音声認識過程での発話分割のための統計的言語モデル

中嶋秀治† 山本博史†

自然な話し言葉での対話においては，1回の発話（または発声）で複数の文が話されることがしばしば起こる．音声認識では，1回の発話を単位として処理が行われるが，複数の文を含んだ発話をそのまま1つの単位にして理解や翻訳や要約などの言語処理を行うことは困難であり，音声認識の後か言語処理の前に発話を文などへ分割することが必要となる．このため，本稿では通常の単語と同様に文境界としての句点を音声認識することによって複数の文が含まれる発話を各文に分割する手法を提案する．評価実験の結果，発話から文への分割性能の点では，最高で再現率94%適合率100%という性能が得られた．また，言語モデルに句点を含むか否かの違いによる句点以外の単語認識率の劣化はないという結果が得られ，本手法の有効性が確認された．

The Statistical Language Model for Utterance Splitting in Speech Recognition

Hideharu Nakajima† and Hirofumi Yamamoto†

In spontaneous dialogs, there are utterances containing several sentences. Although speech recognizers process utterances one by one, language processing such as understanding, translation or summarization needs to split utterances into sentences. This paper presents utterance splitting by recognizing periods, i.e., sentence boundaries, as well as usual words. We evaluate the performance of the model in terms of splitting and word (except for periods) accuracy. Experimental results show high recall/precision rates of splitting (the highest scores are 94%/100%) and no reduction of other word accuracy, proving the applicability of the proposed method.

1. はじめに

自然な話し言葉による対話においては，1回に，あるいはひと息に複数の文が発話（発声）されることがしばしば起こる．この様子を**図1**に示す．図1の話者Aから話者Bへと話者交替が起こるまでの区間（turn1）の中の第1番目の音声区間，すなわち，発話utt1の中の文sent1と文sent2や，話者Bから次の話者へ話者交替が起こるまでの区間（turn2）の中の第2番目の音声区間，すなわち，発話utt4の中の文sent5と文sent6がその例である．従来の音声認識では，発話（発声，図1のutt）を単位として認識処理が行われる．しかし，音声対話システム内の理解処理や，音声翻訳処理や，話し言葉での対話を要約する処理においては，複数の文を含んだ発話そのものを単位として処理することは困難であり，音声認識において，あるいは，その後の言語処理のどこかで，発話を文などに分割する処理が必要となることが指摘されている[1),2),4),7)]．たとえば，話し言葉を対象とした翻訳では処理単位は発話よりも小さく，文相当であった[8)]．つまり，翻訳単位の区切りとしての句点の位置（図1のtarget）を認識する処理が必要となる．そして，音声翻訳対話システムにおいて，待ち時間の少ない円滑な会話を実現さ

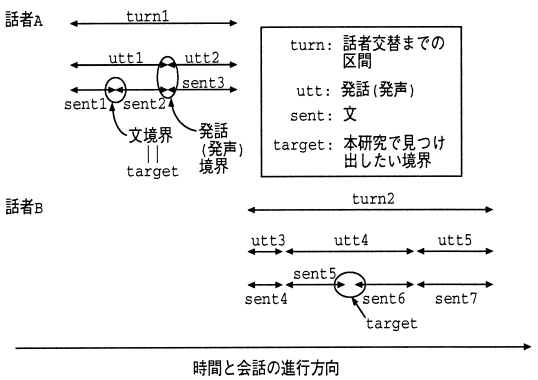

図1 対話の中の1回の発話（発声）で複数の文が話される様子
Fig. 1 A situation where several sentences are uttered in one utterance.

† ATR音声言語通信研究所
ATR Spoken Language Translation Research Laboratories

せるためには，発話の中の文の境界（図 1 の target）を，音声認識の過程で発話の終了（utt の右端）を待たずに発見し，発話を分割して，分割された時刻までの音声認識結果の単語列を翻訳処理へ送ることが必要となる．以上のような目的から，本研究では音声認識過程で発話内に含まれる文の境界（図 1 の target）を検出することを問題とする．

