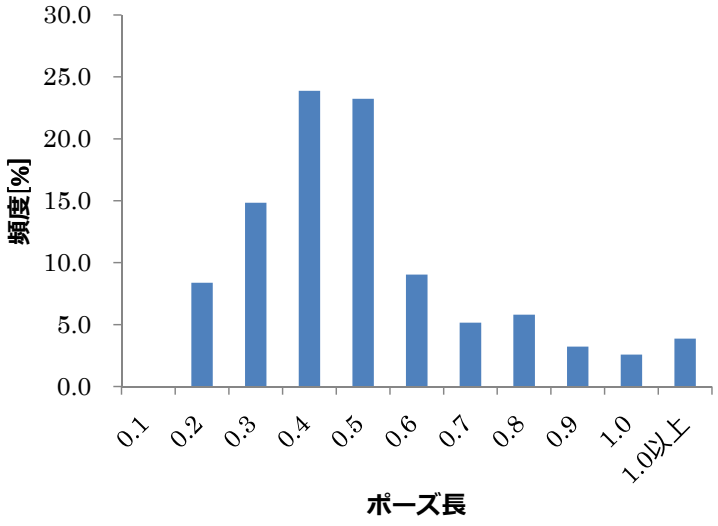


図 5.3 フィラーのポーズ置換発話のポーズ長の分布

表 5.2　フィラーのポーズ置換発話のポーズ長

No.	目的地	文全体の時間長 (sec)	ポーズの時間長 (sec)	ポーズの割合 (%)
1	夕張	12.7	2.9	22.8
2	登別	15.7	4.1	26.4
3	小樽	16.8	4.4	25.9
4	網走	23.2	4.6	20.0
5	釧路	23.0	4.8	21.0
6	知床	24.0	6.5	27.1
7	宗谷	30.6	6.5	21.1
8	旭山	16.7	2.9	17.5
9	美瑛	14.3	3.1	21.4
10	洞爺湖	16.1	3.6	22.2
11	函館	23.0	4.6	19.9
12	襟裳	23.2	6.3	27.2
13	五稜郭	31.9	8.0	25.0
14	松前	37.0	9.1	24.7

の正答率で評価する．道順の回答は，案内文を聴取した後に行い，聴取中にメモを取ったり回答することは認めない．

聞き易さと自然さは，アンケートによる 5 段階評価で評価する．絶対評価ではなく，以下のように，前半と後半の対比較で評価する．

- 5: 前半の方が自然だった (聞き易かった)
- 4: どちらかと言えば前半の方が自然だった (聞き易かった)
- 3: どちらも変わらない
- 2: どちらかと言えば後半の方が自然だった (聞き易かった)
- 1: 後半の方が自然だった (聞き易かった)

被験者には，阿南工業高等専門学校の学生 24 名を用いた．各条件ごとの被験者数を表 5.3 に示す．

表 5.3　各条件における被験者数

条件 / 前半/後半	フィラーあり	ポーズあり	フィラーのポーズ置換
前半 (No.1-7)	8	8	8
後半 (No.8-14)	8	8	8

また，各案内文の諸元を表 5.1 に示す．ここで，「情報スロット数」は，被験者が回答する経由地名，移動手段，移動時間の総数である．単語数を求める際には，形態素解析器と

して，MeCab ver. 0.963[*2] (+ UniDic ver. 1.3.12[*3]) を用いた．

表に示す通り，計算問題 (2nd task) が正答率 80% 程度の簡単なタスクであるのに対し，道順の回答 (1st task) の正答率は 40% 程度である．このことから，計算問題 (2nd task) が，案内文を理解する (1st task) 際の負荷として機能していることがわかる．なお，表 5.1 (b) に示す通り，前半と後半で，1st task の難しさにはやや違いが見られるが，2nd task の難しさにはほとんど差はない．

5.3.2 実験結果

(a) フィラー・ポーズの有無による効果

理解度，聞き易さ，自然さの評価結果をそれぞれ表 5.4, 図 5.4, 図 5.5 に示す．

表 5.4 に示すように，「フィラーあり」の条件が最も理解度が高かった．このことから，フィラーには，ユーザが文を理解し易くなる効果があるといえる．一方で，「ポーズあり」の条件では，理解度が低かった．被験者によると，ポーズが挿入されていると，発話の終わりが分かりづらいという感想があった．このことが被験者を混乱させ，理解の妨げになった可能性がある．発話の終わりが分かりやすくなるよう，ビープ音等を鳴らす必要があると考えられる．なお，以上の結果の傾向は，Watanabe による被験者実験の結果とも一致する．

表 5.4 フィラー・ポーズが理解度に与える効果

条件	フィラーのポーズ置換	フィラーあり	ポーズあり
理解度 (%)	36.8	43.4	34.0

図 5.4 に示すように，聞き易さに関しては，やや混みいった結果が得られた．「フィラーのポーズ置換」と「フィラーあり」の比較では，理解度の結果と異なり，被験者の回答が分かれた．被験者の感じる聞き易さは，実際の理解度とは異なるといえる．一方で，「フィラーあり」と「ポーズあり」の比較では，理解度の結果と同様の傾向が見られた．

図 5.5 に示すように，自然さに関しては，「フィラーあり」の条件が最も良い結果となった．また，「ポーズあり」と「フィラーのポーズ置換」の結果にほとんど差はなかった．しかし，ここで注意しなければならないのは，「フィラーあり」の文の自然さは，フィラー部分の発声の自然さに強く依存するということである．予備実験において，演劇経験のない発話者の録音音声を使用した結果，「フィラーあり」の文の自然さは非常に低く評価された．Adel[77] らも述べているように，音声合成器を用いる場合では，一層自然なフィラー

[*2] http://mecab.sourceforge.net/
[*3] http://www.tokuteicorpus.jp/dist/

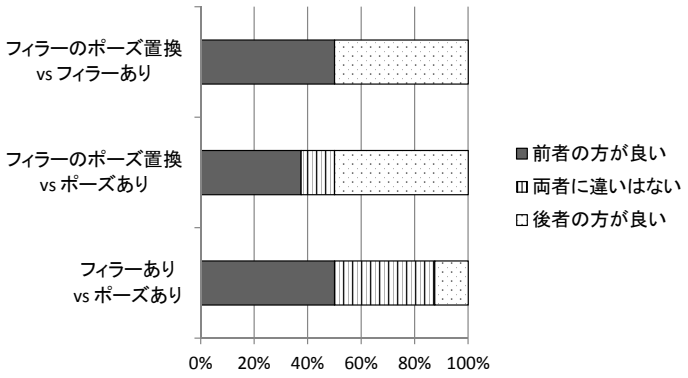


図 5.4　フィラー・ポーズが聞き易さに与える効果

の発声を生成することは極めて困難である．

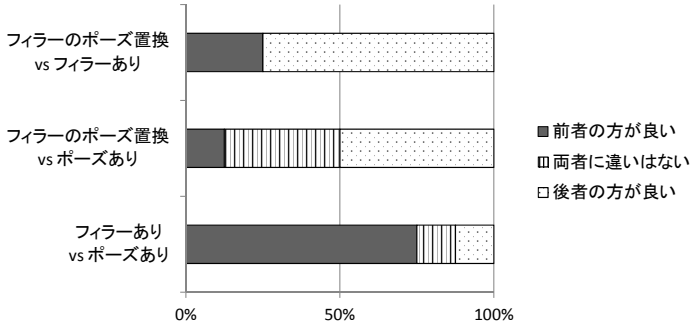


図 5.5　フィラー・ポーズが自然さに与える効果

(b) フィラー・ポーズの時間長の影響

フィラーやポーズの時間長と理解度の関係を調査した．案内文全体の時間長と，その内のフィラー・ポーズの時間長を表 5.5 に示す．フィラーの割合は，前半の文 (No.1〜No.7) では平均して 5.1% ，後半の文 (No.8〜No.14) では平均して 5.2% であり，両者に差はなかった．フィラーの割合と理解度の相関係数は -0.01 であり，ポーズの割合と理解度の相関係数は -0.10 であった．従って，フィラーやポーズの時間長比と，応答文の理解のし易さに相関は見られない．フィラーの長さと理解度の散布図を図 5.6 に，ポーズの長さと理解度の散布図を図 5.7 にそれぞれ示す．

(c) フィラー・ポーズの頻度の影響

フィラーやポーズの頻度と理解度の関係についても調査した．各案内文のフィラー・ポーズの頻度を表 5.5 に示す．フィラーの頻度と理解度の相関係数は -0.335 であり，ポーズの頻度と理解度の相関係数は -0.341 であった．また，フィラーの頻度を案内文の総単

表 5.5　各案内文の時間長

No.	目的地	文全体の時間長 (sec)	フィラーの時間長 (sec)	フィラーの時間長 (%)	フィラーの頻度
1	夕張	12.7	0.53	4.2	1
2	登別	15.7	0.31	2.0	1
3	小樽	16.8	0.95	5.7	2
4	網走	23.2	1.77	7.6	3
5	釧路	23.0	1.29	5.6	2
6	知床	24.0	0.94	3.9	2
7	宗谷	30.6	2.04	6.7	4
8	旭山	16.7	1.34	8.0	1
9	美瑛	14.3	0.83	5.8	2
10	洞爺湖	16.1	0.44	2.7	1
11	函館	23.0	1.00	4.4	2
12	襟裳	23.2	1.01	4.4	2
13	五稜郭	31.9	2.30	7.2	4
14	松前	37.0	1.48	4.0	3

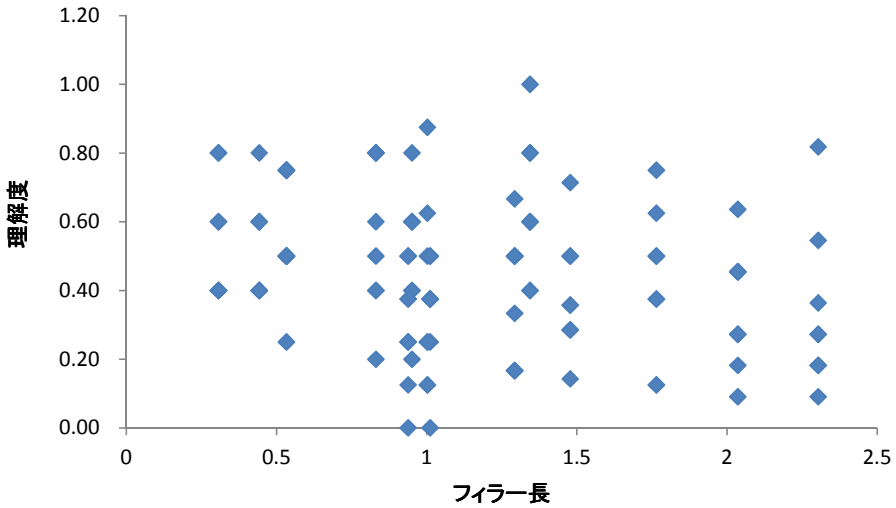


図 5.6　フィラーの長さと理解度

語数で正規化した場合，相関係数は 0.191 であり，ポーズの頻度を案内文の総単語数で正規化した場合，相関係数は-0.241 であった．フィラーの頻度と理解度の散布図を図 5.8 に，ポーズの頻度と理解度の散布図を図 5.9 にそれぞれ示す．

(d) 文の長さと理解度の関係

情報スロットの数や単語数，モーラ数と理解度の関係を調査した．情報スロットの数と理解度の相関係数は-0.264，単語数と理解度の相関係数は-0.288，モーラ数と理解度の

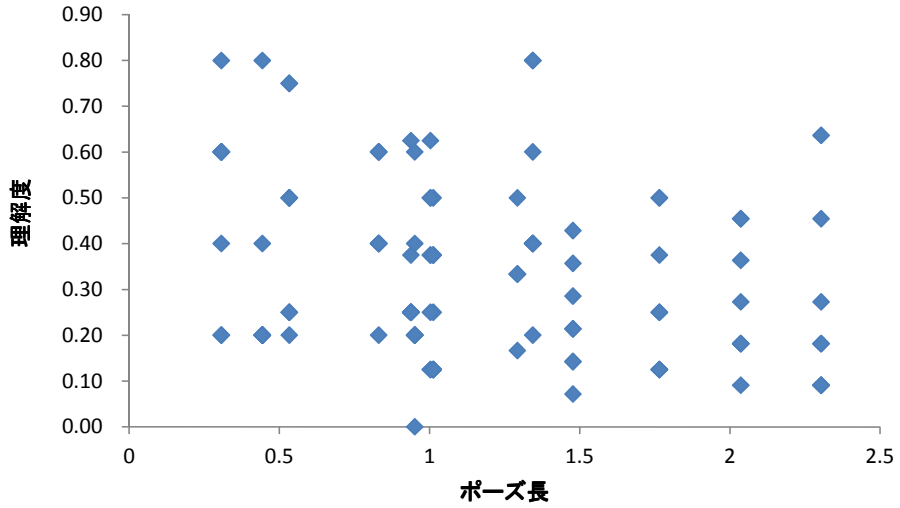


図 5.7　ポーズの長さと理解度

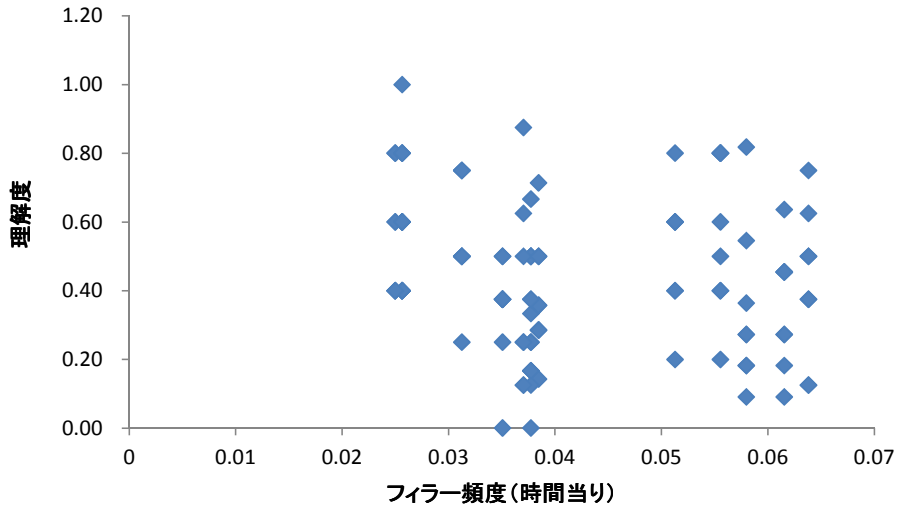


図 5.8　フィラーの頻度と理解度

相関係数は-0.300 であった．情報スロットの数と理解度の散布図を図 5.10 に，単語数と理解度の散布図を図 5.11 に，モーラ数と理解度の散布図を図 5.12 にそれぞれ示す．従って，文が長く複雑になればなるほど，ユーザは文を理解し難いことが示された．

5.4　おわりに

本研究では，音声対話システムの応答文におけるフィラーやポーズが，応答文の理解のし易さや聞き易さ，および自然さに与える影響を調査した．観光案内タスクの被験者実験を行った結果，文中のフィラーには，ユーザの理解度や応答文の自然さを改善する効果があることが示された．しかし，このようなフィラーの効果は，フィラーそれ自体の発声の

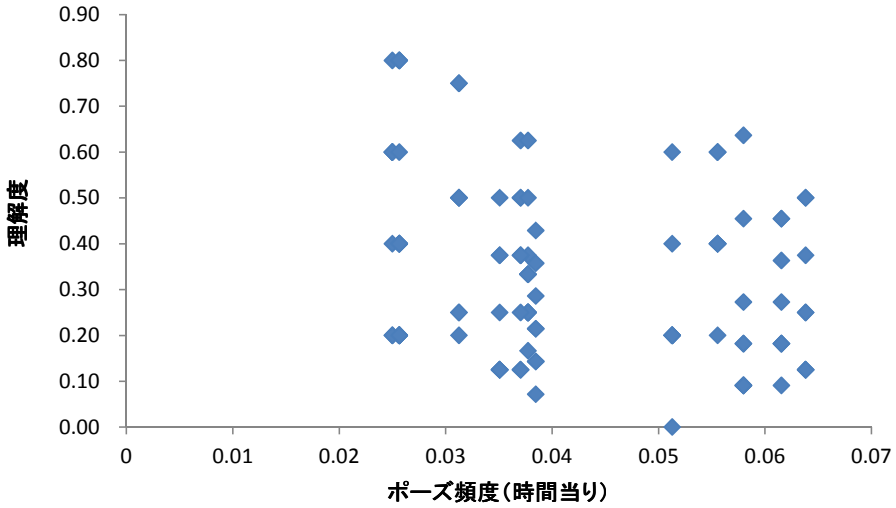


図 5.9　ポーズの頻度と理解度

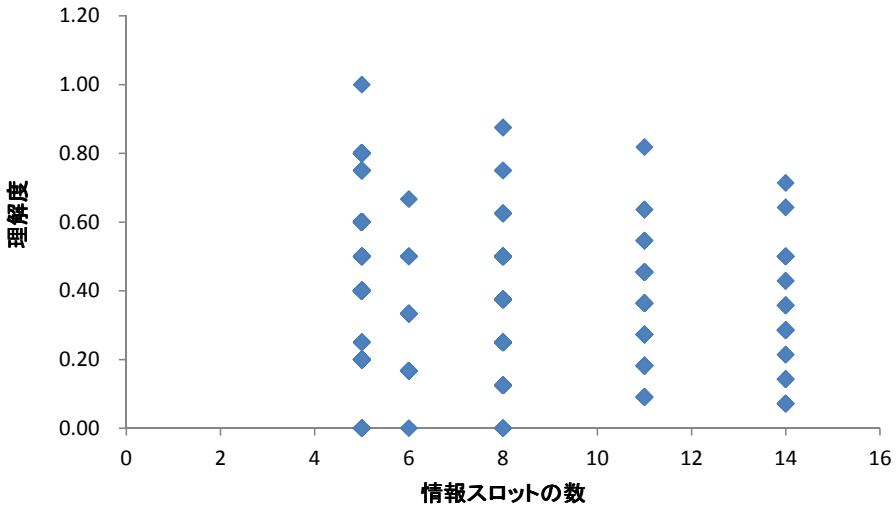


図 5.10　情報スロットの数と理解度

自然さに大きく依存することに注意する必要がある．

今後の課題としては，まず，2nd task を課さなかった場合との比較実験が考えられる．北原らによると，軽作業の環境下のように，聞き取りに集中できる受聴環境下と，そうでない環境下では，韻律やポーズの与える効果が異なる．従って，フィラーの与える影響も，2nd task のような負荷がある場合と，ない場合とで異なる可能性がある．次に，音声応答文から韻律を除去した場合との比較も行う必要があると考えられる．北原らの示すように，韻律情報は，聞き手にとって発話内容の理解の準備を促す効果がある．我々のタスクでも，韻律情報の有無によって，聞き手の理解度や聞き易さが影響を受ける可能性がある．また，合成音声を用いた場合との比較も考えられる．音声の理解度や聞き易さは，音

