情報処理学会論文誌 Vol.53 No.2 889–900 (Feb. 2012)

ポーズを考慮した話し言葉言語モデルの構築

太田 健吾[1,a)] 土屋 雅稔[2] 中川 聖一[1]

受付日 2011年1月21日, 採録日 2011年11月7日

概要：本論文では，話し言葉の音声認識タスクで問題となる，入力音声中のポーズと言語モデルの学習コーパス中の句読点との不一致に対処する方法を提案する．自発的に発声される話し言葉の音声では，言語的な区切りとは無関係な位置に多数のポーズが出現するため，言語モデルを学習する際に，コーパス中の句読点をポーズと見なして学習を行うことは適切でない．この問題に対処する最も簡単な方法は，ポーズの情報を含むコーパスから，実際のポーズを反映した言語モデルを構築することであるが，そのようなコーパスが利用できるドメインはきわめて稀である．そこで本論文では，ポーズ情報の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，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ポーズの情報を補うことにより，ポーズを考慮した言語モデルを作成する．提案手法によって構築された言語モデルを国会審議の認識実験によって評価したところ，従来の句読点に基づく認識処理単位を用いた言語モデルと比較して，認識精度を改善することができた．

キーワード：音声認識，話し言葉，書き言葉，言語モデル，ポーズ挿入

Construction of Spoken Language Model Considering Pause

Kengo Ohta[1,a)] Masatoshi Tsuchiya[2] Seiichi Nakagawa[1]

Received: January 21, 2011, Accepted: November 7, 2011

Abstract: This paper addresses the mismatch between pauses in input speech and punctuations in training corpora of language model. In a spontaneous speech recognition task, it is inadequate to train a language model with regarding punctuations as pauses, because there is an inevitable gap between pauses in input speech and punctuations in corpora. The simplest approach to address this problem is to build a language model that considers pauses from a corpus that includes pause information. However, such corpora can only be available in a limited domain. In this paper, we propose a method to build a language model that considers pauses from a corpus that does not include pause information. In our method, a pause insertion model is trained from spontaneous speech corpora, and then the language model that considers pauses is built by using this model. Our proposed model achieved an improvement over the conventional model in the recognition task of committee meetings of Japanese National Diet.

Keywords: speech recognition, spoken language, written language, language model, pause insertion

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など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 [18] 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）[11] とは異なった位置に出現する．また，中川ら [16] 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

[1] 豊橋技術科学大学情報・知能工学系
Department of Computer Sciences and Engineering, Toyohashi University of Technology, Toyohashi, Aichi 441–8580, Japan
[2] 豊橋技術科学大学情報メディア基盤センター
Information Media Center, Toyohashi University of Technology, Toyohashi, Aichi 441–8580, Japan
[a)] kohta@slp.cs.tut.ac.jp

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 [21] 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，1,000 msec 以上のポーズを認識処理単位を区切るロングポーズ，1,000 msec 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 [9] は，30 msec 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できるドメインはきわめて稀である．

そこで，本論文では，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）[1] 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 [9] 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 [24] 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論文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 [12] 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とき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（4 章参照）．

[*1] たとえば，西村ら [9] は，言い直しに起因する語断片やフィラーなどの不用語を透過単語として扱っている．

以下，本論文の構成を述べる．まず，2 章では，ポーズと句読点の不一致の問題について述べる．次に，3 章では，CRF に基づくポーズの挿入モデルを提案し，この挿入モデルを用いて，ポーズを積極的に考慮した言語モデルを構築する手法について述べる．続いて，4 章では，3 章で述べた提案手法を，国会審議を対象とした認識実験によって評価する．最後に，5 章では，本研究のまとめと今後の課題について述べる．

2. ポーズと句読点の不一致の問題

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）[2] 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 1** に示す．CSJ の転記基本単位は，200 msec 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．CSJ の節単位は，節境界と呼ばれる，発話の統語的・意味的な境界によって区切られる単位である．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 [11]．

