Children's Speech Recognition Based on Clustering Techniques

Atsunori Ogawa† *and Satoshi Takahashi*

Abstract

One of the problems in recognizing children's speech is that the acoustic features are quite different from those of adults and they vary with age. Another is that intra-speaker variations are especially large for the speech of young children. To solve these problems, we applied clustering techniques to children's speech and improved the recognition performance. We constructed a gender- and grade-balanced speech database containing the utterances of elementary school children and used it to investigate the word accuracies of children's speech as a function of their school grade by using a grade-independent child acoustic model and three adult acoustic models. The experimental results revealed that there were roughly three clusters of elementary school children. To capture these three clusters, we applied an automatic clustering algorithm to the children's speech and reduced the word error rate by 8.51% from the baseline. In addition to the automatic clustering, by applying adaptation based on the adult acoustic models to the children's cluster that was closest to the adults, we reduced the word error rate by 19.86% from the baseline.

1. Introduction

Speech recognition techniques that can effectively treat children's speech would obviously be beneficial in the educational field, such as foreign language learning programs [1], [2], and the entertainment industry, such as home video games [3], [4]. Several previous studies, for example [5], [6], have reported that the acoustic features of children's speech are quite different from those of adults. The variations in phoneme and sentence durations, fundamental frequency, formant frequencies, and spectral envelopes of children's speech are much larger than those of adults and they tend to vary rapidly as a function of a child's age. Furthermore, it has also been reported that intra-speaker variations are especially large for the speech of young children. These variations make automatic recognition of children's speech a more challenging problem than the recognition of adults' speech.

To compensate for these variations and to improve the performance of the automatic recognition of children's speech, previous studies have applied normalization techniques, such as the frequency warping approach or vocal tract length normalization, to children's speech [6]-[8]. While these normalization techniques have brought great improvements, the recognition performance is still not as high as that for adults' speech.

In contrast to the normalization techniques, we applied clustering techniques to children's speech to capture their variations and improved the recognition performance. The remainder of paper is organized as follows. Section 2 briefly describes the children's speech database used in our experiments. Section 3 describes the recognition experiments to investigate the baseline recognition performance for children's speech. Section 4 describes how the clustering techniques were applied to children's speech. Section 5 presents the conclusion of this work.

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: ogawa.atsunori@lab.ntt.co.jp

2. Children's speech database

To carry out the recognition experiments on children's speech, we used an animation-character-based recording system (**Fig. 1**) similar to that in Ref. 9 to construct a speech database uttered by a gender- and grade-balanced group of 354 Japanese elementary school children in grades 1 to 6 (i.e., ages 6 to 12). For the training data, 150 phonetically balanced simple Japanese words were selected and divided into two equal sets. 294 children uttered 75 words in either of the sets, and 21,514 utterances were collected (**Table 1**). The recordings were difficult especially for children in lower grades. Many re-utterances were needed, and the task was very time consuming. After the recordings had been made, all utterances were checked and unacceptable ones such as noisy or mispronounced utterances were removed. For the evaluation data, 99 words that were different from the ones in the training data were selected. The remaining 60 children uttered them, and 5865 utterances were collected (**Table 2**). For comparison, we also collected utterances from adults that consisted of the 99 words in the children's evaluation data. Twelve male and 12 female speakers uttered them, and 2376 utterances were collected as the adults' evaluation data. All utterances were recorded with 16-bit quality at a sampling frequency of 12 kHz.

