DEVELOPMENT AND PORTABILITY OF ASR AND Q&A MODULES FOR REAL-ENVIRONMENT SPEECH-ORIENTED GUIDANCE SYSTEMS

Tobias Cincarek, Hiromichi Kawanami, Hiroshi Saruwatari and Kiyohiro Shikano

Graduate School of Information Science
Nara Institute of Science and Technology, Japan
cincar-t@is.naist.jp

ABSTRACT

In this paper, we investigate development and portability of ASR and Q&A modules of speech-oriented guidance systems for two different real environments. An initial prototype system has been constructed for a local community center using two years of human-labeled data collected by the system. Collection of real user data is required because ASR task and Q&A domain of a guidance system are defined by the target environment and potential users. However, since human preparation of data is always costly, most often only a relatively small amount real data will be available for system adaptation in practice. Therefore, the portability of the initial prototype system is investigated for a different environment, a local subway station. The purpose is to identify reusable system parts. The ASR module is found to be highly portable across the two environments. However, the portability of the Q&A module was only medium. From an objective analysis it became clear that this is mainly due to the environment-dependent domain differences between the two systems. This implicates that it will always be important to take the behavior of actual users under real conditions into account to build a system with high user satisfaction.

Index Terms— Guidance System, Real-Environment, ASR, Q&A, Portability, Domain Analysis

1. INTRODUCTION

Many researchers have been trying to build spoken dialogue systems by implementing a truly intuitive human-machine interface, integrating speech recognition and language understanding technology. However, only very few systems have been accepted widely 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 user-driven, open-domain, access-oriented systems. Examples for goal-oriented systems are flight reservation [1], train reservation [2] or bus information [3]. Their drawback is that the system's scope is most often only defined by the developer ideas ignoring needs and behavior of potential users under realistic conditions.

Access-oriented dialogue systems, e.g. for call routing [4], speech-activated text retrieval [5] or speech-oriented guidance [6] suffer less from this problem since they are mainly user-driven. The user can formulate his request freely in natural language and can immediately obtain a response from the system after the first query. The domain of such systems is open from the beginning and is in the end determined by potential users and the target environment. Therefore, system development only based on engineers' ideas is unlikely to succeed. Consequently, it is imperative to collect real speech data in the target environment for development.

Reports on development and portability of real-environment open-domain speech dialogue system are rare. Therefore, the purpose of this paper is to investigate the amount of real data required for initial system development. Furthermore, for practical purposes it is important to know how far components of an existing system can be reused and/or adapted to build a new system for a new target environment.

The speech-oriented guidance system *Takemaru* [6] installed since November 2002 at a public facility is employed as development platform for the initial prototype. For investigating portability, the adaptation of the *Takemaru* system to a different environment, a local subway station, is considered. The outcome are the *Kita* systems [7] installed since March 2006.

This paper is organized as follows. The architecture and purpose of each guidance system is described in Section 2. The contents of the real-environment speech database collected with both systems are described in Section 3. Development of the initial prototype system *Takemaru* is discussed in Section 4. Portability of the *Takemaru* system for the subway station environment of the *Kita* systems is investigated in Section 5. Since the portability of the ASR module is found to be higher than the Q&A module, a detailed comparison of Q&A domains between the *Takemaru* and the *Kita* systems is conducted in Section 6. A conclusion is given in Section 7.

2. SYSTEM OVERVIEW

The purpose of a speech-oriented guidance system is to offer a certain group of users convenient access to proper information in a certain environment. While the information society is at the verge to an ubiquitous society, there is growing demand for this kind of services in any place. Although entering search queries via keyboard is still the prevailing method for accessing information, formulating one's question freely in natural language and using speech is a far more natural way to human-machine communication

Figure 1 shows a block diagram of the main components of the speech-oriented guidance system. User input is recorded via a directivity microphone. After voice activity detection and rejection of non-verbal inputs, speech input is recognized in parallel using the open-source LVCSR engine Julius [8] with an adult and a child acoustic (AM) and language model (LM), respectively.

After age group classification using the acoustic likelihood, response generation is carried out. There is one question and answer

This work is supported by the MEXT e-Society project.

