1J1-OS13-4

# Detection of Robot-Directed Speech by Situated Understanding in Physical Interaction

Xiang Zuo[*1*2] Naoto Iwahashi[*1*5] Taguchi Ryo[*1*4] Shigeki Matsuda[*5]
Komei Sugiura[*5] Kotaro Funakoshi[*3] Mikio Nakano[*3] Natsuki Oka[*2]

[*1]Advanced Telecommunication Research Labs [*2]Kyoto Institute of Technology
[*3]Honda Research Institute Japan Co., Ltd. [*4]Nagoya Institute of Technology
[*5]National Institute of Information and Communications Technology

In this paper, we propose a novel method for a robot to detect robot-directed speech from other speech: to distinguish speech that users speak to a robot from speech that users speak to other people or to himself. The originality of this work is the introduction of Multimodal Semantic Confidence measure, which is used for domain classification of input speech based on deciding 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 achieved a high performance of 94% and 96% in average recall and precision rates for robot-directed speech detection.

## 1. INTRODUCTION

Robots are now being designed to be a part of the everyday lives of 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. For such an interface, the capability to detect 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. For example, [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 unrelated to the robot even while they are looking at it.

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 [Takiguchi 08]. However, this kind of methods also raised the issue of requiring users to adjust their prosody to fit the system, which causes them an additional burden.

In this work, we introduce a multimodal semantic confidence (MSC) measure for domain classification. MSC has the key advantage that it is based on semantic features that determine whether the speech can be interpreted as a feasible action under the current physical situation.

The target task of this work is an object manipulation task in which a robot manipulates objects according to a user's speech. An example of such a task in a home environment is a user telling a robot to "Put the dish in the cupboard." Solving this task requires robots to deal with speech and image signals and to carry out a motion in accordance with the speech. Therefore, the MSC measure is calculated by integrating information obtained from speech, object images, and robot motion.

連絡先: Xiang Zuo, Kyoto Institute of Technology, d8821502@edu.kit.ac.jp

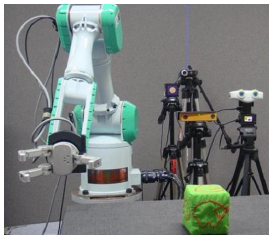


Figure 1: Robot used in the object manipulation task.

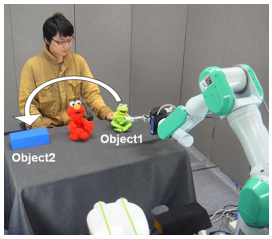



Figure 2: Example of object manipulation tasks.

## 2. Object Manipulation Task

In this work, we assume that humans use a robot to perform an object manipulation task. In this task, users command the robot (Fig. 1) by speech to manipulate objects on a table located between the robot and the user. In the example shown by Figure 2, the robot is told to place Object 1 (Kermit) on Object 2 (big box) by the command speech "Place-on Kermit, and the robot executes an action according to this speech. The solid line in Fig. 2 shows the trajectory of the moving object manipulated by the robot.

The 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 Fig. 2, the phrases for the motion, trajector, and landmark are "Place-on," "Kermit," and "big box," respectively. To execute a correct action according to such a command, 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 or noise. Consequently, an RD speech detection method should be used.

## 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 Fig. 3. First, a Gaussian mixture model based

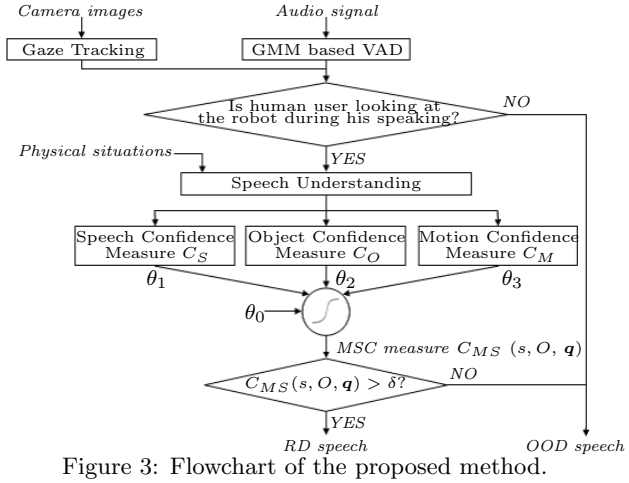



Figure 3: Flowchart of the proposed method.

voice activity detection method 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. If the proportion of the user's gaze at the robot during her/his speech is higher than a certain threshold $\eta$, 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. 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. 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 $\delta$, 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.

## 3.1 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 $O = \{(o_{1,f}, o_{1,p}), (o_{2,f}, o_{2,p}) \dots (o_{m,f}, o_{m,p})\}$, 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 $\xi$ denote the index of the 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.