3. Baseline recognition performance

To investigate the baseline recognition performance

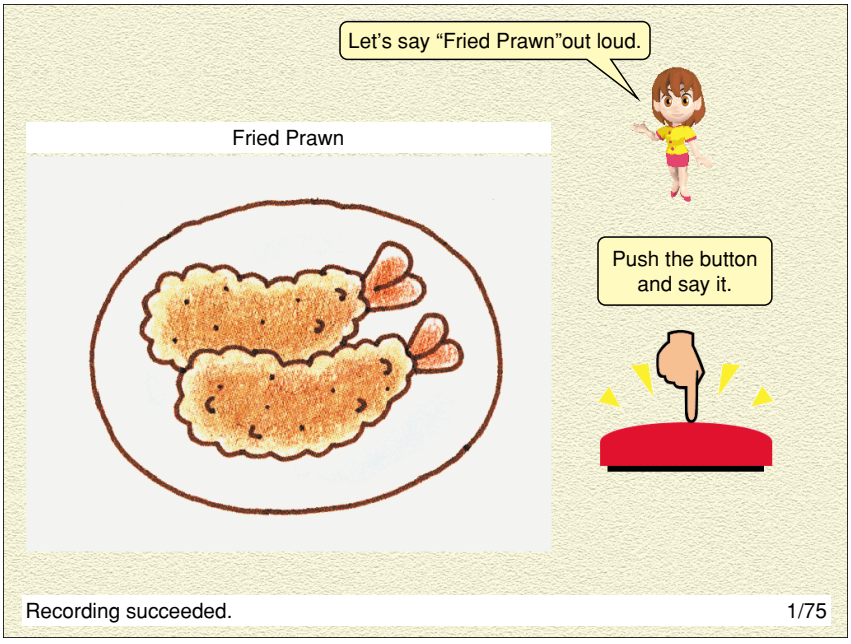

Fig. 1. Screen snapshot of the animation-character-based recording system. The original text was in Japanese.

Table 1. Training data.

Grade	Number of speakers			Number of utterances		
	Male	Female	Total	Male	Female	Total
1	20	20	40	1353	1409	2762
2	28	28	56	2072	2058	4130
3	29	21	50	2139	1547	3686
4	24	20	44	1775	1481	3256
5	26	26	52	1923	1916	3839
6	25	27	52	1846	1995	3841
Total	152	142	294	11,108	10,406	21,514

Table 2. Evaluation data.

Grade	Number of speakers			Number of utterances		
	Male	Female	Total	Male	Female	Total
1	5	5	10	495	493	988
2	5	5	10	495	475	970
3	5	5	10	493	491	984
4	5	5	10	493	458	951
5	5	5	10	495	495	990
6	5	5	10	487	495	982
Total	30	30	60	2958	2907	5865
Adult	12	12	24	1188	1188	2376

for children's speech, we carried out recognition experiments using a grade-independent child acoustic model and three adult acoustic models.

3.1 Experimental setup

Each utterance in the speech database was windowed every 10 ms with a 30-ms Hamming window to derive frame-by-frame acoustic feature vectors. The acoustic feature vector consisted of 12th-order Mel-frequency cepstral coefficients (MFCC), their first-order derivatives (ΔMFCC), and the first-order derivative of logarithmic power (Δpower). The order of each vector was 25.

All the acoustic models used in the following recognition experiments were 31-Japanese-phoneme-based left-to-right hidden Markov models (HMMs). Each HMM had 3 states and each state had 8 continuous-density diagonal-covariance Gaussian mixture components. As the baseline grade-independent child model ("child-HMM"), we trained tied-state triphone HMMs with 900 physical states by using the children's training data shown in Table 1. For comparison, we also trained three adult acoustic models—a gender-independent model ("adult-HMM") and two gender-dependent models ("adult-male-HMM" and "adult-female-HMM")—by using about 77 hours of adults' word utterances from another adult speech database. All of them had the structure of tied-state triphone HMMs with 2200 physical states.

For the evaluation vocabulary, 99 words in the evaluation data shown in Table 2 and the other 401 words were used. Thus, the vocabulary size for the evaluation experiments was 500. The speech recognition system was VoiceRex, which has been developed at NTT Cyber Space Laboratories [10].

3.2 Experimental results

Using the four acoustic models described in the previous section, we carried out 500-isolated-word recognition experiments on the evaluation data shown in Table 2. The results are shown in **Figs. 2** and **3**. In these figures, the word accuracies of the models are plotted as a function of school grade ("child gr.1–6"). The average accuracies for all of the children's speech ("child all"), adult male, and female speech ("adult ma." and "fe.") are also plotted. Figure 3 is an enlarged version of Fig. 2 for the data with word accuracy ranging from 70 to 100%.

