Detection of Utterances Directed to Robots by Situated Understanding in Object Manipulation Tasks

Xiang Zuo[14], Naoto Iwahashi[13], Ryo Taguchi[15], Shigeki Matsuda[3],
Komei Sugiura[3], Kotaro Funakoshi[2], Mikio Nakano[2] and Natsuki Oka[4]

[1]Advanced Telecommunication Research Labs, Japan, [2]Honda Research Institute Japan, Japan,
[3]National Institute of Information and Communication Technology, Japan,
[4]Kyoto Institute of Technology, Japan, and [5]Nagoya Institute of Technology, Japan
sasyou@atr.jp

***Abstract*— In this paper, we propose a novel method to detect robot-directed speech by situated understanding in human-robot physical interaction. The originality of this work is the introduction of a Multimodal Semantic Confidence measure based domain classification method, which is used to decide whether the speech can be interpreted as a feasible action under the current physical situation in an object manipulation task. This measure is calculated by integrating speech, image, and motion confidence with weightings that are optimized by logistic regression. Then we integrated this method with human attention, and conducted experiments under the conditions of natural human-robot interaction.**

keywords: robot-directed speech detection, multimodal semantic confidence, human-robot interaction

I. INTRODUCTION

Robots are now being designed to be a part of the lives of ordinary people in social and home environments. One of the key issues for practical use is the development of user-friendly interfaces. Speech recognition is one of our most effective communication tools for use in a human-robot interface. In recent studies, many systems using speech-based human-robot interfaces have been implemented, such as [1]. For such an interface, the functional capability of detecting robot-directed (RD) speech is crucial. For example, user's speech directed to another human listeners should not be recognized as commands directed to a robot. To resolve this issue, methods have been implemented by many studies, mainly based on two approaches: (1) using the characteristics of the acoustic features of speech and, (2) using human attention such as gaze tracking or body-orientation detection.

As examples of the first approach, Itoh et al. [2] and Yamada et al. [3] have mentioned that some acoustic and linguistic features were affected by whether the dialogue partner is a human or a machine, while some acoustic features alone were affected by concurrent tasks such as a car-driving task. Yamagata et al. [4] have mentioned that acoustic differences of system requests and spontaneous speech usually appears on the head and the tail of the speech, and proposed acoustic based methods for discriminating RD speech from other speech by using power and pitch features [4], [5]. In these works, robot/system directed speech detection is performed based on analyzing the differences in acoustic and linguistic features between robot/system directed speech and other speech. However, this kind of method requires human users to adjust their speaking style or accent to fit the robot/system, which causes an additional burden to them.

On the other hand, human attention based systems have been discussed for a long time as natural, easy, ambient, or subconscious interactions [6]. Schilit et al. [7] have mentioned about context awareness in human-computer interaction, various systems using computers have been developed for providing appropriate services or interactions corresponding to the user's current situation by detecting various kinds of modal information. Lang et al. [8] proposed a multimodal attention system based by using a camera for human face recognition, two microphones for sound source localization, and a laser range finder for leg detection for a mobile robot to automatically recognize when and how long a person's attention is directed towards it for communication. Yonezawa et al. [9] proposed "Crossmodal Awareness" for a daily-partner robot to communicate with human based on the detection of human attention direction by the proportion of the user's gaze at the robot during her/his speech. These ideas were aimed toward ambient sensing of smart interactions for human-robot communication by human attention. However, they also raises an issue. That is, human users may say something irrelevant to the robot/computer while their attention is focused on it. Consider the following conversation where users A and B are talking while looking at the robot in front of them (Fig. 1).

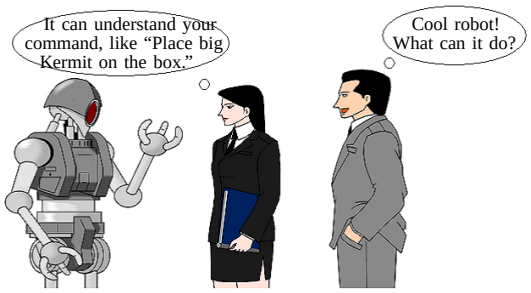

Fig. 1. People talking while looking at a robot.

A: Cool robot! What can it do?

B: It can understand your command, like "Place big Kermit on the box."

