2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Learning Novel Objects Using Out-of-Vocabulary Word Segmentation and Object Extraction for Home Assistant Robots

Muhammad Attamimi, Akira Mizutani, Tomoaki Nakamura, Komei Sugiura,
Takayuki Nagai, Naoto Iwahashi, Hiroyuki Okada and Takashi Omori

Abstract— **This paper presents a method for learning novel objects from audio-visual input. Objects are learned using out-of-vocabulary word segmentation and object extraction. The latter half of this paper is devoted to evaluations. We propose the use of a task adopted from the RoboCup@Home league as a standard evaluation for real world applications. We have implemented proposed method on a real humanoid robot and evaluated it through a task called "Supermarket". The results reveal that our integrated system works well in the real application. In fact, our robot outperformed the maximum score obtained in RoboCup@Home 2009 competitions.**

I. Introduction

In order to realize robots that can naturally behave and interact with humans in our surrounding environments, the integration of robust hardware/software components, such as navigation, manipulation, speech processing, image processing and so forth, is required.

Here, if we focus our attentions on the spoken dialog technology of the robot for supporting our daily life, it turns out that many systems relies on the top-down method with given linguistic knowledge. Since it is implausible for the top-down system to equip all linguistic knowledge in advance, the robot cannot be expected to utter and/or recognize out-of-vocabulary(OOV) words. For example, even if a guiding robot can detect and learn a new person's face, it is impossible to recognize and call the person's name if the name is not registered in the system's lexicon.

On the other hands, robots that can acquire language from audio-visual input in a bottom-up manner have been proposed[1][2]. But, these bottom-up methods have a problem in practical performance.

Under this circumstance, we present a hybrid system which makes it possible to utter and recognize OOV words with the help of linguistic knowledge. More specifically, a robotic system that can learn novel objects is proposed in this paper. The proposed method utilizes template sentences for detecting and learning the OOV words and a rules-based dialogue management is used for usual interactions.

Now, let us consider the situation that a set of images and template sentence of a novel object is given. For learning/recognition of novel objects, following four major problems arise. In the learning phase, we need 1)noise robust speech recognition and 2)object extraction from cluttered scenes. In the recognition phase, we need 3)robust object recognition under various illumination conditions and 4)OOV words recognition and utterance. These problems are solved by integrating the sequential noise estimation, the noise suppression, the OOV word segmentation from audio input, and the voice conversion for the audio system. For the vision system, we develop the motion-attention based object extraction, and integrate it with the Scale-Invariant Feature Transform(SIFT)-based image recognition. By integrating these subsystems we can develop a system for learning novel objects, which is a requisite functionality for assistant robots.

In order to evaluate the proposed object learning system, we refer to the tasks of RoboCup@Home league[4] in a home environment. RoboCup@Home is a new RoboCup league that focuses on real-world applications and human-machine interaction with autonomous robots. The aim is to foster the development of useful robotic applications that can assist humans in everyday life. Since the clearly-stated rules exist, we think that these tasks are suited for evaluation standards of home assistant robots. We choose the "Supermarket" task, in which the robot is told to go and fetch some objects from a shelf. Before the task starts, a user shows an object to the robot and, at the same time, utters the template sentence such as "This is X." to make the robot learns the appearance and name (OOV word) of each object. And then the user, who could differ from the teacher, orders the robot using the specific sentence such as "Bring me X from the shelf."

Related works are included language acquisition[1]-[3] and audio-visual integrated robot systems[5][6]. Language acquisition is a fundamental solution for the problem, however, there are some practical problems as we mentioned earlier. In [3], OOV words acquisition using Hidden Markov Model(HMM) has been examined. Since OOV word segmentation is not involved in [3], the user has to utter only the OOV word. This makes the acquisition system extremely sensitive to noise. Moreover, they do not consider the speech synthesis.