First, we looked at the average word accuracy of each model. For all of the children's speech, the child-HMM gave the highest word accuracy (92.60%), whereas the adult-female- and adult-HMM gave rates about 5% lower, and the adult-male-HMM gave a very low rate. In the same way, the child-HMM gave very low word accuracy for the adult male speech, while it gave a high rate for the adult female speech. These results indicate that, on the

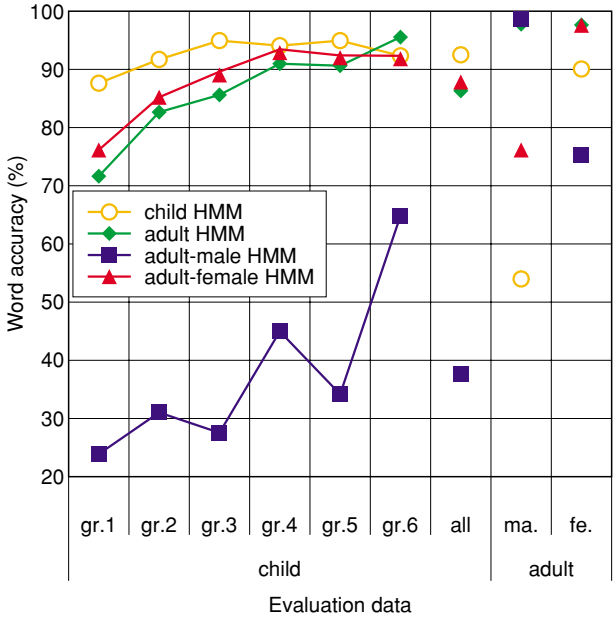

Fig. 2. Baseline word accuracies of children's speech as a function of their school grade.

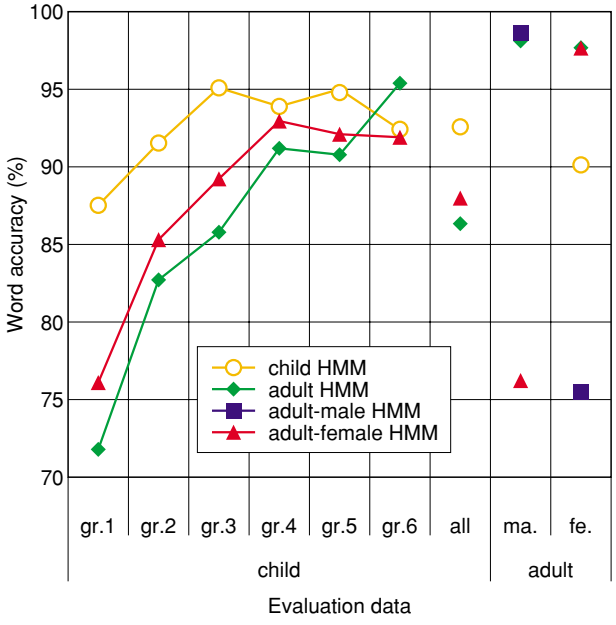

Fig. 3. Baseline word accuracies of children's speech as a function of their school grade (70–100%).

average, the acoustic features of children's speech are closest to those of adult female speech.

Next, we looked into the word accuracies of each model as a function of school grade. The word accuracies of the child-HMM were very high for the middle range of grades (gr. 3 to 5), but they were lower on either side of this range (i.e., gr. 1, 2, and 6). The accuracies for the adult-male-HMM remained very low through grades 1 to 5, but rapidly improved for grade 6. Those for the adult-female-HMM improved proportionally through grades 1 to 4, but fell slightly for grades 5 and 6. Those for the adult-HMM improved proportionally throughout the whole range of grades (with a slight fall at grade 5), and for grade 6 they exceeded the rate for the child-HMM. These results indicate that, as reported in [5], [6], the acoustic features of children's speech vary rapidly as a function of school grade and gradually approach those of adults' speech. Furthermore, differences in acoustic features caused by the gender difference appeared at higher grades. It seemed that there were roughly three clusters of children's speech, as follows: (1) infants, (2) typical elementary school children, (3) children whose speech had acoustic features close to those of adults. Moreover, we can expect the use of the adult-HMM to improve the recognition performance for the speech of children in higher grades.

