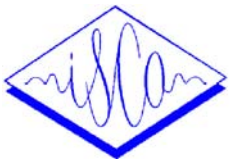
ISCA Archive
http://www.isca-speech.org/archive



EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

A KOREAN SPEECH CORPUS FOR TRAIN TICKET RESERVATION AID SYSTEM BASED ON SPEECH RECOGNITION

Woosung Kim and Myoung–Wan Koo
Multimedia Technology Research Laboratory
Korea Telecom
17 Umyon-dong, Seocho-gu, Seoul, 137 – 792, Korea.
Tel. +82–2–3290–5020, Fax : +82–2–3290–5007, E-mail:{sung,mwkoo}@smm.kotel.co.kr

ABSTRACT

This paper describes the Korean speech corpus for train ticket reservation aid system based on speech recognition. Two sets of speech corpus were collected. One was based on human-human(H-H) dialogues and the other was based on human-computer(H-C) dialogues. WOZ(Wizard of Oz) experiment was carried out to collect speech corpus based on H-C spoken dialogue. A total of 298 speaker data was collected for H-C corpus and a total of 100 speaker data was collected for H-H corpus. Since the basic unit of grammar in Korean is a morpheme, Korean-language model based on a morpheme was designed in addition to a word-based language model. Linguistic analysis results show that people respond differently when they are talking to a computer compared to when talking to a human. Also language-model analysis results reveal that a morpheme-based language model gives 50% reduction in perplexity(PP) over a word-based one.

1. INTRODUCTION

The recent advances in spoken language technology have made it possible for the automatic dialogue systems [1, 2] to communicate with human users as freely as if two humans were communicating with each other. We are currently developing a spoken dialogue system for the application of train ticket reservation. The system consists of a speech recognizer, a dialogue manager, and a speech synthesizer, and it runs on the public telephone network. As the beginning of this development we constructed speech corpus from H-C dialogues collected by WOZ simulation [3]. We also constructed H-H dialogue corpus for comparison reasons.

The basic unit of grammar in Korean which belongs to an agglutinative language is not a word but a morpheme. In other words, each word in Korean can change its form according to its use. For the very large vocabulary or unlimited speech recognition in Korean, it is very difficult to construct a language model with words because the possible number of words would be large. This is not to say that morpheme-based language is without problems. Although the basic unit of grammar in Korean is a morpheme, the short duration of some of the morphemes makes it difficult to be recognized by a speech recognizer. Here, our approach is based on a morphologically analyzed corpus which overcomes the difficulties posed by both a word and a morpheme.

In this paper, we present Korean speech corpus for train ticket reservation aid system based on speech recognition. Section 2 presents the H-H and H-C dialogues. Section 3 presents the basics of language models. Section 4 presents the WOZ procedure for data collection. Section 5 presents the linguistic analysis and language model analysis results. Finally section 6 makes conclusions.

2. HUMAN-COMPUTER AND HUMAN-HUMAN DIALOGUES

To improve the quality of the spoken dialogue system, researchers have begun to focus on H-C dialogues. One of their main concerns is the applicability of the H-H dialogue model to H-C dialogue. It is unclear whether this will improve the performance of the system. In general, people respond differently when interacting with a computer as opposed to a human. For example, people use smaller size vocabulary, shorter utterances, make fewer exchanges and so forth. [4] showed that the main differences between the two dialogues are turn-taking and grounding. We have observed that compared to the H-C dialogue case, in the H-H dialogue case turn-taking was more rapid and groundings such as "Yes." and "No." were more frequent.

3. LANGUAGE MODELS

Statistical language model assigns a probability value to string $W = w_1, \cdots, w_n$ where w_i is the i-th word in the string W as

$$p(W) = \prod_{i=1}^{i=n} p(w_i|w_1, \cdots, w_{i-1}) \approx p(w_i|w_{i-M}, \cdots, w_{i-1}) \ .$$

The main drawback of statistical language model is that it requires very large training data(data sparseness problem). To overcome this problem, various

methods have been proposed. They can be classified into two categories, smoothing and clustering.

