発話状態の予測に基づく対話音声認識手法とその効果

鈴木 雅実† 松崎 克郎‡ 井ノ上 直己† 谷戸 文廣†

†KDD 研究所 ‡慶応義塾大学 (現在 ソニー株式会社)

〒356 上福岡市大原 2-1-15 〒223 横浜市港北区日吉 3-14-1

E-mail: {msuzuki, inoue, yato}@lab.kdd.co.jp, katsuro@vsp.cpg.sony.co.jp

対話音声認識の精度向上を図る手法として，著者らは発話状態の遷移モデルに基づいて，発話状態予測を行なう効率的な手法を提案した。発話状態の一重遷移モデルおよび二重遷移モデルを使用した場合の予測効果を，2種類の音響モデルについて実際に音声認識実験を行なった結果，両者とも音声認識率の向上が見られたが，基本性能の良い音響モデルでは，効果はそれほど顕著ではないことが観察された。また，各発話状態毎の認識率の差や，発話状態の予測のみでは認識性能の向上に寄与しない場合等を分析し，さらに対話音声認識性能を向上させるために利用可能な情報源について考察した。

An Efficient Method of Dialogue Speech Recognition based on Dialogue States Prediction

Masami SUZUKI† Katsuro MATSUZAKI‡ Naomi INOUE† Fumihiro YATO†

†KDD Research and Development Laboratories (2-1-15, Ohara, Kamifukuoka-shi, 356 JAPAN)
‡Keio University (3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, 223 JAPAN)

E-mail: {msuzuki, inoue, yato}@lab.kdd.co.jp, katsuro@vsp.cpg.sony.co.jp

In this article, we report some enhancement on the proposed method for dialogue speech recognition that uses dialogue states transition model and the grammar rules which have relevance with the model. The effects on speech recognition result with different acoustic models were compared. Based on the result, relationships between speech recognition rates and various dialogue states or grammatical complexity were examined. Finally, potential information sources to be used for more effective recognition were discussed in our framework.

1 はじめに

対話音声認識の精度向上と処理の効率化を同時に達成するために，著者らは発話状態の遷移テーブルと，発話状態との関連性を各規則に付与した認識文法を用いて，認識時に探索空間を絞り込む手法を提案した[2]。本稿では，比較的規模の大きな認識文法を用いて，提案した手法が異なる音響モデルの下で，どのような効果を持つかを測定し，その結果について行なった分析に基づいて，局所的な発話状態予測自体の限界や，さらに利用可能な情報源について考察する。

2 背景と目的

対話音声認識においては，ある時点の発話状態から次の発話として後続する可能性のある発話状態を予測することにより，認識候補を絞り込む手法が提案されている。たとえば，永田 [1] は，各発話の発話者と発話行為タイプの組合せによる発話状態から，発話行為タイプのラベルを付与した対話コーパスに基づいて次の発話状態を統計的に予測する対話モデルとその情報理論的な有効性を報告している。このほか，プラン認識等の対話知識のモデルに基づいて，予測を行なう手法も提案されているが，発話タイプ等の表層的なモデルと比較して，知識の記述量や計算量等のコストが増大する問題点が指摘されている。また，これまでの予測モデルは，音声認識候補が出力された後で，その中から適合度の高いものを選択する方式が多い。

一方，小規模な対話システムでは，発話状態を記述するプリミティブの数自体が少なく，有限小数個の発話状態間の遷移が固定的で，各状態毎に対応する認識文法 (CFG, n-gram 等) を作成することが比較的容易である場合は，効率良く対話音声認識を実行することが可能である。しかし，その拡張性には疑問がある。また，これまで，次発話予測の仕組みがある程度の規模以上の対話音声認識システム上に実装され，評価された例はあまり見られない。

