Abstract

We implemented a practical speech guidance system for public use. It is called "Takemaru-kun", and is located daily at the entrance hall of Ikoma Community Center to inform visitors about the center and Ikoma city via speech human-machine interface which includes a Takemaru animating agent. This system also aims at a field test of a speech application as a real world research platform, and collecting actual utterance data in the framework of human-machine interaction. The system has been in operation daily for the last 6 months, and a large number of user utterances have been recorded. In this paper, we describe an overview and the speech related parts of the system, and the current status of the field test and the analysis results of collected data. As a result of the field test, we obtained 1,362 minute length speech data including actual utterances from a wide variety of ages and genders. Although the tendency of the utterance contents is different between adults and children, it is shown that this system is used actively.

In a recognition experiment with extracted samples of adult voices, 86% word accuracy and 76% response correct rate are obtained. However for child voices, the recognition rate and response rate are not adequate. As future works, various developments of the improved components are planned in order to realize the practical speech interface.

Keyword: Speech interface system, Field test in real environment, Collection and Analysis of user's utterances

1. Introduction

Computers are becoming indispensable in daily life. Many people, however, do not find them friendly, because their communication interfaces (mouse, keyboard, ...) are very different from those of human being. As the spoken language interface is one of the most familiar means for humans to communicate with others, an intelligent human-machine speech interface will enhance a chance for us to use a computer in daily life.

Recently, large vocabulary continuous speech recognition has accomplished remarkable improvements in performance. Now we can recognize reading style speech with high accuracy. More than a 94% word recognition rate can be obtained even for a several ten thousand word task such as a newspaper article dictation. Although natural speaking style utterances are still difficult to recognize, human-machine dialogue utterances can also be recognized correctly. People can adapt their speaking style to a listener. It is known that people speak clearly and slowly using basic sentences and vocabulary when speaking to a machine. In our investigation, the recognition rate of 90.9%, which is equal to a reading-style recognition rate, was obtained in the human-robot dialogue task[1].

However, there are few practical speech interface systems running for public use. Why are they not used even with the high recognition rate? In order to investigate the reasons that prevent a speech interface system from being used as an interface utility, the development of a speech interface verification platform is indispensable. A long-term field test of the system is also needed in order to observe actual human-machine interactions in practical environments.

As an example, we have developed a reception guidance humanoid robot, called "ASKA"[1]. ASKA is a prototype research oriented robot which uses face recognition, speech recognition, voice response, robotics technology and internet communication. As a speech application, ASKA can recognize a user's question utterance on a university related topic, and answer in synthesized voice along with hand and head movement.

However, it becomes clear that it is still insufficient for natural human-machine interaction data collection. As
ASKA needs much effort for operation, it is not suitable for practical field test purposes. To develop and research speech interfaces further, the large-scale collection of natural interaction data is indispensable. Field tests of speech applications with a large vocabulary speech recognition have been carried out in few cases\[2\][3][4]. The evaluation and development of speech applications requires large scale speech data collection through real environment field tests.

For further speech data collection in natural speech interface, we have implemented a practical speech guidance system for public use. It is called “Takemaru-kun”, and is located daily at the entrance hall of Ikoma Community Center (Figure 1). Takemaru-kun can inform visitors about the center and Ikoma-city via speech human-machine interface which includes a Takemaru animation character. This system aims at a field test of a robust speech interface system in a practical environment and collecting actual utterance data in the framework of human-machine interaction.

In this paper, we describe an overview of the system and the speech related parts of the system in Sections 2 and 3. The current status of the field test and the analysis results of collected data are presented in Section 4. In Section 5, we evaluate speech recognition rates and response correct rates by using actual user’s utterances. The performance improvements by reconstruction of the models using collected data are also investigated in Section 5.4. We conclude this paper in Section 6, and future works on the speech interface application are also described.

2. Overview of Takemaru-kun

Figure 1 shows the hardware configuration of Takemaru-kun for the setup in the community center. The system has two display monitors, a mouse, a microphone and loud speakers. They are controlled by two Linux-based PCs which have an Internet connection.

