KOREAN SPEECH DIALOG SYSTEM FOR HOTEL RESERVATION

Joon Hyung Ryoo†, Katunobu Itou‡, Satoru Hayamizu‡, and Kazuyo Tanaka‡

†Automatic Interpretation Section,
Electronics and Telecommunications Research Institute, Korea

‡Speech Processing Section,
Electrotechnical Laboratory, Japan

ABSTRACT - This paper describes an experimental Korean speech dialog system for hotel reservation task. The system consists of a language-dependent dialog manager and a language-independent speech recognizer which is based on the techniques used in Japanese continuous speech recognizer. We describe these two modules of the system and how to use the sentence pattern information to reduce the search space. We performed speaker independent recognition experiments to evaluate the performance of the dialog system. Sentence recognition rate is 62.3% with perplexity 13.1 for 1203 utterances uttered by 11 male persons. Response time of the system is 3-5 seconds when we run the whole process on HP9000.

INTRODUCTION

Recently, speech recognition technology has been improved so that a real working system can be implemented for a specific task domain. The speech and text databases for Korean have also been increased in size and quality as a result of recent research activities on continuous speech recognition(Hahn *et al.*, 1994). Under these circumstances, we have developed an experimental Korean speech dialog system for limited hotel reservation task which is of practical use. A user is assumed to have a conversation with the system through the telephone network. The system carries out four principal functions : hotel reservation, change of reservation, cancel of reservation, and call transfer to a certain guest room. Fig. 1 shows some part of dialog for reservation.

(translated in English)

	.
	.
System	Hello! This is a front desk.
	Which do you want to do, reservation, change, or cancel?
User	I want to make a reservation.
System	You want to make a reservation, don't you?
User	Yes.
System	When do you want to stay?
User	For one week from Thursday next week.
System	It is for seven days from March 31 to April 6, isn't it?
User	That's right.
System	O.K. You can make a reservation.
	Would you please spell your last name for me?
User	It is K,I,M.
	.
	.

Figure 1: Part of dialog for reservation

The system has the initiative about dialog. A user is asked to respond to system's requests about

his family name in alphabet, period of stay, grade of room, bedtype, room number(1-99) as well as the confirmation of user's previous response. A user is restricted to respond to the system's request using only one sentence. The system can not handle unknown words and unnecessary words. Fig. 2 shows the block diagram of the system which consists of two modules, a speech recognizer and a dialog manager. We describe both modules as well as how the sentence pattern information is applied to LR table to reduce the search space.

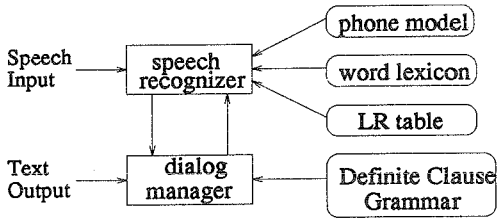

Figure 2: Block diagram of the speech dialog system

SPEECH RECOGNIZER

In order to extract feature parameters, input speech data are sampled at 16kHz, quantized to 16 bit resolution, and Hamming-windowed for each frame. The window is moved 5ms between frames.

We use 29-dimensional feature vectors which consist of 14 LPC mel-cepstral coefficients, 14 differential LPC mel-cepstral coefficients and differential log power. These vectors are converted into VQ code sequences using 1024 size codebook.

Basically the speech recognizer is niNja continuous speech recognition system(Itou *et al.*, 1992). It is an HMM-based system with an N-best search algorithm incorporated with LR-parsing technique. It is based on frame synchronous One-Pass DP algorithm. Recognition is carried out by driving phone-HMMs in accordance with LR table and lexicon. The recognizer regards each utterance as one sentence. The ouput of speech recognizer is a sentence hypothesis with the highest probability.

Recognition is performed by generating and pruning the partial hypotheses at each frame. Each partial hypothesis is handled by a hypothesis cell which has information about a score of the hypothesis and pointers corresponding to local hypotheses such as phone, word, and syntactic category hypotheses.