4. Clustering of children's speech

Based on the above experimental results, we applied clustering techniques to the children's speech to improve the recognition performance for it.

4.1 Clustering algorithms

The experimental results indicated that it is difficult to represent the variations in acoustic features of children's speech by using a grade-independent child acoustic model. Therefore, we applied clustering techniques to the children's speech to represent the acoustic feature variations more accurately and to improve the recognition performance, especially for the speech of children in lower and higher grades. We used two clustering algorithms: 1) clustering based on the school grade and 2) automatic clustering. Based on the discussion in the previous section, we fixed the number of clusters to three.

For the grade-based clustering, we carried out two types of clustering. One was equal clustering for which the children's training data shown in Table 1 was divided into three equal clusters for: (1) grades 1 and 2, (2) grades 3 and 4, and (3) grades 5 and 6. We denote this clustering result as "gr.{1+2}{3+4}{5+6}". The other was clustering based on the word accuracies of the child-HMM as a function of school grade shown in Figs. 2 and 3. In this clustering, we divided the children's training data as follows: (1) grades 1 and 2, (2) grades 3, 4, and 5, (3) grade 6. We denote this clustering result as "gr.{1+2}{3+4+5}{6}".

For the automatic clustering algorithm, we used the clustering technique based on the anchor acoustic models described in **Fig. 4**. This algorithm considers how far the acoustic features of the speech are from each anchor HMM as well as how close the acoustic features of the speech are to each anchor HMM. For the anchor HMMs, we trained 8-mixture monophone HMMs of each grade by using the training data of each grade shown in Table 1. In addition, based on the discussion in the previous section, we also added the adult gender-independent HMM ("adult-HMM") to the anchor HMMs. The clustering procedure works on each utterance in the speech database to be clustered to capture the intra-speaker variation, which as reported in [5], [6] is especially large for young children. The speakers, especially the young children,

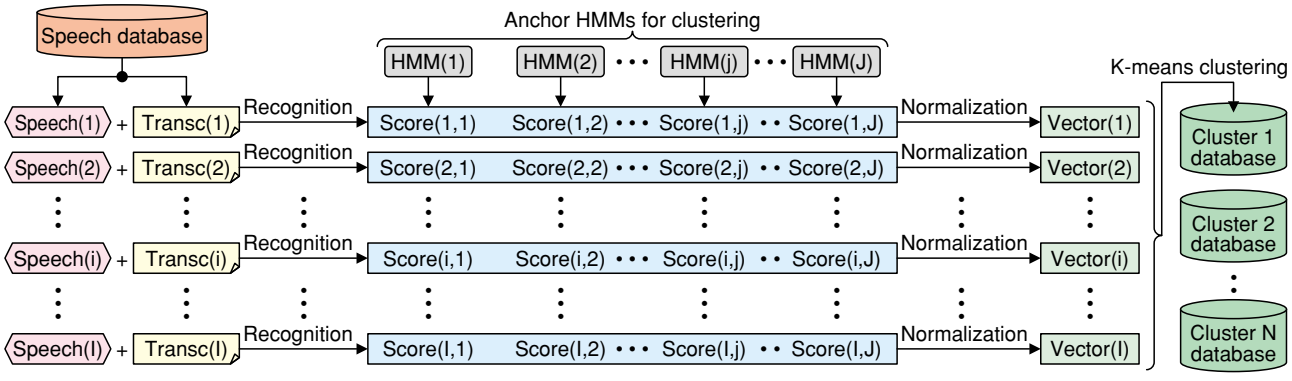

Fig. 4. Flow of anchor-acoustic-model-based automatic speech clustering.

could utter some words fluently like adults but could not utter other words steadily. Applying the clustering procedure to each utterance, we classified the utterance into the appropriate cluster, thereby generating a more accurate cluster-based model. The detailed clustering procedure was as follows.

