特集論文 HAI (Human-Agent Interaction)

Detecting Robot-Directed Speech by Situated Understanding in Physical Interaction

Xiang Zuo — Advanced Telecommunication Research Labs and Kyoto Institute of Technology
d8821502(at)edu.kit.ac.jp

Naoto Iwahashi — Advanced Telecommunication Research Labs and National Institute of Information and Communications Technology
naoto.iwahashi(at)nict.go.jp

Kotaro Funakoshi — Honda Research Institute Japan Co., Ltd.
funakoshi(at)jp.honda-ri.com

Mikio Nakano — (affiliation as previous author)
nakano(at)jp.honda-ri.com

Ryo Taguchi — Advanced Telecommunication Research Labs and Nagoya Institute of Technology
taguchi.ryo(at)nitech.ac.jp

Shigeki Matsuda — National Institute of Information and Communications Technology
shigeki.matsuda(at)nict.go.jp

Komei Sugiura — (affiliation as previous author)
komei.sugiura(at)nict.go.jp

Natsuki Oka — Kyoto Institute of Technology
nat(at)kit.ac.jp

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

Summary

In this paper, we propose a novel method for a robot to detect robot-directed speech: to distinguish speech that users speak to a robot from speech that users speak to other people or to themselves. The originality of this work is the introduction of a multimodal semantic confidence (MSC) measure, which is used for domain classification of input speech based on the decision on 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, object, and motion confidence with weightings that are optimized by logistic regression. Then we integrate this measure with gaze tracking and conduct experiments under conditions of natural human-robot interactions. Experimental results show that the proposed method achieves a high performance of 94% and 96% in average recall and precision rates, respectively, for robot-directed speech detection.

1. INTRODUCTION

Robots are now being designed to be a part of the everyday lives of ordinary people in social and home environments. One of the key issues for practical use of such robots 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 works, many systems using speech-based human-robot interfaces have been implemented, such as [Asoh 99, Ishi 06]. For such an interface, the functional capability of detecting robot-directed (RD) speech is crucial. For example, a user's speech directed to another human listener should not be recognized as commands directed to a robot.

To resolve this issue, many works have used human physical behaviors to estimate the target of the user's speech. Lang et al. [Lang 03] proposed a method for a robot to detect the direction of a person's attention based on face recognition, sound source localization, and leg detection. Mutlu et al. [Mutlu 09] conducted experiments under conditions of human-robot conversation, and they studied how a robot could establish the participant roles of its conversational partners using gaze cues. Yonezawa et al. [Yonezawa 09] proposed an interface for a robot to communicate with users based on detecting the gaze direction during their speech. However, this kind of method raises the possibility that users may say something irrelevant to the robot while they are looking at it. Consider a situation where users A and B are talking while looking at

the robot in front of them (Figure 1).

A: Cool robot! What can it do?

B: It can understand your command, like "Bring me the red box."

Note that the speech here is referential, not directed to the robot. Moreover, even if user B makes speech that sounds like RD speech ("Bring me the red box"), she does not really want to give such an order because no red box exists in the current situation. How can we build a robot that responds appropriately in this situation?

To settle such an issue, the proposed method is based not only on gaze tracking but also on domain classification of the input speech into RD speech and out-of-domain (OOD) speech. Domain classification for robots in previous works were based mainly on using linguistic and prosodic features. As an example, a method based on keyword spotting has been proposed by [Kawahara 98]. However, in using such a method it is difficult to distinguish RD speech from explanations of system usage (as in the example of Figure 1). It becomes a problem when both types of speech contain the same "keywords." To settle this problem, a previous work [Takiguchi 08] showed that the difference in prosodic features between RD speech and other speech usually appears at the head and the tail of the speech, and they proposed a method to detect RD speech by using such features. However, their method also raised the issue of requiring users to adjust their prosody to fit the system, which causes them an additional burden.

In this work, the robot executed an object manipulation task in which it manipulates objects according to a user's speech. An example of this task in a home environment is a user telling a robot to "Put the dish in the cupboard." Solving this task is fundamental for assistive robots. In this task, we assume that a user orders the robot to execute an action that is feasible in the current situation. Therefore, the word sequences and the object manipulation obtained as a result of the process of understanding RD speech, should be possible and meaningful in the given situation. In contrast, word sequences and the object manipulation obtained by the process of understanding OOD speech would not be feasible. Therefore, we can distinguish between RD and OOD speech by using the feasibility for the corresponding word sequence and the object manipulation obtained from a speech understanding process as a measure. Based on this concept, we developed a multimodal semantic confidence (MSC) measure. A key feature of MSC is that it is not based on using prosodic features of input speech as with the method described above; rather, it is based on semantic features that determine whether the speech can be interpreted as a feasible action under the current physical situation. On the other hand, for an object manipulation task robots should deal with speech and image signals and to carry out a motion according to the speech. Therefore, the MSC measure is calculated by integrating information obtained from speech, object images and robot motion.

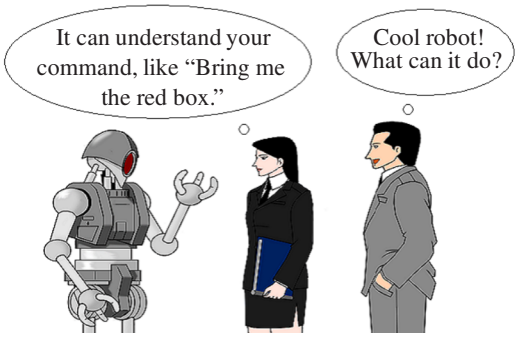


Fig. 1 People talking while looking at a robot.

The rest of this paper is organized as follows. Section 2 gives the details of the object manipulation task. Section 3 describes the proposed RD speech detection method. The experimental methodology and results are presented in Section 4, and Section 5 gives a discussion. Finally, Section 6 concludes the paper.

2. Object Manipulation Task

In this work, humans use a robot to perform an object manipulation task. Figure 2 and Figure 3 show the robot used in this task. It consists of a manipulator with 7 degrees of freedom (DOFs), a 4-DOF multi-fingered grasper, a SANKEN CS-3e directional microphone for audio signal input, a Point Grey Research Bumblebee 2 stereo vision camera for video signal input, a MESA Swiss Ranger SR4000 infrared sensor for 3-dimensional distance measurement, a Logicool Qcam Pro 9000 camera for human gaze tracking, and a head unit for robot gaze expression.

In the object manipulation task, users sit in front of the robot and command the robot by speech to manipulate objects on a table located between the robot and the user. Figure 4 shows an example of this task. In this figure, the robot is told to place Object 1 (Kermit) on Object 2 (big box) by the command speech "Place-on Kermit big box"*1, and the robot executes an action according to this speech. The solid line in Figure 4 shows the trajectory of the moving object manipulated by the robot. Figure 5 shows the sequential photographs of the robot executing an action according to command speech "Place-on Bar-

*1 Commands made in Japanese have been translated into English in this paper.

