Noise Adaptive Spoken Dialog System based on Selection of Multiple Dialog Strategies

Akinori Ito, Takanobu Oba, Takashi Konashi, Motoyuki Suzuki and Shozo Makino

Graduate School of Engineering
Tohoku University, Japan
{aito,bacchi,konashi,moto,makino}@makino.ecei.tohoku.ac.jp

Abstract

Speech recognition under noisy environment is one of the hottest topic in the speech recognition research. In this paper, we propose a method to improve accuracy of spoken dialog system from a dialog strategy point of view. In the proposed method, the dialog system automatically changes its dialog strategy according to the estimated recognition accuracy in noisy environment in order to keep the performance of the system constant. In a noisy environment, the system restricts its grammar and vocabulary to improve recognition accuracy. On the other hand, the system accepts any utterance from a user in a noise-free environment. To realize this strategy, we investigated a method to avoid user’s out-of-grammar utterances through an instruction given by the system to a user. Furthermore, we developed a method to estimate recognition accuracy from features extracted from noise signal. Finally, we constructed a proposed dialog system and confirmed its effectiveness.

1. Introduction

Spoken dialog based user interface is one of the most promising way to communicate with a mobile robot[1] or an electronic appliance. The spoken dialog based interface has many advantages. It is easy to use, no special device like a remote controller is needed, and there is no need to contact the equipment. However, there are several factors that degrade the performance of speech recognition in a real environment. The biggest one among these factors is environmental noise.

Many efforts has been made for speech recognition under noise [2]. Improvement of recognition accuracy under noisy environment is very important to realize noise-robust spoken dialog system. However, recognition accuracy under noise is not still enough. Here, we propose a new approach to improve robustness of a spoken dialog system from dialog strategy point of view. The basic idea of this work is how a user can help a spoken dialog system. If a user knows that the dialog system has difficulty under noisy environment, the user can help the system by limiting vocabulary or using simple expression.

In the next section, we show the overview of the proposed spoken dialog system. To realize the proposed system, two issues have to be solved. One issue is how to reduce out-of-task utterances. This issue is discussed in section 3. The other one is how to estimate speech recognition performance under a certain environmental condition. This is discussed in section 4. In section 5, the result of an experiment is presented to prove the effectiveness of the proposed system.

2. Overview of the system

We assume relatively small-vocabulary dialog system as the target of this work. The system aims at a user interface for mobile robots or home appliances such as TV, air conditioner etc. The vocabulary size is assumed to be up to several hundred, and the task is assumed to have up to six slots to be filled through the dialog. The proposed system relies on the fact that the simpler the grammar of the recognition system is, the higher the recognition accuracy is. Under a noise-free environment, recognition accuracy is expected to be high. In such environment, the dialog system can accept wide variety of utterances, namely, an utterance can contain any number of words to be filled into slots and their order can be scrambled. On the other hand, recognition accuracy degrades under a noisy environment. In this case, the system can keep recognition accuracy high by limiting variety of acceptable expressions. Therefore, if recognition accuracy under a certain environment can be estimated, the dialog system can change its recognition grammar according to the estimated accuracy. When the system is used in a clean environment, the system uses a large vocabulary and a complex grammar. If the environment is noisy, the system uses a smaller vocabulary and a simpler grammar.

To realize this system, we have to solve the following two issues.

1. How to avoid out-of-grammar utterances. Even if the system employs a small vocabulary and a simple grammar, dialog will fail if the user doesn’t know what he/she can utter to the system. There-
fore, we should develop a method to inform a user of the grammatical limitation, in other words, a method to control a user to follow the system’s instruction and to suppress out-of-grammar utterances.

2. How to estimate recognition accuracy under a certain environment. There are many factors that affects the recognition accuracy. We have to find an effective set of features for the estimation, as well as to develop a method to estimate the accuracy. At the same time, a method to choose proper vocabulary and grammar should be developed too.

3. Suppression of out-of-grammar utterances

In this section, we examine dialog strategies to suppress a user’s out-of-grammar (OOG) utterances. To avoid a user’s OOG utterances, we tried to change the system’s dialog strategy. Here we call two factors ‘dialog strategy.’ First one is an instruction the system first gives to the user. The second one is the form of questions the system gives to the user.

We compared the following five kinds of dialog strategies.

- ‘One item (OI)’ strategy is to ask one question to obtain one item.
- ‘Two items (TI)’ strategy is to ask one question to obtain up to two items.
- Questions in ‘answer the question (AQ)’ strategy are same as that in OI strategy, but a system using this strategy first asks a user to answer the question only to suppress utterances the system is not ready to receive.
- ‘Answer using a word (AW)’ strategy is similar to AQ strategy, but the system asks a user to answer using an isolated word. If a user obeys this instruction, the system can keep its grammar quite simple.
- ‘choice(CH)’ strategy is to tell a user all available choices and their numbers.

We carried out a dialog experiment. Subjects are 22 males and 10 females. One subject carried out two dialogs per strategy. The task domain of the dialog was operation of electric appliances. Goal of a dialog was given to the user before the dialog started. The dialog was performed with Wizard-of-Oz method.