When a user speaks to a microphone on the desk, a synthesized voice response is outputted from the loud speakers. The animation agent “Takemaru”, which is the symbolic character of Ikoma-city, is displayed on the left-hand monitor. The Takemaru animation agent performs animation gestures created using Macromedia[5] Flash. It now has 38 gesture patterns. Moreover, the agent can indicate the detection of utterance start to a user by nodding. The visual informations such as a Web page, a map and a timetable etc, are also displayed on the right-hand monitor, corresponding to the synthesized speech response. Figure 3 shows an example of a facility map of the center. Further Web retrieval is also possible with a mouse.

The speech interface program of Takemaru-kun can answer questions on the following topics.

- Guidance to the center’s facilities
- Guidance to the center’s services
- Town information for the neighborhood of the center
- Traffic guidance
Figure 4. Examples of response sentence candidates

- Weather report and news
- Sightseeing information for Nara and Ikoma
- Self-introduction of Takemaru-kun
- Greetings and others

These tasks are determined based on the results from preliminary collected questionnaires, and will be expanded further.

### 3. Configuration of speech interface

In the speech interface program of Takemaru-kun, a response sentence is generated by choosing a suitable one among the prepared response sentence candidates. A response sentence is chosen according to recognition results based on simple one-question and one-response principle[6]. The configuration of the speech interface program should have a simple design to enable the generation of a guide response to a user without time delay.

#### 3.1. Response generation

Figure 4 shows examples of response sentence candidates. There are two kinds of response sentence candidates. One is a fixed form text, and the other is a slot filling text. The number of registered response sentence candidates is now 202. Among these, three sentences about inquiries of time and date are registered as slot filling type sentences.

In order to make an appropriate response choice, the question-example text based scoring method which uses a large number of prestored questions and response examples is adopted. The outline of the question-example text based scoring method is illustrated in Figure 5. The question-example texts consist of 2309 sentences collected from user’s utterance transcriptions and city office’s journals (Figure 6). Each question-example text is attached to a suitable response text beforehand. The number of matched morphemes of independent part-of-speech between a question-example text and a recognized text is calculated for all prestored question-example texts, and totaled to a score of a response candidate corresponding to a question-example text. Then, the response candidate which has the highest score among the 202 response candidates will be selected as a result. At this time, only independent word morphemes (part-of-speech) are used for scoring, such as nouns, verbs and adjectives that are important for understanding of intentions. When two or more response candidates get the same score, one response will be chosen at random.

This procedure is able to generate a response text flexibly for various inputs without keyword definition. Various expression styles of utterances, such as ungrammatical expressions and different word orders, can be treated by increasing the number of question-example texts.

For the input to this response generation program, N-best recognized results are used. In our score calculation, the 100-best recognized results from the speech recognition engine are inputted into the response generation program.

#### 3.2. Speech recognition program

In this speech interface program, our speech recognition engine Julius[7] is used to obtain N-best recognized outputs. As the task-description approach in speech recognition, a
word trigram language model is adopted in order to cope with the wide task variation. Finite state network grammar was usually adopted for restricted small task recognition. The variation in our reception guidance task is too wide to write the whole grammar by using finite state networks. By using a statistical language model instead of a network grammar, some responses are correctly generated even for out-of-domain utterances. Utterances including various expression styles can also be recognized more flexibly than with the network grammar based recognition. Moreover, we can use collected logs to retrain the trigram language model. This framework enables improvement of the speech recognition accuracy at a low cost.

The task-suited word trigram language model for Takemaru-kun has been built according to the following method. In order to construct the language model, we collected the following training texts:

1. Web) Texts from Ikoma-city related web pages collected via the Web search site (texts are extracted by a statistical filter program[8]), including 1,080,272 sentences, 31,265,487 words and 218,723 difference words.

2. QA) Question sentence texts for the Takemaru-kun task collected by hands, including 6,488 sentences, 56,108 words and 3,231 difference words.