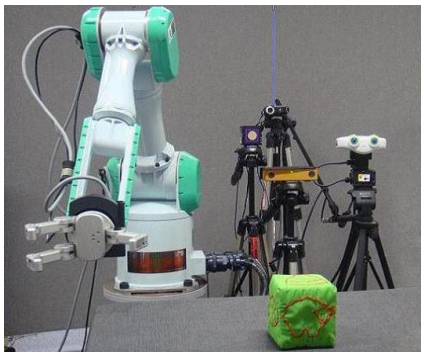


Fig. 2 Robot used in the object manipulation task.

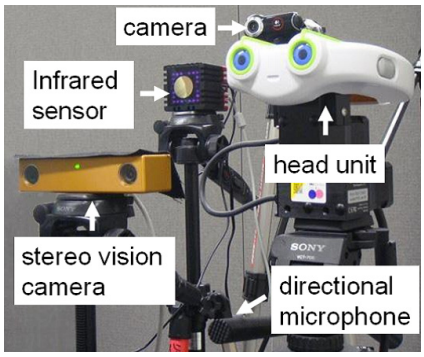


Fig. 3 Cameras, microphone, sensor and head unit of the robot.

bazoo*2 red box".

Commands used in this task are represented by a sequence of phrases, each of which refers to a motion, an object to be manipulated ("trajector"), or a reference object for the motion ("landmark"). In the case shown in Figure 4, the phrases for the motion, trajector, and landmark are "Place-on," "Kermit," and "big box," respectively. Moreover, fragmental commands without a trajector phrase or a landmark phrase, such as "Place-on big box" or just "Place-on," are also acceptable.

To execute a correct action according to such a command, the robot must understand the meaning of each word in it, which is grounded by the physical situation. The robot must also have a belief about the context information to estimate the corresponding objects for the fragmental commands. In this work, we used the speech understanding method proposed by [Iwahashi 07] to interpret the input speech as a possible action for the robot under the current physical situation. However, for an object manipulation task in a real-world environment, there may exist OOD speech such as chatting, soliloquies, or noise. Consequently, an RD speech detection method should be used.

*2 Kermit and Barbazoo are the stuffed toy's names used in our experiment.

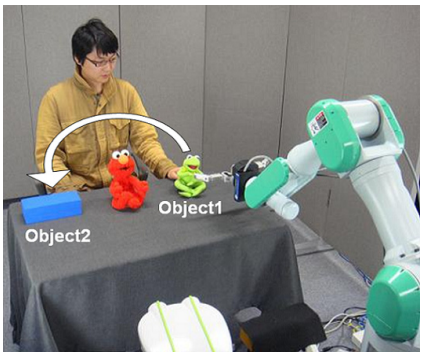


Fig. 4 Example of object manipulation tasks.

3. Proposed RD Speech Detection Method

The proposed RD speech detection method is based on integrating gaze tracking and the MSC measure. A flowchart is given in Figure 6. First, a Gaussian mixture model based voice activity detection method (GMM-based VAD) [Lee 04] is carried out to detect speech from the continuous audio signal, and gaze tracking is performed to estimated the gaze direction from the camera images*3. If the proportion of the user's gaze at the robot during her/his speech is higher than a certain threshold η, the robot judges that the user was looking at it while speaking. The speech during the periods when the user is not looking at the robot is rejected. Then, for the speech detected while the user was looking at the robot, speech understanding is performed to output the indices of a trajector object and a landmark object, a motion trajectory, and corresponding phrases, each of which consists of recognized words. Then, three confidence measures, i.e., for speech (C_S), object image (C_O) and motion (C_M), are calculated to evaluate the feasibilities of the outputted word sequence, the trajector and landmark, and the motion, respectively. The weighted sum of these confidence measures with a bias is inputted to a logistic function. The bias and the weightings $\{\theta_0, \theta_1, \theta_2, \theta_3\}$, are optimized by logistic regression [Hosmer 09]. Here, the MSC measure is defined as the output of the logistic function, and it represents the probability that the speech is RD speech. If the MSC measure is higher than a threshold δ, the robot judges that the input speech is RD speech and executes an action according to it. In the rest of this section, we give details of the speech understanding process and the MSC measure.

*3 In this work, gaze direction was identified by human face angle. We used faceAPI (http://www.seeingmachines.com) to extract human face angles from images captured by a camera.

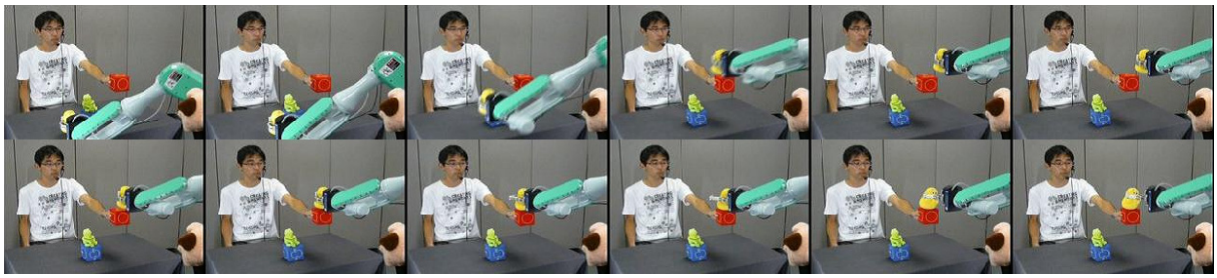


Fig. 5 Sequential photographs of the robot executing an action according to utterance "Place-on Barbazoo red box".

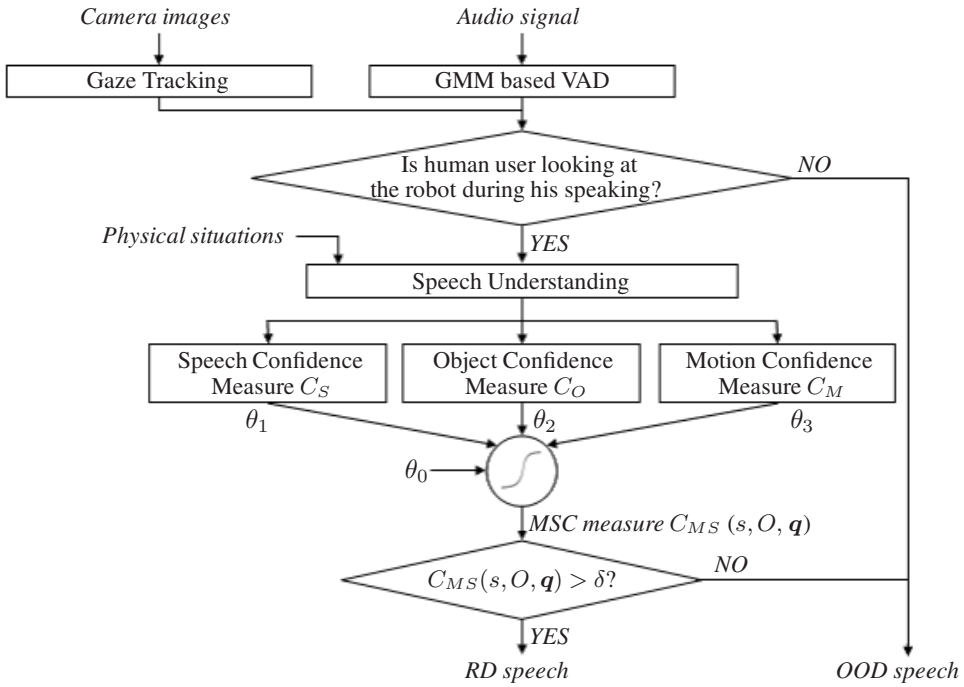


Fig. 6 Flowchart of the proposed RD speech detection method.

