PUBLIC SPEECH-ORIENTED GUIDANCE SYSTEM WITH ADULT AND CHILD DISCRIMINATION CAPABILITY

Ryuichi NISIMURA, Akinobu LEE, Hiroshi SARUWATARI, Kiyohiro SHIKANO

Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN
E-mail: {ryuich-n, ri, sawatari, shikano }@is.aist-nara.ac.jp

ABSTRACT

Takemaru-kun system is a real world speech-oriented guidance system located at the Ikoma-city North Community Center. The system has been operated daily from November, 2002 to provide visitors a speech interface for information retrievals. This system also aims at the field test of a speech interface and collecting actual utterance data. By analyzing and evaluation of the collected utterances, necessities of flexible processing according to the user's age group are discovered. It becomes impossible to disregard the increase of child users when the system is installed in a public place. This paper proposes an automatic approach discriminating speakers between adult and child users, which is based on a statistical learning. This proposal realizes a flexible spoken dialogue to both adult and child users. As for parameter vectors in machine learning, acoustic and linguistic properties extracted from speech recognition logarithm likelihood scores are adopted to discriminate user's age group. Although GMM-based recognition uses only acoustic properties, this method can also consider linguistic properties. In the experiments with the SVM-based screening, we obtained 92.4% discrimination rate to the actual users' utterances. The advantage of using linguistic properties is also shown. This paper also describes an overview of the Takemaru-kun system and the data collection status via the field test. Performances of child speech recognition are evaluated using collected utterances.

1. INTRODUCTION

Speech-related interface systems are regarded as favorable human-machine interface. In order to use it for many users, the system must work even with diversified users' utterances. With few exceptions, speech interface systems have been developed only for adult users[1]. It is known that a speech recognition program developed using adult voices can not recognize child's voices correctly[2, 3].

As a research platform of speech interface, our practical speech-oriented guidance system "Takemaru-kun"[3] have been operated daily from November, 2002 at the entrance hall of Ikoma-city North Community Center (Figure 1). By analyzing and evaluation of the collected utterances via the long-term field test, necessities of flexible processing according to the user's age group are discovered. It becomes impossible to disregard the increase of child users when the speech interface system is installed in a public place. The system needs providing child recognition models specialized in child voices. A response generation routine suitable for children is also planned in our system.

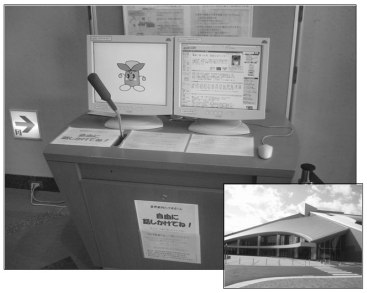

Fig. 1. Speech-oriented guidance system "Takemaru-kun".

For these realization, discriminating a user's age group from the voice is necessary. We propose an automatic approach discriminating users between adults and children, which is based on a statistical learning. It realizes a flexible spoken dialogue to both adult and child users. As for parameter vectors in machine learning, we use acoustic and linguistic properties which are extracted from speech recognition logarithm likelihood scores. A linguistic property of child utterances is different from adult one. Although conventional GMM-based (Gaussian Mixture Model) recognition[4] uses only acoustic properties, this proposed method can also consider linguistic properties on the basis of speech recognition scores.

In this paper, we describe an overview of the Takemaru-kun system and the data collection status via the field test in Section 2. The current performance of child speech recognition is examined in Section 3. In Section 4, the detail procedures of proposed method are explained. Evaluation is carried out using the collected actual users' utterances in Section 5. Finally, we conclude this paper in Section 6.

2. TAKEMARU-KUN SYSTEM AND ITS FIELD TEST

The Takemaru-kun system can answer questions on guidance topics such as facilities, services, town information and so on. When a user speaks to a microphone on the desk, a synthesized voice response is outputted. The animation agent "Takemaru" is displayed on the left-hand monitor (Figure 2). The visual information using Web pages are also displayed on the right-hand monitor.