発話を分割する従来の手法として，隣り合う単語間での分割の有無の判定を，その単語の並びの生起頻度とその単語並びの中で分割が生じる頻度と人手で決められた閾値からなる識別関数を用いて行う方法[1),2)]，文の先頭の語であるか否かという 2 つの状態を導入した確率モデルを用いる方法[3)]，ニューラルネットを使って隣り合う単語間での分割の有無を判定する方法[4),5)]が研究されている．これらは，a) 音声認識用の言語モデルとは異なる知識として，分割のためのモデルが作成され利用されていること，b) 音声認識の後処理として分割が行われること，を特徴としている．しかし，音声の認識と分割のための統合されたモデルを使って，音声認識過程で分割を行う方法とその評価を行った研究は例があまりない．ディクテーション結果の読みやすさの向上を目的として，英語や日本語の言語モデル（単語 3 gram）に，ピリオド，カンマ，セミコロンまたは句点などを組み込んで，音声認識時にそれらを認識する研究は行われている[6),14)]．それらの研究では認識対象文は対話文ではない．文献 6) は分割の評価を行っているが対象が英語であり，文献 14) では発話の分割の点での評価は行われていない．また，句点，読点，息継ぎを同一の記号にして学習された言語モデルが研究されているが，目的が単語認識精度の向上にあり，分割の観点からの評価はない[15)]．このように，我々が認識の対象としている対話音声での分割とその評価は行われていない．

そこで本研究では，日本語の対話音声を対象とし，複数の文が含まれる発話内での文の切れ目である句点の認識を問題とする．その解決のために，音声認識用の統計的言語モデルに分割の情報を統合する．本稿では，言語モデルの構成を述べ，分割の性能と，句点以外の単語の認識性能の点で議論する．以下，2 章では，分割が必要となる発話の事例とその頻度について簡単に説明し，本手法を 3 章で説明する．4 章では評価実験について述べ，5 章で評価実験のまとめを行い，結論を 6 章で述べる．

2. 分割が必要となる発話

自然な会話では 1 回の発話の中に複数の文が含まれている場合があり，そのような発話は文への分割が必要となる[2)]．本章ではそのような発話の例を示す．たとえば，ATR から公開されている ATR 自然発話音声言語データベース[9)]を考える．これはホテルの予約やサービスの問合せに関するホテルの従業員と宿泊客との会話を想定して集められたデータベースであるが，その音声データの中には次のような発話音声がある．

表 1 発話中の句点の分布
Table 1 Distribution of period marks in utterances.

片側会話数	7,262
発話数	89,152
発話内句点数	32,258
総単語数	1,392,557

例 1：複数の文を含む発話音声

宿泊客： もしもし交通手段についてちょっと教えていただきたいんですが

ホテル： はいかしこまりましたどちらへお出かけでしょうか

この例では，ホテル側の 1 つの発話に，“はいかしこまりました．” と “どちらへお出かけでしょうか．” のように 2 つの文が含まれており，分割の必要な発話となっている．このように，発話全体が必ずしも 1 文にはなっていない．したがって，音声認識の後または翻訳などの言語処理の前に，そのような発話を文へ分割する処理が必要となる．

このデータベースの書き起こしテキストデータは，発話を単位として作成されている（たとえば，ホテル側の発話は “はい” から “しょうか” までを 1 つとして扱われている）．発話の末尾には句点が付けられている．さらに，その発話の中に複数の文があるかどうか（句点を置くべきかどうか）については，あらかじめ定められた一定の基準（話し始めの決まり文句の後や挨拶文の後以外の箇所のうち，一般的な文法において文末になる可能性のある箇所と，話し始めの決まり文句や挨拶文と同じ単語列でも決まり文句や挨拶文とは異なる意味を持つ場合にはその末尾の箇所に句点を置く，という意味まで考慮したうえでの判定基準）に基づいて専門の作業者によって判定され，句点が付与されている．翻訳でもまた句点を単位として翻訳処理が行われてきた[8)]．

ATR 自然発話音声言語データベースでの発話中の句点の頻度を**表 1** に示す．1 会話がおよそ 12 発話からなり，その 1 発話がおよそ 16 単語を含んでいる．そのような全発話のうちのおよそ 36%が発話中に句点を含んでいる．

このデータベースでは，文間の無音区間の長さは様々であり，無音区間に関する物理量（音響的特徴量）のみに基づいて文を定義し，発話を分割することは難しい[2]．

そのため本研究では，3章以降で説明と評価を行うように，音響的特徴と言語的特徴の両方を利用してもっともらしい分割が行えるよう，音声認識を利用する．

本研究では，モデルの学習および評価用のデータとして，会話の書き起しデータである ATR 自然発話音声言語データベースを利用する．そして，この中の句点を翻訳単位の境界と定義し，この句点で区切られている単位を翻訳単位，すなわち，文と定義する．このため文は，発話の終端または句点までの単語列となる．本稿では発話を上記のような文へ分割する問題を研究する．