However, the speech here is not direct to robot. Moreover, even if user B makes a speech that resembles an RD speech ("place big Kermit on the box."), he does not really want to give such an order because the box and Kermit do not exist in the current situation. How can we build a robot that responds appropriately in this situation? Attention based methods are ineffective here. To address this kind of problem, in this work,

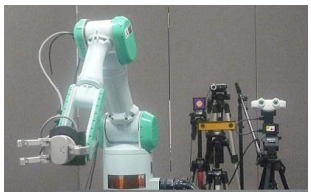

Fig. 2. Robot used in the object manipulation task.

Fig. 3. Scene corresponding to "Place big Kermit on the box."

we proposed a novel method to detect RD speech that is not only based on human attention detection but also based on domain classification of input speech between (1) the RD domain of RD speech and (2) out-of-domains (OOD) of other speech. The proposed method takes following steps:

(1) For each input speech, human behaviors is used to detect the direction of human attention during the speech. The speech without human attention focused on robot is rejected.

(2) For the speech with human attention focused on robot, domain classification is performed by calculating Multimodal Semantic Confidence (MSC) measure.

The main contribution of this work is the introduction of a MSC based domain classification method. MSC is a measure which decides whether the speech can be interpreted as a feasible action under the current physical situation in an object manipulation task. On the other hand, conventional studies on domain classification have typically focused on using speech recognition confidences [10], [11], or topic classification [12]. However, for a domain classification problem to be solved by a robot, we assume that in addition to speech signals, non-speech information would also be helpful because robots communicate in the real world not only with hearing but also with sight, touch, and so on. Therefore, in this work, domain classification is based on MSC measure, which is calculated by using both speech inputs and physical situations.

The rest of this paper is organized as follows: Section II gives the details of the object manipulation task. Section III describes the proposed method. The experimental methodology and results are presented in Section IV. Section V gives a discussion. Finally, Section VI concludes the paper.

II. Object Manipulation Task

The target task of this work is called an object manipulation task, in which the robot shown in Fig. 2 manipulates objects according to a user's speech. Figure 3 depicts a camera image of the current physical situation under the command utterance "Place big Kermit on the box." Here, the robot is told to place object 3 (big Kermit) on object 2 (box). The solid line shows the trajectory intended by the user. The trajectory can be interpreted by the positional change of the relationship between the moved object (trajector) and the reference object (landmark). In the case shown in Fig. 3, the trajector and landmark are objects 3 and 2, respectively.

III. The Method

An overview of our method is shown in Fig. 4. First, the audio signal is used for detecting the human speech by a GMM-based end-point detection (GMM-EPD) method which is used in ATRASR [13], and the camera images are converted to the human face angles. Both of these detected inputs are used to

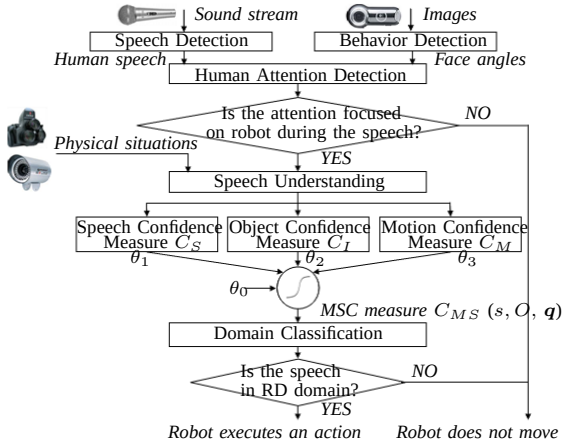

Fig. 4. An overview of the proposed method.

determine the direction of human attention. The human attention is estimated by the proportion of the human face toward to the robot during her/his speech. If the proportion is higher than fifty percent, the system judges the robot to be the focus of human attention during the speech. Then, for the speech with human attention focused on robot, speech understanding (as described later for detail) is performed to interpret the meaning of it as a feasible action by using the information on current physical situation. To evaluate the feasibility of the action, three confidence measures are calculated: C_S for speech, C_I for the static images of the objects, and C_M for the trajectory of motion (as described later for detail). The weighted sum of these confidence measures with a bias is inputted to a sigmoid function. The bias and the weightings, $\{\theta_0, \theta_1, \theta_2, \theta_3\}$, are optimized by logistic regression. Here, the MSC is defined as the output of the sigmoid function, and represents the probability that the speech is RD speech. Finally, domain classification is performed based on this probability. If the input speech is decided to be in RD domain, robot executes an action according to this speech.