PHONE MODELS, LEXICON, AND GRAMMAR

The recognizer uses 46 phone models which include 3 models for the start and the end of each utterance and the pause. Phone models are context-independent discrete density HMM models, each of which has a structure with 3 state, 6 arc, 3 loops and 3 density functions. These models are trained using 4367 utterances spoken by 40 Korean male speakers in sound-proof room. We use only phoneme sequence information to train the models because speech data are not labeled yet.

Vocabulary size in lexicon is 182. Each word is related to one of 47 terminal symbols in CFG for user's response. Phoneme sequences about words in lexicon are stored in 47 LR-dictionaries each of which has structure equivalant to trie structure. An LR-dictionary is constructed with phoneme sequences about the words related to a terminal symbol in CFG. It is similar to making an LR table with word sequences about sentences.

The lexicon includes 52 Korean family names which all are one syllable words. To recognize well one syllable names, we adopt the pronunciation of the alphabet which corresponds to the name. For example, 'KIM', a representative Korean family name, is registered as /k e i a i e m/, not as /k i m/. A Korean syllable has a CVC structure and its pronunciation is frequently changed by syllable context regularly or irregularly. For change of pronunciation between words, all the pronunciation patterns are prepared. But syllable context information is not written in the lexicon. In recognition, therefore, the pronunciation pattern with the best score is chosen regardless of the actual context. All digits are hard to recognize because all of them are composed of one syllable and their pronunciations are frequently changed when they are connected to another digits or one syllable words. For example, '26' and 'room no. 1' are changed as follows.

```
 /i/    /s i bx/  /yu gx/  --->  /i s i m n yu gx/
(two)    (ten)      (six)            (twenty six)

/i lx/    /h o/     --->    /i r o/
(one)  (Room no.)       (Room no. one)
```

In these cases, the combined words are also given as entries in lexicon.

Context free grammar necessary to make an LR table is written by hand considering each sentence pattern. The perplexity of CFG for user's response is 13.1

REDUCTION OF SEARCH SPACE

We use two methods to reduce search space in recognition. The first is beam search by which the number of partial hypotheses is constrained at each frame. The second is to use the information on dialog situation. The sentence pattern of user's next response can be predicted by dialog manager and we search only space related to the pattern. The system uses 9 sentence patterns according to dialog situation. To describe how the sentence pattern is used to reduce the search space, that is, the number of partial hypotheses, we show a simple CFG in Tab. 1. The CFG has two sentence patterns, P1 and P2. They are represented by nonterminals p1 and p2 respectively in Tab. 1.

rule no.	grammar rule
1	S : bgn ptn end
2	ptn : p1
3	ptn : p2
4	p1 : a b
5	p1 : b c
6	p2 : c d
7	a : w
8	b : x
9	c : y
10	d : z

Table 1: A simple CFG

In speech recognition, LR table is used to predict the next phone or word. It has the information about the next state(at shift) and grammar rule(at reduce) which are applied at the next time. In recognition we can reduce search space by handling only actions related to the sentence pattern. LR table corresponding to the the simple CFG is shown in Tab. 2.

	bgn	w	x	y	z	end	$	S	ptn	p1	p2	a	b	c	d
0	**s01**	.	.	.	.	.	.	g02	.	.	.	.	.	.	.
1	.	**s03**	**s04**	**s05**	.	.	.	.	g06	g07	g08	g09	g10	g11	.
2	.	.	.	.	.	.	acc	.	.	.	.	.	.	.	.
3	.	.	**r07**	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	**r08**	.	**r08**	.	.	.	.	.	.	.	.	.
5	.	.	.	.	**r09**	**r09**	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	**s12**	.	.	.	.	.	.	.	.	.
7	.	.	.	.	.	**r02**	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	r03	.	.	.	.	.	.	.	.	.
9	.	.	**s04**	.	.	.	.	.	.	.	.	.	g13	.	.
10	.	.	.	**s05**	.	.	.	.	.	.	.	.	.	g14	.
11	.	.	.	.	s15	.	.	.	.	.	.	.	.	.	g16
12	.	.	.	.	.	.	**r01**	.	.	.	.	.	.	.	.
13	.	.	.	.	.	**r04**	.	.	.	.	.	.	.	.	.
14	.	.	.	.	.	**r05**	.	.	.	.	.	.	.	.	.
15	.	.	.	.	.	r10	.	.	.	.	.	.	.	.	.
16	.	.	.	.	.	r06	.	.	.	.	.	.	.	.	.