1) For the anchor acoustic models, arrange J HMMs, HMM(j), j = 1, 2, ..., J.
2) Extract a speech speech(i) with its transcription transc(i) from the database to be clustered. Put these into all of the anchor HMMs to get the score vector of order J, score(i,j), j = 1, 2, ..., J, by recognition with the correct transcription.
3) Normalize the score score(i,j) at every vector index j with the maximum value max_j{score (i,j)}, j = 1, 2, ..., J, to get the normalized score vector vector(i) of order J.
4) Repeat steps 2) and 3) for all of the speech in the database and get I normalized score vectors vector(i), i = 1, 2, ..., I.
5) Carry out k-means clustering for the I normalized score vectors vector(i), i = 1, 2, ..., I, to get the desired number of speech clusters N.

The normalization procedure in step 3) is done to compensate for the differences in range of score vectors resulting from the differences in the lengths and transcriptions of the speeches. Using the seven anchor HMMs (J = 7) described above, we applied this automatic clustering procedure to the training data shown in Table 1 (I = 21,514) to generate the three clusters of children's speech (N = 3).

The automatic clustering result is shown in **Fig. 5**. In this figure, the classification ratios of each cluster are plotted as a function of school grade. It can be seen that cluster 1, the smallest cluster, mainly consists of speech of children in lower grades and contains very little speech of children in grades higher than 4. Cluster 2, the largest cluster, contains speech from all grades, but mainly consists of speech of children in middle grades. In contrast to cluster 1, cluster 3 mainly consists of speech of children in higher grades and contains very little speech of children in grades lower than 2. These three clusters are similar to the clusters discussed in section 3.2. We denoted this clustering result as "automatic clustering".

4.2 Recognition using cluster-based models

Based on the three clustering results, we trained four cluster-based child acoustic models. The cluster-based model was based on a clustering result consisting of triphone HMMs and Gaussian mixture models (GMMs) [11] corresponding to each cluster. Due to

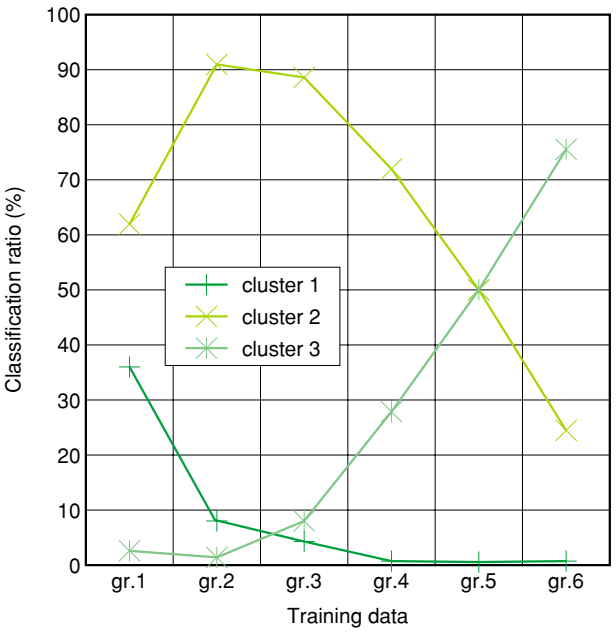

Fig. 5. Classification ratios of three automatic generated clusters as a function of school grade.

the limited amount of speech for each cluster, especially cluster 1 of the automatic clustering, each cluster's triphone HMMs were adapted from the baseline grade-independent child model ("child-HMM") by using maximum *a posteriori* (MAP) adaptation [12] using the speech of each cluster (**Fig. 6(a)**). The GMM of each cluster had 1 physical state of 64 continuous-density diagonal-covariance Gaussian mixture components trained by conventional maximum likelihood estimation (MLE) using only the speech of each cluster. Based on the model structure described above, we trained three cluster-based models corresponding to the three clustering results described in the previous section. Considering the discussion in section 3.2, for the case of the automatic clustering result, we also trained a cluster-based model whose triphone HMMs of cluster 3 were adapted from the adult gender-independent HMM ("adult-HMM"). We denote this cluster-based model as "cluster 3 from adult".