3. 実現手法

本研究では，音声認識の過程で発話の分割を行う．つまり，音声認識の過程で，音響モデルから得られる音響尤度と以下に説明する言語モデルから得られる言語尤度とを使って，一般の単語と同時に，分割の位置を示す句点を認識する．本研究では，音声認識と分割のための統一された言語モデルとして統計的言語モデル（単語 Ngram，多重クラス複合 Ngram）を用いる．従来の音声認識では句点の認識は考慮されていなかったため，句点を取り除いて言語モデルが作成されていた．しかし，本研究では，発話中の句点を学習データの中に残し，発話中の句点への遷移確率，および，発話中の句点からの遷移確率の推定を行うことによって，モデルの統合を行う．さらに，音声認識用の辞書には句点とその発音記号を登録する．

そして，入力された音声に対して，音声認識過程で一般の単語と同様に句点の認識を行うことによって，単語グラフ内のもっともらしい位置に句点が埋め込まれる．発話の分割は選ばれた単語グラフ内のパスの中の句点の位置で行う．単語グラフのパスに句点がなければ，分割しないものと判定する．

全体の単語グラフをリアルタイム処理で文ごとの単語グラフに分割し，早期に後段の処理に送ることが可能となるが，その実装手法は通常の探索の問題であり，ここでは議論しない．

4. 評価実験

音声認識過程で本統計的言語モデルを使うことによって，発話の中の文境界位置での句点の認識と句点以外の単語の認識とが正確に行えるかどうかを確かめるために，以下の実験を行う．

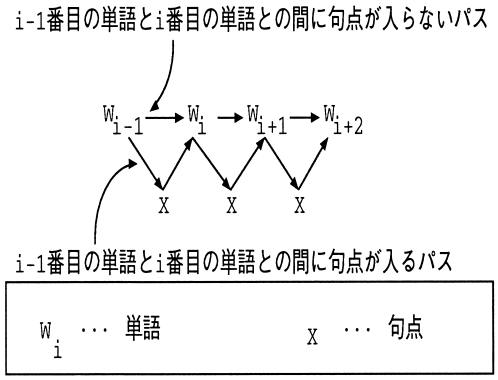

図 2 テキスト入力に対する分割実験での全単語間に句点が入る場合までを想定した単語グラフの一部分
Fig. 2 A part of word graph which includes all possible period marks between all words for the splitting experiment for text input.

4.1 実験用データ

まず最初に実験に用いるデータを説明する．本研究では ATR 自然発話音声言語データベースを用いる．本実験ではそれを以下に説明するように 4 つのサブセットに分けてモデルの学習と評価に利用した．

まず最初に，句点を含まないテキストを入力として，本言語モデルによって文への分割を行う（文に句点の付与を行う）性能を検査する．句点を含まないテキストを入力とすることにより，音声認識において句点以外の単語の系列が完全に認識された場合を模擬する．すなわち，単語の位置が既知であるが句点の位置は未知とし，すべての単語間に句点が入る場合までを想定して単語グラフを作り，その中の各単語と句点に本言語モデルによって言語尤度付けを行い，尤度の最も高いパスを選択する．その結果に含まれる句点の位置で発話の分割を行うものとする．例を**図 2** に示す．W_{i-1} と W_i との間に句点 X が入りうるか否かを “W_{i-1}, X, W_i” と “W_{i-1}, W_i” との尤度から判定する．これにより，句点以外の単語が完全に認識される場合だけではなく，句点の音響的特徴（ポーズ）が完全で音声認識の候補に含められる場合を模擬した状況下での発話の分割性能を評価する．

そのために以下で説明するデータセット ED1 と ED4 を用意した．ED1 は評価用のデータで ATR 自然発話音声言語データベースの一部である音声翻訳研究用 ATR 音声言語データベース[10]の 9 会話（通常の 2 人による会話を，話者の役割（ホテル側/客）ごとに区別してそれぞれを「片側会話」と呼ぶことにすると，18 片側会話）である．そして，音声翻訳研究用 ATR

表 2 実験で用いるデータセット
Table 2 Training and test data.