- 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．

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

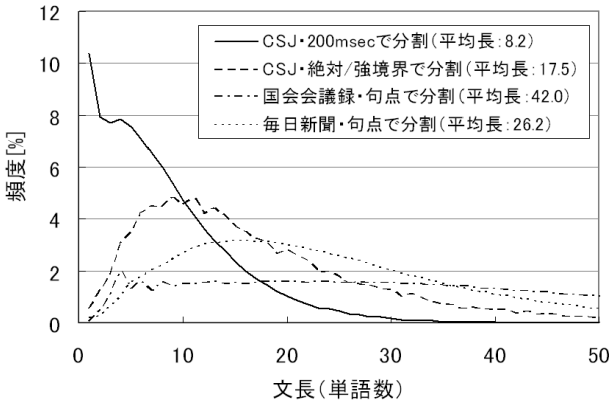


図 1 認識処理単位と文単位の長さの比較

Fig. 1 Comparison between inter-pausal units and sentence units.

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

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 国会会議録における句読点とポーズの例（<p> は 200 msec 以上のポーズ）
Fig. 3 Example of commas in National Diet Records (<p> represents a pause that is longer than 200 msec).

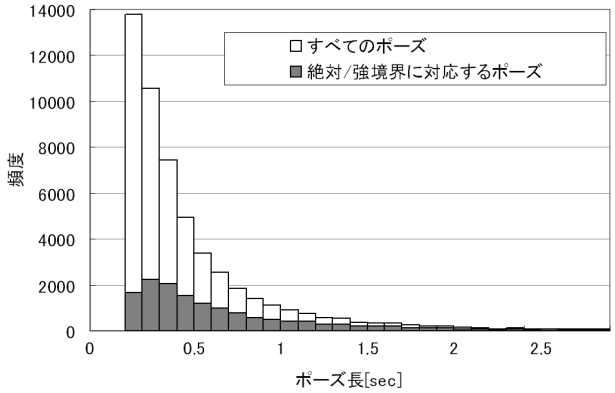


図 2 CSJ に出現するポーズの長さの分布
Fig. 2 Pause length distribution in CSJ.

点が挿入されている．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 1 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よく似ている．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く，しかも実際のポーズ位置と異なる [6], [7], [22]．

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（200 msec 以上）の分布を**図 2** に示す．図 2 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（200 msec 以上）と句読点の位置を比較した結果を**図 3** に示す．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．ポーズは 452 カ所，読点は 357 カ所，句点は 88 カ所に出現したが，ポーズのうち，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 81.8%がポーズと一致していた．よって，読点のみでも，句読点の組合せでも，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．

そこで，本論文では，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する．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すでに含まれている．そこで，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する．なお，以下の議論では，言語的まとまり以外の要因に基づくポーズを，ショートポーズと呼ぶ．

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

本章では，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）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．

3.1 ショートポーズ挿入モデルの定式化

本論文では，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 [20]．具体的には，**図 4** 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．

本論文では，このような問題を解くショートポーズ挿入モデルを，CRF [1] を用いて構築する．CRF は，隠れマル

[*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．

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	0	0	SP	0	0	0 ···

図 4 ショートポーズ挿入モデルの学習用ラベル
Fig. 4 Example of short pause insertion labelling.

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．

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

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++[*4] 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} の組合せに加え，x_i の読みに対応するモーラ列 m_i のうちの終端 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 5** 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．

3.2 ショートポーズ挿入モデルに基づく言語モデルの構築

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 [14] 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．

$$f(w) = \sum_i \delta(x_i = w) \quad (2)$$

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 の値をとる．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> の 1-gram 頻度 $f(<sp>)$ を，次式のように定義する．

$$f(<sp>) = \sum_i P(y_i = SP|x_1^L) \quad (3)$$

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．

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

i	形態素（x） 表層形	品詞	モーラ（m）	ラベル（y）
1	それ	代名詞	ソ, レ	0
2	で	助詞	デ	0
3	ハワイ	名詞	ハ, ワ, イ	0
4	と	助詞	ト	0
5	いう	動詞	イ, ウ	0
6	の	助詞	ノ	0
7	は	助詞	ハ	SP
8	火山	名詞	カ, ザ, ン	0
9	の	助詞	ノ	0
10	噴火	名詞	フ, ン, カ	0
11	で	助詞	デ	0
12	だんだん	副詞	ダ, ン, ダ, ン	0
13	でき	動詞	デ, キ	0
14	てっ	助動詞	テ, ッ	0
15	た	助動詞	タ	0
16	島	名詞	シ, マ	0
17	が	助詞	ガ	SP
18	こう	副詞	コ, ウ	0

図 5 学習データの例
Fig. 5 An example of training data.