本研究では，従来のこのような研究を踏まえた上で，ある発話状態から次の発話状態への遷移可能性に基づいて，次発話として予測される文に関する文法情報のみを音声認識時に利用することにより，計算量の削減と認識精度の向上を図った。すなわち，文献 [2] で，この考えに基づく対話音声認識の手法を提案し，実際にホテル予約をタスクとする自動通訳システムの中に次発話を予測する仕組みを実装した [3]。この段階では，発話状態として利用したのは，発話者の役割 (ホテル予約係／利用者) と 15 種類の発話タイプであった。さらに文献 [5] では，発話タイプに加えて，タスクにおける話題のクラスを用いること，発話状態の一重遷移を用いる基本手法を拡張して二重遷移を用いることにより，予測の精度が増し文認識率も向上することを確認した。今回の報告は，以上の研究結果に基づいて，さらに認識精度の向上と効率化を目的として，次の項目に関する実験および検討を行なった結果をまとめたものである。

1. **使用する音響モデルの違いによる効果の比較と実装上の効率化**

   これまで認識処理系に用いた音響モデルは離散型の HMM であったが，これと比較して基本性能において優れている，連続型の HMM を用いて行なった認識実験的の結果を報告する。また，実装上の工夫として，各発話状態に対応した認識用の単語予測 LR テーブルを予め用意する従来方法の改良による記憶空間の削減を図った。

2. **認識実験結果の分析に基づく発話状態予測の寄与に関する検討**

   音声認識実験結果から得られた認識率と文の特徴との関係について述べる。また，認識対象文の発話タイプによる認識率の差についても分析する。

3. **認識精度向上のために利用可能な情報源の検討**

   認識実験結果を分析した結果に基づき，さらに認識率を向上させるための新たな枠組の導入について述べ，期待される効果等について検討する。

3 音響モデルの違いによる効果の比較と実装上の効率化

本稿で述べる対話音声認識の手法については，すでに文献 [2], [5] で紹介した通りであるが，ここでその基本的な枠組を説明する。

この手法の特徴は，文脈自由文法 (CFG) で与える文法規則群内の各規則に対して発話状態との関連を示すヘッダーを付与しておくことにある。従って，ある発話状態において予測される次の発話状態に対応する規則のみを抽出した文法サブセットによる単語予測 LR テーブルを作ることが可能となり，これにより効率的にかつ精度良く対話音声認識が実行される。

また，この発話予測のモデルは当初，比較的狭い範囲の「ホテル予約」のタスクについて人手で記述した発話状態の一重遷移モデルを使用していたが，筆者らの前の報告 [5] では，対象とするホテル予約に関して模擬対話を収録して，極力制約を付けない状態で，どのような流れで対話が進行するか，また，どのような

言語現象が現れるかを調査した[1]。その結果，文を発話タイプ，話題によって，40種類に分類し，収録された対話にタイプ，話題のヘッダを付与した。この分析結果を基に，発話状態の一重遷移可能性および二重遷移可能性の予測モデルを作成した [5] [6]。

次発話の予測の仕組みは次の通りである。

- 発話状態の一重遷移モデルに基づく予測

  「ある発話状態からの可能な遷移先」を「今の状態から予測される遷移可能な話題，タイプの範囲」とし，現発話の認識結果から話題，タイプを判別し，発話状態の一重遷移テーブルと照合して次発話の話題，発話タイプを予測し，認識対象候補を限定する。ただし，対話の開始時については，「会話のはじめの挨拶に類する文」(G000) からはじまるものとする。

- 発話状態の二重遷移モデルに基づく予測

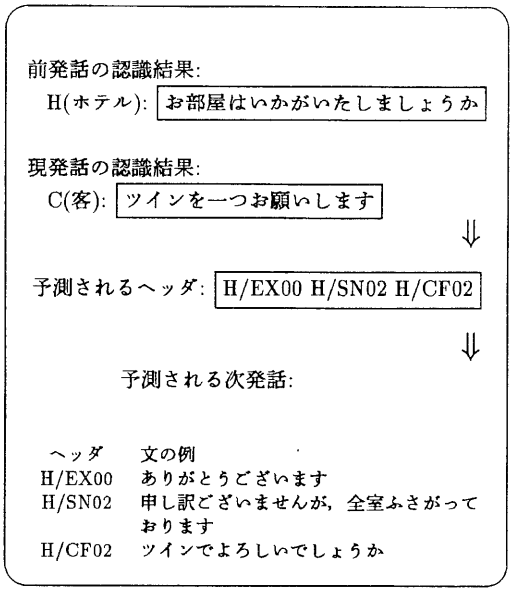
  考え方は，一重遷移モデルの場合と同じである。現発話と前発話の認識結果から話題，タイプを判別し，発話状態の二重遷移テーブルと照合して次発話の話題，タイプを予測し，認識対象候補とする。ただし，2番目の発話までは，一重遷移モデルによる予測を利用する。