As for assistant robots, many audio-visual integrated systems have been developed in the past. For example, in [5] a

Muhammad Attamimi, Akira Mizutani, Tomoaki Nakamura and Takayuki Nagai are with Department of Electronic Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan, {m_att, akira, naka_t}@apple.ee.uec.ac.jp, tnagai@ee.uec.ac.jp

Komei Sugiura and Naoto Iwahashi are with National Institute of Information and Communications Technology, 2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan, {komei.sugiura, naoto.iwahashi}@nict.go.jp

Hiroyuki Okada and Takashi Omori are with Department of Electronic Engineering, Tamagawa University, 6-1-1 Tamagawa-gakuen, Machida-shi, Tokyo 194-8610, Japan, h.okada@eng.tamagawa.ac.jp, omori@lab.tamagawa.ac.jp

978-1-4244-5040-4/10/$26.00 ©2010 IEEE

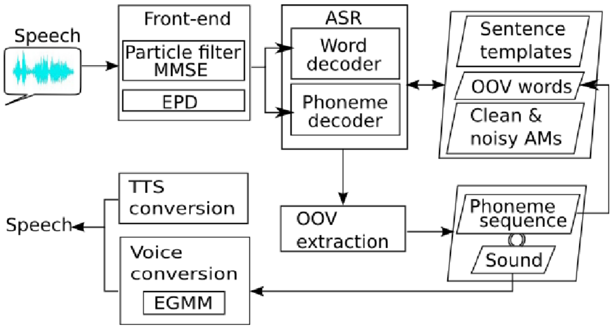

Fig. 1. Schematic of the out-of-vocabulary word extraction.

robot called "Jijo-2" has been developed. Jijo-2 can answer the location of persons and guide persons in office environments using multimodal interaction. In [6], a communication robot "Robovie" has been developed to realize interactive communication. Robovie uses similar speech recognition system as the one we use, however, OOV word segmentation is not considered since the robot recognizes only limited words. In any case, few systems are designed to deal with novel objects by integrating audio-visual information.

This paper is organized as follows; In Section II, the proposed system is divided into the audio and visual systems and each of these is explained separetely. The robot that we have used for the experiment and implementation issues are discussed in section III. After that experimental results are shown. In section IV, results of OOV segmentation are given and, in section V, we show some results on object detection/recognition followed by the results of the integrated system in section VI. Finally, section VII concludes this paper.

II. Proposed method

This section explains the speech and image processing parts of the proposed method.

A. Speech Processing

Fig.1 shows the schematic of the speech processing of the method. The proposed method uses an ASR (Automatic Speech Recognition) system called ATRASR[7]. ATRASR is an HMM (hidden Markov model)-based speech recognition system, and it is used as the front-end and word/phoneme decoder. The phoneme decoder is used for obtaining the phoneme sequence of out-of-vocabulary (OOV) words. Therefore, word-level and phoneme-level speech recognition are possible.

In order to suppress the noise, a particle filter is first applied to the online estimation of non-stationary noise, and then MMSE (Minimum Mean Square Error) estimation is used for noise reduction[8]. Voice activity detection is conducted by EPD (Endpoint Detection) based on frame's energy. This noise reduction part is of critical importance in @Home tasks since the noise condition ranges from 60 dBA to 85 dBA.

Acoustic models for the speech recognizer consist of "clean AMs" (male and female voices), which are trained using only clean voices, and "noisy AMs" (male and female voices), which are trained clean voices mixed with noise. This makes the speech recognition system robust in the noisy environment. By integrating these subsystems, we can solve Problem 1).

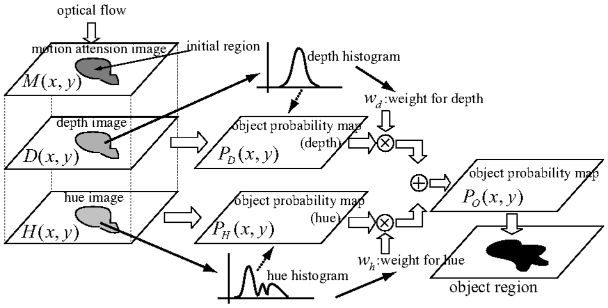

Fig. 2. The object extraction method based on the motion attention.

We use a template-based segmentation of words. In order to teach an OOV word, the user is supposed to say template sentences like "This is X." In terms of practical use, using a standard template sentence is reasonable since it is easy for users to understand how to teach a word to the robot. A set of segmented voice and phoneme sequence is registered in a database. The phoneme sequence is used for the recognition of an utterance with an OOV.