Using these four cluster-based child acoustic models, we carried out 500-isolated-word recognition experiments on the children's evaluation data shown in Table 2. In the evaluation stage, each input speech was assigned to the cluster whose GMM gave the highest matching score to the input speech and was recognized by the HMMs of the assigned cluster (**Fig. 6(b)**). **Table 3** shows the average word accuracies ("ACC") and error reduction rates ("ERR") relative to

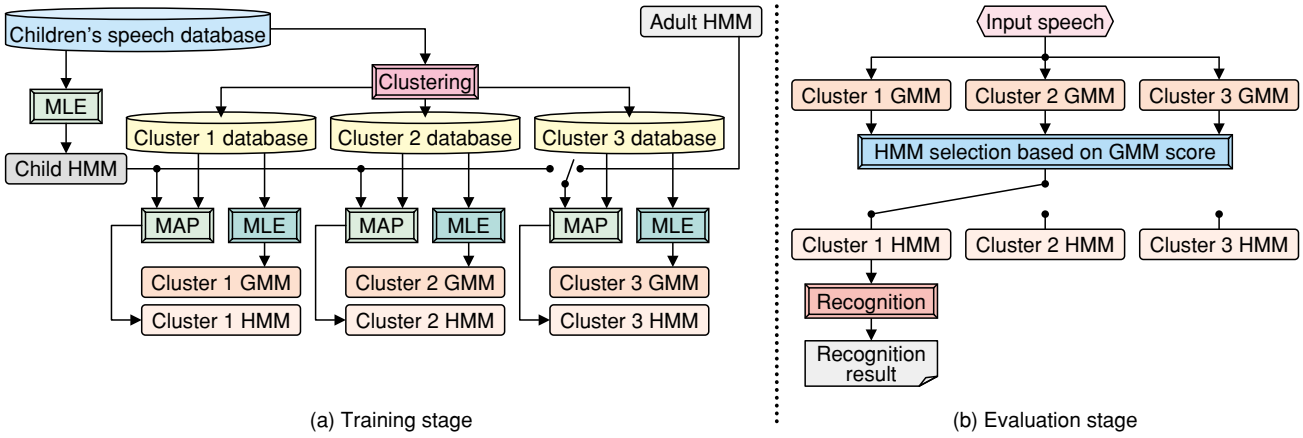

Fig. 6. Training and evaluation flows for cluster-based acoustic models.

Table 3. Average word accuracies and error reduction rates of cluster-based models.

Acoustic model	ACC (%)	ERR (%)
child-HMM	92.60	Baseline
gr.{1+2}{3+4}{5+6}	92.98	5.14
gr.{1+2}{3+4+5}{6}	92.72	1.62
Automatic clustering	93.23	8.51
Cluster 3 from adult	94.07	19.86

the child-HMM. We can see that all cluster-based models gave better accuracy than the baseline child model and that the automatic clustering based models gave better accuracy than the models based on school grade. Furthermore, comparing the two automatic clustering based models, "cluster 3 from adult" gave better accuracy than "automatic clustering". The word accuracies of the child-HMM and the two automatic-clustering-based cluster models are shown in **Fig. 7** as a function of school grade. Compared with child-HMM, "automatic clustering" gave better accuracy for the speech of the children of all grades and especially for that of children in lower grades. We can also see that "cluster 3 from adult" gave better accuracy than "automatic clustering" for the speech of children in the middle and higher grades.