### 3.1.1 Lexicon and Grammar

The robot initially had 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 HMMs. The concepts of object images and motions are represented by Gaussian functions and HMMs, respectively.

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

### 3.1.2 Belief Modules and Integrated Belief

Each of the five belief modules in the integrated belief is defined as follows. First, the belief module of speech, $\boldsymbol{B_S}$, is represented as the log probability of $s$ conditioned by $z$, under lexicon $L$ and grammar $G_r$. The belief module of object image, $\boldsymbol{B_O}$, 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 $o_{l,f}$. The belief module of motion, $\boldsymbol{B_M}$, is represented as the log likelihood of $\boldsymbol{w_M}$ given trajectory $\xi$. The belief module of motion-object relationship, $\boldsymbol{B_R}$, 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, under a parameter set $R$. The belief module of behavioral context, $\boldsymbol{B_H}$, represents the belief that the current speech refers to object $o$, given behavioral context $\boldsymbol{q}$, with a parameter set $H$.

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 $\Psi$, written as

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

where $l$ denotes the index of landmark, $o_t$ and $o_l$ denote the trajector and landmark, respectively, and $o_{t,p}$ and $o_{l,p}$ denote the positions of $o_t$ and $o_l$, respectively. Then, as the meaning of speech $s$, corresponding action $\hat{a}$ is determined by maximizing $\Psi$:

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

Finally, $\hat{a} = (\hat{t}, \hat{\xi})$, $\hat{l}$, and $\hat{z}$ are outputted from the speech understanding process.

## 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

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

Given a threshold $\delta$, speech $s$ with an MSC measure higher than $\delta$ 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.

#### 3.2.1 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

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

where $n(s)$ denotes the analysis frame length of the input speech, $P(s|\hat{z}; L)$ 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, $A$ denotes the phoneme acoustic model used in $\boldsymbol{B_S}$, and $L(G_p)$ denotes a set of possible phoneme sequences accepted by Japanese 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.

#### 3.2.2 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 used as these baselines. Then, the object confidence measure $C_O$ is written as

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

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 $o_{\hat{l},f}$ and are given by $\boldsymbol{B_O}$; furthermore, $\max_{o_f} P(o_f|\ \hat{\boldsymbol{w}}_{\boldsymbol{T}})$ 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.

#### 3.2.3 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

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

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. 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

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

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

#### 3.2.4 Optimization of Weights

We now consider the problem of estimating weight $\boldsymbol{\Theta}$ in Eq. (3). The $i$th 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 is denoted as $\mathbb{T}^N = \{(s^i, O^i, \boldsymbol{q}^i, d^i)|i = 1, ..., N\}$, where $d^i$ is 0 or 1, which represents OOD speech or RD speech, respectively. The likelihood function is written as

$$P(\boldsymbol{d}|\boldsymbol{\Theta}) = \prod_{i=1}^{N} (C_{MS}(s^i, O^i, \boldsymbol{q}^i))^{d^i} (1 - C_{MS}(s^i, O^i, \boldsymbol{q}^i))^{1-d^i}, \quad (8)$$

where $\boldsymbol{d} = (d^1, ..., d^N)$. $\boldsymbol{\Theta}$ is optimized by the maximum-likelihood estimation of Eq. (8).

## 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 was not used and speech was extracted manually without using the GMM based VAD in order to avoid its detection errors. The weighting set $\boldsymbol{\Theta}$ and the threshold $\delta$ were also optimized in this experiment. Then we performed an on-line experiment with the robot to evaluate the entire system.

The robot lexicon $L$ used in both experiments has 50 words, including 31 nouns and adjectives representing 40 objects and 19 verbs representing 10 kinds of motions. $L$ also includes five Japanese postpositions. Different from other words in $L$, none of the postpositions is associated with a concept. By using the postpositions, users can speak a command in a more natural way. The parameter set $\boldsymbol{\Gamma}$ in Eq. (1) was $\gamma_1 = 1.00$, $\gamma_2 = 0.75$, $\gamma_3 = 1.03$, $\gamma_4 = 0.56$, and $\gamma_5 = 1.88$.

### 4.2 Off-line Experiment by Simulation

#### 4.2.1 Setting

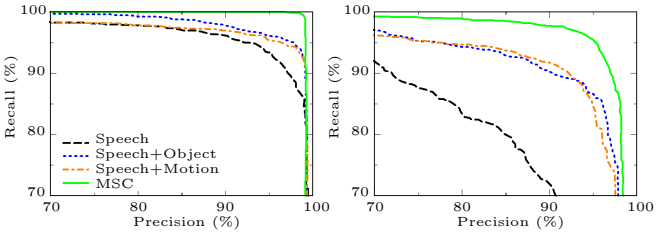
The off-line experiment was conducted under both clean and noisy conditions using a set of pairs of speech $s$ and a scene file, which included $O$ and $\boldsymbol{q}$. We prepared 160 different speech-scene pairs. The speech was recorded under both clean and noisy conditions as follows.