	片側会話数	発話数	発話内句点数	総単語数
ED1	18	216	73	2,437
ED2	42	551	89	4,990
ED3	7,202	88,385	32,096	1,385,130
ED4	1,218	15,891	5,585	244,049

音声言語データベース[10]の残り（609 会話（1,218 片側会話））はモデルの学習用のデータ ED4 とした．

次に，音声入力に対する発話の分割性能を評価するためにデータセット ED2 と ED3 とを用意した．ED2 は音声認識性能の評価用のデータで，提案の言語モデルに句点を含めることによって句点以外の他の単語の認識性能への影響を評価するために利用する．ED2 は ED1 とは別の 42 片側会話である．また，このデータの話者は音響モデルの学習データには含まれていない．ED3 は，前述の ED2 以外の ATR 自然発話音声言語データベースの残りのデータであり，言語モデルの学習に用いられる．音声認識用の言語モデルの学習にはできるだけ多くのデータが必要であるため，このデータセットを用意した．なお，ED4 は完全に ED3 に含まれている．

以上のデータセットのそれぞれの片側会話数，発話数（発話末の句点の数はこれに一致），発話内の句点の総数，および，総単語数を**表 2** に示す．

4.2 評価実験の設定

次に，実験で用いる言語モデル，音声認識時の探索，音響モデル，音響分析条件について説明する．

4.2.1 評価対象の言語モデルと音声認識時の探索

本実験では，句点を単語に含めた統計的言語モデルとして，句点を単語に含めた単語 3 gram と単語 4 gram，および，多重クラス複合 2 gram[12]を用いる．多重クラス複合 2 gram は，音声認識の探索の第 1 パス内での計算コストの高い高次の言語モデル（単語 Ngram（$N \geq 3$））の利用を避けるために用いる．

多重クラス複合 2 gram では次式で単語の予測確率が計算される．

$$P(w_i|w_{i-1}) = P(w_i|C^t_{w_i})P(C^t_{w_i}|C^f_{w_{i-1}}) \quad (1)$$

ここで，w_i や w_{i-1} は単語，または単語系列である．$C^t_{w_i}$ は w_i が属するクラス，$C^f_{w_{i-1}}$ は w_{i-1} が属するクラスである．右辺第 1 項はクラスから単語または単語系列が出現する確率，右辺第 2 項は発声の頭側のクラス $C^f_{w_{i-1}}$ から末尾側のクラス $C^t_{w_i}$ への遷移確率である．そして，発話中の句点 (x) は，C^t_x は発話終了記号と同じクラスとして登録し，C^f_x は発話開始記号と同じクラスとして登録した．その他の点では多重クラス複合 2 gram[12]と同じ方法で，単語の自動クラスタリング，単語系列抽出，パラメータ推定を行った．

すべてのモデルにおいて，発話中の句点は発話末の句点とは別の単語として登録した．本実験では，最短 20 ミリ秒のポーズと同じ発音記号を句点に付与した．

次に音声認識時の探索の方法であるが，発話中の句点が発話開始記号や発話終了記号とは別の単語として登録されているので，従来の認識システムをそのまま用いることができる．本実験では ATRSPREC[11]を用いる．

前節で述べた学習データと上記のモデルとの組合せを変化させることにより数種の言語モデルを作成して，以下のような評価実験を行った．

まず，句点を含まないテキスト入力に対する分割性能の評価のために，ED4 を学習データとして用いた発話中の句点を含む言語モデルとして，単語 3 gram（3-SPLT-S）と単語 4 gram（4-SPLT-S）と多重クラス複合 2 gram（M-SPLT-S）との 3 種類のモデルを作成した．全モデルの語彙サイズは約 5,600 で，M-SPLT-S のクラス数は 700 とした．また，獲得された単語系列数はおよそ 960 であった．

続いて，音声入力に対する性能を確認する目的で，発話中の句点を含む多重クラス複合 2 gram（M-SPLT）を作成した．さらに，M-SPLT が句点以外の単語の認識性能において劣化がないかどうかを確認するために，発話中の句点を含まない多重クラス複合 2 gram（M-BASE）を作成した．これら 2 つのモデルの学習には，ED3 の全部と ED1 の半分のデータを用いた．評価には ED1 と ED2 とを用いる．M-SPLT と M-BASE との間で語彙サイズは同等で約 14,000 であり，句点を含むか否かだけが異なる．また，多重クラス複合 2 gram の M-SPLT と M-BASE の，クラス数は 700 クラスのモデルを採用した．また単語系列数はおよそ 4,700 であった．

さらに，音声入力に対する分割性能において，発話中の句点を含む多重クラス複合 2 gram（M-SPLT）と他の単語 Ngram モデルとを比較するために，M-SPLT の学習で用いたものと同じ学習データ（ED3 の全部と ED1 の半分）を用いて，発話中の句点を含む単語 3 gram（3-SPLT）と同単語 4 gram（4-SPLT）とを作成した．評価データには ED1 を用いる．単語 3 gram と単語 4 gram に関しては M-SPLT で得られた単語グラフをリスコアリングすることにより評価した．

以上の言語モデルとその学習データと評価データとをまとめて**表 3** に示す．

表 3 評価実験で用いる言語モデルと学習および評価データ
Table 3 Language models and their training and test data used in experiments.

