Operating A Public Spoken Guidance System In Real Environment

Ryuichi Nisimura[†], *Akinobu Lee*[‡], *Masashi Yamada*[††], *Kiyohiro Shikano*[††]

[†] Faculty of Systems Engineering, Wakayama University
nisimura@sys.wakayama-u.ac.jp
[‡] Department of Computer Science, Nagoya Institute of Technology
ri@nitech.ac.jp
[††] Graduate School of Information Science, Nara Institute of Science and Technology
{masasi-y, shikano}@is.naist.jp

Abstract

Takemaru-kun system is a practical speech-oriented guidance system developed to examine spoken interface through long-term operation in a public place that collected natural human-machine interaction data. In 2004 the following advances improving reliability of the system were introduced, which conduced acquiring positive increase of access from users: (1) Rejection of unintended speech based on Gaussian Mixture Models (GMMs); (2) Removal of short, unnecessary inputs of impulsive noise; (3) Child or adult user discrimination; (4) Web-based monitoring mechanisms. This paper summarizes the Takemaru-kun system and analysis of 177,789 data collected by two-years actual operation. Experiments with the collected data proved that a combination of GMM-based verification and short input removal can excise 85% of the invalid inputs, including laughter, incomprehensible utterances, and even some background utterances.

1. Introduction

In recent years, spoken dialog system developments have actively sought to realize human-intimate interfaces. The range of applications is spreading widely into such areas as communication robots, car navigation systems, and telephone call systems. A few examples have actually succeeded in being utilized in our daily lives. Further evaluations and developments of spoken interfaces require large-scale speech data collection in actual environments. However, field testing spoken interfaces have rarely been carried out[1][2][3][4][5][6]. Thus, the amount of data is still clearly insufficient to consider natural human-machine interaction.

We have developed a practical speech-oriented guidance system for public use called the "Takemaru-kun"[1] system whose development started in November 2002[7]. It aims at a long-term field test of a robust spoken interface system in a practical environment and collecting actual utterances in a framework of human-machine interaction. The system has been located daily at the entrance hall of the Ikoma Community Center for two years to inform visitors about the center and Ikoma City via a speech human-machine interface with graphical animations and related Web pages (Figure 1).

In 2004, we implemented the following advances to the system that improve its reliability and usefulness:

[1] "Takemaru-kun" character is the symbol mascot of Ikoma City, and its name originates from the fairy of the bamboo in Japanese.

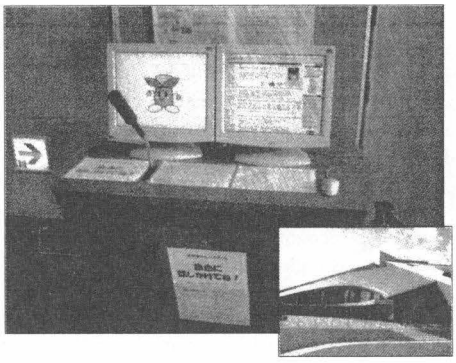


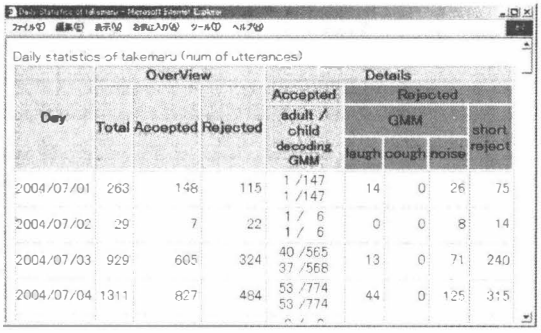
Figure 1: Speech-oriented guidance system "Takemaru-kun", and the Ikoma North Community Center.

