音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理

駒谷和範† 河原達也†

音声対話システムにおいて音声認識誤りへの対処は不可欠であり，必要に応じてユーザに質問したり，ユーザを誘導したりできるような対話管理戦略が望ましい．本稿では，音声認識結果のスコアから発話内容に関する信頼度を計算し，それを用いてシステム側から効果的な確認・誘導を行う方法について述べる．音声認識器の 10-best 出力とそのスコアから内容語に対する事後確率を計算し，内容語に対する信頼度とする．これを用いて必要な場合にのみ効率良く確認発話を行うことができることを示す．また概念レベルでも，発話内容の意味カテゴリについて信頼度を計算し，内容語がうまく認識できなかった場合でも，適切にユーザの誘導を行う方法を提案する．これらを実装し，初心者 24 名から収録したホテル検索の発話データと ATIS，Communicator タスクの認識結果を用いて，その有効性を確認した．

Flexible Dialogue Management for Generating Efficient Confirmation and Guidance Using Confidence Measures of Speech Recognition Result

Kazunori Komatani† and Tatsuya Kawahara†

In order to realize a robust spoken dialogue system, it is inevitable to handle recognition errors, and it is desirable that the system can make confirmation and guidance if necessary. We present a method to realize mixed-initiative dialogue, in which the system makes confirmation and guidance using confidence measures (CMs) derived from speech recognition scores. We define word-level confidence measure as a posteriori probability using 10-best outputs of speech recognizer. Using this confidence measure, the system can efficiently make confirmation. We also define confidence measure of semantic categories, which makes effective guidance even when successful interpretation is not obtained. We evaluate our method on the hotel information system using data collected with 24 novice users, and also on ATIS and Communicator system.

1. はじめに

音声認識技術の向上を受けて，その応用である音声対話システムも実用化へ向けて研究が行われている[1),4),9),11),15),19)]．計算機で音声言語を扱う際には，音声認識に誤りが生じたり，ユーザがシステムの想定していない発話を行ったりするなどといった問題が生じる．これらの問題は，システムの受理できる語彙や文法の範囲を広げたとしても，計算機で人間の音声や言語を扱う場合には本質的に避けられないものである．したがって，実用可能な音声対話システムを構築するためには，音声認識の精度を高めるとともに，どのように音声認識誤りに対処する対話管理を行うかが重要な課題である．すなわち，音声認識誤りが生じた場合の対処や，さらなる音声認識誤りが生じないようにするための対話戦略が不可欠である．

頑健な音声対話システムを実現するための対話戦略の 1 つとして，すべてのやりとりをシステム主導で行うことが考えられる．たとえば列車時刻案内タスクのように，必要な項目（日時，出発地，目的地など）があらかじめ決まっているタスクでは，それらをシステムが順にたずねた後に逐次確認を行うことで，タスクを遂行することが可能である．しかしデータベース検索のようなタスクでは，ユーザによって必須である項目は異なる．このため，単純にシステムが順に項目を質問していく戦略では効率的にタスクを遂行することは不可能であり，任意の項目を任意の順番で入力できることが望ましい．

したがって，ある程度自由度のあるタスクで，効率的かつ頑健な音声対話システムを構築するには，ユーザに自由な発話を許しながらも，必要なときにはユーザへの質問やユーザの誘導を行う混合主導対話（mixed-initiative dialogue）[2),4),9)] が望ましい．混合主導対話

† 京都大学大学院情報学研究科知能情報学専攻
Graduate School of Informatics, Kyoto University

では，語彙外の発話などにより音声認識誤りが起こった場合には，必要に応じてシステム側から発話内容に対する確認を行ったり，さらなる認識誤りを防ぐためにユーザを誘導したりする．

対話システムの入力に誤りがある場合に直接/間接的に確認を行う戦略やその有効性に関して，文献 17) では数式を用いて，文献 10) では計算機どうしのシミュレーションを用いて示している．また文献 18) では，文献 17) をスロットフィリングタスクにおいて発展させている．これらの研究では，対話全体ですべての発話に対して一定の性能（平均音声認識精度）を仮定している．しかし，実際の対話管理においては，確認を行うかどうかは個々の発話に対して動的に決定されるべきである．実際，人間は相手の発話内容をうまく聞き取れなかった場合にのみ確認を行う．同様に，音声対話システムにおいても対話管理を行うための指標として，音声認識結果の各項目に対する信頼度が必要である．

音声認識結果の信頼度を内容語に対して計算し，その受理/棄却に用いることは自然な考え方であり，過去にもそのような研究は行われている[3]．本研究では，単語レベルだけでなく意味カテゴリレベルにおいても信頼度を定義・計算し，この 2 レベルの信頼度を用いて効率良く確認や誘導を行う対話管理法について述べる．さらに，対話における損失関数を定義することにより，信頼度に対するしきい値を最適に設定する方法も提案し，余分な確認の回数を抑えながら意味理解誤り率を削減できることを示す．

2. 音声認識結果の信頼度（CM）の計算

音声認識結果に対する信頼度（Confidence Measure，以下 CM）に関する研究は，音声認識の後処理として認識結果を受理/棄却する発話検証（utterance verification）の問題として行われている[5]．計算機による音声認識は，入力された音声に対して最も尤度の高い単語列を出力するという過程であるため，正しい認識結果と認識誤りとを判別するためには何らかの尺度が必要である．一般に発話検証では，小語彙では音節モデル，競合音素モデルと比較することが有効であるが，大語彙では他の候補（N-best）と比較することが有効であると知られている[12]．本研究では他の N-best 候補と比較する方法を用いる．

