音声言語情報処理 30－9
（2000．2．5）

音声認識結果の信頼度を用いた頑健な混合主導対話の実現法

駒谷 和範　　河原 達也

京都大学 情報学研究科 知能情報学専攻
komatani@kuis.kyoto-u.ac.jp

概要

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。頑健な音声対話システムを実現するためには、音声認識誤りへの対処は不可欠であり、それに加えて、必要なときにはユーザに質問したりユーザを誘導できるような対話管理戦略が望ましい。そこで、音声認識器の 10-best 出力とそのスコアから部分文に対する事後確率を計算し、その部分文に対する信頼度とした。これを用いて内容に関する確認発話を適切に行うことができる。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することにより、単語の認識がうまくいかなかった場合でも、適切な発話の誘導を行う。これらの評価実験をホテル検索をタスクとして初心者 24 名の発話を収録して行い、その有効性を確認した。

A Robust Mixed-Initiative Dialogue System using Confidence Measures of Speech Recognition Results

Kazunori Komatani　　Tatsuya Kawahara

Graduate School of Informatics, Kyoto University

Abstract

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. In order to realize a robust spoken dialogue system, it is inevitable to handle recognition errors, and it is desired for the system to make confirmation and guidance if necessary. We calculate confidence measures by a posteriori probability with 10-best outputs of speech recognizer, and the system can make effective confirmation. We also calculate CMs of semantic categories of utterances, which makes effective guidance utterances even when the word-level CMs are unreliable. Evaluations on the hotel retrieval task shows the effectiveness of this method, which are collected from 24 novice users.

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の研究が行なわれている。本研究ではホテル検索をタスクとして、音声対話を通じて情報検索を行なうことを目標としている。

計算機と音声で対話を行う際には、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なうなどといった問題が頻繁に生じる。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た

め、その対処は不可欠である。現在実用的に広く使われている音声対話システムが存在しないのは、この頑健性の欠如が大きな原因の一つであると考えられる。

したがって、実際の話し言葉を対象とした対話システムを構築するには、基本的に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に質問したりユーザを誘導したりする混合主導対話 (mixed-initiative dialogue) を、認識誤りを考慮した上で実現する必要がある。

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、[1] では数式を用いて、[2] では計算機同士のシミュレーションを用いて示している。また [3] では、[1] をスロットフィリングタスクにおいて発展させている。

本稿では、実際の対話音声とその認識結果から信頼度を計算し、それを用いて確認や誘導を効果的に行うことで、頑健に対話を進めていく方式及びその評価について述べる。2 章では、実際に混合主導対話を行うのに用いる音声認識結果の信頼度 (Confidence Measure: CM) の計算法について述べる。次に 3 章では、この CM を用いて認識誤りに対処する対話管理に関して述べる。4 章では、CM と対話管理戦略の有効性に関して行った評価実験について報告する。

2 音声認識結果の信頼度（CM）の計算

「よく聞き取れなかった言葉について確認を行なう」ということは、人間同士の対話でもよく行なわれることである。したがって、音声認識の結果が「認識誤りである可能性が高い」とわかることは、確認を行う戦略を考える上で非常に有用である [4]。しかし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プロセスであるため、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。そこで、この章では認識結果に対する信頼度（Confidence Measure: CM）を計算する方法についてまず考える。

2.1 単語に関する CM の計算

音声認識では、入力音声に対して尤度の高い順に n-best 解を求めることができる。本研究では、認識エンジンとして本研究室で開発された Julian[5] を用いており、解ごとにスコアが計算される。そこで、この n-best 解のスコアを用いて、単語ごとの CM を求める。本研究では、$n = 10$ とした。

