音声言語情報処理 5－7
（1995. 2. 4）

音声認識の誤りを考慮した対話制御方式のモデル化

新美康永　小林豊
京都工芸繊維大学

近年隠れマルコフモデルを基礎とした連続音声の認識技術が急速に発達してきており、この技術を用いた音声対話システムの研究も活発に行われている。しかし、朗読音声のようにきれいでない対話音声に対して、現在の音声認識技術はまだ不十分で、若干の誤りは避け難い。従って、ユーザの入力に対して何等かの確認作業が必要になる。本稿では、ユーザの発話内容を直接確認する場合（直接確認）と、ユーザの発話に対するシステムの応答の中にユーザ発話の内容をリフレイズして含める場合（間接確認）を考え、それぞれの対話制御方式の数学的なモデル化を行った結果について報告する。

Modeling Dialogue Control Strategies to Relieve Speech Recognition Errors

Yasuhisa Niimi and Yutaka Kobayashi
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, 606 Japan

Much effort has been made to study spoken dialogue systems. However, current technology for speech recognition, which has made remarkable progress, is still insufficient for recognition of utterances in spoken dialogue. So dialogue systems need to confirm recognized utterances. This paper considers two strategies for confirmation: direct confirmation and indirect confirmation. Using the first strategy, the dialogue system makes direct confirmation of user's utterances, and based on the second strategy, it includes rephrased utterances as a part of its responses. This paper reports mathematical models of these two strategies.

1 まえがき

近年、隠れマルコフモデルを基礎とした連続音声技術が急速に発達してきた。この技術を用いて、音声対話システムの研究も活発に行われている [1, 2, 3, 4]。しかし、朗読音声のようにきれいでない対話音声に対して、現在の音声認識技術は不十分で、若干の誤りは避け難い。したがって、ユーザの入力に対して何等かの確認作業が必要になる [5]。例えば、ユーザの入力が、「金閣寺の拝観料はいくらですか」といった質問であったとしよう。確認の手段としては、「ご質問は金閣寺の拝観料ですね」といった直接的な確認質問をする方法と、ユーザの質問に単に「500 円です」と答える代わりに、「金閣寺の拝観料は 500 円です」というふうに、システムの認識結果を応答に含ませる間接的な方法とが考えられる。本稿では、前者を直接確認、後者を間接確認と呼ぶ。このような確認に対するユーザからの応答の認識にも誤りを考慮しなければならない。以下では、このような確認質問を含んだ対話制御方式の数学的なモデル化について報告する。

2 対話制御方式のモデル化

2.1 簡単なモデル — モデル 0

最も単純な制御方法として、ユーザの発話を認識したとき、その認識スコアがある閾値以上であれば無条件で受理し、スコアが閾値より小さければ再入力を促すという方式を考えよう。このとき、ユーザの発話を受理する（認識のスコアが閾値以上になる）確率を α、受理した発話が正しく認識されている確率を p（α に依存する）とする。このようなシステムの動作を状態遷移図で表すと図 1のようになる。この図で、太い矢印はユーザまたはシステムの発話による遷移を、実線の矢印はシステムの判断による遷移を、点線は確率的な場合分けを示す。状態 get はユーザ発話に含まれているメッセージが正しくシステムに伝達された状態を、状態 loss はそのメッセージが誤って伝達された状態を示す。点線の部分は、システムには認識されていない。

このモデルにおいて、ユーザの発話に含まれていたメッセージがシステムに正しく伝達される確率を P_{ac} とすると、P_{ac} は次式によって計算できる。

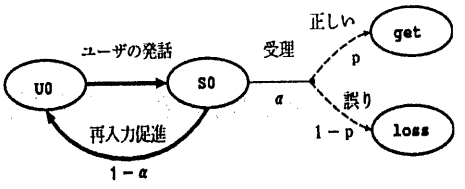


図 1. 最も簡単な対話制御のモデル

$$\begin{aligned} P_{ac} &= \alpha p + (1-\alpha) P_{ac} \\ P_{ac} &= p \end{aligned} \tag{1}$$

また、メッセージの伝達が終了するまでに要する平均発話数（ユーザとシステムの双方発話）を $N^{(0)}$ とすると、

$$\begin{aligned} N^{(0)} &= \alpha + (N^{(0)} + 2)(1-\alpha) \\ N^{(0)} &= \frac{2}{\alpha} - 1 \end{aligned} \tag{2}$$