1. Rejection of unintended speech based on likelihood measurements using Gaussian Mixture Models (GMMs).
2. Removal of unnecessary input that have short input length for impulsive noise.
3. Discrimination between child and adult users on the basis of speech recognition scores using two parallel decoders.
4. Internet Web-based monitoring mechanisms that check the system's condition while gathering numerical statements including the number of recorded utterances, rejected inputs, and user's age group. (Figure 2)

This paper consists of five major sections. We will first describe the architecture of the Takemaru-kun system in Section 2. Section 3 reports the results of the field test and analysis of collected data. Improvements of unnecessary input rejections brought by a combination of GMM-based verification and short input removal are examined in Section 4. We conclude this paper in Section 5 and also describe about future works.

2. Overview of Takemaru-kun system

The architecture of Takemaru-kun system is shown in Figure 3. Its spoken interface was designed as a simple one question one response strategy to accomplish daily guidance for users without time delays. A user's speech is captured by a single microphone, and a response is output by a synthesized voice



Daily statistics of takemaru (num of utterances)

Day	OverView			Details				
	Total	Accepted	Rejected	Accepted	Rejected			
				adult / child decoding GMM	GMM			short reject
					laugh	cough	noise	
2004/07/01	263	148	115	1 /147 1 /147	14	0	26	75
2004/07/02	29	7	22	1 / 6 1 / 6	0	0	8	14
2004/07/03	929	605	324	40 /565 37 /568	13	0	71	240
2004/07/04	1311	827	484	53 /774 53 /774	44	0	125	315

Figure 2: Snapshot of the Takemaru-kun monitoring mechanism through the Internet.

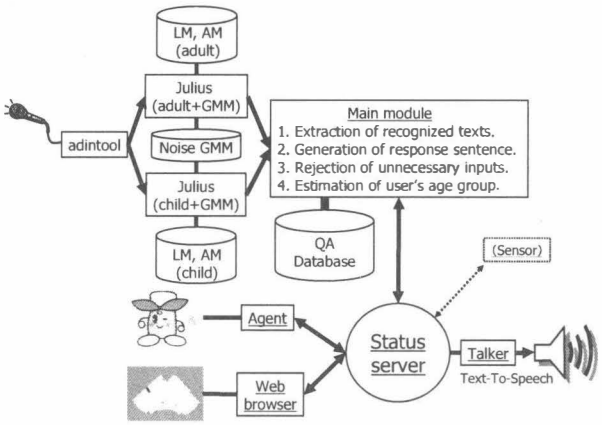


Figure 3: Takemaru-kun system architecture.

generated by a Text-To-Speech program that also provides animated gestures and related Web pages on a screen.

2.1. Software configuration

We designed this architecture based on a blackboard model comprised of four modules: Main, Agent, Web browser, and Talker. They communicate with each other through a status server that shares the states of all modules via TCP/IP. Each module operates independently and can stop and start at anytime. This modularity simplifies the development of each part.

The main module recognizes input speech to texts and generates responses automatically by choosing a suitable one from the prepared response sentence candidates. In addition, rejection of unnecessary inputs and estimation of user age groups are also performed in the same module. The talker module synthesizes the response speech according to the generated response. The agent module displays animation gestures synchronized with response speech created on Macromedia Flash. Agents can also indicate the detection of the start of an utterance to a user by nodding. Visual information such as Web pages, maps, and timetables are also displayed by Web browser module. For further Web retrieval, manual operation with a mouse is also possible.

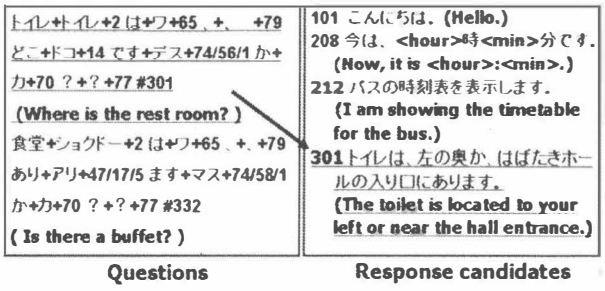


Figure 4: Examples of QA database.