ショートポーズ挿入モデルとして CRF を用いる場合は，式 (1) を用いて次式のように求める．

$$P(y_i = SP|x_1^L) = \sum_{\{y_1^L|y_i=SP\}} P(y_1^L|x_1^L) \quad (4)$$

式 (2) と式 (3) を用いると，0-gram 頻度総数 $f(\cdot)$ は，次式のように定義できる．

$$f(\cdot) = f(<sp>) + \sum_w f(w) \quad (5)$$

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．

$$P(w) = \frac{f(w)}{f(\cdot)} \quad (6)$$

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すべての事象の頻度であり，次式のように定義する．

$$f(w, w') = \sum_i \delta(x_{i-1} = w)\delta(x_i = w') \times (1 - P(y_{i-1} = SP|x_1^L)) \quad (7)$$

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．

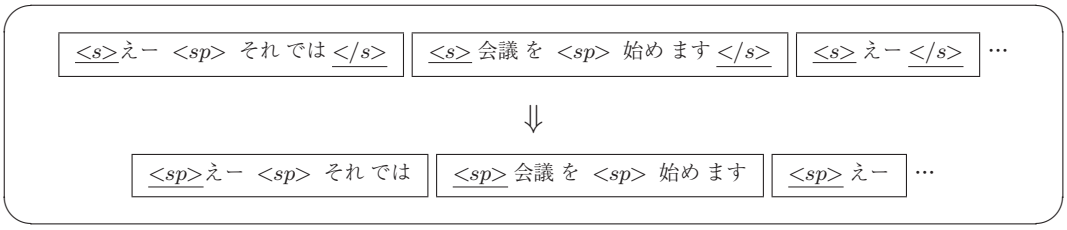


図 6 ショートポーズの認定
Fig. 6 Example of inserting <*sp*> into an Inter-Pausal Unit boundary.

$$f(w, <sp>) = \sum_i \delta(x_i = w) P(y_i = SP | x_1^L) \tag{8}$$

$$f(<sp>, w) = \sum_i \delta(x_i = w) P(y_{i-1} = SP | x_1^L) \tag{9}$$

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．

$$f(w, w', w'') = \sum_i \delta(x_{i-2} = w)\delta(x_{i-1} = w') \times \delta(x_i = w'')(1 - P_{i-2})(1 - P_{i-1}) \tag{10}$$

$$f(w, w', <sp>) = \sum_i \delta(x_{i-1} = w) \times \delta(x_i = w')(1 - P_{i-1})P_i \tag{11}$$

$$f(w, <sp>, w') = \sum_i \delta(x_{i-1} = w)\delta(x_i = w')P_{i-1} \tag{12}$$

$$f(<sp>, w, w') = \sum_i \delta(x_{i-1} = w) \times \delta(x_i = w')P_{i-2}(1 - P_{i-1}) \tag{13}$$

$$f(<sp>, w, <sp>) = \sum_i \delta(x_i = w)P_{i-1}P_i \tag{14}$$

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．本論文では，Witten-Bell バックオフ [5] を適用して，次式のように $P(w|h)$ を求める．

$$P(w|h) = \begin{cases} \frac{f(h,w)}{f(h)+r(h)} & \text{if } f(h,w) > c \\ \frac{r(h)}{f(h)+r(h)} \cdot P(w|h') & \text{otherwise} \end{cases} \tag{15}$$