2.1 内容語に関する CM の計算

N-best 文とそのスコア（各 N-best 文に付与されている対数スケールの尤度）を用いて，内容語に関する CM を定義する．すなわち CM は文全体に対してではなく内容語ごとに求める．本研究では，予備実験の結果 $N = 10$ とした☆．

入力発話：「付帯施設にレストランのある宿」

i	認識結果	p_i
1	あー 施設 に レストラン の 加悦町	.24
2	あー 施設 に レストラン の 桂 の	.24
3	あー 施設 に レストラン の 上賀茂	.20
4	<g> 施設 に レストラン の 加悦町	.08
5	<g> 施設 に レストラン の 桂	.08
6	<g> 施設 に レストラン の 上賀茂	.06
7	あー 施設 に レストラン の カフェ	.05
8	<g> 施設 に レストラン の カフェ	.02
9	<g> 設備 を レストラン の 加悦町	.01
10	<g> 設備 を レストラン の 桂 の	.01

（<g> はフィラーモデルにマッチした部分）

CM_w	（内容語）	@	（意味カテゴリ）
1	レストラン	@	施設
0.33	加悦町	@	所在
0.33	桂	@	所在
0.25	上賀茂	@	所在
0.07	カフェ	@	施設

図 1 内容語の信頼度（CM_w）の計算例
Fig. 1 An example of confidence measure calculation.

(1) N-best 文の対数スケールの各スコア $score_i (1 \leq i \leq N)$ に定数 $\alpha (\alpha < 1)$ を乗じた後，指数化し，i 番目の文の事後確率 p_i を求める[3]．α はスムージング係数で，triphone モデルに対しては予備実験の結果 $\alpha = 0.05$ とした．

$$p_i = \frac{e^{\alpha \cdot score_i}}{\sum_{j=1}^{N} e^{\alpha \cdot score_j}}$$

(2) ある内容語 w が i 番目の文に含まれるとき $\delta_{w,i} = 1$，含まれないとき $\delta_{w,i} = 0$ とすると，入力音声に w が含まれていた確率 p_w は，

$$p_w = \sum_{i=1}^{N} p_i \cdot \delta_{w,i}$$

となる．この事後確率 p_w を内容語 w の CM（CM_w）とする．

具体例として，「付帯施設にレストランのある宿」という発話に対する認識結果の 10-best 文と，文ごとの事後確率（p_i），内容語の CM を**図 1** に示す．

2.2 意味カテゴリに関する CM の計算

内容語レベルのみでなく，概念レベルにおいても信頼度を定義する．これはユーザの意図を推定するのに有益である．

認識された内容語を含む各フレーズは，意味解釈部

☆ 複数の N に対して予備実験を行ったが，N を 10 以上に設定しても，認識結果とその信頼度はほぼ変化しない．これは $N = 10$ 程度で $e^{\alpha \cdot score_i}$ の値が十分に小さくなり確率の総和が収束するため，解析結果がほぼ変化しなくなるからである．

によりそのフレーズに対する意味カテゴリが付与される．意味カテゴリは各検索項目に対応するように定めた意味解析用文法により規定する．意味解析用文法は，有限状態オートマトン（FSA）で記述されている認識文法を，前もって検索項目ごとにわけておいたものである☆．ホテル検索タスクにおける意味カテゴリは「所在」「付帯施設」など 7 種類である．ここで，各フレーズに対する意味カテゴリの信頼度を定義するために，N-best の i 番目の文において，カテゴリ c を規定する意味解析用文法によって受理されたフレーズ $P_{c,i}$ が，カテゴリ c についてのフレーズである度合い $\beta_{c,i}$（最大値 1）を定義する．

まず，一般的な単語（数値など）からなるフレーズよりも，特定のカテゴリでしか現れない単語を含むフレーズの方が，意味カテゴリ c である度合いは高い．これを情報検索で用いられる idf 値（inverse document frequency）を用いて表す．ある単語 w_j の idf 値（$idf(w_j)$）は以下のように求める．

$$idf(w_j) = \log \frac{M}{df(w_j)}$$

M はカテゴリの総数で，$df(w_j)$ は単語 w_j が出現するカテゴリの数である．すなわち，特定のカテゴリにしか現れない単語の idf 値（$idf(w_j)$）は大きくなる．

次に,「京都府の」のように内容語と助詞のみからなる短いフレーズよりも,「所在が京都府の」のように項目名を指定したフレーズの方が,「所在」カテゴリに関して述べている度合いは高い．つまり，$\beta_{c,i}$ は該当フレーズが長い（含まれる単語数が多い）ほど大きくなるように定義する．

これらにより，i 番目の文中のカテゴリ c のフレーズ $P_{c,i}$ に含まれる単語の idf 値の和を求めることにより，そのフレーズがカテゴリ c である度合いとし，これをカテゴリごとに設定した値 γ_c で正規化したものを $\beta_{c,i}$（最大値 1）として定義する．

$$\beta_{c,i} = \frac{1}{\gamma_c} \sum_{w_j \in P_{c,i}} idf(w_j)$$

γ_c はカテゴリ c のフレーズに出現する $\beta_{c,i}$ の最大値を前もって計算しておいたものである．この $\beta_{c,i}$ と i 番目の文の事後確率 p_i の積を用いて，意味カテゴリ c の信頼度 CM_c とする．すなわち，

$$CM_c = \sum_{i=1}^{N} \beta_{c,i} \cdot p_i$$

とする．

図 2 の例において，ほぼすべての N-best 解に所在やホテルタイプに関するフレーズがあるのに，意味カテゴリ［所在］や［ホテルタイプ］の信頼度が高くないのは，各フレーズが内容語と一般的な助詞の 2 単語のみからなり，$\beta_{c,i}$ が小さいためである．一方，**図 3** の例では，シングル料金に関するフレーズは単語数が多いため $\beta_{c,i}$ が大きくなる．所在に関するフレーズも「にある」という単語の idf 値が大きいため，意味カテゴリの CM は高くなる．このように内容語だけでなく付属語の情報も考慮に入れて，意味カテゴリ CM を計算する．

