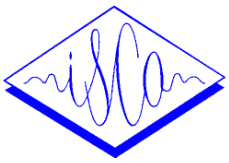

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

MODELING DIALOGUE CONTROL STRATEGIES TO RELIEVE SPEECH RECOGNITION ERRORS

Y. Niimi & Y. Kobayashi
e-mail: niimi@dj.kit.ac.jp
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, 606
JAPAN

ABSTRACT

This paper considers three dialogue control strategies to relieve speech recognition errors. These are the prompt to speak again (basic strategy), the direct confirmation and the indirect confirmation. The purpose of modeling the dialog control strategies is to estimate two quantities P_{ac} and N, given the performance of the speech recognizer used in a dialogue system. P_{ac} is the probability that information included in user's utterance is conveyed to the system correctly, and N is the average number of turns taken between the user and the system until terminating subdialogue on user's first utterance. The analysis has proven that the direct confirmation can increase P_{ac} and the indirect confirmation can reduce N in comparison with the basic strategy. Since the mathematical analysis described in this paper can easily be extended to more sophisticated strategies, it will contribute to the quantitative design of a dialogue control strategy, give the performance of a speech recognizer.

1. INTRODUCTION

A number of attempts have been made to study spoken dialogue systems[1, 2, 3, 4]. However,current technology for speech recognition, which has made a remarkable progress, is still insufficient for complete recognition of utterances in spoken dialogue, which are not so clean and grammatical as ones in read speech. Since misrecognitions are inevitable for such utterances, dialogue systems need to confirm recognized utterances[5]. This paper considers three dialogue control strategies to relieve speech recognition errors. These are the prompt to speak again (basic strategy), the direct confirmation and the indirect confirmation.

Here assume that the dialogue system have recognized an utterance as the sentence, "Please tell me the entrance fee of Kinkakuji temple." If the system cannot accept the sentence reliably, it has three options; it prompts the user to speak again, confirms directly by saying, "You mean the entrance fee of Kinkakuji temple?", or makes an indirect confirmation by answering, "You can enter Kinkakuji temple by 500 yen," instead of answering, "It's 500 yen."

The purpose of modeling the dialog control strategies is to estimate two quantities P_{ac} and N, given the performance of the speech recognizer. P_{ac} is the probability that information included in user's utterance is conveyed to the system correctly, and N is the average number of turns taken between the user and the system until terminating subdialogue on user's first utterance.

The analysis has proven that the direct confirmation can increase P_{ac} and the indirect confirmation can reduce N in comparison with the basic strategy. Since the mathematical treatment of dialogue control strategies can easily be extended to more sophisticated strategies, it will contribute to the quantitative design of a dialogue control strategy, given the performance of the speech recognizer.

2. THE BASIC STRATEGY

2.1 The Simplest Model

The first dialogue control strategy, the simplest of the three, is that the dialogue system accepts user's utterances when their recognition scores are greater than a threshold value, but rejects them otherwise and prompts the user to speak again. The dialogue system using this strategy is called model 0. Now assume we know the probability, denoted by α, that user's utterances are accepted, and the probability, denoted by p, that accepted utterances have been recognized correctly. We call these parameters recognizer parameters below in that they represent the performance of a speech recognizer. How to estimate them will be explained later.

How the dialogue system using the basic strategy works can be described by the state transition diagram as shown in Fig. 1. The state U0 indicates the situation the user is to speak something, and the state S0 the situation the system has recognized user's utterance. The state GET represents the situation the information contained in user's utterance has been conveyed to the system correctly, and the state LOSS the situation the system has misunderstood what the user said. The thick arrows show state transitions induced by an utterance of the user or the system, the thin arrow shows that the system has accepted user's utterance, and the dotted arrows probabilistic events which the system cannot realize.