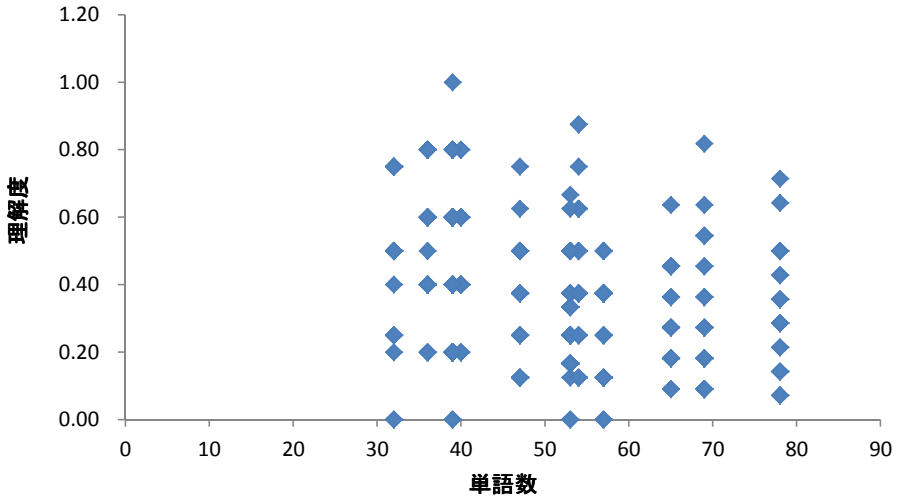


図 5.11　単語数と理解度

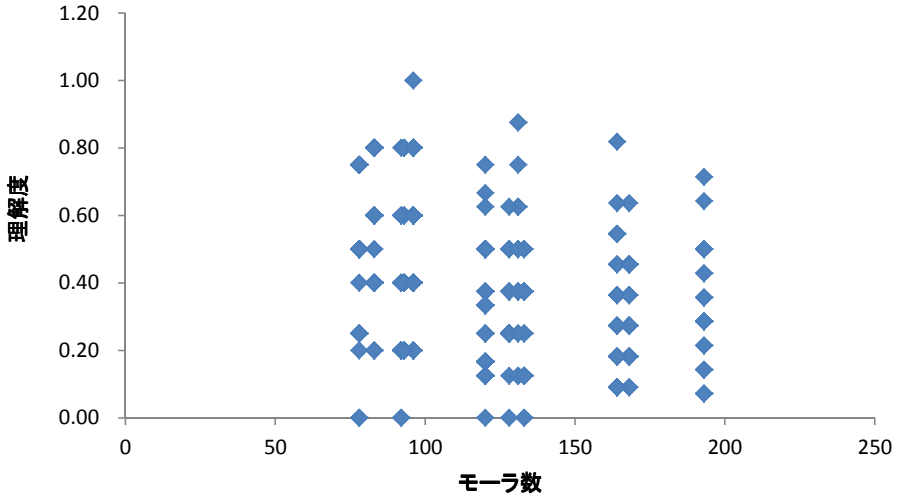


図 5.12　モーラ数と理解度

声それ自体の自然性に影響されると考えられる．従って，合成音声を用いることにより，フィラーやポーズの効果がどのように変化するかを検証する必要がある．さらに，フィラーをポーズに置き換えた実験のやり直しも必要である．発声時間の長いフィラーは長いポーズに置き換えたため，不自然になったと考えられる．

第 6 章

話し言葉音声認識のためのポーズの統計的モデリング

6.1 はじめに

話し言葉の音声においては，読み上げ原稿が用意されていない状況で自発的に発声がなされることから，話者の思考状態や息継ぎ等の生理的現象により，句読点が付与されるような言語的な区切りとは無関係な位置にポーズが出現する．たとえば，西光ら [78] によれば，日本語話し言葉コーパスに収録されている講演音声において，約半数のポーズが言語的な区切り（節境界）[66] とは異なった位置に出現する．また，中川ら [65] によれば，対話音声において，約 13% のポーズが文節の内部に出現する．このように，話し言葉の音声では，ポーズの出現位置と句読点とは必ずしも対応しない．さらに，言語的な区切りとは無関係な位置に出現するポーズは，しばしば音声認識システムの誤認識を引き起こす要因となる．たとえば，一般的な音声認識システムは，発話音声をポーズによって分割してから音声認識を行うため，言語的な区切りと無関係な位置のポーズによって発話音声が分割されてしまうと，言語的な制約が効きにくくなり誤認識が増加する．また，ポーズ周辺の単語の予測には，ポーズを単語履歴として使用するため，ポーズのモデリングが悪いと単語予測にも悪影響が及ぶ．

こうした問題に対処するためには，句読点よりもポーズを考慮した言語モデルが必要である．そのような言語モデルを構築する最も簡単な手法は，ポーズ情報を含むコーパスから，ポーズの生起確率を含むような言語モデルを学習する手法である．たとえば南條ら [37] は，日本語話し言葉コーパスから言語モデルを学習する際の，言語モデル上のポーズのモデル化について検討を行っており， 1000msec 以上のポーズを認識処理単位を区切るロングポーズ，1000msec 未満のポーズを認識処理単位として区切らないショートポーズとして扱った場合に最も良い認識率を得ている．また，西村ら [24] は，30msec 以上の無

音区間を読点として書き起こした講義音声コーパスから単語 3-gram モデルを学習し，この 3-gram 確率を用いて，音声認識時にポーズの出現予測を行っている．これらの手法はいずれも，ポーズ情報が付与されたコーパスが利用できることを前提としている．しかし，実際には，そのようなコーパスが利用出来るドメインは極めて稀である．

そこで，本章では，ポーズ情報が付与されていないコーパスからポーズを積極的に考慮した言語モデルを構築する手法を提案する．提案手法では，ポーズ情報を補うためのポーズ挿入モデルを条件付き確率場（Conditional Random Field．以下，CRF）[67] に基づいて構築し，この挿入モデルを用いて，ポーズを考慮した言語モデルを構築する．本手法は，確率モデルに基づいてポーズの予測を行うという点では西村ら [24] の手法と共通しているが，ポーズ情報が付与されたコーパスが利用できないドメインの話し言葉を対象とする音声認識用言語モデルの構築において広く適用が可能である．増村ら [42] は，Web から収集した話し言葉に近いコーパスを対象として本章の提案手法を適用し，話し言葉の音声認識に有用であると報告している．また，話し言葉の音声からのポーズの検出は，句読点や節境界などの言語的な区切りの検出に比べると容易である．そのため，ポーズを積極的にモデル化する提案手法は，入力音声とモデルの学習データからポーズを除去して音声認識を行う方法 [79] や，ポーズを透過単語 [*1] として処理して後続単語を予測する時の単語履歴に含めない方法よりも効果的であることを示す（6.4 節参照）．

6.2　ポーズ予測モデルの定式化

本章では，ショートポーズの挿入モデルを，形態素列を対象とし，個々の形態素に対して，その直後にショートポーズを挿入するべきかどうかという二値のラベルを付与する，系列ラベリング問題として定式化する [80]．具体的には，図 6.1 のように，個々の形態素に対して，直後にショートポーズを挿入すべきである場合にはラベル SP を，そうでない場合にはラベル O を付与する系列ラベリング問題を考える．

本章では，このような問題を解くショートポーズ挿入モデルを，CRF[67] を用いて構築

形態素列	そこ	で	本	研究	の	···
	（文頭） 代名詞	助詞	接頭辞	名詞	助詞	
	（文頭） ソコ	デ	ホン	ケンキュウ	ノ	
ラベル列	O	O	SP	O	O	O ···

図 6.1　ショートポーズ挿入モデルの学習用ラベル

[*1] たとえば，西村ら [24] は，言い直しに起因する語断片やフィラーなどの不要語を透過単語として扱っている．

する．CRF は，隠れマルコフモデルなどのモデルと比較して柔軟な素性設計が可能であり，また，比較的少量の学習データでも高い性能を示すことが知られている識別モデルである．CRF では，形態素列 x_1^L に対するラベル列 y_1^L の条件付き確率 $P(y_1^L|x_1^L)$ を，次式のように表す．

$$P(y_1^L|x_1^L) = \frac{1}{Z(x_1^L)} \exp\left(\sum_a \lambda_a f_a(x_1^L, y_1^L)\right) \tag{6.1}$$

ここで，L は系列の長さ，f_a は素性関数，λ_a は素性関数に対する重み，$Z(x_1^L)$ は正規化項をそれぞれ表す．なお，CRF の学習用プログラムとしては CRF++[*2] を用い，CRF の学習時には，事前分布として Gaussian Prior を用いて事後確率を最大化することによってパラメータを正則化した．また，素性情報としては，各形態素の表層形や品詞，読みなどを用いた．具体的には，学習データとして与えられる形態素列中の i 番目の形態素 x_i に対するラベル y_i を決定する際には，周囲の 5 つの形態素（表層形と品詞の組）x_{i-2},x_{i-1},x_i,x_{i+1},x_{i+2} の組み合わせに加え，x_i の読みに対応するモーラ列 m_i の内の終端 2 モーラ（もしくは 1 モーラ）を素性として用いた．たとえば，図 6.2 のようなデータの場合，図中の網掛け部が y_7 に対する素性となる．

6.3 コーパスへのショートポーズ挿入に基づく言語モデルの構築

本節では，言語的まとまり以外の要因に基づくポーズ（ショートポーズ，SP）の出現位置を，話し言葉音声コーパスに基づいて学習したモデルによって補う方法について述べる．

6.3.1 ショートポーズ挿入モデルに基づく言語モデルの構築方法

ショートポーズの挿入モデルに基づいて言語モデルを構築するには，(a) コーパスからのモデル化と (b) 出現確率からのモデル化の 2 通りの方法が考えられる．

(a) コーパスからのモデル化

4.3.2 節と同様の方法により，コーパスに対してあらかじめショートポーズを挿入し，ショートポーズの挿入されたコーパスから言語モデルを学習する．コーパスへのショートポーズの挿入は，ショートポーズの出現が確率的な振舞いを取ることを考慮し，以下の手順で行う．まず，形態素列 X 中のそれぞれの形態素 x_i の直後にショートポーズが挿入さ

[*2] http://chasen.org/~taku/software/CRF++/

i	形態素（x）		モーラ	ラベル
	表層形	品詞	（m）	（y）
1	それ	代名詞	**ソ, レ**	O
2	で	助詞	**デ**	O
3	ハワイ	名詞	ハ, **ワ, イ**	O
4	と	助詞	**ト**	O
5	いう	動詞	**イ, ウ**	O
6	の	助詞	**ノ**	O
7	は	助詞	**ハ**	SP
8	火山	名詞	カ, **ザ, ン**	O
9	の	助詞	**ノ**	O
10	噴火	名詞	フ, **ン, カ**	O
11	で	助詞	**デ**	O
12	だんだん	副詞	ダ, ン, **ダ, ン**	O
13	でき	動詞	**デ, キ**	O
14	てっ	助動詞	**テ, ッ**	O
15	た	助動詞	**タ**	O
16	島	名詞	**シ, マ**	O
17	が	助詞	**ガ**	SP
18	こう	副詞	**コ, ウ**	O

図 6.2　学習データの例

れる確率 $P(y_i = SP|X)$ を次式により求める．

$$P(y_i = SP|X) = \sum_{\{Y|y_i=SP\}} P(Y|X) \tag{6.2}$$

次に，一様でランダムな確率変数 $Q_i (0 \leq Q_i \leq 1)$ を導入し，これが $Q_i \leq P(y_i = SP|X)$ を満たすとき，形態素 x_i の直後にショートポーズを挿入する．こうしてショートポーズの挿入されたコーパスから，言語モデルを構築する．

なお，実際のモデル化においては，ショートポーズの挿入を M(=1,10,100) 回行い，その結果得られる M 個のコーパスから N-gram カウントを求めた．ただし，M 個のコーパスの N-gram カウントの総和をそのまま用いると，低頻度事象について，式 (6.17) 中の $f(h, w)$ と $r(h)$ の大小関係に不整合が生じ，過度に大きな確率が割り当てられてしまう問題がある．よって，本研究では式 (6.3) のように，M 個のコーパスの N-gram カウントを平均して用いた．ここで，$f_k(w_{i-N+1}^i)$ は，k 番目のコーパスにおける単語列 w_{i-N+1}^i の出現頻度である．また，N-gram のカットオフは，N-gram カウントを平均した後に適用した．

$$f(w_{i-N+1}^{i}) = \frac{1}{M}\sum_{k=1}^{M} f_k(w_{i-N+1}^{i}) \tag{6.3}$$

(b) 出現確率からのモデル化

ショートポーズの挿入モデルが与える挿入確率に基づいて，N-gram 言語モデルを直接推定することができる．まず，森ら [81] の方法を参考にして，N-gram 形態素列の出現頻度を推定する．学習コーパスの形態素列を x_1^L とすると，形態素 w の 1-gram 頻度 $f(w)$ は，次式のように書くことができる．

$$f(w) = \sum_{i} \delta(x_i = w) \tag{6.4}$$

ただし，δ はクロネッカーのデルタであり，括弧内の条件が満たされれば 1，満たされなければ 0 の値を取る．これに対し，形態素列 x_1^L には明示的には出現しないショートポーズ <*sp*> [*3] の 1-gram 頻度 $f(<sp>)$ を，次式のように定義する．

$$f(<sp>) = \sum_{i} P(y_i = SP|x_1^L) \tag{6.5}$$

ここで，$P(y_i = SP|x_1^L)$ は，形態素列 x_1^L の i 番目の形態素 x_i の直後にショートポーズが挿入される確率である．ショートポーズ挿入モデルとして CRF を用いる場合は，式 (6.1) を用いて次式のように求める．

$$P(y_i = SP|x_1^L) = \sum_{\{y_1^L|y_i=SP\}} P(y_1^L|x_1^L) \tag{6.6}$$

式 (6.4) と式 (6.5) を用いると，0-gram 頻度 $f(\cdot)$ は，次式のように定義できる．

$$f(\cdot) = f(<sp>) + \sum_{w} f(w) \tag{6.7}$$

以上を用いると，形態素 w の 1-gram 確率 $P(w)$ は，次式によって求められる．

$$P(w) = \frac{f(w)}{f(\cdot)} \tag{6.8}$$

また，2-gram 頻度 $f(w, w')$ は，学習コーパス中に形態素 w と w' が連続して出現し，かつ，その形態素間に <*sp*> が挿入されなかったという全ての事象の頻度であり，次式のように定義する．

$$f(w, w') = \sum_{i} \delta(x_{i-1} = w)\delta(x_i = w')(1 - P(y_i = SP|x_1^L)) \tag{6.9}$$

[*3] ここで，"SP"はショートポーズの挿入モデルが与えるラベルであり，"<*sp*>"は 1 形態素としてのショートポーズである点に注意されたい．

これに対し，2-gram 頻度 $f(w, <sp>)$ は形態素 w の直後にショートポーズ <sp> が挿入される頻度であり，2-gram 頻度 $f(<sp>, w)$ は形態素 w の直前にショートポーズ <sp> が挿入される頻度である．それぞれ以下のように定義する．

$$f(w, <sp>) = \sum_i \delta(x_i = w) P(y_i = SP | x_1^L) \tag{6.10}$$

$$f(<sp>, w) = \sum_i \delta(x_i = w) P(y_{i-1} = SP | x_1^L) \tag{6.11}$$

同様に，3-gram 頻度はそれぞれ以下のように定義する．ここで，簡単のために $P(y_i = SP|x_1^L)$ を P_i と表記する．

$$f(w, w', w'') = \sum_i \delta(x_{i-2} = w)\delta(x_{i-1} = w')\delta(x_i = w'')(1 - P_{i-2})(1 - P_{i-1}) \tag{6.12}$$

$$f(w, w', <sp>) = \sum_i \delta(x_{i-1} = w)\delta(x_i = w')(1 - P_{i-1})P_i \tag{6.13}$$

$$f(w, <sp>, w') = \sum_i \delta(x_{i-1} = w)\delta(x_i = w')P_{i-1} \tag{6.14}$$

$$f(<sp>, w, w') = \sum_i \delta(x_{i-1} = w)\delta(x_i = w')P_{i-2}(1 - P_{i-1}) \tag{6.15}$$

$$f(<sp>, w, <sp>) = \sum_i \delta(x_i = w)P_{i-1}P_i \tag{6.16}$$

以上の N-gram 頻度から，N-gram 確率 $P(w|h)$ を求めることができる．本章では，Witten-Bell バックオフ [58] を適用して，次式のように $P(w|h)$ を求める．

$$P(w|h) = \begin{cases} \frac{f(h,w)}{f(h)+r(h)} & \text{if } f(h,w) > c \\ \frac{r(h)}{f(h)+r(h)} \cdot P(w|h') & \text{otherwise} \end{cases} \tag{6.17}$$

ここで，h は履歴であり，h' は 1 形態素バックオフした履歴である．また，$r(h)$ は，履歴 h の直後に出現する形態素の種類数である．ただし，$r(h)$ を $f(h, w) > 0$ であるような形態素 w の種類数と定義すると，$f(h, w) < 1$ の場合も含まれることになるため，$r(h)$ が極端に大きくなるという問題が生じる．そこで，本章では，$r(h)$ を，N-gram カットオフの閾値 c を用いて $f(h, w) > c$ を満たすような形態素 w の種類数とする．

なお，式 (6.6) では，ショートポーズ挿入モデルとして CRF を用いる場合について述べたが，実際には，ショートポーズ生起確率を十分な精度で予測できる任意の確率モデルを，ショートポーズ挿入モデルとして用いることができる．例として，ショートポーズ挿

入モデルとして，ショートポーズを語彙に含む形態素 3-gram モデルを用いる場合を考える．この場合は，ショートポーズが挿入される確率が直前 2 形態素のみによって定まると仮定し，次式のように近似する．

$$P(y_i = SP|x_1^L) \simeq P_{trigram}(SP|x_{i-2}, x_{i-1}) \tag{6.18}$$

この式を，式 (6.7)∼ 式 (6.16) に対して適用すると，ショートポーズ挿入モデルとして形態素 3-gram モデルを用いて言語モデルを構築することができる．

6.3.2 従来法によるポーズ予測モデル

本研究では，従来法によるポーズ予測モデルとして，以下の手法を比較する．

- 形態素トライグラムに基づくポーズ挿入モデル
- ユニグラム確率に基づくフィラー挿入モデル (単語間にランダムにポーズを挿入)

形態素トライグラムに基づくポーズ予測モデルでは，直前 2 形態素を考慮したトライグラムモデルに基づいてポーズの挿入位置を決定する．なお，トライグラムモデルの平滑化のために，Witten-Bell バックオフを用いる．ユニグラム確率に基づくフィラー挿入モデルでは，文脈を考慮せず，単純に，ユニグラム確率に基づいてフィラーの挿入位置を決定する．

6.3.3 ポーズ単位とショートポーズを利用した音声認識

多くの音声認識システムでは，まずロングポーズによって入力音声を分割し，分割された音声を処理単位として，処理単位毎に独立に音声認識を行う．すなわち，処理単位の先頭の語は，直前の処理単位の末尾に依存しないと仮定されている．しかし，実際の話し言

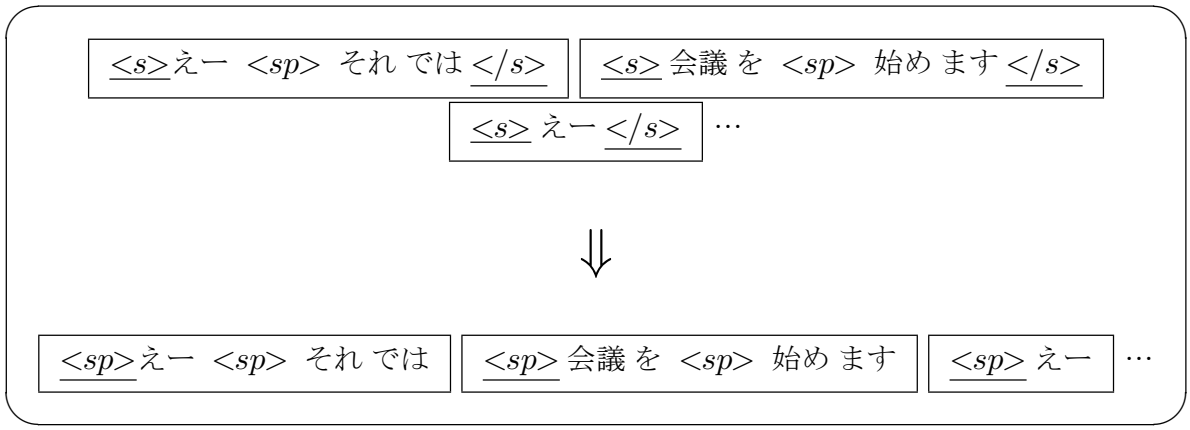


図 6.3 ショートポーズの認定

葉の音声では，言語的な区切りと無関係な位置にも多数のポーズが出現するので，この仮定は成り立たないことも多い．そのため，話し言葉の音声認識においては，処理単位間の依存性を考慮した手法が必要となる．

処理単位間の依存性を考慮するには，処理単位の境界のポーズを，処理単位の始終端 <s>,</s> ではなく，ショートポーズ <sp> として扱う方法がある．例を図 6.3 に示す．単語 3-gram 言語モデルを使う場合，図 6.3 の単語「会議」の確率として，P(会議 | </s>,<s>) ではなく P(会議 | では,<sp>) を使うことになる．南條ら [37] は，1000msec 未満の長さのポーズを <sp> として扱って処理単位間の依存性を考慮し，1000msec 以上の長さのポーズを <s>,</s> として扱って処理単位が独立であると仮定する方法を提案している．予備実験の結果，一部の処理単位境界を <s>,</s> とするモデルよりも，全ての処理単位境界を <sp> とするモデルの方が，パープレキシティが低かった．これは，図 3.5 に示したように，閾値よりも長いロングポーズであっても言語的区切り以外の場所に出現している場合があり，処理単位間の言語的依存関係を無視できない場合があるからだと考えられる．そのため，本章では，全ての処理単位境界を <sp> として扱うことにする．

なお，実際の認識の際には，各認識処理単位（図 6.3 の各枠）ごとに認識結果を確定させながら認識処理を進める．この認識処理単位は，パワーが閾値以下であるようなフレームが 200msec 以上継続したことを手がかりとして決定する．