3. 信頼度を用いた柔軟な対話戦略

3.1 内容語の CM を利用した確認発話

2.1 節で述べた内容語に関する信頼度を用いて確認発話を生成する．確認を行うことによって認識誤りを受理する回数を削減できるが，すべての発話に対して確認を行うのは冗長であり協調的ではない．したがって，音声認識誤りに対処するための確認は，認識結果が信頼できない場合にのみ行われるべきである．つまり，正解である可能性が高い場合にはそのまま受理し，誤りである可能性が高い場合には確認せずに棄却した方が，余分な確認により対話が冗長になるのを防ぐことができる．これを実現するために，2 つのしきい値 θ_1，$\theta_2(\theta_1 > \theta_2)$ を設定し，確認発話を以下の手順で生成する．

- $CM_w \geq \theta_1$ のとき
  そのまま受理する（確認は行わない）
- $\theta_1 > CM_w \geq \theta_2$ のとき
  直接的に確認を行う
  例：「○○でよろしいですか？」
- $\theta_2 > CM_w$ のとき
  棄却する

θ_1 は認識された内容語の候補をそのまま受理するか確認を行うかの信頼度のしきい値であり，θ_2 は確認を行うか棄却するかのしきい値である．信頼度 CM_w は内容語ごとに計算しているため，1 発話内に複数の内容語が含まれている場合には，その内容語ごとに受理/確認/棄却を決定できる．すべての内容語が棄却された場合は，再発話を促すことになる．

具体例を図 2 に示す．複数の内容語が認識されているが，$\theta_1 = 0.9$，$\theta_2 = 0.6$ とすると，まず $CM_w \geq \theta_1$

☆ 一フレーズが複数の意味解析用文法により受理されることもありうる．このような場合には後段の処理で曖昧性解消を行うことになる．

発話：「旅館を探してください」
正解：旅館@ホテルタイプ

i	認識結果
1	旅館 の 大阪市 で 下さい
2	旅館 の 大阪府 で 下さい
3	旅館 の 大阪市 で 下さい <g>
4	八鹿町 の 大阪市 で 下さい
5	旅館 の 三田市 で 下さい
6	旅館 の 泉南市 で 下さい
7	旅館 の 大阪市 です 下さい
8	旅館 の 堺市 で 下さい
9	<g> 大阪市 で 下さい
10	八鹿町 の 三田市 で 下さい

（<g> はフィラーモデルにマッチした部分）

CM_w	（内容語）@（意味カテゴリ）
0.93	旅館@ホテルタイプ
0.80	大阪市@所在
0.10	大阪府@所在
0.05	八鹿町@所在
0.05	三田市@所在
0.03	泉南市@所在
0.02	堺市@所在

CM_c	意味カテゴリ
0.50	所在
0.47	ホテルタイプ

図 2 複数の内容語が認識された場合の例
Fig. 2 An example of multiple content words recognized in a sentence.

発話：「京都市にあるシングル料金が 18,000 円以下の宿」
正解：京都市@所在，18000@シングル料金:以下

i	認識結果
1	京都市 にある シングル料金 が 18,000 円 以下 の 宿
2	京都市 にある シングル料金 が 16,000 円 以下 の 宿
3	京都市 にある シングル料金 が 17,000 円 以下 の 宿
4	京都市 にある シングル料金 は 18,000 円 以下 の 宿
5	京都市 にある シングル料金 の 18,000 円 以下 の 宿
6	京都府 にある シングル料金 が 18,000 円 以下 の 宿
7	京都市 にある シングル料金 は 16,000 円 以下 の 宿
8	京都市 にある シングル料金 の 16,000 円 以下 の 宿
9	京都府 にある シングル料金 が 16,000 円 以下 の 宿
10	京都府 にある シングル料金 が 17,000 円 以下 の 宿

CM_w	（内容語）@（意味カテゴリ）
0.87	京都市@所在
0.51	18000@シングル料金:以下
0.30	16000@シングル料金:以下
0.19	17000@シングル料金:以下
0.13	京都府@所在

CM_c	意味カテゴリ
1	所在
1	シングル料金:以下

図 3 意味カテゴリの信頼度による誘導が行われる例
Fig. 3 An example of getting high semantic attribute confidence in spite of low word confidence.