The speech interface program of the Takemaru-kun system has simple one-question and one-response principle[3]. This interface program consists of our speech recognition engine Julius[5], which can recognize wide task domain utterances with over 20k vocabulary trigram language model.

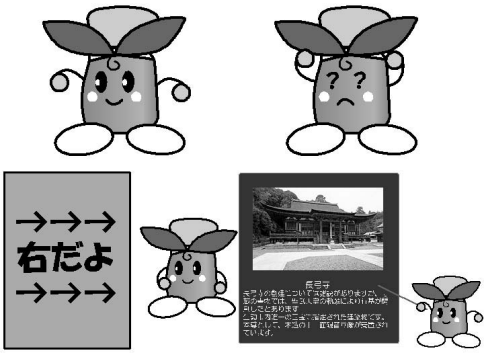

Fig. 2. Examples of Takemaru animation agent.

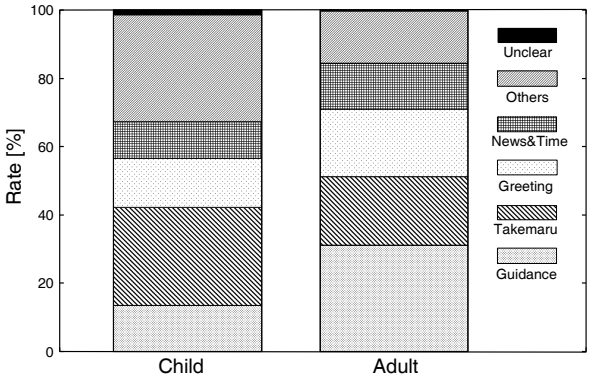

Fig. 3. Topic population for different age groups.

Table 1. Age and gender classification of collected data.

		Gender			
	Age Group	Male	Female	Uncertain	Total
a)	Infant	76	1421	585	2082
b)	Lower Grade Child	1920	7961	2843	12724
c)	Higher Grade Child	934	1154	498	2586
d)	Adult	5520	2496	70	8086
e)	Elderly Person	8	12	0	20
Total		8458	13044	3996	25498

The Takemaru-kun system also aims at a field test in a practical environment and collecting actual utterance data in the framework of human-machine interaction. It has been operated during every business day from the opening of the center (November 6, 2002). 46,754 speech inputs have been recorded during the field test for 125 days till March 31, 2003. The collected data, which are classified and transcribed manually, contain gender-unknown and invalid inputs such as noise, level overflowed shouts, and other unclear inputs. The classification result of the data except those unclear utterances is shown in Table 1, where 68.2% of collected data is uttered by children.

The topics of the utterances are also classified. Figure 3 shows a comparison between adults and children, for the percentages of the labels in the database. In this figure, the label "Guidance" is given to the utterances that include queries about rooms, facilities, timetables and other questions about the center and its surroundings. The label "Takemaru" indicates utterances asking about personal matters for the agent Takemaru. "Greeting" label is given to common greetings. "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. The "Unclear" label is assigned to other types of meaningless speech, including unclear utterances, shouts and so on. As this figure is shown, most adults and children have different interests towards the Takemaru-kun system. That is, the linguistic properties on the utterance differs between adults and children. This result suggests an effective use of linguistic properties is important when classifying actual user utterances.

In our previous evaluations[3] using adult recognition models, the system could answer appropriate responses to 73.4% query utterances by adult users. For child utterances, however, only 37.4% of generated responses was appropriate.

3. EVALUATIONS OF CHILD SPEECH RECOGNITION

In this section, the performance of child speech recognition is investigated, and compared with recognition accuracy of adult voice.

For constructing back-off word trigram language models, we collected the following training texts.

Web) Texts extracted from Ikoma-city related web pages, including 1,080,272 sentences, 31,265,487 words and 218,723 different words.