6.3.4　評価基準

提案手法に基づいて構築された言語モデルの評価は，テストセットパープレキシティと補正テストセットパープレキシティに加え，音声認識結果の単語正解率および単語認識精度に基づいて行う．

なお，ショートポーズ <sp> および認識処理単位の始終端 <s>,</s> の頻度の違いによるテストセットパープレキシティの変化を無視し，一般の単語についてのテストセットパープレキシティのみを評価するため，ショートポーズ <sp> および認識処理単位の始終端 <s>,</s> は，単語予測の履歴にのみ使用し，テストセットパープレキシティの計算には含めない．そのため，履歴 h に対する単語 w の確率として，式 (6.20) によって定義された $P'(w|h)$ を用いて正規化を行う．

$$P'(w|h) = \begin{cases} 0 & \text{if } w \in ccs \\ \alpha(h) \cdot P(w|h) & \text{otherwise} \end{cases}, \tag{6.19}$$

$$\alpha(h) = \frac{1}{1 - \sum_{w \in ccs} P(w|h)}, \tag{6.20}$$

$$ccs = \{<s>, </s>, <sp>\} \tag{6.21}$$

ここで，$P(w|h)$ は従来の言語モデル確率である．また，ccs はコンテキストキューの集合であり，ショートポーズ <sp> および認識処理単位の始終端 <s>,</s> がこれに含まれる．

6.4 国会会議録を対象とする評価実験

本節では，6.3 節で述べた手法によって構築される言語モデルを，国会審議の音声認識実験によって評価する．さらに，ポーズの挿入とフィラーの挿入の関係についても分析を行う．

6.4.1 実験条件

6.3 節で述べたショートポーズ挿入モデルを，国会会議録を対象とした実験によって評価した．具体的には，CSJ の学会・模擬講演 (表 6.1 の学習セット) から学習したショートポーズ挿入モデルと，フィラー予測モデル [82] によってフィラーを自動挿入した国会会議録 (表 6.1 の開発セット) を組み合わせて，出現頻度の上位 20,000 語の語彙からなる形態素 3-gram モデルを構築した．平滑化のため，式 (6.17) によって定義される Witten-Bell バックオフを適用し，N-gram カットオフの閾値 c は 1 とした．また，ショートポーズ挿入モデルの学習には，200msec 以上・1000msec 未満のポーズを用いた．これは以下の 2 つの理由による．第 1 に，図 3.5 より，1000msec 以上のポーズは，過半数が言語的なまとまりに対応するポーズである．第 2 に，200msec 未満のポーズについては，CSJ に位置情報が付与されていなかった．この影響については，6.4.2 節で述べる．

この他に，読点の一部あるいは全部をショートポーズとして扱ったモデル，および，単語間にランダムにショートポーズを挿入したモデルと比較した．さらに，句点（文境界）を <sp> としたモデルについても評価を行った．

各言語モデルは，6.3.4 節で述べた通り，テストセットパープレキシティ（PP），補正テストセットパープレキシティ（$PP*$），および音声認識実験における単語正解率（$Cor.$）と単語正解精度（$Acc.$）で評価した．音声認識用のデコーダは我々の研究室で開発している SPOJUS（3-gram 言語モデルに基づく 1 パスデコーダ）[83] を用い，音響モデルは CSJ から学習した左コンテキスト依存音節モデル（left-to-right 型 HMM，5 状態 4 出力分布，全共分散行列からなる 4 混合ガウス分布/出力分布）を用いた．左コンテキスト依存音節モデルでは，116 種類の音節に対して 8 種類の先行音素（/a/, /i/, /u/, /e/, /o/, /N/, /文頭/, /促音/）を考慮しているため，モデル数は 928 である [84]．音響分析条件は表 4.8 と同じである．言語重みと挿入ペナルティは各言語モデルで共通の固定値を用いた．テストセットには，2007 年に衆議院で行われた会議から，それぞれ異なる議題につ

いての会議 4 件を選び，100 秒以上の発話が記録されている 12 名の男性話者 [*4] による発話 (88 分) を人手で書き起したテキストを用意した．形態素数 · 語彙サイズを表 6.1 のテストセット欄に，各会議の詳細を表 6.2 に示す．

表 6.1　実験データ諸元

	学習セット		開発セット	テストセット
種類	CSJ (学会講演)	CSJ (模擬講演)	国会会議録 (衆議院, 1999 年 ∼2007 年)	国会会議録 (衆議院，2007 年)
講演数	967	1,705	1,083	4
形態素数	3,194K	3,584K	38,668K	21K
語彙サイズ	37K	48K	58K	2K

表 6.2　テストセット諸元

ID	委員会名称	開催日	議題	概要	話者数
T1	第 166 回国会 総務委員会 第 9 号	H19 年 3 月 15 日	放送法第三十七条第二項の規定に基づき，承認を求めるの件	日本放送協会の受信料義務化	3
T2	第 166 回国会 予算委員会 第 8 号	H19 年 2 月 14 日	平成十九年度一般会計予算，平成十九年度特別会計予算，平成十九年度政府関係機関予算	北海道夕張市の財政再建	3
T3	第 166 回国会 予算委員会 第 12 号	H19 年 2 月 20 日	平成十九年度一般会計予算，平成十九年度特別会計予算，平成十九年度政府関係機関予算	輸入牛肉の BSE 対策，北朝鮮に関する六カ国協議	4
T4	第 166 回国会 内閣委員会 第 10 号	H19 年 4 月 4 日	株式会社日本政策金融公庫法案及び株式会社日本政策金融公庫法の施行に伴う関係法律の整備に関する法律案	中小企業の金融支援，起業支援，セーフティネット	4

[*4] 表 6.2 の話者数の合計は 14 名であるが，2 名が重複しているためである．

6.4.2 学習データの制限による影響

前述したように，本実験では，200msec 以上・1000 msec 未満のポーズに基づいてショートポーズ挿入モデルを学習しており，200msec 未満のポーズは学習に含まれていない（学習に使用した CSJ では，200msec 未満のポーズの位置情報は付与されていないため）．それにもかかわらず，テストセットに出現した 200msec 以上のポーズと，200msec 未満のポーズに対して，ショートポーズ挿入モデルによるポーズの予測確率を比較してみたところ，前者が平均して $p = 0.192$ 程度であったのに対し，後者は $p = 0.364$ 程度であった．このように，200msec 以上のポーズに基づいてショートポーズ挿入モデルを学習した場合でも，200msec 未満のポーズに対して，200msec 以上のポーズよりも高い予測確率が割り当てられていた．よって，200msec 以上のポーズに基づいてショートポーズ挿入モデルを学習しても影響はないといえる．

6.4.3 言語モデルの構築方法の比較

予備実験として，6.3 節で述べた 2 つの手法 ((a) コーパスからのモデル化と (b) 出現確率からのモデル化) の比較を行った．なお，本予備実験のテストセットは，表 6.2 に示したものとはやや異なる (100 秒未満の発話も用いている)．

各手法の比較結果を表 6.3 に示す．表 6.3 より，コーパスからのモデル化 (M=1,10,100) と，出現確率からのモデル化のそれぞれにおいて，パープレキシティ，単語正解率，単語認識精度に大きな差は見られない．よって，以後の実験では，出現確率からのモデル化の手法を用いる．

表 6.3 コーパスからのモデル化と出現確率からのモデル化の比較

No.	言語モデルの構築方法	倍率	PP	Cor.	Acc.
1	コーパスからのモデル化	1	55.3	67.0	59.1
2		10	55.3	67.2	59.2
3		100	55.3	67.2	59.3
4	出現確率からのモデル化	-	54.7	67.6	59.2

6.4.4 パープレキシティによる従来手法との比較

ここでは，句読点をショートポーズとして扱う従来手法と，ショートポーズ挿入モデルを用いて言語モデルを構築する提案手法とを比較する．ショートポーズ挿入モデルとしては，式 (6.6) のように CRF を用いる場合と，式 (6.18) のように形態素 3-gram モデルを

用いる場合を検討する．この形態素 3-gram モデルは，CSJ(表 6.1 の学習セット) から学習したモデルであり，200msec 以上，1000msec 未満のポーズを語彙に含む．まず，各手法をパープレキシティで比較した結果を図 6.4 に示す．図 6.4 より，読点の一部あるいは全部をショートポーズとして利用する言語モデル (○) や単語間にランダムにショートポーズを挿入した言語モデル (×) よりも，3-gram や CRF を用いて文脈を考慮してショートポーズ生起確率を割り当てる提案手法が，より小さいパープレキシティを示している．したがって，文脈を考慮してショートポーズ生起確率を割り当てる提案手法は，従来手法に比べて，パープレキシティの点で優れている．さらに，直前後の単語・モーラを考慮する CRF をショートポーズ挿入モデルとして用いた言語モデルのパープレキシティ (◇) は，直前の単語文脈だけを考慮する 3-gram を用いた言語モデルのパープレキシティ (△) より小さかった．よって，直後の文脈は，適切なショートポーズ生起確率の予測に効果的である．

また，<*sp*> を含まない言語モデルを用いて，<*sp*> を除外したテストコーパスに対するパープレキシティを求めたところ，提案手法よりも大きい値が得られた（$PP = 59.0$）．これは，緒方ら [79] が試みているような，入力音声とモデルの学習データから無音区間を除去して認識を行う手法に対応する．この結果より，<*sp*> を除去して音声認識を行う手法や，ポーズを透過単語として扱う手法よりも，提案手法のように <*sp*> を積極的にモデル化する方が効果的である．

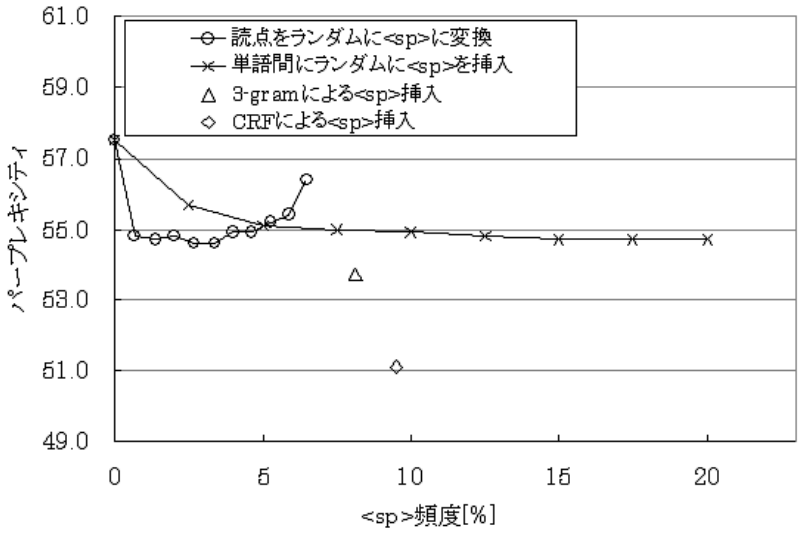


図 6.4　ショートポーズの挿入方法の比較

6.4.5 音声認識による従来手法との比較

次に，音声認識実験による各手法の評価結果を表 6.4 に示す．なお，ここでは，テストデータ全体に対する認識率に加え，ポーズの周辺のみにおける認識率も評価した．ここで，ポーズの周辺とは，ポーズの直前 2 単語およびポーズの直後 2 単語を含む．まず，<*sp*> なしのモデル（No.1）では，<*sp*> にまったく対応しておらず，また，実際の認識の際には入力音声におけるショートポーズがすべて別の単語として誤認識（湧き出し誤り）されてしまうことから，パープレキシティと単語認識精度では最も悪い結果となっている．これに対し，<*sp*> を何らかの形で考慮することで，パープレキシティや認識率を改善することができると考えられるが，<*sp*> が任意の位置に現れうると仮定した単純なモデル（No.2）では大きな改善は得られなかった．また，従来の読み上げ音声の認識と同様にコーパス中の読点をショートポーズとして扱うこと（No.3）を試みたが，こちらも認識性能の改善は小さかった．話し言葉においては，コーパスの作成者が付与した読点と，実際の音声におけるショートポーズとは必ずしも対応していないからである．これに対し，3-gram や CRF を利用し，文脈を考慮してショートポーズ生起確率を割り当てたモデル (No.5 および No.7) では，パープレキシティ，補正パープレキシティ，単語正解率，単語認識精度のすべてにおいてベースラインよりも高い性能が得られた．特に，直前の単語文脈だけを考慮する 3-gram をショートポーズ挿入モデルとして用いた言語モデル (No.5) よりも，直前後の単語・モーラを考慮する CRF を用いた言語モデル (No.7) の方が高い性能が得られた．これは，フィラーの挿入 [82] の場合と同様の結果である．

また，句点をショートポーズとして扱うことで，パープレキシティ，補正パープレキシティ，単語正解率，単語認識精度にそれぞれ改善が見られた（No.6 および No.8）．最終的に，ベースラインと比べ，提案手法はテストデータ全体に対する単語正解率で 2.5%，単語認識精度で 4.0% の相対的な改善を得た．このように，No.3 と No.7，および No.4 と No.8 の比較より，読点の情報よりもショートポーズの情報の方がコンテキスト情報として有効である．なお，符号検定 [3] により，提案手法 (No.8) は，ベースライン手法 (No.4) より危険率 1% で性能が有意に高いと言えた．

ショートポーズを考慮した言語モデルを作成するための従来手法として，句読点を用いる手法以外に，ポーズ情報を含まない (または不十分な) コーパスから作成した言語モデルと，ポーズ情報を含むコーパスから作成した言語モデルを N-gram カウント混合する手法がある [42]．表 6.5 に，N-gram カウント混合法を用いた認識実験の結果を示す．No.9 は，ポーズ情報を含まない国会会議録から作成した言語モデル (No.1) と，ポーズ情報を含む CSJ(表 6.1 の学習セット) から作成した言語モデルを，1 : 1 の重みで N-gram カウント混合した言語モデルであり，従来の一般的な N-gram カウント混合法に相当する．

表 6.4 認識実験によるショートポーズ挿入方法の評価

No.	手法	PP	$PP*$	<*sp*> 頻度 (%)	テストデータ全体 Cor. (%)	テストデータ全体 Acc. (%)	ポーズ周辺 Cor. (%)	ポーズ周辺 Acc. (%)
1	<*sp*> なし（参考）	57.5	63.5	0.0	68.5	60.6	69.0	59.6
2	ランダム ($p = 0.1$) に <*sp*> を挿入	54.9	60.7	10.0	68.5	62.6	68.6	61.4
3	全ての読点を <*sp*> として扱う	56.4	62.3	6.5	68.0	61.8	68.4	60.6
4	全ての読点を <*sp*> として扱う ＋全ての句点を <*sp*> として扱う	55.8	61.7	7.6	68.7	61.9	69.2	61.3
5	3-gram を用いて <*sp*> を挿入	53.7	59.4	8.2	69.1	63.1	69.6	62.3
6	3-gram を用いて <*sp*> を挿入 ＋全ての句点を <*sp*> として扱う	52.8	58.3	9.6	69.7	63.4	70.6	63.2
7	CRF を用いて <*sp*> を挿入	51.1	56.5	9.5	69.2	63.4	69.4	62.1
8	CRF を用いて <*sp*> を挿入 ＋全ての句点を <*sp*> として扱う	50.9	56.2	10.9	70.4	64.4	71.3	64.1

表 6.5 国会会議録と CSJ を N-gram カウント混合した言語モデルの音声認識実験

No.	手法	PP	$PP*$	<*sp*> 頻度 (%)	テストデータ全体 Cor. (%)	テストデータ全体 Acc. (%)	ポーズ周辺 Cor. (%)	ポーズ周辺 Acc. (%)
9	<*sp*> を含まない国会会議録 (No.1)+CSJ	49.2	54.3	1.4	68.9	61.7	69.4	60.9
10	提案手法 (No.8)+CSJ	45.0	49.7	10.0	70.6	64.4	71.7	64.2

No.9 は，提案手法 (No.8) に比べて，パープレキシティおよび補正パープレキシティは改善しているが，単語正解率および単語認識精度は低下している．これは，従来の N-gram カウント混合法 (No.9) では，国会会議録にのみ出現する語とショートポーズが同時に含まれる 3-gram や 2-gram が考慮されないため，テストコーパスにおける 3-gram ヒット率が 78.6% から 76.6% に低下したことが原因と考えられる．よって，提案手法 (No.8) は，従来の N-gram カウント混合法 (No.9) よりも効果的であると言える．次に，No.10 は，提案手法により作成された言語モデル (No.8) と，CSJ(表 6.1 の学習セット) から作成した言語モデルを，1 : 1 の重みで N-gram カウント混合した言語モデルである．No.10 は，提案手法 (No.8) に比べて，パープレキシティおよび補正パープレキシティは大きく改善しているが，単語正解率および単語認識精度はほとんど変わっていない．No.10 の 3-gram ヒット率は 81.0% で，提案手法 (No.8) の 3-gram ヒット率 78.6% よりも高い．しかし，No.10 に含まれる 3-gram によって新たにカバーされた 500 個所の音声認識結果を，提案手法 (No.8) の音声認識結果と比べると，誤認識されていた個所が正しく認識されるように変化した個所はわずか 33 個所だった．このように，音声認識の性能という観

点から見ると，提案手法 (No.8) は，提案手法と CSJ を N-gram カウント混合したモデル (No.10) に近い性能を持つと言える．

CRF によるショートポーズ挿入モデルと句点を組み合わせた言語モデル (No.8) と，ベースラインの言語モデル (No.2 および No.4) による音声認識性能を，話者別に比較した結果を表 6.6 に示す．話者 S5 の単語正解率と話者 S8 の単語認識精度という 2 つの例外を除いて，提案手法 (No.8) は，ベースライン手法 (No.2 および No.4) よりも高い性能を示している．したがって，提案手法は，殆んどの話者に対して有効である．

提案手法の言語モデル (No.8) とベースラインの言語モデル (No.2 および No.4) による音声認識性能を，会議別に比較した結果を表 6.7 に示す．各会議の議題と内容は互いに大きく異なる (表 6.2) にも関わらず，テストセットに含まれる全ての会議について，提案手法 (No.8) は，ベースライン手法 (No.2 および No.4) よりも高い性能を示している．したがって，提案手法は，話題によらず有効である．

なお，図 6.3 で説明した単位間の依存性を考慮した認識手法と，各単位を独立に扱う従来の認識手法とを比較したところ，パープレキシティにおいては前者の手法による改善が

表 6.6 話者別の比較

ID	発話数	発話時間 (秒)	形態素数	未知語率 (%)	ベースライン				提案手法	
					No.2		No.4		No.8	
					Cor. (%)	Acc. (%)	Cor. (%)	Acc. (%)	Cor. (%)	Acc. (%)
S1	676	1342.3	5197	0.29	72.0	65.4	71.0	64.6	**74.2**	**68.2**
S2	54	142.3	642	0.16	58.1	53.6	58.9	53.0	**60.6**	**55.1**
S3	70	139.6	477	3.98	70.7	58.1	70.7	58.1	**73.1**	**60.3**
S4	72	165.1	724	1.24	63.1	57.9	61.8	56.6	**67.0**	**62.2**
S5	226	645.4	2621	0.95	49.0	43.7	51.3	43.3	51.3	**44.9**
S6	234	949.9	3630	0.99	84.0	80.5	83.8	79.5	**85.3**	**82.0**
S7	18	103.2	464	0.43	69.6	66.4	70.5	65.5	**73.9**	**71.1**
S8	59	186.7	590	1.19	70.5	**58.8**	71.5	58.3	**71.7**	56.3
S9	104	373.5	1112	2.16	74.8	66.6	76.8	67.1	**77.2**	**69.7**
S10	316	808.8	3416	1.26	61.1	55.8	60.9	55.1	**62.4**	**56.6**
S11	49	175.6	747	1.47	79.1	76.0	79.0	75.4	**79.3**	**76.2**
S12	65	244.6	952	0.21	63.7	53.2	64.3	51.3	**64.7**	**54.2**
合計	1943	5276.9	20572	0.94	68.5	62.6	68.7	61.9	70.4	64.4

表 6.7　会議別の比較

ID	発話数	発話時間 (秒)	形態素数	未知語率 (%)	ベースライン				提案手法	
					No.2		No.4		No.8	
					Cor. (%)	Acc. (%)	Cor. (%)	Acc. (%)	Cor. (%)	Acc. (%)
T1	360	1041.6	4038	0.99	54.6	48.4	56.0	47.6	**56.9**	**49.3**
T2	333	760.5	3201	1.97	61.4	54.9	61.6	55.0	**63.3**	**56.5**
T3	845	1872.5	7380	0.24	69.6	62.8	68.8	61.7	**71.4**	**65.0**
T4	405	1602.2	5953	1.23	80.5	76.2	80.9	75.6	**82.1**	**78.1**
合計	1943	5276.9	20572	0.94	68.5	62.6	68.7	61.9	70.4	64.4

