INSIGHTS GAINED FROM DEVELOPMENT AND LONG-TERM OPERATION OF A REAL-ENVIRONMENT SPEECH-ORIENTED GUIDANCE SYSTEM

Tobias Cincarek[1], *Ryuichi Nisimura*[2], *Akinobu Lee*[3], *Kiyohiro Shikano*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]Faculty of Systems Engineering, Wakayama University, Japan
[3]Department of Computer Science, Nagoya Institute of Technology, Japan
cincar-t,shikano@is.naist.jp

ABSTRACT

This paper presents insights gained from operating a public speech-oriented guidance system. A real-environment speech database (300 hours) collected with the system over four years is described and analyzed regarding usage frequency, content and diversity. Having the first two years of the data completely transcribed, simulation of system development and evaluation of system performance over time is possible. The database is employed for acoustic and language modeling as well as construction of a question and answer database. Since the system input is not text but speech, the database enables also research on open-domain speech-based information access. Apart from that research on unsupervised acoustic modeling, language modeling and system portability can be carried out. A performance evaluation of the system in an early stage as well as late stage when using two years of real-environment data for constructing all system components shows the relative importance of developing each system component. The system's response accuracy is 83% for adults and 68% for children.

Index Terms— Guidance System, Real Environment, ASR Models, Q&A Database

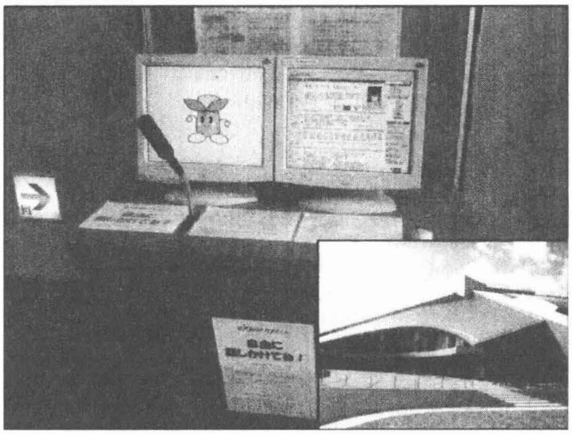

Fig. 1. Speech-oriented guidance system "Takemaru-kun" installed at the North Community Center in Ikoma City, Nara Prefecture, Japan.

1. INTRODUCTION

Many researchers have been trying to build spoken dialogue systems by implementing a truely intuitive human-machine interface, integrating speech recognition and language understanding technology. However, only very few systems have been accepted widely as as a convenient and user-friendly interface to the underlying service.

Spoken dialogue systems may be categorized into rather system-driven, goal-oriented systems and rather open-domain, access-oriented systems. Examples for goal-oriented systems are flight reservation [1], train reservation [2] or bus information [3]. The system's scope is most often defined by the developer and the dialogue emerging from using such systems is most often system-driven. The dialogue grows in length mainly due to misunderstanding by the system or system strategies to avoid misunderstanding. The convenience of such systems for the user may be questionable in general. There are immense research efforts for automatically learning dialogue strategies from dialogue corpora, e.g. [4]. Since the optimality of a dialogue strategy itself depends on the user model [5], it is hard to say which strategy to prefer.

Access-oriented dialogue systems such as for call routing [6], speech-activated text retrieval [7] or speech-oriented guidance [8] are mainly user-driven and give a much more direct and immediate response to the users input. This avoids lengthy misunderstanding and clarification dialogues. If the system's response is not as expected, the user can immediately reformulate his request or just give it a second try. An important aspect of such systems is, that they are open-domain. The scope of such systems is in the end more determined by the users' inputs and less influenced by the developers ideas. From these circumstances arises the necessity to update the models of the ASR engine and the content database to generate a satisfying response for each user input.