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

The collection data shown in Table 1 are also used in order to prepare the models suitable for adults and children,

Adult) Texts from d) and e) groups in Table 1, including 7,606 sentences, 39,913 words and 1,747 different words.

Child) Texts from a), b) and c) groups, including 16,892 sentences, 88,614 words and 3,355 different words.

For adults, a trigram language model built from **Adult** texts is merged into the base model from **Web** and **QA** texts[3]. As for a child model, the **Web**, **QA** and **Child** merged model are also created. Furthermore, the Takemaru-kun task network grammar (441 different words) is applied to each models for acquiring higher accuracy recognition[3]. The vocabulary size is set to 40k words.

PTM (Phonetic Tied Mixture)[6] triphone HMM (Hidden Markov Model) models are trained from the following speech data as acoustic models.

JNAS-Male) Reading style speech by male speakers, extracted from the JNAS (Japanese Newspaper Article Sentences)[7] database, 20,063 utterances.

JNAS-Female) Reading style speech by female speakers from the JNAS, 20,023 utterances.

Takemaru-Adult) Natural utterances by adults, extracted from d) and e) groups in Table 1, 7,515 utterances.

Takemaru-Child) Natural utterances by children, from a), b) and c) groups, 17,294 utterances.

We used the **JNAS-Male**, **JNAS-Female** and **Takemaru-Adult** utterances data for training an adult PTM acoustic model. The MAP (Maximum *A Posteriori*)[8] or MLLR (Maximum Likelihood Linear Regression)[9] adaptation to the created adult acoustic model are performed using the **Takemaru-Adult** utterances. The

Table 2. Word accuracy. (%)

Model Type	Acoustic Model	Test Sets	
		Adult	Child
Adult	No Adaptation	90.7	(53.1)
	MAP	92.2	(57.3)
	MLLR	93.1	(56.1)
Child	No Adaptation	(61.7)	60.8
	MAP	(64.3)	70.9
	MLLR	(64.1)	70.7

JNAS-Female and **Takemaru-Child** speech data are used to train a child acoustic model. The MAP or MLLR adapted child models are also prepared using the **Takemaru-Child** utterances.

As for the test sets, 500 adult and 500 child utterances are extracted from collected data shown in Table 1. The test sets are excluded from the model training data.

Table 2 shows experimental results in word accuracy. In this table, results with the original age-group acoustic model without MAP and MLLR adaptation are shown in "No adaptation". "MAP" and "MLLR" indicate results when using each adapted models. It is found that the child recognition accuracy using adult models is worse than that of adult voices. The general speech recognition system based on adult voices can not recognize child voices correctly. However, remarkable improvements are obtained by using the child suitable models. The MAP adapted child model brings 13.6% improvement in accuracy compared with that of the adult model. Thus, it is confirmed that offering of the recognition models suitable for user's age group is essential.

4. CHILD AND ADULT DISCRIMINATION ON THE BASIS OF SPEECH RECOGNITION SCORES

The system has two parallel speech recognition decoders. Each has an age-group-dependent acoustic model and a language model suitable for adult or child users. An output is chosen on the basis of speech recognition logarithm likelihood scores from each recognized results. The properties to be taken are average acoustic score (AP) and language score (LP), which are given by

$$AP = \frac{Acoustic\ \ Model\ \ Logarithm\ \ Likelihood\ \ Score}{Number\ \ of\ \ Input\ \ Speech\ \ Frames} \tag{1}$$

and

$$LP = \frac{Language\ \ Model\ \ Logarithm\ \ Likelihood\ \ Score}{Number\ \ of\ \ Output\ \ Words}. \tag{2}$$

Then, the scores obtained from adult recognition models in Section 3 are treated as parameters, AP_{adult} and LP_{adult}. On the other hand, the scores from the child models are indicated to be AP_{child} and LP_{child}. For robustness to the change of recognition models and speech record conditions, difference scores between each age group models, $AP_{adult} - AP_{child}$ and $LP_{adult} - LP_{child}$, are also introduced.