2.2. Creating response

To make appropriate response choices automatically, we prepared an example question-answer (QA) database beforehand. It consists of actual questions queried to the system, which are morphologically analyzed from transcriptions of user utterances. Each question was attached to a suitable answer. Figure 4 shows examples of the QA database, where the answer to "Where is the rest room?" is defined as #301. After recognized text is inputted, the number of matched morphemes of independent parts of speech between a question and recognized text is totaled for all prestored questions. In this procedure, N-best output is used as speech recognized result that complement recognition errors. Then each score is determined by dividing the number of matched morpheme by words in the questions. A response candidate attached to the best matched example will be selected as a response. An advantage of this approach is that it generates a certain response concerned with a guidance task if the QA database is satisfied. The system never produces such meaningless responses as "I don't know" or "Could you repeat that please?," which confuses users and inhibits the operation of the system. We prepared the QA database that could cover a wide area related of the topic of users' interests. In addition, after the system was operated, we examined the log and added additional the QA database as judged necessary.

2.3. Rejection of unnecessary inputs

To reject unnecessary inputs, we investigated speech verification to determine whether the inputted voice was intended by comparison of acoustic likelihoods given by GMMs[8]. GMMs have proven to be powerful tools for text-independent speaker verifications[9]. Although conventional speech verification studies have only focused on the rejection of environmental noise, our proposed method can also consider more utterance-like wrong inputs such as laughter, coughing, and wrongly triggered background speech. In addition to GMM-based verification, the removal of short length inputs is incorporated into the 2004 Takemaru-kun system. The improvements of rejection derived from combining GMM-based verification and short input removal will be evaluated in Section 4.

2.4. Child or adult user discrimination

Two parallel speech recognizers were installed to estimate user age groups while achieving sufficient speech recognition accuracy. Each has an age group-dependent language model (LM) and an acoustic model (AM) suitable for adult or child users. Outputs are chosen based on comparisons between the two like-

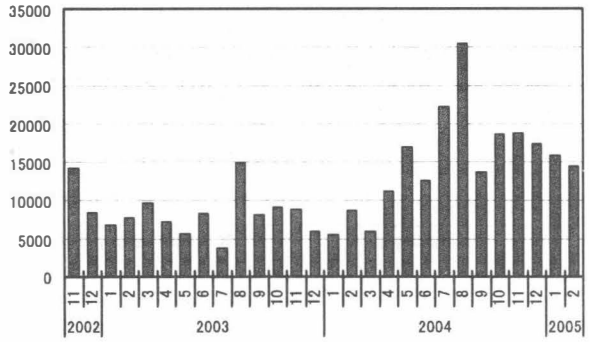


Figure 5: Number of data collected per month.

lihood scores from each decoder. The speech recognizer is our open source speech recognition engine Julius[10]. For the language model, task-dependent word 3-grams are trained by Web texts related to guidance tasks and transcriptions of questions collected through the operation of the system. The vocabulary size was set to 40,000 words. Speaker independent phonetic tied-mixture (PTM) triphone models were adopted as acoustic models and trained by collected utterances and reading style speech extracted from the JNAS newspaper database. See [11] for details.

3. Data collection and analysis

The operation of Takemaru-kun system started on November 6, 2002, and operated every business day at the entrance hall of the Ikoma Community Center. 328,288 inputs were recorded in 28 months by February 2005; voice segmentation from inputs using a raw level threshold and zero cross counts method were performed by Julius. This means an average of about 490 inputs were recorded per day. The amount of data reached about 17.9 gigabytes and 158 hours.

Figure 5 shows the number of data collected per month. The system regularly obtained actual utilization by general citizens for a long period without special promotions or human assistance. After implementing the new functions in April 2004, the number of access from users were increased because the system acquired improvements of reliability. The system attained twice growth of access in comparison with the previous year.

All collected data were manually transcribed, classified, and tagged for smooth detailed analysis by operators who subjectively listened to it. We have completed 177,789 data from the beginning to May 2004. Age group classifications of the data are illustrated in Figure 6, showing that 58.1% was uttered by children. Thus, we should accept the necessities of service for children when a spoken interface is located in a public place.