Two types of language models for Web) and QA) are created from each group of training texts. We use the back-off N-gram model construction procedure of the IPA Japanese free dictation program project[9] to train a trigram model. The vocabulary size for the Web) model is set to the 40,000 words which appear most frequently in the training texts. For the QA) model, all words in the training texts (3,231 words) are included. These two generated models are merged by using the N-gram model merging tool[10] (merge rate 1:1). Hereafter, this merged model is called the base language model.

We also created a grammar adapted language model in order to realize higher accuracy recognition. Word-pair constraints of the written grammar which strengthen the N-gram logarithm probability are applied to the base language model. By enhancing the words connection, in-task utterances that are expected by the system can be recognized more accurately, while keeping the acceptability of N-gram against unexpected utterances. The network grammar used for this adaptation is created by hands with 441 difference words. The logarithm probability value of a 2-gram entry with the grammar acceptable word-pair is increased 0.55 times. A back word-pair is used for grammar acceptance for a 3-gram entry (Figure 7). When the word is defined in the network grammar, the 1-gram word logarithm probability is increased 0.55 times.

Forty-three words are unknown in the base language model although they are defined in the grammar. These words are added to the model by giving output notations and readings to the unknown word class in the word dictionary.

As an acoustic model, we use the speaker-independent PTM[11] triphone HMM (Hidden Markov Model). The acoustic model is trained using the JNAS[12] speech data. To cope with noisy environments, exhibition hall noise is superimposed on the training speech by 25dB SNR.

4. Field test and utterance data collection

For a study of user's actual utterance to a public speech interface system, we have been recording all of the speech inputs to the system that are automatically segmented by the recognition engine, including clear sentence utterances, fragmentary utterances and invalid inputs such as noises and background voices. All of the collected data are manually transcribed, classified and tagged. The summary of the collected data is presented in Section 4.1.

4.1. Collected utterances

This system has been operated during every business day from the opening of the center (November 6, 2002) till March 31, 2003, for a total 125 days of operation, and 46,754 speech inputs have been recorded. This means that about 374 utterances were recorded per day in average. The total amount of the data is about 2.5 gigabytes in byte size, and 1,362 minutes in total.

Figure 8 shows the averaged numbers of collected data per day, averaged for each week. The number of clean utterances is also plotted in Figure 8. Our system has been utilized by general citizens regularly for a long period without any special promotions or human guidances.
Table 1. Age and gender classification of collected data

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Male</th>
<th>Female</th>
<th>Unknown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Small Child</td>
<td>115 / 76</td>
<td>1940 / 1421</td>
<td>1044 / 585</td>
<td>3099 / 2082</td>
</tr>
<tr>
<td>b) Lower Grade Child</td>
<td>2709 / 1920</td>
<td>9793 / 7961</td>
<td>4579 / 2843</td>
<td>17081 / 12724</td>
</tr>
<tr>
<td>c) Higher Grade Child</td>
<td>1142 / 934</td>
<td>1290 / 1154</td>
<td>673 / 498</td>
<td>3105 / 2586</td>
</tr>
<tr>
<td>d) Adult</td>
<td>5867 / 5520</td>
<td>2778 / 2496</td>
<td>103 / 70</td>
<td>8748 / 8086</td>
</tr>
<tr>
<td>e) Senior</td>
<td>10 / 8</td>
<td>12 / 12</td>
<td>0 / 0</td>
<td>22 / 20</td>
</tr>
<tr>
<td>x) Noise</td>
<td>2 / 0</td>
<td>6 / 0</td>
<td>14691 / 16</td>
<td>14699 / 16</td>
</tr>
<tr>
<td>Total</td>
<td>9845 / 8458</td>
<td>15819 / 13044</td>
<td>21090 / 4012</td>
<td>46754 / 25514</td>
</tr>
</tbody>
</table>

Figure 8. Average number of collected data per day

The collected data were then classified and transcribed manually. Each speech data is transcribed, and the gender and the age group of the speaker is also labeled. The collected data contain a large number of irrelevant and invalid inputs such as noises, coughs, laughter, level-overlowed shouts, and other unclear inputs. Clearness is also tagged for each speech input. All of the tagging work has been carried out by a single operator subjectively.