A. Speech Understanding

We previously proposed a machine learning method called LCore that enables robots to acquire the capability of linguistic communication from scratch through verbal and nonverbal interaction with users [14]. In this study, we employ the speech understanding method used in LCore.

In the process of the speech understanding, we assume that speech s can be interpreted with conceptual structure $z =$ [(Motion: $\boldsymbol{w_W}$), (Trajector: $\boldsymbol{w_T}$), (Landmark: $\boldsymbol{w_L}$)], where $\boldsymbol{w_M}$, $\boldsymbol{w_T}$, and $\boldsymbol{w_L}$ represent the phrases describing motion, a trajector, and a landmark, respectively. (Or $z =$ [(Motion: $\boldsymbol{w_M}$), (Trajector: $\boldsymbol{w_T}$)] for an action that does not need a landmark). The order of the components in z represents the word sequence of s. For example, in Fig. 3, utterance "Place big Kermit on the box" is interpreted as [(Motion: "Place-on"), (Trajector: "big Kermit"), (Landmark: "box")].

Given speech s and current physical situation, speech understanding selects the optimal action a based on the conceptual structure z by a multimodal integrated user model that is trained by the interaction between the user and the robot. In this paper, a is defined as $a = (t, \xi)$, where t and ξ denote a trajector

and a trajectory of motion, respectively. The physical situation consists of current scene O, which includes the visual features and positions of all objects in it, and behavioral context $\boldsymbol{q}$, which includes information on which objects were a trajector and a landmark in the previous action and on which object the user is now holding. A user model integrates the five belief modules – (1) speech, (2) motion, (3) vision, (4) motion-object relationship, and (5) behavioral context – and is called shared belief. Each of the five belief modules in the shared belief is defined as follows:

Speech $\boldsymbol{B_S}$: This module is represented as the log probability of s conditioned by z, under lexicon L and grammar G_r. It is written as $\log P(s|z;L)P(z;G_r)$, where L includes pairs of a word and a concept, each of which represents the static image of the object and the motion as well as particles. G_r is represented by the statistical language model for possible robot commands. In this paper, the word is represented by HMMs using mel-scale cepstrum coefficients and their delta parameters (25-dimensional). ATRASR [13] was used for phoneme recognition.

Concept of static image of object $\boldsymbol{B_I}$: This module, which is represented as the log likelihood of Gaussian distributions in a multi-dimensional visual feature space (size, color (L^*, a^*, b^*), and shape), is written as $\log P(o_{t,f}|\boldsymbol{w_T};L)$ and $\log P(o_{l,f}|\boldsymbol{w_L};L)$, where $o_{t,f}$ and $o_{l,f}$ denote the visual features of trajector t and landmark l in scene O.

Concept of motion $\boldsymbol{B_M}$: This module is represented as the log likelihood of HMM using a sequence of vertical and horizontal coordinates of the trajectory ξ, given motion word $\boldsymbol{w_M}$. It is written as $P(\xi|o_{t,p},o_{l,p},\boldsymbol{w_M};L)$, where $o_{t,p}$ and $o_{l,p}$ denote the positions of t and l.

Motion-object relationship $\boldsymbol{B_R}$: This module represents the belief that in the motion corresponding to motion word $\boldsymbol{w_M}$, features $o_{t,f}$ and $o_{l,f}$ of objects t and l are typical for a trajector and a landmark, respectively. This belief is represented by a multivariate Gaussian distributions, $P(o_{t,f},o_{l,f}|\boldsymbol{w_M};R)$, where R is its parameter set.

Behavioral context $\boldsymbol{B_H}$: This module represents the belief that the current speech refers to object o, given behavioral context $\boldsymbol{q}$. It is written as $B_H(o,\boldsymbol{q};H)$, where H denotes its parameter set.

