ROBOT-DIRECTED SPEECH DETECTION USING MULTIMODAL SEMANTIC CONFIDENCE BASED ON SPEECH, IMAGE, AND MOTION

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

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

ABSTRACT

In this paper, we propose a novel method to detect robot-directed (RD) speech that adopts the Multimodal Semantic Confidence (MSC) measure. The MSC measure is used to decide whether the speech can be interpreted as a feasible action under the current physical situation in an object manipulation task. This measure is calculated by integrating speech, image, and motion confidence measures with weightings that are optimized by logistic regression. Experimental results show that, compared with a baseline method that uses speech confidence only, MSC achieved an absolute increase of 5% for clean speech and 12% for noisy speech in terms of average maximum F-measure.

Index Terms— robot-directed speech detection, multimodal semantic confidence, human-robot interaction

1. INTRODUCTION

Robots are now being designed to be a part of the lives of ordinary people in social and home environments. One of the key issues for practical use is the development of user-friendly interfaces. Speech recognition is one of our most effective communication tools for use in a human-robot interface. In recent studies, many systems using speech-based human-robot interfaces have been implemented, such as [1]. For a speech-based interface, the functional capability of detecting robot-directed (RD) speech is crucial. For example, user's utterances directed to another human listeners should not be recognized as commands directed to a robot.

To resolve this issue, many methods have been implemented, mainly based on two approaches: (1) using the characteristics of the acoustic features of speech, and (2) using human physical behaviors such as gaze tracking or body-orientation detection.

As examples of the first approach, methods based on acoustic features have been proposed for RD speech detection in [2], and for computer-directed speech detection in [3]. In these works, robot/computer-directed speech detection is performed based on analyzing the differences in acoustic features between robot/computer-directed speech and other speech. However, this kind of method requires humans to adjust their speaking style or accent to fit the robot/computer, which causes an additional burden to human users.

On the other hand, methods based on detecting human physical behaviors have been proposed. In [4], RD speech are detected by the proportion of the user's gaze at the robot during her/his speech. In [5], RD speech are detected by a multimodal attention system, which detects the direction of a person's attention based on a method for multimodal person tracking that uses face recognition, sound source localization, and leg detection. However, this kind of method raises two issues: (1) humans must adjust their behaviors to fit the robot while trying to give an order, which causes an additional burden to humans, and (2) humans may say something irrelevant to the robot while their behaviors are fitting it.

In contrast, the goal of this work is to implement a no-burden method. We defined the RD speech detection problem as a domain classification problem between (1) the RD domain of RD speech and (2) out-of-domains (OOD) of other speech. Different from recent works, our method does not require humans to adjust their behaviors to fit the robot; rather, it is based on deciding whether the speech can be interpreted as a feasible action under the current physical situation for a robot in an object-manipulation task by calculating the Multimodal Semantic Confidence (MSC) measure.

Conventional studies on domain classification have typically focused on using speech recognition confidences or topic classification [7]. However, for a domain classification problem to be solved by a robot, we believe that in addition to speech signals, non-speech information would also be helpful because robots communicate in the real world not only with hearing but also with sight, touch, and so on. Therefore, in our method, the MSC measure is calculated using both speech inputs and physical situations. The features of this work can be summarized as follows:

(1) The RD speech detection problem is defined as a domain classification problem.

(2) Domain classification is based on the MSC measure, which is calculated by using not only speech inputs but also physical situations.

The remainder of this paper is organized as follows: Section 2 gives the details of the object manipulation task. Section 3 describes the proposed method. The experimental methodology and results are presented in Section 4. Finally, Section 5 gives our conclusions.

2. OBJECT MANIPULATION TASK

The target task of this work is called an object manipulation task in which the robot shown in Fig. 1 manipulates objects according to a user's utterances under current physical situation. Fig. 2 depicts a camera image of the current physical situation under the command utterance "Place big Kermit on the box." In this example, the robot is told to place object 3 (big Kermit) on object 2 (box). The solid line shows the trajectory intended by the user. The trajectory can be interpreted by the positional change of the relationship between the

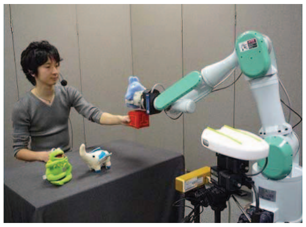


Fig. 1. Robot used in the object manipulation task.

