音声言語情報処理 35−14
(2001. 2. 3)

話者の心の状態遷移モデルに基づく対話音声認識

藤崎 博也 [1], 阿部 賢司 [1], 黒川 一滋 [1], 武田 和也 [1], 成澤 修一 [1], 大野 澄雄 [2]

[1] 東京理科大学 基礎工学部
〒278 野田市山崎 2641
Tel: 0471-24-1501, Fax: 0471-22-9195
E-mail: fujisaki@te.noda.sut.ac.jp

[2] 東京工科大学 工学部
〒192-0982 東京都八王子市片倉町 1404-1
Tel: 0426-37-2111, FAX: 0426-37-2112
E-mail: ohno@cc.teu.ac.jp

あらまし 音声認識において話題や話者の内部状態に適応した言語モデルを用いる場合、生起し得る単語列の候補数が減少するため、認識率は高くなることが予想される。このような観点から、筆者らは、一般的なコーパスから作成した平均的な言語モデルを用いた場合を基準とし、話題を "学術情報検索" に限定した場合のユーザの発話の言語モデル、および、ユーザの内部状態に適応した言語モデルの採用による認識率の改善の効果を実験により検証した。その結果、平均的な言語モデルよりも、話題に適応し、さらに状態に適応した言語モデルの採用が最も効果的であることを確認した。

キーワード 話者の心のモデル, 状態遷移モデル, 状態推定, 言語モデル, 音声認識

Speech Recognition in Spoken Dialogue Based on State Transition Model of Speaker's Mind

Hiroya Fujisaki [1], Kenji Abe [1], Kazushige Kurokawa [1], Kazunari Taketa [1], Shuichi Narusawa [1], Sumio Ohno [2]

[1] Science University of Tokyo
2641 Yamazaki, Noda, 278-8510
Tel: 0471-24-1501, Fax: 0471-22-9195
E-mail: fujisaki@te.noda.sut.ac.jp

[2] Tokyo University of Technology
1404-1 Katakura, Hachioji, 192-0982
Tel: 0426-37-2111, FAX: 0426-37-2112
E-mail: ohno@cc.teu.ac.jp

Abstract The performance of a speech recognition/understanding system is expected to be higher if the language model can be adapted to the current topic of the utterance. It is expected to be still higher if the model can be adapted to the current state of the mind of the speaker. From this point of view, the present study examines the merits of these adapted language models over a language model obtained from a general corpus, by restricting the topic to "academic information retrieval" and by adopting a representation of the speaker's mind in terms of a probabilistic finite-state automaton. The experimental results confirmed the advantages of these models in quantitative terms.

Key words model of speaker's mind, state transition model, state inference, language model, speech recognition

1. はじめに

近年の音声認識では、平均的な言語モデル (例えば単語列の生起確率モデル) を用いることが多いが、これは特定の話題や場面における言語の生起確率とは必ずしも合致しないため、最適なものとはいえない [1]。これに対し、言語モデルを話題に適応的に変化させることにより、認識率の向上が得られる [2]。さらに、明確な目的をもつ対話の場合には、言語モデルを対話の状態に適応的に変化させることにより、認識率の一層の向上が期待される [3]。

筆者らは、さきに音声対話に基づく知的情報検索システムにおいてユーザおよびシステムをそれぞれ確率的有限状態オートマトンとしてモデル化し、それらの状態を考慮して対話を効率的に管理する手法を提案した [4,5]。本稿では、一般的なコーパスから作成した平均的な言語モデルを用いた場合を基準とし、このシステムにおいて、話題を "学術情報検索" に限定した場合のユーザの発話の言語モデル、および、ユーザの内部状態に適応した言語モデルの採用による認識率の改善の効果を実験により検証した結果について述べる。