Given weighting parameter set $\boldsymbol{\Gamma}=\{\gamma_1...,\gamma_5\}$, the degree of correspondence between speech s and action a is represented by shared belief function Ψ written as

$$\begin{aligned}
&\Psi(s,a,O,\boldsymbol{q},L,G_r,R,H,\boldsymbol{\Gamma}) = \\
&\max_{z,l}\Big(\gamma_1 \log P(s|z;L)P(z;G_r) && [\boldsymbol{B_S}]\\
&\quad+\gamma_2\Big(\log P(o_{t,f}|\boldsymbol{w_T};L)+\log P(o_{l,f}|\boldsymbol{w_L};L)\Big) && [\boldsymbol{B_I}]\\
&\quad+\gamma_3 \log P(\xi|o_{l,p},o_{t,p},\boldsymbol{w_M};L) && [\boldsymbol{B_M}]\\
&\quad+\gamma_4 \log P(o_{t,f},o_{l,f}|\boldsymbol{w_M};R) && [\boldsymbol{B_R}]\\
&\quad+\gamma_5\Big(B_H(t,\boldsymbol{q};H+B_H(l,\boldsymbol{q};H)\Big)\Big), && [\boldsymbol{B_H}]
\end{aligned} \tag{1}$$

where conceptual structure z and landmark l are selected to maximize the value of Ψ. As the meaning of speech s under scene O, corresponding action $\hat{a}$ is determined by maximizing Ψ:

$$\hat{a}=(\hat{t},\hat{\xi})=\underset{a}{\operatorname{argmax}}\,\Psi(s,a,O,\boldsymbol{q},L,G_r,R,H,\boldsymbol{\Gamma}). \tag{2}$$

Finally, action $\hat{a}=(\hat{t},\hat{\xi})$, selected landmark $\hat{l}$, and conceptual structure $\hat{z}$ are outputted. Then the MSC measure is calculated based on these outputs.

B. MSC Measure

Next, we describe the proposed MSC measure. MSC measure C_{MS} is a measure of the feasibility for action $\hat{a}$ under the current physical situation and represents an RD speech probability. For input speech s, current scene O, and behavior context $\boldsymbol{q}$, C_{MS} is calculated based on the outputs of speech understanding $(\hat{a},\hat{l},\hat{z})$ and is written as

$$\begin{aligned}
C_{MS}(s,O,\boldsymbol{q}) &= P(domain=RD|s,O,\boldsymbol{q})\\
&= \frac{1}{1+e^{-(\theta_0+\theta_1 C_S+\theta_2 C_I+\theta_3 C_M)}},
\end{aligned} \tag{3}$$

where C_S, C_I, and C_M are the confidence measures of the speech, the object images, and the trajectory of motion. $\boldsymbol{\Theta}=\{\theta_0,\theta_1,\theta_2,\theta_3\}$ is applied to these confidence scores. Then, given a threshold δ, speech s with a MSC measure higher than δ is treated as in RD domain.

1) Speech Confidence Measure: The confidence measure of speech C_S is calculated by weighting the observed phoneme sequence's likelihood against the one of an unconstrained model sequence. It is conventionally used as a confidence measure for speech recognition [15], and is calculated as

$$C_S(s,\hat{z};A,G_p)=\frac{1}{n(s)}\log\frac{P(s|\hat{z};A)}{\max_{y\in L(G_p)}P(s|y;A)}, \tag{4}$$

where $n(s)$ denotes the analysis frame length of the input speech, $P(s|\hat{z};A)$ denotes the likelihood of word sequence $\hat{z}$ for input speech s by a phoneme acoustic model A, y denotes a phoneme sequence, and $L(G_p)$ denotes a set of possible phoneme sequences accepted by phoneme network G_p. For speech that matches robot command grammar G_r, C_S has a greater value than speech that does not match G_r.

The basic concept of this method is that it treats the likelihood of the most typical (maximum-likelihood) phoneme sequences for the input speech as a baseline. Based on this idea, the confidence measures of image and motion are defined as follows.

2) Image Confidence Measure: As a baseline of the image confidence measure, the likelihood of the most typical visual features for selected objects can be obtained by maximizing Gaussians of the objects. For visual features $o_{\hat{t},f}$ and $o_{\hat{l},f}$ of $\hat{t}$ and $\hat{l}$, which are represented by $\hat{\boldsymbol{w}}_{\boldsymbol{T}}$ and $\hat{\boldsymbol{w}}_{\boldsymbol{L}}$, respectively, the image confidence measure is calculated by the summed log-likelihood ratios of likelihood and baseline. It is written as

$$\begin{aligned}
&C_I(o_{\hat{t},f},o_{\hat{l},f},\hat{\boldsymbol{w}}_{\boldsymbol{T}},\hat{\boldsymbol{w}}_{\boldsymbol{L}};L)=\\
&\qquad\log\frac{P(o_{\hat{t},f}|\hat{\boldsymbol{w}}_{\boldsymbol{T}};L)P(o_{\hat{l},f}|\hat{\boldsymbol{w}}_{\boldsymbol{L}};L)}{\max_{o_f}P(o_f|\hat{\boldsymbol{w}}_{\boldsymbol{T}})\max_{o_f}P(o_f|\hat{\boldsymbol{w}}_{\boldsymbol{L}})},
\end{aligned} \tag{5}$$