1. n-best 解の対数スケールの各スコア $score_i (1 \leq i \leq n)$ から、最尤解のスコア $score_1$ をひいたものに、定数 $\alpha (\alpha < 1)$ をかける。

$$scaled_i = \alpha \cdot (score_i - score_1)$$

2. 1. で求めた $scaled_i$ を対数スケールから元に戻し、i 番目の解の事後確率 p_i を求める [6]。

$$p_i = \frac{e^{scaled_i}}{\sum_{i=1}^{n} e^{scaled_i}}$$

3. ある単語 w が i 番目の解に含まれるとき $\delta_{w,i} = 1$ とすると、w が正しい確率 p_w は、

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

で求める。

この p_w を単語 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」という発話の認識結果と、その内容語 (content word) の CM を計算したものを図 1 に示す。

2.2 意味カテゴリに関する CM の計算

認識された各内容語には意味カテゴリを付与している。これは、有限状態オートマトン (FSA) で記述されている認識文法を意味カテゴリごとにわけておき、どの FSA に受理されたかによって判定したものである。本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。

この意味カテゴリ (Concept Category) に関しても CM を求める。まず単語の CM の場合と同様に、n-best 解の i 番目の解の事後確率 p_i を求める。ある意味カテゴリ c の内容語が i 番目の解に含まれるとき $\delta_{c,i} = 1$ とすると、c が正しい確率 p_i は、

$$p_c = \sum_{i=1}^{n} p_i \cdot \delta_{c,i}$$

で求める。この p_c を意味カテゴリの CM(CM_c) とする。この CM_c は誘導発話の生成の際に用いる（3.3.2 節）。

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

ただし、<g>: filler model

CM_w	(単語)	@	(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: 単語の信頼度 (CM_w) の計算例

3 音声認識誤りに頑健な対話の実現法

3.1 フィラーモデルの導入

入力音声には、検索条件の指定以外にも、ユーザのつぶやきや間投詞、雑音などが含まれる。これらが認識文法中の単語に誤認識されると、システムの誤作動を引き起こす原因となる。

そこで、文頭と文末にフィラーモデル [7] を導入して、これらとマッチングさせる。

3.2 確認発話の生成

3.2.1 内容語の CM を利用した確認発話

2.1 節で述べた内容語 (content word) に関する信頼度を用いて確認発話を生成する。2 つのしきい値 $\theta_1, \theta_2 (\theta_1 > \theta_2)$ を設定すると、確認発話は以下のようになる。内容語に関する信頼度 CM_w は内容語ごとに計算しているので、一発話内に複数の内容語が含まれている場合でも、その内容語ごとに受理／確認／棄却を決定することができる。したがって、全ての内容語が棄却された場合に、再発話をうながすことになる。

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

しきい値 θ_1, θ_2 は、誤受理率 (False Acceptance) や誤棄却率 (False Rejection) を考慮して、実際のデータから決定した。これについては 4.2 節で述べる。

3.2.2 対話レベルの知識を利用した確認発話

情報検索というタスクにおいては、検索条件を追加／削除しながら対話が進んでいくため、このような対話の進行から大きく離れた発話は誤りである可能性が高い。その一例として、検索条件がすでに入力されている項目に対して、さらに上書きを行なうような発話は、(CM が高いとしても) 認識誤りである可能性が考えられる。実際、短い地名の湧き出し誤りでは、音響的に CM が高くなってしまうことがある。このようなものに対しても確認を行うことにより、誤受理率を抑えることが期待できる。

3.3 誘導発話の生成

対話においては、ユーザの発話内容に関して確認を行うだけではなく、ユーザの発話を誘導することも必要となる。

3.3.1 ユーザの練度の推定

システムを使い慣れないユーザに対しては、検索できる項目を提示することが有用である場合がある。したがって、無音（沈黙）が一定以上続く場合 [8] には、システム側から検索可能な項目についての誘導を行うことも考えられる。

3.3.2 意味カテゴリの CM を用いた誘導発話

内容語 (content word) の CM_w が低い場合でも、意味カテゴリの CM_c が高い場合にはその推定された概念に基づいて、誘導発話を生成することができる。その例を図 2 に示す。