2. 音声認識における言語モデル

音声言語の生成過程を音声認識の立場からモデル化する試みは 1970 年代から行われている。すなわち F. Jelinek [6] は図 1(a) のモデルに基づいて、音響パラメータベクトルの時系列 A に対する単語列 W の事後確率 $P(W|A)$ を最大化する W を選択する問題として音声認識を定式化した。これは、現在の確率的手法による音声認識の基礎となっている。一方、B.-H. Juang [7] は図 1(b) のモデルに示すように言語モデルの前段階としてメッセージ源 (概念モデル) を設け、音響パラメータベクトルの時系列 A に対するメッセージ M の事後確率を最大化する問題として定式化した。この場合、言語モデルはメッセージ M に適応して変化するため、認識率の向上が期待されるが、その程度は僅少であったと報告されている [8]。

これに対して筆者らは、特定の発話の生起確率を左右する最大の要因が話者の心の状態であると考える。従って本稿では、図 1(c) に示すように、話者の心のモデルの状態に依存する言語モデルを導入する。この場合、図 1(d) のように、心のモデルと言語モデルとの間に概念モデルの存在を想定することも可能であるが、ここでは、図 1(c) のモデルに従って論を進める。

以下、本稿では、このモデルに基づき、筆者らが先に提案した対話管理手法を用いて話者の内部状態を推定し、対話音声認識を高度化する手法について述べる。

3. 新しい対話管理手法

3.1 対話管理手法の特徴

筆者らが提案した対話管理手法の特徴は以下の 2 つである。

(1) ユーザモデルおよびシステムモデルを別々に定義するため、それぞれのモデルが簡単かつ明解になり、その構築・修正が容易である。

(2) ユーザおよびシステムの内部状態の推定が可能であり、対話を効率良く管理することができる。

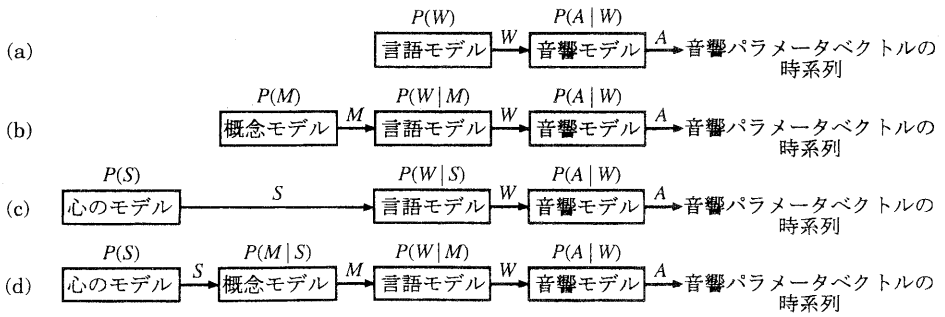


図 1. 音声認識の立場からみた音声言語の生成過程のモデル

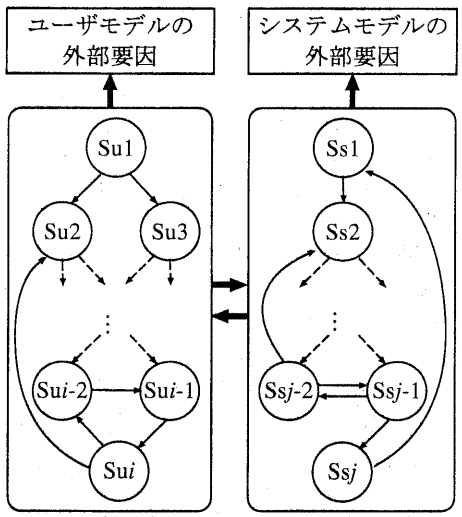
3.2 ユーザモデルとシステムモデルの構築方針

この研究では、情報検索におけるユーザおよびシステムの内部状態を、

a) 「検索要求を伝えたい状態」や「検索要求受付け状態」といった、処理手続き上の状態 (以下、これを "処理手続き上の状態" と呼ぶ)

b) 「検索速度を重視したい場合」や「インターネット上の回線が混雑している場合」といった、ユーザの検索処理全体に対する要求や、システム内外の状況など、処理形態に影響を及ぼす外部要因 (以下、これを "外部要因" と呼ぶ)