となる。

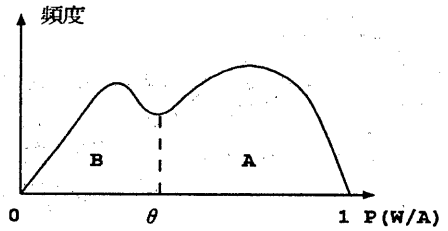
2.2 モデル・パラメタの決定法

図 1に示したモデル 0 には、2 つのパラメタ α と p とが含まれている。これらの決定法について説明する。

(1) ユーザ発話の受理基準 — いまユーザ発話の音響データ系列 A を単語列 W と認識したとしよう。認識スコアとして条件つき確率 $P(W/A)$ を採用するものとする。ベイズの定理により、

$$P(W/A) = P(A/W)P(W)/P(A) \tag{3}$$

となる。与えられた言語的制約を満たす単語列のうちで式 (3) を最大にする単語列 W を認識結果とする。通常の認識では W を変化させるだけなので、式 (3) の右辺の $P(A)$ を除いたものが用いられるが、ここでは異なった音響データ間の比較が必要になるので、式 (3) を用いる。この式の右辺に含まれる 3 つの項のうち、$P(A/W)$ は隠れマルコフモデルなどを用いた認識機構によって計算される。また、$P(W)$ は与えられた言語モデルによって計算できる。$P(A)$ を評価するには、次の 2 つの方法が考えられる。



$$\alpha = \frac{A}{A+B}$$

図 2. $P(W/A)$ のヒストグラム

(a) X を音素系列として、$P(X)P(A/X)$ で代用する [6]。

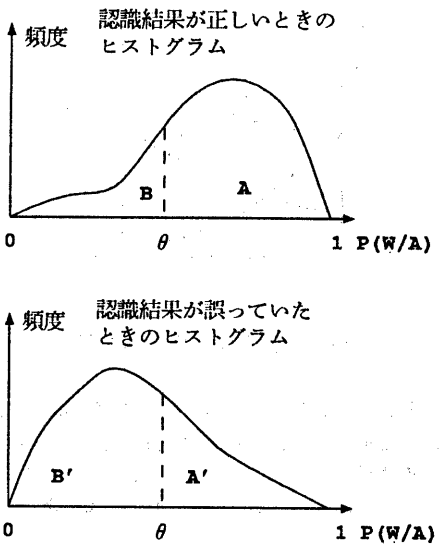
$$\begin{aligned} P(A) &= \sum_X P(A,X) \\ &= \sum_X P(X)P(A/X) \\ &\approx \max_X P(X)P(A/X) \end{aligned}$$

ここで、$P(X)$ は $P(W)$ と同様、音素単位の言語モデルから計算され、$P(A/X)$ は音素レベルの制約にのみ従った音素系列の認識結果から得られる。

(b) $P(A)$ を直接計算する。例えば、系列 A を生成する隠れマルコフモデルを用意しておき、それを用いて $P(A)$ を計算する。

以上のような方法で計算される $P(W/A)$ を多量の音響データ系列について計算し、ヒストグラムを作成する。ユーザの発話を受理するための閾値として θ をとり、$P(W/A) \geq \theta$ であれば発話を受理すると仮定する。図2に示したように、$P(W/A) \geq \theta$ となる部分のヒストグラムの面積を A、$P(W/A) < \theta$ となる部分のヒストグラムの面積を B とすると、上記のモデル中のパラメタ α は、$\alpha = A/(A+B)$ となる。

(2) 受理された発話が正しく受理されている確率 p — (1) で作成したヒストグラムを、認識結果の正しかったものと誤っていたものとに分けて描くと図3のようになる。各ヒストグラムの面積を図3のようにとると、モデルのパラメタ p は次式で与えられる。



$$\begin{aligned} p &= \frac{A'}{A'+A''} = \frac{A'}{A} \\ q &= \frac{B'}{B'+B''} = \frac{B'}{B} \end{aligned}$$