For this model $P_{ac}^{(0)}$ (the upper script indicate the model index) is computed by the following recursive formula,

$$P_{ac}^{(0)} = \alpha + (1-\alpha)P_{ac}^{(0)}$$

The first term of the right hand side of the above equation represents the probability that the system accept user's utterance with no rejection, and the second term represents the probability that the system accept it after at least a rejection. From the above equation, we have,

$$P_{ac}^{(0)} = p. \quad (1)$$

$N^{(0)}$ can be computed in the same way.

$$N^{(0)} = \alpha + (1-\alpha)(N^{(0)} + 2).$$
$$N^{(0)} = 2/\alpha - 1. \quad (2)$$

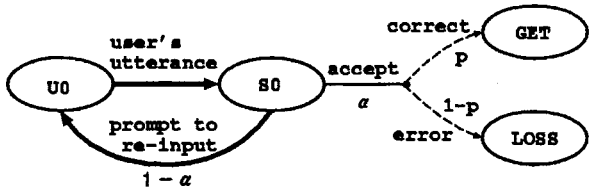

Fig.1 The state transition diagram for the model 0

Since p is expected to be inversely proportional to α, it is necessary for α to make small in order to increase $P_{ac}^{(0)} = p$. This, however, makes $N^{(0)}$ large. Some trade-off is then needed between $P_{ac}^{(0)}$ and $N^{(0)}$. In order to increase α and p in user's second turn, the dialogue system might analyze the cause of failure to accept user's first utterance, and suggest the user to speak more loudly or more clearly.

2.2 Estimation of the recognizer parameters

In this section we consider how to estimate α and p, and add a new parameter to the set of recognizer parameters. Let A denote the acoustic data stream of an utterance, and W denote a string of words. We can adopt the posterior probability $P(W/A)$ of W given A as a recognition score[6]. The recognized string of words is such a string that maximizes $P(W/A)$ under the given linguistic constraint. By Bayes' theorem,

$$P(W/A) = P(A/W)P(W)/P(A)$$

The quantity $P(A/W)P(W)$, which is used as a conventional criterion in speech recognition, is computed by using the hidden Markov model and the language model. Two methods can be considered for estimating $P(A)$; the first is to approximate $P(A)$ by $\max_X P(X)P(A/X)$where X is a string of phonemes, and the second is to use the HMM to compute $P(A)$ directly. Using one of these schemes to compute $P(W/A)$'s for many training utterances, we can create a distribution of $P(W/A)$. Selecting a threshold value θ, we can estimate α as the area of the portion of the histogram in which the inequality $P(W/A) \geq \theta$ is satisfied. p is also estimated in the similar way by using separate distribution created from correct recognitions and incorrect recognitions.

Here we introduce a new parameter denoted by q, the probability that an utterance whose reliability $P(W/A)$ is less than θ has been recognized correctly. We add it to the set of recognizer parameters, which will be used to describe the dialogue control strategies explained latter.

2.3 Generalization of the model 0

Generalizing the model 0, we introduce some formulae useful to analyze more complex dialogue control strategies than the basic one explained in section 2.1. Paying the attention to whether the system has understood user's utterances correctly, we can change the state transition diagram illustrated in Fig. 1 into the one as shown by solid lines in Fig. 2. In the new diagram g_1 denotes the probability that the system accept user's utterances and go to the state GET, and l_1 denotes the probability that the system accept user's utterances and go to the state LOSS. Then the probability that the system reject

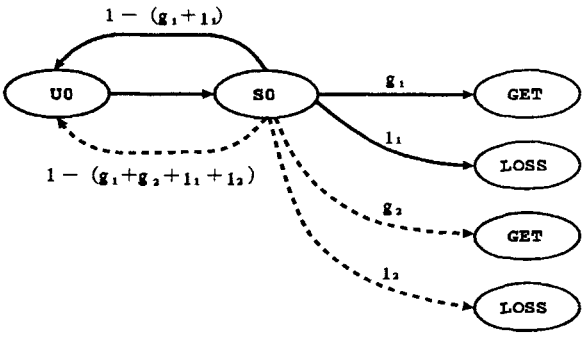

Fig.2 The state transition diagram for the generalized model

user's utterances and prompt the user to reinput can be represented by $1 - (g_1 + l_1)$. Using those notation we have,