見られた [*5] が，認識率においては有意な改善が得られなかった．単位間の依存性を考慮した認識手法は，特に，認識処理単位の先頭単語が機能語である場合に有効と考えられるが，今回のテストデータではそのような認識処理単位は全体の 7.2% のみであったため，有意な性能差として現れなかったものと考えられる．

また，本節の実験では最大で 60% 程度の認識精度が達成されたが，国会答弁を対象とした音声認識では，たとえば秋田ら [85] のように，80〜90% 程度の認識精度を達成している例もある．これは，第一に，音響モデルの違いによるものと考えられる．秋田らは音響モデルを国会答弁のデータを用いた音素誤り最小化（MPE）学習によって学習し，さらに，話者区間ごとに話者適応化を行っている．これに対し，我々の音響モデルは CSJ から最尤（ML）学習したものであり，話者適応化は行っていない．第二に，秋田らは話し言葉特有の発音変動を考慮した統計的変換手法を発音辞書に適用している．これに対し，我々は CSJ にて観測された発音変動を発音辞書に加えたのみで，話し言葉特有の発音変動への対処としては限定的である．第三に，秋田らは音声認識用のデコーダとして 2 パス方式のデコーダ（Julius rev4.1）を使用しているのに対し，我々が用いたデコーダは 1 パス方式であり，リスコアリングは行っていない．第四に，認識の難しさは会議によって大きく異なる．我々がこれまでに行った実験では，認識精度の最も低い会議と最も高い会議とで 30% 程度の認識精度の違いを確認している．

[*5] たとえば，No.7 のモデルでは，単位間の依存性を考慮しない場合のパープレキシティは 60.5 であった．また，認識処理単位の先頭単語のみのパープレキシティは，単位間の依存性を考慮した場合には 449.0，単位間の依存性を考慮しない場合には 981.3 であった．

6.5 フィラー挿入とショートポーズ挿入の関係の分析

一般に話し言葉において，フィラーとポーズは互いに隣接して出現し易いことが知られている．たとえば，中川ら [65] の模擬対話音声を対象とした分析によれば，フィラーの直前・直後のいずれかにポーズ（10msec 以上の無音区間）が現れる割合は 81% と非常に高い．また，Gabrea ら [86] や Stouten ら [87] も，Switchboard コーパスを対象とした分析において，ほぼ同様の分析結果を得ている．本研究で使用した CSJ の学会・模擬講演においても，フィラーの直前・直後のいずれかにポーズ（200msec 以上の無音区間）が現れる割合は 56.0% であり，また，ポーズ（200msec 以上の無音区間）の直前・直後のいずれかにフィラーが現れる割合は 35.2% であった．このように，フィラーとポーズは隣接して出現する割合が高いことから，ポーズの挿入においてもフィラーの情報は重要なコンテキストであると考えられる．国会会議録にはフィラー情報が含まれていないため，本節の実験では，自動挿入したフィラーをコンテキストとして参照して，ポーズの予測を行っている．

しかし，フィラー挿入は，非常にランダム性が高いプロセスであり，自動挿入したフィラーは，現実のフィラーとは大きな異なりがある．CSJ の模擬講演 (表 6.1) から学習したフィラー挿入モデルを用いて，CSJ の学会講演 (表 6.1) に対してフィラーを自動挿入する実験を行ったところ，フィラーが自動挿入された位置に現実のフィラーが存在する割合は 27.5% であり，自動挿入されたフィラーの種類が現実のフィラーと一致する割合は 10.2% だった．そのため，自動挿入されたフィラーを参照してポーズの予測を行うと，不適切な位置にポーズ生起確率を割り当ててしまうことが懸念される．そこで，本節では，この点について，日本語話し言葉コーパスを用いた実験によって調査・分析する．また，フィラーとショートポーズの同時挿入についても検討する．

6.5.1 フィラーとショートポーズの挿入手法の比較

CSJ の模擬講演を学習コーパスとして，(1) フィラーを参照しないポーズ挿入モデル，(2) 自動挿入 [82] したフィラーを参照するポーズ挿入モデル，(3) 実際のフィラーを参照するポーズ挿入モデルという 3 通りのポーズ挿入モデルを作成し，それぞれを CSJ の学会講演に適用して，3 通りの言語モデルを作成した．加えて，(4)CSJ の学会講演に含まれる実際のポーズを用いた言語モデルを作成した．作成した 4 通りの言語モデルを，CSJ の音声認識テストセット **test-set 2**[88] をテストコーパスとして比較したところ，それぞれの補正パープレキシティ $PP*$ は表 6.8 の通りであった． 表 6.8 より，(1) フィラーを参照しないポーズ挿入モデルと，(2) 自動挿入 [82] したフィラーを参照するポーズ挿入

モデルに性能差はない．よって，フィラーの自動挿入は，ポーズ予測に悪影響を与えていないと考えられる．しかし，効果はないと言える．

表 6.8　フィラーとショートポーズの挿入手法の比較

言語モデルの構築方法	PP*
フィラーを参照しないポーズ挿入モデル	86.9
自動挿入したフィラーを参照するポーズ挿入モデル	86.9
実際のフィラーを参照するポーズ挿入モデル	85.5
CSJ の学会講演に含まれる実際のポーズを利用	83.7

6.5.2　フィラーとショートポーズの同時挿入手法の検討

本章ではこれまでフィラーの挿入とショートポーズの挿入を別々に行ってきた．しかし，このような手法では，たとえばフィラー，ショートポーズの順に挿入を行う場合に，ショートポーズの情報をフィラーの挿入のための素性情報として利用することはできない．また，フィラーやショートポーズの出現は確率的な現象であることから，決定的に挿入されたショートポーズの情報を利用するよりも，ショートポーズを確率的に考慮する方がより適切であると考えられる．

そこで，フィラーとショートポーズを同時に挿入するモデルを，CRF を用いて構築した．具体的には，個々の形態素に対し，下記の 5 種のラベルのいずれかを付与するラベリング問題に CRF を適用した．なお，フィラーの種類は CRF によるラベリングを行った後に，形態素 3-gram に基づいて決定した．

- F … 形態素の直後にフィラーのみが挿入される．
- P … 形態素の直後にショートポーズのみが挿入される．
- FP … 形態素の直後にフィラー，ショートポーズがこの順で挿入される．
- PF … 形態素の直後にショートポーズ，フィラーがこの順で挿入される．
- O … ショートポーズもフィラーも挿入されない．

このようにフィラーとショートポーズを単一の CRF でモデル化することで，両現象を同時に，かつ確率的に考慮することができる．

しかし，本手法を前節と同様に CSJ の学会・模擬講演を用いた実験により評価したところ，従来のフィラーとショートポーズを別々に挿入した場合と比べ，性能の改善は得られなかった．これは，従来と比べラベリングの問題が二値ではなく多値であることからモデルの学習が難しくなったことが原因と考えられる．

6.6 まとめ

本章では，言語的なまとまり以外の要因に基づくポーズを積極的に考慮した言語モデルを構築することによって，ポーズ周辺の単語の音声認識を改善する方法を提案した．言語モデルを作成するためのコーパス（国会会議録や新聞など）には，句点という形で，言語的まとまりに起因するポーズの位置情報は既に含まれている．まず，書き起こしコーパスの句読点とポーズの一致率を分析し，大きな不一致があることを明らかにした．そこで，句読点の同定は難しく，ポーズの検出は容易な点を考慮して，言語的まとまり以外の要因に基づくポーズの出現位置を，話し言葉音声コーパスに基づいて学習したモデルによって補うことにより，ポーズ出現位置の情報を考慮した言語モデルを作成した．国会審議の音声認識実験において，提案手法に基づくポーズを考慮した言語モデルを用いて認識を行ったところ，従来の句読点をポーズに対応させた言語モデルと比較して，パープレキシティおよび音声認識精度を改善することができた．

さらに，以前報告したフィラー挿入モデルと併用することにより，フィラーやポーズが正確に書き起こされていないコーパスに対し，フィラーとポーズを挿入することにより，話し言葉の言語モデルを構築できることを示した．

今後の検討課題として，提案手法によるパープレキシティの改善について，特にポーズの周辺におけるパープレキシティで評価することが挙げられる．

第 7 章

整形された会議録を用いた話し言葉音声認識のための音響モデリング

7.1 はじめに

各種の音声言語処理アプリケーションについて，大量の話し言葉の正確な書き起こしを含む大規模コーパスが利用できることは非常に重要である．たとえば，話し言葉を対象とする音声認識システムには，対象音声とドメインが一致し，かつ，話し言葉特有の言い回しに対応した言語モデルが不可欠である．そのような言語モデルを構築する最も単純な方法は，対象音声と同一ドメインの大規模な話し言葉コーパスから言語モデルを学習するという方法である．しかし，話し言葉を正確に書き起こす作業は極めて高いコストを必要とするため，あらゆるドメインに対して，そのようなコーパスが入手できると仮定することは現実的ではない．

それに対し，速記録や会議録は，正確な書き起こしより広く作成されており，比較的容易に入手が可能である．ただし，速記録や会議録では，可読性を高めるために，フィラーや言い淀み，言い直しなどの話し言葉特有の現象は削除され，話し言葉特有の言い回しは適切な書き言葉に置き換えられていることが一般的である．本章では，このような書き起こしを，**整形された書き起こし**と呼ぶ．たとえば，国立国会図書館は，1947 年以降の全ての国会の会議録を公開している [*1]．例を図 7.1 に示す．図 7.1 に見られるように，国会会議録では，可読性を高めるために，フィラー（例．”えー”, ”いー”）や言い直し（例．”け”），および冗長な言い回し（例．”ですね”, ”と”）はすべて削除されており，また，口語体の言い回し（例．”てる”）は文語体（例．”ている”）に置き換えられている．さらに，助詞の脱落が補われている（例．”を”）ほか，一部の読点は速記者の判断によって追加あ

[*1] http://kokkai.ndl.go.jp/

るいは削除されている．また，Williams らは，多数の非熟練作業者を利用して書き起こしを作成する方法を提案している [43]．この方法は，熟練作業者による書き起こしに比べて，作業コストの面では非常に安価ではあるものの，多数の書き起こしミスが含まれる欠点がある．このように，整形された書き起こしは，今後大量に利用できるようになる可能性が高い．

言語モデルや音響モデルの学習のために忠実な書き起こしが利用できない条件において，Lamel ら [89] による lightly supervised training のほか，様々な学習法が検討されている．

三村ら [90] は，会議音声の大規模なアーカイブとそれらに対する整形された書き起こし (会議録) が利用できる場合に，会議録のテキストデータに統計的話し言葉変換を適用して，会議のターンごとに制約の強い言語モデルを作成し，この言語モデルを音声認識に用いることにより，音響モデル学習のためのラベルを生成している．また，Lecouteux ら [91] は，三村らと同様の条件において，整形された書き起こしに基づいた A* 探索によるデコーディングを行うことにより，音響モデル学習のためのラベルを生成している．

このほか，Web におけるクラウドソーシングを用いた学習データの効率的な収集法として，緒方ら [92] は，音声認識結果を不特定多数のオンラインユーザに訂正してもらう枠組みを提供し，訂正された書き起こしに基づいて音声認識の音響モデルの改善を行なっている．また，Lane ら [93] は，Amazon Mechanical Turk を用いて，有償だが安価なコストで，不特定多数のユーザによる効率的な音声・言語データの収集を行なっている．ただし，Amazon Mechanical Turk では，高いコストをかけてデータを収集する場合とは異なり，習熟度の低いユーザによって誤りを含んだデータが収集される場合があり，従ってデータ収集の際の品質管理が重要であることが報告されている．

これらの先行研究に対し，本研究では，整形された書き起こしを正確な書き起こしに半

ところ が ですね， えー， この 資料， 見 て み ます と 神奈川 県 の 場合 は， け， 結果 と し て 財政 的 に いー 豊か に なっ てる と .

(i) 正確な書き起こし

ところ が， この 資料 を 見 て み ます と， 神奈川 県 の 場合 は， 結果 と し て 財政 的 に 豊か に なっ て いる .

(ii) 整形された書き起こし

図 7.1 正確な書き起こしと整形された書き起こしの例

自動的に変換する枠組みの構築について検討する．この枠組みでは，まず整形された書き起こし中の整形箇所を自動検出し，検出された整形箇所のみを人手で書き起こすことにより，できるだけ小さいコストで，整形された書き起こしと正確な書き起こしのパラレルコーパスを作成する．先に述べた通り，整形された書き起こしは，話し言葉的現象を削除・整形して書き言葉に近づけたテキストであるから，このパラレルコーパスは，話し言葉特有の言い回しから書き言葉特有の言い回しへの変換パターンの構築 [94] に利用可能である．このようにして構築された言い回し変換パターンと，整形された書き起こしを対象としてフィラーおよびポーズを復元する手法 [75][82] を組み合わせることにより，任意の書き言葉コーパスから，そのコーパスのドメインについての話し言葉に対応した言語モデルを自動構築することが，本研究の目標である．また，本手法により，整形されていない部分，すなわち正確に書き起こされた部分も同時に検出されることから，これらを音響モデルの教師なし学習 [89] に用いることも可能である．

本章では，この枠組みのために，整形された書き起こしから整形箇所を自動検出する方法を提案する．本手法は，2 つのステップからなる．第 1 に，整形された書き起こしとその原音声とでアラインメントを行い，第 2 に，アラインメントによって得られた素性に基づく Support Vector Machine (SVM) を用いて，整形箇所を検出する．第 2 段階において，音節を単位とする音響的素性を用いる．

テキストと音声とのアラインメントの応用例として，たとえば Lamel ら [89] は，音響モデルの教師あり学習において，信頼できない学習データを除去するためにアラインメントを利用している．また，Roy ら [95] は，アラインメントから得られた音響スコアに基づいて，対象音声の書き起こし作業の難しさを推定している．さらに，丸山ら [96] は，ドキュメンタリー番組を対象として，字幕の送出タイミングを検出するためにアラインメントを利用している．これに対し，提案法では，整形された書き起こしと原音声とでアラインメントを行うため，両者の不一致部分（＝整形箇所）では，アラインメントにおける音響スコアが低下し，また，各音節の対応区間にばらつきが生じる．この音響スコアの低下や音節区間のばらつきを手がかりとして，整形箇所を自動検出することが提案法のねらいである．

7.2 整形された書き起こしと原音声のアラインメント

提案法の第 1 段階として，整形された書き起こしとその原音声との強制アラインメントを行う．最初に，整形された書き起こしの単語列 $w_1 w_2 \cdots w_n$ から図 7.2 のような 2-gram 言語モデルを作成する．図 7.2 において，w_i は整形された書き起こしに含まれる i 番目の単語であり，$Filler_i$ および sp_i は w_i の直後に出現するフィラーとショートポーズである．明らかに，図 7.2 の言語モデルは，整形された書き起こし，または単語間にフィラー

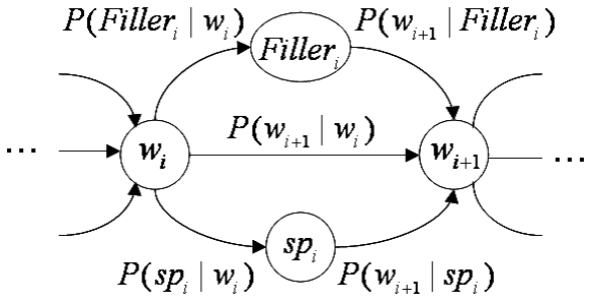


図 7.2　2-gram 言語モデルに基づく制約

· ポーズが挿入された単語列のみを受理する．よって，図 7.2 の言語モデルを用いて原音声の連続音声認識を行うと，単語間にフィラーとショートポーズの挿入を許すという制約のもとで，整形された書き起こしに含まれる各単語を原音声にアラインメントできる

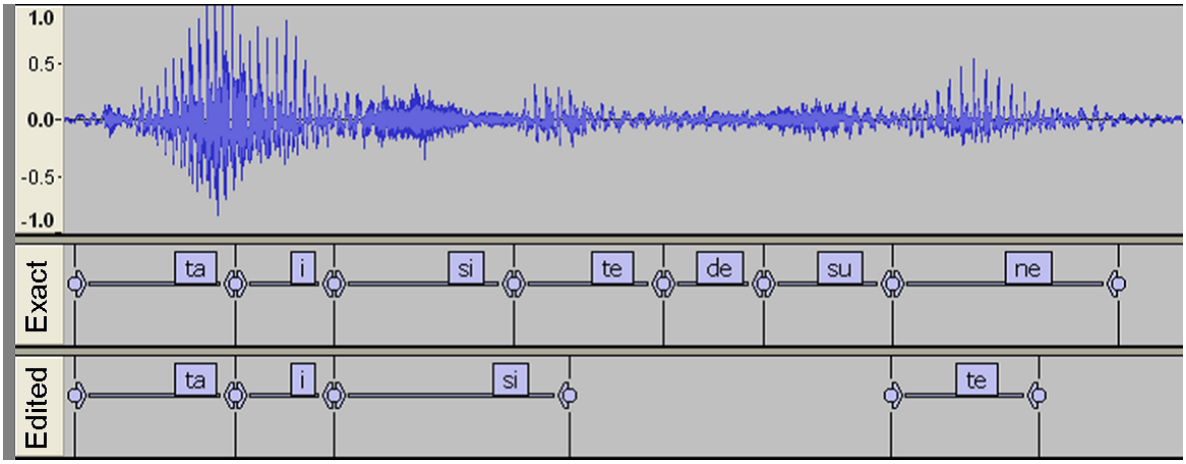


図 7.3　正確な書き起こし/整形された書き起こしと原音声とのアラインメント例

図 7.3 に，正確な書き起こしと原音声のアラインメント例，および整形された書き起こしと原音声のアラインメント例を示す．図 7.3 に見られるように，整形された書き起こしと原音声とのアラインメントを行うと，両者の不一致部分（すなわち，整形箇所）において，実際の発声とは異なるモデルが強制的に対応付けられることになる．この結果，周囲の音節の対応区間が歪められ，また，モデルの不一致によって音響スコアが低下する．図 7.3 の例では，音節/*si*/の区間が不適切に延びており，また，モデル/*te*/が音節/*ne*/に，ショートポーズモデル/*sp*/が音節/*te de su*/にそれぞれ強制的に対応付けられている．よって，音節の区間が極端に長い/短い部分や，音響スコアが通常よりも低い部分は，整形箇所である可能性が高いといえる．

なお，多くの整形された書き起こしにおいて，書き起こし中の各文と原音声中の各発話の対応に関する情報は付与されていないことが多い．これを考慮し，本章では，連続音節認識結果の音節数に基づいて，各発話の単語数を推定することにより，両者の対応付けを

自動で行った．この対応付けに基づいて，書き起こし区間に対し，原音声の発話区間を切り出した上で，書き起こし中の各文とアラインメントを行う．予備実験により，このように各発話の単語数を自動で推定した場合でも，各発話の単語数が正確に得られた場合と同等の精度でアラインメントが行えることを確認している．

7.3 非整形箇所と整形箇所の自動検出

本章では，整形・非整形部分の検出を 2 通りの方法で行う．第 1 の方法は，提案法で，強制アラインメントと Support Vector Machine (SVM) を用いて整形箇所を検出する．第 2 の方法は，Paulik らによって提案されている方法 [97] で，大語彙連続音声認識 (LVCSR) の認識結果を用いる方法である．以下では，それぞれの方法について詳細を述べる．

7.3.1 Support Vector Machine に基づく整形・非整形部分の検出

提案法の第 2 段階として，7.2 節の強制アラインメントによって得られた素性に基づく SVM を用いて，整形された書き起こしから整形箇所を検出する．本研究では，整形箇所の検出を，整形された書き起こしに含まれる各単語に対する二値分類問題として定式化する．つまり，整形された書き起こしに含まれる各単語を，整形された単語か，整形されていない単語かの 2 種類に分類する． SVM のツールキットとしては，TinySVM（ver 0.09）[98] を用いる．カーネル関数には多項式カーネルを採用する．

本章では，各々の単語について，以下の特徴量を用意する．

- 音響的素性
  - 単語単位の音響尤度
  - 音節単位の音響的素性
- 言語的素性
  - 単語
  - 品詞
  - 単語に含まれる音節数

判定対象となる単語，その直前 2 単語，その直後 2 単語に関する上記の特徴量を組み合わせて，判定用の素性として用いる．素性の選択にあたっては，Huang らの文献 [99] を参考にした．以下では，音響的素性について詳細を述べる．

