A DIALOG CONTROL STRATEGY BASED ON THE RELIABILITY OF SPEECH RECOGNITION

Yasuhisa NIIMI and Yutaka KOBAYASHI

Department of Electronics and Information Science,
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, 606 JAPAN
e-mail: niimi@dj.kit.ac.jp

ABSTRACT

This paper considers a dialog control strategy based on the reliability of speech recognition. The spoken dialog system using this strategy accepts an utterance of the user if the reliability is high, while it rejects the utterance and asks the user to speak again (the basic strategy), or confirms the content of the utterance if the reliability is low. The purpose of this paper is to estimate two quantities P_{ac} and N, given the performance of the speech recognizer used in a dialog 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 the subdialog on user's first utterance terminates. 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. The paper also reported a numerical evaluation of the dialog control strategy conducted by using a real speech recognition system.

1. INTRODUCTION

A number of attempts have been made to study spoken dialog systems[1, 2]. However, current technology for speech recognition, which has made a remarkable progress, is still insufficient for complete recognition of utterances in spoken dialog, which are not so clean and grammatical as ones in read speech. Since misrecognitions are inevitable for such utterances, dialog systems need to confirm recognized utterances[3].

This paper considers a dialog control strategy based on the reliability of speech recognition. The spoken dialog system using this strategy accepts an utterance of the user if the reliability is high, while it selects one of rejection of the utterance (the basic strategy), direct confirmation and indirect confirmation of the content of the utterance if the reliability is low. Here assume that the dialog 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, there would be 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 this paper 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 the subdialog on user's first utterance terminates.

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. We also conducted a numerical evaluation of the control strategy, using a speech recognition system. The evaluation showed that the reduction of N by the indirect confirmation was of significance while the improvement of P_{ac} by the direct confirmation was numerically negligible.

2. MODELING OF DIALOG CONTROL STRATEGY

2.1. Dialog control strategy

In this section we propose a dialog control strategy to relieve speech recognition errors. Let $R(u)$ be a function to measure reliability in recognizing an utterance u. We assume that $R(u)$ is non-negative and reliability is as high as $R(u)$ is small. The proposed dialog control strategy, which is illustrated in Fig. 1, can be stated as follows.

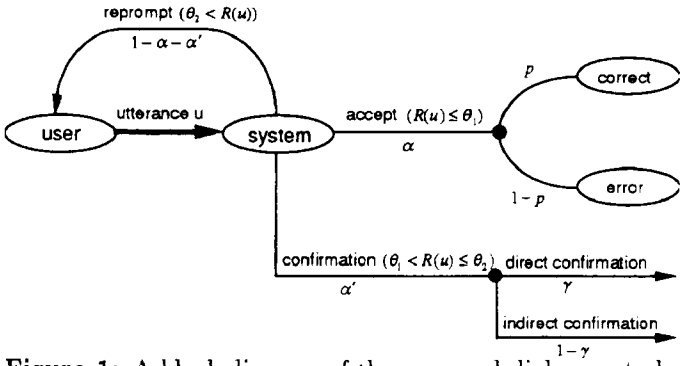

Figure 1: A block diagram of the proposed dialog control strategy

(1) If $R(u) \leq \theta_1$, the dialog system accepts an utterance u. Let α be the probability that an utterance is accepted, and p the probability that the accepted utterance has been recognized correctly. α and p are generally dependent on θ_1, α being proportional but p inversely proportional to θ_1.

(2) It confirms the content of u, if $\theta_1 < R(u) \leq \theta_2$. Let α' be the probability that the confirmation is made, and q be the probability that the confirmed utterance has been recognized correctly.

(3) It asks the user to speak again if $R(u) > \theta_2$. The probability for this is $1-\alpha-\alpha'$. We call these four probabilities α, α', p and p recognizer's parameters below.

We use two kinds of confirmations, that is, direct confirmation and indirect confirmation as stated in the previous section. We select a direct confirmation with the probability γ and an indirect confirmation with the probability $1-\gamma$. We assume the followings for user's response to the confirmation.

(1) User's response to the direct confirmation is either "yes" or "no" for simplicity. Following what is stated in (2), an affirmative response occurs with the probability q, and a negative response with the probability $1-q$.

(2) The user proceeds to a new utterance without making 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 corrections occur with the probability $1-q$.