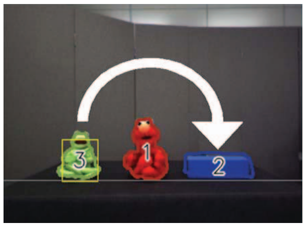


Fig. 2. Scene corresponding to "Place big Kermit on the box."

moved object (trajector) and the reference object (landmark). In the case shown in Fig.2, the trajector and landmark are objects 3 and 2, respectively.

3. MSC-BASED RD SPEECH DETECTION METHOD

An overview of our method is shown in Fig. 3. First, using the information on current scene O and behavioral context $\boldsymbol{q}$, speech understanding is performed to interpret the meaning of speech s as a possible action. Second, to evaluate the feasibility of the action, three confidence measures are calculated: C_S for speech, C_I for the static images of the objects, and C_M for the trajectory of motion. Then the weighted sum of these confidence measures with a bias is inputted to a sigmoid function. The bias and the weightings, $\{\theta_0, \theta_1, \theta_2, \theta_3\}$, are optimized by logistic regression. Here, the MSC is defined as the output of the sigmoid function, and represents the probability that s is RD speech.

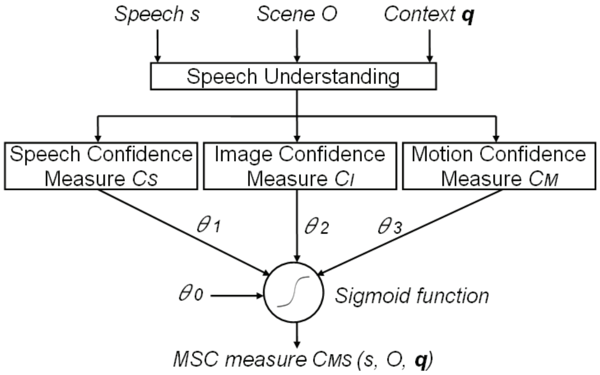
3.1. Speech Understanding

We previously proposed a machine learning method called LCore that enables robots to acquire the capability of linguistic communication from scratch through verbal and nonverbal interaction with users [6]. In this study, we employ the speech understanding method used in LCore.

In the process of the speech understanding, we assume that s can be interpreted with conceptual structure z = [(Motion: $\boldsymbol{w}_{\boldsymbol{W}}$), (Trajector: $\boldsymbol{w}_{\boldsymbol{T}}$), (Landmark: $\boldsymbol{w}_{\boldsymbol{L}}$)], where $\boldsymbol{w}_{\boldsymbol{M}}$, $\boldsymbol{w}_{\boldsymbol{T}}$, and $\boldsymbol{w}_{\boldsymbol{L}}$ represent the phrases describing motion, a trajector, and a landmark, respectively. (Or z = [(Motion: $\boldsymbol{w}_{\boldsymbol{M}}$), (Trajector: $\boldsymbol{w}_{\boldsymbol{T}}$)] for an action that does not need a landmark). The order of the components in z represents the word sequence of s. For example, in Fig. 2, the user's utterance, "Place big Kermit on the box" is interpreted as [(Motion: "Place"), (Trajector: "big Kermit"), (Landmark: "box")].

Given speech s, current scene O, which includes the visual features and positions of all objects in it, and behavioral context $\boldsymbol{q}$, speech understanding selects the optimal action a based on the conceptual structure z by a multimodal integrated user model that is trained by the interaction between the user and the robot. In this paper, a is defined as $a = (t, \xi)$, where t and ξ denote a trajector and a trajectory of motion, respectively. A user model integrates the five belief modules – (1) speech, (2) motion, (3) vision, (4) motion-object relationship, and (5) behavioral context – and is called shared belief. Each of the five belief modules in the shared belief is defined as follows:

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



Speech s
Scene O
Context q
Speech Understanding
Speech Confidence Measure Cs
Image Confidence Measure CI
Motion Confidence Measure CM
θ1
θ2
θ3
θ0
Sigmoid function
MSC measure CMS (s, O, q)

Fig. 3. MSC measure calculated by the outputs of speech understanding.

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

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

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

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

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

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

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

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

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

3.2. MSC Measure

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

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

where C_S, C_I, and C_M are the confidence measures of the speech, the object images, and the trajectory of motion. $\boldsymbol{\Theta}= \{\theta_0, \theta_1, \theta_2, \theta_3\}$ is applied to these confidence scores.