From the data classification, it was found that 21,240 inputs (almost 45.4%) were not clear, or were invalid inputs from which no semantic meaning can be extracted at all. We divided the collected data into two classes, and call the invalid inputs as "noise inputs", and the remaining 25,514 data set as "clean utterances" below.

The resulting classification according to age and gender is shown in Table 1. In the table, the left-hand count in each cell shows the number of whole collected data, and the right-hand count shows that of clean utterances. It is evident that we succeeded in collecting utterances from a wide variety of ages and both genders.

Among the inputs, 14,691 data were classified as both gender-undeterminable and age-undeterminable, and most of them contains only noises which are not intended as an input to the system and are triggered wrongly by input level, such as the noise of rubbing the microphone with hands, collision noises of facilities, distant background speech spoken by other visitors, and so on. As the number of such unintended inputs is not small and it may affect system interaction, it reveals that the rejection of such meaningless inputs is an essential issue which must be addressed for the realization of a practical speech interface.

The ratio of clean utterances differs remarkably between adults and children. The ratio of clean utterances was 92.4% in adults, and 74.7% in children. Most of the unclear inputs are inarticulate mumbling speech, and level-overlowed speech. It was found that 2,683 utterances of children are overlowed, while only 50 utterances are overlowed in adults. This result suggests that children are likely to behave with a rough attitude towards such an artificial agent system, while adults tend to speak politely.

4.2. Analysis of utterance topic

The topics of the utterances are also classified. All of the collected data, excluding gender-unknown ones, are labeled according to their intended topic category.

Figure 9 shows a comparison between adult speakers and child speakers, for the percentages of the labels in the database.

In this figure, the label "Guidance" is given to the utterances that include queries about a room, a facility, a location, the timetables of buses and trains and other questions about the center and its surroundings. The label "Take-maru" indicates utterances asking or speaking about personal matters for the software agent Take-maru itself, including questions about name, age, favorite foods, birthday, and so on. “Greeting” label is given to common greetings, that are usually uttered when users contact Take-maru for the first time. “News & Time” is a label for query utterances relating to current news topics or the current time. “Others” is a label for out-of-task utterances to which the system cannot
reply. It also includes utterances whose speech is clear but the meaning is wholly uncertain. The “Others” label also includes the user’s response utterances to the system, such as “I see” or “repeat, please”, since such utterances are not dealt with properly. The “Unclear” label is assigned to other types of meaningless speech, including unclear utterances, shouts and so on.

In adult speakers, 31.2% of utterances are classified as “Guidance”, while the child speakers’ utterances contain only 13.5% of guidance query utterances. On the other hand, the ratio of children’s utterances labeled “Takemaru” is higher by 8.7% than that of adults. This indicates that adults are likely to attempt to utilize our system to obtain information according to our system design. However, children are much more interested in the personality of the agent than what the system can provide.

From this comparison, it is apparent that most adults and children have different attitudes towards a spoken interface system.

5. Experiments

In the experiments, we evaluate speech recognition rate and response correct rate of our system in order to evaluate

the baseline performance of our speech interface program. For the test set, 1,000 utterances are extracted from the 25,514 clean utterances described in Section 4.1. To keep the variety of utterance contents in the test set, utterances of the same transcription are discarded. Specifications of the test set sentences are shown in Table 2. Figure 10 and Figure 11 show examples of utterance sentences for the evaluation test sets in Japanese.

5.1. Evaluation of speech recognition

We carried out the large vocabulary continuous speech recognition experiment with Julius[7] on the test set. The grammar adapted language model described in Section 3.2 was used as a trigram language model. For the comparison of acoustic model, exhibition hall noise superimposed PTM triphone HMM model, JNAS clean acoustic model, CSRC female model and CSRC child model are used. The JNAS clean acoustic model is a gender-independent PTM triphone HMM model trained by the clean voice of the JNAS database. The CSRC female model and the CSRC child model are also PTM triphone models distributed by the Continuous Speech Recognition Consortium (CSRC)[13]. Table 3 shows experimental results in word correct rate (Corr) and word accuracy (Acc). It is found that the recognition accuracy of a child voice is significantly worse than that of an adult voice. In this experiment, a 1.6% improvement in word accuracy is achieved by using the female acoustic model, which is closer to the child model than the JNAS model. Furthermore, using the child model...
Table 3. Results of speech recognition experiment (%)  