$$P_{ac} = g_1/(g_1 + l_1)$$

For more sophisticated control strategies than the basic one, in which the system makes confirmations and the user responds to them, new paths reaching the state GET or the state LOSS would be added to the diagram shown in Fig. 2. A new path to the state GET and one to the state LOSS are illustrated by dotted lines in Fig. 2, being associated with the probabilities g_2 and l_2 respectively. The probability of the feedback path from the state S0 to the state $U0$ changes to $1 - (g_1 + g_2 + l_1 + l_2)$ by adding the new paths as shown by the dotted arrow. For the new diagram we have,

$$P_{ac} = (g_1 + g_2)/(g_1 + g_2 + l_1 + l_2) \qquad (3)$$

In order to increase P_{ac} by adding the new paths, the following inequality is necessary to hold,

$$\begin{aligned} (g_1 + g_2)/(g_1 + g_2 + l_1 + l_2) &> g_1/(g_1 + l_1) \\ g_2/g_1 &> l_2/l_1 \end{aligned} \qquad (4)$$

Next we consider N, that is, the average number of turns taken between the system and the user. In order to emphasize how many turns are taken along each path, we change the diagram in Fig. 2 to the one as shown in Fig. 3, in which N_i's (or N_j's) denote the number of turns taken during passing along each path, and α_i's (or β_j's) denote the probability that each path be taken. For this diagram we can compute N as follows,

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

Since $\sum \alpha_i + \sum \beta_j = 1$, then

$$N = \frac{\sum_i \alpha_i N_i + \sum_j \beta_j N_j}{\sum_i \alpha_i} \qquad (5)$$

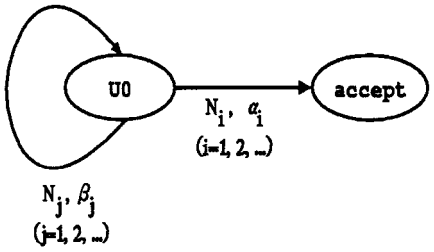

Fig.3 The simplified diagram to compute N

3. CONFIRMATION

3.1 Direct Confirmation

Here we consider the second dialogue control strategy, that is, the direct confirmation. The system using this strategy is called model 1. By this strategy the system confirms the content of recognized utterance, instead of rejecting it and asking the user to speak again, if its recognition score is less than the threshold value θ.

In the following analysis we assume for simplicity that user's response to the confirmation be either 'yes' or 'no' although some corrective information would follow the response 'no', and that the system ask the user to tell again what he has said first if it cannot accept the response reliably. The probability that the response 'yes' occur is q and the probability that the response 'no' occur is $1-q$, where the parameter q was introduced in the section 2.2.

It is illustrated in Fig. 4 how the system using the direct confirmation works. At the state $S2$ in Fig. 4, in other words, for user's response to the confirmation, the system selects one of the following three decisions;

(1) to reject the response,

(2) to accept the response which has been recognized as 'no', and

(3) to accept the response which has been recognized as 'yes'.

If it selects the first decision, the system goes to the state $S3$ where it prompts the user to tell again what he has said first. If it selects the second decision, it can realize that it should have misrecognized either user's first utterance or user's response to the confirmation. So it goes to the state $S3$. The probability that this happen is $\alpha(1-q)p + q(1-p)$. The third decision means that the system ends in accepting the unreliable recognition of user's first utterance. As defined in the section 2.2, we denote by q the probability that this recognition be correct. Then by this decision, the system obtains the correct information with the probability αpq and the wrong information with the probability $\alpha(1-p)(1-q)$.