図 3. 認識結果毎の $P(W/A)$ のヒストグラム

$$p = \frac{A'}{A'+A''} = \frac{A'}{A} \tag{4}$$

ユーザ発話を受理するための $P(W/A)$ の閾値 θ を変化させたときの、p, α, N の想定される変化の様子を図4に示す。モデル0では、メッセージが正しく伝達される確率は式 (1) により p に等しいから、式 (4) の p を最大にするように θ (従って α) を定めることになるが、その結果 α が小さくなりすぎると、式 (2) によって平均発話回数 $N^{(0)}$ が大きくなるので適当な妥協が必要である。

2.3 一般化されたモデル

上で述べたモデル0を一般化することを考える。ユーザの発話に含まれているメッセージがシステムに正しく伝達された状態 (get) に至る経路と、誤って伝達された状態 (loss) に至る経路にのみ着目すると、図1の状態遷移図は図5の実線で示した部分のように抽象化される。図5で g_1 は、システムがユーザ発話を認識したとき、これを受理して状態 get に至る確率、l_1 は同じく状態 loss に至る

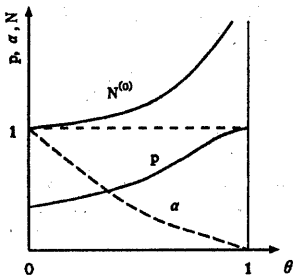


図 4. θ と $P, \alpha, N^{(0)}$ の関係

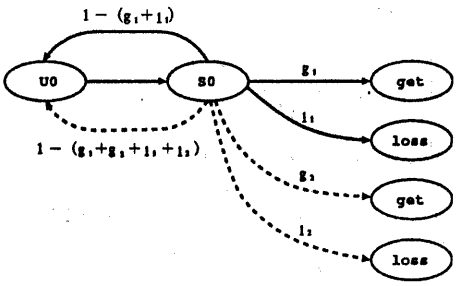


図 5. 抽象化されたモデルの状態遷移図

確率を表す。このとき、ユーザ発話を棄却して再入力を促す確率は、$1-(g_1+l_1)$ となる。従って、ユーザ発話に含まれているメッセージがシステムに正しく伝達される確率 P_{ac} は、次式によって計算される。

$$P_{ac} = \frac{g_1}{g_1+l_1} \tag{5}$$

このモデルに、確認用の質問を行う部分を追加すると、新たに状態 get や状態 loss に至る経路（システムの確認発話とそれに対するユーザの応答を含む）が追加される。この部分を図 5の点線で示している。新たに生じた経路の確率を図のように g_2, l_2 とする。この場合再入力が促進される確率は、$1-(g_1+g_2+l_1+l_2)$ となる。このとき、

$$P_{ac} = \frac{g_1+g_2}{g_1+g_2+l_1+l_2} \tag{6}$$

従って、モデル 0 に新しく点線の経路を追加することによってメッセージが正しく伝達される確率が増加するためには、

$$\begin{aligned} \frac{(g_1+g_2)}{(g_1+g_2+l_1+l_2)} &> \frac{g_1}{g_1+l_1} \\ \therefore \quad \frac{g_2}{g_1} &> \frac{l_2}{l_1} \end{aligned} \tag{7}$$

であることが必要である。

次に平均発話回数について考える。この場合もモデル 0 を一般化すると図 6のようになる。図 6には二種類の経路が存在する。一つはユーザの発話（システムに伝えたいメッセージを含む）をシステムが再入力を促すことなく受理する経路で、その経路の生起確率を α_i、その経路をたどる間に交わされる発話の数を $N_i(i=1,2,\cdots)$ とする。他はシステムがユーザの発話を棄却して再入力を促す