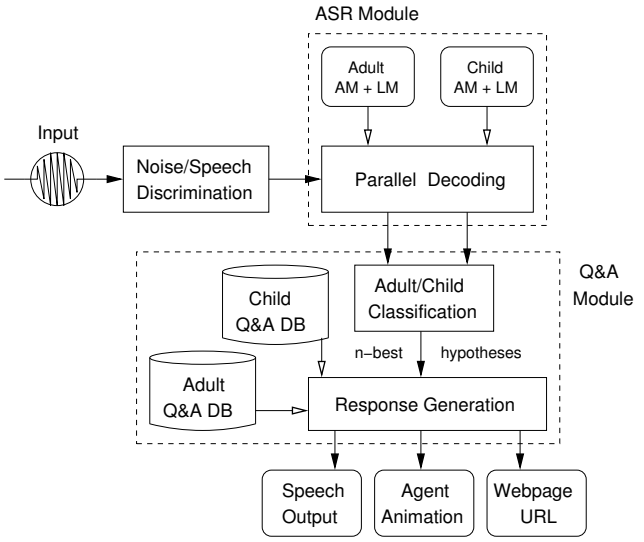

Fig. 1. The main building blocks of a speech-oriented guidance system.

database (QADB) per age group. Each QADB contains a large number of question and answer pairs to cope with the wide variety of user questions. The response sentence corresponding to the example question most similar to the recognition result is selected based on the n-best recognition hypotheses.

Besides voice-based response message output, each system uses an extra screen to display a computer graphics agent (or moving eyes in case of *Kita-robo*) and to display web pages. The presence of the agent gives the human user a virtual opponent to talk to in order to realize a more lively and natural human-machine interaction. The purpose of displaying web pages from the Internet is to give the user complementary information to the voice-based response.

More details about system architecture, adult/child discrimination, rejection of non-speech input (accuracy $\geq$ 85%) and preliminary results for recognition accuracy have been reported in [6, 9].

2.1. Takemaru

Takemaru is installed inside the entrance hall of the North community center in Ikoma city, Nara Prefecture, Japan since November 2002 (cf. Figure 2). The indoor environment is relatively calm with a background noise level of approx. 50 dB(A). The place is frequently visited by adults and children because it is a public facility with a library, a branch office for residental services and there are weekly events. The system uses the mascot character of Ikoma city, *Takemaru*, as agent. The *Takemaru* system can handle questions related to the agent, general information such as time, date, weather and news, the facility itself, surrounding area and sightseeing.

2.2. Kita-chan and Kita-robo

The *Kita* systems are installed near the passenger gate of a subway station since March 2006. There is *Kita-chan*, a terminal-based system similar to Takemaru, and *Kita-robo*, a robot with moving eyes (cf. Figure 3). The agent's character and the robot's appearance are an imitation of the mascot of the subway station itself. No difference between the *Kita* systems will be made for portability investigations

Fig. 2. The speech-oriented guidance system *Takemaru*.

Fig. 3. The speech-oriented guidance systems *Kita-robo* and *Kita-chan* (from left to right).

in this paper. Although there is a roof above both systems, the environment is partly open-air. This is the main reason for a background level of approx. 60 dB(A), about 10 dB(A) higher than for *Takemaru*. Fortunately, this is less problematic because a directivity microphone is employed for sensing speech input. The contents of the *Kita* systems are an extension of the Takemaru system. They can also handle train information queries and display maps of certain areas or show the location of places of interest around the station, e.g. restaurants, shops, post offices, etc.

3. SPEECH DATABASE

The total number of speech and noise inputs collected by end of June 2007 with the *Takemaru* and *Kita* systems is shown in Table 1. Takemaru has been collecting data for almost five years. The first two years are completely transcribed, labeled with tags (e.g. noisy, incomplete, invalid) and classified into five speaker groups (preschool children, lower grade school children, higher grade school children, adults, elderly and noise). Furthermore, utterances forming valid

Table 1. Speech data collected with the *Takemaru* and the *Kita* systems by end of June 2007

	Takemaru		Kita	
Classification	# Inputs	Time	# Inputs	Time
Transcribed	273,698	121.2 h	62,463	30.1 h
Preschool Children	27,535	14.3 h	7,674	4.0 h
Lower Grade	106,797	57.7 h	19,038	10.2 h
Higher Grade	31,402	15.8 h	7,900	3.7 h
Adults, Elderly	31,100	14.1 h	19,428	8.7 h
Noise, Non-Verbals	76,864	19.3 h	8,423	3.5 h
Untranscribed	553,930	265.4 h	147,517	67.5 h
Total	827,628	386.6 h	209,980	97.6 h

queries to the system, have been labeled with one or more possible system responses. The *Kita* systems have been collecting data for more than one year. Inputs from seven months have been transcribed and labeled by humans. If all systems are taken together more than one million inputs and more than 480 hours of real-environment data have been collected since their operation began.