の 2 つの要素で記述することとし、これらの内部状態間の関係を確率的有限状態オートマトンモデルとして表現したものを、ユーザモデルおよびシステムモデルと呼ぶ。ただし、ここで用いるオートマトンは、"現状態"、"現状態への入力"、"次状態に遷移する際の出力"、"次状態" の 4 つの要素によって記述するものであり、"現状態"、"次状態に遷移する際の出力"、"次状態" の 3 つの要素で記述する従来のものとは異なる。ここで、上記 a) は、モデルを構築する際の基本要素として、また、b) は、モデルにおける状態遷移確率に影響を及ぼすパラメータとして定義する。これらのことを考慮したユーザモデルおよびシステムモデルは図 2 のようになる。これらのモデルを構築する際には、まず、処理手続き上の状態およびそれらの状態遷移規則を決定し、次に、外部要因を考慮した場合の状態遷移確率を決定する。



Su1~Sui : ユーザの処理手続き上の状態
Ss1~Ssj : システムの処理手続き上の状態
➡ : 参照の方向

図 2. ユーザモデル・システムモデル

3.3 構築の手順

ユーザモデルおよびシステムモデルを構築するために、まず、学術情報検索を目的とした多数の模擬対話を収集・分析し、各発話の意味的なタイプ (以下、本稿では、これを "発話タイプ" と呼ぶ)、および、それらの発話が生起する際のユーザおよびシステムの内部状態を調査した。つぎに、その結果に基づいて、発話タイプとその発話が生起したときのユーザあるいはシステムの内部状態に関する情報を各発話に付加したタグつき対話コーパスを作成し、そこから求められる状態遷移規則に従ってモデルを構築した。

3.4 模擬対話の収集・分析

学術情報検索を目的とした対話における各発話のタイプ、および、それらの発話が生起する際のユーザおよびシステムの内部状態を把握するため、ユーザ役 (10 名) とシステム役 (7 名) を設定し、Wizard of Oz 法に基づいて模擬対話 100 対話 (3417 発話) を収集した。ユーザ役およびシステム役に関しては、

(1) ユーザ役は学術論文を検索するという目的を把握しており、検索対象について最低限の知識を持っている。また、発話は自由に行なえる。

(2) システム役はシステムの機能および限界を把握している。また、発話は予め定めたマニュアルに従って行なう。

という条件を設けた。

つぎに、収集した対話を分析し、各発話をタイプ毎に分類した。発話タイプの種類を表 1に示す。また、各発話が生起する際のユーザ・システムの内部状態を人間の判断に基づいて推定・分類した。ただし、対話から推定できる内部状態は、一般に、処理手続き上の状態となる。これらの状態の種類

を表2に示す。さらに、発話から直接推定することはできないが、情報検索において想定すべき外部要因の例を表3に示す。

表 1. 発話タイプの種類

発話者	発話タイプ
ユーザ	Ou1: 検索開始の挨拶
	Ou2: 検索要求
	Ou3: 検索式訂正要求
	Ou4: 検索実行の許可
	Ou5: 検索やり直し要求
	Ou6: 他論文表示要求
	Ou7: 詳細表示要求
	Ou8: 欲しい論文の指定
	Ou9: 欲しい論文の指定解除
	Ou10: 指定した論文の要求
	Ou11: 新規検索要求
	Ou12: 検索終了要求
システム	Os1: 検索要求の催促
	Os2: 検索式および検索実行の確認
	Os3: 検索式の訂正内容の催促
	Os4: 検索結果の提示 1(ヒット件数: 1 以上)
	Os5: 検索結果の提示 2(ヒット件数: 0)
	Os6: 論文の詳細情報の提示
	Os7: 論文指定の確認
	Os8: 論文指定解除の確認
	Os9: 指定された論文の提示
	Os10: 検索継続意図の確認
	Os11: 検索終了の挨拶