である「旅館@ホテルタイプ」が受理される．次に，$\theta_1 > CM_w \geq \theta_2$ である「大阪市@所在」に関して「所在が大阪市ですか？」のように確認を行う．これにはユーザが「いいえ」と応答することにより棄却される．これ以外の内容語も $\theta_2 > CM_w$ であるため確認されずに棄却される．結果として湧き出し誤りが棄却され，正解である「旅館@ホテルタイプ」のみを受理することになる．

3.2 意味カテゴリの CM を用いた誘導発話

人間は言われたことを完全に理解できなくても，部分的に理解できた場合には，それに基づいて聞き取れなかった部分を聞き返すことができる．たとえば，内容語がきちんと聞き取れなかった場合でも，相手の発話が所在に関するものであると分かれば，「すみません．所在はどこでしたか？」のように聞き直すことで，地名のみを丁寧に再発声するように暗黙に要求する．このように，内容語の候補が十分に高い信頼度で得られなかった場合でも，解釈した結果を単純に棄却してユーザに再発話を要求するのではなく，どのカテゴリに関する発話であるかを推察することによって，次のユーザ発話を導いたり，次発話を予測することで認識の対象を狭め，より高い精度で認識を行ったりすることが可能となる．

このような戦略が有効となる例を図 3 に示す．図 3 の例では，前節で述べた内容語に関する CM の処理に基づき，「所在が京都市ですか？」のような確認が行われる．その後，前節の枠組みではさらなる受理や確認は行われないが，意味カテゴリの信頼度を考慮に入れると，受理・確認を行った［所在］カテゴリ以外にも［シングル料金：以下］の信頼度（CM_c）が高いことを利用できる．これは人間の場合では，相手の発話が「シングル料金が……円以下」としか聞き取れなかった場合に相当する．このような場合には，単純に単語レベルの信頼度が低い残りの候補を棄却するよりも，「シングル料金がいくら以下ですか？」とユーザ発話を誘導する方が，ユーザは金額のみを単独で発声したりするので，次発話の認識が容易になる．上で述べたシステムの対話戦略の概要を**図 4** に示す．

さらに，意味カテゴリの CM が高いのに内容語の認識ができない状況が続く場合には，内容語が未知語である可能性がある．未知語が発話されている場合には，ユーザが何度発話しても正しく認識することはできないため，ユーザの発話をシステムの受理できる表現に変えてもらう必要がある．たとえば意味カテゴリ［所在］の信頼度が高い場合には，「都道府県名（あるいは市町村名）から指定してください．」とユーザの発話をシステムの語彙内に誘導し，さらなる認識誤りを防ぐことも可能である．

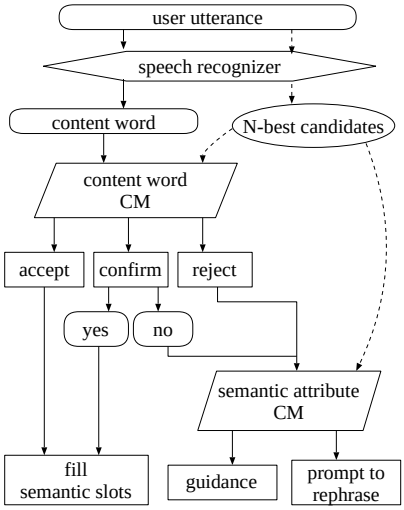

図 4 対話管理戦略の概要
Fig. 4 Overview of proposed dialogue strategy.

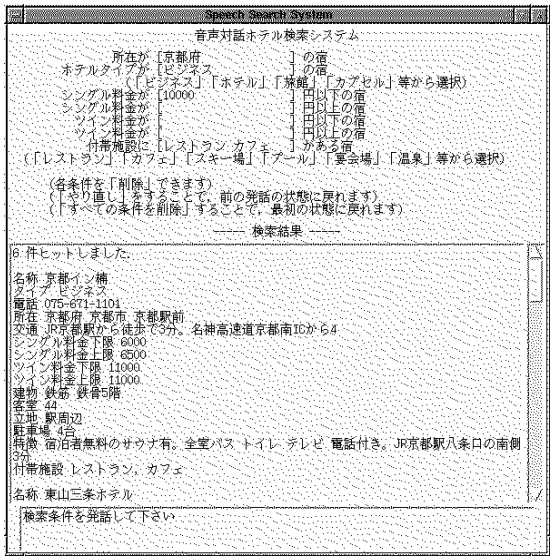

図 5 ホテル検索システムの GUI[20)]
Fig. 5 Outlook of GUI of hotel information system.

4. 評価実験

4.1 タスクと実験データ

提案する対話戦略をホテル検索をタスクとして評価実験を行った．用いたシステムは，関西地区の 2,040 件のホテルが収録されているデータベースを検索対象として,「所在」,「シングル料金上限」,「付帯施設」などの 7 項目について，任意の時点で任意の条件を追加・削除できるものである．また 1 発話で複数の検索項目を指定・削除することも可能である．

音声認識エンジンには，本研究室で開発された Julian[21)] を用いた．認識部の語彙数は 982，言語制約はフレーズ単位の繰返しであるような有限状態文法である．Julian は A*探索を行っているため，正しい N-best 解を効率良く求めることができる．文頭と文末にはフィラーモデル[13)] を導入して，発話中に含まれる検索条件指定部分以外の，ユーザのつぶやきや間投詞，雑音などとマッチングさせる．音響モデルは，3000 状態 16 混合分布の triphone[14)] である．入力はすべて音声で行っているが，ユーザへの出力には GUI[20)] を用いている．GUI の外観は**図 5** のようになっており，入力可能な項目や入力された検索条件，検索結果が表示される．

評価用データとして，音声対話インタフェースを使用したことのない被験者 24 名による発話を収集した．ユーザに対して，まず関西地区のホテル検索システムであることや検索可能な項目，項目の削除の方法などを教示したのち，希望の結果が得られるまで使用してもらった．GUI の画面には，入力されている条件とその検索結果が発話のたびに表示される．ユーザは GUI を見ながら条件の追加・削除を行った．この結果，全体で約 120 分間の音声データが得られた[16)]．この音声データを 1.25 秒以上の無音区間で区切り，雑音や息の音だけの部分を取り除いた結果，全部で 705 発話（約 29 発話/人，最大 64，最小 11）となった．その 705 発話に対して，書き起こし文と意味解釈結果を人手で付与し，正解とした．評価実験では収録した音声データの認識結果に対して信頼度を計算し，設定されるしきい値により正解と認識誤りを判別できるかを判定する．