4. INITIAL SYSTEM DEVELOPMENT

In this Section, we report about the development of the initial *Takemaru* prototype system. After describing the data to construct the acoustic model (AM), language model (LM) and question and answer database (QADB), the influence of increasing amounts of training data on system performance is analyzed.

4.1. Speech Data

The data employed for actual development is shown in Table 2. Only valid user inputs which have been transcribed and labeled by humans with a correct system response were employed for system development. Invalid inputs, i.e. meaningless, unintelligible, too noisy utterances, etc. were excluded since they would not bring any benefit for constructing AM, LM and QADB. The data collected during November 2002 and August 2003 were put aside as validation and evaluation data, respectively.

4.2. ASR Module

Experimental conditions for AM training, LM training and speech recognition are given in Table 3. The LMs for each training period is constructed by linear interpolation of the adult-dependent or child-dependent LM with the all data LM. A weight for linear LM interpolation is determined automatically so that the perplexity of the validation data set is minimized.

Since a user expects an immediate response from a dialogue system, speech recognition may not cause a delay until response generation. Consequently, a context-dependent, state-clustered, phonetic-tied mixture [12] acoustic model with relatively few parameters (8,192 Gaussians) is employed for real-time speech recognition. The Japanese Newspaper Article Sentences (JNAS) database [13] was employed to build the initial AM. This initial model is retrained with *Takemaru* speech data using either Baum-Welch training, or MLLR-MAP [14, 15] adaptation depending on the amount of available training data.

Table 2. Data for developing the *Takemaru* ASR module

Takemaru	Collection	Adult		Child	
Data Sets	Period	# Utter	Time	# Utter	Time
Training	22 months	16,332	8.2 h	75,315	41.4 h
Validation	1 months	3,069	1.5 h	4,115	2.3 h
Evaluation	1 months	1,085	0.5 h	6,568	3.7 h

Table 3. Experimental Conditions

AM Training	HTK 3.2 [10]
LM Training	SRILM 1.5.0 [11]
Acoustic Model	PTM [12], 2,000 states
Acoustic Features	12 MFCC, 12 Δ MFCC, Δ E
AM Training	Baum-Welch, 3 Iterations
AM Adaptation	MLLR-MAP, 256 Classes, 3 Iterations
Language Model	3-gram, Kneser-Ney Smoothing
ASR Engine	Julius 3.5 [8]

Table 4. Number of distinct example questions and system responses in the question and answer database (QADB)

Takemaru	# Questions		# Responses	
QADB (Set)	Adult	Child	Adult	Child
(All Data)	6,671	32,992	275	285
(Training)	4,052	17,891	265	282

4.3. Q&A Module

Transcribed user utterances (= questions) are labeled by humans with one correct system response (= answer). During the first months of operating *Takemaru* new responses were added continuously if necessary to improve user satisfaction. The number of distinct Q&A pairs in the QADB for building the Q&A module is shown in Table 4. Pairs for utterances with a transcription appearing only once and which are linguistically unintelligible or out-of-domain are excluded from the QADB because they had a negative effect on response accuracy.

4.4. Performance Evaluation

System performance is evaluated in case of short-term (one month, 4k data), medium-term (six months, 23k data) and long-term (22 months, 91k data) development. Word and response accuracy are given in Table 5. It is clear, that the performance improvement from short-term to medium-term development is quite large but relatively small from medium-term to long-term development. Children performance saturates earlier than adult performance because the num-

Table 5. ASR and Q&A performance of *Takemaru* measured by word accuracy (WA) and response accuracy (RA)

Training	# Utterances		Adult [%]		Child [%]	
Period	Ad	Ch	WA	RA	WA	RA
1 mon	1k	3k	68.9	52.9	52.1	43.0
6 mon	6k	17k	77.4	67.8	60.8	53.7
22 mon	16k	75k	79.5	72.1	62.0	55.8

Table 6. Vocabulary Size [words], OOV rate [%] and test set perplexity (PP) of *Takemaru* LM

Training Period	Adult			Child		
	Vocab	OOV	PP	Vocab	OOV	PP
1 mon	0.6k	8.2	20.7	1.2k	10.3	35.9
6 mon	1.6k	3.0	12.0	3.6k	4.1	21.6
22 mon	3.1k	1.6	9.9	8.5k	1.7	16.5

ber of available adult data is only small. Consequently, long-term development may be required for developing the initial prototype system in practice.