Smoothing assumes that every events unseen or rarely seen while training can be seen while testing. So it prevents any probability from being zero, but it also attempt to improve the accuracy of the model as a whole. We used two representative smoothing methods, linear interpolation and backing-off. Linear interpolation tries to estimate the probabilities as a linear combination of lower order models. For example, linear interpolated trigram(M=2) can be estimated as

$$p(w_i|w_{i-2},w_{i-1}) = q_3\frac{C([w_{i-2},w_{i-1},w_i])}{C([w_{i-2},w_{i-1}])} + q_2\frac{C([w_{i-1},w_i])}{C([w_{i-1}])} + q_1\frac{C([w_i])}{N} \quad ,$$

where the function $C(\Omega)$ counts the frequency of the string Ω in the W, the nonnegative weights $\{q_i\}_{i=1}^3$ satisfy $q_1+q_2+q_3=1$ and N is the total number of words in the training corpus. Another smoothing method, back-off model extends the intuitions of Good-Turing by adding the interpolation of higher-order models with lower-order models [5, 6].

The other category is clustering, which reduces search space for speech recognizer as well as solves the data sparseness problem. The model from this technique is called class-based M-gram [7]. Let $G: w \rightarrow G(w) = g_w$ be a function that maps each word w to its class $G(w) = g_w$. We can model the probability of w_i given $w_1, \cdots, w_{i-1}$ as

$$p(w_i|w_1,\cdots,w_{i-1}) \approx p(w_i|w_{i-M},\cdots,w_{i-1})$$
$$= p(G(w_i)|G(w_{i-M}),\cdots,G(w_{i-1})) \times p(w_i|G(w_i)) \quad .$$

The main difficulty in using this model is the determination of the optimal class mapping function which reduces the overall perplexity.

In developing our dialogue system, each of the smoothing techniques mentioned above was used in a bigram model for comparison reasons. A class-based bigram model with linear interpolation which is constructed by automatic clustering algorithm to find optimal cluster mapping functions was also incorporated in our dialogue system [8] .

4. DATA COLLECTION

4.1. Wizard of Oz Simulation

To develop a spoken dialogue system which can interact with a human, we must determine the system specifications such as vocabulary, language model, dialogue model and so forth. In order to do so, we generally rely on the analysis results of H-H dialogues. Unfortunately, it is not known whether people interact with a computer the same way as they do with a human. For this reason, two sets of speech corpus were collected. One was based on H-H dialogues and the other was based on H-C dialogues. But it is impossible to collect speech data based on H-C dialogue since to do so requires the dialogue system itself which we are trying to develop. And the system is not available yet. Thus for the collection of speech data based on H-C dialogue, a human(wizard) plays the role of a computer, thus simulating. This method of collection speech data through simulation is called Wizard of Oz.

The most important issue in the WOZ experiment is to make the users feel that they are interacting not with a human but with a computer. We let the responses of a human wizard be similar to those of a computer. So we limited the wizard's responses to some predefined templates. Also we used the Korean unlimited TTS(Text-To-Speech) system [9] to simulate wizard's voice.

Figure 1 shows the block diagram of our system used for data collection. When a user call comes through the system via the PSTN(Public Switched Telephone Network), the system receives it automatically. User's speech is then automatically saved to several speech files after it has been segmented accordingly by an automatic end point detection process. The speech is also transmitted to the wizard at the same time so as to keep the dialogue going. After listening to the user's speech, the wizard checks the train ticket database to see whether the user's requests are available. We implemented a train-ticket database query system for simulation. From a set of predefined templates, the wizard makes its response. Also the wizard's responses are transmitted to the user as speech over the telephone line after the TTS process. We designed and implemented this database collection system which runs on IBM-PC with the telephone interface equipments.

4.2. Database

First, we collected speech data based on the H-C dialogues. A total of 298 speakers was chosen for the WOZ experiments. The statistical distributions of the speakers' ages, the telephone units(cellular phone, regular phone, speaker phone, etc.) and the background environment from where the telephone call was made were chosen to simulate the actual scenarios. Initially a speaker is given only the information on names of departure and arrival stations(sometimes departure station is not given), type of train, date, time, and identification numbers, and the speaker is free to change his/her ticket according to the wizard's responses.