<table>
<thead>
<tr>
<th></th>
<th>Adult</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exhibition noise added</td>
<td>85.6</td>
<td>59.5</td>
</tr>
<tr>
<td>JNAS clean</td>
<td>84.2</td>
<td>55.4</td>
</tr>
<tr>
<td>Female (CSRC)</td>
<td>58.4</td>
<td>39.5</td>
</tr>
<tr>
<td>Child (CSRC)</td>
<td>63.2</td>
<td>46.7</td>
</tr>
</tbody>
</table>

Brings about an 8.8% improvement of recognition accuracy. These results confirm the known fact that the recognition accuracy of a child voice can be improved by re-training an acoustic model using a child voice corpus in order to reduce the acoustic feature mismatch between the adult and child voices[14]. However, the accuracy of child speech recognition is still not enough. In the recognition of a child’s voice, the difference between the word correct rate and the word accuracy is larger than in the recognition of an adult voice. The insertion errors in child speech recognition indicates that their spontaneous utterances causes much mismatch. Children cannot speak smoothly, and their utterances contain many hesitations and unnecessary words. Therefore, in addition to the reduction of acoustic feature mismatch, a correction technique for false-start utterances is also required for child speech recognition.

As compared with clear model, the exhibition noise added model shows better performance. It confirms that adaptation against noise is necessary.

5.2. Evaluation of response correct rate

Next, the correct rate of response generation was investigated. Here we define the response correct rate to be the rate from which the satisfactory response result is obtained. The response sentences were generated from the 100-best speech recognition results. The decision of whether the response contents is satisfactory or not was performed on the basis of subjective judgment by one person.

Response correct rates for 500 test set utterances are shown in Table 4. A response correct rate of 76% is observed for adult utterances. However, it is found that a child’s response correct rate is much lower than the adult as well as a speech recognition result. By using the child acoustic model, a 2.6% recovery of response correct rate is obtained.

5.3. Summary of baseline performance experiments

From the experiment results described above, it can be said that this system is reasonably equipped with baseline performance as a real environmental research platform required for interaction observation. However, the performance is not sufficient especially for children. Although only clear utterances were used in this experiment, many of children’s utterances include level-overlusted and unclear inputs. In order to continue the field test, introduction of dialogue processing for children and child speech recognition technology is necessary as well as improvement of recognition performance.

5.4. The reconstruction of the models using collected data

We also investigated how speech recognition performance can be improved by retraining the models with collected data. Both the acoustic model and the trigram language model were newly trained using the collected data. Specifications of training data for the model-reconstruction are shown in Table 5 and Table 6.

The training data for the trigram language model reconstruction consists of 24,514 sentences, which are made from transcriptions of the 25,514 clean utterances described in Section 4.1. 1,000 sentences for the evaluation test sets in Section 5 are excluded from these training data. The trigram model created from these training data is merged into the base language model in Section 3.2 (merge rate 1:1). And we also adapted the grammar constraints to the created model using the word-pair enhancing procedure in Section 3.2. This consummated model can be considered as a speaker-independent trigram language model specialized in the Takemaru-kun task.

A child-matched acoustic model is trained for acquiring sufficient accuracy in child speech recognition. For training the acoustic model, children’s clear utterances collected by the Takemaru-kun system, and children’s speech database offered by the NTT-AT Corporation[15] were prepared. In order to obtain a sufficient quantity of training data, the JNAS female speech data were also used to train the acoustic model.