The probability associated with each path in Fig. 6 can be obtained from those considerations. Using the formula (3), we can compute $P_{ac}^{(1)}$ for the model 1.

$$P_{ac}^{(1)} = \frac{p\{1+(1-\alpha)q\}}{1+(1-\alpha)(1+2pq-p-q)} \qquad (6)$$

Now we simplify the diagram shown in Fig. 4 to the one in Fig. 5 to compute $N^{(1)}$, the average number of turns taken between the user and the system until the

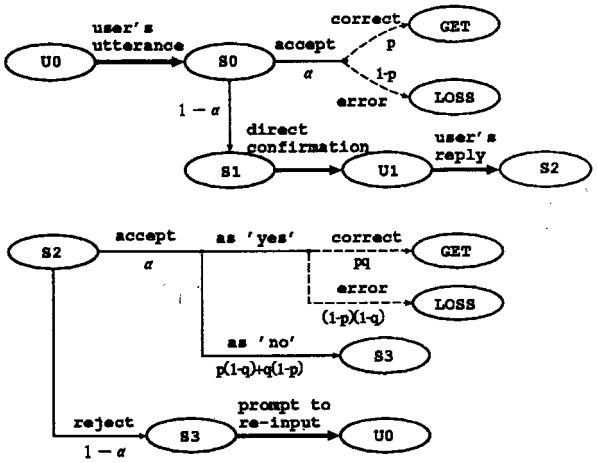

Fig.4 The state transition diagram for the model 1

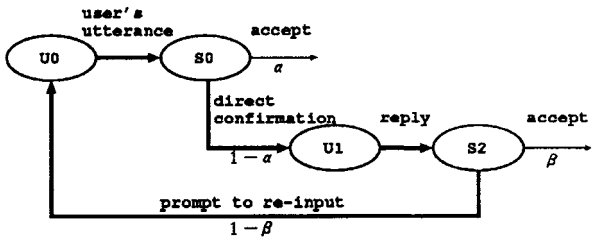

Fig.5 The simplified diagram for the model 1

subdialogue on user's first utterance terminates. The parameter β in Fig. 5 denotes the sum of the probabilities associated with the paths from the state S2 to the states GET and LOSS in Fig. 4, being $\alpha(1+2pq-p-q)$. Applying the formula (5) to Fig. 5, we have

$$N^{(1)} = \frac{\alpha+(1-\alpha)(4-\beta)}{\alpha+(1-\alpha)\beta} \qquad (7)$$

It is proven by simple calculation that $N^{(1)} > N^{(0)}$, and $P_{ac}^{(1)} > P_{ac}^{(0)} = p$ if $q > 1/2$.

3.2 Indirect Confirmation

Finally we consider the third strategy, that is, indirect confirmation. If it cannot accept user's utterance reliably, the dialogue system confirms its recognition indirectly by embedding the content of the utterance in its response. We call the system using only the indirect confirmation model 2, and the system using the direct and indirect confirmations model 3.

We assume the followings for the performance of the system and user's response to indirect confirmations.

(1) The system selects a direct confirmation with the probability γ and an indirect confirmation with the probability $1-\gamma$.

(2) The user proceeds to a new utterance without any comment to the indirect confirmation of correct recognition, but makes some correction to incorrect recognition. Therefore, of user's responses new utterances occur with the probability q and correction occur with the probability $1-q$.