Figure 4 and 5 illustrate distributions of $AP_{adult} - AP_{child}$ when inputting the whole collected utterances in Table 1 into the parallel speech recognition decoder. In Figure 4, the distribution of adult male data differs from child utterances greatly by using no adaptation models as the acoustic model. However, since the adult female utterances have overlapped with the child, we can not dissociate them. Because of features from the JNAS female training data included in child acoustic model[1], adult female utterances have high probabilities for the child model. In view of this property, reductions of female adult features in the child model by MAP or MLLR adaptation are effective. Figure 5 shows that the MAP adaptation separates the adult female distribution from the child. Distributions of $LP_{adult} - LP_{child}$ are also illustrated in Figure 6. Slight differences in distribution tendencies between each groups are observed.

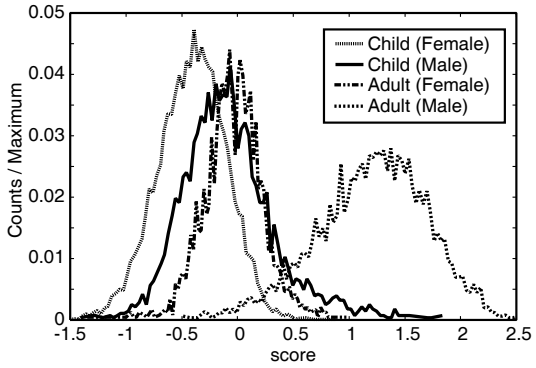

Fig. 4. Distributions of $AP_{adult} - AP_{child}$ without acoustic model adaptation.

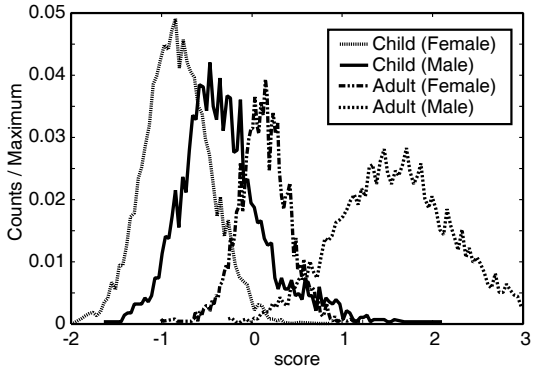

Fig. 5. Distributions of $AP_{adult} - AP_{child}$ using MAP adapted acoustic models.

To classify utterances into adult and child voices, we used the SVM (Support Vector Machines)[10], which are two-value classification algorithms based on a statistical learning, and have been applied to the natural language processing, such as text categorization and so on[11, 12]. The parameter vectors which consists of above-mentioned AP and LP are given to the SVM-based screening. Gaussian basis function is adopted as the kernel function in the SVM algorithm.

5. EXPERIMENTS

We evaluated the proposed method using collected actual users' utterances. For training data of the SVM, 7536 adult and 7536 child utterances are extracted from collected data.

[1]PTM triphone model training needs phonetic balanced utterances. We have to include the JNAS female data for training a child acoustic model because collected utterances are not well balanced phonetically.

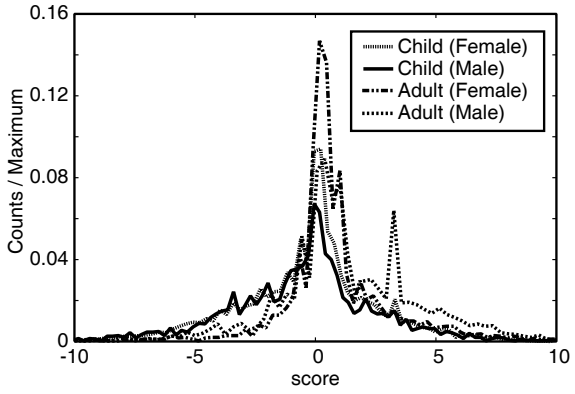

Fig. 6. Distributions of $LP_{adult} - LP_{child}$.