図1に示すのは，発話状態の二重遷移を用いた場合の発話予測の例である。予めすべての発話状態から予測される次の状態に対応する文法のサブセットおよびそれに基づく単語予測LRテーブルを用意することは，大きな記憶スペースの確保を必要とする。そこで，実装上の工夫として，各発話状態での認識に用いる単語予測テーブルを共有化し，処理用の資源の節約を図った。

音響的なモデルとしてはHMMを採用しているが，本手法の考案時は，日韓自動通訳システムにおける準リアルタイムの対話音声認識処理を実現するため，処理時間の低減を最優先し，離散型のHMMを使用していた [3] [2]。また文献 [5] では，収録した模擬会話に現れる主要な表現をカバーすべく，文法を拡張した。さらに，より優れた音響モデルを使用した場合の発話状態予測の効果を測定する目的で，連続型のHMM(MFCC分析を使用) を用いた認識評価実験を行ない，上記の発話状態の一重遷移および二重遷移可能性に基づく発話予測手法の効果を測定した。認識対象は，無音室で収録した18〜25歳の男女各12名による，各12対話

[1]音声対話処理等について予備知識のない被験者24人による，80対話，全2383文からなる対話コーパス。被験者には予約内容や予約状況に関するメモをあらかじめ渡しておき，相手が話し終ってから発声を始めるように指示した以外は，自由発話に近い対話として収録された。ただし，実際にはある程度，発話の重なりが見られた。

[2]この時点では文法規則で記述する対象範囲も限定されていた。



(注) 発話タイプの説明
上の図中の発話タイプは次のような意味を持つ

EX: 謝意等の感情表現

SN: 否定的な陳述

CF: 確認　　02: 部屋に関する話題

図1: 発話状態の二重遷移可能性による予測の具体例

(計約300文) の発声データである。図2に示す対話例のように，認識対象文にはいわゆる不要語は含まれない。また，使用した認識文法は，これらのすべての対象文を受理可能な (一部として包含する) 規則集合である。次の表1に，離散型のHMMを用いた場合の認識率，認識時間を，また，表2に連続型のHMMを用いた場合の認識率，認識時間を示す[3]。

表1: 離散型のHMMの場合の認識率，認識時間

	話者情報のみ利用	一重遷移	二重遷移
1位認識率	46.0%	58.2%	62.8%
上位3位認識率	54%	67%	72%
上位10位認識率	60%	73%	79%
1文平均認識時間	8.7(sec.)	3.9(sec.)	3.3(sec.)

表2: 連続型のHMMの場合の認識率，認識時間

	話者情報のみ利用	一重遷移	二重遷移
1位認識率	76.4%	82.9%	83.5%
上位3位認識率	83%	90%	90%
上位10位認識率	87%	92%	92%
1文平均認識時間	12.6(sec.)	8.8(sec.)	8.2(sec.)