For generating an utterance with an OOV, the proposed method first converts the segmented voice recorded when the OOV is registered. The other part of the utterance is synthesized by XIMERA[9], which is a text-to-speech (TTS) conversion system. The OOV part is converted into the robot's voice since the original sound is the user's voice which is unnatural to be concatenated with synthesized voice. For the voice conversion is based on Eigenvoice Gaussian Mixture models (EGMMs)[10]. The recognized phoneme sequence of the OOV word is not used for synthesis since phoneme recognition accuracy is less than 90% and the number of utterances for teaching an OOV word is virtually constrained to one owing to the time constraint of RoboCup@Home. Although the evaluation of OOV synthesis is beyond the scope of this paper, it is practically important to generate a confirmation sentence such as "You said X. Is this correct?"

B. Image Processing

There are two problems left regarding the image processing system as we mentioned in the previous section. One of these problems is the object extraction from complex background. Since we assume that the user shows a target object to the robot, the object can be segmented out by paying attention to the motion cue. This fact motivates us to use the object extraction based on motion attention.

The idea behind the motion-attention based method is that the segments with synchronous motion are parts of an identical object. Hence the motion detector is first employed in the object detection subsystem. The motion detector extracts the initial object region at first. Then, the object information such as color (hue) and depth is taken from the region. In particular, hue and depth histograms are taken from the region and normalized. Since these two histograms can be considered as probability density functions of the target object, the object probability map of each component at each pixel location can be easily obtained. The weighted sum of these two object probability maps results in the object probability map. The weights are automatically assigned inversely proportional to the variance of each histogram. The map is binarized, and then final object mask is obtained by the connected component analysis. These processes are summarized in Fig.2. Although the stereo processing is included for obtaining the depth information, the image processing system still works around 10 fps for the object extraction.

In the learning phase, object images are simply collected, and then color histograms and SIFT features are extracted. These are used for the object detection and recognition.

When the system recognizes an object, the target object should be extracted from the scene. However, the same method in the learning phase is not applicable because the user cannot hold the object at this time. Therefore, the modified active search which uses color histogram and depth information for speeding-up the search time is applied for region extraction in the object recognition phase. We use SIFT descriptors for the recognition. In this time, we narrow down the candidates at first by using color information followed by the matching of SIFT descriptors, which are collected during the learning phase. It should be noted that the SIFT descriptors are extracted from multiple images taken from different viewpoints. Moreover, number of object images are reduced for speeding up the SIFT matching process by matching among within-class object images and discarding similar ones. This process is also useful for deciding the threshold on the SIFT matching score.

III. Robot platform

Fig.3 shows the robot, which is used in this research. The robot is based on the Segway RMP200 and consists of the following hardware components:

- Laser range finder (HOKYO UTM-30LX) is used for environmental mapping.
- Two iARMs (6DOF robotic arm manufactured by Exact Dynamics) and 1DOF grippers are mounted for object manipulation.
- Four on board PCs (Intel Core2Duo processor) are communicated each other through LAN.
- Sanken shotgun microphone CS-3e for audio input and YAMAHA speaker NX-U10 for audio output.
- A stereo camera is used for obtaining depth information.
- The camera and microphone are mounted on Directed Perception pan-tilt unit PTU-46-47.

The abilities of the robot other than the learning novel objects are listed below[11]:

1) Online SLAM(Simultaneous Localization and Mapping) and path planning
2) Object manipulation (RRT-based path planning)
3) Simple speech interaction in English and Japanese
4) Human detection and tracking using visual information
5) Searching objects in the living room environments
6) Face recognition using 2D-HMM
7) Gesture recognition

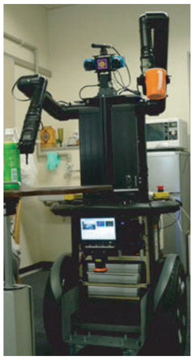

Fig. 3. The robot used in the experiment.

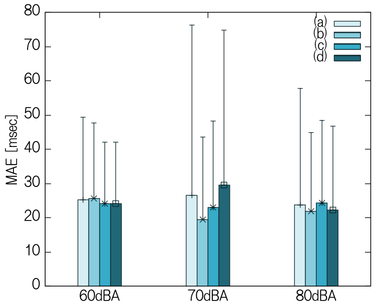

Fig. 4. The MAE of out-of-vocabulary word segmentation. (a) clean AMs (manual), (b) clean & noisy AMs (manual), (c) clean AMs (EPD), (d) clean & noisy AMs (EPD)

All of these abilities are required for completing tasks for the RoboCup@Home league. For example, in the ”who’s who?” task, the robot is expected to find unknown persons who are standing in the living room, and then has to learn their faces and names. Obviously, the proposed learning framework is also applicable to this task.