■単語単位の音響尤度 強制アラインメントによって得られた音響尤度を素性として利用する．ただし，この音響尤度（対数尤度）は単語の時間長によって正規化し，さらに，連

続音節認識によって得られた音響スコア（対数尤度）との差分を取る．これは，対数事後確率を素性として用いることに相当する．

■音節単位の音響的素性 筆者らの方法 [100, 101] では，整形・非整形を判定しようとしている単語を単位として，その単語に含まれる全ての音節の音節長の平均や分散を素性として用いていた．しかし，単語に含まれる全ての音節に基づく素性は，音節単位での異常さを検出するには不十分な場合がある．例として，国会会議録では「ところが」と書き起こされている部分について，原音声を正確に書き起こすと「ところがですね」となっていて，以下のようにアラインメントされた場合を考える．

国会会議録 (整形された書き起こし)	to	ko	ro	ga
正確な書き起こし	to	ko	ro	ga-de-su-ne

この場合，最初の 3 音節/*to*/ /*ko*/ /*ro*/の音響尤度や音節長には一切の異常はなく，最後の音節/*ga*/の音響尤度と音節長のみに異常が現れる．そのため，単語に含まれる全ての音節の音節長の平均や分散という素性では，異常さが 1/4 に薄められてしまい，整形部分を取り出す素性として使いにくくなることが予想される．この問題を回避するには，単語に含まれる全ての音節に基づく素性の代わりに，単語に含まれる音節から最も異常と推測される音節を 1 つ代表として取り出し，その代表音節に基づく素性を用いる方法が有効と考えられる．本章では，代表音節を取り出す尺度として，**正規化音節長**と**正規化音響尤度**という 2 つの尺度を用いる．

正規化音節長の定義は，以下の通りである．最初に，国会会議録の原音声を正確に書き起こした音節列 s_1^N と原音声をアラインメントし，各音節の音節長 $d(s_i)$ を求める．次に，音節の種類 x 毎に，音節長の平均 $E_d(x)$ と分散 $V_d(x)$ を求める．

$$E_d(x) = \frac{\sum_{i=1}^{N} \delta(s_i = x) d(s_i)}{\sum_{i=1}^{N} \delta(s_i = x)} \tag{7.1}$$

$$V_d(x) = \frac{\sum_{i=1}^{N} \delta(s_i = x) \left(d(s_i) - E_d(x)\right)^2}{\sum_{i=1}^{N} \delta(s_i = x)} \tag{7.2}$$

実際に出現した音節 s_j の音節長について，平均 $E_d(s_j)$ と分散 $V_d(s_j)$ を用いて，次式のように平均 0，分散 1 に正規化する．

$$\tilde{d}(s_j) = \frac{d(s_j) - E_d(s_j)}{\sqrt{V_d(s_j)}} \tag{7.3}$$

このように正規化した音節長 $\tilde{d}(s_j)$ を，**正規化音節長**と呼ぶ．正規化音節長は，音節の種類毎に自然な音節長があり，かつ，音節長は正規分布に従うという仮定の下で，実際に出

現した音節が，どの程度に異常な音節長となっているかの尺度として使うことができる．

正規化音響尤度の定義は，以下の通りである．最初に，国会会議録の原音声を正確に書き起こした音節列 s_1^N と原音声をアラインメントし，各音節の音響尤度 $L(s_i)$ を求める．次に，音節の種類 x 毎に，音響尤度の平均 $E_L(x)$ と分散 $V_L(x)$ を求める．

$$E_L(x) = \frac{\sum_{i=1}^{N} \delta(s_i = x) L(s_i)}{\sum_{i=1}^{N} \delta(s_i = x)} \tag{7.4}$$

$$V_L(x) = \frac{\sum_{i=1}^{N} \delta(s_i = x) \left(L(s_i) - E_L(x)\right)^2}{\sum_{i=1}^{N} \delta(s_i = x)} \tag{7.5}$$

次に，実際に出現した音節 s_j の音響尤度について，平均 $E_L(s_j)$ と分散 $V_L(s_j)$ を用いて，次式のように平均 0，分散 1 に正規化する．

$$\tilde{L}(s_j) = \frac{L(s_j) - E_L(s_j)}{\sqrt{V_L(s_j)}} \tag{7.6}$$

このように正規化した音響尤度 $\tilde{L}(s_j)$ を，**正規化音響尤度**と呼ぶ．正規化音響尤度は，正規化音節長と同様に，音節の種類毎に自然な音響尤度があり，かつ，音響尤度は正規分布に従うという仮定の下で，実際に出現した音節が，どの程度に異常な音響尤度となっているかの尺度として使うことができる．

国会会議録においては，言い直し・言い淀みの削除が最も多い整形処理であることを考慮すると，正規化音節長が最大であるような音節，または，正規化音響尤度が最小であるような音節が，その単語内において最も異常な音節である可能性が高いと考えられる．そのため，以下の素性を用いる．

1. 当該単語中の正規化音節長の最大値
2. 当該単語中の正規化音響尤度の最小値
3. 当該単語中で正規化音節長が最大な音節の正規化音響尤度
4. 当該単語中で正規化音響尤度が最小な音節の正規化音節長

7.3.2　大語彙連続音声認識結果による整形・非整形部分の検出

Paulik らは，整形された書き起こしと LVCSR の認識結果を，音声認識精度を求める場合と同じ方法を用いて単語単位でアラインメントを行い，人手で与えた閾値以上に連続して一致した部分を発音ラベルとして用いて話者適応する方法を提案している [97]．本章では，Paulik らの方法を利用して，以下の基準により整形・非整形部分の検出を行う．

(a) 閾値以上に連続して一致した部分を，非整形部分とする．
(b) 非整形部分以外の全てを，整形部分とする．

7.3.3　整形・非整形部分の検出性能の目標値

本章では，整形箇所の検出性能として，精度 33%，再現率 50% を目標値とする．ここで例として，100 単語からなる会議録があり，この内の 10 単語が整形箇所である場合を考える．この場合，目標値通りの性能が達成できたとすると，15 単語が整形箇所として検出され，この内の 5 単語が真の整形箇所である．従って，書き起こし作業者は，200 単語から整形箇所を 30 ヶ所自動抽出して，計 30 単語のみを確認することで，10 単語の整形箇所を見つけることができる．これは，100 単語すべてを確認するのと比べ，約 3 倍以上の作業効率である．

また，本手法を逆に適用して，非整形箇所，すなわち正確な書き起こし部分を検出することも可能である．上述の場合，100 単語の会議録の内，90% が非整形箇所である．これに対し，本手法を適用すると，85 単語が非整形箇所として検出され，この内の 80 単語，すなわち 94% が真の非整形箇所である．従って，より正確な書き起こしを抽出できたことになる．

7.4　自動検出された発音ラベルを用いた音響モデル学習

自動検出された発音ラベルを教師ラベルとして，最大事後確率推定法 (MAP 推定) を用いた逐次連結学習 [52] によって音響モデルの話者適応を行う．最大事後確率推定法 (MAP 推定) を用いた逐次連結学習では，適応化文に対する音節ラベル系列だけを与えることで，文発話データから連続出力分布型 HMM の最適なパラメータを求めることができる．

連結学習では，文発話データに対応する音節ラベル系列に従って音節 HMM を連結することで文 HMM を作成し，文 HMM を文発話データによって学習する．逐次連結学習では，MAP 推定によるパラメータ推定を 1 文ごとに行う手法であり，すべての適応文について，以下の手順を繰り返す．

1. 文発話データの音節ラベル系列に従って音節 HMM をヌルアークで連結し，文 HMM を作成する．このとき，同じカテゴリの音節 HMM はそれぞれ対応する状態で結びにする．
2. 文発話と文 HMM の間で Viterbi アルゴリズムによる照合を行った後，セグメンテーションを行い，各音節 HMM の各状態に対応するフレームサンプルを見つける．それらのサンプルを用いて，MAP 推定に基づくパラメータ推定を行う．
3. 文 HMM を再びヌルアークの位置で音節 HMM ごとに切り離す．

なお，文発話データでは，無音区間を削除し，無音・促音カテゴリは音節ラベル系列から外して，推定の対象外とする．

7.5 国会会議録を対象とする評価実験

実験には，国会会議録の一部を用いた．実験データの諸元を表 7.1 に示す．なお，テストコーパスの話者 11 名の内 1 名は女性のため，実験結果に悪影響を与えている可能性がある．

提案法の強制アラインメントおよび連続音節認識を行うデコーダには，SPOJUS++[102] を用いた．音響モデルは，CSJ[8] から学習した音節モデル (left-to-right 型 HMM，5 状態 4 出力分布，単混合全共分散行列，男性話者モデル) を用いた．モデル数は 116 である．音響分析条件は表 4.8 と同じである．予備実験より，図 7.2 の各状態の遷移確率として，$P(Filler_i|w_i) = 0.05$, $P(w_{i+1}|w_i) = 0.475$, $P(sp_i|w_i) = 0.475$ という値を用いた．

LVCSR の認識結果を用いる方法 (7.3.2 節) のデコーダとしても，SPOJUS++ を用いた．音響モデルは，CSJ から学習した左コンテキスト依存音節モデル（left-to-right 型 HMM，5 状態 4 出力分布，64 混合対角分散行列，男性話者モデル）を用いた．この左コンテキスト依存音節モデルは，116 種類の音節に対して 8 種類の先行音素 (/a/, /i/, /u/, /e/, /o/, /N/, /文頭/, /促音/) をコンテキストとして考慮しているため，モデル数は 928 である．音声分析条件は表 4.8 と同じである．言語モデルは，国会会議録に収録された 38,668K 語の整形された書き起こし (会議数としては 1,083 件に相当) から構築した形態素 3-gram モデルを用いた．国会会議録はフィラーおよびポーズが削除されているため，CSJ から学習したフィラー予測モデルおよびポーズ予測モデルを用いて，フィラーおよびポーズを考慮した言語モデルを作成した [75][82]．スムージングには，一般的な Witten-Bell バックオフを用いた．

7.5.1 実験条件

7.5.2 整形部分の自動検出

図 7.4 に，提案法による整形部分の自動検出結果を示す．図 7.4 の実線は，音響的素性と言語的素性の両方を使って整形部分を自動検出した場合の結果，図 7.4 の破線は，言語的素性のみを使って整形部分を自動検出した場合の結果である．この 2 つの結果の違いから，音響的素性は，整形部分の自動検出に対して効果的であることが分かる．また，図 7.4 より，正確に整形部分を検出することは困難だが，7.3.3 節で述べた性能の目標値は達成できているから，パラレルコーパス作成の作業効率は 7.3.3 節の想定通りに改善で

表 7.1　実験データ諸元

	学習	テスト
時間長 (min)	42	60
話者数	7	11
単語数	7.2k	10.8k
整形単語数	604	426
整形単語率 (%)	8.4	3.9

きると考えられる．

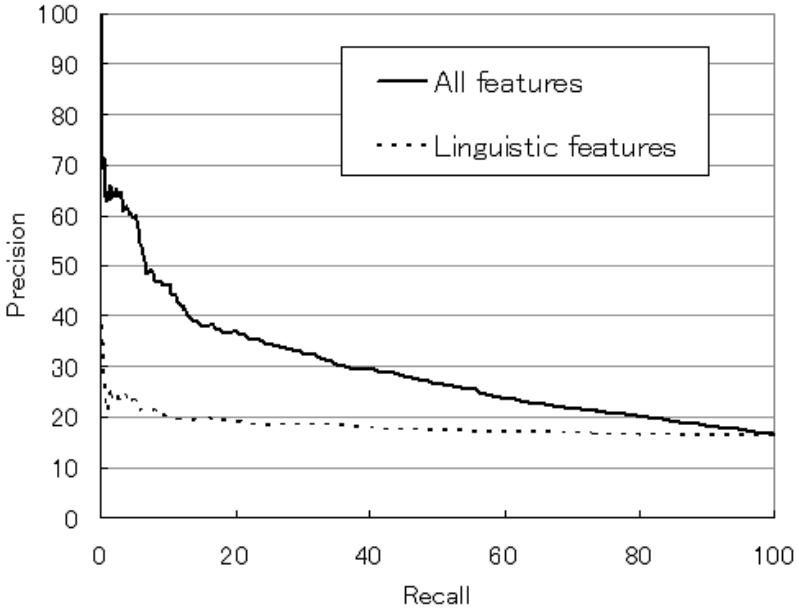


図 7.4　提案法による整形部分検出

図 7.5 に，提案法，LVCSR の認識結果を用いる方法，および両者を併用した結果を示す．両者の併用とは，以下のような方法である．

(a) 整形された書き起こしと LVCSR の認識結果が，閾値以上に連続して一致した部分を，非整形部分とする．
(b) それ以外の部分を対象として，提案法による判定結果を採用する．

国会会議録の場合は，LVCSR の認識精度が比較的低いため [*2]，LVCSR の認識結果を用いる方法による性能は，提案法とほとんど変わらないことが分かる．そのため，両者を併用する方法についても，ほとんど変わらない性能となっている．

[*2] 表 7.3 の「話者適応なし」行を参照．

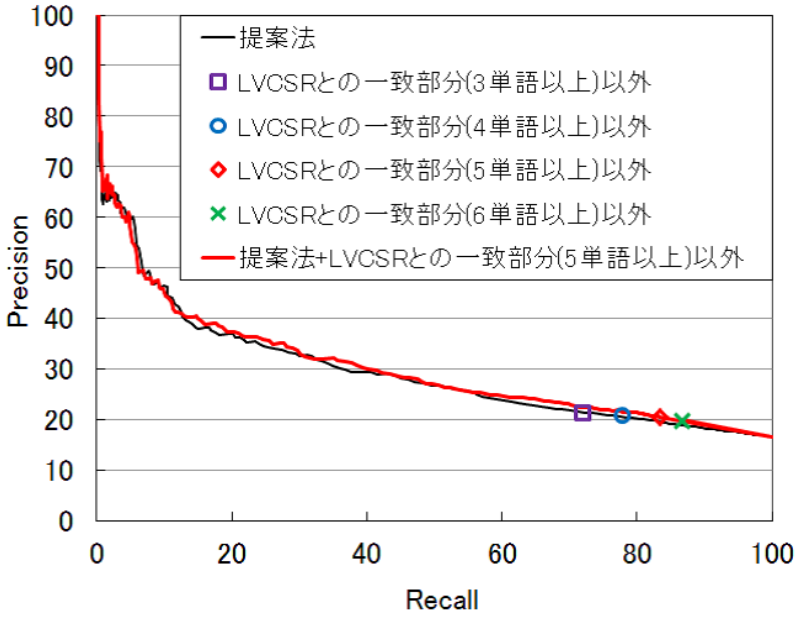


図 7.5　大語彙認識結果と組み合わせた整形部分検出

7.5.3　非整形部分の自動検出

整形された書き起こしから原音声と一致する部分 (すなわち，非整形部分) を取り出す実験を行った結果を図 7.6 に示す．図 7.6 より，整形された書き起こしに含まれている非整形部分は，音節単位で 80.1% である．よって，提案法を用いず，整形された書き起こし全体を非整形部分と見なすと，非整形部分の自動検出精度は 80.1% になる．提案法により，整形された書き起こしの 60% を選択すると，非整形部分の自動検出精度は 86.5% まで改善された．なお，この結果は，単語単位では 83.7% から 88.9% への改善に相当する．また，図 7.6 より，LVCSR の認識結果を用いる方法は，特に一致部分の長さの閾値が小さい場合には，提案法とほとんど性能が変わらないことが分かる．そのため，提案法と LVCSR の認識結果を用いる方法を併用した場合については，再現率が 40% ～ 60% の区間において，やや改善効果が見られる．

7.5.4　自動検出された発音ラベルを用いた話者適応

本節では，自動検出された非整形部分を発音ラベルとして用いて話者適応を行った実験結果について述べる．

HTK HMM toolkit ver 3.4.1[103] を利用し，MAP[52] による話者適応を行う．音声認識用デコーダ，音響モデルおよび言語モデルは，LVCSR の認識結果に基づいて整形・

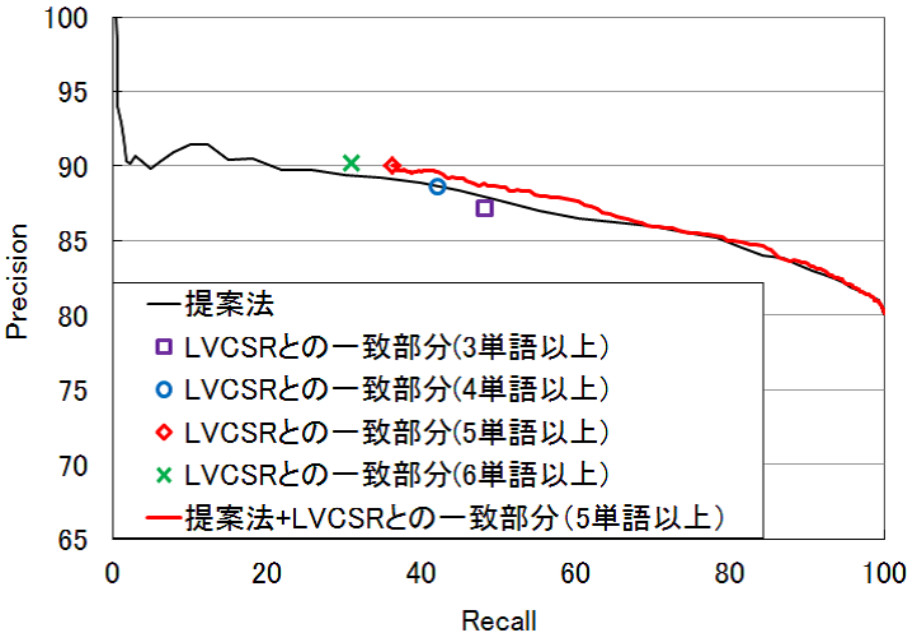


図 7.6　非整形部分の検出 (音節単位)

表 7.2　話者適応実験（発音ラベル諸元）

話者適応用発音ラベルの作成方法	ラベル数	精度 (%)	カバー率 (%)
話者適応なし	0	—	0
国会会議録全体を発音ラベルとして使用	5,155 (91.4%)	80.1	84.3
国会会議録と LVCSR の一致部分 (5 単語以上) を使用	1,728 (30.6%)	90.0	69.3
提案法によって取り出した発音ラベル ($recall = 90\%$) を使用	4,056 (71.9%)	83.0	83.0
提案法によって取り出した発音ラベル ($recall = 90\%$) と，国会会議録と LVCSR の一致部分 (5 単語以上) の併用	4,212 (74.7%)	83.5	83.2
正確な書き起こしの音節ラベルを使用	5,642 (100.0%)	100.0	92.6

表 7.3　話者適応実験（音声認識性能）

話者適応用発音ラベルの作成方法	Cor. (%)	Acc. (%)
話者適応なし	71.5	67.2
国会会議録全体を発音ラベルとして使用	74.3	70.6
国会会議録と LVCSR の一致部分 (5 単語以上) を使用	73.9	70.1
提案法によって取り出した発音ラベル ($recall = 90\%$) を使用	75.1	70.7
提案法によって取り出した発音ラベル ($recall = 90\%$) と，国会会議録と LVCSR の一致部分 (5 単語以上) の併用	75.0	71.0
正確な書き起こしの音節ラベルを使用	76.2	71.6

非整形部分を検出する時に用いたもの (詳細は，7.5.1 節を参照) と同じである．

表 7.3 に，話者適応の実験結果を示す．なお，音響モデルの話者適応用学習コーパスは，男性話者 3 名の 5,642 音節からなり，テストコーパスは同じ話者 3 名の 4,411 音節からなる．表 7.3 の「精度」列は，学習用データに含まれる正しい発音ラベルの比率を，「カ

バー率」列は，テスト用データに含まれる音節が学習用データに含まれる比率 (token 単位) を示す．「国会会議録全体を発音ラベルとして使用」行は，提案法を用いずに，国会会議録 (整形された書き起こし) 全体を発音ラベルとして用いて話者適応を行った結果である．表 7.3 の「提案法によって取り出された発音ラベルを使用」行は，提案法を用いて $recall = 90\%$ の条件で取り出した非整形部分を発音ラベルとして用いて話者適応を行った結果である．「正確な書き起こしの音節ラベルを使用」行は，原音声を正確に書き起こした発音ラベルを用いて話者適応を行った結果であり，この学習コーパスに基づいて話者適応を行った場合の性能の上限を示す．表 7.3 より，提案法は，国会会議録全体を利用する方法よりも話者適応に用いる発音ラベル数が少ないにも関わらず，ベースライン手法 (国会会議録全体を利用する方法) よりも話者適応後の音声認識性能が良いことが分かる．これは，提案法によって取り出された発音ラベルが，話者適応に対して有効であることを示している．国会会議録と LVCSR による認識結果の一致部分 (5 単語以上) の発音ラベルを使用する方法は，話者適応用発音ラベルの精度は高いものの，カバー率が低いため，話者適応の結果が悪くなっている．また，図 7.6 より，提案法と，提案法と国会会議録と LVCSR による認識結果の一致部分 (5 単語以上) を併用する方法の 2 つは，$recall = 90\%$ の条件では，非整形部分検出の性能がほとんど同じである．そのため，話者適応の結果も若干の改善にとどまった．