少しの雑音や助詞の省略などを含めて，システムが受理可能であると考えられる発話は 581 発話で，全体の 82.4%であった．残りの 124 発話（17.6%）はシステムの想定外である発話（語彙外・文法外・タスク外・発話の断片など）である．以降の実験では，これらの想定外である発話を含むすべてのデータを用いて評価を行う．実際の音声対話システムでは，整った発話を正しく認識できるとともに，想定外である発話に対して誤動作を起こさないことや，部分的に解釈しその結果を利用して対話を続けることが必要となるからである．正解数は発話単位ではなく，内容語（＝意味スロット）を単位として数えた．全正解数は 804 である．

4.2 信頼度に対するしきい値の決定

収録したホテル検索タスクの音声データに対する，内容語の信頼度の分布（CM_w）を**表 1** に示す．CM_w が高い値を得ている範囲では，出力結果の適合率（正

表 1 認識結果の信頼度（CM_w）の分布（ホテル検索タスク）
Table 1 Distribution of CM_w (hotel task).

CM_w	出力数	正解数	適合率 (%)
0.0 - 0.1	158	2	1.3
0.1 - 0.2	39	2	5.1
0.2 - 0.3	25	2	8.0
0.3 - 0.4	24	1	4.2
0.4 - 0.5	20	6	30.0
0.5 - 0.6	29	10	34.5
0.6 - 0.7	20	9	45.0
0.7 - 0.8	27	13	48.1
0.8 - 0.9	39	19	48.7
0.9 - 1.0	137	110	80.3
1.0	530	455	85.8
合計	1,048	629	60.0

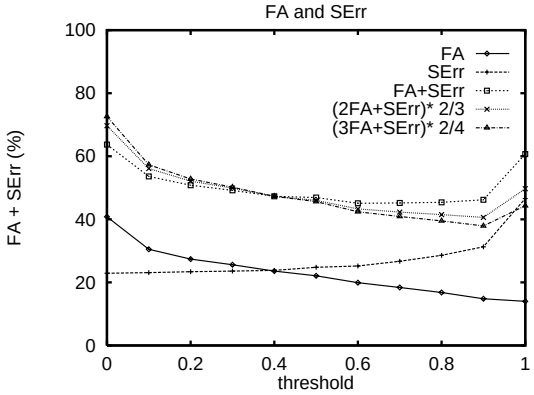

図 6 θ_1 に対する FA+SErr の変化（ホテル検索タスク）
Fig. 6 Operation curve of FA+SErr against threshold θ_1 (hotel task).

解数/出力数）も高くなっており，定義した CM_w が音声認識の信頼度として適切な尺度になっていることが分かる．この CM_w に関して，2 つのしきい値 θ_1, $\theta_2(\theta_1 > \theta_2)$ を定める．

本稿では，しきい値 θ_1，θ_2 の推定に，収集したすべてのデータを用いた．これは，CM の値の分布が主として文法と音響モデルに依存し，しかも θ_1 と θ_2 の変化が 0.1 刻み程度と細かくないので，closed test と open test でしきい値は大きく変化しないと考えられるためである．したがって，同一の文法や音響モデルを用いている限りは，新たなデータに対しても同等の性能での動作が可能である．

4.2.1 しきい値 θ_1 の決定

θ_1 は，受理と確認との境界を定めるしきい値であるため，認識誤りを誤って受理してしまう割合（誤受理率 False Acceptance；FA）と，正解が受理する候補に含まれていない割合（Slot Error；SErr）の両方を考慮して決定する．人手で与えた正解数を C，$CM_w \geq \theta_1$ で受理される候補の数を A，そのうちの正解数を CA とすると，FA および SErr は以下のように定義される．

$$FA = 1 - \frac{CA}{A}$$

$$SErr = 1 - \frac{CA}{C}$$

この FA と SErr は，それぞれ適合率（precision）と再現率（recall）の補数になっており，互いにトレードオフの関係にある．たとえば，しきい値 θ_1 を高く設定すると，A と CA はともに減少するが，CA の値が A の値に近づくため FA が減少し，一方で C は一定であるため SErr が増加する．

ホテル検索タスクのデータに対して，しきい値 θ_1 を 0.1 刻みで変化させたときの FA と SErr の値を**図 6** に示す．θ_1 が 0.6 から 0.9 の間で FA+SErr の値はほぼ一定だが，ここで実際の対話を考えると，誤って受理された場合には，ユーザはその項目をまず削除したうえで，さらに再入力しなければならない．これに対して，受理される範囲に入らない場合は，確認が行われるか，棄却されたとしても再発話すればよいため，誤って受理される方が対話全体に対する損失は大きい．したがって，$(\lambda \cdot \mathrm{FA} + \mathrm{SErr}) \times 2/(1+\lambda)$（$\lambda \geq 1$）のように FA に重みを与え，これを損失関数と定義した．λ を種々の値に変えた結果，図 6 のとおり θ_1 は一貫して 0.9 で最小値となることから $\theta_1 = 0.9$ とする．