where $P(o_{\hat{t},f}|\hat{\boldsymbol{w}}_{\boldsymbol{T}};L)$ and $P(o_{\hat{l},f}|\hat{\boldsymbol{w}}_{\boldsymbol{L}};L)$ denote the likelihood of $o_{\hat{t},f}$ and $o_{\hat{l},f}$, $\max_{o_f}P(o_f|\hat{\boldsymbol{w}}_{\boldsymbol{T}})$ and $max_{o_f}P(o_f|\hat{\boldsymbol{w}}_{\boldsymbol{L}})$ denote the maximum likelihood for object image models that are treated as baselines, and o_f denotes the visual features in object image models.

3) Motion Confidence Measure: As a baseline of the motion confidence measure, the likelihood of the most typical trajectory for motion word $\hat{\boldsymbol{w}}_{\boldsymbol{M}}$, given positions $o_{\hat{t},p}$ and $o_{\hat{l},p}$ of trajector $\hat{t}$ and landmark $\hat{l}$, can be obtained by maximizing HMMs of the

motion, while treating the trajector position as a variable. Then the motion confidence measure is calculated as

$$C_M(\hat{\xi}, \hat{\boldsymbol{w}}_{\boldsymbol{M}}; L) = \log \frac{P(\hat{\xi}|o_{\hat{t},p}, o_{\hat{l},p}, \hat{\boldsymbol{w}}_{\boldsymbol{M}}; L)}{\max_{\xi,o_p} P(\xi|o_p, o_{\hat{l},p}, \hat{\boldsymbol{w}}_{\boldsymbol{M}}; L)}, \quad (6)$$

where $P(\hat{\xi}|o_{\hat{t},p}, o_{\hat{l},p}, \hat{\boldsymbol{w}}_{\boldsymbol{M}}; L)$ denotes the likelihood for trajectory $\hat{\xi}$ and $\max_{\xi,o_p} P(\xi|o_p, o_{\hat{l},p}, \boldsymbol{w}_{\boldsymbol{M}}; L)$ denotes the likelihood of the maximum likelihood trajectory ξ of motion word $\hat{\boldsymbol{w}}_{\boldsymbol{M}}$, when the trajector position is variable, o_p denotes this variable.

4) Optimization of Weights: We now consider the problem of estimating weight $\boldsymbol{\Theta}$ in Eq. 3. The ith training sample is given as the pair of $C^i_{MS} = C_{MS}(s^i, O^i, \boldsymbol{q}^i)$ and teaching signal d^i. Thus, the training set $\mathbb{T}^N$ contains N samples:

$$\mathbb{T}^N = \{(C^i_{MS}, d^i)|i = 1, ..., N\}, \quad (7)$$

where d^i is 0 or 1, which represents OOD speech or RD speech, respectively.

A logistic regression model [16] is used for optimizing $\boldsymbol{\Theta}$. The likelihood function is written as

$$P(\boldsymbol{d}|\boldsymbol{\Theta}) = \prod_{i=1}^{N} (C^i_{MS})^{d^i} (1 - C^i_{MS})^{1-d^i}, \quad (8)$$

where $\boldsymbol{d}$= $(d^1, ..., d^N)$. $\boldsymbol{\Theta}$ is optimized by the maximum-likelihood estimation of Eq. 8 using Fisher's scoring algorithm [17].

IV. Experimental Methodology

A. Hardware Setting and Preparation of the Belief Modules

We conducted experiments using the robot shown in Fig. 2. This robot consists of a manipulator with 7 degrees of freedom (DOFs), a four-DOF multifingered grasper, a directional microphone for audio signal input, a web camera for face tracking, a stereo vision camera and an infrared camera for objects tracking, and a head unit for robot gaze expression.

The five belief modules ($\boldsymbol{B_S}$, $\boldsymbol{B_I}$, $\boldsymbol{B_M}$, $\boldsymbol{B_H}$ and $\boldsymbol{B_R}$) and shared belied function Ψ were learned beforehand by a method described in [14]. During the learning, 56 words, including 40 nouns and adjectives, 19 verbs representing 10 motions, and 7 particles were used. After learning, the values of parameter set $\boldsymbol{\Gamma}$= $\{\gamma_1 ..., \gamma_5\}$ in Eq. 1 were set to: $\gamma_1 = 1.00$, $\gamma_2 = 0.75$, $\gamma_3 = 1.03$, $\gamma_4 = 0.56$, and $\gamma_5 = 1.88$.