7.5.5 話者適応用発音ラベル数の比較

自動検出 (または，自動検出と LVCSR の認識結果の一致部分を用いる方法の併用) を用いて非整形部分を取り出す場合，非整形部分の精度を高くすると，再現率が低くなる (図 7.6)．そのような非整形部分を発音ラベルとして用いて話者適応する場合，発音ラベルの質は良いが，カバー率が低いため，話者適応の効果は小さくなる．本節では，より大規模なコーパスを用意することによって，カバー率の低下を補えるか検討する．

2007 年に行われた 2 件の委員会 [*3] から，新たな話者適応実験用データを収集した．データ量を表 7.4 に示す．話者 A〜F の 6 名について，実験条件を揃えるために，1 人あたり約 2,000 音節を話者適応用データとして取り出し (表 7.4 の「2k」列)，1 人あたり約 2,000 音節をテスト用データとして，話者適応実験を行った結果を表 7.6 に示す．音声認識性能は話者 6 名の平均値である．表 7.6 の「話者適応なし」の音声認識精度は，表 7.3 の「話者適応なし」の音声認識精度に比べてかなり悪い．音声認識結果を見ると，置換誤りが非常に多く発生しており，言語モデルのドメインが異なっている可能性が考えられる．表 7.6 より，国会会議録から検出した発音ラベルを用いる方法と，国会会議録から検

[*3] 国土交通委員会および環境委員会

出した発音ラベルと国会会議録と LVCSR の一致部分を併用する方法は，双方ともに他のベースライン手法を上回っており，この 2 つの方法によって取り出された発音ラベルが話者適応に対して有効であることを示している．次に，話者 A,B の 2 名について，話者適応用データの 1 人あたり分量を，約 2k 音節，約 4k 音節，約 6k 音節の 3 通りに変化させて，話者適応の効果がどのように変化するか実験した結果を図 7.7 に示す．音声認識性能は話者 2 名の平均値である．図 7.7 より，国会会議録から検出した発音ラベルを用いる方法と，国会会議録から検出した発音ラベルと国会会議録と LVCSR の一致部分を併用する方法は，1 つの例外を除いて，ベースライン手法を上回っている．例外は，4k 音節を適応用データとして，国会会議録全体を発音ラベルとして話者適応を行った場合である．この時，話者 A では，2 つの提案手法がベースライン手法を上回っているものの，話者 B では逆転しており，2 人の話者の平均値としては逆転した結果となっている．そのため，この事例については例外とすると，全体としては 2 つの提案手法の有効性が示されていると考えられる．また，LVCSR による認識結果全体を用いる方法と，国会会議録と LVCSR の一致部分を用いる方法の 2 つを比較すると，1 人あたり 4k 音節までは認識結果全体を用いる方法が上回っているが，1 人あたり 6k 音節に適応用データを増やすと，一致部分を用いる方法が逆転している．よって，話者適応用データを増やすことによって，高品質な非整形部分を使うことによるカバー率の低下を補うことができると考えられる．

表 7.4 追加データ

話者	テスト用音節数	話者適応用音節数 2k	4k	6k	総音節数
話者 A	2,729	1,994	3,997	5,998	8,727
話者 B	2,000	1,997	3,979	6,000	8,000
話者 C	2,108	1,996	3,992	0	6,100
話者 D	3,125	1,997	0	0	5,122
話者 E	2,818	1,997	0	0	4,815
話者 F	2,413	1,992	0	0	4,405

表 7.5 追加データに対する話者適応実験（発音ラベル諸元）

話者適応用発音ラベルの作成方法	ラベル数	精度 (%)	カバー率 (%)
話者適応なし	0	—	0
国会会議録全体を発音ラベルとして使用	22,171 (92.6%)	80.1	95.3
LVCSR による認識結果全体を発音ラベルとして使用	25,385 (106.1%)	63.1	95.2
国会会議録と LVCSR の一致部分 (5 単語以上) を使用	7,510 (31.4%)	89.3	80.9
提案法によって取り出した発音ラベル ($recall = 90\%$) を使用	18,018 (75.3%)	88.2	92.2
提案法によって取り出した発音ラベル ($recall = 90\%$) と，国会会議録と LVCSR の一致部分 (5 単語以上) の併用	18,921 (79.1%)	89.0	92.5
正確な書き起こしの音節ラベルを使用	23,932 (100.0%)	100	96.4

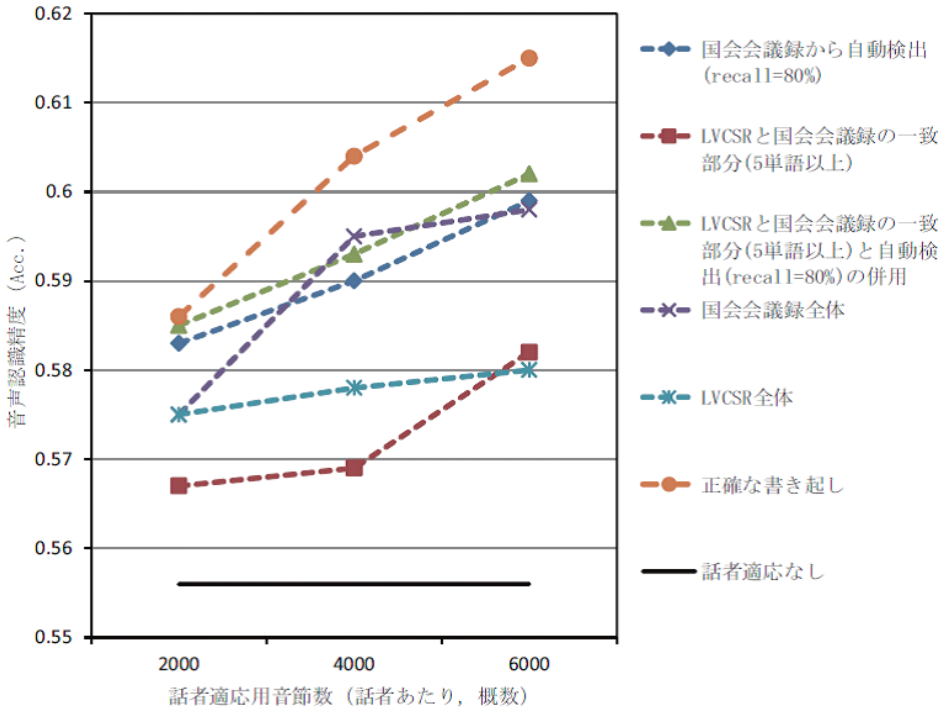


図 7.7　話者適応用音節数と音声認識精度の比較

表 7.6　追加データに対する話者適応実験（音声認識性能）

話者適応用発音ラベルの作成方法	Cor. (%)	Acc. (%)
話者適応なし	58.5	54.3
国会会議録全体を発音ラベルとして使用	61.7	56.7
LVCSR による認識結果全体を発音ラベルとして使用	60.7	55.4
国会会議録と LVCSR の一致部分 (5 単語以上) を使用	60.7	55.9
提案法によって取り出した発音ラベル ($recall = 90\%$) を使用	62.0	57.1
提案法によって取り出した発音ラベル ($recall = 90\%$) と，国会会議録と LVCSR の一致部分 (5 単語以上) の併用	62.0	57.1
正確な書き起こしの音節ラベルを使用	62.7	57.4

7.6　まとめ

本章では，整形された書き起こしから整形箇所を自動検出する手法を提案した．提案法は 2 段階からなる．第 1 に，整形された書き起こしと原音声のアラインメントを行い，第 2 に，アラインメントによって得られた素性に基づく SVM を用いて，整形箇所を検出する．国会会議録を用いた評価実験により，提案法は，整形された書き起こしと正確な書き起こしのパラレルコーパスをできるだけ少ないコストで用意するための目標性能を達成できることを示した．また，提案法によって取り出された非整形部分は，話者適応の発音ラベルとして有効であることを示した．

提案法 (または，提案法と LVCSR の認識結果を用いる方法の併用) を用いて非整形部

分を取り出す場合，非整形部分の精度を高くすると，再現率が低くなる (図 7.6)．そのような非整形部分を発音ラベルとして用いて話者適応する場合，発音ラベルの質は良いが，カバー率が低いため，話者適応の効果は小さくなる．今後の検討課題として，提案手法の有効性を有意差検定によって検証することが挙げられる．

第 8 章

結論

本研究では，話し言葉コーパスと書き言葉コーパスの比較・分析に基づき，話し言葉特有の現象を統計的にモデル化する手法について検討した．

1 章では，本研究の背景について述べた．

2 章では，現在の音声認識の基本原理と，隠れマルコフモデルに基づく音響モデル，N-gram に基づく統計的言語モデル，および，音声認識性能の評価方法について説明した．

3 章では，話し言葉コーパスや書き言葉コーパスの比較・分析を行った。まず，日本語話し言葉コーパスを用いて，フィラーや言い直しといった話し言葉特有の現象の出現頻度を調査し，フィラーが特に出現頻度が高く，優先的に対処する必要があることを示した．また，フィラーの語彙について調査を行い，特に頻度の高い 6 種類のフィラー語彙のみで，フィラーの出現総数の約 9 割がカバーされること，および，フィラーの直前ではモーラや品詞等のコンテキストに偏りが生じることを明らかにした．

続いて，言語モデルの学習に用いるコーパスの文単位に比べ，ポーズに基づく認識処理単位は非常に短く，文・発話における単語の結束性を十分に利用することができないことを示した．また，こうした文・発話の境界を，ポーズの情報に基づいて弁別することが非常に困難であることを示した．

さらに，会議録や速記録といった整形された書き起こしの例として，国会会議録における様々な整形処理について調査・分析した．

4 章以降では，3 章での分析に基づき，話し言葉特有の現象をモデル化する方法について検討した。

まず，4 章では，話し言葉特有の現象の中でも特に出現頻度の高いフィラーに対処するため，フィラー予測モデルを用いる方法を提案した．提案手法は 2 段階からなり，最初に，正確な話し言葉コーパスからフィラー予測モデルを作成し，次に，このモデルから与えられる確率に基づいてフィラーを挿入したコーパスから言語モデルを構築した．日本語話し言葉コーパスを対象とした実験により，提案手法は，実際の正確な話し言葉コーパス

から作成された言語モデルにかなり近い言語モデルを作成できることを示した．また，国会会議録を対象とした認識実験により，提案手法は，従来の手法よりも高い認識率を達成することができることを示した．

5 章では，音声対話システムの応答音声にフィラーを挿入することで，応答音声の自然性や聞き易さを向上させることについて検討した．特に，文頭や文節境界にランダムにフィラーを挿入するのではなく，文脈等を考慮してより適切な位置に挿入することで，自然性や聞き易さ, 理解度等の改善を行った．

また，6 章では，ポーズに基づく単位や文単位の境界をショートポーズとして扱い，前後で単語履歴を引き継ぐことにより，パープレキシティが改善することを示した．さらに，コーパスの文単位と実際の認識における認識単位との不一致を考慮し，コーパス中にポーズを挿入する手法を提案した．国会会議録に対する実験の結果，言語モデルのパープレキシティおよび音声認識精度を改善することができた．

さらに，7 章では，整形された書き起こしから整形箇所を自動検出する手法を提案した．提案法は 2 段階からなる．第 1 に，整形された書き起こしと原音声のアラインメントを行い，第 2 に，アラインメントによって得られた素性に基づく SVM を用いて，整形箇所を検出する．国会会議録を用いた評価実験により，提案法は，整形された書き起こしと正確な書き起こしのパラレルコーパスをできるだけ少ないコストで用意するための目標性能を達成できることを示した．また，提案法によって取り出された非整形部分は，話者適応の発音ラベルとして有効であることを示した．

今後の課題としては，まず，個々の手法のさらなる改善として，フィラーの個人性を考慮したフィラー予測モデルの検討や，ショートポーズの予測モデルの高精度化のほか，これらの予測モデルを拡張して，音声対話システムの応答文の聞き易さ・自然の改善に利用することが挙げられる．すなわち，音声認識用言語モデルの観点ではなく，音声対話システムにおける聞き易さ・自然さの観点から，応答文中の最適な位置にフィラーやポーズを挿入するモデルについて検討を行う．特に，構文木や依存木などの長距離の素性に基づいて，言語的な区切りを考慮するようなモデル化を検討する．

また，整形された書き起こしから整形箇所を自動検出する手法において，アラインメントを行う際のフィラーやポーズの扱いを改善することも考えられる．特に，フィラー予測モデルやショートポーズの予測モデルをアラインメントに組み込むことで，アラインメントの精度を改善し，より高精度な音響モデルの学習ラベルを生成できる可能性がある．このほか，言語モデリングへの応用として，提案法を用いてパラレルコーパスを作成し，書き言葉から話し言葉への変換規則を収集することにより，任意のドメインに対して，対応する書き言葉コーパスから話し言葉コーパスを作成し，ドメイン依存の話し言葉言語モデルを構築することが考えられる．

さらに，これまでに提案した各現象のモデル化手法をより統合的に扱うための枠組みに

ついても検討していきたい.

謝辞

本研究を進めるにあたり，終始熱心かつ適切なご指導・ご鞭撻を賜りました豊橋技術科学大学情報・知能工学系中川聖一教授に深く感謝の意を表します．

本論文をまとめるにあたり多くの御助言をいただきました豊橋技術科学大学情報メディア基盤センター井佐原均教授，同大学情報・知能工学系秋葉友良准教授に厚く御礼申し上げます．

名古屋大学大学院情報科学研究科北岡教英准教授には，音声対話の研究に関して大変有益な御助言をいただきました．厚く御礼申し上げます．

豊橋技術科学大学情報・知能工学系山本一公准教授，同大学情報メディア基盤センター土屋雅稔助教には，音声認識や言語モデルの研究に関して日頃から多くの御助言をいただきました．厚く御礼申し上げます．

中川研究室博士・OB 研究報告会参加者の方々には，博士課程における研究に関して様々な御助言を頂きました．厚く御礼申し上げます．

また，中川研究室の皆様からは，日頃から昼夜問わず様々なご助力をいただきました．ここに感謝の意を表します．

参考文献

[1] 藤井康寿. 音声ドキュメントの音声認識、整形、要約に関する研究. 豊橋技術科学大学大学院工学研究科電子・情報工学専攻平成 24 年度博士論文, 2012.

[2] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄. 音声認識システム. オーム社, 2001.

[3] 中川聖一, 高木英行. パターン認識における有意差検定と音声認識システム評価法. 日本音響学会誌, Vol. 50, No. 10, pp. 849–854, 1994.

[4] 河原達也. 議会の会議録作成のための音声認識—衆議院のシステムの概要—. 情報処理学会研究報告, 2012-SLP-93-5, 2012.

[5] T.Kawahara. Transcription system using automatic speech recognition for the japanese parliament (diet). In *Proc. of AAAI/IAAI*, pp. 2224–2228, 2012.

[6] 猿谷豊. 衆議院における音声認識を利用した会議録作成業務. 情報管理, Vol. 55, No. 6, pp. 392–399, 2012.

[7] 佐藤庄衛. 音声認識を用いた生放送番組への字幕付与. メディア教育研究, Vol. 9, No. 1, S9-S18, 2012.

[8] Kikuo Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, pp. 7–12, Tokyo, Japan, 2003.

[9] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilayet. Recent progress in the mit spoken lecture processing project. In *Proc. of Interspeech*, pp. 2553–2556, 2007.

[10] 南條浩輝, 秋田祐哉, 河原達也. 音声認識を利用した会議録・講演録の作成支援システムの設計と評価. 日本音響学会秋季研究発表会講演論文集, 17-13, 2005.

[11] Thomas Hain, John Dines, Giulia Garau, Martin Karafiat, Darren Moore, Vincent Wan, Roeland Ordelman, and Steve Renals. Transcription of conference room meetings: An investigation. In *Proc. of Interspeech*, pp. 1661–1664, 2005.

[12] J.J. Godfrey, E.C. Holliman, and J. McDaniel. Switchboard: telephone speech

corpus for research and development. In *Proc. of ICASSP*, pp. 517–520, 1992.

[13] C. Cieri, D. Miller, and K. Walker. The fisher corpus: a resource for the next generations of speech-to-text. In *Proc. of LREC*, pp. 69–71, 2004.

[14] Jonathan Fiscus, John Garofolo, Mark Przybocki, William Fisher, and David Pallett. 1997 english broadcast news speech (hub4). In *Linguistic Data Consortium*, 1997.

[15] A.Park, T.Hazen, and J.Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In *Proc. of ICASSP*, pp. 497–500, 2005.

[16] Xiaojin Zhu and R. Rosenfeld. Improving trigram language modeling with the world wide web. In *Proc. of ICASSP*, pp. 533–536, 2001.

[17] Ivan Bulyko, Mari Ostendorf, and Andreas Stolcke. Getting more mileage fromweb text sources for conversational speech language modeling using class-dependent mixtures. In *Proc. of HLT*, Vol. 2, pp. 7–9, 2003.

[18] 翠輝久, 河原達也. ドメインとスタイルを考慮した web テキストの選択による音声対話システム用言語モデルの構築. 電子情報通信学会論文誌. D, Vol. 90, No. 11, pp. 3024–3032, 2007.

[19] 南條浩輝, 河原達也, 山田篤, 内元清貴. 講演音声認識のための言語モデルの教師なし適応. 電子情報通信学会技術研究報告, NLC2002-75, pp. 25–30, 2002.

[20] 秋田祐哉, 河原達也. 言語モデルと発音辞書の統計的話し言葉変換に基づく国会音声認識. 情報処理学会研究報告, 2007-SLP-69-11, 2007.

[21] 渡邉真人, 秋田祐哉, 河原達也. 予稿の話し言葉変換に基づく言語モデルによる講演音声認識. 情報処理学会研究報告, SLP-89-1, 2011.

[22] 黒橋禎夫, 大泉敏貴, 柴田知秀, 鍜治伸裕, 河原大輔, 岡本雅史, 西田豊明. 会話型知識プロセスのための言語情報のメディア変換. 社会技術研究論文集, Vol. 2, pp. 173–180, 2004.

[23] 太田健吾, 土屋雅稔, 中川聖一. 講義・講演音声におけるフィラー，言い淀み，倒置の発生頻度の分析. 日本音響学会秋季研究発表会講演論文集, 2-P-30, 2006.

[24] 西村雅史, 伊東伸泰. 講義コーパスを用いた自由発話の大語彙連続音声認識 (音声情報処理 : 現状と将来技術論文特集). 電子情報通信学会論文誌. D-II,, Vol. 83, No. 11, pp. 2473–2480, 2000.

[25] Young-Hee Park and Minhwa Chung. Style-specific language model adaptation for korean conversational speech recognition. In *Proc. of the International Conference of Natural Language Processing and Knowledge Engineering*, pp. 591–596, 2003.

[26] Andreas Stolcke and Elizabeth Shriberg. Statistical language modeling for speech disfluencies. In *Proc. of ICASSP*, pp. 405–408, 1996.

[27] 稲垣貴彦, 廣瀬啓吉, 峯松信明. 話し言葉音声認識における韻律的特徴を利用したフィラー検出. 日本音響学会春季研究発表会講演論文集, 3-10-20, 2008.

[28] Michiko Watanabe, Keiichi Hirose, Yasuharu Den, and Nobuaki Minematsu. Filled pauses as cues to the complexity of following phrases. In *Proc. of INTERSPEECH 2005*, pp. 37–40, 2005.

[29] M.Somiya, K.Kobayashi, H.Nishizaki, and Y.Sekiguchi. The effect of filled pauses in a lecture speech on impressive evaluation of listeners. In *Proc. of INTERSPEECH2007*, pp. 2673–2676, 2007.

[30] 伊藤敏彦, 峯松信明, 中川聖一. 間投詞の働きの分析とシステム応答生成における間投詞の利用と評価. 日本音響学会誌, 第 55 巻 of *No.5*, pp. 333–342, 1999.

[31] Toshiyuki Shiwa, Takayuki Kanda, Michita Imaia, Hiroshi Ishiguro, and Norihiro Higata. How quickly should a communication robot respond? *International Journal of Social Robotics*, Vol. 1, No. 2, pp. 141–155, 2009.

[32] R. B. Miller. Response time in man-computer conversational transactions. In *Proc. of Spring Joint Computer Conference*, pp. 267–277, 1968.