3·1 Speech Understanding

Given input speech s and a current physical situation consisting of object information O and behavioral context $\boldsymbol{q}$, speech understanding selects the optimal action a based on a multimodal integrated user model. O is represented as $O = \{(o_{1,f}, o_{1,p}), (o_{2,f}, o_{2,p}) \ldots (o_{m,f}, o_{m,p})\}$, which includes the visual features $o_{i,f}$ and positions $o_{i,p}$ of all objects in the current situation, where m denotes the number of objects and i denotes the index of each object that is dynamically given in the situation. $\boldsymbol{q}$ 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 is defined as $a = (t, \xi)$, where t and ξ denote the index of trajector and a trajectory of motion, respectively. A user model integrating the five belief modules – (1) speech, (2) object image, (3) motion, (4) motion-object relationship, and (5) behavioral context – is called an integrated belief. Each belief module and the integrated belief are learned by the interaction between a user and the robot in a real-world environment.

§1 Lexicon and Grammar

The robot has basic linguistic knowledge, including a lexicon L and a grammar G_r. L consists of pairs of a word and a concept, each of which represents an object image or a motion. The words are represented by the sequences of phonemes, each of which is represented by HMM using mel-scale cepstrum coefficients and their delta parameters (25-dimensional) as features. The concepts of object images are represented by Gaussian functions in a multi-dimensional visual feature space (size, color (L^*, a^*, b^*), and shape). The concepts of motions are represented by HMMs using the sequence of three-dimensional positions and their delta parameters as features.

The word sequence of speech s is interpreted as a conceptual structure $z = [(\alpha_1, \boldsymbol{w_{\alpha_1}}), (\alpha_2, \boldsymbol{w_{\alpha_2}}), (\alpha_3, \boldsymbol{w_{\alpha_3}})]$, where α_i represents the attribute of a phrase and has a value among $\{M, T, L\}$. $\boldsymbol{w_M}$, $\boldsymbol{w_T}$ and $\boldsymbol{w_L}$ represent the phrases describing a motion, a trajector, and a landmark, respectively. For example, the user's utterance "Place-on Kermit big box" is interpreted as follows: [(M, Place-on), (T, Kermit), (L, big box)]. The grammar G_r is a statisti-

cal language model that is represented by a set of occurrence probabilities for the possible orders of attributes in the conceptual structure.

§2 Belief modules and Integrated Belief

Each of the five belief modules in the integrated belief is defined as follows.

Speech B_S: This module is represented as the log probability of speech s conditioned by z, under grammar G_r.

Object image B_O: This module is represented as the log likelihood of $\boldsymbol{w_T}$ and $\boldsymbol{w_L}$ given the trajector's and the landmark's visual features $o_{t,f}$ and $o_{l,f}$.

Motion B_M: This module is represented as the log likelihood of $\boldsymbol{w_M}$ given the trajector's initial position $o_{t,p}$, the landmark's position $o_{l,p}$, and trajectory ξ.

Motion-object relationship B_R: This module represents the belief that in the motion corresponding to $\boldsymbol{w_M}$, features $o_{t,f}$ and $o_{l,f}$ are typical for a trajector and a landmark, respectively. This belief is represented by a multivariate Gaussian distribution of vector $[o_{t,f}, o_{t,f} - o_{l,f}, o_{l,f}]^T$.

Behavioral context B_H: This module represents the belief that the current speech refers to object o, given behavioral context $\boldsymbol{q}$.

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

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

where l denotes the index of landmark, o_t and o_l denote the trajector and landmark, respectively. Conceptual structure z and landmark o_l are selected to maximize the value of Ψ. Then, as the meaning of speech s, corresponding action $\hat{a}$ is determined by maximizing Ψ:

$$
\hat{a} = (\hat{t}, \hat{\xi}) = \operatorname*{argmax}_a \Psi(s,a,O,\boldsymbol{q},\boldsymbol{\Gamma}). \tag{2}
$$

Finally, action $\hat{a} = (\hat{t}, \hat{\xi})$, index of selected landmark $\hat{l}$, and conceptual structure (recognized word sequence) $\hat{z}$ are outputted from the speech understanding process.

§3 Learning the Parameters

In the speech understanding, each belief module and the weighting parameters $\boldsymbol{\Gamma}$ in the integrated belief are learned online through human-robot interaction in a natural way in an environment in which the robot is used [Iwahashi 07]. For example, a user shows an object to the robot while uttering a word describing the object to make the robot learn the phoneme sequence of the spoken word which refers to the object and the Gaussian parameters representing the object image concept based on Bayesian learning. In addition, the user orders the robot to move an object by making an utterance and a gesture, and the robot acts in response. If the robot responds incorrectly, the user slaps the robot's hand, and the robot acts in a different way in response. The weighting parameters $\boldsymbol{\Gamma}$ are learned incrementally, online with minimum classification error learning [Katagiri 98], through such interaction. This learning process can be conducted easily by a non-expert user. In contrast, other speech understanding methods need an expert to manually adjust the parameters in the methods, and the operation is not practical for ordinary users. Therefore, in comparison with other methods, the speech understanding method used in this work has an advantage in that it adapts to different environments, depending on the user.

3·2 MSC Measure

Next, we describe the proposed MSC measure. MSC measure C_{MS} is calculated based on the outputs of speech understanding and represents an RD speech probability. For input speech s and current physical situation (O, $\boldsymbol{q}$), speech understanding is performed first, and then C_{MS} is calculated by the logistic regression as

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

Logistic regression is a type of predictive model that can be used when the target variable is a categorical variable with two categories, which is quite suitable for the domain classification problem in this work. In addition, the output of the logistic function has a value in the range from 0.0 to 1.0, which can be used directly to represent an RD speech probability.

Finally, given a threshold δ, speech s with an MSC measure higher than δ is treated as RD speech. The $\boldsymbol{B_S}$, $\boldsymbol{B_O}$, and $\boldsymbol{B_M}$ are also used for calculating C_S, C_O, and C_M, each of which is described as follows.

§1 Speech Confidence Measure

Speech confidence measure C_S is used to evaluate the reliability of the recognized word sequence $\hat{z}$. It is calculated by dividing the likelihood of $\hat{z}$ by the likelihood of a maximum likelihood phoneme sequence with phoneme

network G_p, and it is written as

$$C_S(s,\hat{z}) = \frac{1}{n(s)} \log \frac{P(s|\hat{z})}{\max_{u \in L(G_p)} P(s|u)}, \quad (4)$$

where $n(s)$ denotes the analysis frame length of the input speech, $P(s|\hat{z})$ denotes the likelihood of $\hat{z}$ for input speech s and is given by a part of $\boldsymbol{B_S}$, u 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 speech confidence measure is conventionally used as a confidence measure for speech recognition [Jiang 05]. The basic idea 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 object and motion confidence measures are defined as follows.