To evaluate the proposed method, we conducted two kinds of experiments: (1) learning of the MSC function (Eq. 3) by a batch processing, and (2) evaluating the proposed method in the object manipulation task.

B. Experiment 1

1) Experimental Setting: We conducted a batch experiment to learn the MSC function. The training and test data was obtained by taking following steps. First, we prepared 160 speech samples and manually labeled them as either RD or OOD (80 RD and 80 OOD). Then we gathered this speech from 16 subjects (8 males and 8 females) in a soundproof room with a SANKEN-CS5 directional microphone without noise. All of these subjects were native Japanese speakers, and each of them sat on a bench one meter from the microphone and produced the pre-determined speech in Japaneses. As a result, we obtained a clean speech corpus including 2560 speech samples. Finally, we paired each speech with a scene file, which was captured by the stereo vision camera. Each scene file included three objects in average. Figure 3 shows an example shot of a scene file. In this figure, the yellow box on object 3 represents the behavioral context $\boldsymbol{q}$, which means object 3 was manipulated most recently.

By using these clean speech-scene pairs, we performed leave-one-out cross-validation: 15 subjects' data was used as a training set, and the remaining 1 subject's data was used as a test set and repeated 16 times. During cross-validation, $\boldsymbol{\Theta}$ was optimized, and the averages were: $\hat{\theta}_0 = 5.9$, $\hat{\theta}_1 = 0.00011$, $\hat{\theta}_2 = 0.053$, and $\hat{\theta}_3 = 0.74$.

Then we tested these averages under noisy conditions. We obtained a noisy speech corpus by mixing each speech sample in the clean speech corpus with dining hall noise at a level from 50 to 52 dBA and then performed noise suppression [18]. The same scene files which were used to pair with the speech in clean corpus were also used here to produce the noisy speech-scene pairs. The evaluation under noisy conditions was performed by using these noisy speech-scene pairs without cross-validation.

The human attention was not used in this experiment. For each speech-scene pair, speech understanding was performed directly, then the MSC measure was calculated. During speech understanding, accuracies of 83% and 67% in phoneme recognition were obtained for the clean speech corpus and the noisy speech corpus, respectively.

For comparison, we used a baseline that performs RD speech detection based on the speech confidence measure.

2) Results: Figures 6 and 7 show the precision-recall curves for the clean and noisy speech corpora. The MSC measure and baseline performances are shown by "MSC" and "Baseline." The two lines clearly show that the MSC measure outperforms the baseline for RD speech detection, for both clean and noisy speech corpora. Moreover, the performances using the partial MSC measure are shown by "Speech-Image" (using the confidence measures of speech and image) and "Speech-Motion" (using the confidence measures of speech and motion). These lines show that both image and motion confidences helped to improve performance. The average maximum F-measures of MSC and baseline were 99% and 94% for clean speech corpus, respectively, and 95% and 83% for noisy speech corpus, respectively. MSC achieved an absolute growth of 5% with the clean speech corpus and 12% with the noisy speech corpus for average maximum F-measure. Then we performed the paired t-test and found that there were statistical differences ($p < 0.01$) between MSC and baseline for both clean and noisy speech corpora. Note that MSC obtains a high performance of 95% even for the noisy speech corpus, while the baseline obtains 83%. This means that MSC is particularly effective under noisy conditions.

To perform an RD domain classification by MSC, a threshold could be set to $\hat{\delta} = 0.79$, which maximized the average F-measure for the clean speech corpus. This means that a speech with a high RD speech probability of more than 79% will be treated as being in the RD domain and the robot will execute an action according to this speech.