[33] T. Starner. The challenges of wearable computing: Part 2. *IEEE Micro*, Vol. 21, No. 4, pp. 54–67, 2001.

[34] 尾嶋憲治, 秋田祐哉, 河原達也. 局所的な係り受けと韻律の素性を用いた話し言葉の節・文境界推定. 情報処理学会研究報告, 2007-SLP-67-3, 2007.

[35] Y.Liu and et.al E.Shriberg. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. Audio, Speech and Language Process*, Vol. 14, No. 5, pp. 1526–1539, 2006.

[36] Y.Liu and et al N.V.Chawla. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech and Language*, Vol. 20, pp. 468–494, 2006.

[37] 南條浩輝, 加藤一臣, 李晃伸, 河原達也. 大規模な日本語話し言葉データベースを用いた講演音声認識. 電子情報通信学会論文誌, Vol. J86-DII, No. 4, pp. 450–459, 2003.

[38] 寺尾真, 広瀬啓吉, 峯松信明. アクセント句境界情報を利用した n-gram 言語モデルの高精度化. 情報処理学会研究報告, 2001-SLP-39-18, 2001.

[39] 上西康太, 広瀬啓吉, 峯松信明. アクセント句境界を考慮した言語モデルの日本語話し言葉コーパスへの適用. 日本音響学会春季研究発表会講演論文集, 3-9-2, 2007.

[40] 鄭聖曄, 広瀬啓吉, 峯松信明. 文節境界情報を利用した n-gram 言語モデルの高精度

化. 情報処理学会研究報告, SLP2003-47, pp. 13–18, 2003.

[41] 細田聖人, 広瀬啓吉, 峯松信明. 話し言葉認識における文節境界情報を用いた言語モデルに関する検討. 日本音響学会秋季研究発表会講演論文集, 3-1-8, 2008.

[42] 増村亮, 咸聖俊, 伊藤彰則. Web データを用いた話し言葉用言語モデルの作成. 第 5 回音声ドキュメント処理ワークショップ講演論文集, pp. 77–82, 2011.

[43] Jason D. Williams, I. Dan Melamed, Tirso Alonso, Barbara Hollister, and Jay Wilpon. Crowd-sourcing for difficult transcription of speech. In *Proc. of the Automatic Speech Recognition and Understanding Workshop*, pp. 535–540, 2011.

[44] I. Makhoul. Spectral linear prediction: properties and applications. *IEEE Trans. ASSP-23*, pp. 283–296, 1975.

[45] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, Vol. 87, No. 4, pp. 1738–1752, 1990.

[46] 山本一公, 中川聖一. 長時間位相特徴と振幅スペクトル特徴の併用による音声認識の検討. 日本音響学会秋季研究発表会講演論文集, 2-Q-13, 2011.

[47] H. Hermansky, D. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional hmm systems. In *Proc. ICASSP*, pp. 1635–1638, 2000.

[48] H. Hermansky and S. Sharma. Temporal patterns(traps) in asr of noisy speech. In *Proc. ICASSP*, pp. 289–292, 1999.

[49] S. Shimizu, M. Suzuki, N. Minematsu, and K. Hirose. An experimental study on dynamic features of speech structure. *Journal of Research Institute of Signal Processing*, Vol. 16, No. 4, pp. 319–322, 2012.

[50] Narpendyah Ariwardhani, Yurie Iribe, Kouichi Katsurada, and Tsuneo Nitta. Phoneme recognition based on af-hmms with an optimal parameter set. In *Proc. of NCSP12*, pp. 170–173, 2012.

[51] 古井貞熙. ディジタル音声処理. 東海大学出版会, 1985.

[52] Yutaka Tsurumi and Seiichi Nakagawa. An unsupervised speaker adaptation method for cotinuous parameter hmm by maximum a posteriori probability estimation. In *Proc. of ICSLP'94*, pp. 431–434, 1994.

[53] 中川聖一. 確率モデルによる音声認識. 電子情報通信学会, 1988.

[54] Lee C.H., Lin C.H., and Juang B.H. A study on speaker adaptation of the parameters of continuous density hidden markov models. In *IEEE Trans. Signal Processing*, 39, pp. 806–814, 1991.

[55] Fukunaga K. *Introduction to Statistical Pattern Recognition.* 2nd Edition, Academic Press, 1990.

[56] NilssonN.J.(渡辺茂訳). 学習機械. コロナ社, 1967.

[57] 王龍標. 遠隔発話の音声認識と話者認識に関する研究. 豊橋技術科学大学大学院工学研究科情報工学専攻 平成 16 年度修士論文, 2004.

[58] I. H. Witten and T. C. Bell. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, Vol. 37, No. 4, pp. 1085–1094, Jul 1991.

[59] 中川聖一, 赤松裕隆. 未知語を含む文集合のパープレキシティの算出法—新補正パープレキシティ—. 日本音響学会秋季研究発表会講演論文集, 2-1-3, 1998.

[60] 土屋雅稔, 小暮悟, 西崎博光, 太田健吾, 山本一公, 中川聖一. 日本語講義音声コンテンツコーパスの作成と分析. 情報処理学会論文誌, Vol. 50, No.2, pp. 448–450, 2008.

[61] 前川喜久雄. 『日本語話し言葉コーパス』の概観. `http://www.ninjal.ac.jp/corpus_center/csj/manu-f/overview.pdf`.

[62] 小磯他. 転記テキストの仕様. `http://www.ninjal.ac.jp/corpus_center/csj/manu-f/transcription.pdf`.

[63] 高梨他. 『日本語話し言葉コーパス』における節単位認定. `http://www.ninjal.ac.jp/corpus_center/csj/manu-f/clause.pdf`.

[64] 内元他. 『日本語話し言葉コーパス』における係り受け構造附与. `http://www.ninjal.ac.jp/corpus_center/csj/manu-f/dependency.pdf`.

[65] 中川聖一, 小林聡. 自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質. 日本音響学会誌, Vol. 51, No. 3, pp. 202–210, 1995.

[66] 高梨克也, 丸山岳彦, 内元清貴, 井佐原均. 話し言葉の文境界 —CSJ コーパスにおける文境界の定義と半自動認定—. 言語処理学会第 9 回年次大会発表論分集, pp. 521–524, 2003.

[67] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of ICML*, pp. 282–289, 2001.

[68] 水上悦雄, 山下耕二. 対話におけるフィラーの発話権保持機能の検証. 認知科学, Vol. 14, No. 4, pp. 588–603, 2007.

[69] 南條浩輝, 河原達也, 篠崎隆宏, 古井貞熙. 音声認識のための音響モデルと言語モデルの仕様. `http://www.ninjal.ac.jp/corpus_center/csj/manu-f/asr.pdf`.

[70] 北岡教英, 新宮将久, 中川聖一. 言語的・音響的コンテキストが講演音声の聴き取りおよび認識に及ぼす効果. 電子情報通信学会技術研究報告, SP2003-33, 2003.

[71] Wataru Naito, Hiromitsu Nishizaki, and Yoshihiro Sekiguchi. Evaluation and advice system for improving the manner of speaking in lectures using features of filled pauses. In *Proc. of APSIPA ASC 2011*, p. 4 pages, 2011.

[72] Nobuaki Minematsu and Seiichi Nakagawa. Correlation between acoustic pauses and perceptual pauses in speech. In *Proc. of ASR and ASJ Third Joint Meeting*, pp. 1193–1198, 1996.

[73] 北原義典, 武田昌一, 市川熹, 東倉洋一. 音声言語認知における韻律の役割. 電子情報通信学会論文誌, Vol. J70-D, No. 11, pp. 2095–2101, 1987.

[74] Michiko Watanabe. *Features and Roles of Filled Pauses in Speech Communication.* 羊書房, 2009.

[75] 太田健吾, 土屋雅稔, 中川聖一. ポーズを考慮した話し言葉言語モデルの構築. 情報処理学会論文誌, Vol. 53, No. 2, pp. 889–900, 2012.

[76] David L. Strayer and William A. Johnston. Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science*, Vol. 12, No. 6, pp. 462–466, 2001.

[77] Jordi Adell, David Escudero, and Antonio Bonafonte. Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, Vol. 54, No. 3, pp. 459 – 476, 2012.

[78] 西光雅弘, 高梨克也, 河原達也. 係り受けとポーズ・フィラーの情報を用いた話し言葉の段階的チャンキング (session-8 ポスターセッション: 一般, 第 7 回音声言語シンポジウム). 情報処理学会研究報告. SLP, 音声言語情報処理, pp. 247–252, 2005.

[79] 緒方淳, 後藤真孝, 伊藤克亘. 有声・無声休止区間の自動検出を考慮したデコーディングによる自由発話音声認識の性能改善. 電子情報通信学会論文誌, Vol. J92-D, No. 2, pp. 226–235, 2009.

[80] 太田健吾, 土屋雅稔, 中川聖一. 音声認識言語モデルにおけるポーズ情報の扱いに関する検討. 第 3 回音声ドキュメント処理ワークショップ講演論文集, pp. 77–82, 2009.

[81] 森信介, 笹田鉄郎, Neubig Graham. 確率的タグ付与コーパスからの言語モデル構築. 情報処理学会研究報告, 2010-NL-196, pp. 1–7, 2010.

[82] 太田健吾, 土屋雅稔, 中川聖一. フィラー予測モデルに基づく話し言葉言語モデルの構築. 情報処理学会論文誌, Vol. 50, No. 2, pp. 477–487, 2009.

[83] J. Zhang, L. Wang, and S. Nakagawa. Lvcsr based on context dependent syllable acoustic models. In *Proc. of Asian Workshop on Speech Science and Technology*, SP2007-200, pp. 81–86, 2008.

[84] 高橋伸寿, 中川聖一. コンテキスト依存音節単位 HMM の評価. 日本音響学会春季研究発表会講演論文集, 3–3–2, 2001.

[85] 秋田祐哉, 三村正人, 河原達也. 会議録作成支援のための国会審議の音声認識システム. 日本音響学会春季研究発表会講演論文集, 3–5–7, 2009.

[86] M.Gabrea, D.O' Shaughnessy. Detection of filled pauses in spontaneous conversational speech. In *In Proc. of ICSLP*, pp. 678–681, 2000.

[87] Frederik Stouten, Jacques Duchateau, Jean-Pierre Martens, and Patrick Wambacq. Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. *Speech Communication*, Vol. 48, No. 11, pp. 1590 – 1606, 2006.

[88] 南條浩輝, 河原達也, 篠崎隆宏, 古井貞熙. 音声認識のための音響モデルと言語モデルの仕様 Ver.1.0, 2004. (CSJ コーパス付属文書).

[89] Lori Lamel, Jean Luc Gauvain, and Gilles Adda. Investigating lightly supervised acoustic model training. In *Proc. of ICASSP*, pp. 477–480, 2001.

[90] 三村正人, 秋田祐哉, 河原達也. 統計的言語モデル変換を用いた音響モデルの準教師つき学習. 電子情報通信学会論文誌, Vol. J94–D, No. 2, pp. 460–468, 2011.

[91] Benjamin Lecouteux, Georges Linars, and Stanislas Oger. Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. *Computer Speech and Language*, Vol. 26, No. 2, pp. 67 – 89, 2012.

[92] 緒方淳, 後藤真孝. Podcastle: 集合知を活用した音響モデル学習による音声認識の性能向上. 日本音響学会春季研究発表会講演論文集, 2–5–1, 2009.

[93] I. Lane, A. Waibel, M. Eck, and K. Rottmann. Tools for collecting speech corpora via mechanical-turk. Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon' s Mechanical Turk, pp. 184–187, 2010.

[94] Graham Neubig, Yuya Akita, Mori Shinsuke, , and Kawahara Tatsuya. Improved statistical models for smt-based speaking style transformation. In *Proc. of ICASSP*, pp. 5206–5209, 2010.

[95] Brandon C. Roy, Soroush Vosoughi, and Deb Roy. Automatic estimation of transcription accuracy and difficulty. In *Proceeding of Interspeech*, pp. 1902–1905, 2010.

[96] 丸山一郎, 阿部芳春, 江原暉将, 白井克彦. ドキュメンタリー番組における字幕送出タイミング検出の一検討. 日本音響学会秋季研究発表会講演論文集, 3–Q–30, pp. 177–178, 1999.

[97] Matthias Paulik and Panchi Panchapagesan. Leveraging large amounts of loosely transcribed corporate videos for acoustic model training. In *Proc. of the Automatic Speech Recognition and Understanding Workshop*, pp. 95–100, 2011.

[98] Taku Kudoh. TinySVM. `http://chasen.org/~taku/software/TinySVM/`.

[99] X. Huang, A. Acero, A. Acero, and H.W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, 2001.

[100] 太田健吾, 土屋雅稔, 中川聖一. 整形された会議録とその原音声のアラインメントに基づく整形箇所の自動検出. 第 5 回音声ドキュメント処理ワークショップ講演論文集, 5-16, 2011.

[101] Kengo Ohta, Masatoshi Tsuchiya, and Seiichi Nakagawa. Detection of precisely transcribed parts from inexact transcribed corpus. In *Proc. of the Automatic Speech Recognition and Understanding Workshop*, pp. 541–546, 2011.

[102] Y. Fujii, K. Yamamoto, and S. Nakagawa. Large vocabulary speech recognition system: Spojus++. In *Proceeding of 11th WSEAS International Conference MUSP–11*, 2011.

[103] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, 2006.

発表論文

- 学会論文誌
  1. 太田健吾, 土屋雅稔, 中川聖一, 『ポーズを考慮した話し言葉言語モデルの構築』, 情報処理学会論文誌, Vol.53, No.2, pp.889-900, 2012
  2. 太田健吾, 土屋雅稔, 中川聖一, 『フィラー予測モデルに基づく話し言葉言語モデルの構築』, 情報処理学会論文誌, Vol.50, No.2, pp.477-487, 2008
  3. 土屋雅稔, 小暮悟, 西崎博光, 太田健吾, 山本一公, 中川聖一, 『日本語講義音声コンテンツコーパスの作成と分析』, 情報処理学会論文誌, Vol.50, No.2, pp.448-450, 2008.
- 国際会議
  1. Kengo Ohta, Norihide Kitaoka, Seiichi Nakagawa, “Analyzing Effects of Filled Pauses and Silences in Responses of a Spoken Dialogue System”, The 6th World Congress: Applied Computing Conference, 7pages, 2013.
  2. Seiichi Nakagawa, Turmunkh Erdenebat, Hiroshi Kibishi, Kengo Ohta’, Yasuhisa Fujii, Masatoshi Tsuchiya, Kazumasa Yamamoto, “Development of large vocabulary continuous speech recognition system for Mongolian language”, Proc of SLTU12 , pp.19-23, 2012.
  3. Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa, “Developing Partially-Transcribed Speech Corpus from Edited Transcriptions”, Proc. of LREC, pp.3399-3404, 2012.
  4. Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa, “Detection of Precisely Transcribed Parts from Inexact Transcribed Corpus”, Proc. of ASRU, pp.541-546, 2011.
  5. Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa, “Automatic Detection of Edited Parts in Inexact Transcribed Corpora Based on Alignment between Edited Transcription and Corresponding Utterance”, The 11th WSEAS International Conference on Multimedia Systems & Signal Pro-

cessing, pp.155-162, 2011.

6. Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa, “Construction of Spoken Language Model from Written Language Corpora “, International Symposium on Electronics-Inspired Interdisciplinary Research, 1 page, 2010.
7. Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa, “Effective use of pause information in language modelling for speech recognition”, Proc. of Interspeech, pp.2691-2694, 2009.
8. Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa, “Evaluating Spoken Language Model Based on Filler Prediction Model in Speech Recognition”, Proc. of Interspeech, pp.1558-1561, 2008.
9. Masatoshi Tsuchiya, Satoru Kogure, Hiromitsu Nishizaki, Kengo Ohta and Seiichi Nakagawa, “Developing Corpus of Japanese Classroom Lecture Speech Contents”, Proc. of LREC, 5 pages, 2008.
10. Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa, “Construction of spoken language model including fillers using filler prediction model”, Proc. of Interspeech, pp.1489-1492, 2007.

- 国内学会・研究会

1. 太田健吾, 北岡教英, 中川聖一, 『音声対話システムの応答文におけるフィラーとポーズがユーザに与える影響の分析』, 日本音響学会, 講演論文集, 2-8-10, 4 pages, 2013.
2. 太田健吾, 土屋雅稔, 中川聖一, 『整形された書き起こしからの整形・非整形部分の自動検出』, 第 6 回音声ドキュメント処理ワークショップ講演論文集, 8 pages, 2012.
3. 太田健吾, 土屋雅稔, 中川聖一, 『整形された会議録からの整形箇所の自動検出』, GCOE シンポジウム東京, 1 page, 2012.
4. 太田健吾, 土屋雅稔, 中川聖一, 『整形された会議録とその原音声のアラインメントに基づく整形箇所の自動検出』, 第 5 回音声ドキュメント処理ワークショップ講演論文集, 8 pages, 2011.
5. 太田健吾, 土屋雅稔, 中川聖一,, 『書き言葉コーパスからの話し言葉用言語モデルの構築』, 第 3 回 GCOE センシングアーキテクト・シンポジウム, 1 page, 2010.
6. 太田健吾, 『書き言葉コーパスからの話し言葉用言語モデルの構築』, 第 2 回 GCOE センシングアーキテクト・シンポジウム, 1 page, 2009.
7. 太田健吾, 土屋雅稔, 中川聖一, 『音声認識用言語モデルにおけるポーズ情報の有効利用』, 日本音響学会春季講演論文集, 2-5-8, pp.59-62, 2009.

8. 太田健吾, 土屋雅稔, 中川聖一, 『音声認識言語モデルにおけるポーズ情報の扱いに関する検討』, 第 3 回音声ドキュメント処理ワークショップ講演論文集, pp.77-82, 2009
9. 太田健吾, 張健, 土屋雅稔, 山本一公, 中川聖一, 『国会答弁の音声認識の検討』, 平成 20 年度電気関係学会東海支部連合大会予稿集, pp.1-6 , 2008.
10. 太田健吾, 土屋雅稔, 中川聖一, 『フィラー予測モデルを用いた話し言葉言語モデルの音声認識による評価』, 第 2 回音声ドキュメント処理ワークショップ講演論文集, pp.1-6 , 2008.
11. 太田健吾, 土屋雅稔, 中川聖一, 『フィラーの書き起こしのないコーパスからのフィラー付き言語モデルの構築』, 情報処理学会, 音声言語情報処理研究会, SLP-67-1, 6 pages, 2007.
12. 太田健吾, 土屋雅稔, 中川聖一, 『フィラーの挿入モデルを用いた話し言葉の言語モデルの構築』, 日本音響学会, 講演論文集, 1-9-16, pp.31-32, 2007.
13. 土屋雅稔, 太田健吾, 中川聖一, 『フィラー予測モデルに基づくフィラー付き言語モデルの構築』, 第 1 回音声ドキュメント処理シンポジュウム, pp.81-88, 2007.
14. 太田健吾, 土屋雅稔, 中川聖一, 『講義・講演音声におけるフィラー、言い淀み、倒置の発生頻度の分析』, 日本音響学会, 講演論文集, 2-P-30, pp.153-154, 2006.