We have developed modular network architecture for the robot. The whole system is divided into four client modules (Vision, Audio, Robot and Task Modules) and a server. All modules are connected through the “server” with GigE and have subscription information that describes required information for the processing in each module. All information is gathered in the server and then the server forwards information to each module according to its subscription information. The “Task Module” works as a controller for each scenario of the task. This modular network architecture makes it relatively easy to share the job in the development stage of the robot system.

IV. Experiment 1: OOV word segmentation from audio input

The objectives of this experiment are the evaluation of (a) the voice activity detection, and (b) the error in the OOV word segmentation.

A. Experimental Conditions

First, we constructed database for evaluating the proposed method. In RoboCup@Home competitions, spoken dialogue between a user and robot is conducted in noisy conditions. Such noise ranging from 60 dBA to 85 dBA arises from announcements or music. In this situation, main difficulties are low SNR and the Lombard effect (human utterances are affected by noise).

In order to examine the robustness against these difficulties, the method was evaluated under the similar condition as used in the RoboCup@Home competitions. The noise source recorded in an exhibition hall was played in an anechoic chamber to duplicate a realistic environment. The noise level was either of 60 dBA, 70 dBA or 80 dBA, and a microphone was placed 30 cm away from a subject.

The utterances of subjects between the ages of 20 to 40 were recorded in the environment. Each subject is asked to utter eight sentences at intervals of 2 seconds. At this

TABLE I

RECOGNITION ACCURACY OF TEMPLATE SENTENCES.

	Clean AMs	Clean & noisy AMs
Manual	99.5	99.5
EPD	82.8	83.3

time, the subject is instructed to utter: "Eraser, This is *X*." The OOV word is any of "slippers", "stuffed lion", "stuffed tiger", "pen holder", "photo album", "wet tissue", "green tea" and "garbage" (in Japanese). We set up non-speech interval for 20 seconds before the first utterance for each noise level for adaptation.

The speech signals are sampled at 16 kHz with 16 bits/sample, and 25-dimensional features are computed. We use 12-dimensional MFCC(Mel Frequency Cepstral Coeffients), 12-dimensional Δ MFCC and logarithmic power as the features. The length of each frame and shift length were 20 msec and 10 msec, respectively. We use following two template sentences for the recognition.

"Eraser, This is *X*."
"Eraser, What is this?"
Here, *X* is represented as the free transition of phonemes.

B. Evaluation Procedure

In the evaluation, voice activity detection accuracy is first examined. The evaluation is conducted not as the coverage rate of detection, but as the recognition accuracy of template sentences. This is because an evaluation based on the coverage rate may overestimate the performance if an utterance is detected as many small pieces of signals. On the other hand, the evaluation should be done as the integrated system in a home assistant robot. Here, we use the following accuracy Acc' for the evaluation:

$$Acc' = \frac{number\ of\ correctly\ recognized\ utterances}{number\ of\ all\ utterances}, \quad (1)$$

where the numerator indicates that the segmented utterance is recognized correctly.

The accuracy of OOV word segmentation is evaluated by Mean Absolute Error(MAE) between detected start point and manually labeled one. It should be noted that the comparison is carried out only when the voice activity detection is succeeded. The OOV part can be obtained by cutting out the waveform from OOV start position to the end of the sentence.

C. Experimental Results

Table I shows accuracy of voice activity detection. In the table, "Manual" and "EPD" represent the recognition accuracy of speech segmented by manual labeling and EPD, respectively. The columns compare the conditions regarding acoustic models.

From the table, it can be seen that Acc' of almost 100% is obtained in manual condition. This is reasonable since the task is regarded as an isolated word (sentence) recognition. On the other hand, Acc' was 83% in EPD condition. This is due to the fact incorrect voice activity detection deteriorated the accuracy. Specifically, an utterance is rejected when it is detected as multiple pieces. In other words, the problem lay in the EPD rather than the speech recognition. This is also supported by the fact that Acc' was almost 100% under the manual condition.

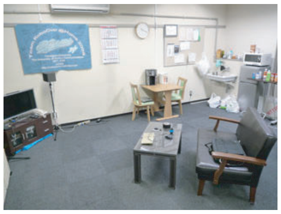

Fig. 5. Experimental environment.