C. Experiment 2

1) Experimental Setting: Next, by using the weighting set $\hat{\boldsymbol{\Theta}}$ and the threshold $\hat{\delta}$ optimized in experiment 1, we conducted an experiment with the object manipulation task. In this experiment, 2 subjects stay in front of the robot and ordered the robot to manipulate objects according to the current physical situation by Japaneses. The subjects were also allowed to chat with each

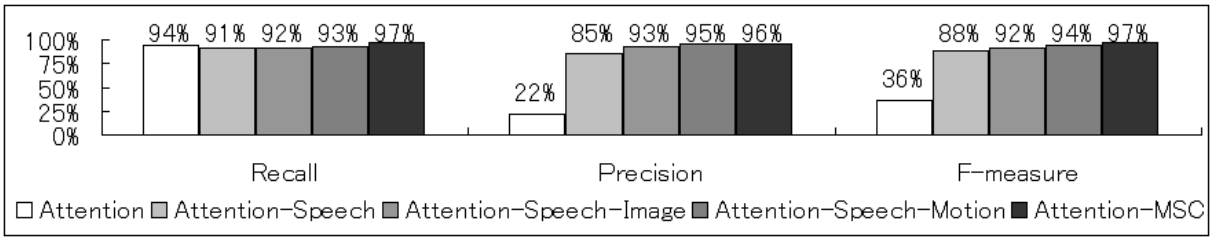

Fig. 5. The recall rate, precision rate and F-measure of five criteria obtained from the experiment.

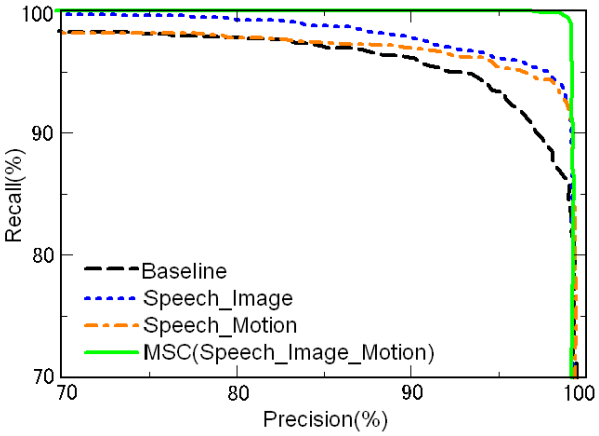

Fig. 6. Precision-recall curve for clean speech.

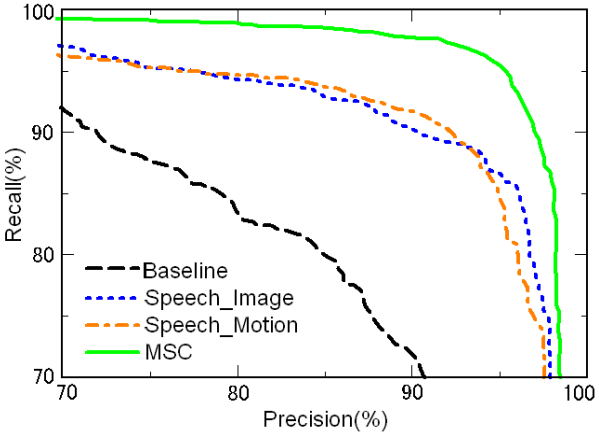

Fig. 7. Precision-recall curve for noisy speech.

other freely during the experiment. In the experiment, human attention was detected by face angles. All human speech as well as noise was input into the system.

We conducted a total of 4 sessions of this experiment by 4 pairs of subjects, each session lasted for 50 minutes. All subjects were adult males. During the experiment, surrounding noise about 48dBA from the robot's power always existed. In the experiments, a total of 983 speech was made, each of which was manually labeled as either RD or OOD after the experiments.

2) Result: First, the result of human attention detection is shown in table I. "RD" and "OOD" represent the quantities of RD speech or OOD speech that was manually labeled after

TABLE I
THE RESULT OF HUMAN ATTENTION DETECTION IN EXPERIMENT 2.

	With attention	Without attention	Total
RD	155	10	165
OOD	553	265	818
Total	708	275	983

the experiments, respectively. "With attention" and "Without attention" represent the quantities of speech during which human attention was focused on robot or not, respectively. "Total" represents the total speech made in the experiments. In this table, we can see that (1) almost all RD speech was made while subjects were facing to the robot and, (2) there was also a lot of OOD speech with the human attention focused on the robot. This caused a high recall rate and low precision rate.