§ 2 Object Confidence Measure

Object confidence measure C_O is used to evaluate the reliability that the outputted trajector $o_{\hat{t}}$ and landmark $o_{\hat{l}}$ are referred to by $\hat{\boldsymbol{w}}_{\boldsymbol{T}}$ and $\hat{\boldsymbol{w}}_{\boldsymbol{L}}$. It is calculated by dividing the likelihood of visual features $o_{\hat{t},f}$ and $o_{\hat{l},f}$ by a baseline obtained by the likelihood of the most typical visual features for the object models of $\hat{\boldsymbol{w}}_{\boldsymbol{T}}$ and $\hat{\boldsymbol{w}}_{\boldsymbol{L}}$. In this work, the maximum probability densities of Gaussian functions are treated as these baselines. Then, the object confidence measure C_O is written as

$$C_O(o_{\hat{t},f}, o_{\hat{l},f}, \hat{\boldsymbol{w}}_{\boldsymbol{T}}, \hat{\boldsymbol{w}}_{\boldsymbol{L}}) = \log \frac{P(o_{\hat{t},f}|\hat{\boldsymbol{w}}_{\boldsymbol{T}}) P(o_{\hat{l},f}|\hat{\boldsymbol{w}}_{\boldsymbol{L}})}{\max_{o_f} P(o_f|\hat{\boldsymbol{w}}_{\boldsymbol{T}}) \max_{o_f} P(o_f|\hat{\boldsymbol{w}}_{\boldsymbol{L}})}, \quad (5)$$

where $P(o_{\hat{t},f}|\,\hat{\boldsymbol{w}}_{\boldsymbol{T}})$ and $P(o_{\hat{l},f}|\,\hat{\boldsymbol{w}}_{\boldsymbol{L}})$ denote the likelihood of $o_{\hat{t},f}$ and $o_{\hat{l},f}$ and are given by $\boldsymbol{B_O}$, and $\max_{o_f} P(o_f|\,\hat{\boldsymbol{w}}_{\boldsymbol{T}})$ and $\max_{o_f} P(o_f|\,\hat{\boldsymbol{w}}_{\boldsymbol{L}})$ denote the maximum probability densities of Gaussian functions, and o_f denotes the visual features in object models.

For example, Figure 7(a) describes a physical situation under which a low object confidence measure was obtained for input OOD speech "There is a red box." The examples in Figure 7 are selected from the raw data of the experimental results. Here, by the speech understanding process, the input speech was recognized as a word sequence "Raise red box." Then, an action of the robot raising object 1 was outputted (solid line) because the "red box" did not exist and thus object 1 with the same color was selected as a trajector. However, the visual feature of object 1 was very different from "red box," resulting in a low value of C_O.

§ 3 Motion Confidence Measure

The confidence measure of motion C_M is used to evaluate the reliability that the outputted trajectory $\hat{\xi}$ is referred to by $\hat{\boldsymbol{w}}_{\boldsymbol{M}}$. It is calculated by dividing the likelihood of $\hat{\xi}$ by a baseline that is obtained by the likelihood of the most typical trajectory $\tilde{\xi}$ for the motion model of $\hat{\boldsymbol{w}}_{\boldsymbol{M}}$. In this work, $\tilde{\xi}$ is written as

$$\tilde{\xi} = \underset{\xi, o_p^{traj}}{\operatorname{argmax}} P(\xi|o_p^{traj}, o_{\hat{l},p}, \hat{\boldsymbol{w}}_{\boldsymbol{M}}), \quad (6)$$

where o_p^{traj} denotes the initial position of the trajector. $\tilde{\xi}$ is obtained by treating o_p^{traj} as a variable. The likelihood of $\tilde{\xi}$ is the maximum output probability of HMMs. In this work, we used the method proposed by [Tokuda 95] to obtain this probability. Different from $\hat{\xi}$, the trajector's initial position of $\tilde{\xi}$ is unconstrained, and the likelihood of $\tilde{\xi}$ has a greater value than $\hat{\xi}$. Then, the motion confidence measure C_M is written as

$$C_M(\hat{\xi}, \hat{\boldsymbol{w}}_{\boldsymbol{M}}) = \log \frac{P(\hat{\xi}|o_{\hat{t},p}, o_{\hat{l},p}, \hat{\boldsymbol{w}}_{\boldsymbol{M}})}{\max_{\xi, o_p^{traj}} P(\xi|o_p^{traj}, o_{\hat{l},p}, \hat{\boldsymbol{w}}_{\boldsymbol{M}})}, \quad (7)$$

where $P(\hat{\xi}|o_{\hat{t},p}, o_{\hat{l},p}, \hat{\boldsymbol{w}}_{\boldsymbol{M}})$ denotes the likelihood of $\hat{\xi}$ and is given by $\boldsymbol{B_M}$.

For example, Figure 7(b) describes a physical situation under which a low motion confidence measure was obtained for input OOD speech "Bring me that Chutotoro." Here, by the speech understanding process, the input speech was recognized as a word sequence "Move-away Chutotoro." Then, an action of the robot moving away object 1 from object 2 was outputted (solid line). However, the typical trajectory of "move-away" is for one object to move away from another object that is close to it (dotted line). Here, the trajectory of outputted action was very different from the typical trajectory, resulting in a low value of C_M.

§ 4 Optimization of Weights

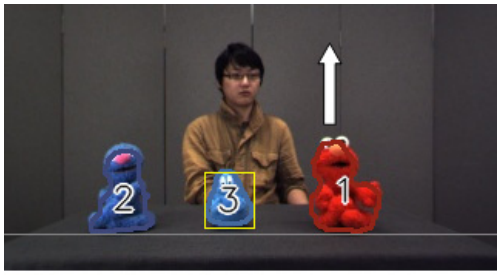
We now consider the problem of estimating weight $\boldsymbol{\Theta}$. The ith training sample is given as the pair of input signal $(s^i, O^i, \boldsymbol{q}^i)$ and teaching signal d^i. Thus, the training set $\mathbb{T}^N$ contains N samples:

$$\mathbb{T}^N = \{(s^i, O^i, \boldsymbol{q}^i, d^i)|i = 1, ..., N\}, \quad (8)$$

where d^i is 0 or 1, which represents OOD speech or RD speech, respectively. The likelihood function is written as

$$P(\boldsymbol{d}|\boldsymbol{\Theta}) = \prod_{i=1}^{N} (C_{MS}(s^i, O^i, \boldsymbol{q}^i))^{d^i} (1 - C_{MS}(s^i, O^i, \boldsymbol{q}^i))^{1-d^i}, \quad (9)$$

where $\boldsymbol{d} = (d^1, \ldots, d^N)$. $\boldsymbol{\Theta}$ is optimized by the maximum-likelihood estimation of Eq. (9) using Fisher's scoring algorithm [Kurita 92].



Input speech: “There is a red box.”
Recognized as: [Raise red box.]

(a) Case for object confidence measure

Input speech: “Bring me that Chutotoro.”
Recognized as: [Move-away Chutotoro.]