This paper reports about insights gained from development and long-term operation of the speech-oriented guidance system Takemaru-kun [8]. After briefly describing the system's architecture, an analysis of speech collected with the system installed in a real environment over four years is given. The database opens up research opportunities on problems such as language modeling, acoustic modeling, speech-based information access, system portability and unsupervised model construction. With the first two years of the data being completely labeled and transcribed by humans, performance simulation of the system over time for various development techniques of the system components is also possible. An evaluation of the system when using a few month vs. two years of labeled data shows the relative importance of effort for constructing the acoustic model, language model and the question and answer database.

A part of this work is supported by the MEXT COE and e-Society project, Japan.

2. SYSTEM ARCHITECTURE

Figure 2 shows a block diagram of the speech-oriented guidance system. User input is recognized in parallel with a children and adult-dependent acoustic (AM) and language model (LM). Voice activity detection and GMM-based rejection of too noisy or nonverbal inputs is integrated in the open-source LVCSR engine Julius [9]. The age group of the speaker is determined based on the acoustic likelihood. A multimodal response using voice, graphics, text and animation is selected from a separate Q&A database for each age group. Response selection is based on a similarity measure between the n-best recognition hypotheses and example questions. The response score is defined by the number of matching words divided by the maximum number of words in a hypothesis and the example question. More details about the architecture, a performance evaluation of adult/child discrimination, rejecting unusable input (both having an accuracy greater than 85%) and preliminary results of ASR performance have been reported in [8, 10].

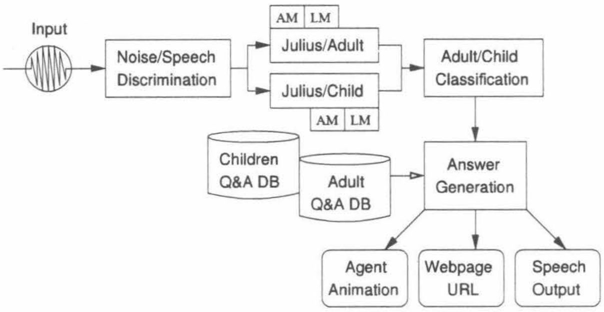

Fig. 2. The system's main building blocks.

3. DATA AND STATISTICS

The speech data collected over almost four years from November 2002 to August 2006 is shown in Table 1. Inputs collected from the beginning until September 2004 are are completely transcribed, labeled with tags (e.g. noise) and classified subjectively into five speaker groups: preschool children, elementary school children, junior-high school children, adults and elderly persons. Given the ratio of speech vs. noise is 5:1 for the labeled data as given in the table, it may be assumed that approx. 250 hours of speech and approx. 50 hours of noise data have been collected in total.

Table 1. Speech data collected with Takemaru-kun: relative share, total number and total duration noise and speech inputs for each speaker group.

Category	rel.share	# data	Time
Preschool Children	10.1%	27,535	14.3 hrs
Elementary School	39.0%	106,797	57.7 hrs
Junior-high School	11.5%	31,402	15.8 hrs
Adults, Elderly	11.3%	31,100	14.1 hrs
Noise, Non-Verbals	28.1%	76,864	19.3 hrs
Labeled Inputs		273,698	121.2 hrs
Unlabeled Inputs		388,249	174.1 hrs
Total # Inputs		661,947	295.3 hrs

Figure 3 shows the variation of collected inputs over time. Local peaks in the number of inputs are reached during the summer holidays in August 2003 and August 2004. Most of the inputs are from children showing that the Takemaru system is an easy way to collect spontaneous children speech. The average number of daily inputs during 2005 and 2006 is larger than in 2003 and 2004.

