情報処理学会論文誌 Vol. 50 No. 2 477–487 (Feb. 2009)

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

太田健吾[†1] 土屋雅稔[†2] 中川聖一[†1]

本論文では，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って，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．

Construction of Spoken Language Model Based on Filler Prediction Model

Kengo Ohta,[†1] Masatoshi Tsuchiya[†2] and Seiichi Nakagawa[†1]

We have been proposing a method to build a language model that includes fillers from a corpus without fillers using a filler prediction model. In our method, a filler prediction model is trained from a corpus that does not cover domain-relevant topics; it restores fillers in the inexact transcribed corpora in the target domain, and then the language model that includes fillers is built from the corpora. The results of evaluation on the Corpus of Spontaneous Japanese showed that language models constructed by the proposed method achieve quite near performance of the traditional 3-gram language model constructed from the exact spontaneous speech corpus including fillers. Additionally, the results of evaluation on the Japanese National Diet Record showed that our method achieves higher recognition performance than conventional ones.

1. はじめに

近年，講義音声の自動要約やニュース音声のインデキシングなどの技術に対する需要が高まってきている[9),12)]．これらを実現するには，対象となる音声の発話スタイルとドメインが一致し，かつ，フィラーなどの話し言葉特有の現象にも対応した言語モデルを備えた大語彙音声認識器が必要である．そのような言語モデルを構築する最も単純な方法は，対象とする音声と同一ドメインの大規模な話し言葉コーパスから，言語モデルを構築するという方法である．しかし，そのようなコーパスを整備する作業はきわめて高コストであり，あらゆるドメインに対して，条件を満たすコーパスを入手できると仮定することは非現実的である．

このような状況に対処するため，対象とする音声とは異なるドメインの話し言葉言語モデルと，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，Hain ら[4)] は，会議音声の認識のために，Switchboard コーパス[3)] や Fisher コーパス[2)]（電話での対話），HUB4 コーパス（放送ニュース），Web コーパス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いる．また，Park ら[1)] 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．一方で，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[13)] 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[16)] 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．

これらに対し，我々は，書き言葉コーパスと話し言葉コーパスの中間的なコーパスとして，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が省略されている**不正確な話し言葉コーパス**に注目する．このようなコーパスは，議事録や速記録の形で広く作成されており，話し言葉特有の現象も正確に書き起されている話し言葉コーパスに比べて，比較的容

†1 豊橋技術科学大学情報工学系
Department of Information and Computer Sciences, Toyohashi University of Technology
†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

易に入手可能である．たとえば，国立国会図書館は，1947 年以降のすべての国会の会議録を公開している[⋆1]．このようなコーパスは，書き言葉コーパスよりも話し言葉に近いコーパスと考えられるので，話し言葉特有の現象に対応した言語モデルを作成するという目的には，書き言葉コーパスよりも適していると期待される．ただし，不正確な話し言葉コーパスは，話し言葉特有の現象のほとんどが省略されているため，言語モデルの学習を行う前にそれらを復元する必要がある．

本論文では，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[8]，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不正確な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する．

2. フィラー予測モデル

2.1 フィラー予測モデルの定式化

フィラーを含まないコーパスから，フィラーに対応した言語モデルの作成方法を考える．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[16] 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，いくつかの欠点がある．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．

⋆1 http://kokkai.ndl.go.jp/

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あげられる．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[16] 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．したがって，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[16] 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よいので，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いえる．

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）[5] を適用する．

本論文ではこれ以降，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．実際の正確な話し言葉コーパス[6] を対象とする分析から，フィラーには多様な派生形が存在することが分かっており，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 つのモデルの組合せとしてフィラー予測モデルを定式化する．

2.2 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ときに，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．本論文では，このモデルを，形態素列

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 1　フィラー挿入モデルの学習用ラベル
Fig. 1　Training labels for filler insertion model.

