Robot-Command Utterance Detection in Object Manipulation Task Using Multimodal Semantic Confidence Based on Speech, Image, and Motion

*Xiang Zuo1,4, Naoto Iwahashi1,3, Ryo Taguchi1,5, Shigeki Matsuda3, Komei Sugiura3, Kotaro Funakoshi2, Mikio Nakano2, and Natsuki Oka4

1Advanced Telecommunication Research Labs, 2Honda Research Institute Japan, 3National Institute of Information and Communication Technology, 4Kyoto Institute of Technology, and 5Nagoya Institute of Technology

sasyou@atr.jp

Abstract—In this paper, we propose a method to detect commands directed to robots in an object manipulation task. Multimodal Semantic Confidence in speech is estimated in terms of speech, the static image of objects in the current scenes, and the trajectory of motion. Experimental results show the effectiveness of the method under conditions of natural human-robot interaction.

Key Words: robot-command utterance detection, multimodal semantic confidence, human-robot interaction

1. Introduction

For a speech-based human-robot interface, detecting robot-command (RC) utterances is important. In recent researches, many systems based on eye gaze detection or body orientation detection have been implemented [1][2]. However, people may say something irrelevant to the robot while looking at it or may order the robot to do something without fixing their eye gaze on it. Consider the following conversation where users A and B are talking while looking at the robot in front of them (Fig.1).

A: Cool robot! What can it do?
B: It can understand your command, like “big Kermit box move-onto.”

However, the utterances here are not RC utterances. Moreover, even if B makes an utterance that resembles a command (“big Kermit box move-onto”), 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? Eye gaze detection or body orientation detection based systems are ineffective here. To address this kind of problem, we propose a novel method to detect RC utterances. Different from previous researches, in our method, MSC, which leads to RC utterance detection in an object manipulation task performed by a robot, is calculated by using both speech inputs and physical situations.

2. Object Manipulation Task

Fig.2 shows the hardware platform used in the object manipulation task. Fig.3 depicts a camera image in which the robot is told to place object 1 (Kermit1) on object 3 (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 (tra-
jector) and the reference object (landmark). The reference object can be the trajector itself or a landmark characterizing the trajectory of the trajector. In the case shown in Fig.3, the trajector and landmark are objects 1 and 3, respectively.

![Fig.2 Hardware platform used in object manipulation task](image1)

![Fig.3 Corresponding scene to “big Kermit box move-onto”](image2)

3. **MSC-Based Utterance Detection Method**

An overview of our method is shown in Fig.4. First, using the information on current scene $O$ and behavioral context $q$, utterance understanding is performed to interpret the meaning of utterance $s$ as a possible action that should be accepted. The output of the utterance understanding includes a possible action and the conceptual structure of the utterance and the action. Second, using the output from the utterance understanding, three confidence measures are calculated: those for speech, the static images of the objects, and the trajectory of motion. Then, the MSC measure is defined as the weighted sum of these confidence measures with weight optimization. Finally, the MSC value is input to a logistic sigmoid function, and RC utterance probability $P(class = RC|s, O, q)$ is obtained as an output.

3-1 **Utterance 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 [4]. In this study, we employ the utterance understanding method used in LCore. The method selects the optimal action based on a multimodal integrated user model trained by the interaction between the user and the robot when a user’s utterance is input. A user model corresponding to each user model that integrates the five belief modules – (1) speech, (2) motion, (3) vision, (4) motion-object relationship, and (5) behavioral context – is called shared belief $\Psi$.