発話：「所在が大阪府の宿」
正解：大阪府@所在
認識結果：
1: 所在 が ポートアイランド の <g>
2: 所在 が ポートアイランド の <g>
3: 所在 が 大阪府 の <g>
4: 所在 が 大阪府 の <g>
5: 所在 が 大阪市 の <g>
6: 所在 が 大阪市 の <g>
7: 所在 が 岡崎 の <g>
8: 所在 が 岡崎 の <g>
9: 所在 が 大原 の <g>
10: 所在 が 大原 の <g>

category [所在]　CMc = 1
ポートアイランド@所在　CMw = 0.38
大阪府@所在　CMw = 0.30
大阪市@所在　CMw = 0.13
岡崎@所在　CMw = 0.11
大原@所在　CMw = 0.08

図 2: CM を用いた誘導発話が有効な実例

図 2 の例では、内容語に関する CM がどれもしきい値（θ_2）よりも低いため、受理や確認を行うことにはならないが、意味カテゴリ [所在] の CM が高い。このような場合には、単純に入力を全て棄却して「もう一度言ってください。」と言うよりも、「所在がどこですか？」とユーザ発話を誘導する方が、次発話の認識が容易になり、より頑健に対話を進めることができる。

また、意味カテゴリの CM が高いのに内容語の認識が出来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高いと推測できる。例えば、未知の地名を発声されたと推定できる場合には、「都道府県名（あるいは市町村名）から指定してください。」とユーザ発話を誘導することも出来る。

なお、さらに認識誤りが続く場合には、ユーザが全く受理できないことを言っている可能性も考えられるので、システムの使い方や対象としているタスクについての説明を出力することも有効であると考えられる。

4 評価実験

4.1 実験データ

音声対話システムを使ったことのない 24 名の話者に対して、関西地区のホテル検索システムであるこ

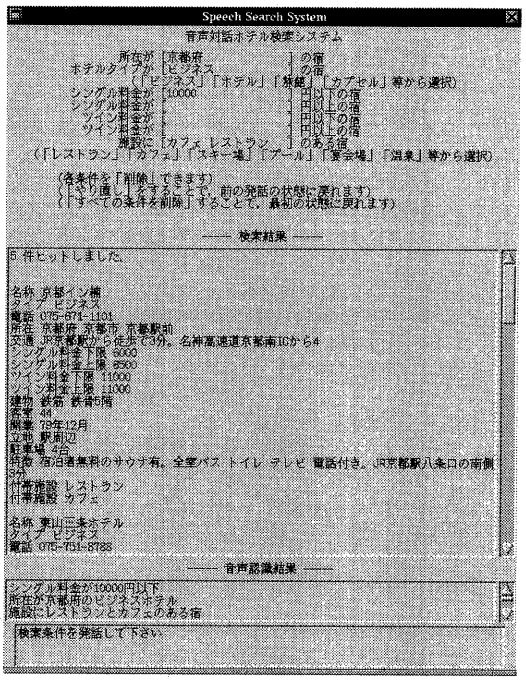

図 3: ホテル検索システムの GUI[9]

とや検索可能な項目、項目の削除の仕方などを教示し、全体で約 120 分間、GUI[9]（図 3）の付いたシステムを使用してもらった音声を収録した。その音声データを 1.25 秒毎に区切り、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。

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。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）は 124 発話存在した。

4.2 確認発話のしきい値の決定

3.2 節では、単語に関する CM にしきい値 $\theta_1, \theta_2 (\theta_1 > \theta_2)$ を設定することでシステムの応答戦略を定めた。このしきい値を 4.1 節のデータを全て用いて定める。また、正解数は発話単位ではなく、内容語（＝スロット）を単位として数えている。全