**Clean condition:** We recorded the speech in a soundproof room without noise. A subject sat on a chair one meter from a SANKEN CS-3e directional microphone and read out a text in Japanese.

**Noisy condition:** We added dining hall noise, having a level from 50 to 52 dBA, to each speech record gathered under a clean condition.

We gathered the speech records from 16 subjects, including 8 males and 8 females. 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 manually labeled as either RD or OOD and then inputted into the system. For each pair, speech understanding was first performed, and then the MSC measure was calculated. During the speech understanding experiment, a Gaussian mixture model based noise suppression method 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 1 subject's data were used as a test set and repeated 16 times. The values of the weighting $\hat{\boldsymbol{\Theta}}$




(a) Clean condition (b) Nosy condition

Figure 4: Average precision-recall curves under clean and noisy conditions.

learned by using 16 subjects' data 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. 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.

#### 4.2.2 Results

The average precision-recall curves over 16 subjects under clean and noisy conditions are shown in Fig. 4. The performances of each of four cases are shown in the figure. From the figure, 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 conditions are MSC: 99%, Speech+Object: 97%, Speech+Motion: 97%, Speech: 94%; those for noisy condition are MSC: 95%, Speech+Object: 92%, Speech+Motion: 93%, and Speech: 83%. By comparison 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 and found that there were statistical differences between Speech and all other cases under both conditions.

The values for $\hat{\mathbf{\Theta}}$ optimized under the clean condition were: $\hat{\theta}_0 = 5.9$, $\hat{\theta}_1 = 0.00011$, $\hat{\theta}_2 = 0.053$, and $\hat{\theta}_3 = 0.74$. The threshold $\delta$ of domain classification was set to $\hat{\delta} = 0.79$, which maximized the F-measure of MSC under the clean condition. The $\hat{\mathbf{\Theta}}$ and $\hat{\delta}$ were used in the on-line experiment.

### 4.3 On-line Experiment Using the Robot

#### 4.3.1 Setting

In the on-line experiment, the entire 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. The threshold $\eta$ of gaze tracking was set to 0.5.

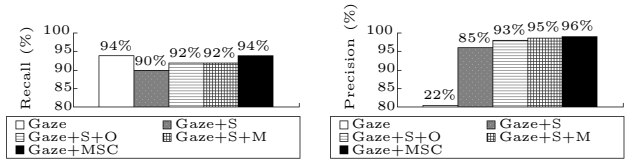
We conducted a total of four sessions of this experiment using four pairs of subjects, and each session lasted for about 50 minutes. All subjects were adult males. There was constant ambient noise of about 48 dBA from the robot's power module in all sessions. For comparison, five cases were evaluated 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.

#### 4.3.2 Results

During the experiment, a total of 983 pieces of speech were made, each of which was manually labeled. The numbers of them are shown in 1. There were 708 pieces of speech which were made while the operator was looking at the robot, including 115 and 553 pieces of RD and OOD speech, respectively. This means that in addition to the RD speech, there was also a lot of OOD speech made while the subjects were looking at the robot.

Table 1: 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 |




(a) Recall (b) Precision

Figure 5: Average recall and precision rates obtained in the on-line experiment.

The average recall and precision rates for each of the above five cases are shown in Fig. 5. By using gaze only, an average recall rate of 94% was obtained, 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 integrating gaze with the speech confidence measure, which means that some RD speech was rejected erroneously by the speech confidence measure. However, by integrating gaze with MSC, the recall rate returned to 94% because the mistakenly rejected RD speech was correctly detected by MSC. In (b), the average precision rate by using gaze only was 22%. However, by using MSC, the instances of OOD speech were correctly rejected, resulting in a high precision rate of 96%, which means the proposed method is particularly effective in situations where users make a lot of OOD speech while looking at a robot.

## 5. Conclusion

This paper described a robot-directed (RD) speech detection method that enables a robot to distinguish the speech to which it should respond in an object manipulation task The novel feature of this 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 show that the method is very effective and provides an essential function for natural and safe human-robot interaction.

## References

[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)

[Nakamura 06] Nakamura, S., 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)

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

[Yonezawa 09] Yonezawa, T., et al.: Evaluating Crossmodal Awareness of Daily-partner Robot to User's Behaviors with Gaze and Utterance Detection, in *Proc. CASMS*, pp. 1–8 (2009)