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．たとえば，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 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%程度ときわ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[10]）．したがって，今回はそのような状況は扱わない．

本論文では，このような問題を解くフィラー挿入モデルを，CRF 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良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．

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

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，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．

2.3　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ときに，挿入すべき適当なフィラーを選択するモデルである．本論文では，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[7]，次式のように推定する．

$$P_s(f|h) = \begin{cases} \frac{c(h,f)}{c(h)+r(h,f)} & \text{if } c(h,f) > 0 \\ \frac{r(h,f)}{c(h)+r(h,f)} \cdot P_s(f|h') & \text{otherwise} \end{cases} \tag{2}$$

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）．

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

本論文では，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，さらに以下の 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．

本章では，この処理の詳細について述べる．

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うえで，フィラーを取り除く．たとえば，文 (2) を学習コーパス中の文とすると，図 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

⋆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 2　学習データの例
Fig. 2　An example of training data.

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 モーラを素性として用いる．たとえば，**図 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．本論文では，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 (2) に基づいて求められる．ただし，フィラーは，発音上の揺れによる派生形が生じやすい．たとえば，日本語話し言葉コーパス（以下，**CSJ** と略記する）[6] には 151 種類のフィラーが出現しているが，これらの多くは，長音・促音の有無や語尾音節の繰返しなどの発音上の揺れによる派生形である．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，今回は，発音が類似しているフィラーは同一のものと見なした．これにより，151 種類のフィラーを 58 のグループにまとめた．

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．

$$P(y_i = \mathtt{F}|X) = \sum_{\{Y|y_i=\mathtt{F}\}} P(Y|X). \tag{3}$$

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

$$\sum_{j=1}^{k-1} P_s(f_j|h_i) \le Q'_i < \sum_{j=1}^{k} P_s(f_j|h_i) \tag{4}$$

ただし，Q'_i は一様でランダムな確率変数（$0 \le Q'_i \le 1$），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.　日本語話し言葉コーパスを対象とする実験

本章では，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．CSJ は，話し言葉特有の現象を含めて正確に書き起された話し言葉コーパスである．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．

表 1　学会講演と模擬講演の比較
Table 1　Comparison between APS and SPS.
（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%

表 2　実験データ諸元
Table 2　Data sets.

	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	1,715	937	50
収録時間（hour）	329.9	258.4	16.0
総文数	498 k	363 k	22 k
総単語数	3,606 k	3,109 k	170 k
語彙サイズ	41 k	29 k	8 k
フィラー発生頻度	175 k	174 k	11 k
フィラー発生率	4.8%	5.6%	6.7%

4.1　実 験 条 件

CSJ は，学会講演・模擬講演・対話・朗読という 4 種類の部分コーパスに分けることができる．このうち，模擬講演の一部（1,665 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 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．

そこで，本章の実験では，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 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．

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．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率を考慮した尺度であり，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，次式によって定義される[11]．

$$\log_2 PP^* = \log_2 PP + \frac{o}{n}\log_2 m \tag{5}$$

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．

$$H_F = -\frac{1}{n_F}\log \prod_{w_i \in F} P(w_i|w_{i-2}w_{i-1}) \tag{6}$$

$$PP_F = 2^{H_F} \tag{7}$$

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．

$$H_O = -\frac{1}{n_O}\log \prod_{w_i \in O} P(w_i|w_{i-2}w_{i-1}) \tag{8}$$

$$PP_O = 2^{H_O} \tag{9}$$

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，すべてのフィラーを同一視した実験を行った．結果を**表 3** に示す．表 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，実際の話し言葉にきわめて近い言語モデルを再現できるといえる．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いえる．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 3　フィラー挿入モデルの性能比較
Table 3　Performance comparison among filler insertion models.

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			品詞						
	名詞	動詞/形容詞	その他							
CRF	○	○	○	○	○	152,614	60.5	68.3	13.7	67.7
	×	○	○	○	○	151,269	60.7	68.5	14.0	67.8
	×	×	○	○	○	153,722	60.9	68.7	14.0	68.0
形態素トライグラム						134,234	62.9	70.7	17.1	69.3
品詞トライグラム						155,463	63.5	71.7	16.3	70.4
ユニグラム						148,452	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						175,253	59.5	67.1	10.9	67.6

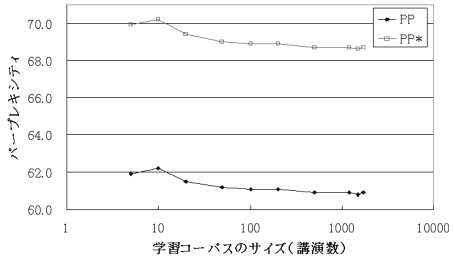


図 3　フィラー挿入モデルの学習曲線
Fig. 3　Training curve of the filler insertion model.