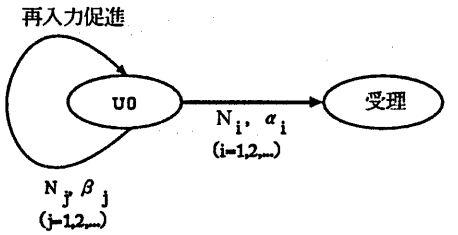


図 6. 抽象化されたモデルの平均発話回数を求めるための図

経路で、その生起確率を β_j、その経路をたどる間に交わされる発話の数を $N_j(j=1,2,\cdots)$ とする。ただし、$\sum\alpha_i+\sum\beta_j=1$ である。このとき、ユーザの最初の発話が最終的に受理されるまでに必要な平均発話回数 N は、次式によって計算できる。

$$N = \sum_i \alpha_i N_i + \sum_j \beta_j (N+N_j)$$

$\sum\alpha_i+\sum\beta_j=1$ に注意すると、

$$N = \frac{\sum_i \alpha_i N_i + \sum_j \beta_j N_j}{\sum_i \alpha_i} \tag{8}$$

となる。以下具体的な場合について考えよう。

2.4 直接確認を行うモデル — モデル 1

モデル 0 より少し複雑なモデルを考えよう。

$P(W/A)$ が θ より小さい場合、直ちに再入力を促すのではなく、認識内容を直接確認で確かめるものとする。この場合のシステムの動作は、図 7の

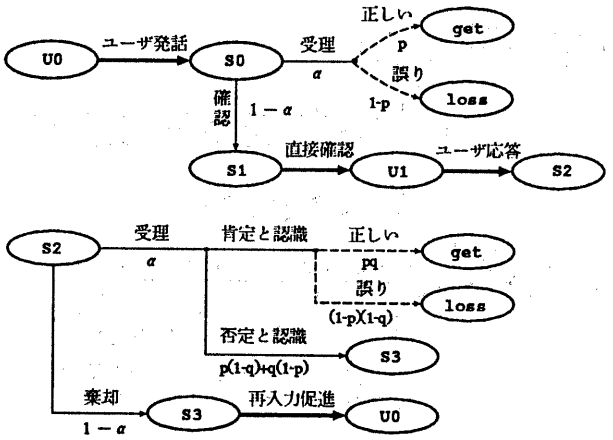


図 7. 直接確認を行なう場合の対話制御のモデル

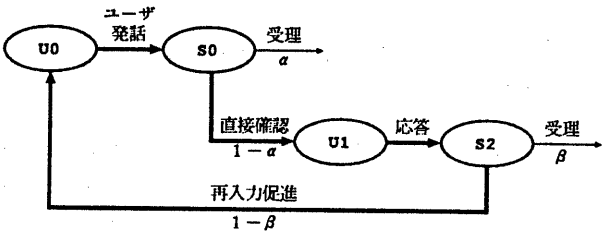


図 8. 図 7 を簡略化して示したダイアグラム

状態遷移図によって表現される。状態 S2 以降は、ユーザの応答に対するシステムの動作を場合分けして示したものである。ここに新しいパラメタ q が導入されている。これは、ユーザの発話を受理できなかった（システムが確認質問をした）場合、ユーザ発話の認識結果が正しかった確率（当然 p より小さい）を表し、図 3で定義した量を用いて、次式によって与えられる。

$$q = \frac{B'}{B' + B''} \tag{9}$$

図 7の点線の部分は、システムには認識できない。しかし、システムがユーザの応答を受理したとき、それを肯定応答と認識するか、否定応答と認識するかはシステムの判断であるので、状態 get へ至るものと状態 loss へ至るものとの和と、状態 S3 へ至るものとの区別は可能である。

直接確認に対するユーザの応答は、確率 q で肯定応答、確率 $(1-q)$ で否定応答になっている筈である。否定応答の場合には、単純な否定の他に訂正用の情報が含まれている場合があるが、ここでは簡単のためそれを無視している。また、直接確認に対するユーザの応答を棄却した場合は、この応答ではなく、最初の発話の再入力を促すものとしている。

このモデルでは、メッセージが正しく伝達される確率 P_{ac} は式 (6) を参照して次のようになる。

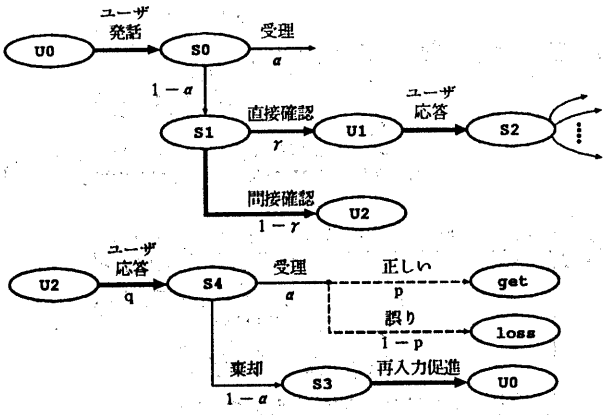


図 9. 間接確認を併用した場合の対話制御のモデル

$$P_{ac} = \frac{\alpha p + (1-\alpha)\alpha qp}{\alpha p + \alpha(1-p) + (1-\alpha)\alpha qp + (1-\alpha)\alpha(1-q)(1-p)}$$