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，本論文では，$r(h)$ 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 (4) 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．

$$P(y_i = SP|x_1^L) \simeq P_{trigram}(SP|x_{i-2}, x_{i-1}) \tag{16}$$

この式を，式 (5)〜式 (14) 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

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，処理単位ごと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** に示す．単語 3-gram 言語モデルを使う場合，図 6 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 [21] は，1,000 msec 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，1,000 msec 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，すべ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 2 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，本論文では，すべ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 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 200 msec 以上継続したことを手がかりとして決定する．

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．

テストセットパープレキシティ（PP）は，テストコーパスの単語列 w_1^n に対して，次式のように定義される．

$$PP = P(w_1^n)^{-\frac{1}{n}} \tag{17}$$

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 (19) によって定義された $P'(w|h)$ を用いて正規化を行う．

$$P'(w|h) = \begin{cases} 0 & \text{if } w \in ccs \\ \alpha(h) \cdot P(w|h) & \text{otherwise} \end{cases}, \tag{18}$$

$$\alpha(h) = \frac{1}{1 - \sum_{w \in ccs} P(w|h)}, \tag{19}$$

$$ccs = \{<s>, </s>, <sp>\} \tag{20}$$

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．

また，w_i が未知語の場合には，唯一の未知語クラス UNK を導入して，$P(w_i|w_{i-2}, w_{i-1}) = P(UNK|w_{i-2}, w_{i-1})$ とする．ただし，この定義では，言語モデルの語彙サイズが小さくなると，未知語率が増加し，テストセットパープレキシティは小さくなるという矛盾が生じる．そこで，この問題を考慮した補正テストセットパープレキシティ $PP*$ が提案されている [17]．補正テストセットパープレキシティでは，未知語クラス UNK 内の各未知語の生起確率は一様に分布すると仮定し，未知語の異なり数が m のとき，未知語 w_i の生起確率を $\frac{P(UNK)}{m}$ として計算する．

単語正解率（$Cor.$）および単語認識精度（$Acc.$）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．

$$Cor. = \frac{H}{H + D + S} \tag{21}$$

$$Acc. = \frac{H - I}{H + D + S} \tag{22}$$

4. 国会会議録を対象とした評価実験

本章では，3 章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．

4.1 実験条件

3 章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 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 [19] によってフィラーを自動挿入した国会会議録（表 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 (15) 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，200 msec 以上かつ 1,000 msec 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 2 より，1,000 msec 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，200 msec 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いため，利用できなかった．この影響については，4.2 節で述べる．

このほかに，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，3.4 節で述べたとおり，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）[8] 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 [13]．音響分析条件は**表 3** のとおりである．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

表 1 実験データ諸元
Table 1 Statistics of data sets.

	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	3,194 K	3,584 K	38,668 K	21 K
語彙サイズ	37 K	48 K	58 K	2 K

表 2 テストセット諸元
Table 2 Detail of test set.

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

表 3 音響分析条件
Table 3 Conditions of acoustic analysis for input speech.

サンプリング周波数	16 kHz
プリエンファシス	0.98
分析窓	Hamming 窓
分析窓長	25 ms
窓間隔（フレームシフト）	10 ms
特徴パラメータ	MFCC（12 次）+ ΔMFCC（12 次）+ ΔΔMFCC（12 次）+ Δ パワー + ΔΔ パワー（計 38 次）

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[*5]による発話（88 分）を人手で書き起こしたテキストを用意した．形態素数 · 語彙サイズを表 1 のテストセット欄に，各会議の詳細を**表 2** に示す．

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，200 msec 以上かつ 1,000 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，200 msec 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，200 msec 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 200 msec 以上のポーズと，200 msec 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，200 msec 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，200 msec 未満のポーズに対して，200 msec 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，200 msec 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．

4.3 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 (4) のように CRF を用いる場合と，式 (16) 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 1 の学習セット）から学習したモデルであり，200 msec 以上かつ 1,000 msec 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

[*5] 表 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．

表 4 認識実験によるショートポーズ挿入方法の評価
Table 4 ASR evaluation of <*sp*> insertion methods.