Table 3. Discrimination rate. (%)

Parameter Vector Type	Acoustic Model		
	No Adapt.	MAP	MLLR
- Acoustic 1. $AP_{adult} - AP_{child}$	85.4	91.6	91.4
- Acoustic and Linguistic 1. $AP_{adult} - AP_{child}$ 2. $LP_{adult} - LP_{child}$	87.4	92.4	92.0
- Acoustic and Linguistic with real decoder scores 1. AP_{adult} 2. AP_{child} 3. LP_{adult} 4. LP_{child}	86.9	92.0	91.8
GMM based recognition (baseline)			86.4

Table 3 shows experimental results of evaluations on 500 adult and 500 child test sets. The proposed method with MAP adapted acoustic models acquires 92.4% discrimination rate. Combinations of acoustic and linguistic properties conduce to certain improvements in accuracy compared with that of using only acoustic property. We confirmed improvements reflected by MAP and MLLR adaptations. The accuracy when inputting real decoder scores of AP and LP as parameter vectors are almost same as that of difference scores.

For comparison with the conventional speaker identification method, we evaluated GMM likelihood based recognition. Five-class GMMs for a) to e) age groups (Table 1) are built from the whole collected utterances, and each has 64 Gaussian mixtures. The utterances classified into d) or e) by GMM likelihood comparison are judged to be adult utterances, and others are judged to be child utterances. As an experimental result, 86.4% discrimination rate is observed, and indicates 6.0% fall from results in Table 3. These results confirm the advantage of the proposed method.

6. CONCLUSIONS

We described the overview of our speech interface research platform Takemaru-kun system and the status of its field test. By analyzing and evaluation of the collected utterances, necessities of flexible processing according to the user's age group are confirmed.

The automatic approach discriminating speakers between adults and children is proposed. We developed it on the basis of acoustic and linguistic properties obtained from speech recognized results. It can realize a flexible spoken dialogues to both adult and child users in public speech interface systems. The advantage of this method is that the improvement of accuracy is obtained by including linguistic properties in its algorithms. In the experiments using the SVM-based screening, the proposed method achieved 92.4% discrimination rate to the actual users' utterances. It means 6.0% improvement compared with the conventional GMM likelihood based speaker recognition.

As future works, other parameters and machine learning algorithms to determine users' age group will be considered. We intend to implement the flexible spoken dialogue according to users' age group to the Takemaru-kun system. Further evaluations of the system via the field test are planned.

7. ACKNOWLEDGMENT

The authors greatly appreciate the support provided by the Ikomacity office and the Ikoma North Community Center.

8. REFERENCES

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

[2] K. Shobaki et al., "The OGI Kid's Speech Corpus and Recognizers," Proc. *ICSLP2000*, vol.4, pp.258–261, October 2000

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

[4] D.A. Reynolds et al., "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol.3, no.1, pp.72–83, January 1995.

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

[6] A. Lee et al., "A New Phonetic Tied-Mixture Model for Efficient Decoding," Proc. *ICASSP2000*, SP-P1-8, June 2000.

[7] K. Itou et al., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of the Acoustical Society of Japan (E)*, vol.20, no.3, pp.199–206, May 1999.

[8] J.L. Gauvain et al., "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, vol.2, no.2, pp.291–298, April 1994

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

[10] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995

[11] T. Kudo et al., "Fast Methods for Kernel-based Text Analysis," Proc. *ACL2003*, pp.24–pp.31, July 2003

[12] V. Wan et al., "SVMSVM: Support Vector Machine Speaker Verification Methodology," Proc. *ICASSP2003*, vol.2, pp.221–224, 2003