モデル名	学習データ	評価データ	モデルタイプ	句点	実験対象
3-SPLT-S	ED4	ED1	単語 3 gram	有	テキスト
4-SPLT-S	ED4	ED1	単語 4 gram	有	テキスト
M-SPLT-S	ED4	ED1	多重クラス複合 2 gram	有	テキスト
M-SPLT	ED3，ED1 の半分	ED1，ED2	多重クラス複合 2 gram	有	音声
M-BASE	ED3，ED1 の半分	ED1，ED2	多重クラス複合 2 gram	無	音声
3-SPLT	ED3，ED1 の半分	ED1	単語 3 gram	有	音声
4-SPLT	ED3，ED1 の半分	ED1	単語 4 gram	有	音声

表 4 音響モデル学習条件
Table 4 Training conditions for acoustical models.

音響分析
サンプリング周波数 16 kHz，preemphasis 0.98
フレーム周期 10 ms，フレーム長 20 ms（Hamming 窓）
logpower，Δlogpower，12 次-MFCC，12 次-ΔMFCC
ケプストラム平均，パワーを正規化
音響モデル
男性モデル：音声 1400 状態 5 混合，無音 3 状態 10 混合
（学習データ 167 話者，総発話時間 約 2 時間）
女性モデル：音声 1400 状態 15 混合，無音 3 状態 10 混合
（学習データ 240 話者，総発話時間 約 3 時間）

また本認識実験で用いられる辞書内の各単語の発音記号の末尾には，一部の動詞の活用形の末尾（たとえば，動詞 “いただく” の未然形，“いただか” の末尾など）を除いて，ポーズが選択的に付与されている．

4.2.2 音響モデルと音響分析条件

音声認識実験で用いられる音響モデルの設定は，**表 4** のとおりで，ATR 自然発話音声言語データベースに含まれる音声で学習されたモデルである．

4.3 テキスト入力に対する分割性能の評価

句点なしのテキストを入力として文への分割実験を行った．

本実験では，評価データの中の話者が交替するまでの間の形態素解析済の単語列（図 1 の turn1 などの単語列に相当）を対象として，単語既知，しかし，句点未知のもとで，隣り合うすべての単語間に句点が入る場合までを想定して単語グラフを構成し，これに本言語モデルを使って言語尤度付けを行い，そのうちの最も尤度が高いパスを選択した（図 2）．上のように句点が入りうるすべての可能性を想定することにより，句点の認識に必要な音響的要素であるポーズが与えられている場合を仮定したうえでの分割性能の評価を行った．

ここでは，ED4 を用いて作成された言語モデル（3-SPLT-S，4-SPLT-S，M-SPLT-S）を使って，ED1 のテキストを評価対象の入力として，実験を行った．評価対象の句点は，話者が交替するまでの区間に話したすべての文の境界を示す句点のうち，話者が交替する直前の最後の句点（図 1 の turn1 の右端の句点など）を除いた句点（図 1 の sent1，sent2 の各末尾の句点など）とした．これに該当する句点は ED1 の中に 123 個あった．

表 5 テキスト入力に対する発話分割の性能
Table 5 Splitting performance (text inputs).

方式	再現率 [%]	適合率 [%]	F 値
3-SPLT-S	77.9	88.8	83.0
4-SPLT-S	76.2	89.4	82.3
M-SPLT-S	86.9	46.5	60.6

実験結果を**表 5** に示す．数値は左から再現率，適合率，およびこれらの調和平均である F 値の順で書かれている．F 値は $\frac{2.0\times\text{再現率}\times\text{適合率}}{\text{再現率}+\text{適合率}}$ で計算した．テキスト入力は，音声入力において発話内の全単語が認識され句点の音響的条件が満たされ，言語的尤度によって句点が採用されるか否かを決めるだけが残された状態を模擬している．この状況では，単語ベースの統計的言語モデルである単語 3 gram（3-SPLT-S）および単語 4 gram（4-SPLT-S）が F 値の点で高い性能を示した．多重クラス複合 2 gram（M-SPLT-S）では，再現率は高いが，適合率が低かった．

単語ベースの言語モデル（3-SPLT-S や 4-SPLT-S）では，固有の単語列に引き続き句点が現れる場合を直接表現できるため，精度が高く適合率が高いが，学習データに現れた組合せしか記憶されないためスパースネスからカバレッジは低く再現率が低くなる．一方，クラス Ngram に基礎をおく多重クラス複合 2 gram では，精度は低いもののカバレッジは高いため，単語 Ngram の場合とは逆の傾向となる．このため表 5 のような結果になったと考えられる．