[3]なお，ここでいう認識率は完全一致正解の他かに意味的に等価な認識候補を含めたものである。

H: ありがとうございます。
H: こちらはホテルアイリスでございます。
C: お部屋の予約をお願いしたいんですが。
H: ありがとうございます。
H: お泊まりの日にちはいつでしょうか。
C: 12月10日，1泊でお願いします。
H: お部屋の方はいかがいたしましょうか。
C: ツインを，1部屋お願いします。
H: ツインにはスタンダードタイプとデラックスタイプがございますが。
C: スタンダードの方をお願いします。
H: ただいま確認いたします。
H: 少々お待ちください。
H: 申し訳ございません。
H: あいにくですがその日は，スタンダードの方が全室ふさがっております。
C: デラックスのお部屋はありますか。
H: ええ，デラックスの方お取りできます。
C: では，デラックスにしてください。
H: それでは，お客様のお名前をお願いいたします。
C: カタオカと申します。
H: カタオカ様。
H: フルネームでお願いいたします。
C: カタオカミヤコと申します。
H: ご連絡先の電話番号をお願いいたします。
C: 03，xxxx，xxxx です。
H: それでは繰り返します。
H: カタオカ様，12月10日，ご1泊でよろしいでしょうか。
H: お部屋はデラックスツインを1部屋でございますね。
H: お電話番号が，東京03，xxxx，xxxx でございますね。
C: はい，その通りです。
C: よろしくお願いいたします。
H: ご予約どうもありがとうございました。
H: それでは，お待ちいたしております。
C: どうもありがとうございました。

図 2: ホテル予約の対話例 (認識実験対象の一部)

この結果より，発話状態の予測による認識率の向上は，発話者情報を用いた場合，発話状態の一重遷移可能性を用いた場合，二重遷移可能性を用いた場合の順に向上することが分かる。しかし，離散型の HMM を使用した場合と比較して，連続型の HMM を用いた場合では，基本性能レベルが向上し，予測による効果はそれほど顕著なものではないことが明らかとなった。一方，認識時間の短縮については，両モデルの場合とも同様な効果が得られた。

4 認識実験結果の分析

4.1 認識率と文の複雑さの関係

発話者情報を用いる場合，発話状態の一重遷移モデルで予測する場合，二重遷移モデルで予測する場合のいずれにおいても，文ごとの認識率は最高で 100%，最低で 0%と大きな開きがあった。このことは，認識し易い文と認識の難しい文が存在することを示している。

- そうですか
- はい
- クドウです
- はい，けっこうです
- イトウ様
- そうです
- 少々お待ちくださいませ
- ツインでございますね
- ご予約どうもありがとうございました

図 3: 次発話を予測しなくても認識率が 100%の文

- 申し訳ないんですが，両日ともにツインは全室ふさがっております
- 4月の19日にシングルを一部屋お願いしたいんですが
- 4月の19日にシングルを一部屋でよろしいでしょうか
- 大変申し訳ございませんが，両日ともシングルは満室でございます
- シングルユースですと，ツインのお部屋をおとりできますが
- イノウエマリコさま，2月の8日のご宿泊で，お部屋がデラックスツイン
- クドウチエコ様，8月の28日に，シングル，ご一泊ですね
- あいにくですが，その日は，シングルのほうは全室ふさがっております
- ヨコヤマアユミ様，6月の2日と3日，2泊ですね

図 4: 二重遷移モデルで予測しても認識率が低い文

図 3 に認識しやすい文の例を，図 4 に 認識しにくい文の例を示す。図 3，図 4 の例を見てわかることは以下のとおりである。

1. 短い文は認識しやすい
2. 長い文は認識しにくい
3. 複雑な文は認識しにくい

3. について説明する。例えば，

- ヨコヤマアユミ様，6月の2日と3日，2泊ですね

という文を，文法で記述された単位ごとに区切ると次のようになる。

- [ヨコヤマ] - [アユミ] - 様 - [6] - 月の - [2] - 日と - [3] - 日 - [2] - 泊ですね

このうち，[　　]で囲ってあるところが，変数部分[4]であり，苗字と名前が文法中にそれぞれ100種類あるとすると「[ヨコヤマ]-[アユミ]」という部分までに100×100で10,000通りの候補があることになる。さらに，意味的に同じものも含めれば色々な場合が考えられる。このような文は，構文的複雑さも高く，特に変数部分では，一時的に分岐数が非常に大きくなる。そこで，認識率が0%である群と100%である群の構文的複雑さ，認識対象文の長さ[5]について調査した(表3)。

表3: 認識率と構文的複雑さ，文の長さについて

	予測なし		一重遷移モデル		二重遷移モデル	
認識率	0%	100%	0%	100%	0%	100%
静的平均分岐数	18.7		17.3	13.8	20.3	13.4
重み付き分岐数	1128.1		484.0	175.6	291.6	116.2
文の長さ	44.5	14.8	48.9	19.8	46.7	21.2