A similar tendency can be observed for vocabulary size and LM quality as given in Table 6. When comparing medium-term and long-term development, the vocabulary size more than doubles, the OOV rate is reduced by more than 50% and there is a significant reduction in perplexity.

5. CROSS-ENVIRONMENT PORTABILITY

In this section the portability of *Takemaru* for the *Kita* environment is investigated. A system or a module of a system can be considered as portable if system performance is high without adaptation, it can be improved remarkably even with moderate amounts of adaptation data and performance improvement shows signs of stagnation. Consequently, portability comprises both reusability and adaptability which means that cost-effective adaptation of a system to a new environment is possible in practice.

5.1. Speech Data for System Update

Speech utterances forming valid queries collected during one (short-term) to six months (medium-term) of system operation are employed for updating AM, LM and QADB. 14 days of user inputs collected during the first half of May 2006 are employed for performance evaluation (cf. Table 7).

For AM adaptation to the acoustic environment and due to the comparably low amount of adaptation data, MLLR-MAP is employed. Baum-Welch training using all available *Takemaru* and *Kita* training data could not outperform MLLR-MAP adaptation. The LM is reconstructed completely using the transcriptions of all valid *Takemaru* utterances with the transcriptions of the *Kita* training utterances added.

The contents of the initial QADB for the *Kita* systems are mostly human-labeled Q&A pairs collected during the first months of operating *Takemaru*. They have partially been edited by humans for the *Kita* systems. This database is updated with Q&A pairs obtained during six months of operating the *Kita* systems. The number of distinct Q&A pairs in the QADB after adding the newly collected pairs almost triples. 75 new response sentences have also been added for user questions with no appropriate counterpart available in the existing response set (cf. Table 8).

Table 7. Training and evaluation data collected in *Kita* environment

Kita Data Sets	Collection Period	Adult		Child	
		# Utter	Time	# Utter	Time
Training	6 months	11,276	5.5 h	18,720	10.5 h
Evaluation	14 days	1,699	49 m	2,732	91 m

Table 8. Number of distinct example questions and system responses in the question and answer database

Kitachan QADB (Set)	# Questions		# Responses	
	Adult	Child	Adult	Child
(Initial)	2,761	5,062	183	179
(Update)	5,505	10,091	288	294
(Final)	7,018	13,022	315	320

Table 9. ASR and Q&A performance of the *Kita* systems measured by word accuracy (WA) and response accuracy (RA)

Training Period	# Utterances		Adult [%]		Child [%]	
	Ad	Ch	WA	RA	WA	RA
Takemaru	20k	86k	73.1	51.3	56.2	44.8
+ 1 mon	+ 5k	+ 7k	76.8	66.2	58.9	53.4
+ 6 mon	+ 11k	+ 19k	78.0	69.7	60.1	57.1

5.2. Performance Evaluation

The ASR and Q&A performance before and after system update is given in Table 9. With the short-term (one month, 12k data) update the absolute improvement in word accuracy is only moderate (3.7% and 2.7%). The medium-term (six months, 30k data) update yields only very small additional improvements (1.2%) over the short-term update. The absolute improvement for adults is higher than for children. This is likely to be due to the fact that more child data has been available for constructing the models of the *Takemaru* ASR module. The quality for initial and updated LMs is shown in Table 10. There is only a small increase in vocabulary size and small decrease in perplexity. These results show that the *Takemaru* ASR module has a high portability.

Considering simultaneous adaptation of ASR and Q&A modules, there are remarkable improvements in response accuracy after short-term (14.9% and 8.6%) and further significant improvements after medium-term (3.5% and 3.7%) update. Although performance when using the initial, human-edited QADB is low, the response accuracy rebounds with the short-term update and reaches a level comparable to the *Takemaru* system with the medium-term update. This indicates a medium portability of the *Takemaru* Q&A module in the *Kita* environment.