We carried out speech recognition experiments with Julius using the 500 children’s test sets. Figure 12 shows the experimental results for word accuracy. For comparison, the word accuracies in the case of the original grammar adapted language model and the CSRC child acoustic model are also shown.
Table 5. Specifications of training data for reconstructing the trigram language model

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sentences</td>
<td>24,514</td>
</tr>
<tr>
<td>Number of total words</td>
<td>127,836</td>
</tr>
<tr>
<td>Number of difference words</td>
<td>3,962</td>
</tr>
</tbody>
</table>

Table 6. Specifications of training data for reconstructing the acoustic model

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children’s utterances</td>
<td>15,952</td>
</tr>
<tr>
<td>collected using Takemaru-kun</td>
<td></td>
</tr>
<tr>
<td>Children’s utterances</td>
<td>10,000</td>
</tr>
<tr>
<td>provided by NTT-AT</td>
<td></td>
</tr>
<tr>
<td>JNAS female utterances</td>
<td>20,016</td>
</tr>
</tbody>
</table>

Figure 12. Results of speech recognition using reconstructed models

It is found that the reconstructed models bring high accuracy in child speech recognition. In this experiment, a 62.4% of word accuracy is achieved by reconstructing both the language model and the acoustic model with training data collected within the task. Especially, over 10% improvements in word accuracy are obtained by adopting the reconstructed language model.

However, we gained only 4% of accuracy improvement by the acoustic model adapted training data. Although this adapted acoustic model attains the improvement of recognition accuracy, a development of more effective adaptation method using collected data will be considered in future works.

6. Conclusions and future works

6.1. Conclusions

We have developed the speech application system “Takemaru-kun” for research platforms of a practical speech interface. Takemaru-kun can inform visitors about the Ikoma Community Center and Ikoma city via a speech human-machine interface and a funny animating agent. The question-answer speech interface of this system is implemented on the basis of large vocabulary continuous speech recognition with the trigram language model. A response is generable with high precision and robustness by using the task-suited language model and the question-example text based response selection method. In a recognition experiment with actual adult users’ utterances, 86% word accuracy and 76% response correct rate are obtained. This system is set up daily in the community center entrance hall and realizes field test in a real environment. Users always can enjoy their communication with Takemaru-kun.

In this research, the collection of actual users’ utterance data has been performed through the field test for about five months. The natural utterances of users in a broad age range were recorded on a large scale. By analyzing the collected data, we showed that this system is used regularly by citizens. Furthermore, the result of classification according to utterance topics shows that this system is practically helpful for people. Therefore, the usefulness of this system is proved.

The models reconstructed using collected data realize a 16% improvement of recognition accuracy as experimental results. This result shows the importance of actual data collection.

6.2. Future works

As future works, various improvements of the system and developments of the component engineering are planned. In this process, ensuring the effective use of collected data is important.

From the evaluation using collected data, the necessity for the improvement of child voice recognition is indicated although this system is equipped with the required baseline performance for adults. A child’s natural utterances have many unclear utterances, for example level-overlarded inputs. We plan to improve the accuracy of child speech recognition by applying an acoustic model adaptation using collected data. However, an effective countermeasure against correction or unclear utterances is also required to obtain high-precision child speech recognition. As a solution to this problem, a new response generation routine which asks again for an unclear utterance is needed. For the realization of this routine, a confidence measure method of generated responses is considered essential.

To deal with the difference in the utterance contents of an adult and a child, we develop a switching method for the response generation according to the age group. The technique of distinguishing speaker’s age group is currently under examination.

First International Workshop on Language Understanding and Agents for Real World Interaction, 2003
Regarding the noise problem, investigation of an automatic method of discarding the noise input is essential as a future work.

Finally, we consider that the most important element in our work is the continuation of the field test. The users' treatment of the system is very rough, and hardware failure sometimes occurs. Continuing the service at the public place requires a great effort. We intend to overcome these difficulties and continue a long-term field test, aiming at the realization of a useful speech interface. With these efforts and the collected data, we will be able to improve our speech oriented information system further.

Acknowledgment We greatly appreciate the support provided by the Ikoma-city personnel in this project, with special thanks to Mr. Kazuhsa Shiramoto (Ikoma City Office) and Mr. Shuichi Yoneda (Superintendent of Ikoma North Community Center).

References