注) **重み付き分岐数**

$$WP = \sum_k P(k)N(k)$$

k : 節

$P(k)$: 文頭から k に到達する確率

$N(k)$: k から分岐している枝の数

重み付き分岐数は，文法全体の大きさと節ごとの複雑さを総合した尺度であり，値が大きいほど，複雑な文法ということになる。

どの場合においても，認識率の高い文は，構文的複雑さが低く，文の長さは短く，認識率の低い文は，構文的複雑さが高く，文は長かった。また，次発話を予測することによって，より長い文章を認識できるようになることが分かった。

次発話を予測しない場合においても100%の認識率になる文は，短い文であった。逆に言うと，たとえ，文法が構文的に複雑であったとしても(つまり，探索空間が広くても)，短く，単純な文であれば，認識は可能であるといえる。

4.2 発話状態の予測の認識率への寄与

認識候補として上位に現れた文の発話状態(発話タイプと話題の組合せ)の分布から，発話状態の予測自体がどの程度認識率に寄与しているかを分析する。

[4]文の中で構文上同じ役割を果たしている並列的な候補がたくさんある場合，その候補一つ一つをここでは，変数と呼ぶ．固有名詞などが変数になる．プログラミング言語や数学における変数の概念とは必ずしも一致しない．

[5]認識対象文のバイト数．音素数とほぼ一致する．

表4: 認識対象文内の発話状態の分布(最頻10位)

発話状態	頻度(割合)	認識正解率*
RP00	34 (10.9%)	98%
GC00	31 (10.0%)	99%
GO00	24 (7.7%)	100%
AC00	23 (7.4%)	100%
SP00	17 (5.5%)	67%
SN00	15 (4.8%)	77%
EX00	14 (4.5%)	99%
DA02	13 (4.2%)	97%
QP02	11 (3.5%)	80%
DA01	10 (3.2%)	78%
IN00	10 (3.2%)	99%
CF00	10 (3.2%)	33%
WS00	10 (3.2%)	88%

* 連続型HMM, 二重遷移モデルでの上位3以内

表5: 発話状態の予測が認識成功に寄与しなかった割合

予測モデル	A／B
話者情報のみ	8.5%／13%
一重遷移モデル	6.5%／8%
二重遷移モデル	6.3%／8%

A = 上位10位までの候補の発話状態のみが正解に一致

B = 上位10位以内に正解のない場合(表2参照)

まず，認識実験対象文における発話状態の分布は表4の通りである。表4において，(予測される)発話状態に対応する認識率を比較すると，発話状態CF(確認)，SP(肯定の陳述)，SN(否定の陳述)に属する発話の認識率が他と比較して，低下している。

また，予測された発話状態が上位10位までのすべての認識候補について一致して現れたたにもかかわらず，その中に正解候補がなかった場合は表5の通りである(連続型HMMを使用した場合)。

この表5を見ると，一重遷移および二重遷移のどちらの予測モデルの場合も，正解候補が10位までに得られなかった場合の大半(80%程度)が，発話状態の予測だけでは上位に正解候補が得られなかったことが分かる。これらの場合のほとんどは，前述したように変数的な認識対象を(多く)含む文であり，そのような対象の認識精度の向上には，ここで用いている発話状態の予測以外の効果的な手法(音響モデルの改善を含む)を加える必要がある。一方，認識候補の最上位に，正解の発話状態とは異なる状態に属する文がランクされ

た割合は，10位までに正解候補がない場合のうち，一重遷移モデルで27%(全体の2.2%)，二重遷移モデルで17%(全体の1.4%)であった。

このことから，今回用いた発話状態の遷移可能性に基づく予測モデルは，局所的な対話の流れに着目するだけで，正解認識候補を絞り込む上ですでに相当に有効であり，現在の枠組みでは，これをさらに向上させ得る余地はわずかであることが分かった。