4.2.2 しきい値 θ_2 の決定

θ_2 は，$CM_w \geq \theta_1$ で受理されなかった残りの候補の集合に対して，確認と棄却との境界を定めるしきい値である．したがって，正しい候補を棄却してしまう割合（誤棄却率 False Rejection；FR）と，誤った内容に関して確認を行う割合（$\theta_1 > CM_w \geq \theta_2$ での誤受理率 conditional False Acceptance；cFA）を考慮して決定する．$0 < CM_w < \theta_1$ の範囲内での正解数を C'，$0 < CM_w < \theta_1$ の認識結果に含まれる候補数を O'，$\theta_2 \leq CM_w < \theta_1$ で確認されることになる候補の数を A'，そのうちの正解数を CA' とすると，cFA および FR は以下のように定義される．

$$cFA = 1 - \frac{CA'}{A'}$$

$$FR = \frac{C' - CA'}{O' - A'}$$

しきい値 θ_2 の値が小さいほど誤って棄却される候補の割合（FR）は減少するが，誤った内容に関して確認を行う割合（cFA）が増加することになる．FR+cFA を損失関数とし，ホテル検索タスクにおけるその値の変化を**図 7** に示す．ここで最小値を示している $\theta_2 = 0.6$

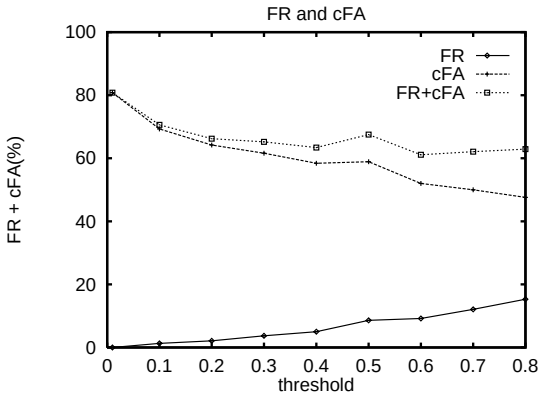

図 7 θ_2 に対する FR+cFA の変化（ホテル検索タスク）
Fig. 7 Operation curve of FR+cFA against threshold θ_2 (hotel task).

表 2 ホテル検索タスクにおける従来手法との比較
Table 2 Comparison of methods (hotel task).

	FA+SErr	FA(%)	SErr(%)
第 1 候補をそのまま受理	48.6	24.7	23.9
CM で受理，確認なし	44.1	19.5	24.6
CM で受理，確認あり	39.9	15.3	24.6

とする．

4.3 従来手法との比較

音声認識結果の第 1 候補をそのまま受理する従来の方法と，本研究で提案する手法の誤り率を比較した．結果を**表 2** に示す．誤り率は FA+SErr で比較している．ただし，FA は挿入誤り数および置換誤り数，SErr は削除誤り数および置換誤り数を含んでいるため，置換誤りを二重に計数していることになる．

表中の「確認なし」はしきい値 θ を 1 つだけ設定し，θ 以上は受理，それに満たない候補は棄却する場合である．FA+SErr が最小となるときの FA(θ)+SErr(θ) の値を示してある（$\theta = 0.6$）．「確認あり」は $\theta_1 > CM_w \geq \theta_2$ で確認を行い，それが誤りなく受理/棄却されると仮定して，FA(θ_1)+SErr(θ_2) の値を求めている．ここでは $\theta_1 = 0.9$，$\theta_2 = 0.6$ である．なお確認に対する応答（肯定・否定）は誤りなく認識できると仮定している．CM_w を計算して候補を選択的に受理することにより，SErr は若干増えるが FA が大幅に削減され，確認を行うことによってさらに誤りが 4.2 ポイント削減されている（表 2）．

確認発話を行うと全体の発話数は増加するが，しきい値 θ_2 を設定することにより，無駄な確認，すなわち誤った候補に対する確認の生成を抑えることができる．θ_2 を設定しない場合には，信頼度が θ_1 以下で受理されない候補に対しては，すべてを確認するか，すべてを棄却することになる．**表 3** に示すように，θ_1 以下のすべての候補に対して確認を行うと，大半が誤りの候補への確認となる．一方すべての候補を棄却する場合には表 1 に示されるように，正解が多く含まれている部分（$0.9 > CM_w \geq 0.6$）を棄却することになるため SErr が増加し，正解を認識しているにもかかわらずユーザに再発話を求める回数が増加してしまう．

表 3 θ_2 設定の効果（ホテル検索タスク）
Table 3 The effect of setting θ_2 (hotel task).

	出力数	正解数	適合率
$\theta_1 > CM_w > 0$	400	68	17%
$\theta_1 > CM_w \geq \theta_2$	102	49	48%

（ただし，$\theta_1 = 0.9$，$\theta_2 = 0.6$）

これに対して θ_2 を設定した場合には，ホテル検索タスクでは確認する 102 個のスロットのうち約半数（48%）が正しい内容語を含んだものとなる．これは，正しいか誤りかが識別しにくい候補に対してのみ確認を行い，正誤がはっきりしている候補に対しては確認を行っていないことを示している．これにより，必要以上に発話数を増加させることなく，効果的な確認が行われていることが分かる．

4.4 他タスクにおける評価

提案手法の有効性と一般性を確認するために，ATIS[6] と Communicator[8] の両タスクで評価を行った．これらの実験においてはベル研究所の認識システムの結果を利用した．