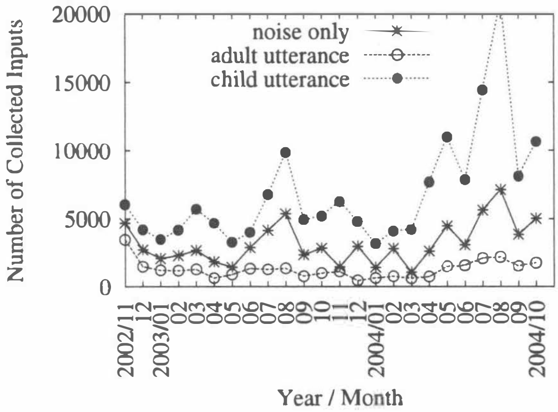

Fig. 3. Statistics of speech and noise inputs collected during the first two years which is transcribed completely.

How many inputs are from the same user or from the same group of users is shown in Figure 4. This statistic was determined automatically using only valid speech inputs and assuming a maximum inter-utterance pause of 20 seconds between every usage. The graph indicates that most of the users have more than one question.

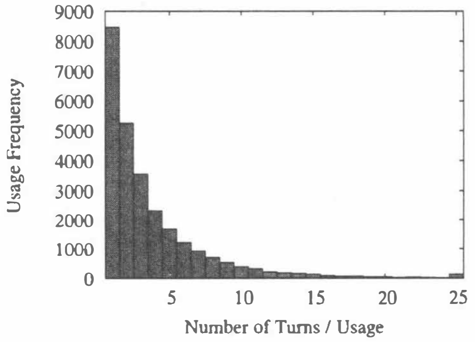

Fig. 4. Automatically obtained usage statistic. Frequency of utterance sequences with inter-utterance pauses less than 20 seconds and from the same speaker group.

Table 2 shows a classification of valid speech inputs into eight major categories. The statistic is given for the training data (the first two years) based on automatically determined responses and for the test data (one month) based on human-made responses. Most of the inputs are of general nature, greetings or related to the animated character. The possible reasons for this large share (about 80% for children and 65% for adults) are that the system is user-driven and it employs a computer graphics agent to give the user a virtual opponent to talk to. Apart from that, the user inputs are mainly information requests related to the community center, Ikoma city (the place

Table 2. Classification of collected speech inputs by response sentences into sub-domains, which were obtained from an automatic evaluation result of the training, and from a human evaluation of the test data.

Utterance Classification	2 years / Auto. Adult	 Child	Test / Human Adult	 Child
General, Greeting	41.6%	48.0%	38.6%	43.5%
Takemaru-related	24.0%	32.2%	26.2%	34.2%
Web-Search, HP	2.4%	2.1%	1.5%	0.7%
Weather, News	7.8%	3.5%	9.0%	4.4%
Time, Date	2.6%	3.6%	5.1%	5.5%
Ikoma City	7.4%	3.0%	5.8%	2.9%
Community Center	11.6%	7.1%	12.4%	8.3%
Other	2.5%	0.5%	1.5%	0.5%

and area were the system is installed), weather, news, time, date and webpage access.

The diversity of user inputs can be measured by their entropy. Figure 5 shows the overall word entropy when pooling the words of all transcribed speech inputs from the beginning of data collection up to certain points in time. Although entropy is increasing steadily, the relative increase in the last months of the first two years is low. Furthermore, while the relative share of words occurring the first time is between one and two percent during the first months, the share drops below a half percent after two years. This indicates, that the relative scope of the open-domain guidance system installed in a public place is almost determined after about two years in operation.

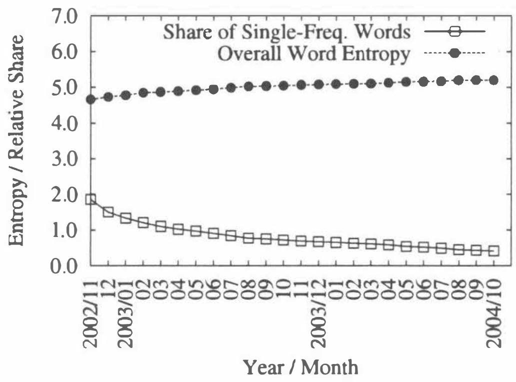

Fig. 5. Word entropy of transcribed sentences. Relative share of words occurring the first time.