付録 A

フィラーの出現とモーラの関係

表 A.1: フィラーの直前に出現したモーラ

No.	モーラ	出現数	[%]	累積 [%]
1	デ	42669	9.60	9.60
2	ワ	36912	8.30	17.90
3	テ	32910	7.40	25.31
4	ノ	30841	6.94	32.25
5	ガ	27170	6.11	38.36
6	ス	26466	5.95	44.31
7	ニ	24005	5.40	49.71
8	モ	22068	4.96	54.68
9	ト	21857	4.92	59.60
10	ネ	14476	3.26	62.85
11	タ	13654	3.07	65.93
12	エ	13410	3.02	68.94
13	オ	13342	3.00	71.94
14	ラ	11664	2.62	74.57
15	ン	11348	2.55	77.12
16	カ	9418	2.12	79.24
17	マ	8892	2.00	81.24
18	イ	7834	1.76	83.00
19	ナ	6214	1.40	84.40
20	リ	5927	1.33	85.73

表 A.1: フィラーの直前に出現したモーラ

No.	モーラ	出現数	[%]	累積 [%]
21	ル	5734	1.29	87.02
22	ド	5517	1.24	88.27
23	シ	4911	1.10	89.37
24	ユ	4702	1.06	90.43
25	ク	4376	0.98	91.41
26	ア	3916	0.88	92.29
27	バ	3227	0.73	93.02
28	ズ	2092	0.47	93.49
29	コ	1958	0.44	93.93
30	ウ	1939	0.44	94.37
31	ツ	1795	0.40	94.77
32	キ	1537	0.35	95.12
33	レ	1500	0.34	95.45
34	ダ	1244	0.28	95.73
35	チ	1151	0.26	95.99
36	head	1140	0.26	96.25
37	ロ	1033	0.23	96.48
38	ヤ	947	0.21	96.70
39	ッ	800	0.18	96.88
40	ケ	753	0.17	97.04
41	ジ	721	0.16	97.21
42	メ	684	0.15	97.36
43	ゴ	656	0.15	97.51
44		649	0.15	97.65
45	ヨ	597	0.13	97.79
46	セ	595	0.13	97.92
47	ショ	563	0.13	98.05
48	ム	484	0.11	98.16
49	ホ	472	0.11	98.26
50	ソ	468	0.11	98.37

表 A.1: フィラーの直前に出現したモーラ

No.	モーラ	出現数	[%]	累積 [%]
51	φ	389	0.09	98.46
52	ヒ	373	0.08	98.54
53	サ	356	0.08	98.62
54	ミ	351	0.08	98.70
55	ブ	348	0.08	98.78
56	ジョ	327	0.07	98.85
57	ハ	320	0.07	98.92
58	ジャ	299	0.07	98.99
59	シュ	287	0.06	99.06
60	フ	281	0.06	99.12
61	グ	262	0.06	99.18
62	ジュ	223	0.05	99.23
63	ウォ	215	0.05	99.28
64	ビ	210	0.05	99.32
65	シャ	190	0.04	99.37
66	ヘ	177	0.04	99.41
67	キョ	173	0.04	99.45
68	ギ	172	0.04	99.48
69	キュ	160	0.04	99.52
70	ポ	152	0.03	99.55
71	ボ	133	0.03	99.58
72	チョ	128	0.03	99.61
73	ゼ	121	0.03	99.64
74	ヌ	121	0.03	99.67
75	プ	117	0.03	99.69
76	トゥ	117	0.03	99.72
77	チュ	102	0.02	99.74
78	パ	91	0.02	99.76
79	ザ	88	0.02	99.78
80	リョ	85	0.02	99.80

表 A.1: フィラーの直前に出現したモーラ

No.	モーラ	出現数	[%]	累積 [%]
81	ゲ	80	0.02	99.82
82	ティ	78	0.02	99.84
83	ベ	69	0.02	99.85
84	ヒョ	64	0.01	99.87
85	ゾ	62	0.01	99.88
86	ギョ	53	0.01	99.89
87	ニュ	44	0.01	99.90
88	ディ	40	0.01	99.91
89	ピ	35	0.01	99.92
90	×	35	0.01	99.93
91	チャ	31	0.01	99.94
92	リュ	31	0.01	99.94
93	ドゥ	28	0.01	99.95
94	ピョ	26	0.01	99.95
95	ヒュ	18	0.00	99.96
96	ビョ	17	0.00	99.96
97	ペ	16	0.00	99.97
98	リャ	16	0.00	99.97
99	ニョ	13	0.00	99.97
100	ジェ	12	0.00	99.98
101	ヒャ	11	0.00	99.98
102	ニャ	10	0.00	99.98
103	ファ	9	0.00	99.98
104	ウェ	9	0.00	99.98
105	キャ	8	0.00	99.99
106	ギュ	7	0.00	99.99
107	ビュ	6	0.00	99.99
108	フィ	6	0.00	99.99
109	スィ	6	0.00	99.99
110	ニェ	4	0.00	99.99

表 A.1: フィラーの直前に出現したモーラ

No.	モーラ	出現数	[%]	累積 [%]
111	ミュ	4	0.00	99.99
112	シェ	3	0.00	99.99
113	ツェ	3	0.00	99.99
114	ウィ	3	0.00	100.00
115	イェ	3	0.00	100.00
116	フォ	3	0.00	100.00
117	ギャ	2	0.00	100.00
118	クヮ	2	0.00	100.00
119	チェ	2	0.00	100.00
120	フェ	2	0.00	100.00
121	デュ	2	0.00	100.00
122	ツァ	1	0.00	100.00
123	ヴィ	1	0.00	100.00
124	ピュ	1	0.00	100.00
125	ツォ	1	0.00	100.00
126	ミャ	1	0.00	100.00
127	ミョ	1	0.00	100.00

表 A.2: モーラのユニグラム統計

No.	モーラ	出現数	[%]	累積 [%]
1	ン	745928	5.72	5.72
2	イ	645148	4.94	10.66
3	ト	576029	4.41	15.07
4	ノ	559506	4.29	19.36
5	カ	470397	3.60	22.96
6	テ	434972	3.33	26.30
7	デ	420469	3.22	29.52
8	シ	394829	3.03	32.54

表 A.2: モーラのユニグラム統計

No.	モーラ	出現数	[%]	累積 [%]
9	タ	393722	3.02	35.56
10	ス	383961	2.94	38.50
11	マ	372173	2.85	41.35
12	ッ	349257	2.68	44.03
13	ナ	339944	2.60	46.64
14	オ	339269	2.60	49.24
15	コ	336643	2.58	51.81
16	エ	315840	2.42	54.23
17	ニ	312064	2.39	56.63
18	ワ	286922	2.20	58.82
19	ア	279158	2.14	60.96
20	ク	276394	2.12	63.08
21	ガ	273310	2.09	65.18
22	モ	265020	2.03	67.21
23	ル	236659	1.81	69.02
24	キ	205811	1.58	70.60
25	レ	198589	1.52	72.12
26	ソ	184263	1.41	73.53
27	リ	175489	1.34	74.87
28	ラ	171436	1.31	76.19
29	ツ	170532	1.31	77.49
30	ケ	168419	1.29	78.79
31	ユ	162424	1.24	80.03
32	ド	138091	1.06	81.09
33	サ	133047	1.02	82.11
34	セ	129544	0.99	83.10
35	ヨ	126093	0.97	84.07
36	ダ	121906	0.93	85.00
37	ジ	113120	0.87	85.87
38	チ	112144	0.86	86.73

表 A.2: モーラのユニグラム統計

No.	モーラ	出現数	[%]	累積 [%]
39	ホ	80839	0.62	87.35
40	ネ	80441	0.62	87.96
41	ハ	80350	0.62	88.58
42	ゴ	77413	0.59	89.17
43	ミ	75789	0.58	89.75
44	ロ	73431	0.56	90.31
45	ブ	66348	0.51	90.82
46	ヤ	64480	0.49	91.32
47	バ	64323	0.49	91.81
48	ヒ	63931	0.49	92.30
49	メ	63857	0.49	92.79
50	ウ	63806	0.49	93.28
51	フ	54840	0.42	93.70
52	ショ	52450	0.40	94.10
53	ジョ	41058	0.31	94.41
54	ジュ	40712	0.31	94.73
55	ズ	40346	0.31	95.04
56	ゲ	38186	0.29	95.33
57	シュ	36141	0.28	95.61
58	ム	34682	0.27	95.87
59	チョ	31202	0.24	96.11
60	キョ	29771	0.23	96.34
61	グ	29482	0.23	96.56
62	×	28672	0.22	96.78
63	パ	27249	0.21	96.99
64	ベ	25985	0.20	97.19
65	シャ	24218	0.19	97.38
66	ゼ	24131	0.18	97.56
67	ギ	20843	0.16	97.72
68	ザ	20067	0.15	97.88

表 A.2: モーラのユニグラム統計

No.	モーラ	出現数	[%]	累積 [%]
69	ビ	19497	0.15	98.02
70	ヘ	18417	0.14	98.17
71	リョ	17419	0.13	98.30
72	キュ	17290	0.13	98.43
73	ゾ	16703	0.13	98.56
74	ボ	16690	0.13	98.69
75	プ	16439	0.13	98.81
76	チュ	12923	0.10	98.91
77	ジャ	12900	0.10	99.01
78	ヒョ	12078	0.09	99.10
79	チャ	9499	0.07	99.18
80	ɸ	9176	0.07	99.25
81	ポ	8339	0.06	99.31
82	ピ	7725	0.06	99.37
83	ティ	6775	0.05	99.42
84	ギョ	6727	0.05	99.47
85	ヌ	6602	0.05	99.52
86	ニュ	6331	0.05	99.57
87	ヒャ	5511	0.04	99.62
88	ペ	5229	0.04	99.66
89	ディ	4839	0.04	99.69
90	head	3965	0.03	99.72
91	キャ	3892	0.03	99.75
92	ピョ	2770	0.02	99.77
93	ビョ	2525	0.02	99.79
94	リュ	2379	0.02	99.81
95	ジェ	2363	0.02	99.83
96	ファ	2192	0.02	99.85
97	ギャ	1902	0.01	99.86
98	フィ	1900	0.01	99.88

表 A.2: モーラのユニグラム統計

No.	モーラ	出現数	[%]	累積 [%]
99	トゥ	1788	0.01	99.89
100	××	1406	0.01	99.90
101	フォ	1370	0.01	99.91
102	ウォ	1066	0.01	99.92
103	リャ	990	0.01	99.93
104	ウェ	804	0.01	99.93
105	フェ	714	0.01	99.94
106	ピュ	704	0.01	99.94
107	チェ	631	0.00	99.95
108	ミュ	566	0.00	99.95
109	ビャ	540	0.00	99.96
110	ミャ	533	0.00	99.96
111	ピャ	495	0.00	99.96
112	ギュ	469	0.00	99.97
113	スィ	452	0.00	99.97
114	ミョ	432	0.00	99.97
115	テュ	413	0.00	99.98
116	ニョ	411	0.00	99.98
117	ドゥ	385	0.00	99.98
118	ビュ	363	0.00	99.99
119	ニャ	234	0.00	99.99
120	ウィ	234	0.00	99.99
121	ツォ	225	0.00	99.99
122	シェ	224	0.00	99.99
123	ヒュ	212	0.00	100.00
124	デュ	173	0.00	100.00
125	ツァ	86	0.00	100.00
126	ツェ	73	0.00	100.00
127	ズィ	72	0.00	100.00
128	ヴィ	57	0.00	100.00

表 A.2: モーラのユニグラム統計

No.	モーラ	出現数	[%]	累積 [%]
129	ツィ	30	0.00	100.00
130	イェ	27	0.00	100.00
131	ニェ	27	0.00	100.00
132	フュ	21	0.00	100.00
133	クヮ	11	0.00	100.00
134	グヮ	11	0.00	100.00
135	ヴァ	7	0.00	100.00
136	ヴェ	5	0.00	100.00
137	ミェ	3	0.00	100.00
138	ヴォ	3	0.00	100.00
139	ヴ	3	0.00	100.00
140	ヒェ	2	0.00	100.00

付録 B

案内文リスト

- 富良野：
  - 【フィラーなし】札幌から富良野 (ふらの) へ行くには，JR 特急で 2 時間です.
  - 【フィラーあり】 札幌から富良野 (ふらの) へ行くには，えっと，JR 特急で 2 時間です.
- 旭山動物園：
  - 【フィラーなし】 札幌から旭山 (あさひやま) 動物園へ行くには，札幌駅から JR 特急で旭川 (あさひかわ) 駅まで 1 時間 30 分，旭川駅からバスに乗って旭山動物園まで 40 分です.
  - 【フィラーあり】 札幌から旭山 (あさひやま) 動物園へ行くには，札幌駅から JR 特急で旭川 (あさひかわ) 駅まで 1 時間 30 分，えー，旭川駅からバスに乗って旭山動物園まで 40 分です.
- 夕張：
  - 【フィラーなし】 札幌から夕張 (ゆうばり) へ行くには，札幌駅から JR 特急「スーパーとかち」で新夕張まで 1 時間，新夕張で乗り換えて夕張まで 20 分です.
  - 【フィラーあり】 札幌から夕張 (ゆうばり) へ行くには，札幌駅から JR 特急「スーパーとかち」で新夕張まで 1 時間，それで，新夕張で乗り換えて夕張まで 20 分です.
- 美瑛：
  - 【フィラーなし】 札幌から美瑛 (びえい) へ行くには，札幌駅からスーパーカムイに乗って旭川 (あさひかわ) 駅まで 1 時間 30 分，旭川駅から直行バスで美瑛駅まで 50 分です.
  - 【フィラーあり】 札幌から美瑛 (びえい) へ行くには，えー，札幌駅からスーパーカムイに乗って旭川 (あさひかわ) 駅まで 1 時間 30 分，あと，旭川駅から

直行バスで美瑛駅まで 50 分です.

- 登別温泉：
  - 【フィラーなし】 札幌から登別 (のぼりべつ) 温泉へ行くには，札幌駅から特急列車で登別駅まで 1 時間 10 分，登別駅で道南 (どうなん) バスの登別温泉行きに乗れば 15 分で行けます.
  - 【フィラーあり】 札幌から登別 (のぼりべつ) 温泉へ行くには，札幌駅から特急列車で登別駅まで 1 時間 10 分，で，登別駅で道南 (どうなん) バスの登別温泉行きに乗れば 15 分で行けます.
- 洞爺湖：
  - 【フィラーなし】 札幌から洞爺湖 (とうやこ) へ行くには，札幌駅から特急「スーパー北斗」で洞爺駅まで 1 時間 50 分，洞爺駅から道南 (どうなん) バス・洞爺湖温泉行きで洞爺湖温泉まで 20 分です.
  - 【フィラーあり】 札幌から洞爺湖 (とうやこ) へ行くには，札幌駅から特急「スーパー北斗」で洞爺駅まで 1 時間 50 分，えっと，洞爺駅から道南 (どうなん) バス・洞爺湖温泉行きで洞爺湖温泉まで 20 分です.
- 小樽運河：
  - 【フィラーなし】 札幌から小樽 (おたる) 運河へ行くには，札幌駅から快速列車「いしかりライナー」で小樽駅まで 50 分，小樽駅から小樽散策バスに乗れば小樽運河ターミナルまで 6 分です.
  - 【フィラーあり】 札幌から小樽 (おたる) 運河へ行くには，えっと，札幌駅から快速列車「いしかりライナー」で小樽駅まで 50 分，で，小樽駅から小樽散策バスに乗れば小樽運河ターミナルまで 6 分です.
- 函館：
  - 【フィラーなし】 札幌から函館 (はこだて) へ行くには，札幌駅から北都 (ほくと) 交通バスに乗って丘珠 (おかだま) 空港まで 30 分，丘珠空港から飛行機で函館 (はこだて) 空港まで飛んで 40 分，函館空港から空港連絡バスに乗り換えると函館駅まで 20 分で行けます.
  - 【フィラーあり】 札幌から函館 (はこだて) へ行くには，札幌駅から北都 (ほくと) 交通バスに乗って丘珠 (おかだま) 空港まで 30 分，それで，丘珠空港から飛行機で函館 (はこだて) 空港まで飛んで 40 分，えっと，函館空港から空港連絡バスに乗り換えると函館駅まで 20 分で行けます.
- 網走：
  - 【フィラーなし】 札幌から網走 (あばしり) へ行くには，札幌から北都 (ほくと) 交通バスで丘珠 (おかだま) 空港まで 30 分，丘珠空港から女満別 (めまんべつ) 空港まで飛行機で 50 分，女満別空港から連絡バスに乗れば網走 (あばし

り) バスターミナルまで 30 分です.

- 【フィラーあり】 札幌から網走 (あばしり) へ行くには，あのー，札幌から北都 (ほくと) 交通バスで丘珠 (おかだま) 空港まで 30 分，それで，丘珠空港から女満別 (めまんべつ) 空港まで飛行機で 50 分，えー，女満別空港から連絡バスに乗れば網走 (あばしり) バスターミナルまで 30 分です.

- 襟裳岬：
  - 【フィラーなし】 札幌から襟裳 (えりも) 岬へ行くには，札幌駅から快速・普通列車を乗り継いで約 1 時間の苫小牧 (とまこまい) 駅から日高 (ひだか) 本線に乗り換えて様似 (さまに) 駅まで 3 時間 10 分，様似から JR バス襟裳岬・広尾 (ひろお) 行きに乗って 1 時間で襟裳岬に着きます.
  - 【フィラーあり】 札幌から襟裳 (えりも) 岬へ行くには，札幌駅から快速・普通列車を乗り継いで約 1 時間の苫小牧 (とまこまい) 駅から，あのー，日高 (ひだか) 本線に乗り換えて様似 (さまに) 駅まで 3 時間 10 分，それで，様似から JR バス襟裳岬・広尾 (ひろお) 行きに乗って 1 時間で襟裳岬に着きます.
- 釧路湿原：
  - 【フィラーなし】 札幌から釧路 (くしろ) 湿原へ行くには，札幌駅から JR 特急「スーパーおおぞら」で南千歳 (みなみちとせ) 駅を経由して釧路駅まで 3 時間 46 分，釧路駅から JR くしろ湿原ノロッコ号で釧路湿原駅まで 23 分です.
  - 【フィラーあり】 札幌から釧路 (くしろ) 湿原へ行くには，あのー，札幌駅から JR 特急「スーパーおおぞら」で南千歳 (みなみちとせ) 駅を経由して釧路駅まで 3 時間 46 分，それで，釧路駅から JR くしろ湿原ノロッコ号で釧路湿原駅まで 23 分です.
- 五稜郭：
  - 【フィラーなし】 札幌から五稜郭 (ごりょうかく) へ行くには，札幌駅から北都 (ほくと) 交通バスに乗って丘珠 (おかだま) 空港まで 30 分，丘珠空港から飛行機で函館 (はこだて) 空港まで 40 分，函館空港から空港連絡バスに乗ると函館駅まで 20 分，函館駅から函館市電・湯 (ゆ) の川 (かわ) 行きで五稜郭公園前まで 15 分です.
  - 【フィラーあり】 札幌から五稜郭 (ごりょうかく) へ行くには，えー，札幌駅から北都 (ほくと) 交通バスに乗って丘珠 (おかだま) 空港まで 30 分，それで，丘珠空港から飛行機で函館 (はこだて) 空港まで 40 分，えっと，函館空港から空港連絡バスに乗ると函館駅まで 20 分，えー，函館駅から函館市電・湯 (ゆ) の川 (かわ) 行きで五稜郭公園前まで 15 分です.
- 知床五湖：
  - 【フィラーなし】 札幌から知床 (しれとこ) 五湖 (ごこ) へ行くには，札幌バス

ターミナルから高速バス「イーグルライナー」で斜里 (しゃり) バスターミナルまで 5 時間 50 分，斜里バスターミナルからバスで知床自然センターまで 1 時間，知床自然センターからバスで知床五湖まで 15 分です.

- 【フィラーあり】 札幌から知床 (しれとこ) 五湖 (ごこ) へ行くには，札幌バスターミナルから高速バス「イーグルライナー」で斜里 (しゃり) バスターミナルまで 5 時間 50 分，それで，斜里バスターミナルからバスで知床自然センターまで 1 時間，えっと，知床自然センターからバスで知床五湖まで 15 分です.

- 松前：
  - 【フィラーなし】 札幌から松前 (まつまえ) へ行くには，札幌駅から北都 (ほくと) 交通バスで丘珠 (おかだま) 空港まで 30 分，丘珠空港から飛行機で函館 (はこだて) 空港まで 40 分，函館空港から空港連絡バスで函館駅まで 20 分，函館駅から津軽 (つがる) 海峡線の特急で木古内 (きこない) 駅まで 38 分，木古内駅から函館バス・松前出張所行きで松前まで 1 時間 30 分です.
  - 【フィラーあり】 札幌から松前 (まつまえ) へ行くには，えー，札幌駅から北都 (ほくと) 交通バスで丘珠 (おかだま) 空港まで 30 分，丘珠空港から飛行機で函館 (はこだて) 空港まで 40 分，それで，函館空港から空港連絡バスで函館駅まで 20 分，函館駅から津軽 (つがる) 海峡線の特急で木古内 (きこない) 駅まで 38 分，えー，木古内駅から函館バス・松前出張所行きで松前まで 1 時間 30 分です.

- 宗谷岬：
  - 【フィラーなし】 札幌から宗谷 (そうや) 岬へ行くには，札幌から北都交通バスで丘珠 (おかだま) 空港まで 30 分，丘珠空港から飛行機に乗り換えて稚内 (わっかない) 空港まで 50 分，稚内空港から連絡バスで稚内駅前ターミナルまで 30 分，稚内駅前ターミナルから宗谷バス・宗谷岬行きに乗って 50 分で行けます.
  - 【フィラーあり】 札幌から宗谷 (そうや) 岬へ行くには，えっと，札幌から北都交通バスで丘珠 (おかだま) 空港まで 30 分，で，丘珠空港から飛行機に乗り換えて稚内 (わっかない) 空港まで 50 分，それで，稚内空港から連絡バスで稚内駅前ターミナルまで 30 分，えっと，稚内駅前ターミナルから宗谷バス・宗谷岬行きに乗って 50 分で行けます.