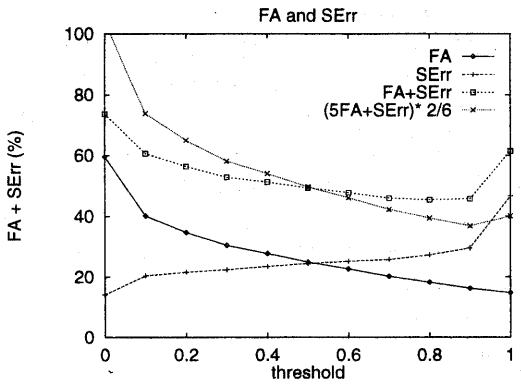

図 4: θ_1を変えたときの FA+SErr の変化

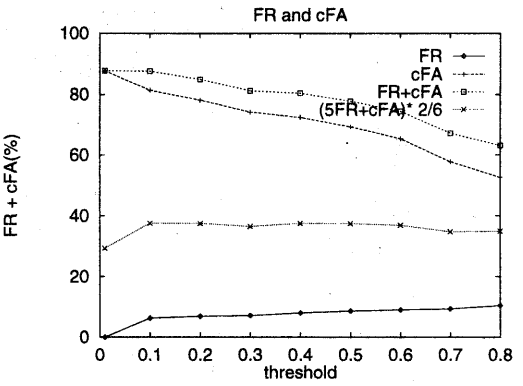

図 5: θ_2を変えたときの FR+cFA の変化

正解数は 804 である。

θ_1について

θ_1は、$CM_w > \theta_1$でそのまま受理するという境界を定めるしきい値なので、認識誤りを誤って受理してしまう率（誤受理率 (False Acceptance; FA)）と、正解が受理する部分に含まれていない率 (Slot Error; SErr) のバランスがとれている必要がある。

$$FA = \frac{\text{受理した中で誤っていたスロット数}}{\text{受理したスロット数}}$$

$$SErr = 1 - \frac{\text{受理した正解スロット数}}{\text{実際の正解スロット数}}$$

誤って受理された場合には、その項目を削除してから再入力する必要が生じる。これに対して、ここで受理される範囲には入らない場合は、確認が行われるか、棄却されて再発話することになるため、実際の対話の状況から考えると、誤って受理される方が問題が大きい。図 4 に、0.1 刻みでしきい値（θ_1）を変化させたときの FA+SErr の値と、FA に対して重みを与えた場合を示す。この結果から $\theta_1 = 0.9$ が最適であることが分かる。

θ_2について

θ_2は確認を行う境界を定めるしきい値で、CM_wがθ_1との間となる内容語に関して確認を行う。したがって、認識出来ているのに棄却してしまう率（誤棄却率 (False Rejection; FR)）と、確認を行う範囲（$\theta_2 < CM_w \leq \theta_1(=0.9)$）での誤受理率 (conditional False Acceptance; cFA) のバランスを考えなければならない。

$$FR = \frac{\text{誤って棄却したスロット数}}{\text{棄却したスロット数}}$$

この FR+cFA の値の変化を図 5 に示す。誤受理の場合は確認に対して「いいえ」と答えるだけで済むが、誤棄却の場合には再発話する必要が生じるため、誤棄却率に重みを加えた。図 5 からは有意に最適となる点は見つからないが、重みを与えたときに極小値を示している$\theta_2 = 0.7$を、以降θ_2として用いる。

表 1: 1-best 解のみを用いた手法との比較

	1best 解のみ	確認無	確認有
FA+SErr(%)	51.5	45.7	41.9

4.3 1-best 解のみを用いる場合との比較

従来の最尤解を認識結果とする方法と本手法の精度を比較した。その結果を表 1 に示す。但し、精度は FA+SErr で比較している。「確認無」はθ_1以上は受理、これ以下は棄却したもので、$\theta_1 = 0.9$ のときの FA(θ_1)+SErr(θ_1) の値である。このとき、FA(θ_1) = 16.2%, SErr(θ_1) = 29.5%である。「確認有」は$\theta_1 \geq CM_w > \theta_2$で確認を行い、それが誤り無く受理／棄却された場合で、$\theta_1 = 0.9$, $\theta_2 = 0.7$ のときの FA(θ_1)+SErr(θ_2) の値である。このときは FA(θ_1) = 16.2%, SErr(θ_2) = 25.7%となっている。