The collected data included such invalid inputs as unintended or unclear speech, fragmentary utterances, noises, and background speech. We classified all inputs as valid or invalid by hand, and categorized them by age groups in Table 1. Valid inputs were about 80%, except for the infants. The system succeeded in being utilized with obvious intention of talking to agents by many users, particularly adults. 96% of the inputs whose age group was deemed uncertain by operators were classed as invalid. They consisted of unnecessary inputs that should have been rejected. 37% of the invalid inputs were caused by mis-triggered recordings that only contained noise.

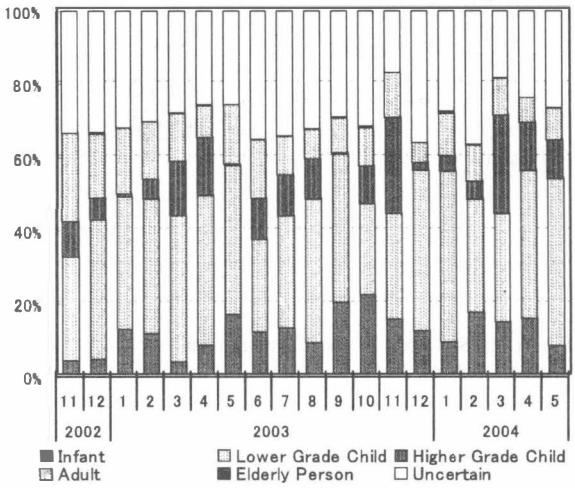


Figure 6: Populations by age group per month.

Table 1: Number of valid/invalid inputs.

Age group	Valid	Invalid	Undetermined	Validity rate
Infant	13,535	5,534	743	68%
Lower Grade Child	53,422	9,605	2,003	82%
Higher Grade Child	15,605	2,548	358	84%
Adult	19,193	2,024	455	89%
Elderly Person	244	67	3	78%
Uncertain	2,214	50,034	200	4%

4. Experiments of unnecessary rejection

The rejection of unnecessary inputs was carried out experimentally. Test set samples included the 8,248 data extracted from the collected inputs and excluded from the GMMs training.

Table 2 shows the experimental results and the number of inputs by category. Because of category duplication, the sum of the inputs and the number of test sets do not agree.

"Rejected by short" indicates the number of inputs removed due to their short length. In these experiments, a short input was defined as under 0.8 seconds, and they were discarded as unnecessary.

Next, GMM-based verification was applied to the remnants of short input removal to extirpate invalid inputs. Table 3 shows the conditions of GMM training and the number of data set. We prepared 5-class GMMs with 128 Gaussian mixtures from the training data classified: child, adult, laughter, coughing, and other. These classes were defined according to the amount of each categorized data existing in the training sets. That is, invalid inputs which had the few amount of data were treated as "Other".

This combined procedure rejected 85.1% of the invalid inputs as experimental results. The majority were rejected by short input removal. GMMs assisted rejection of utterance like wrong inputs such as laughter, coughing, and background speech. Mis-rejection of valid inputs was only 16.6%, which of-

Table 2: Results of rejection experiments.

Category	# of inputs	Rejected by short	Rejected by GMMs	Rejection rate [%]
Valid inputs				
Total	4,450	567	171	16.6
Invalid inputs				
- Out of domain utterances	381	99	45	37.8
- Incomprehensible utterances	455	248	175	93.0
- Background utterances	567	244	220	81.8
- Laughter	160	47	111	98.8
- Crying	11	0	10	90.9
- Breathing	69	11	58	100.0
- Coughing	38	16	22	100.0
- Beginning of sentence omitted	309	178	33	68.3
- Contained noise	167	65	44	65.3
- Noise	1,139	720	419	100.0
- Level underflow	300	299	1	100.0
- Unclear voice	244	125	67	78.7
Total	3,654	1,957	1,153	85.1

Table 3: Training conditions of GMMs

Amount of Training data	Valid inputs	Child	20,016
		Adult	4,065
	Invalid inputs	Laughter	849
		Coughing	98
		Other	6,413
Sampling rate/bit	16 kHz, 16 bit		
Window width/shift	25/19 msec		
Parameter	MFCC (12 dim.), ΔMFCC, ΔPower		

"Other" includes background speech and miscellaneous noises.

ten occurred when dealing with rapid utterances such as greetings and slanders.