まず，ATIS-3 コーパスの中から各ユーザの最初の発話のみを対象とした．発話数は 669 である．意味スロットの抽出には，統計的手法により自動的に学習した FSM（finite state machine）パーザ[7] を用いている．音声認識部の語彙数は 1047 で，テストセットには未知語も含まれる．ATIS タスクでは大量のコーパスが利用可能で，タスクに依存した音響モデルおよび統計的言語モデルを構築できるため，音声認識率はホテル検索タスクよりも良い．結果を**表 4** に示す．「確認なし」の場合のしきい値は $\theta = 0.4$,「確認あり」の場合の 2 つのしきい値は $\theta_1 = 0.7$，$\theta_2 = 0.4$ となり，ホテル検索タスクよりも小さい値となった．ATIS タスクでは，認識誤りがホテル検索タスクと比較して少ないため，確認を行わない場合ではほぼ誤り率は変化しないが，確認を行う戦略と組み合わせることにより誤り率は 1.7 ポイント減少する（表 4）．

さらに，開発途上の DARPA Communicator システムを用いた評価を行った．プロトタイプシステムにより電話を通じて収集された 1,395 発話を対象としている．しきい値は $\theta_1 = 0.8$，$\theta_2 = 0.6$ となった．結

表 4 ATIS タスクにおける従来手法との比較
Table 4 Comparison of methods (ATIS).

	FA+SErr	FA(%)	SErr(%)
第 1 候補をそのまま受理	10.9	3.7	7.3
CM で受理，確認なし	11.1	4.1	7.0
CM で受理，確認あり	9.2	2.2	7.0

表 5 Communicator タスクにおける従来手法との比較
Table 5 Comparison of methods (Communicator).

	FA+SErr	FA(%)	SErr(%)
第 1 候補をそのまま受理	39.7	19.4	20.3
CM で受理，確認なし	36.8	15.1	21.7
CM で受理，確認あり	34.0	12.3	21.7

果を**表 5** に示す．このタスクでは提案手法による有効性がホテル検索タスクの場合と同様に示されている．

4.5 意味カテゴリ推定の有用性

ホテル検索タスクにおける内容語の CM と意味カテゴリの CM に関する FA+SErr の分布を**図 8** に示す．これより，内容語の CM よりも意味カテゴリの CM の方が，全般に精度が良いことが分かる．これは，内容語の CM で十分に信頼度の高い候補が得られない場合でも，意味カテゴリの CM から有用な情報が得られる場合がありうることが示されている．

ホテル検索タスクで $\theta_1 = 0.9$, $\theta_2 = 0.6$ とした場合，単語レベルでは正しい認識結果が得られなかったスロット（$\theta_2 > CM_w$ で棄却されたスロットと，$\theta_1 > CM_w \geq \theta_2$ で確認されたが回答が「いいえ」となるスロット）の合計は 142 個であった．つまりこれらのスロットは，単語レベルの信頼度だけでは単純に棄却されるものである．このような場合，意味カテゴリの信頼度 CM_c を求めて，単語レベルの信頼度 CM_w と同様にしきい値を設定する．つまり，CM_c が 1 の場合にはそのまま誘導を行い，$1 > CM_c \geq 0.5$ の場合にはその意味カテゴリを確認してから誘導を行う．たとえば,「シングル料金が，んと，12,000 円以下で.」という発話に対して，単語レベルでは信頼度の高い候補が得られなかった場合でも意味カテゴリが「シングル料金：以下」であるという信頼度が高い場合には，単純に棄却して再発話を促すのではなく，システム側から「シングル料金がいくら以下ですか？」のような発話を行う．これにより，単語レベルでは候補が得られていなかった前述の 142 個のスロットの 17%に対して，有効な誘導発話を行える．この数字は低いようにみえるが，従来は完全に棄却されていたユーザ発話に対して有効な誘導が行えており，有意義であるといえる．

ATIS タスクにおいては，単語レベルの信頼度では受理されなかった 66 個のスロットのうち 40 個（61%）に対して，意味カテゴリを推定できた．この中には，出発時間を表す [FROM_TIME] の代わりに単純に [TIME] が認識されているといった不十分なものも含まれるが，これらの情報を後段の対話管理部で用いれば有用な応答が可能となる．

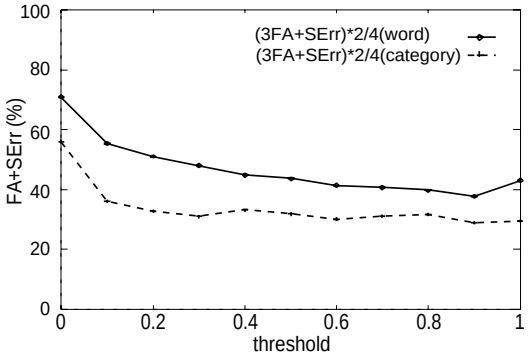

図 8 内容語 CM と意味カテゴリ CM の性能の比較（ホテル検索タスク）
Fig. 8 Performance of word-level and concept-level CMs (hotel task).

5. ま と め

音声認識誤りに頑健な対話システムを実現するために，音声認識結果に対して内容語と意味カテゴリの 2 レベルで信頼度（Confidence Measure; CM）を定義し，それを用いた対話管理の方法について述べた．

対話管理に用いる信頼度のしきい値は，対話における損失関数を定義し，それを最適化することによって定めた．意味理解における 2 種の誤り率の和は，CM を用いることにより，ホテル検索タスクでは確認を行わない場合で 4.5 ポイント，確認を行う場合では 8.7 ポイントの向上が得られた．音声認識誤りの少ない ATIS タスクでも，CM を用いて確認を行うことにより 1.7 ポイント改善し，Communicator タスクでも 5.7 ポイントの向上が得られた．確認発話は最適化されたしきい値を用いることにより，正しい候補と誤った候補が約半数ずつとなる部分に対してのみ行われており，無駄な確認による発話数の増加は抑えられている．さらに，意味カテゴリに関する信頼度を用いることで，単語レベルの CM では棄却されていた語彙外の発話などを含むスロットに対しても，有効な誘導発話を生成できることを示した．

参 考 文 献

1) Allen, J., Miller, B., Ringger, E. and Sikorski, T.: A Robust System for Natural Spoken Dialogue, *Proc. ACL*, pp.62–70 (1996).