Table 10. Vocabulary size [words], OOV rate [%] and test set perplexity (PP) of the LM for the *Kita* systems

Training Period	Adult			Child		
	Voc.	OOV	PP	Voc.	OOV	PP
Takemaru	3.7k	2.9	21.1	10.3k	2.4	31.6
+ 1 mon	4.3k	1.8	19.1	10.7k	1.9	30.0
+ 6 mon	4.8k	1.5	17.5	11.2k	1.7	28.7

Table 11. Domain comparison using response statistics

Responses A ↔ B	COR	A∩B
Takemaru ↔ Takemaru (Subsets)	1.00	0.93
Takemaru ↔ Kita (All)	0.59	0.56

Table 12. Domain comparison using language models

Set	Take (A) # words	Kita (B) # words	A∪B # words	A∩B # words	KLD [bit]
All	11,192	4,768	12,430	3,530	8.57
Ad.	3,782	2,625	4,865	1,542	0.52
Ch.	10,344	3,696	11,172	2,868	10.13

6. DOMAIN COMPARISON

Although the portability of the Q&A module can be assessed based on evaluating the response accuracy after system update, an approach which gives more insight into the actual circumstances would be preferable. Therefore, the domain difference between the *Takemaru* and the *Kita* environment is assessed by directly comparing the contents of the QADB and domain-specific language models. The purpose is to identify a metric which can measure the domain difference and degree of portability objectively.

6.1. Approach

The domains of the *Takemaru* (A) and *Kita* systems (B) are compared based on the set of system responses and the language models trained on utterance transcriptions. For this analysis and in order to determine that part of the *Takemaru* QADB which is reusable for the *Kita* systems, a mapping of corresponding responses was established between the two systems.

Based on the definition of a random variable x which takes as values either system response identifiers or words, the inter-system domain distance can be measured in the following three ways:

- COR
  correlation between response frequencies $f_A(x)$ and $f_B(x)$ of each system
- P(A∩B)
  probability of the response intersection set of both systems, counting repeated occurrences of the same response x for different utterances and different users
- KLD
  symmetric Kullback-Leibler divergence between probabilities $P_A(x)$ and $P_B(x)$ of n-grams x of domain-specific language models.

Two domains are the more similar the higher the value of COR and P(A∩B), or the lower the value of KLD. For comparison using P(A∩B), resampling was carried out to obtain data sets of equal cardinality.

6.2. Results

The result of domain comparison is shown in Tables 11-12. The difference between two random subsets of *Takemaru* data is shown as reference. Since there is a very high correlation (1.00) and a high probability for the intersection set A∩B (0.93) the proposed metric can be considered as valid. When comparing *Takemaru* and *Kita* domain, a value of 0.56 for P(A∩B) indicates that at least half of the users' inputs to *Takemaru* have also been observed for the *Kita* systems and vice versa. The correlation has a similar value of 0.59. These values can be taken as a more concrete measure for the notion of 'medium' portability of the *Takemaru* Q&A database for the *Kita* domain as discussed in the previous section.

The number of words in the intersection and union of domain vocabularies as well as the KL distance between uni-gram language model probabilities is shown in Table 12. It is clear that both domains have many words in common. The KL distance is difficult to interpret. It would be of more practical interest if more than two domains are compared.

Finally, an insight into the actual data is given. The ten most probable system responses, i.e. responses x in the intersection set with maximum probability $P_A(x)P_B(x)$ are shown in Table 13. Among them are mostly greetings, agent-related information and weather forecast.

A ranklist of relatively frequent responses in e.g. domain A can be obtained by sorting the responses by the weighted probability ratio $P_A(x)\log[P_A(x)/P_B(x)]$. The list of relatively frequent *Takemaru* and *Kita* responses is given in Tables 14 and 15, respectively.

While users of *Takemaru* are often concerned about current time, bus timeable and local information, users of *Kita* are mainly interested in the local map, location of restaurants, post office, etc.