OOG rates for each strategy is shown in Figure 1. OOG rate of the CH strategy was 0% (no OOG utterance was observed). For the AQ strategy, there was only one OOG utterance. The OI strategy gave higher OOG rate than the TI strategy because users tend to utter an answer that contains more than one items even if the system requires only one item. From this result, CH, AQ and TI strategies were found to be effective to avoid OOG utterances.

4. Estimation of recognition accuracy under noisy environment

The second issue to realize the proposed system is to develop a method to estimate recognition accuracy under certain environment. To tackle this problem, we employed neural network based method.

We examined the following 11 parameters.

- S/N ratio ($R$)
- Standard deviation of power divided by mean of power ($\sigma_p/\mu_p$)
- Mean and variance of correlations between contiguous two frames ($\mu_c, \sigma^2_c$)
- Mean and variance of spectral slope of each frame ($\mu_s, \sigma^2_s$)
- Number of phonemes obtained from the speech recognizer when noise is input to the recognizer ($N_p$)
- Vocabulary size ($V$)
- Number of sentences generated from the recognition grammar ($N_s$)
- Number of words included in the above sentences ($N_w$)
- Branching factor of the grammar ($B$)

Correlation between two frames is used to detect voiced period. Number of phoneme from the recognizer is used as a practical index how the noise is similar to speech. $N_s$ and $N_w$ can be calculated because we use finite state automata without loop as grammars.

Then we reduced number of parameters to investigate which parameters are essential for estimation of recognition accuracy. The selection procedure is as follows. When $N$ kinds of parameters are given, we can create $N$ models, each of them uses $N - 1$ parameters by removing one parameter out of $N$ parameters. Then estimation
errors are measured for all $N$ models, and the model with the minimum error is chosen. In this way, we can reduce number of parameters one by one.

In the experiment, 52 grammars were examined. Minimum and maximum values of each parameter for all grammars are shown in Table 1. Utterances for recognition experiment were chosen from utterances gathered through the experiments described in the previous section. For each grammar, five utterances that can be accepted by the grammar were chosen.

For noise signals, we chose eight environmental sounds from the JEIDA noise database[4] and four musical pieces from the RWC music database[5]. 135 noise data are chosen among these databases. First 1 or 2 second noise data are used to estimate parameters of the noise. The observed noise signal is analyzed using 25msec Hamming window with 25msec frame shift. Five speech data are mixed to each noise signal respectively, with 1 second interval between each data. Seven kinds of SNR (40, 30, 20, 15, 10, 5, 0dB) were examined for each frame ($\mu_s$), number of phonemes obtained from the speech recognizer when noise is input to the recognizer ($N_p$) and the vocabulary size ($V$).

<table>
<thead>
<tr>
<th>property</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V$</td>
<td>30</td>
<td>340</td>
</tr>
<tr>
<td>$N_s$</td>
<td>155</td>
<td>$8.0 \times 10^8$</td>
</tr>
<tr>
<td>$N_v$</td>
<td>860</td>
<td>$1.6 \times 10^6$</td>
</tr>
<tr>
<td>$B$</td>
<td>1.7</td>
<td>3.6</td>
</tr>
</tbody>
</table>

5. Realization and evaluation of the dialog system

In this section, we describe realization of the proposed dialog system and the result of the dialog experiment.

The proposed system works as follows. First, the system observes noise signal and extracts the noise parameters. The dialog strategy selector receives the parameters and estimates recognition accuracy for each grammar. Then it chooses the optimum grammar and dialog strategy. The selected grammar is passed to the speech recognizer, and the speech recognizer gets ready to hear user’s utterance.

In this system, the estimation of accuracy is carried out only once before the dialog begins. It would be possible to estimate the accuracy before each utterance, but it is still unclear whether it is useful to switch dialog strategy utterance by utterance. Therefore, we decided to determine the strategy at the very first of the dialog.

Next we explain how the strategy selector determines the optimum dialog strategy. The basic criteria to select a strategy is number of user utterances and recognition accuracy. The strategy selector chooses a strategy whose estimated number of user utterance is minimum and estimated recognition accuracy is higher than a certain threshold.

First, the noise parameter extractor calculates three parameters $R$, $\mu_s$, and $N_p$ that are selected in the last section. The dialog strategy selector receives these parameters, and estimates the recognition accuracy using the vocabulary size $V$ of one of the prepared dialog strategy. If a strategy uses more than one grammar, average size of the vocabularies is regarded as $V$.

Let the strategies be $S_1$, $S_2$, $S_k$. Let $V_i$ be vocabulary size of the grammar used in strategy $S_i$. Using tuple ($R, \mu_s, N_p, V_i$), the neural network estimates the recognition accuracy $\alpha_i$. Then estimated number of user utterance $N_w(S_i, \alpha_i)$ is calculated as follows.

Let $k_i$ be number of items expected to be contained
A noise-robust dialog system is proposed. This system changes its dialog strategy according to the environmental noise. To achieve this, the following to issues are investigated: dialog strategy to avoid out-of-grammar utterances and estimation method of recognition accuracy under a certain noise environment. Finally, we realized the proposed system and carried out dialog experiment. From the results, it is confirmed that the proposed system achieves a task with smaller number of user utterances compared to conventional system.

7. References