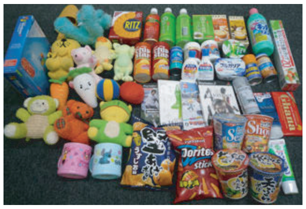

Fig. 6. Objects used for experiments.

Fig.4 shows the MAE of the OOV word segmentation. The segmentation accuracy is not plotted against the SNR since basically we cannot observe the SNR in RoboCup@Home competitions. For comparison, the absolute levels 60, 70, and 80 dBA were corresponding to 9.6dB, 2.6dB, and 2.0dB in SNR, respectively. In the figure, (a) and (b) are under the manual segmentation condition, and (c) and (d) are under the EPD condition. The error bars in the figure shows standard deviations.

Fig.4 shows that MAE was around 20-30 msec in three conditions. This result indicates that the accuracy is practically sufficient since mean duration of OOV words were approximately 670 msec.

V. EXPERIMENT 2: LEARNING OBJECT FROM IMAGES

Here, two experiments have been conducted to evaluate the vision system.

A. Evaluation of Object Extraction

We will discuss accuracy of the object extraction. The experiment has been carried out in an ordinary living room shown in Fig.5. We use fifty ordinary objects that can be roughly classified into stuffed toys, plastic bottles, canned coffee, cups, packages of DVD, cup noodles, snacks, boxes, and slippers as shown in Fig.6. A user teaches every object by showing and telling the name to the robot. The robot acquires fifty consecutive frames for each object and extracts the target object region from each image. Fig.7 shows some examples of object extraction. Accuracy of the detection is measured by recall and precision rates as shown in Fig.8. In the figure, "object region" indicates the manually labeled object region. Fig.9 shows the 2-D plot of recall vs. precision. Each point represents an average of a single object (50 frames). The averages of all objects are 0.89 for recall and 0.886 for precision, respectively. This fact implies that about 90% of the detected region contains the target object and covers about 90 % of the object. In the worst case, about 70% recall/precision rate has been obtained. As shown in the figure, the plastic bottle gives the worst F-measure. It turns out that the transparent part is responsible for the low recall rate.

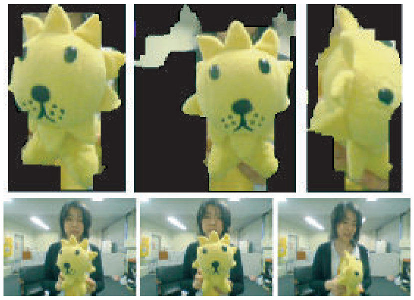

Fig. 7. Examples of object segmentation.

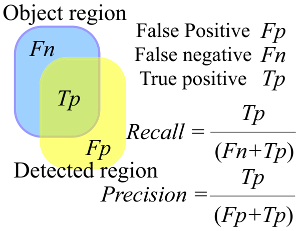

Fig. 8. Recall and precision rates.

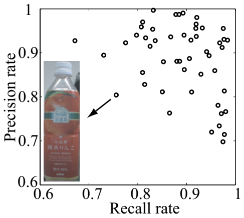

Fig. 9. Results of the object detection.

In this experiment, names of all objects have been taught to the robot simultaneously so that the learnt results can be used for the remaining experiments.

B. *Experiment of Object Recognition*

We use 50 objects, which have been learnt by the robot in the previous subsection. Four different locations(different lighting conditions) in the living room are selected and each object is recognized twice (with different poses) at one location. The results are listed in Table II. The average recognition rate is about 90%. A Major problem is that highly reflective surfaces of DVD packages cause saturation effect of the CCD.

VI. Experiment 3: Evaluation of Integrated System

We have implemented intregated audio-visual processing system on the robot and performed experiment in the living room. The purpose of this experiment is to show the proposed method is really useful in our everyday life scenario and to evaluate the total performance. We choose a task called "Supermarket" in the competition of RoboCup@Home league. A scenery of the RoboCup@Home competition in 2009 is shown in Fig.10 for reference.

A. *Task(Supermarket) and Score System*

The official rule book[4] describes the task as follows: A random person selected by the referees is using natural interaction (gestures, speech) without prior knowledge on how to use the robot, to get the robot to deliver a maximum number of three objects from one or more shelves within ten minutes. The robot is allowed to give instructions on how it can be operated. The three objects are taken from the set of standard objects. The team can choose one of the objects itself, the other two objects are chosen by the referees (respecting the physical constraints of the robot). The objects are then put on one or more shelves by the referees. A team has to announce whether the robot is able to get objects from different levels before the test starts.