また，この結果から，探索ではまず多重クラス複合 2 gram を用い，句点の候補が得られたならば，局所的に単語 4 gram を用いることにより，高い再現率と適合率を得ることが期待できる．

以上の実験では，話者が交替するまで（図1のturn1などの右端まで）を単位として処理を行った．一方，音声認識では，話者が次の話者へ交替するまでの区間を単位とするのではなく，それよりも小さな単位である発話（図1のutt1など）を単位としている．それは，比較的長いポーズが入る箇所で，話者が次の話者へ交替するまでの区間（図1のturn）が複数に分けられた区間である．この対話における発話境界，または発話という音声区間の開始と終了は高い割合で正しく認識することが可能である[13]．

また，この発話と発話の境界の認識に失敗した場合，たとえば，図1で文sent2と文sent3とが別の発話（utt）に分かれているが実は1つの文であるような場合，つまり，音声区間の切り出しによって元々1つの文が複数の発話に分割されている場合でも，本モデルによって誤って認定された発話末が文末であるかどうかを，すなわち，その位置に句点が来るかどうかを高い精度で判定できる．それが可能であることは，話者が交替するまでの区間を単位として，句点を含まないテキストを入力とした場合の句点の認識結果である表5から確認されている．

このため，以下では，発話を入力単位として発話の中の文境界としての句点（図1のtarget）を探索する．以下では，上記の評価対象であった句点のうちの発話の末尾の句点を除き，発話内の句点のみ（図1のtarget）を対象として実験を行う．

4.4 音声入力に対する分割性能の評価

できるだけ多くのデータを音声認識の言語モデルの学習に用いることによって，認識精度を高くできる見込みがある．そのためED3とED1の半分を学習データとして作られた言語モデルであるM-SPLTを用いた．そして，分割の評価データにED1を用いて発話を入力単位とした音声を入力として，音声認識過程での発話分割実験を行った．さらに，M-SPLTを使った音声認識で得られた単語グラフのリスコアリングを，単語3gram（3-SPLT）と単語4gram（4-SPLT）とを用いてそれぞれ行うことによって，単語3gram（3-SPLT）と単語4gram（4-SPLT）とを用いた音声認識過程での分割を模擬した．

ここでは発話内の句点が対象であり，その総数は73であった．分割の評価については，認識結果の第1位候補での句点の再現率と適合率とF値の観点から評価を行った．

結果を**表6**に示す．表6から，単語3gramよりも単語4gramの方がF値の点で分割性能が高かった．単語NgramのNが大きいモデルの方の性能が高い結果となった．

表6 音声入力の場合の発話分割の再現率と適合率
Table 6 Recall and precision of splitting (speech input).

	再現率 [%]	適合率 [%]	F値
3-SPLT	64.4	100.0	78.3
4-SPLT	94.5	100.0	97.2
M-SPLT	78.1	90.5	83.8

発話内の句点に関しては多様性が少ないため再現率が高くなり，単語4gramでの結果が他の2つのモデルよりも良くなったと考えられる．また，多重クラス複合2gramでは，やはり適合率が単語3gramや単語4gramに比べて低くなったけれども，辞書に登録された単語系列によって，単語3gramよりも再現率の高い結果となったと考えられる．

4.3節の後半で，句点を含まないテキスト入力に対して，まず多重クラス複合2gram，次に単語4gramを利用することによって性能向上を期待したように，M-SPLTの出力結果を4-SPLTでリスコアリングすることによって再現率94.5%，適合率100.0%という結果が得られた（表6）．

4.4.1 分割誤り

ED1の音声入力に対するM-SPLTを使った分割結果のうち，分割誤りの事例のいくつかをあげる．

削除誤り（分割漏れ）には例2のような事例があった．「×」が分割位置であるにもかかわらず正しく分割されなかった分割位置である．

例2

削除誤： ‥ 申し訳ございません × シングルは ‥
削除誤： 東京シティーホテルご滞在 × 零三の …
削除誤： 調べます × しばらくお待ち下さい

“申し訳ございません”の後にはポーズが認識されていなかった．次の単語である“シングル”の頭の音/sh/に句点の発音が吸収されたか，または20ミリ秒以下のポーズであったためとらえられなかったと考えられる．

“ご滞在”のような体言止めの例は少ない恐れがある．そのため，その語の後の句点に対して多重クラス複合2gramから与えられる言語尤度が高くなかったことが考えられる．本実験用の認識用の辞書の各単語の発音記号の末尾には，動詞の一部の活用形以外にはポーズが選択的に付与されている．句点の言語尤度が十分に高くない場合，句点のポーズが単語の末尾のポーズとして単語に吸収されてしまうことが考えられる．実際に，“ご滞在”の後の句点の削除誤りでは，句点の直前の単語（ご滞在）の方に，ポーズ（615ミリ秒）が吸収されてしまっており，句点の発音としての

表 7 単語認識率
Table 7 Word accuracy.