Table 2: LR table about the simple CFG

set no.	**0**		**1**									**2**	**3**	**4**	**5**
items	**0**	**1**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**	**7**	**8**	**9**
	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1

set no.	**6**	**7**	**8**	**9**		**10**		**11**		**12**	**13**	**14**	15	16
items	**1**	**2**	**3**	**4**	**8**	**5**	**9**	6	10	**1**	**4**	**5**	10	6
	2	1	1	1	0	1	0	1	0	**3**	**2**	**2**	1	2

Table 3: LR(0) items about the simple CFG

The following is how to obtain the rules and states corresponding to each sentence pattern. In Tab. 1, grammar rules corresponding to sentence pattern P1 (P1 rules, hereafter) are the ones that are induced from nonterminal symbol p1. That is, those are rules that have p1 in the left side and rules that are induced from nonterminals found in the right side of the former rules. Therefore, P1 rules are 0, 1, 2, 4, 5, 7, 8, 9 as shown with boldfaces in Tab. 1. Here rule number 0 is an augmented rule which is applied in acceptance. We can obtain states corresponding to pattern P1(P1 states, hereafter) using P1 rules. The left-most column in LR table shown in Tab. 2 represents a state number and corresponds to the set number in Tab. 3. Each state in LR table consists of LR(0) items found at lower side in Tab. 3.

Each item consists of two numbers which represent the rule number(upper) and the position(lower) already processed in that rule. we collect sets which include one or more P1 rules. In Tab. 3, we can obtain set numbers 0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 14 as shown in boldfaces in Tab. 3. In recognition, we utilize P1 rules and P1 states to process Only actions in boldfaces in Tab. 2.

DIALOG MANAGER

Dialog manager takes the hypothesis with best score from recognizer and performs semantic processing, response generation, and dialog control on Prolog. In case that a posteriori probability

of recognized sentence is much lower than the threshold fixed previously, system repeat the previous question to user. This is because the sentence pattern of user's response may be different from the one predicted by the dialog manager. we choose a fixed threshold which minimizes the number of repeated questions, between two average a posteriori probabilities which correspond to correctly recognized sentence and misrecognized one. Semantic processing is implemented easily in Prolog by converting two CFGs for user's and system's responses to corresponding definite clause grammars(DCG). Contents of reservation are saved in a list form as follows.

```
Reservation list: [Name, Date, Roomgrade, Bedtype]
```

Flow of dialog can be represented by the finite state network. Text response is generated to a screen according to the dialog state. Dialog can be divided into four cases: reservation, change of reservation, cancelation of reservation and a call transfer to a certain guest room. The number of interaction to complete the reservation procedure is about 13. It is about 15, 7, or 4 in each case of change, cancel, or call-transfer respectively.

Dialog manager and speech recognizer run independently and communicate with each other through the communication file in which a list is written as follows.

```
Speech recognizer: [recog, probability, recognized sentence with list form]
Dialog manager:    [dialog, the next sentence pattern]
```

Each module always investigates the attribute of the first item of the list in the communication file and proceeds its own work with the list when the list is generated by the other side.

The dialog manager sends to the speech recognizer the sentence pattern which would be uttered by user next time.