(b) Case for motion confidence measure

Fig. 7 Example cases where object and motion confidence measures are low. These examples are selected from the raw data of the experimental results.

4. Experiments

4·1 Experimental Setting

We first evaluated the performance of MSC. This evaluation was performed by an off-line experiment by simulation where gaze tracking is not used, and speech is extracted manually without the GMM based VAD to avoid its detection errors. The weighting set $\mathbf{\Theta}$ and the threshold δ were also optimized in this experiment. Then we performed an on-line experiment with the robot to evaluate the whole system.

The robot lexicon L used in both experiments has 50 words, including 31 nouns and adjectives representing 40 objects and 19 verbs representing ten kinds of motions. Figure 8 shows some of the objects used in the experiments. Figure 9 shows the examples for each motion. The solid line in each example represents the motion trajectory. L also includes five Japanese postpositions. Different from other words in L, each of the postpositions is not associated with a concept. By using the postpositions, users can speak a command in a more natural way. The parameter set $\mathbf{\Gamma}$ in Eq. (1) was $\gamma_1 = 1.00$, $\gamma_2 = 0.75$, $\gamma_3 = 1.03$, $\gamma_4 = 0.56$, and $\gamma_5 = 1.88$.

The speech detection algorithm was run on a Dell Precision 690 workstation, with an Intel Xeon 2.66GHz CPU and 4GB memory for speech understanding and the calculation of MSC measure. In the on-line experiment, we added another Dell Precision T7400 workstation with an Intel Xeon 3.2GHz CPU and a 4GB memory for the image processing and gaze tracking.

4·2 Off-line Experiment by Simulation

§1 Setting

The off-line experiment was conducted under both clean and noisy conditions using a set of pairs of speech s and scene information $(O, \boldsymbol{q})$. Figure 7(a) shows an example

Fig. 8 Some of the objects used in the experiments.

of scene information. The yellow box on object 3 represents the behavioral context $\boldsymbol{q}$, which means object 3 was manipulated most recently. We prepared 160 different such scene files, each of which included three objects on average. We also prepared 160 different speech samples (80 RD speech and 80 OOD speech) and paired them with the scene files. The RD speech samples included words that represent 40 kinds of objects and ten kinds of motions, which were learned beforehand in lexicon L. Each RD and OOD speech sample included 2.8 and 4.1 words on average, respectively. Table 1 shows examples of the speech spoken in the experiment. In addition, a correct motion phrase, correct trajectory, and landmark objects are given for each RD speech-scene pair. We then recorded the speech samples under both clean and noisy conditions as follows.

- Clean condition: We recorded the speech in a soundproof room without noise. A subject sat on a chair one meter from the SANKEN CS-3e directional microphone and read out a text in Japanese.
- Noisy condition: We added dining hall noise whose level was from 50 to 52 dBA to each speech record gathered under a clean condition.

We gathered the speech records from 16 subjects, including eight males and eight females. All subjects were native Japanese speakers. All subjects were instructed to

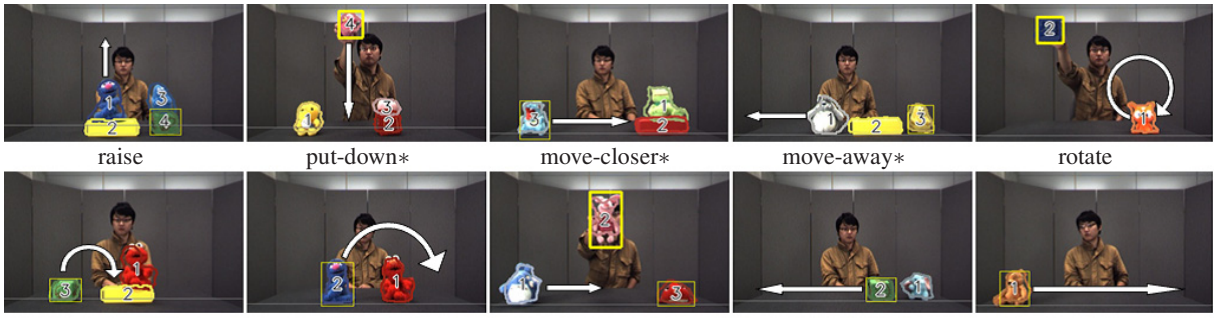


Fig. 9 Examples for each of the 10 kinds of motions used in the experiments. "∗" means that synonymous verbs are given in the lexicon for this motion.

speak naturally as if they were speaking to another human listener. As a result, 16 sets of speech-scene pairs were obtained, each of which included 320 pairs (160 for clean and 160 for noisy conditions). These pairs were inputted into the system. For each pair, speech understanding was first performed, and then the MSC measure was calculated. During speech understanding, a Gaussian mixture model based noise suppression method [Fujimoto 06] was performed, and ATRASR [Nakamura 06] was used for phoneme and word sequence recognition. With ATRASR, accuracies of 83% and 67% in phoneme recognition were obtained under the clean and noisy conditions, respectively.

The evaluation under the clean condition was performed by leave-one-out cross-validation: 15 subjects' data were used as a training set to learn the weighting $\boldsymbol{\Theta}$ in Eq. (3), and the remaining one subject's data were used as a test set and repeated 16 times. By cross-validation, the generalization performance for different speakers was evaluated. The average values of the weighting $\hat{\boldsymbol{\Theta}}$ learned by the training set in cross-validation were used for the evaluation under the noisy condition, where all noisy speech-scene pairs collected from 16 subjects were treated as a test set.

System performances was evaluated by recall and precision rates, which were defined as follows:

$$Recall = \frac{N^{cor}}{N^{total}}, \tag{10}$$

$$Precision = \frac{N^{cor}}{N^{det}}, \tag{11}$$

where N^{cor} denotes the number of RD speech correctly detected, N^{total} denotes the total number of existing RD speech, N^{det} denotes the total number of speech detected as RD speech by the MSC measure.

Finally, for comparison, four cases were evaluated for RD speech detection by using: (1) the speech confidence measure only, (2) the speech and object confidence measures, (3) the speech and motion confidence measures and, (4) the MSC measure.

Table 1 Examples of the speech spoken in the experiments.