	ED1	ED2
M-SPLT	92.9	85.6
M-BASE	93.1	85.3

ポーズが認識されていなかった．

“調べます” の後でも，直前の “ます” の方にポーズ（170 ミリ秒）が吸収されてしまっていた．

一方，ポーズが短い場合でも，句点の認識ができているものもあった．このため，句点を認識できるか否かはポーズの長短に必ずしも依存するものではなく，句点に与えられる言語尤度が重要と考えられる．

同様に，挿入誤り（過分割）には，例 3 のような事例があった．「※」が誤って挿入された分割位置を示す．

例 3

挿入誤： そうですか※料金はそれぞれおいくら ….

挿入誤： そうですか※じゃバス付の方でお願いしたい ‥

データベース内の「そうですね」や「そうですか」は話し始めの決まり文句に相当し，それらの後の位置には，句点ではなく読点がおかれているが，一般に終助詞のあとに句点が置かれることが多い．その結果挿入誤りとなったと考えられる．これは M-SPLT のようなバイグラムでは正しくとらえることは困難である．

4.4.2 単語認識率の比較

句点を組み込まない従来の言語モデルを句点を組み込んだ言語モデルに変更することによって，句点以外の単語の認識性能の劣化がないかどうかを確認するために，M-SPLT と M-BASE を使った場合の認識率の比較を行う．これまで用いてきた評価データ ED1 に加えて，主に音声認識の評価用として使われるデータである ED2 に対する実験も行った．音声認識結果の第 1 位候補での単語認識率（%Accuracy）を**表 7** に示す．表 7 の「M-SPLT」の行は，M-BASE と比べるために算出した，認識結果の第 1 位候補と正解との間での句点以外の単語の認識率である．表 7 のように，M-SPLT と M-BASE とのそれぞれの結果の比較によれば，M-SPLT モデルは M-BASE と句点以外の単語の認識性能においてほとんど違いがなく，句点を学習することによる性能の劣化はない．

5. 評価実験のまとめ

5.1 分割の観点から

句点を含まないテキストの分割においては，オープンな評価実験の結果（表 5），単語 3 gram（3-SPLT-S）および単語 4 gram（4-SPLT-S）が F 値の点で高い性能を示した．多重クラス複合 2 gram（M-SPLT-S）では，再現率は高いが，適合率が低かった．

音声入力に対する分割においては，評価実験の結果（表 6），多重クラス複合 2 gram（M-SPLT）では，8 割近くの再現率と 9 割の適合率が得られた．多重クラス複合 2 gram（M-SPLT）を使った場合の分割の削除誤り（例 2）や挿入誤り（例 3）の事例の分析から，句点の認識には句点に与えられる言語尤度が重要であると考えられる．また，多重クラス複合 2 gram では，話し始めの決まり文句の末尾の終助詞の後に，句点が来る確率を正しく推定することが困難であり，実際よりも大きな確率が与られ挿入誤りが生じたと考えられる．

M-SPLT の単語グラフのリスコアリングによって，単語 3 gram（3-SPLT）や単語 4 gram（4-SPLT）を用いることにより，認識過程の第 1 パスで計算コストの高い高次 Ngram（$N \geq 3$）の使用を避けながら，単語 4 gram で最高 94%の再現率と 100%の適合率（F 値で 97%）という再現率と適合率とがともに高い結果が得られた．

5.2 音声認識率の観点から

句点以外の単語認識率の点では，表 7 のように，句点を含むモデルと含まないモデルとの間での性能の差はない．すなわち，句点を言語モデルに含めることによる句点以外の単語の音声認識への悪影響はない．

6. おわりに

本稿では，発話内の文境界の記号としての句点を他の単語と同様に音声認識することによって発話の分割を行うための統計的言語モデルを提案し，音声対話データでの分割性能の評価と音声認識性能の評価とを行った．従来の音声認識のための統計的言語モデルとの違いは，句点を単語として扱うこと，句点にはポーズの発音を与えることだけであるが，その結果，