Given utterance $s$, current scene $O$, which includes the visual features and positions of all objects in it, and behavioral context $q$, possible action $a = (t, \xi)$ should be accepted under $O$, where $t$ and $\xi$ denote a trajector and a trajectory of motion, respectively. In the process of the utterance understanding, we assume that RC utterance $s$ can be interpreted with conceptual structure $z$ = [(Trajector: $W_T$), (Landmark: $W_L$), (Motion: $W_M$)], where $W_T$, $W_L$, and $W_M$ represents the phrases describing a trajector, a landmark, and motion, respectively. (Or $z = [(\text{Trajector}: W_T), (\text{Motion}: W_M)]$ 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, user’s utterance, “big Kermit box move-onto,” is interpreted as follows: [(Trajector: “big Kermit”), (Landmark: “box”), (Motion: “move-onto”)]. Each of the five belief modules is defined as follows:

**Speech $B_s$:** 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 the concepts of the static images of the objects and the motions and $G_r$ represents the probabilistic language model for possible robot commands.

**Static Image of Object $B_l$:** This module, which is represented as the log likelihood of Gaussian distributions, is written as $\log P(o_{i,f}|W_T; L)$ and $\log P(o_{i,l}|W_T; L)$, where $o_{i,f}$ and $o_{i,l}$ denote the feature of trajector $o_i$ and landmark $o_l$ in scene $O$.

**Motion $B_M$:** This module is represented as the log likelihood of a probabilistic model given trajectory $\xi$ referred to by motion word $W_M$. It is written as $P(\xi|o_{i,f}, o_{i,l}, W_M; L)$, where $o_{i,p}$ and $o_{i,p}$ denote the positions of trajector $t$ and landmark $l$, respectively.

**Motion-object relationship $B_R$:** This module represents the belief that in the motion corresponding to motion word $W_M$, feature $o_{i,f}$ of trajector $t$ and feature $o_{i,l}$ of landmark $l$ are typical. This belief is represented by a conditional multivariate Gaussian probability density function, $P(o_{i,f}, o_{i,l}|W_M; R)$, where $R$ is its parameter set.

**Behavioral context $B_B$:** This module represents the belief that the current utterance refers to object $o$, given behavioral context $q$. It is written as $B_B(o, q; H)$, where $q$ includes information on which objects were a trajector and a landmark in the previous action and which object the user is holding; $H$ is its parameter set.

Given parameter set $\Gamma = \{\gamma_1, ..., \gamma_8\}$, the degree of correspondence between utterance $s$ and action $a$ is determined as Eq.1:

$$\Psi(s,a, O, q, L, G_r, R, H, \Gamma) =$$

$$\max_{\xi} \left\{ \gamma_1 \log P(s|z; L)P(z; G_r) \right\}$$

$$\left\{ \begin{array}{l}
\gamma_2 \left( \log P(o_{i,f}|W_T; L) + \log P(o_{i,l}|W_T; L) \right) \\
\gamma_3 \log P(\xi|o_{i,p}, o_{i,p}, W_M; L) \\
\gamma_4 \log P(o_{i,f}, o_{i,l}|W_M; R) \\
\gamma_5 B_B(t, q; H + B_B(l, q; H)) \right\}$$

where conceptual structure $z$ and landmark $l$ are selected to maximize the value of $\Psi$. As the meaning of utterance $s$ under scene $O$, corresponding action $a$, which is represented by trajector $t$ and the trajectory of motion $\xi$, is
determined by maximizing $\Psi$:
\[
\hat{a} = (\hat{i}, \hat{\xi}, \hat{l}) = \arg\max_a \Psi(s, a, O, q, L, G_r, R, H, \Gamma).
\] (2)

Finally, action $\hat{a} = (\hat{i}, \hat{\xi}, \hat{l})$, selected landmark $\hat{l}$, and conceptual structure $\hat{z}$ are outputted from the utterance understanding process.

3.2 MSC Measure

Next, we describe the proposed MSC measure. MSC measure $C_{MS}$ is a measure of the reliability for action $\hat{a}$ under the current scene and leads to a RC utterance decision. For input utterance $s$ and current scene $O$, $C_{MS}$ is calculated based on the output of utterance understanding $(\hat{a}, \hat{l}, \hat{z})$ and is written as
\[
C_{MS}(s, O, \hat{a}, \hat{l}, \hat{z}) = \theta_1 C_S + \theta_2 C_l + \theta_3 C_M,
\] (3)
where $C_S$, $C_l$, and $C_M$ are the confidence measures of the speech, the object images, and the trajectory of motion. $\Theta = [\theta_1, \theta_2, \theta_3]$ is applied to these confidence scores, and the weighted sum is defined as the MSC measure for the RC utterance decision.