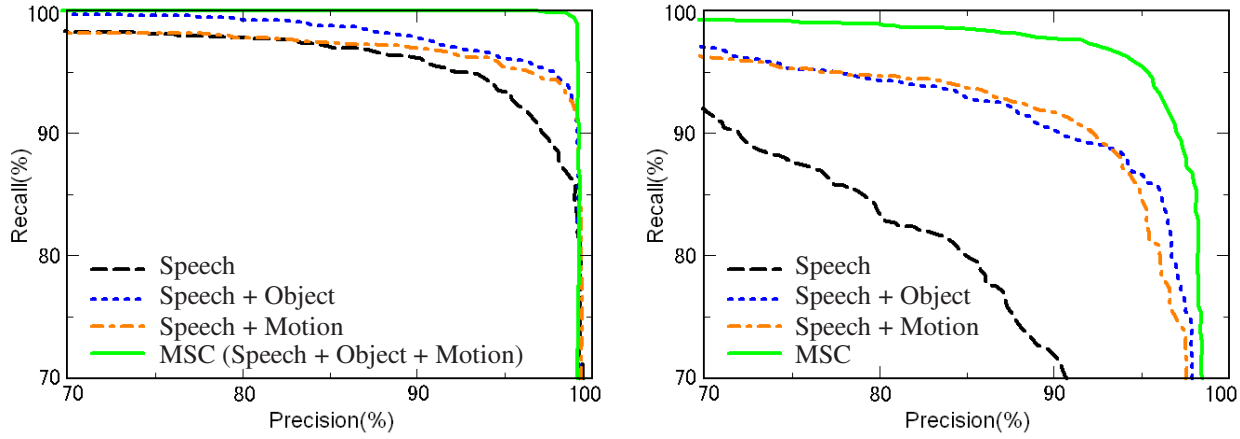
RD speech	OOD speech
Move-away Grover.	Good morning.
Place-on Kermit small box.	How about lunch?
Rotate Chutotoro.	There is a big Barbazoo.
Raise red Elmo.	Let's do an experiment.

We also evaluated the speech understanding using the RD speech-scene pairs. Differences between the output motion phrase, trajectory, and landmark objects and the given ones were treated as an error in speech understanding.

§2 Results

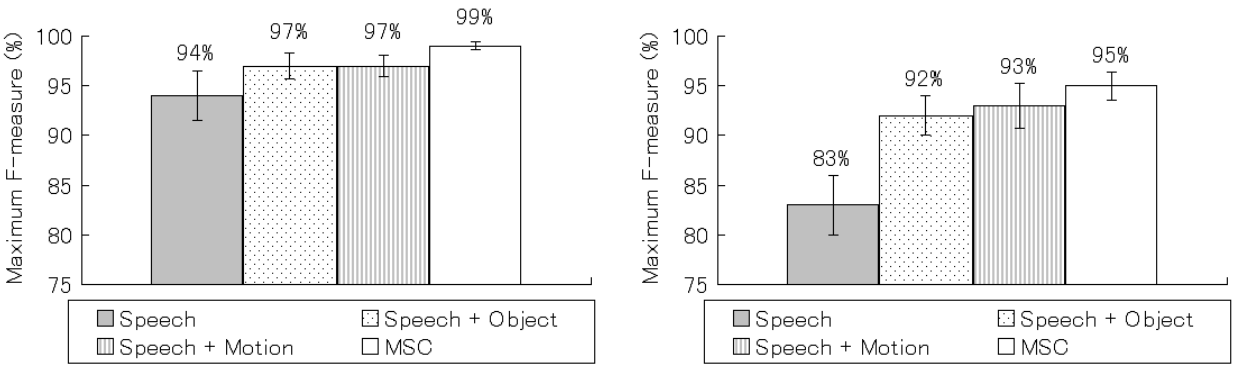
The average precision-recall curves for RD speech detection over 16 subjects under clean and noisy conditions are shown in Figure 10. The performances of each of four cases are shown by "Speech," "Speech + Object," "Speech + Motion," and "MSC." From the figures, we found that (1) the MSC outperforms all others for both clean and noisy conditions and, (2) both object and motion confidence measures helped to improve performance. The average maximum F-measures under clean and noisy conditions are shown in Figure 11. By comparing it with the speech confidence measure only, MSC achieved an absolute increase of 5% and 12% for clean and noisy conditions, respectively, indicating that MSC was particularly effective under the noisy condition. We also performed the paired t-test. Under the clean condition, there were statistical differences between (1) Speech and Speech + Object ($p < 0.01$), (2) Speech and Speech + Motion ($p < 0.05$), and (3) Speech and MSC ($p < 0.01$). Under the noisy condition, there were statistical differences ($p < 0.01$) between Speech and all other cases.

Examples of the raw data of the experimental results are shown in Figure 7 and Figure 12. The examples in Figure 7 are for OOD speech and have been explained in Sections 3.2.2 and 3.2.3. The examples in Figure 12 are for RD speech "Place-on Elmo big box" and "Jump-over



(a) Under clean condition (b) Under noisy condition

Fig. 10 Average precision-recall curves obtained in the off-line experiment.



(a) Under clean condition (b) Under noisy condition

Fig. 11 Average maximum F-measures obtained in the off-line experiment.

(a) “Place-on Elmo big box” (b) “Jump-over Barbazoo Totoro”

Fig. 12 Examples selected from the raw data of the experiment.

Table 2 Means (m) and variances (v) of weighted confidence measures for all RD and OOD speech obtained under noisy conditions.

	RD			OOD		
	C_S	C_O	C_M	C_S	C_O	C_M
m	-0.71	-0.88	-0.30	-3.8	-6.0	-3.3
v	1.1	0.55	0.72	6.4	130	23

Table 3 Accuracy of RD speech understanding.

	Total	Detected
Clean	99.8%	100%
Noisy	96.3%	98.9%

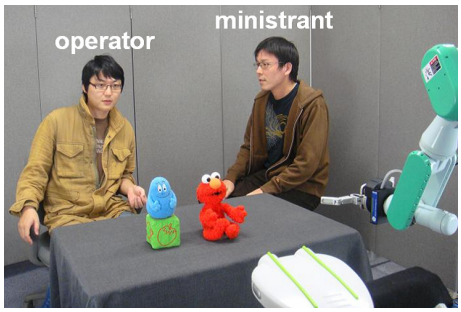


Fig. 13 Example of on-line experiment.

Barbazoo Totoro". These utterances were successfully detected by the MSC measure. The processing times (seconds) spent on the speech understanding process and the MSC-based domain classification was 1.09 and 1.36 for the examples shown in Figure 6(a) and (b), respectively, 1.39 and 1.36 for the examples shown in Figure 11(a) and (b), respectively. These times indicated that our method could respond quickly in practical human-robot interactions in real time. Table 2 shows the means and variances of the weighted confidence measures for all RD and OOD speech obtained under the noisy condition. Notice that the variances of C_O and C_M have large values for OOD speech, which means it is difficult to perform RD speech detection using C_O or C_M only.

In the experiment, weight $\mathbf{\Theta}$ and threshold δ were optimized under the clean condition. The optimized $\hat{\mathbf{\Theta}}$ were: $\hat{\theta}_0 = 5.9$, $\hat{\theta}_1 = 0.00011$, $\hat{\theta}_2 = 0.053$, and $\hat{\theta}_3 = 0.74$. The optimized $\hat{\delta}$ was set to 0.79, which maximized the average F-measure. This means that a piece of speech with an MSC measure of more than 0.79 will be treated as RD speech and the robot will execute an action according to this speech. The above $\hat{\mathbf{\Theta}}$ and $\hat{\delta}$ were used in the on-line experiment.

Finally, the accuracies of speech understanding using all RD speech and RD speech detected with the proposed method are shown in Table 3, where "Total" and "Detected" represent all RD speech and the detected RD speech, respectively, and "Clean" and "Noisy" represent clean and noisy conditions, respectively.