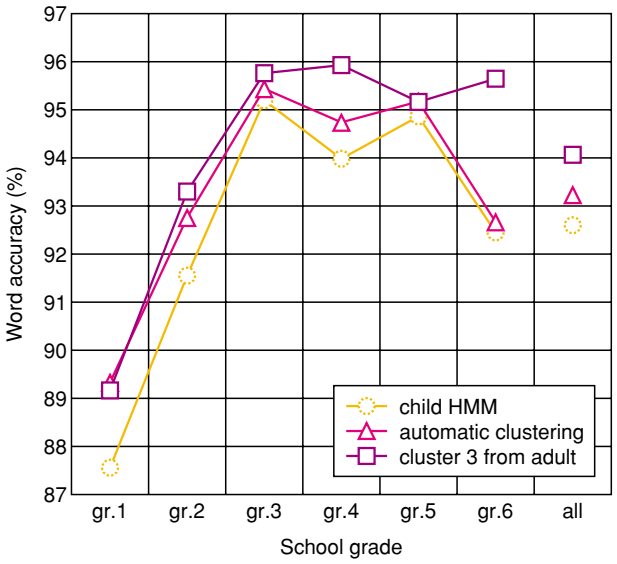

Fig. 7. Word accuracies of cluster-based models as a function of school grade.

5. Conclusion

In this paper, we have experimentally shown the difficulty of achieving accurate recognition of children's speech, which varies with age and among speakers, and obtained improved recognition performance of children's speech by applying clustering techniques. From baseline recognition experiments, we found that there were roughly three clusters of elementary school children and that the use of the adult gender-independent model could improve the recognition performance for the speech of children in higher school grades. To capture the three clusters of children, we applied an automatic clustering algorithm to the children's speech and reduced the error rate for the speech of children of all grades and especially for those of children in lower grades. In addition to the automatic clustering, when we applied adult-acoustic-model-based adaptation to the children's cluster that was closest to the adults, we got further error reduction for the speech of children in middle and higher grades. Our techniques improved the recognition performance of children's speech and we hope this will contribute to the creation of new prod-

ucts, especially in the educational field and the entertainment industry. However, the performances for children in lower grades remain low, so we must continue our efforts to improve them.

References

[1] http://www.benesse.co.jp/english/index.html
[2] http://www.unbalance.co.jp/nova_usagi/speaking/ (in Japanese).
[3] http://www.vivarium.co.jp/ (in Japanese).
[4] http://www.nintendo.co.jp/ds/ (in Japanese).
[5] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am., Vol. 105, No. 3, pp. 1455-1468, Mar. 1999.
[6] S. Narayanan and A. Potamianos, "Creating Conversational Interfaces for Children," IEEE Trans. on Speech and Audio Processing, Vol. 10, No. 2, pp. 65-78, Feb. 2002.
[7] A. Potamianos, S. Narayanan, and S. Lee, "Automatic Speech Recognition for Children," Proc. Eurospeech'97, Vol. 5, pp. 2371-2374, 1997.
[8] D. Giuliani and M. Gerosa, "Investigating Recognition of Children's Speech," Proc. ICASSP'03, Vol. 2, pp. 137-140, 2003.
[9] K. Shobaki, J. P. Hosom, and R. A. Cole, "The OGI Kids' Speech Corpus and Recognizers," Proc. ICSLP'00, Vol. 4, pp. 258-261, 2000.
[10] A. Ogawa, Y. Noda, and S. Matsunaga, "Novel Two-Pass Search Strategy Using Time-Asynchronous Shortest-First Second-Pass Beam Search," Proc. ICSLP'00, Vol. 4, pp. 290-293, 2000.
[11] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 1, pp. 72-83, Jan. 1995.
[12] J. L. Gauvain and C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, Apr. 1994.

Atsunori Ogawa
Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
He received the B.E. and M.E. degrees in information engineering from Nagoya University, Aichi in 1996 and 1998, respectively. Since joining NTT Laboratories in 1998, he has been engaged in research on speech recognition. In 2003, he received the Acoustical Society of Japan (ASJ) best poster presentation award. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and ASJ.

Satoshi Takahashi
Senior Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
He received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in information and computer science from Waseda University, Tokyo in 1987, 1989, and 2002, respectively. Since joining NTT Laboratories in 1989, he has been engaged in research on speech recognition. He is a member of IEICE and ASJ.