5 認識精度向上のために利用可能な情報源

前章では発話状態の予測の効果を分析したが，ここでは，さらに認識精度を向上させるための手法の改良と，他の情報(知識)源の利用方法について検討する。

5.1 本手法の改良と問題点

認識候補の最上位(1〜3位)に，正解の発話状態とは異なる状態に属する文がランクされた典型例は，正解候補「はい」(発話状態RP = 肯定的な短い応答表現)に対して，同時に予測される発話状態SP (= 肯定的な陳述一般)に属する，「2日です」等の候補が上位に現れるような場合である。これについては，発話タイプSPに属する文の種類の多さが問題となっていることが分かった。すなわち，実際には依頼／意思表示等に該当しない中立的な肯定文がすべて含まれるため，SPの発話状態に属し得る文の数が非常に多くなる。そこで，SPを下位分類することが考えられるが，発話状態／タイプを分けたことによる効果と副作用を考慮する必要がある。すなわち，下位分類した各状態毎に，予測される次の発話状態を指定しなければならないが，本手法のように発話状態ラベル付きのコーパスを利用する場合は，そのデータ量の不足による偏りを補うため，小規模な遷移テーブル作成時にはそれほど問題にならなかった，人手によるチェックが困難となる。

5.2 対話に関する知識の利用法

本手法では，一重遷移モデルないし二重遷移モデルを用いて，発話状態の予測のみを行なっているが，対話に関する種々の知識を利用した認識性能の向上が考えられる。一つは単純な発話状態遷移モデルを拡張して，プラン認識等の手法で用いられているような，より大きな対話の構造を把握する方向に一歩近づくことである。しかし，処理の効率化をも目指す本手法本来の利点を生かすためには，文法記述との関連を明示的に与えることが望ましい。この点では，比較的浅い対話構造を利用する巌寺らの手法([4]等)が参考になる。

もう一つの選択肢は，タスク内の対話に関する履歴情報を利用して，非正解候補を棄却する方法である。これに関する典型例は，発話タイプCF(確認)に属する発話である。すなわち，確認の発話では確認すべき内容が一度以前に出現していることを利用して，それに関する情報フィルター的な処理を追加することが可能である。もっとも，このように音声を用いて直接内容の確認を行なうことの必然性については疑問が残る。対話内容の履歴が記録されていれば，それを指示するだけの複雑性の低い発話が認識されれば十分であるような対話インタフェースを，タスク依存で設計することが可能であろう。これは，やや極端な例であるが，これ以外にもタスク依存の比較的単純な知識を利用して，同様に認識候補の中から正しいものを絞り込む処理を，提案した手法と組み合わせる戦略が考えられる。

6 まとめ

本報告では，発話状態の一重遷移可能性および二重遷移可能性に基づく，局所的な対話モデルを用いた場合の対話音声認識率の向上について，2種類の音響モデルを用いて実験を行ない，十分な効果が得られることを示した。さらに，実験結果の分析を踏まえて，さらに音声認識精度を向上させるための手段を検討した。

謝辞

本研究の遂行に際してご指導，ご助言を頂いたKDD研究所の村上所長・浦野前所長，ならびに慶應義塾大学理工学部の中西教授・斎藤先生に感謝いたします。また，有益なコメントを寄せられた学会関係の方々に深謝いたします。さらに，システム作成の過程でお世話になった桶谷・関の両氏，評価実験等に際してご協力頂いた社内外の方々に，この場で厚くお礼申し上げます。

参考文献

[1] Nagata, M. and Morimoto, T.: "An Information-Theoretic Model of Discourse for Next Utterance Type Prediction", *Transaction of IPSJ*, Vol.35 No.6, pp.1050-1061, 1994.

[2] 鈴木・井ノ上・谷戸: "発話タイプの予測を用いた対話音声認識方式", 日本音響学会講演論文集, 1995.

[3] Suzuki, M. et al.: "A Prototype of a Japanese-Korean Realtime Speech Translation System", *EUROSPEECH '95*, 1995.

[4] 巌寺・石崎・森元: "対話構造の定量的評価", 情報処理学会全国大会, 1995.

[5] 松崎・鈴木・井ノ上・谷戸・斎藤・中西: "対話音声認識における次発話予測の効果", 96年信学会総合大会, 1996.

[6] 松崎克郎: "対話音声認識における次発話予測の効果", 慶應義塾大学修士論文, 1996.