We also assume the followings for the reaction of the dialog system to user's response, which is denoted by u'. It accepts a response u' if $R(u') \leq \theta_1$, while it rejects u' and asks the user to repeat the first utterance u. Let α be the probability that a response is accepted, and p the probability that the accepted response has been recognized correctly. Then when a response u' to a direct confirmation is accepted and recognized as an affirmative one, the correct information is conveyed to the system with the probability αpq and the incorrect information with the probability $\alpha(1-p)(1-q)$. When the dialog system has accepted u' and recognized as a negative response, it could know it has recognized either u or u' incorrectly. Accordingly the system prompts the user to repeat what (s)he said first. When the user's utterance to an indirect confirmation is accepted and recognized as a new utterance (its probability being αpq), we consider the turns taken for the indirect response and user's new utterance are not spent to convey the information of the first utterance u.

2.2. Analysis of the dialog control strategy

The purpose of modeling the dialog control strategy is to estimate two quantities P_{ac} and N; 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 the subdialog on user's first utterance terminates. Using the method reported in [4], we have the following formulae for P_{ac} and N;

$$P_{ac} = \frac{p\{1+\alpha'[1+\gamma(q-1)]\}}{1+\alpha'[1+\gamma(\beta-1)]} \quad (1)$$

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

where $\beta = 1+2pq-p-q$.

Here we consider three special cases of the dialog control strategy mentioned above. These are the case where $\alpha' = 0$, the case where $\alpha' \neq 0$ and $\gamma = 1$, and the case where $\alpha' \neq 0$ and $\gamma = 0$. We call these cases basic strategy, direct confirmation strategy and indirect confirmation strategy, respectively, and distinguish P_{ac}'s and N's for these cases by upperscripts 0, 1 and 2. For example, we have as the two parameters for the basic strategy, $P_{ac}{}^{(0)} = p$ and $N^{(0)} = 2/\alpha - 1$. By simple calculation we have the following inequalities, $P_{ac} > P_{ac}{}^{(0)}$ if $q > 1/2$ and $\alpha' \neq 0$, and $N < N(0)$ if $\gamma(p+q+(\alpha-2)pq) < \alpha pq$. When these conditions are satisfied, we can increase P_{ac} and reduce N by making confirmations.

3. ESTIMATION OF THE RECOGNIZER PARAMETERS

In this section we consider how to estimate the four recognizer parameters a, a', p and q. First recognizing many training utterances by a speech recognizer to be used in the dialog system, we have a recognition result, "correct" or "incorrect", and its reliability measure $R(u)$ for each utterance u. Then we can create two histograms of $R(u)$ for the correct recognition and the incorrect recognition, as shown in Fig. 2. Let $NT(x,y)$ and $NF(x,y)$ denote the accumulated frequency of the correct and the incorrect recognitions respectively for which $x < R(u) \leq y$. By using these notations, the four recognizer parameters a, a', p and q can be defined as follows:

$$\left.\begin{array}{lcl} \alpha & = & N(0,\theta_1)/N(0,\infty) \\ \alpha' & = & N(\theta_1,\theta_2)/N(0,\infty) \\ p & = & NT(0,\theta_1)/N(0,\theta_1) \\ q & = & NT(\theta_1,\theta_2)/N(\theta_1,\theta_2) \end{array}\right\} \quad (3)$$

where $N(x,y) = NT(x,y) + NF(x,y)$.

4. NUMERICAL EVALUATION OF THE CONTROL STRATEGY

4.1. The measure of the reliability of speech recognition

In this section we consider how to estimate $R(u)$, the reliability in recognizing an utterance u. Let A denote the

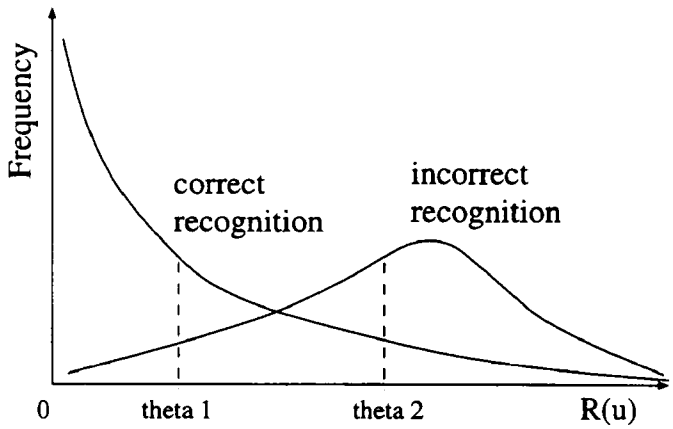

Figure 2: The histograms of $R(u)$'s for the correct and the incorrect recognitions

acoustic data stream of an utterance, and W denote a string of words. We have adopted as $R(u)$ a quantity proportional to the posterior probability $P(W/A)$ of W given A, as used in several speech recognition systems[5]. By Bayes' theorem, we can write $P(W|A)$ as

$$P(W|A) = P(A|W)P(W)/P(A). \quad (4)$$

The quantity $P(A|W)P(W)$, which is computed by using the hidden Markov model and the language model, is used as a conventional criterion in HMM-based speech recognition systems. In such systems the recognized string of words is the one that maximizes $P(W|A)$ under the given linguistic constraint.