3.2.1 Speech Confidence Measure

The confidence measure of speech $C_S$ is calculated as
\[
C_S(s, \hat{z}; A, G_p) = \frac{1}{n(s)} \log \frac{P(s|\hat{z}; A)}{\max_{y \in L(G_p)} P(s|y; A)},
\] (4)
where $n(s)$ denotes the analysis frame length of the input speech, denominator $P(s|\hat{z}; A)$ denotes the likelihood of word sequence $\hat{z}$ for input utterance $s$ by a phoneme acoustic model $A$, and numerator $\max_{y \in L(G_p)} P(s|y; A)$ denotes the likelihood of maximum possible phoneme sequence $y$ in $L(G_p)$ that means a set of possible phoneme sequences accepted by phoneme network $G_p$. For utterances which match robot command grammar $G_r$, $C_S$ has a greater value than utterances that do not match the robot command grammar.

3.2.2 Image Confidence Measure

The model for each image category is represented by a Gaussian function in a multi-dimensional visual feature space. For features $(\alpha_{i,f}$ and $\alpha_{i,l}$) of $\hat{i}$ and $\hat{l}$, the confidence measure of image is calculated as
\[
C_I(\alpha_{i,f}, \alpha_{i,l}, \hat{W}_r, \hat{\tilde{W}}_l; L) = \log \frac{P(\alpha_{i,f}|\hat{W}_r; L)P(\alpha_{i,l}|\hat{W}_l; L)}{\max_{\alpha_{i,f}, \hat{W}_r} P(\alpha_{i,l}|\hat{W}_r) \max_{\alpha_{i,l}, \hat{W}_l} P(\alpha_{i,l}|\hat{W}_l)},
\] (5)
where $P(\alpha_{i,f}|\hat{W}_r; L)$ and $P(\alpha_{i,l}|\hat{W}_l; L)$ denote the likelihood of $\alpha_{i,f}$ and $\alpha_{i,l}$ and $\max_{\alpha_{i,f}, \hat{W}_r} P(\alpha_{i,l}|\hat{W}_r)$ and $\max_{\alpha_{i,l}, \hat{W}_l} P(\alpha_{i,l}|\hat{W}_l)$ denote the maximum-likelihood for object image models. By this maximization, the most typical visual features are obtained for $\hat{W}_r$ and $\hat{W}_l$.

3.2.3 Motion Confidence Measure

For selected trajectory $\hat{\xi}$, the confidence measure of motion is calculated as
\[
C_M(\hat{\xi}, \hat{W}_M; L) = \log \frac{P(\hat{\xi}|\alpha_{i,p}, \alpha_{l,p}; \hat{W}_M; L)}{\max_{\alpha_{i,p}} P(\hat{\xi}|\alpha_{l,p}; \hat{W}_M; L)},
\] (6)
where $P(\hat{\xi}|\alpha_{i,p}, \alpha_{l,p}; \hat{W}_M; L)$ denotes the likelihood for trajectory $\hat{\xi}$ referred to by motion word $\hat{W}_M$, given positions $\alpha_{i,p}$ and $\alpha_{l,p}$ of $\hat{i}$ and $\hat{l}$, and $\max_{\alpha_{i,p}} P(\hat{\xi}|\alpha_{l,p}; \hat{W}_M; L)$ denotes the maximum-likelihood trajectory $\hat{\xi}$ of motion word $\hat{W}_M$, given landmark position $\alpha_{l,p}$ while trajectory position $\alpha_{i,p}$ is variables. By this maximization, the most typical trajectory for $\hat{W}_M$ is obtained.