Then we give a result of the proposed method by Fig. 5. We made comparisons among five criteria: (1) use human attention only, (2) use human attention and speech confidence measure, (3) use human attention and speech-image confidence measures, (4) use human attention and speech-motion confidence measures and, (5) use human attention and MSC measure. The recall rates, precision rates and F-measures of them are shown by "Attention", "Attention-Speech", "Attention-Speech-Image", "Attention-Speech-Motion" and, "Attention-MSC", respectively. In this figure, we can see that comparing to the human attention only, by using the proposed method (human attention and MSC measure), the precision rate was greatly enhanced from 22% to 96%, when the recall rate remained almost unchanged, and leaded to an absolute growth of 61% for F-measure. Moveover, in this experiment, both image and motion confidences also helped to improve performance.

V. DISCUSSION

Human attention is very important to distinguish the target of one's speech in human daily communications. However, it does not work effectively when used for a robot to detect RD speech in human-robot interactions because human does not treat the robot as a real person, and usually talk while pay attention to robot. According to this reason, we implemented MSC measure, and integrated it with human attention for robot to detect RD speech.

Moreover, for a robot in social and home environments, surrounding noise always exists. This decreases the reliability of speech recognition. Consequently, we believe that, in addition to the speech signal, other information should be used to improve the performance for RD speech detection. In this work, we

calculated MSC measure by speech, image and motion, and demonstrated its validity by experiments.

However, MSC has some limitations:

(1) To calculate the image and motion confidence measures, the manipulated object must be in a position that is visible to the robot. For speech that includes objects that are not in the robot's vision, the MSC measure will no longer be effective. This issue can be solved by an active exploration by the robot: when such an object is not visible, the robot will search for it in its surroundings.

(2) MSC is not suitable for tasks such as a dialog task that is not grounded in a physical situation. For such a task, a method that switches between the speech confidence and MSC measures should be implemented.

VI. Conclusion

In this paper, we proposed a novel method for RD speech detection by integrating human attention and MSC based domain classification, and evaluated it by experiments under conditions of human-robot interaction. The contribution of this paper is a new paradigm for a robot to use in distinguishing the information to which it should respond, which is as crucial importance for assistive robots supporting human users in daily environments.

References

[1] H. Asoh et al., "A spoken dialog system for a mobile office robot," in *Proc. on Eurospeech*, 1999, pp. 1139–1142.

[2] T. Itoh et al., "Linguistic and acoustic changes of user's utterances caused by different dialogue situations," in *Proc. on ICSLP*, 2002, pp. 545–548.

[3] S. Yamada et al., "Linguistic and acoustic features depending on different situations - the experiments considering speech recognition rate," in *Proc. on Interspeech*, 2005, pp. 3393–3396.

[4] T. Yamagata et al., "System request detection in conversation based on acoustic and speaker alternation features," in *Proc. on Interspeech*, 2007, pp. 2789–2792.

[5] T. Yamagata et al., "System request detection in human conversation based on multi-resolution gabor wavelet features," in *Proc. on Interspeech*, 2009, pp. 256–259.

[6] D. B. Koons et al., "Integrating simultaneous input from speech, gaze, and hand gestures," *MIT Press*, pp. 257–276, 1993.

[7] B. Schilit et al., "Context-aware computing applications," in *Proc. on WMCSA*, 1994.

[8] S. Lang et al., "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in *Proc. on ICMI*, 2003, pp. 28–35.

[9] T. Yonezawa et al., "Evaluating crossmodal awareness of daily-partner robot to user's behaviors with gaze and utterance detection," in *Proc. on CASEMANS*, 2009, pp. 1–8.

[10] R. A. Sukkar et al., "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 6, pp. 420–429, 1996.

[11] R. San-Segundo et al., "Confidence measures for spoken dialogue systems," in *Proc. on ICASSP*, 2001, pp. 393–396.

[12] I. R. Lane et al., "Out-of-domain utterance detection using classification confidences of multiple topics," in *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, 2007, pp. 150–161.

[13] S. Nakamura et al., "The atr multilingual speech-to-speech translation system," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 365–376, 2006.

[14] N. Iwahashi, "Robots that learn language: A developmental approach to situated human-robot conversations," *Human-Robot Interaction*, pp. 95–118, 2007.

[15] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, pp. 455–470, 2005.

[16] D. W. Hosmer et al., *Applied Logistic Regression*. Wiley-Interscience, 2009.

[17] T. Kurita, "Iterative weighted least squares algorithms for neural networks classifiers," in *Proc. on ALT*, 1992.

[18] M. Fujimoto et al., "Sequential non-stationary noise tracking using particle filtering with switching dynamical system," in *Proc. on ICASSP*, vol. 2, 2006, pp. 769–772.