5. Conclusions

This paper summarized the architecture of the Takemaru-kun system. It has a practical spoken dialog interface to realize a public guidance in the Ikoma Community Center. Two-years actual operation described in Section 3 showed that the system is regularly used while obtaining 490 inputs per day. We also described the following advances introduced to the 2004 system: (1) Rejection of unintended speech based on GMMs; (2) Removal of short, unnecessary inputs of impulsive noise; (3) Child or adult user discrimination; (4) Web-based monitoring mechanisms. These advances contributed improvements of reliability, which conduced acquiring positive rise of the number of access from users. As a result, the system realized twice growth of access in comparison with the previous year. It suggests the necessity of large-scale data collection to investigate the actualities of how users employ a spoken interface.

In experiments, we evaluated the rejection of unnecessary inputs included in actual human-machine utterances. The combination of GMM-based verification and short input removal produced a certain advantage for this scheme and achieved a 85.1% rejection rate for invalid inputs.

In future work, we plan to continue the operation and data collection in various situations. Evaluations and developments using more huge amounts of data will be actualized. Further developments to have flexibility for unanticipated errors are required. It is important to improve acquiring paralanguage information on the basis of statistical methods exploiting large amounts of collected data. We will also perform detailed analysis of the data that resolve difficulty of data collection, which is a major bottleneck in developing dialog systems[12].

6. Acknowledgment

A part of Takemaru-kun project is supported by the e-Society project provided by MEXT (Ministry of Education, Culture, Sports, Science and Technology), Japan. The authors greatly appreciate supports when filed testing the system by the Ikoma City office and the Ikoma North Community Center.

7. References

[1] E. Hurley et al., "Telephone Data Collection Using The World Wide Web," in *Proc. ICSLP96*, vol.3, pp.1898–1901, 1996.

[2] V. Zue et al., "JUPITER: A Telephone-based Conversational Interface for Weather Information," in *IEEE Trans. on Speech and Audio Processing*, vol.8, no.1, pp.100–112, 2000.

[3] L. Bell et al., "Child and Adult Speaker Adaptation during Error Resolution in a Publicly Available Spoken Dialogue System," in *Proc. EUROSPEECH2003*, pp.613–pp.616, 2003.

[4] N. Kawaguchi et al., "Multimedia Corpus of In-Car Speech Communication," in *Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol.36, pp.153–159, 2004.

[5] T. Giorgino et al., "Automated spoken dialog system for hypertensive patient home management," in *International Journal of Medical Informatics*, vol.74, no.2-4, pp.159–167, 2005.

[6] V. Goffin et al., "The AT&T WATSON Speech Recognizer," in *Proc. ICASSP2005*, vol.1, pp.1033–1036, 2005.

[7] R. Nisimura et al., "Takemaru-kun: Speech-Oriented Information System for Real World Research Platform," in *Proc. First International Workshop on Language Understanding and Agents for Real World Interaction*, pp.70–78, 2003.

[8] A. Lee et al., "Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs," in *Proc. INTERSPEECH2004-ICSLP*, 2004.

[9] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," in *Speech Communication*, vol.17, pp.91–108, 1995.

[10] A. Lee et el., "Julius – An Open Source Real-Time Large Vocabulary Recognition Engine," in *Proc. EUROSPEECH2001*, pp.1691–1694, 2001.

[11] R. Nisimura et al., "Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability," in *Proc. ICASSP2004*, vol.1, pp.433–436, 2004.

[12] Y. Gao et el., "Portability Challenges in Developing Interactive Dialogue Systems," in *Proc. ICASSP2005*, vol.5, pp.1017–1020, 2005.