TABLE II

Object recognition rates.

	Place 1	Place 2	Place 3	Place 4
Recognition rate	91%	88%	89%	90%

Fig. 10. A scenery of RoboCup @Home competition 2009.

The score system is defined as follows: 1)Correctly understanding which object to get: For every correctly understood object, 50 points are awarded, i.e. by clearly indicating the object. 2)Recognition: For every correctly found object, 150 points are awarded. 3)Grabbing: For every correct object retrieved from the shelf, 100 points are awarded. If the object was lifted for at least five seconds another 100 points are awarded. 4)Delivery: For every object delivered to the person, 100 points are awarded. 5)Different levels: For getting objects from different levels, 200 points are awarded. 6)Multimodal input: For using gestures in a meaningful way besides speech to communicate with the robot, 300 points are awarded.

Here, 6) is not considered since gesture recognition is not a scope of this paper. Hence the maximum total score is 1700 points in our experiment.

B. *Experimental Setup*

Fig.11 illustrates the map generated by the robot using SLAM and location of the shelf. We designed the task module according to the flowchart in Fig.12.

At first, a user interacts with the robot at the start position. Then the robot navigates to the shelf, recognizes the specified object, grasps it and comes back to the user. This process is repeated for three objects. Five persons have been selected and told to carried out the task twice. Therefore, the robot is supposed to bring 30 objects in total from the shelf. In each task, three objects are randomly chosen from 50 objects, which have been learnt by the robot in section V.

C. *Experimental Results*

Here we evaluate the results from three view points, that is, success rate of each process, process elapsed time, and the score as total performance. Fig.13 shows success rate of each process. From the figure, one can see that high success rates over 90 % are obtained except for the grasping process. In the grasping process, some objects, which have almost equal to the gripper in width, cause failures of grasping. It should be noted that the success rate of the speech recognition was 70.0 % if the retry was forbidden. When the retry was restricted to once, the rate went up to 90.0 %. In practice, the user

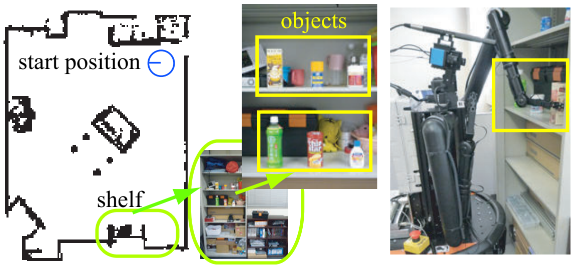

Fig. 11. The map and location of the shelf.

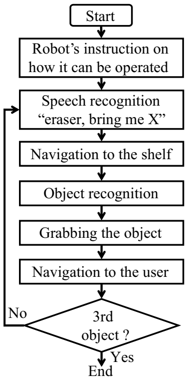

Fig. 12. Flowchart of the supermarket task.

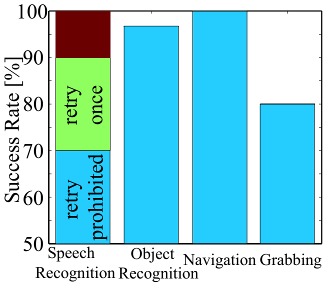

Fig. 13. Success rates.

can freely retry the speech recognition process within the limited time. This leaded to the success rate of 100 % in our experiment.

Fig.14 depicts elapsed time for each process (per object). From Fig. 14, it is confirmed that every trial has been completed within 10 min (elapsed time should be tripled and added 60 sec for the robot's instruction). In fact, average total time was 473 sec. It is interesting to see that second trial was completed faster than first one. More specifically, all users improved the efficiency in the speech interaction process. Finally, we evaluate the score. Fig.15 shows the comparison of scores among teams which have participated in the real competition of 2009. We can see from Fig.15, that the average of all final scores for the proposed method is 1555 points and the highest score was 1750 (50 points bonus is awarded for using onboard microphone) which means perfect score as we mentioned previously. In the real competition, the best score of this task was 1450 points. It should be noted that three objects are selected from ten standard objects set whose name list is given in advance in the real competition. Therefore, it is possible to register names of all objects to the lexicon manually. On the other hands, objects are chosen from fifty objects in our experiment. Moreover, no manual process is included in the learning process. Considering these conditions, the score obtained in this experiment seems good enough.