EXPERIMENTS

We performed speaker independent recognition experiments to measure performance of the dialog system. To evaluate the speech dialog system, we investigated the performance of the speech recognizer. Speech data used in recognition experiment are 1203 utterances by 11 Korean males which are not included in training phoneme HMM. Speech data for training and recognition were recorded in the same condition. Average number of words per utterance is 3.13. Recognition experiments are performed in two cases which are with and without the sentence pattern information to constrain the search space.

Recognition rates under the beam width N=200 are shown in Tab. 4.

sent. pttn info.	word rec. rate	sentence rec. rate	perplexity
no use	83.5%	62.3%	13.1
use	87.2%	64.8%	

Table 4: Recognition rates of the speech recognizer

Although the sentence pattern is applied, the recognition rate does not increase much. Tab. 5 shows word recognition rate about 9 sentence patterns. Most of recognition errors occurred within the same sentence pattern. This means that the differences of scores within the patterns are lower than those between the patterns.

sent. pttn info.	P1	P2	P3	P4	P5	P6	P7	P8	P9
no use	76.7	11.1	87.7	66.7	87.4	75.3	84.8	84.3	94.4
use	90.8	68.9	87.9	86.7	87.4	85.1	88.8	86.8	94.4
no. of sent.	85	72	139	43	405	255	80	95	29
perplexity	4.1	13.3	5.8	5.0	8.2	14.4	4.7	4.0	4.3

Table 5: Word recognition rate for each sentence pattern(beam width N=200, unit:%)

In Tab. 5, we can find relatively low recognition rate in pattern P2. According to additional analysis, most of the error types are that /n e/ (yes) in P2 is misrecognized to /n ae i r yo/ (It is tomorrow). /n e/ is one syllable sentence and usually pronounced long time. So it seems that /n e/ is misrecognized to /n ae i r yo/, a long sentence including similar phonemes.

Response time of the system is 3-5 seconds when we run the whole process with beam width N=100 on HP9000 model 735. Entire process includes speech detection, feature extraction, vector quantization as well as recognition and dialog management.

CONCLUSION

This paper describes speech recognizer and dialog manager of the Korean speech dialog system. The speech recognizer is an HMM-based system with an N-best search algorithm incorporated with LR-parsing technique. We also describe how to utilize the sentence pattern information in LR table to reduce search space. We have performed speaker independent recognition experiments as a measure of performance of the dialog system. Sentence and word recognition rate are 62.3% and 83.5% respectively with perplexity 13.1 for 1203 utterances uttered by 11 male persons. Response time of the system is 3-5 seconds when we run the whole process on HP9000.

As a future work, it is necessary to process systematically the variation of pronunciation between words. In order to prevent one syllable sentence from being misrecognized as similar multi-syllable sentence, context-dependent HMM models may be applied to this system.

ACKNOWLEDGEMENT

We express our appreciation to all members of Automatic Interpretation Section in ETRI who provide Korean speech DB necessary for training and recognition of phoneme model.

REFERENCES

J. H. Ryoo, K. Itou, S. Hayamizu, and K. Tanaka, 'Construction of Korean speech dialog system for hotel reservation', *Proc. ASJ Spring Meeting*, pp.35-36, 1994.3. (in Japanese)

N. Y. Hahn, K. W. Hwang, H. R. Kim, Y. M. Ahn, 'Implementation of a continuous speech recognition system using finite state network and Viterbi beam search', to be published, *Proc. of the 1994 Korean Signal Processing Conference*, 1994.10. (in Korean)

K. Itou, S. Hayamizu, and H. Tanaka, 'Continuous speech recognition by context-dependent phonetic HMM and an efficient algorithm for finding N-best sentence hypotheses', *IEICE Trans. Inf.&Syst.*, Vol.J75-D-II, No.6, pp.1023-1030, 1992.6. (in Japanese)

K. Itou, S. Hayamizu, K. Tanaka, and H. Tanaka, 'System design, data collection and evaluation of a speech dialogue system', *IEICE Trans. Inf.&Syst.*, Vol.E76-D, No.1, pp.121-127, 1993.1.