3.2.4 Optimization of Weights

We now consider the problem of estimating weight $\Theta$ of $C_{MS}$ in Eq.3. The $i$th training sample is given as the pair of $C_i$ and teaching signal $d^i$, $\{(C_i, d^i)\}_{i=1}^N$, where $d^i$ is 0 or 1 and $N$ is the total number of the training samples. Logistic regression model [5] is used for optimizing $\Theta$ and results in a RC utterance decision probability. It is written as
\[
P(\text{class} = \text{RC}|s, O, q) = \sigma(\theta_0 + C_{MS}(s, O, \hat{a}, \hat{l}, \hat{z})),
\] (7)
where $\sigma$ denotes the logistic sigmoid function and $\theta_0$ denotes an assistant parameter in $\sigma$. Then $\theta_0, \theta_1, \theta_2, \theta_3$ are optimized by maximum-likelihood estimation using Fisher’s scoring algorithm.

4. Experimental Evaluation

4.1 Conditions

Our experiment was conducted using a set of pairs of utterance and scenery with objects. Each utterance was manually labeled as either RC or OOD. We gathered a corpus of 1280 utterances (640 RC and 640 OOD) from eight participants (four males and four females) in a soundproof room with a SANKEN-CS5 directional microphone without noise. Each participant sat on a bench one meter from the microphone and made utterances in Japaneses2. To evaluate under noisy conditions, we mixed these utterances with dining hall noise at a level from 50 to 52 dBA.

2In this paper, the utterances were translated into English.
For both clean and noisy speech, all utterances were evaluated by the repeated 8-fold cross validation: utterances gathered from seven participants (1120 utterances) were used as a training set, and the remaining 160 utterances were used as a test set and repeated eight times. Weights \( \theta_0, \theta_1, \theta_2, \theta_3 \) in Eq.7 were optimized during the cross validation. Their averages were: \( \theta_0 = 5.7, \theta_1 = 0.000088, \theta_2 = 0.060, \) and \( \theta_3 = 0.98. \)

Speech was represented by mel-scale cepstrum coefficients and their delta parameters (25 dimensional). For each speech file, noise suppression and speech detection were performed in the first step, and then speech recognition was performed with the ATRASR recognition engine [6]. Static object features captured by the camera device were represented by size, color \((L^*, a^*, b^*)\), and shape \((\text{width/height, squareness})\). Trajectory was represented by a sequence of vertical and horizontal coordinates and 2-dimensional velocity.

To provide a baseline for comparison to the MSC measure, a speech measure conventional using \( A \) and \( G_p \) without non-speech information was implemented and tested with the same data set. Different from the MSC measure, the speech measure does not have the utterance understanding component and performs RC utterance detection only using the speech confidence measure, which is calculated as

\[
C_s(s, W; A, G_p) = \frac{1}{n(s)} \log \frac{P(s|W; A)}{\max_{y \in LG_p} P(s|y; A)}, \tag{8}
\]

where \( W \) denotes the maximum likelihood word sequence for \( s \) using \( G_r \) from ATRASR.

4.2 Result

The phoneme recognition accuracies were 83% for clean speech and 67% for noisy speech. Fig.5 and Fig.6 show the scatter diagrams of precision-recall calculated by cross validation for clean and noisy speech. The MSC measure and baseline performances are shown by “MSC” and “Baseline,” respectively. The two lines clearly show that the MSC measure outperforms the baseline for RC utterance detection both under clean speech and noisy conditions. Moreover, the performances using the partial MSC measure are shown by “Speech-Image” (using the confidence measure of speech and image) and “Speech-Motion” (using the confidence measure of speech and motion). These lines show that both image and motion confidences contributed to performance improvement. The maximum f-measures of MSC and baseline were 99% and 94% under clean speech, respectively, and 97% and 88% under noisy conditions, respectively. MSC achieved an absolute growth of five points for clean speech and nine points for noisy speech compared to the baseline with max f-measure.

5. Conclusion