表 2. ユーザ・システムの処理手続き上の状態

発話者	状態の種類
ユーザ	Su1: 初期状態
	Su2: 検索要求を伝えたい
	Su3: 検索結果待ち
	Su4: 検索式を訂正したい
	Su5: 検索結果を吟味したい
	Su6: 論文を提示してもらいたい
	Su7: 検索を終了したい
	Su8: 終了状態
システム	Ss1: 初期状態
	Ss2: 検索要求受付け
	Ss3: 検索式待ち
	Ss4: 検索実行の命令受付け
	Ss5: 検索式訂正要求受付け
	Ss6: 検索結果待ち
	Ss7: 提供する論文の受付け
	Ss8: 検索継続意図の受付け
	Ss9: 終了状態

表 3. ユーザ・システムの外部要因の例

発話者	外部要因の種類
ユーザ	検索における重視点
	システムに対する習熟度
	検索対象の具体性
システム	インターネット上の回線状況
	システム稼働率
	時間制限

3.5 タグつき対話コーパスの作成

収集した対話をテキスト形式に変換し、発話ごとに、発話タイプとその発話が生起したときのユーザおよびシステムの内部状態に関する情報を付加したタグつき対話コーパスを作成した。その一部を以下に示す(ただし、Ou1~Ou8、Os1~Os4、Su1~Su6、Ss1~Ss6 の内容は、表 1、2に従う)。なお、このコーパスには、システムから情報検索部への応答 (S2, S4)、および、情報検索部からシステムへの応答 (I1, I2) など、ユーザ-システム間の対話に現れない部分の動作も付加した。また、発話をせずに状態が遷移する場合 (U4, U5) の動作も付加した。

<タグつき対話コーパスの一部>

U1 論文を探しているんですけど。/Ou1/Su1/

S1 どのような論文をお探しですか。/Os1/Ss1/

U2 自然言語処理関係の論文が欲しいんですけど。/Ou2/Su2/

S2 検索式生成依頼 (情報検索部へ)。/検索式生成依頼/Ss2/

I1 検索式 (システムへ)。/検索式/情報検索部/

S3 自然言語処理というキーワードで検索してよろしいですか。/Os2/Ss3/

U3 はい。/Ou4/Su3/

S4 検索依頼 (情報検索部へ)。/検索依頼/Ss4/

I2 検索結果 (システムへ)。/検索結果/情報検索部/

S5 該当する論文は以下の通りです。ご希望の論文はございますか。また、論文の詳細情報を表示することもできます。/Os4/Ss6/

U4 状態遷移 /Null/Su3/

U5 状態遷移 /Null/Su5/

U6 5 番目の論文が欲しいんですけど。/Ou8/Su6/

(U1~U6: ユーザの発話・動作、S1~S5: システムの発話・動作、I1~I2: 情報検索部の動作、Null: 発話なし)