A similar list can be obtained for keywords in user utterances using uni-gram language model probabilities (cf. Table 16). It is interesting that among the keywords for the *Takemaru* domain are objects related to the environment (room, library, book), the agent's name and that it looks 'cute'. For the *Kita* domain there is also the agent's name, environment-related objects (station, vending machine), places near to the station (restaurant, SanMarc, NAIST, Kitayamato, Mayumi) and farer locations (Kyoto, Namba).

7. CONCLUSION

In this paper, development and portability of ASR and Q&A modules for two speech-oriented guidance systems was investigated. Since ASR task and Q&A domain of an open-domain dialogue system are defined by actual users of the system, real speech data collected in the target environment are required for development.

The development of a prototype system is simulated for short-term (one month, 4k data), medium-term (six months, 23k data) and long-term (two years, 91k data) data collection periods. From the evaluation it was clear that long-term development may be necessary in practice until performance saturates.

However, it is usually impractical to collect data over a long time span whenever a system for a new environment is to be built. At most short-term development will be possible in practice. Consequently, it is worth considering the reuse of modules of an existing system. Therefore, the performance improvements in case of short-term (one month, 12k data) and medium-term (six months, 30k data) adaptation of the prototype to the new target environment was evaluated to investigate the degree of portability of each module.

Experimental results showed that the *Takemaru* ASR module is highly portable. However, there is only medium portability for the Q&A module. Consequently, short-term adaptation would be sufficient for the ASR module. However, the Q&A module showed the tendency to require more development effort although it seems possible to avoid the costs for long-term development if models and data of a well-trained prototype system can be reused.

Table 13. Responses very common in *Takemaru* and *Kita* domain

Response Meaning	Response Sentence
(greeting)	"Hello."
(weather forecast)	"Tomorrow's weather will be ..."
(agent, self-intro)	"My name is ..."
(place, toilet)	"The toilet is .."
(greeting)	"See you again."
(agent, current age)	"I am ... year(s) old."
(agent, favorite food)	"My most favorite food is .."
(websearch)	"Please tell me the search keyword."
(greeting)	"You're welcome. Please come again."
(newspaper)	"I show you the newspaper page."

Table 14. Responses relatively frequent for *Takemaru* domain

Response Meaning	Response Sentence
(out-of-domain)	"I am sorry, but I do not know."
(current time)	"The time is ..."
(greeting)	"Hello."
(agent, self-intro)	"My name is ..."
(bus information)	"I show you the bus timetable."
(agent, appearance)	"Don't you think I am cute?"
(offer friendship)	"Please become my friend."
(misunderstanding)	"I do not understand you."
(local information)	"You are at the community center."
(offer information)	"Please ask me something about ..."

Table 15. Responses relatively frequent for *Kita* domain

Response Meaning	Response Sentence
(map, local)	"I show you the local map."
(map, restaurant)	"I show you the restaurant map."
(place, toilet)	"The toilet is ..."
(map, post office)	"The nearest post office is ..."
(agent, origin)	"My name is Kita because ..."
(weather forecast)	"Tomorrow's weather will be .."
(user warning)	"Please do not tease me."
(misunderstanding)	"Could you please say that again?"
(general response)	"How may I help you?"
(bus information)	"Please take the south exit."

Table 16. Keywords in user utterances with a high probability for both domains and which are relatively more frequent in one domain than the other domain

Common Domain	where? Takemaru Ikoma what? who? here like news station name today now you toilet stupid when? search weather forecast tomorrow born goodbye cute understand say Mayumi how_old?
Takemaru Domain	Takemaru begin search understand now what? birthday friend when? you stupid pool game live kindergarten bus cute home what? room sleep library front who? noisy sing book
Kita Domain	Kita today vending machine line where? news Kyoto like map restaurant Ikoma Kuragaritoge show weather SanMarc NAIST Kitayamato nearby go tell_me Namba Nara station Mayumi

8. 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] 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.

[5] 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.

[6] 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.

[7] H. Kawanami, M. Kida, N. Hayakawa, T. Cincarek, T. Kitamura, T. Kato, and K. Shikano, "Spoken Guidance Systems Kita-chan and Kita-chan robot. Their Development and Operation in a Railway Station," Tech. Rep., IEICE, SP2006-14, 2006.

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

[9] 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.

[10] "HTK Speech Recognition Toolkit `http://htk.eng.cam.ac.uk/`,".

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

[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] 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.

[14] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[15] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.