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 3** に示す．図 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ことが分かる．

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．

4.3　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，

表 4　フィラー予測モデルの性能比較
Table 4　Performance comparison among filler prediction models.

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	153,722	70.6	79.6
	モーラトライグラム	153,722	70.7	79.8
	品詞トライグラム	153,722	70.5	79.6
	ユニグラム	153,722	71.7	81.0
形態素トライグラム	形態素トライグラム	134,234	72.7	81.8
	モーラトライグラム	134,234	72.8	82.0
	品詞トライグラム	134,234	72.6	81.7
	ユニグラム	134,234	73.8	83.1
品詞トライグラム	品詞トライグラム	155,463	73.2	82.7
ユニグラム	ユニグラム	148,452	79.7	90.1
開発コーパスから言語モデル作成		175,253	67.9	76.6

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** に示す．

表 4 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，フィラー選択モデルにおいて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）を利用する方法についても検討したが，フィラー選択精度は改善されなかった．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．

5. 国会会議録を対象とする実験

5.1 実 験 条 件

前章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章と同様に 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 1,083 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 5** に示す．

表 5 のコーパスを用いて，**表 6** 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**[14]，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，3 章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．

表 5　実験データ諸元
Table 5　Data sets.

	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	3.6 M	36 M	3.6 k
語彙サイズ	41 k	55 k	0.8 k
フィラー発生頻度	175 k	0	0.3 k
フィラー発生率	4.8%	0.0%	8.3%

表 6　比較した言語モデル
Table 6　Compared language models.

言語モデル	フィラー予測モデル		フィラー
	挿入モデル	選択モデル	
CSJ 付属	なし		含む
フィラーなし国会			含まない
フィラーなし国会＋模擬講演			含む
フィラーつき国会（CRF）	CRF	形態素トライグラム	含む
フィラーつき国会（トライグラム）	形態素トライグラム		含む
フィラーつき国会（ユニグラム）	ユニグラム		含む
フィラーつき国会（CRF）＋模擬講演	CRF	形態素トライグラム	含む

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 25,300 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 Ver4.0.1 を用い，音響モデルは，講演音声認識のため

表 7　音響分析条件
Table 7　Conditions of acoustic analysis for input speeches.

サンプリング周波数	16 kHz
プリエンファシス	0.97
分析窓	Hamming 窓
分析窓長	25 ms
窓間隔	10 ms
特徴パラメータ	MFCC（12 次）+ΔMFCC（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

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[14]．これは，合計 2,496 講演（486 時間）の学会講演および模擬講演から学習した，状態数 3,000，16 混合の triphone モデルである[14]．音響分析条件は**表 7** のとおりに設定した．

単語辞書は，Mecab Ver0.96（IPA 辞書 Ver2.7.0）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．

5.2　実験結果

各言語モデルの評価結果を**表 8** 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^* が大幅に改善された．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．

表 8　各言語モデルのパープレキシティと未知語率
Table 8　Perplexity and OOV rates by the language models.

言語モデル	語彙サイズ	PP	PP^*	未知語率
CSJ 付属	25,300	−	−	−
模擬講演	20,000	114.0	226.3	12.75%
フィラーなし国会	19,979	88.7	135.3	9.88%
フィラーなし国会 + 模擬講演	20,000	96.3	113.1	3.86%
フィラーつき国会（ユニグラム）		86.2	101.2	
フィラーつき国会（トライグラム）		86.1	101.1	
フィラーつき国会（CRF）		83.2	97.7	
フィラーつき国会（CRF）+ 模擬講演		78.6	92.3	

なお，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．

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 9** 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

表 9　各言語モデルの認識性能（%）
Table 9　Recognition rates by the language models.