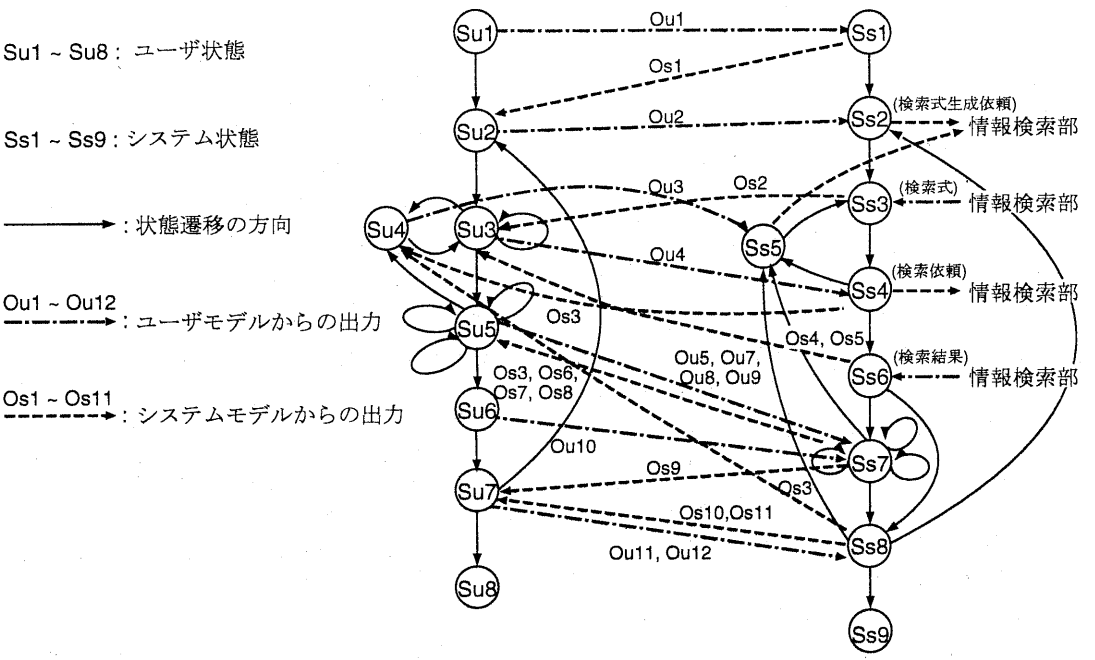


図 3. 対話コーパスに基づいて構築したユーザモデル・システムモデル

3.6 タグつき対話コーパスからのモデルの構築

作成したタグつき対話コーパスから、ユーザおよびシステムの内部状態の遷移規則、すなわち、現状態 i において相手からの応答 j が入力されたとき、相手への応答 k を出力して次状態 l に遷移する規則を抽出し、それを確率的有限状態オートマトンとしてモデル化することにより、ユーザモデルおよびシステムモデルを構築した。

構築したユーザモデルおよびシステムモデルを図 3に示す。Su1~Su8、Ss1~Ss9 は、ユーザおよびシステムの処理手続き上の状態を、Ou1~Ou12、Os1~Os11 は各状態からの出力を表す。この図のように、一般には、ユーザモデルの出力がシステムモデルへの入力となり、システムモデルの出力がユーザモデルへの入力となる。なお、この図では、各状態遷移規則の生起確率に関する記述を省略しているが、これらの生起確率は、タグつき対話コーパス、および、外部要因 (表 3参照) に基づいて求める。

4. 構築したモデルに基づく対話音声認識

4.1 ユーザの状態毎の言語モデルと perplexity

言語モデルとして単語バイグラムを用い、話題を限定した対話コーパスにおけるモデルの perplexity と、状態に適応したモデルの perplexity とを比較した結果を表 4に示す。この表からも明らかなよ

表 4. 単語バイグラムの perplexity

バイグラムの種類	perplexity
1) 話題に適応した単語バイグラム	4.393
2) 状態に適応した単語バイグラム：	
Su1 における単語バイグラム	1.601
Su2 における単語バイグラム	2.646
Su3 における単語バイグラム	1.810
Su4 における単語バイグラム	3.267
Su5 における単語バイグラム	3.606
Su6 における単語バイグラム	1.685
Su7 における単語バイグラム	1.665
状態出現率を考慮した Su1 ~ Su7 の perplexity の平均	3.051