$$\therefore \quad P_{ac} = \frac{p\{1+(1-\alpha)q\}}{1+(1-\alpha)(1+2pq-p-q)} \tag{10}$$

となる。また、式(7)を参照すると、$q > 1/2$ のとき $P_{ac} > p$ となることがわかる。

メッセージの伝達が終了するまでの平均発話回数 N を求めるために、図7を簡単化して示すと図8のようになる。ここで、パラメタ β は図7において状態 S2 から状態 get と状態 loss に至るパスの確率の和であり、

$$\begin{aligned}\beta &= \alpha qp + \alpha(1-q)(1-p) \\ &= \alpha(1+2pq-p-q)\end{aligned}$$

となる。$N^{(1)}$ は式 (8) を用いると次式のようになる。

$$N^{(1)} = \frac{\alpha + (1-\alpha)(4-\beta)}{\alpha + (1-\alpha)\beta} \tag{11}$$

2.5 間接確認を併用したモデル — モデル2

次に状態 S0 でユーザの発話を受理できなかったとき、直接確認だけでなく、間接確認を併用する場合を考えよう。直接確認を確率 γ で、間接確認を確率 $(1-\gamma)$ で選択するものとしよう。ユーザの発話内容によっては、間接確認を用いることができない場合があるが、このような場合もこの確率によって吸収しているものと考える。このモデルの状態遷移図は図9のようになる。ここで、システムの間接確認に対するユーザの応答が直接受理できなかった場合は、もとのユーザ発話（状態 U0 での発話）の再入力を促進することにしている。従って、メッセージが正しく伝達される確率 P_{ac} は、式 (6) によって次のように計算される。

$$P_{ac} = \frac{p\{1+(1-\alpha)[1+(q-1)\gamma]\}}{1+(1-\alpha)\{1+(2pq-p-q)\gamma\}} \tag{12}$$

この場合も簡単な計算により $q > 1/2$ のとき $P_{ac} > p$ となることが示される。また $\gamma = 0$、すなわち、直接確認を用いない場合は、$P_{ac} = p$ となる。

メッセージが伝達されるまでに必要な平均発話回数 $N^{(2)}$ を求めるために図9を書き換えると、図10のようになる。ここで、状態 S4 でユーザの応答を受理した後の確率的な場合分けについて、次のことを仮定している。

(a) 状態 U0 におけるユーザの発話が正しく認識されていれば、ユーザはシステムの間接確認には挨拶せず次の発話に進み、最初の発話が誤って認識されていれば、間接確認に対して訂正用の発話をするものとする。従って、システムの間接確認に対するユーザの応答のう

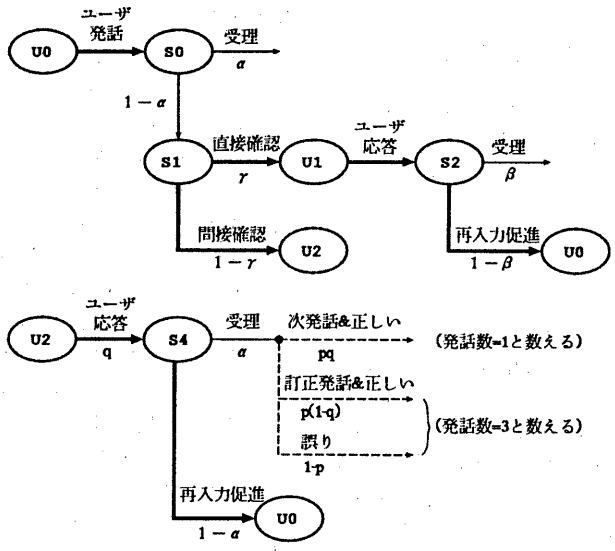


図 10. モデル 2 の平均発話回数を求めるためのダイアグラム

ち、次発話は確率 q で生起し、訂正発話は確率 $(1-q)$ で生起する。

(b) 状態 S4 でシステムがユーザの応答を受理したものに対しても、受理しかつ正しく認識したものに対しても、これらの割合は変わらない。

(c) 状態 S4 でシステムがユーザの次発話を正しく認識したときは、システムの間接確認とそれに対するユーザの発話は、最初の発話のために費やされたのではないと考え、この経路に対しては必要な発話数を 1 と数えている。式 (8) を参照すると次のようになる。