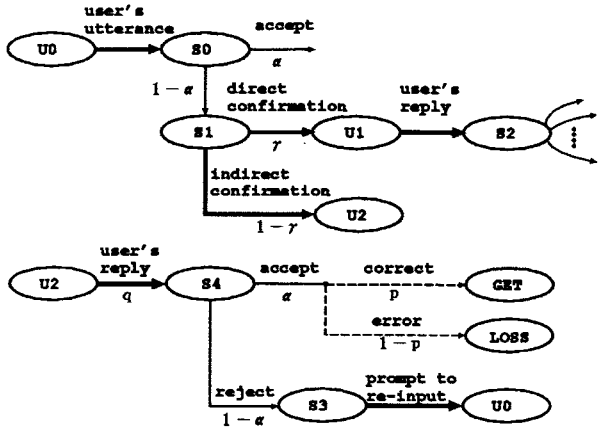

Fig.6 The state transition diagram for the model 3

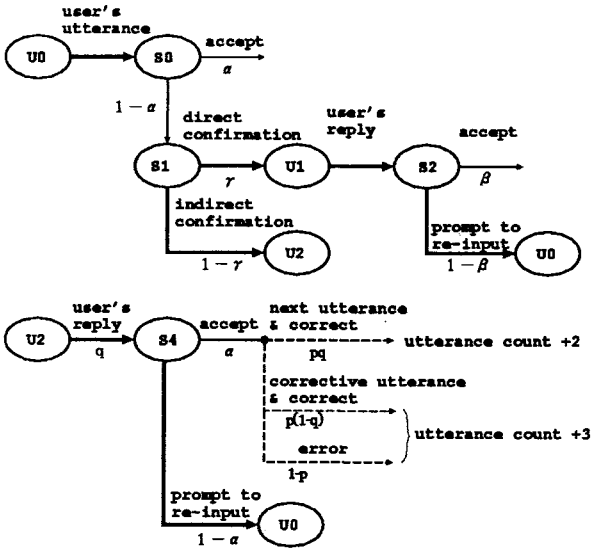

Fig.7 The simplified diagram for the model 3

(3) These two kinds of utterances, new utterances and corrections are equally accepted with the probability α and accepted utterances are recognized correctly with the probability p.

(4) When user's new utterance is recognized correctly, we consider the turns taken for the indirect confirmation and user's new utterance are not spent to convey the information of the first utterance.

The state transition diagram of the model 3 and its simplified diagram are illustrated in Fig. 6 and Fig. 7 respectively. Both new utterances and some corrections are contained in user's responses to the indirect confirmations, but these are treated in the same way by the system according to the above assumption (3). Applying the formula (3) to Fig. 6 and the formula (5) to Fig. 7, we can compute the model parameters $P_{ac}^{(3)}$ and $N^{(3)}$ for the model 3.

$$P_{ac}^{(3)} = \frac{p\{1+(1-\alpha)[1+(q-1)\gamma]\}}{1+(1-\alpha)\{1+(2pq-p-q)\gamma\}} \quad (8)$$

and

$$N^{(3)} = \frac{\alpha+(1-\alpha)[(4-\beta)\gamma+(4-\alpha-2\alpha pq)(1-\gamma)]}{\alpha+(1-\alpha)[\beta\gamma+\alpha(1-\gamma)]} \quad (9)$$

By putting $\gamma = 0$ we have for the model 2,

$$P_{ac}^{(2)} = p \quad (10)$$

and

$$N^{(2)} = \frac{\alpha+(1-\alpha)(4-\alpha-2\alpha pq)}{\alpha+\alpha(1-\alpha)} \quad (11)$$

In this case it is proven that $P_{ac}^{(3)} > P_{ac}^{(0)}$ if $q > 1/2$, and $P_{ac}^{(2)} = P_{ac}^{(0)}$ and $N^{(2)} < N^{(0)}$ if the system adopts only the indirect confirmation. It is proven from the simple calculation that the following approximations hold,

$$P_{ac}^{(3)} \approx \gamma P_{ac}^{(1)} + (1-\gamma)P_{ac}^{(2)} \quad (12)$$

and

$$N^{(3)} \approx \gamma N^{(1)} + (1-\gamma)N^{(2)} \quad (13)$$

4. CONCLUSION

This paper has reported three dialogue control strategies to relieve errors in speech recognition, and analyzed them mathematically. The analysis has proven that the direct confirmation can increase the probability that information included in user's utterances is conveyed to the system correctly, and the indirect confirmation can reduce the average number of turns exchanged between the user and the system. Since the method proposed in this paper can easily be extended to more sophisticated strategies, it will contribute to the quantitative design of a dialogue control strategy, given the performance of a speech recognizer. It should be investigated for future work how large P_{ac} and how small N are necessary to have comfortable conversation with a machine. This will determine the required performance of a speech recognizer.

REFERENCES

[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).