4. OVERALL EVALUATION

The speech-oriented guidance system is evaluated at two stages: The baseline system which was built using transcribed data from almost five months collected until March 2003, and the updated system employing transcribed data from the first two years.

4.1. System Configuration

Table 3 shows the contents of the Q&A database. The questions for the baseline system are taken from the utterance transcriptions of the first five months. The corresponding system responses are devised by humans. The updated database is obtained by adding inputs of the same surface form with an occurrence frequency ≥ 2 among all data, and adding missing response sentences. While there is a relatively large increase in the number of example questions, the number of necessary response types increased only by one fifth from 300 to 358 types.

Table 3. Contents of the question & answer database for the baseline and the updated system.

Q & A Database	Questions Adult	 Child	Responses Adult	 Child
Baseline	3,307	6,573	302	298
Updated	3,948	10,273	358	358

Table 4. Takemaru speech data employed for building the baseline and updated acoustic models.

Training Data	Adult Model	Child Model	Training Method
Baseline	8,106	17,392	MAP Adaptation
Updated	23,417	120,671	F-B Algorithm

The number of training data for building the acoustic model is given in Table 4. An initial acoustic model is built using a large-scale read speech database of Japanese Newspaper Article Sentences [11]. The baseline models are obtained via MAP adaptation employing the first five months of collected data, the updated models by retraining the initial acoustic model with using the forward-backward algorithm. The acoustic model is a context-dependent phonetic-tied mixture model [12] for real-time decoding.

Table 5 shows the training data for building the bigram and trigram language model employed by Julius during first and second pass decoding. Since only relatively few utterance transcriptions were available when constructing the baseline model, a large amount of text data from webpages of Ikoma city and hypothesized user questions were employed additionally. After building a separate model for each text data source, the models are merged with weighting coefficients 0.4 (web), 0.4 (questions) and 0.2 (transcriptions) and performing complementary back-off as implemented in the `mergelm` tool mentioned in [13]. The updated model is built from the utterance transcriptions of two years employing Kneser-Ney smoothing with the SRILM toolkit [14]. Since the number of training data is rather few for language model training, an initial language model using all data is constructed first. The final adult and child model is obtained by merging the adult and child-specific model with the all data model, respectively.

Table 5. Number of training sentences employed for training the baseline and updated language model.

Language Model Training Data	Baseline Adult	 Child	Updated Adult	 Child
Web-Text Data	1,080k	1,080k	-	-
Hyp. Questions	6k	6k	-	-
Transcriptions	8k	17k	21k	115k
Vocabulary Size	$\approx$ 42k		$\approx$ 11k	

Table 6. Evaluation: baseline system, independent update of the acoustic model (AM), language model (LM), AM and LM (A+L), question and answer database (DB), and update of all system components (Updated).

Results	Base	AM	LM	A+L	DB	Updated
Adult Inputs						
Word Acc.	80.7	83.9	81.8	84.2	-	**84.2**
Word Cor.	84.8	87.5	87.9	89.8	-	**89.8**
Resp. Acc.	71.5	73.3	71.3	72.4	74.9	**75.5 (83.1)**
Child Inputs						
Word Acc.	63.2	67.3	62.5	66.9	-	**66.9**
Word Cor.	69.0	71.4	72.3	74.8	-	**74.8**
Resp. Acc.	53.5	55.0	54.0	55.6	57.1	**61.1 (67.8)**

4.2. Experimental Result

The system is evaluated with 1,052 adult and 6,516 children utterances collected August 2003. About one third of the test sentences do not appear in identical form in the Q&A database. The result is shown in Table 6. The overall performance is lowest for the baseline and highest when updating all system components. When independently updating the baseline system's components, the relative importance of developing each system component can be assessed. The largest improvement is obtained by updating the question and answer database. The update of the acoustic model is important in the second place. Least important seems to be an update of the language model. Although ASR performance increases, the system's response accuracy did not improve. The synergetic effect when updating all system components is most remarkable. The system's response accuracy (RA), i.e. the percentage of correct responses, reaches 76% for adults and 61% for children. Here, the RA is calculated automatically based on a one-to-one correspondence of user inputs to system responses. The true RA is actually even higher as determined by a human evaluation and is given in parenthesis in Table 6.