4·3 On-line Experiment Using the Robot

§1 Setting

In the on-line experiment, the whole system was evaluated by using the robot. In each session of the experiment, two subjects, an "operator" and a "ministrant," sat in front of the robot at a distance of about one meter from the microphone. The operator ordered the robot to manipulate objects in Japanese. He was also allowed to chat freely with the ministrant. Figure 13 shows an example of this experiment. The threshold η of gaze tracking was set to 0.5, which means that if the proportion of operator's gaze at the robot during input speech was higher than 50%, the robot judged that the speech was made while the operator was looking at it.

We conducted a total of 4 sessions of this experiment using 4 pairs of subjects, and each session lasted for about 50 minutes. All subjects were adult males. As with the off-line experiment, the subjects were instructed to speak to the robot as if they were speaking to another human listener. There was constant surrounding noise of about 48 dBA from the robot's power module in all sessions. For comparison, five cases were evaluated for RD speech detection by using (1) gaze only, (2) gaze and speech confidence measure, (3) gaze and speech and object confidence measures, (4) gaze and speech and motion confidence measures and, (5) gaze and MSC measure.

§2 Results

During the experiment, a total of 983 pieces of speech were made, each of which was manually labeled as either RD or OOD. The numbers of them are shown in Table 4. "w/ gaze and w/o gaze" show the numbers of speech productions that were made while the operator was looking/not looking at the robot. "RD/OOD " shows the numbers of RD/OOD speech productions that were manually labeled after the experiment. Aside from the RD speech, there was also a lot of OOD speech made while the subjects were looking at the robot (see "w/ gaze" in Table 4).

The accuracies of speech understanding were 97.6% and 98.1% for all RD speech and the detected RD speech, respectively. The average recall and precision rates for RD speech detection are shown in Figure 14. The performances of each of five cases are shown by "Gaze," "Gaze + Speech," "Gaze + Speech + Object," "Gaze + Speech + Motion," and "Gaze + MSC," respectively. By using gaze only, an average recall rate of 94% was obtained (see "Gaze" column in Figure 14(a)), which means that almost all of the RD speech was made while the operator was looking at the robot. The recall rate dropped to 90% by

Table 4 Numbers of speech productions in the on-line experiment.

	w/ gaze	w/o gaze	Total
RD	155	10	165
OOD	553	265	818
Total	708	275	983

integrating gaze with speech confidence measure, which means some RD speech was rejected by the speech confidence measure by mistake. However, by integrating gaze with MSC, the recall rate returned to 94% because the mis-rejected RD speech was correctly detected by MSC. In Figure 14(b), the average precision rate by using gaze only was 22%. However, by using MSC, these instances of OOD speech were correctly rejected, resulting in a high precision rate of 96%, which means the proposed method is particularly effective under situations where users make a lot of OOD speech while looking at a robot.

5. Discussion

5·1 Using in a Real World Environment

Although the proposed method was evaluated in our laboratory, we consider that our method could be used for real world environments because the used speech understanding method is adaptable to different environments. In some cases, however, physical conditions can dynamically change. For example, lighting conditions may change suddenly due to sunlight. The development of a method that works robustly in such variable conditions is future work.

5·2 Extended Applications

This work can be extended in many kinds of ways, and we mention some of them. Here, we evaluated the MSC measure under situations where users usually order the robot while looking at it. However, users possibly order a robot without looking at it under some situations. For example, in such an object manipulation task where a robot manipulates objects together with a user, the user may make an order while looking at the object which he is manipulating instead of looking at the robot itself. For such tasks, the MSC measure should be used separately without integrating it with gaze. Therefore, a method that automatically decides whether to use the gaze information according to the task and user situation should be implemented.

Moreover, aside from the object manipulation task, the MSC measure can also be extended to the multi-task dialog including both the physically grounded and ungrounded tasks. In the physically ungrounded tasks, users' utterances express no immediate physical objects or motions. For such dialog, a method that automatically switches between the speech confidence and MSC measures should be implemented. In the future works, we will evaluate the MSC measure for various dialog tasks.

In addition, we can use the MSC to develop an advanced interface for human-robot interaction. The RD speech probability represented by MSC can be used to provide feedback such as the utterance "Did you speak to me?", and this feedback should be made under situations where the MSC measure has an ambiguous value. Moreover, each of the object and motion confidence measures can be used separately. For example, if the object confidence measures for all objects in a robot's vision were particularly low, an active exploration should be executed by the robot to search for a feasible object in its surroundings, or an utterance such as "I cannot do that" should be made for situations where the motion confidence measure is particularly low.

Finally, in this work, we evaluated the MSC measure obtained by integrating speech, object and motion confidence measures. In addition, we can consider the use of the confidence measure obtained from the object-motion relationship. In the future, we will evaluate the effect of using this confidence measure.

6. Conclusion

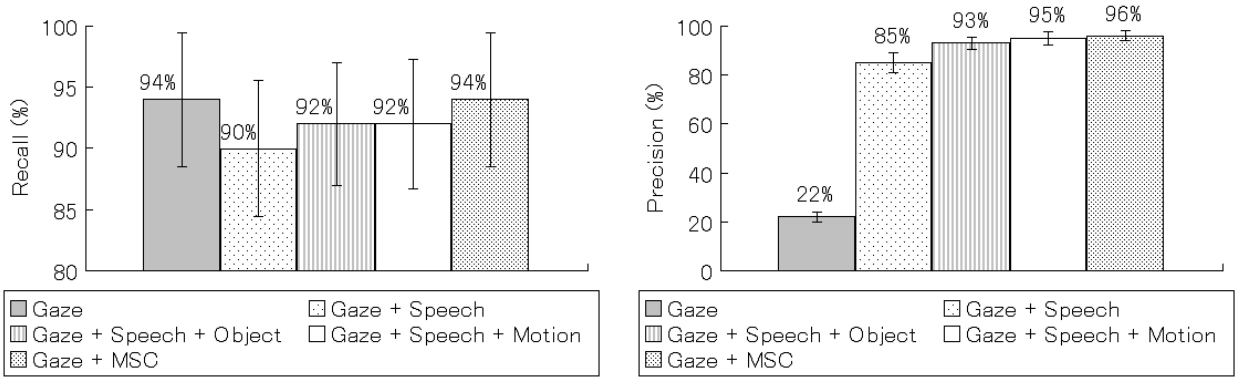
This paper described an RD speech detection method that enables a robot to distinguish the speech to which it should respond in an object manipulation task by combining speech, visual, and behavioral context with human gaze. The remarkable feature of the method is the introduction of the MSC measure. The MSC measure evaluates the feasibility of the action which the robot is going to execute according to the users' speech under the current physical situation. The experimental results clearly showed that the method is very effective and provides an essential function for natural and safe human-robot interaction. Finally, we would emphasize that the basic idea adopted in the method is applicable to a broad range of human-robot dialog tasks.

♢ References ♢

[Asoh 99] Asoh, H., Matsui, T., Fry, J., Asano, F., and Hayamizu, S.: A spoken dialog system for a mobile robot, in *Proc. Eurospeech*, pp. 1139–1142 (1999)

[Fujimoto 06] Fujimoto, M. and Nakamura, S.: Sequential Non-Stationary Noise Tracking Using Particle Filtering with Switching Dynamical System, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 769–772 (2006)

[Hosmer 09] Hosmer, D. W. and Lemeshow, S.: *Applied Logistic Regression*, Wiley-Interscience (2009)



(a) Recall rates (b) Precision rates

Fig. 14 Average recall and precision rates obtained in the on-line experiment.