3.2.1. *Speech Confidence Measure*

The confidence measure of speech C_S is calculated by using the likelihood of an acoustic model, which is conventionally used as a confidence measure for speech recognition [8]. It is calculated as

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

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

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

3.2.2. *Image Confidence Measure*

As a baseline of the image confidence measure, the likelihood of the most typical visual features for selected objects are those that maximize Gaussians of the objects. For visual features ($o_{\hat{t},f}$ and $o_{\hat{l},f}$) of $\hat{t}$ and $\hat{l}$, which are represented by $\boldsymbol{w_T}$ and $\boldsymbol{w_L}$, respectively, the image confidence measure is calculated by the summed log-likelihood ratios of likelihood and baseline. It is written as

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

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

3.2.3. *Motion Confidence Measure*

As a baseline of the motion confidence measure, the likelihood of the most typical trajectory for motion word $\hat{\boldsymbol{w}}_{\boldsymbol{M}}$, given positions $o_{\hat{t},p}$ and $o_{\hat{l},p}$ of trajector $\hat{t}$ and landmark $\hat{l}$, can be obtained by treating the trajector position as a variable. Then the motion confidence measure is calculated as

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

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

3.2.4. *Optimization of Weightings*

We now consider the problem of estimating weighting $\boldsymbol{\Theta}$ of C_{MS} in Eq. 3. The ith training sample is given as the pair of $C^i_{MS} = C_{MS}(s^i, O^i, \boldsymbol{q}^i)$ and teaching signal d^i, $\{(C^i_{MS}, d^i)|i = 1, ..., N\}$, where d^i is 0 or 1, which represents OOD speech or RD speech, respectively, and N is the total number of training samples. A logistic regression model [9] is used for optimizing $\boldsymbol{\Theta}$. The likelihood function is written as

$$P(\boldsymbol{d}|\boldsymbol{\Theta}) = \prod_{i=1}^{N} (C^i_{MS})^{d^i} (1 - C^i_{MS})^{1-d^i}, \tag{7}$$

where $\boldsymbol{d}= (d^1, ..., d^N)$. $\boldsymbol{\Theta}$ is optimized by the maximum-likelihood estimation of Eq. 7 using Fisher's scoring algorithm [10].

4. EXPERIMENTAL EVALUATION

4.1. Data Collection and Experiment Setting

Our experiment was conducted under both clean and noisy conditions by using a set of pairs of speech and scene. We prepared a clean speech corpus by taking the following steps. First, we gathered 2560 speech samples from 16 participants (8 males and 8 females) in a soundproof room with a SANKEN-CS5 directional microphone without noise. All of these participants were native Japanese speakers, and each of them sat on a bench one meter from the microphone and produced speech in Japaneses[1]. Then we paired each speech with a scene, which was captured by the stereo vision camera. Figure 2 shows an example shot of a scene file. Each scene included three objects in average. Finally, each pair was manually labeled as either RD or OOD. For the noisy speech corpus, we mixed each speech sample in the clean speech corpus with dining hall noise at a level from 50 to 52 dBA and then performed noise suppression [11].

The evaluation under the clean speech corpus was performed by leave-one-out cross-validation: 15 participants' data was used as a training set, and the remaining 1 participant's data was used as a test set and repeated 16 times. During cross-validation, $\boldsymbol{\Theta}$ was optimized, and the averages were: $\hat{\theta}_0 = 5.9$, $\hat{\theta}_1 = 0.00011$, $\hat{\theta}_2 = 0.053$, and $\hat{\theta}_3 = 0.74$. Then, the evaluation under the noisy speech corpus was performed using these averages without cross-validation.

The robot lexicon L used in our experiment included 56 words, including 38 nouns and adjectives, 11 verbs representing 7 motions, and 7 particles. For each speech-scene pair, speech understanding

[1] In this paper, the speech was translated into English.

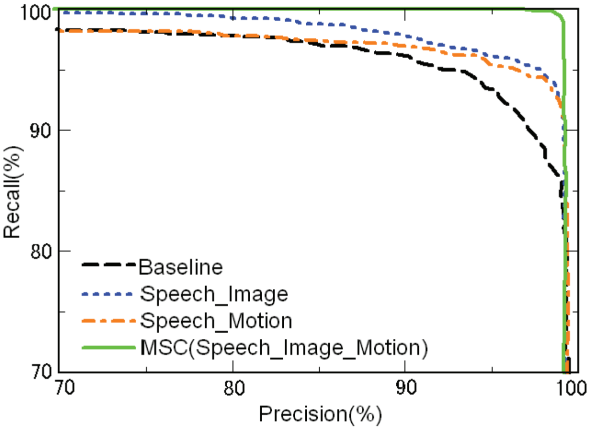


Fig. 4. Precision-recall curve for clean speech corpus.