5. CONCLUSION

Achievements and insights from development and long-term operation of a speech-oriented guidance system were reported. A real-environment speech database with 300 hours real data has been collected. An analysis of the transcribed data shows, that the system is suitable to collect large amounts of spontaneous children speech. The diversity of collected user inputs measured by word entropy does not increase much after two years. Most of the system's users have more than one question, but the share of utterances concerned with greeting and general interaction is large. An evaluation was carried out for comparison of the baseline system in an early stage against systems with independent updates of each system component. Most important seem to be efforts regarding the Q&A database, followed by the acoustic model and the language model.

6. FUTURE WORK

The speech database was already employed for research on unsupervised selective training of acoustic models [15]. It remains to investigate unsupervised language model construction and possibilities for automatic derivation of a satisfying user response from the world wide web instead of costly efforts by human designers.

7. REFERENCES

[1] S. Seneff and J. Polifroni, "Dialogue Management in the Mercury Flight Reservation System," in *Proceedings of NASLP-NAACL Satellite Workshop*, 2000, pp. 1–6.

[2] L. Lamel, S. Rosset, J.L. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Routs, "The LIMSI ARISE system," *Speech Communication*, vol. 4, no. 31, pp. 339–353, 2000.

[3] A. Raux, D. Bohus, B. Langner, A.W. Black, and M. Eskenazi, "Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! experience," in *Proceedings of the International Conference on Spoken Language Processing*, 2006, pp. 65–68.

[4] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human-machine interaction for learning dialogue strategies," *IEEE Trans. on speech and audio processing*, vol. 8, no. 1, pp. 11–23, 2000.

[5] J. Schatzmann, M.N. Stuttle, K. Weilhammer, and S. Young, "Effects of the User Model on Simulation-based Learning of Dialogue Strategies," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 220–225.

[6] A.L. Gorin, G. Riccardi, and J.H. Wright, "How may i help you?," *Speech Communication*, vol. 23, no. 1/2, pp. 113–127, 1997.

[7] S. Ishikawa, T. Ikeda, K. Miki, F. Adachi, R. Isotani, K. Iso, and A. Okumura, "Speech-activated Text Retrieval System for Multimodal Cellular Phones," in *Proc. of ICASSP*, 2004, pp. 453–456.

[8] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano, "Public Speech-oriented Guidance System with Adult and Child Discrimination Capability," in *Proc. of ICASSP*, 2004, pp. 433–436.

[9] "Julius, an Open-Source Large Vocabulary CSR Engine - `http://julius.sourceforge.jp/`,".

[10] R. Nisimura, A. Lee, M. Yamada, and K. Shikano, "Operating a public spoken guidance system in real environment," in *European Conference on Speech Communication and Technology*, 2005, pp. 845–848.

[11] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research," *The Journal of the Acoustical Society of Japan*, vol. 20, pp. 199–206, 1999.

[12] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A New Phonetic Tied-Mixture Model for Efficient Decoding," in *International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. 1269–1272.

[13] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Itou, and K. Shikano, "Continuous Speech Recognition Consortium - an Open Repository for CSR Tools and Models," in *Proc. of LREC*, 2002, pp. 1438–1441.

[14] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. of ICSLP*, 2002, pp. 901–904.

[15] T. Cincarek, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic modeling for spoken dialogue systems based on unsupervised selective training," in *Proc. of ICSLP*, 2006, pp. 1722–1725.