表 1 の結果より、CM_wを計算ししきい値を定めたことで精度が約 6%向上し、またCM_wを利用して確

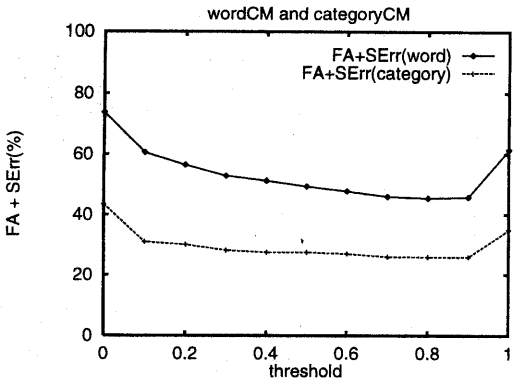

図 6: 単語 CM と意味カテゴリ CM の変化

認発話を行うことにより精度がさらに約 4%向上することが確認できた。

4.4 意味カテゴリ推定の有用性

単語の CM と意味カテゴリの CM の関係を図 6 に示す。これより、単語の CM よりも意味カテゴリの CM の方が、全般にわたって精度が良いことがわかる。すなわち、単語 CM で棄却された場合でも、意味カテゴリの CM から有用な情報が得られる可能性が示されている。

$\theta_2 = 0.7$ とした場合、$CM_w \leq \theta_2$で棄却されたスロットと、$\theta_2 < CM_w \leq \theta_1$で確認したが実際は棄却されるべきであった（回答が「いいえ」となる）スロットの合計は 153 個であった。このうち、意味カテゴリの CM_cが 0.9 以上でかつ正解であるものは 41 個であるため、単語 CM では棄却されたスロットのうち 27%に対して有効な誘導発話を行うことができる（FA は 69%）。

5 まとめ

音声認識誤りに対してより頑健に対話を進めるために、音声認識結果に対して内容語ごとに信頼度（Confidence Measure; CM）を計算し、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（CM）のしきい値は、実際に収録した対話音声に基づいて、受理／棄却のバランスが最適となるように定めた。また意味理解率は n-best 解を用いて CM を計算することにより、確認を行わない場合でも 6%、確認を行った場合では 10%の向上が見られた。意味カテゴリに関する CM を用いることにより、単語レベルの CM では棄却されたスロット（語彙外の発話なども含む）のうちの 27%に対して、単純に棄却してしまうよりも有効な誘導発話を生成できる可能性が示された。今後この意味カテゴリの CM の計算法を改良し、評価する予定である。

謝辞

本研究に対して、小笠原科学技術振興財団の支援を受けた。

参考文献

[1] 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.

[2] Watanabe T., Araki M., Doshita S.. Evaluating Dialogue Strategies under Communication Errors using Computer-to-Computer Simulation. Trans. of IEICE, Info & Syst., Vol.E81-D, No.9, pp.1025–1033, 1998.

[3] 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.

[4] T.Kawahara, C.-H.Lee, B.-H.Juang. 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.

[5] 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, 4, pp.1374–1382, 1999.

[6] G.Bouwman, J.Sturm, L.Boves. Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the ARISE Project. Proc. of ICASSP99, 1999.

[7] 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, 4, pp.1491–1498, 1999.

[8] 西宏之. 音声対話システムにおけるプロンプト音声送出タイミングの評価と制御法. 電子情報通信学会論文誌, Vol.J79-D-II, No.12, pp.2170–2175, 1996.

[9] 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 (98-SLP-24-14), 1998.