VII. Conclusions

In this paper, we have presented a method of learning novel objects in daily life environments. The proposed method is based on the OOV word segmentation and object extraction.

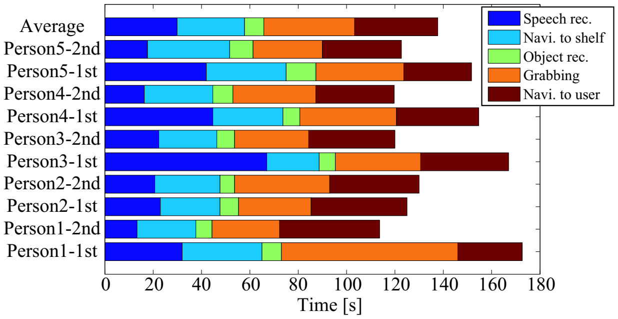

Fig. 14. Elapsed time of each process.

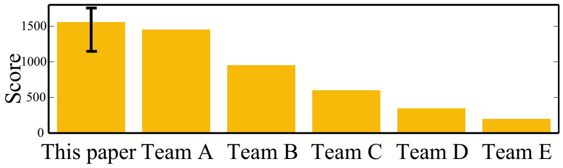

Fig. 15. The score comparison.

The algorithm has been implemented on a real humanoid robot and evaluated its performance through a task of RoboCup@Home league. The results show the validity of the proposed method in real applications.

VIII. Acknowledgements

This work was supported in part by Grant-in-Aid for Scientific Research(C)(20500186).

References

[1] Iwahashi, N., "Robot That Learn Language: Developmental Approach to Human-Machine Conversations", *Human-Robot Interaction* (Sanker, N., et al.(eds.)), I-Tech Education and Publishing, pp 95-118, 2007.

[2] Roy, D., "Grounding Words in Perception and Action: Computational Insights", *Trends in Cognitive Science*, vol.9, no.8, pp 389-396, 2005.

[3] Fujita, M., Hasegawa, R., Costa, G., Takagi, T., Yokono, J. & Shimomura, H. "An autonomous robot that eats information via interaction with human and environment", *in Proc. ROMAN*, pp.383-389, 2001.

[4] RoboCup@Home Rules & Regualtions, available at http://www.ai.rug.nl/robocupathome/documents/rulebook2009_FINAL.pdf, RoboCup@Home league committee 2009.

[5] Asoh, H., Motomura, Y., Asano, F., Hara, I., Hayamizu, S., Itou, K., Kurita, T., Matsui, T., Vlassis, N., Bunschoten, R., Kroese, B., "Jijo-2: An Office Robot that Communicates and Learns", *IEEE Intelligent Systems*, Vol.16, No.5, pp.46-55, 2001.

[6] Ishi, C., Matsuda, S., Kanda, T., Jitsuhiro, T., Ishiguro, H., Nakamura, S. and Hagita, N., "Robust Speech Recognition System for Communication Robots in Real Environments", *in Proc. of Int. Conf. on Humanoid Robots*, pp.340-345, 2006.

[7] Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Zhang, J., Yamamoto, H., Sumita, E. and Yamamoto, S., "The ATR multilingual speech-to-speech translation system", IEEE Transactions on Audio, Speech, and Language Processing, 2006, Vol. 14, pp. 365-376.

[8] Fujimoto, M. and Nakamura, S., "Sequential Non-Stationary Noise Tracking Using Particle Filtering with Switching Dynamical System", *in Proc. of ICASSP*, pp 769-772, 2006.

[9] Kawai, H., Toda, T., Ni, J., Tsuzaki, M. and Tokuda, K., "XIMERA: A New TTS from ATR Based on Corpus-Based Technologies", *in Proc. of Fifth ISCA Workshop on Speech Synthesis*, pp 179-184, 2004.

[10] Toda, T., Ohtani, Y. and Hagita, N., "One-to-Many and Many-to-One Voice Conversion Based on Eigen-voices", *in Proc. of IEEE ICASSP*, pp 1249-1252, 2007.

[11] Okada, H., Omori, T., Iwahashi, N., Sugiura, K., Nagai, T., Watanabe, N., Mizutani, A., Nakamura, T., Attamimi, M., "Team eR@sers 2009 in the @Home League Team Description Paper", *in Proc. of RoboCup International Symposium 2009 CD-ROM*, Graz, Austria, 2009.