According to the eq.(4) we must know $P(A)$ to compute $P(W|A)$. Two methods can be considered for estimating $P(A)$; the first is to use the HMM to compute $P(A)$ directly and the second is to approximate $P(A)$ by $\max_X(P(X)P(A|X))$ where X is a string of phonemes. In this paper we have adopted the second method. In the above approximation we can compute $P(A|X)$ using phoneme HMM's and Viterbi algorithm and $P(X)$ using the probabilities of bigrams or trigram of phonemes under the assumption of the Markovian property of the phoneme string. Thus we can write $P(W|A)$ as

$$P(W|A) = P(A|W)P(W)/\max_X P(A|X)P(X).$$

Since we use a deterministic language model in our speech understanding system which will be described later, we can consider that $P(W)$'s are equal for all word strings and $P(X)$'s are equal for all phoneme strings. In this case we have

$$P(W|A) = P(A|W)/\max_X P(A|X).$$

Since $\log P(A|W)$ and $\log P(A|X)$ can be computed easily in HMM-based speech recognition systems, we have adopted the following as $R(u)$;

$$R(u) = c \log P(W|A)/NFR$$

where c is a negative constant to make $R(u)$ positive and NFR is the length of an acoustic data stream A. It is expected that $P(W|A)$ is close to one if the recognition of an utterance u is correct while it is close to zero if the recognition is incorrect. This means that the closer to zero is $R(u)$, the higher is the reliability in recognizing an utterance u.

4.2. Numerical evaluation of the dialog control strategy

In this section we will demonstrate how the proposed dialog control strategy improves the performance of a spoken dialog system. A preliminary evaluation of the control strategies were conducted by using SUSKIT-II[6], the continuous speech recognition system we developed. SUSKIT-II is based on discrete HMM's of phonemes. A main task of it, of which the perplexity is about eight, is a database query about sightseeing spots in Kyoto. The syntax of the query is described by the definite clause grammar (the augmented context free grammar).

In the present evaluation, SUSKIT-II recognized three hundred utterances, after it was trained in the speaker-dependent mode. Then we created the histograms of $R(u)$'s for the utterances correctly recognized and for the utterances incorrectly recognized, which are illustrated in Fig. 3. Selecting two threshold values θ_1 and θ_2, we can determine the four recognizer parameters α, α', p and q, which were explained in section 3.

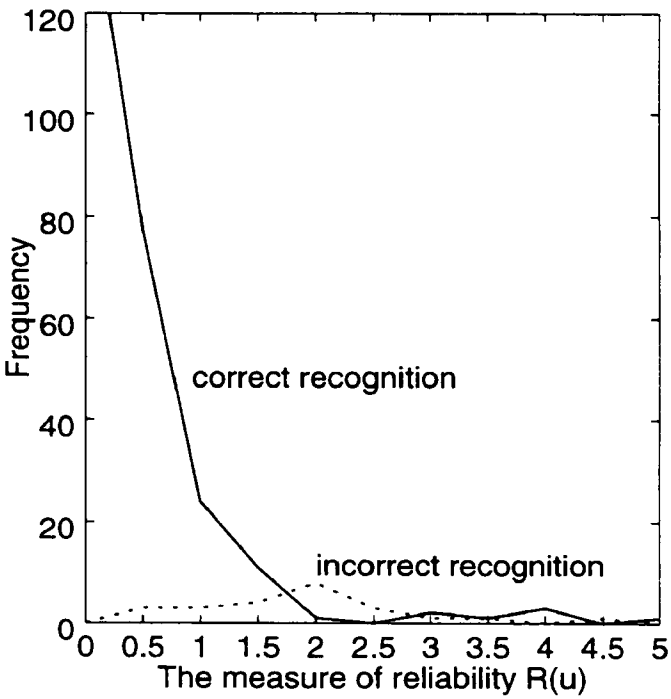

Figure 3: The histograms of $R(u)$'s measured by SUSKIT-II

(1) Evaluation of the basic strategy
The dialog system using the basic strategy accepts an

utterance u if $R(u) \leq \theta_1$, and otherwise rejects it, asking the user to reinput what (s)he has said. Given a threshold value θ_1, we can compute α and p by eq.(3), and thereby compute $P_{ac}{}^{(0)}$ and $N^{(0)}$ by eqs (1) and (2). Table 1 shows the relation of α, $P_{ac}{}^{(0)}(=p)$, and $N^{(0)}$ to the threshold θ_1.