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	すべての読点を <*sp*> として扱う	56.4	62.3	6.5	68.0	61.8	68.4	60.6
4	すべての読点を <*sp*> として扱う ＋すべての句点を <*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 ＋すべての句点を <*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 ＋すべての句点を <*sp*> として扱う	50.9	56.2	10.9	70.4	64.4	71.3	64.1

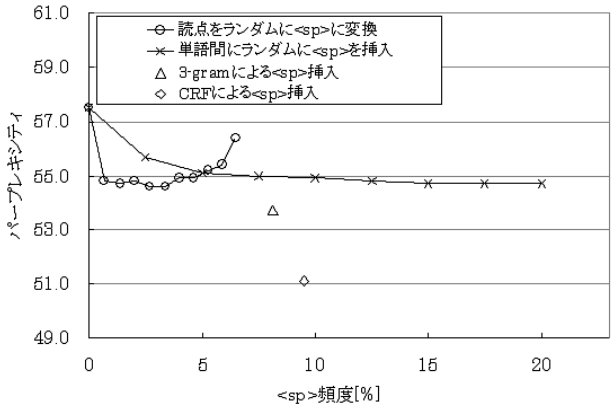


図 7 ショートポーズの挿入方法の比較
Fig. 7 Comparison of <*sp*> insertion methods.

で比較した結果を**図 7** に示す．図 7 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な値となった（$PP = 59.0$）．これは，緒方ら [12] 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．

4.4 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 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 単語のことである．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方法（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

表 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
Table 5 ASR evaluation of language models based on N-gram count marging.

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

· 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 [19] 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加えることで，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 [15] 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．

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 [24]．**表 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 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．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されている（話し言葉表現の補充によると考えられる）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いえる．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 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されているが，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 カ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 カ所だった．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いえる．

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** 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，ほと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 7** に示す．各会議の議題と内容は互いに大きく異なる（表 2）にもかかわらず，テストセットに含まれるすべての会議について，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 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見られた[*6]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，本章の実験では最大で 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

[*6] 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．

表 6 話者別の比較
Table 6 Comparison of ASR performance for speakers.

ID	発話数	発話時間（秒）	形態素数	未知語率（%）	ベースライン No.2 Cor.（%）	ベースライン No.2 Acc.（%）	ベースライン No.4 Cor.（%）	ベースライン No.4 Acc.（%）	提案手法 No.8 Cor.（%）	提案手法 No.8 Acc.（%）
S1	676	1,342.3	5,197	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	2,621	0.95	49.0	43.7	51.3	43.3	51.3	**44.9**
S6	234	949.9	3,630	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	1,112	2.16	74.8	66.6	76.8	67.1	**77.2**	**69.7**
S10	316	808.8	3,416	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	1,943	5,276.9	20,572	0.94	68.5	62.6	68.7	61.9	70.4	64.4

表 7 会議別の比較
Table 7 Comparison of ASR performance for talks.

ID	発話数	発話時間（秒）	形態素数	未知語率（%）	ベースライン No.2 Cor.（%）	ベースライン No.2 Acc.（%）	ベースライン No.4 Cor.（%）	ベースライン No.4 Acc.（%）	提案手法 No.8 Cor.（%）	提案手法 No.8 Acc.（%）
T1	360	1,041.6	4,038	0.99	54.6	48.4	56.0	47.6	**56.9**	**49.3**
T2	333	760.5	3,201	1.97	61.4	54.9	61.6	55.0	**63.3**	**56.5**
T3	845	1,872.5	7,380	0.24	69.6	62.8	68.8	61.7	**71.4**	**65.0**
T4	405	1,602.2	5,953	1.23	80.5	76.2	80.9	75.6	**82.1**	**78.1**
合計	1,943	5,276.9	20,572	0.94	68.5	62.6	68.7	61.9	70.4	64.4

秋田ら [23] 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 1 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 2 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 3 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 4 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．