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	25,300	–	49.0	40.4	53.9	47.0	47.5	39.0	47.9	41.2
模擬講演	20,000	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	19,979	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	20,000	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

(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 え NHK 予算 の お 審議 は ですね え 大臣 が 今 まで おっしゃっ て き た こと そして これから え 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 え NHK 予算 の 審議 は ですね まー 大臣 が 今 まで おっしゃっ て え き た こと そして これから 大臣 が ですね ···

図 4　国会会議録に対するフィラー挿入の例
Fig. 4　An example of filler inserted corpus.

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．

5.3　フィラー挿入の精度

1 章で述べたように，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」など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，フィラーなどの話し言葉特有の現象は省略されている．**図 4** 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 のとおり，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．

フィラー挿入の精度については，**表 10** に示す．表 10 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

表 10　フィラー挿入の精度と再現率
Table 10　Precision and recall of filler insertion.

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

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舞いをすることを考慮に入れる必要がある．たとえば，3-gram による単語の予測精度は約 17%である[15]．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．

6. おわりに

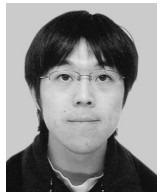
本論文では，フィラーを含む正確な話し言葉コーパスが十分に得られない状況のもとで，フィラーを考慮した言語モデルを構築するための手法として，フィラー予測モデルを用いる方法を提案した．提案手法は 2 段階からなり，最初に，正確な話し言葉コーパスからフィラー予測モデルを作成し，次に，このモデルから与えられる確率に基づいてフィラーを挿入したコーパスから言語モデルを構築した．日本語話し言葉コーパスを対象とした実験により，提案手法は，実際の正確な話し言葉コーパスから作成された言語モデルにかなり近い言語モデルを作成できることを示した．また，国会会議録を対象とした認識実験により，提案手法は，従来の手法よりも高い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

参 考 文 献

1) Park, A., Hazen, T. and Glass, J.: Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling, *Proc. ICASSP*, pp.497–500 (2005).
2) Cieri, C., Miller, D. and Walker, K.: The fisher corpus: A resource for the next generations of speech-to-text, *Proc. LREC*, pp.69–71 (2004).
3) Godfrey, J.J., Holliman, E.C. and McDaniel, J.: Switchboard: Telephone speech corpus for research and development, *Proc. ICASSP*, pp.517–520 (1992).
4) Hain, T., Dines, J., Garau, G., Karafiat, M., Moore, D., Wan, V., Ordelman, R. and Renals, S.: Transcription of conference room meetings: An investigation, *Proc. INTERSPEECH*, pp.1661–1664 (2005).
5) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. ICML*, pp.282–289 (2001).
6) Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (*SSPR2003*), pp.7–12, Tokyo, Japan (2003).
7) 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).
8) 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).
9) 岩野公司，広畑　誠，新中庸介，古井貞熙：重要文抽出による音声自動要約手法とその客観評価法についての検討（要約，検索，認識・理解・対話・一般），電子情報通信学会技術研究報告，SP2005-20, pp.1–6 (2005).
10) 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).
11) 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).
12) 藤井　敦，伊藤克亘，秋葉友良，石川徹也：音声言語データの構造化に基づく講演発表の自動要約，話し言葉の科学と工学ワークショップ講演予稿集，pp.173–177 (2001).
13) 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).
14) 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，『日本語話し言葉コーパス』付属文書（asr.pdf）.
15) 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).

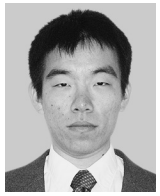
16) 秋田祐哉，河原達也：言語モデルと発音辞書の統計的話し言葉変換に基づく国会音声認識，情報処理学会研究報告，2007-SLP-69-11 (2007).

(平成 20 年 6 月 4 日受付)
(平成 20 年 11 月 5 日採録)



太田　健吾（学生会員）

2007 年豊橋技術科学大学工学部卒業．現在，同大学大学院工学研究科情報工学専攻在学．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．