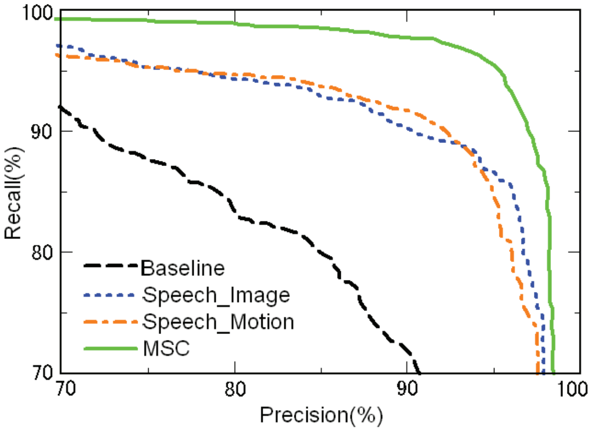


Fig. 5. Precision-recall curve for noisy speech corpus.

was first performed, and then the MSC measure was calculated. ATRASR [12] was used for speech recognition during the speech understanding. By using ATRASR, accuracies of 83% and 67% in phoneme recognition were obtained for the clean speech corpus and the noisy corpus, respectively.

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

4.2. Results

Figures 4 and 5 show the precision-recall curves for clean and noisy speech corpora. The MSC measure and baseline performances are shown by "MSC" and "Baseline." The two lines clearly show that the MSC measure outperforms the baseline for RD speech detection, for both clean and noisy speech corpora. Moreover, the performances using the partial MSC measure are shown by "Speech-Image" (using the confidence 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 improvement in performance. The average maximum F-measures of MSC and baseline were 99% and 94% for clean speech corpus, respectively, and 95% and 83% for noisy speech corpus, respectively. The performance improvement of MSC compared to baseline is 5% for clean speech corpus and 12% for noisy speech corpus. Then we performed the paired t-test and found that there were statistical differences ($p < 0.01$) between MSC and baseline for both clean and noisy speech corpora. Notice that MSC obtains a high performance of 95% even for noisy speech corpus, while the baseline obtains 83%. This means that MSC is particularly effective under noisy conditions.

Finally, to make an RD speech decision by MSC, a threshold could be set to 0.79, which maximized the average F-measure for the clean speech corpus. This means that a speech with a high RD speech probability of more than 79% will be treated as being in the RD domain and the robot will execute an action according to this speech.

5. CONCLUSION

We proposed a novel RD speech detection method based on the MSC measure from speech, the static image of objects, and the motion. Consequently, we showed that the method achieved higher performance compared with a baseline under both clean and noisy conditions.

In future work, we will evaluate our system with a domain selection task and integrate it with methods based on human physical behavior for various applications.

6. REFERENCES

[1] T. Tojo et al., "A conversational robot utilizing facial and body expressions," in *Proc. SMC*, vol. 2, 2000, pp. 858–863.

[2] T. Tetsuya et al., "Human-robot interface using system request utterance detection based on acoustic features," in *Proc. MUE*, 2008, pp. 304–309.

[3] N. Sugimoto et al., "Method for discriminating user-to-system and user-to-human utterances using acoustic features," in *Proc. IEICE*, vol. 2006, no. 1, 2006, p. 133.

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

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

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

[7] I. R. Lane et al., "Out-of-domain utterance detection using classification confidences of multiple topics," in *IEEE Trans. ASLP*, vol. 15, no. 1, 2007, pp. 150–161.

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

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

[10] T. Kurita, "Iterative weighted least squares algorithms for neural networks classifiers," in *Proc. of the Third Workshop on Algorithmic Learning Theory*, 1992.

[11] J. C. Segura et al., "Model-based compensation of the additive noise for continuous speech recognition. experiments using aurora ii database and tasks," in *Proc. Eurospeech*, vol. I, 2001, pp. 221–224.

[12] T. Shimizu et al., "Spontaneous dialogue speech recognition using cross-word context constrained word graph," in *Proc. ICASSP*, 1996, pp. 145–148.