4.5 フィラーとショートポーズの関係

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やすいことが知られている．たとえば，中川ら [16] 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（10 msec 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 [3] や Stouten ら [4] 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（200 msec 以上の無音区間）が現れる割合は 56.0%であり，また，ポーズ（200 msec 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，本章の実験では，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 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 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．

この検討のため，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 [19] 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** [10] 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 86.9，86.9，85.5，83.7 だった．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．

5. おわりに

本論文では，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すでに含まれている．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．ただし，本論文の手法は，述部などの言い換えには対応していない．今後，書き言葉の述部を話し言葉の言い回しに変換する方法が実現できれば，大量の書き言葉コーパスから音声認識用の話し言葉の言語モデルの構築が可能になると考えられる．

参考文献

[1] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. ICML*, pp.282–289 (2001).

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

[3] Gabrea, M. and O'Shaughnessy, D.: Detection of filled pauses in spontaneous conversational speech, *Proc. ICSLP*, pp.678–681 (2000).

[4] Stouten, F., Duchateau, J., Martens, J.-P. and Wambacq, P.: Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation, *Speech Communication*, Vol.48, No.11, pp.1590–1606 (2006).

[5] Witten, I.H. and Bell, T.C.: The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression, *IEEE Trans. Information Theory*, Vol.37, pp.1085–1094 (July 1991).

[6] Liu, Y. and Shriberg, E.: 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).

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

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

[9] 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，電子情報通信学会論文誌，Vol.J83-DII, No.11, pp.2473–2480 (2000).

[10] 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（CSJ コーパス付属文書）(2004).

[11] 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).

[12] 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).

[13] 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).

[14] 森　信介，笹田鉄郎，Graham, N.：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，Vol.2010-NL-196, pp.1–7 (2010).

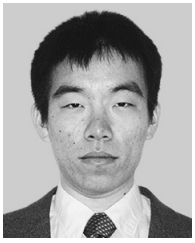
[15] 中川聖一，高木英行：パターン認識における有意差検定と音声認識システム評価法，日本音響学会誌，Vol.50, No.10, pp.849–854 (1994).

[16] 中川聖一，小林　聡：自然な音声対話における間投詞・ポーズ・い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).

[17] 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).
[18] 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，Vol.2005, No.127, pp.247–252 (2005).
[19] 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 (2008).
[20] 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).
[21] 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-D-II, No.4, pp.450–459 (2003).
[22] 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).
[23] 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).
[24] 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 (2011).

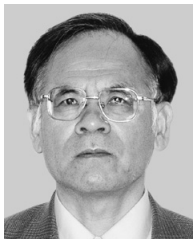
太田 健吾　（学生会員）

2007 年豊橋技術科学大学工学部卒業．2009 年同大学大学院修士課程情報工学専攻修了．現在，同大学院博士後期課程電子・情報工学専攻在学中．2011 年より日本学術振興会特別研究員（DC2）．音声言語処理に関する研究に従事．日本音響学会，電子情報通信学会，人工知能学会各学生会員．



土屋 雅稔　（正会員）

1998 年京都大学工学部卒業．2004 年同大学大学院情報学研究科知能情報学専攻博士課程単位認定退学．博士（情報学）．2004 年豊橋技術科学大学情報処理センター助手．2007 年より同大学情報メディア基盤センター助教．自然言語処理に関する研究に従事．言語処理学会会員．



中川 聖一　（フェロー）

1976 年京都大学大学院博士課程修了．同年京都大学工学部情報工学科助手．1980 年豊橋技術科学大学情報工学系講師．1990 年教授．1985～1986 年カーネギメロン大学客員研究員．音声情報処理，自然言語処理，人工知能の研究に従事．工学博士．1977 年電子通信学会論文賞，1988 年 IETE 最優秀論文賞，2001 年電子情報通信学会論文賞，各受賞．電子情報通信学会フェロー．情報処理学会フェロー．著書『確率モデルによる音声認識』（電子情報通信学会編），『音声聴覚と神経回路網モデル』（共著，オーム社），『情報理論の基礎と応用』（近代科学社），『パターン情報処理』（丸善），『Spoken Language Systems』（編著，IOS Press）等．