$$N^{(2)} = \frac{\alpha + (1-\alpha)[(4-\beta)\gamma + (4-\alpha-2\alpha pq)(1-\gamma)]}{\alpha + (1-\alpha)[\beta\gamma + \alpha(1-\gamma)]} \qquad (13)$$

3 モデルの比較

先に述べたようにモデル 1 においては、$q > 1/2$ であれば $P_{ac} > p$ となる。従って、認識結果が受理できないとき、単に再入力を促すのではなく、直接確認を行うことによって、システムの性能を向上させることができることを示している。その理由は、図 3の状態 S2 においてユーザの応答を受理したとき、認識結果が「否定応答」であれば、システムは最初のユーザ発話か直接確認質問に対する応答のいずれかを誤認識したことを「確信」できるからである。

モデル 2 において、$\gamma \neq 0$ かつ $q > 1/2$ のとき $P_{ac} > p$ であったが、$\gamma = 0$ とすると $P_{ac} = p$ となる。すなわち、間接確認だけではシステムの性能の向上ははかれない。

次に平均発話回数をモデル間で比較してみよう。モデル 0 とモデル 1 とでは、式 (2) と式 (11) とを比較すると、

$$N^{(1)} - N^{(0)} = \frac{\alpha + (1-\alpha)(4-\beta)}{\alpha + (1-\alpha)\beta} - \frac{2\alpha}{\alpha}$$
$$= \frac{2(1-\alpha)\{p(1-q) + q(1-p)\}}{\alpha + (1-\alpha)\beta} > 0$$

となり、モデル 0 の方が少ないことがわかる。また、式 (13) において $\gamma = 0$ (間接確認のみを用いる場合) とすると、

$$N^{(2)}{}_{(\gamma=0)} = N^{(0)} - \frac{2(1-\alpha)pq}{2-\alpha}$$

となり、$N^{(2)}{}_{(\gamma=0)} < N^{(0)}$ であることがわかる。従って、平均発話回数については、

モデル2$_{(\gamma=0)}$ < モデル0 < モデル1

となる。

4 あとがき

音声認識の誤りを考慮して、再入力を促したり確認質問を行ったりする対話制御方式を3つ挙げ、それぞれの動作の解析を行った。その結果、直接確認を行うとこれを行わない場合と比較して、メッセージが正しく伝達される確率が増加すること、また間接確認を行うとメッセージの伝達が終了するまでに必要な平均発話回数が減少することが示された。

今後は、実際の音声認識システムを用いて、モデルの基本パラメタを推定していくことが重要である。

参考文献

[1] Young, S.J. and Proctor, C.E., "The design and implementation of dialogue control in voice operated database inquiry systems," Computer Speech and Language, vol.13, no.4, pp.329-353 (1989).

[2] Young, S.R., Hauptman, A.G., Ward, W.D., Smith, E.T. and Werner, P.,"High Level Knowledge Sources in Usable Speech Recognition Systems," Comm. of ACM, vol.32, no.2, pp.183-194 (1989).

[3] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips. M., Polifroni, J. and Seneff, S, "The Voyager Speech Understanding System: Preliminary Development and Evaluation," Proc. of ICASSP, pp.73-76 (1990)

[4] Peckham, J., "Speech understanding and dialogue over telephone: an overview of progress in the SUNDIAL project," Proc. of the DARPA Speech and Natural Language Workshop, pp.14-27 (1992).

[5] Cozannet, A. and Siroux, J., "Strategies for oral dialogue control," Proc of ICSLP, pp.963-966 (1994).

[6] Young, S., "Detecting misrecognitions and out-of-vocabulary words," Proc. of ICASSP, vol.2, pp.21-24 (1994).