うに、いずれの場合も、状態に適応したバイグラムの方が話題を限定しただけのバイグラムよりもperplexityが低い。

4.2 対話音声認識実験

平均的な言語モデルとして、文献[9]の基本ソフトウエアに収録されている20000語彙からなる単語バイグラム(毎日新聞CD-ROM 1991年版から1994年版を利用して構築)を、また、話題に適応した言語モデルとして、先に述べた対話コーパス(収録語彙数：307)から求めた単語バイグラムを、さらに、状態に適応した言語モデルとして、状態毎(Su1 ～ Su7)の各発話から求めた単語バイグラムを用意し、それらの言語モデルを対話音声認識に適用したときの効果を実験的に検証した。実験では、音声認識器としてJulius [9]を使用し、また、入力データとして、情報検索を目的としたユーザ発話(60発話)を用いた。

実験結果を表5、6に示す。表5は単語認識率に着目して比較した結果を、表6は対話音声理解率に着目して比較した結果を示している。ここで、対話音声理解率に関しては、入力データの意味的内容が正しく再現されているか否かを人間が判断し、正しく再現されている場合を正解とした。

単語認識率、対話音声理解率のいずれにおいても、話題を限定し、さらに状態に適応した言語モデルを用いた場合の結果が最も高い値を示している。

表5. 各言語モデルを用いたときの単語認識率

言語モデル	認識率 [%]
平均的なモデル	71.8
話題に適応したモデル	73.1
状態に適応したモデル	77.1

表6. 各言語モデルを用いたときの対話音声理解率

言語モデル	理解率 [%]
平均的なモデル	61.9
話題に適応したモデル	70.0
状態に適応したモデル	75.0

5. おわりに

本稿では、音声認識に一般的なコーパスから作成した平均的な言語モデルを用いた場合を基準とし、話題を“学術情報検索”に限定した場合のユーザの発話の言語モデル、および、ユーザの内部状態に適応した言語モデルの採用による認識率の改善の効果を実験により検証した結果について述べ、単語認識および対話音声理解のいずれにおいても、話題を限定し、さらに状態に適応した言語モデルを用る方法が最も効果的であることを示した。

参考文献

[1] 大黒 慶久, 中川 聖一: “音韻認識率・文発声法・Perplexity及び文認識率との相互関係,” 信学技報, SP88-113 (1988).

[2] 中川 聖一: “音声認識研究の動向,” 電子情報通信学会論文誌, vol. J83-D-II, no. 2, pp. 433-457 (2000).

[3] F. Wessel, A. Baader and H. Ney: “A comparison of dialogue-state dependent language models,” ESCA Workshop on Interactive Dialogue in Multi-modal Systems, Kloster Irsee, pp. 93-96 (1999).

[4] H. Fujisaki, H. Kameda, S. Ohno, T. Ito, K. Tajima and K. Abe: “An intelligent system for information retrieval over the Internet through spoken dialogue,” *Proceedings of Eurospeech'97*, vol.3, pp.1675-1678 (1997).

[5] K. Abe, K. Kurokawa, K. Taketa, S. Ohno and H. Fujisaki: “A new method for dialogue management in an intelligent system for information retrieval,” *Proceedings of ICSLP2000*, vol. 2, pp. 118-121 (2000).

[6] F. Jelinek: “Continuous speech recognition by statistical methods,” *Proceedings of IEEE*, vol. 64, no. 4, pp. 532-556 (1976).

[7] B.-H. Juang: “Automatic speech recognition: Problems, progress & prospects,” IEEE Workshop on Neural Networks for Signal Processing (1996).

[8] 大附 克年, 古井 貞熙, 桜井 直之, 岩崎 淳, 張 志勝: “ニュース音声認識のための言語モデルと音響モデルの検討,” 信学技報, vol. 98, no. 463, pp. 1-7 (1998).

[9] 河原 達也, 李 晃伸, 小林 哲則, 武田 一哉, 峯松 信明, 嵯峨山 茂樹, 伊藤 克亘, 伊藤 彰則, 山本 幹雄, 山田 篤, 宇津呂 武仁, 鹿野 清宏：日本ディクテーション基本ソフトウェア - 1999年度版 - (2000).