To compare how people change their responses when they are talking to a computer or to a human, we also collected H-H dialogues in the same task domain. Even though the detailed scenarios given to each speaker are different from those cases of H-C

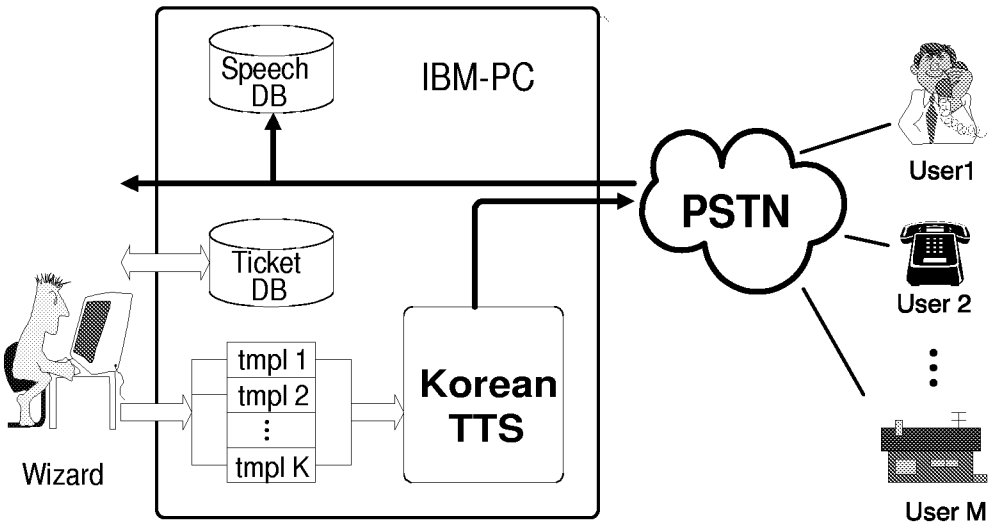


Figure 1: Block Diagram of Data Collection System

Table 1: Characteristics of Users for Database Collection

Ages	Number of Users		
	Male	Female	Total
20's	48	64	112
30's	44	55	99
40's	16	43	59
50's	17	11	28
Total	125	173	298

Table 2: Linguistic Analysis of Users' Responses Between H-H and H-C Dialogues

	H-H	H-C
# of Dialogues	100	900
Avg. Uttrs/Dialogue	20.39	14.53
Avg. Words/Uttr	2.25	2.61
Yes/No Rate(%)	37.47	8.31
Avg. Words/Uttr(NoGr)	3.00	2.80
Interjections Rate(%)	4.49	0.99

dialogues, we can make some comparisons between the two cases. All speech databases are transcribed and confirmed by human experts.

Table 1 shows the distribution of the speakers who participated in database collection with respect to their sex and age. There were more people in their 20's and 30's than in their 40's and 50's.

4.3. Morphological Analysis

In Korean which belongs to an agglutinative language, the basic unit of grammar is not a word but a morpheme. We developed a language model based on a morpheme. By doing so we were able to reduce the vocabulary size of the speech recognizer.

To develop this model the corpus based on a word was morphologically analyzed by an automatic analyzer [10]. It often gave multiple outputs resulted from ambiguity of Korean. Choosing correct one among these outputs was done by the human experts, then we constructed a corpus based on a morpheme with the help of the human experts.

5. ANALYSIS RESULTS

5.1. Linguistic Analysis

We linguistically analyzed the H-H dialogues and H-C ones to find out how human users differentiate their responses between two cases. Table 2 shows the analysis results. At first, the average number of utterances per dialogue in each cases indicates that users seem not to use much utterances while they were talking to computer than to human. The average number of words per utterance are higher in the H-C dialogues than in the H-H ones. This is due to the fact that "Yes/No" grounding is more frequent in the H-H dialogue cases than in the H-C ones. So if we remove "Yes/No" grounding utterances, the average number of words per utterance in each case becomes about equal. Also people use less interjections while interacting with a computer than with a human. These results match with general assumption and approve it.