[Ishi 06] Ishi, C. T., 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. IEEE-RAS Int. Conf. on Humanoid Robots*, pp. 340–345 (2006)

[Iwahashi 07] Iwahashi, N.: Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations, *Human-Robot Interaction*, pp. 95–118 (2007)

[Jiang 05] Jiang, H.: Confidence measures for speech recognition: A survey, *Speech Communication*, Vol. 45, pp. 455–470 (2005)

[Katagiri 98] Katagiri, S., Juangs, B. H., and Lee, C. H.: Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method, in *Proceedings of the IEEE*, Vol. 86, pp. 2345–2373 (1998)

[Kawahara 98] Kawahara, T., Ishizuka, K., Doshita, S., and Lee, C.-H.: Speaking-Style Dependent Lexicalized Filler Model for Key-Phrase Detection and Verification, in *Proc. IEEE Int. Conf. on Spoken Language Processing*, pp. 3253–3259 (1998)

[Kurita 92] Kurita, T.: Iterative weighted least squares algorithms for neural networks classifiers, in *Proc. Workshop on Algorithmic Learning Theory* (1992)

[Lang 03] Lang, S., Kleinehagenbrock, M., Hohenner, S., Fritsch, J., Fink, G. A., and Sagerer, G.: Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot, in *Proc. ACM Int. Conf. on Multimodal Interfaces*, pp. 28–35 (2003)

[Lee 04] Lee, A., Nakamura, K., Nishimura, R., Saruwatari, H., and Shikano, K.: Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs, in *Proc. Interspeech*, pp. 173–176 (2004)

[Mutlu 09] Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., and Hagita, N.: Footing in human-robot conversations: how robots might shape participant roles using gaze cues, in *Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction*, pp. 61–68 (2009)

[Nakamura 06] 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 Trans. on Audio, Speech, and Language Processing*, Vol. 14, No. 2, pp. 365–376 (2006)

[Takiguchi 08] Takiguchi, T., Sako, A., Yamagata, T., and Ariki, Y.: System Request Utterance Detection Based on Acoustic and Linguistic Features, *Speech Recognition, Technologies and Applications*, pp. 539–550 (2008)

[Tokuda 95] Tokuda, K., Kobayashi, T., and Imai, S.: Speech parameter generation from HMM using dynamic features, in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 660–663 (1995)

[Yonezawa 09] Yonezawa, T., Yamazoe, H., Utsumi, A., and Abe, S.: Evaluating Crossmodal Awareness of Daily-partner Robot to User's Behaviors with Gaze and Utterance Detection, in *Proc. ACM Int. Workshop on Context-Awareness for Self-Managing Systems*, pp. 1–8 (2009)

〔担当委員：稲邑 哲也〕

Received January 30, 2010.

Author's Profile

Xiang Zuo (Student Member)

is a Ph.D. student at graduate school of Engineering Design, Kyoto Institute of Technology. He received his B.E. degree in Mechanical Engineering, and M.S. degree in Information Science from University of Science and Technology Beijing and Kyoto Institute of Technology, respectively in 2002 and 2007. Since 2009, he has been an intern researcher at Advanced Telecommunications Research labs. His research interests include robot language acquisition and human-robot interaction. He is a member of the Robotics, Information Processing and Acoustical Society of Japan.

Naoto Iwahashi (Member)

Naoto Iwahashi received the B.E. degree in engineering from Keio University in 1985, Yokohama, Japan. He received Ph.D. degree in engineering from Tokyo Institute of Technology in 2001. In 1985, he joined Sony Corporation, Tokyo, Japan. From 1990 to 1993, he stayed at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. From 1998 to 2003, he was with Sony Computer Science Laboratories Inc., Tokyo, Japan. In 2003, he joined Advanced Telecommunication Research Laboratories International. In 2006, he Joined National Institute of Information and Communications Technology. His research areas include interaction speech system, language acquisition, human-robot interaction. He is a member of IEICE, SOFT, JCSS, and RSJ.

Kotaro Funakoshi (Member)

He is with Honda Research Institute Co., Ltd. since 2006. He received the B.S. degree in 2000 from Tokyo Institute of Technology, the M.S. and the Dr. Eng. degrees from Tokyo Institute of Technology in 2002 and 2005, respectively. His research interests are natural language understanding/generation, spoken dialogue systems and conversational robots. He is a member of AAAI, ACM SIGCHI, IPSJ, and NLP.

Mikio Nakano (Member)

Mikio Nakano, Sc.D. is a Principal Researcher at Honda Research Institute Japan Co., Ltd. (HRI-JP). He received his M.S. degree in Coordinated Sciences and Sc.D. degree in Information Science from the University of Tokyo, respectively in 1990 and 1998. From 1990 to 2004, he worked for Nippon Telegraph and Telephone Corporation. In 2004, he joined HRI-JP, where he engages in research on speech communication. He is a member of ACM, ACL, IEEE, ISCA, AAAI, IPSJ, RSJ, IEICE, and ANLP.

Ryo Taguchi (Member)

He received the B.E., M.E. and Ph.D. degrees in engineering from Toyohashi University of Technology, in 2002, 2004 and 2008, respectively. Currently, he is working at Nagoya Institute of Technology. He is a member of the Robotics Society of Japan, Japanese Cognitive Science Society, Information Processing Society of Japan, and the Acoustical Society of Japan.

Shigeki Matsuda

received his B.S. degree from the Department of Information Science, Teikyo University, in 1997, completed his doctoral program at the Japan Advanced Institute of Science and Technology in 2003. From 2003, he was a researcher at ATR Spoken Language Communication Laboratories, and in 2008 became a senior research scientist. He is an expert researcher at National Institute of Information and Communications Tehnology (NICT). He holds a doctoral degree in information science. He is engaged in research on speech recognition, and is a member of the institute of Electronics, Information and Communication Engineers, the Acoustic Society of Japan, and Information Processing Society of Japan.

Komei Sugiura (Member)

is an expert researcher at Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology (NICT). He received his B.E. degree in electrical and electronic engineering, M.S. and Ph.D. degrees in informatics from Kyoto University in 2002, 2004, and 2007, respectively. From 2006 to 2008, he was a research fellow, Japan Society for the Promotion of Science, and he has been with NICT since 2008. His research interests include robot language acquisition, spoken dialogue systems, machine learning, and sensor evolution. He is a member of the Society of Instrument and Control Engineers, and the Robotics Society of Japan.

Natsuki Oka (Member)

Dr. Oka received the B.E. degree from the University of Tokyo in 1979. After working at Shimadzu Corporation, the University of Tokyo, Institute for New Generation Computer Technology (ICOT), and Panasonic Corporation, he joined Kyoto Institute of Technology as a professor in 2003. His recent concern has been the understanding of and the modeling of cognitive development. He won FIT2007 Best Paper Award.