If an original (first) utterance of the user is accepted after a repetition, the number of turns taken between the user and the system is three. In general if the repetition is required once k original utterances, the average number of turns necessary to accept one of these utterances is $(k+2)/k$. This value is 1.33 for $k=6$ and 1.40 for $k=5$. Taking these facts into consideration, we can interrupt what Table 1 means as follows.

(i) If an utterance is accepted unconditionally, the probability $P_{ac}{}^{(0)}$ is 0.92.

(ii) If the repetition is allowed once six original utterances of the speaker, $P_{ac}{}^{(0)}$ can be improved to over 0.98.

(iii) If the repetition is allowed once five original utterances of the speaker, $P_{ac}{}^{(0)}$ can be improved to 0.99.

(2) Evaluation of the indirect confirmation strategy
Putting $\gamma = 0$ in eqs (1) and (2), we have

$$P_{ac}{}^{(2)} = p$$

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

This means that the indirect confirmation can improve $N^{(2)}$ in comparison with $N^{(0)}$, while the probability $P_{ac}{}^{(2)}$ is kept equal to $P_{ac}{}^{(0)}$. Here we will examine numerically the extent to which $N^{(2)}$ can be reduced. The two recognizer parameters α' and q are necessary for this purpose. Fixing θ_1 to 1 and varying θ_2 from 1.5 to 3.5, for example, we can compute $\alpha'(s)$ and q's by using eq.(3), and Fig. 3. Table 2 shows those values together with $N^{(2)}$'s, which can be computed by eq.(2).

As stated above, fixing θ_1 to 1 in the basic strategy means that $P_{ac}{}^{(0)}$ is 0.99 and $N^{(0)}$ is 1.41. The latter corresponds to the case of $k=5$. In this situation, letting θ_2 be 3.5 in the indirect confirmation strategy, $N^{(2)}$ is improved to 1.34, which corresponds to the case of $k=6$.

(3) Evaluation of the direct confirmation strategy
As stated in section 2.2, the direct confirmation can improve $P_{ac}{}^{(1)}$ if $q > 0.5$. The rate of improvement can be written as $(1+\alpha q)/(1+\alpha'\beta)$ by putting $\gamma = 1$ in eq.(1). For the same set of parameters as used in (2) this rate becomes 0.06%, which is numerically negligible.

5. CONCLUSION

This paper has reported a dialog control strategy based on the reliability of speech recognition, and analyzed it 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.

Using a continuous speech recognition system, we estimated the reliability of speech recognition and thereby the four parameters α, α', p and q. Then we evaluated the proposed dialog control strategy quantitatively by computing P_{ac} and N. It has turned out that while the increase of P_{ac} by the direct confirmation is negligible, the reduction of N by the indirect confirmation is of significance.

Table 1: The relation of α, $P_{ac}{}^{(0)}$ and $N^{(0)}$ to θ_1

θ_1	α	$P_{ac}{}^{(0)}$	$N^{(0)}$
0.5	0.51	1.00	2.93
1.0	0.78	0.99	1.41
1.5	0.87	0.98	1.29
2.0	0.93	0.96	1.16
2.5	0.96	0.93	1.09
3.0	0.97	0.93	1.07
∞	1.00	0.92	1.00

Table 2: The relation of α', q and $N^{(2)}$ to θ_2

θ_2	α'	q	$N^{(2)}$
1.5	0.09	0.89	1.41
2.0	0.14	0.83	1.35
2.5	0.17	0.71	1.35
3.0	0.18	0.67	1.35
3.5	0.20	0.67	1.34

REFERENCES

[1] 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)

[2] 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).

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

[4] Niimi, Y. and Kobayashi, Y.,"Modeling dialogue control strategies to relieve speech recognition errors," Proc. of EUROSPEECH'95, pp.1177-1180 (1995).

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

[6] Kobayashi, Y. and Niimi, Y.,"Evaluation of a speech understanding system — SUSKIT-2," Proc. of ISCLP, pp.725-728 (1990).