We found out that a wizard's voice using Korean TTS was still unfamiliar to the naive users who have heard the sound from the TTS for the first time and it makes the users feel uncomfortable. It is clear that

Table 3: Language Model Analysis Results

	Corpus Based On	
	Word	Morpheme
Vocab(different words)	1476	791
Avg. Words/Uttr	3.96	5.00
Hit Ratio(%)	91	96
Bigram PP with LI	28.28	14.43
Bigram PP with BO	26.12	13.36
Class PP with LI	23.87	14.06

people interact differently when they are interacting with a computer than with a human and that iterative WOZ simulation is important for further update of the system.

5.2. Language Model Analysis

Of the 298 H-C dialogue data, we morphologically analyzed only 148. For language model analysis, we used 140 speakers' data for training and 8 speakers' one for test. Table 3 shows language model analysis results for both corpora. After we changed the word-based corpus into a morpheme-based one, the size of vocabulary was reduced while average words per utterance was increased. Though the bigram hit ratio was increased, perplexity was much reduced for all 3 models, bigram with linear interpolation(LI), bigram with back-off(BO) and class bigram with linear interpolation. A morpheme-based language model showed better results, it reduced the perplexity in half.

6. CONCLUSION

In this paper we described the speech corpus for train ticket reservation aid system based on speech recognition. We have been developing Korean spoken dialogue system for train ticket reservation domain. At first we collected database using WOZ simulation and analyzed it. And we compared it to that of H-H dialogues. The linguistic analysis results showed people responded differently in H-C dialogue case compare to in H-H dialogue case. Furthermore we morphologically analyzed the H-C dialogue corpus, then we constructed a morpheme-based language model. Using a language model based on a morpheme we were able to reduce both the perplexity by 50% and the size of the vocabulary. We are developing continuous speech recognizers based on both language models (morpheme-based and word-based ones) to make comparisons between the two.

7. REFERENCES

[1] L. Lamel, S. Bennacef, H. Bonneau-Maynard, S. Rosset, and J. L. Gauvain, "Recent developments in spoken language systems for information retrieval," in *ESCA Workshop on Spoken Dialogue Systems*, (Viso, Denmark), 1995.

[2] N. M. Fraser and J. H. S. Thorton, "Vocalist : A robust, portable spoken language dialogue system for telephone applications," in *Proc. of European Conf. on Speech Commnunication and Technology*, vol. 3, pp. 1947–1950, September 1995.

[3] N. M. Fraser and G. N. Gilbert, "Simulating speech systems," *Computer Speech and Language*, vol. 5, pp. 81–99, 1991.

[4] A. Johnstone, U. Berry, and T. Nguyen, "There was a long pause: influencing turn-taking behavioour in human-human and human-computer spoken dialogues," *Int. J. Human–Computer Studies*, vol. 41, pp. 383–411, 1994.

[5] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 400–401, March 1987.

[6] R. Rosenfeld, "The cmu statistical language modeling toolkit and its use in the 1994 arpa csr evaluation," in *Proceedings of the Spoken Language Systems Technology Workshop*, pp. 47–50, Morgan Kaufmann Publishers, January 1995.

[7] F. Jelinek, "Self–organized language modeling for speech recognition," in *Readings in Speech Recognition*, pp. 450–506, Morgan Kaufmann Publishers, 1991.

[8] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical language modelling," in *Proc. of European Conf. on Speech Commnunication and Technology*, pp. 973–976, 1993.

[9] E. I. Kim and J. I. Kim, "Hansori: An unlimited synthesis system," in *Proc. on Speech Communication and Signal Processing Workshop*, pp. 342–345, October 1994.

[10] Y. Hong, M.-W. Koo, and G. Yang, "A Korean morphological analyzier for speech translation system," in *Proc. Int. Conf. on Spoken Language Processing*, pp. 676–679, October 1996.