- 対話音声に対して高い再現率と適合率で発話の分割を行えること，
- 句点以外の単語の音声認識性能を劣化させないこと，

が確認された．また，本方法には

- 音声認識の過程で句点を認識し，そこで発話の分割を行うので，リアルタイムで認識結果を出力でき，早期に後段の言語処理に多くの情報を出力できること，
- 分割のための情報が音声認識の統計的言語モデルつまり Ngram の枠組みに統一されているので，モデルの維持管理が容易であること，

という利点があり，有効なモデルとなっている．

謝辞 実験環境の構築にご協力くださった林輝昭氏に感謝します．日頃ご討論ご指導くださる匂坂芳典元室長，比較データの提供と議論を行ってくださった竹澤寿幸主任研究員，および ATR 音声言語通信研究所の皆様，ならびに，NTT の横尾昭男氏，塚田元氏に感謝します．

参 考 文 献

1) Lavie, A., et al.: Input Segmentation of Spontaneous Speech in JANUS: a Speech-to-speech Translation System, *Proc. ECAI-96 Workshop on Dialogue Processing in Spoken Language Systems* (1996).
2) 竹澤寿幸，森元 逞：発話単位の分割または接合による言語処理単位への変換手法，自然言語処理，Vol.6, No.2, pp.83–95 (1999).
3) Stolcke, A. and Shriberg, E.: Automatic Linguistic Segmentation, *Proc. ICSLP-96*, pp.1005–1008 (1996).
4) Zechner, K. and Waibel, A.: DiaSumm: Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains, *Proc. Coling-2000*, pp.968–974 (2000).
5) Gavalda, M., et al.: High Performance Segmentation of Spontaneous Speech Using Part of Speech and Trigger Word Information, *Proc. 5th ANLP*, pp.12–15 (1997).
6) Chen, C.J.: Speech Recognition with Automatic Punctuation, *Proc. Eurospeech-99*, pp.447–450 (1999).
7) Furuse, O., et al.: Splitting Long or Ill-formed Input for Robust Spoken-language Translation, *Proc. COLING-ACL-98*, pp.421–427 (1998).
8) 古瀬 蔵ほか：多言語話し言葉翻訳に関する変換主導翻訳システムの評価，言語処理学会第 3 回年次大会発表論文集，pp.39–42 (1997).
9) Takezawa, T., et al.: Speech and Language Databases for Speech Translation Research in ATR, *Proc. 1st International Workshop on East-Asian Language Resources and Evaluation* (*EALREW '98*), pp.148–155 (1998).
10) Morimoto, T., et al.: A Speech and Language Databases for Speech Translation Research, *Proc. ICSLP-94*, pp.1791–1794 (1994).
11) 内藤正樹ほか：旅行会話タスクにおける ATR-SPREC の性能評価，日本音響学会秋季研究発表会論文集，3-1-9, pp.113–114 (1999).
12) Yamamoto, H., et al.: Multi-class Composite N-gram Based on Connection Direction, *Proc. ICASSP-99*, pp.533–536 (1999).
13) 山本博史ほか：母音および無音の HMM を用いた音声始端検出法，日本音響学会春季研究発表会論文集，1-Q-3, pp.137–138 (2000).
14) 情報処理振興事業協会（IPA）：日本語ディクテーション基本ソフトウエアの開発仕様策定書および説明書—1997 年度 (1998).
15) Imai, T., et al.: A language model for recognition of continuously uttered sentences, *J. Acoust. Soc. Jpn.*(*E*), Vol.21, No.2, pp.111–114 (2000).

(平成 12 年 7 月 21 日受付)
(平成 13 年 9 月 12 日採録)

中嶋 秀治（正会員）

1990 年徳島大学工学部情報工学科卒業，1992 年同大学大学院工学研究科情報工学専攻修士課程修了．同年，日本電信電話（株）入社．1997 年 3 月より国際電気通信基礎技術研究所（ATR）に出向．現在，ATR 音声言語通信研究所に勤務．音声翻訳通信実現のための話し言葉処理の研究に従事．日本音響学会，言語処理学会，電子情報通信学会，日本認知科学会各会員．

山本 博史

1979 年東京大学農学部農業生物学科卒業，1981 年同大学大学院修士課程修了．同年，(株）CSK に入社．1996 年より ATR 音声翻訳通信研究所に出向．現在，ATR 音声言語通信研究所に勤務．音声認識の研究開発に従事．日本音響学会，言語処理学会，電子情報通信学会各会員．