2) Bennacef, S., Devillers, L., Rosset, S. and Lamel, L.: Dialog in the RAILTEL Telephone-Based System, *Proc. ICSLP* (1996).
3) Bouwman, G., Sturm, J. and Boves, L.: Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the ARISE Project, *Proc. ICASSP* (1999).
4) Goddeau, D., Meng, H., Polifroni, J., Seneff, S. and Busayapongchai, S.: A Form-Based Dialogue Manager for Spoken Language Applications, *Proc. ICSLP* (1996).
5) Kawahara, T., Lee, C.-H. and Juang, B.-H.: Flexible Speech Understanding Based on Combined Key-Phrase Detection and Verification, *IEEE Trans. on Speech and Audio Processing*, Vol.6, No.6, pp.558–568 (1998).
6) Pieraccini, R., Levin, E. and Lee, C.-H.: Stochastic representation of conceptual structure in the ATIS task, *Proc. 4th Joint DARPA Speech and Natural Language Workshop* (1991).
7) Potamianos, A. and Kuo, J.: Statistical Recursive Finite State Machine Parsing for Speech Understanding, *Proc. ICSLP* (2000).
8) Potamianos, A., Ammicht, E. and Kuo, H.-K.J.: Dialogue Management in the Bell Labs Communicator System, *Proc. ICSLP* (2000).
9) Sturm, J., Os, E. and Boves, L.: Issues in Spoken Dialogue Systems: Experiences with the Dutch ARISE System, *Proc. ESCA IDS'99 Workshop* (1999).
10) Watanabe, T., Araki, M. and Doshita, S.: Evaluating Dialogue Strategies under Communication Errors using Computer-to-Computer Simulation, *Trans. IEICE, Info & Syst.*, Vol.E81-D, No.9, pp.1025–1033 (1998).
11) 伊藤敏彦，小暮　悟，中川聖一：協調的応答を備えた音声対話システムとその評価，情報処理学会論文誌，Vol.39, No.5, pp.1248–1257 (1999).
12) 河原達也：音声でスライド画面を操作する，bit 4月号，共立出版 (2000).
13) 河原達也，石塚健太郎，堂下修司：発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化，情報処理学会論文誌，Vol.40, No.4, pp.1491–1498 (1999).
14) 河原達也，李　晃伸，小林哲則，武田一哉，峯松信明，伊藤克亘，山本幹雄，山田　篤，宇津呂武仁，鹿野清宏：日本語ディクテーション基本ソフトウェア（98年度版），日本音響学会論文誌，Vol.56, No.4, pp.255–259 (2000).
15) 中野幹生，堂坂浩二，宮崎　昇，平沢純一，田本真詞，川森雅仁，杉山　聡，川端　豪：TV番組の録画予約を受け付ける実時間音声対話システム，情報処理学会研究報告，98-SLP-22-8 (1998).
16) 安達史博，駒谷和範，河原達也：音声対話情報検索システムにおける想定外の発話の分析とその対処，人工知能学会研究会資料，SIG-SLUD-A001-2 (2000).
17) 新美康永，小林　豊：音声認識の誤りを考慮した対話制御方式のモデル化，情報処理学会研究報告，95-SLP-5-7 (1995).
18) 新美康永，西本卓也，荒木雅弘：確認対話の制御方式の効率と音声認識システムの性能との関係，情報処理学会研究報告，99-SLP-27-17 (1999).
19) 竹林洋一：音声自由対話システム TOSBURG II—ユーザ中心のマルチモーダルインタフェースの実現に向けて，電子情報通信学会論文誌，Vol.J77-D-II, No.8, pp.1417–1428 (1994).
20) 田中克明，河原達也，堂下修司：汎用的な情報検索音声対話プラットフォーム，電子情報通信学会技術研究報告，SP98-109, NLC98-45 (1998).
21) 李　晃伸，河原達也，堂下修司：文法カテゴリ対制約を用いた A*探索に基づく大語彙連続音声認識パーザ，情報処理学会論文誌，Vol.40, No.4, pp.1491–1498 (1999).

(平成12年12月21日受付)
(平成14年 9 月 5 日採録)

駒谷　和範（学生会員）

1998年京都大学工学部情報工学科卒業．2000年同大学院情報学研究科知能情報学専攻修士課程修了．現在，同大学院博士後期課程に在学中．言語処理学会会員．

河原　達也（正会員）

1987年京都大学工学部情報工学科卒業．1989年同大学院修士課程修了．1990年同博士後期課程退学．同年京都大学工学部助手．1995年同助教授．1998年同大学情報学研究科助教授．現在に至る．この間，1995年から96年まで米国ベル研究所客員研究員．1998年からATR客員研究員．1999年から国立国語研究所非常勤研究員．2001年から科学技術振興事業団さきがけ研究21研究者．音声認識・理解の研究に従事．京大博士（工学）．1997年度日本音響学会粟屋賞受賞．2000年度情報処理学会坂井記念特別賞受賞．情報処理学会連続音声認識コンソーシアム代表．電子情報通信学会，日本音響学会，人工知能学会，言語処理学会，IEEE各会員．