The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Motion Planning based on Simultaneous Perturbation Stochastic Approximation for Mobile Auditory Robots

Makoto Kumon, Keiichiro Fukushima, Sadaaki Kunimatsu and Mitsuaki Ishitobi

Abstract— **In this paper, a motion planning method for mobile auditory robots is proposed based on an optimization technique. Since it is one of the most important abilities for auditory robots to recognize vocal messages correctly, the proposed method is designed to maximize the confidence measure of a speech recognition since the measure is thought to be strongly related to the accuracy of the speech recognition. However, the cost function to optimize is hard to model explicitly, and it is difficult to obtain the gradient that is normally utilized to derive the motion. In order to overcome this difficulty, simultaneous perturbation stochastic approximation(SPSA) that does not require an explicit model of the cost function is applied to generate robot motion. The effectiveness of the approach was verified through real experiments: the robot could get better speech recognition rate after it approached the sound source by measuring the confidence measure.**

I. Introduction

Auditory information such as conversation, ring tone, buzzers and so on, is widely used in our daily life. It is also necessary for robots that work in our environment to handle these auditory features. Especially for human-machine interface, accurate speech recognition is one of the most important abilities and many researchers proposed various approaches in order to realize robust and accurate speech recognition.

For example, beam forming or noise reduction with plural microphones(e.g.[1]) emphasizes the target signal for better recognition. When the recorded sound signal is disturbed and hard to recover the original information, the lost frequency band can be removed from the recognition process using a missing feature mask[2]. Under the existence of plural sound sources, sound source separation technique such as ICA[3] have been widely studied. Those researches tackled the difficulty by improving signal processing methods that handle recorded auditory signal.

On the other hand, the robot can ask the speaker to provide the information repeatedly if necessary in most of practical cases as we do in our daily life. We can also change our position or configuration if it is hard to recognize auditory signal precisely. This approach is also possible for robots since they usually have an ability to change their location or configuration. The present paper proposes a method to plan the motion for a mobile robot with a microphone to reach the better listening spot in the sense of speech recognition.

Most of motion planning methods for robots are based on optimization techniques to maximize (or minimize) the utility (cost) function. For speech recognition, the recognition accuracy is the target to be maximized, but this information is not available for robots since it requires the ground-truth of the sentence to recognize. Although it is impossible to utilize the recognition accuracy, speech recognition algorithms compute various statistical indexes such as estimated generation probability of a word or a phrase, likelihood of the sentence and so on. It is known that those quantities have positive correlation with recognition accuracy, and they can be utilized as quantities to optimize in order to improve the speech recognition accuracy. It is hard to get an explicit model to evaluate those quantities because they are obtained after complicate computation. Therefore, optimization methods utilize gradient of the target function are not suitable for this case.

In order to overcome this limitation, this paper proposes to apply Simultaneous Perturbation Stochastic Approximation(SPSA) algorithm[4], [5] as an optimizer. SPSA utilizes two (or a small number of) parameters to evaluate the target function and approximates the gradient by a difference of those evaluated costs. Parameter sets are generated by stochastic perturbation and the approximated gradient is used to update parameters to optimize. This method is suitable for the objective because it evaluates the target function only twice (or a few times) at each step, which implies that the robot is able to plan and to execute the motion only with a few trials of speech recognition.

This paper is organized as follows. Fundamentals of SPSA is summarized in the next section. The motion planning method for auditory robots to improve speech recognition is proposed in Section III. Since the optimization requires that the target function is continuous, a preliminary experiment was also conducted to show that the quantity to optimize can be a function of the listening position, and an optimization fits as a motion planner. This is also included in Section III. The proposed method was verified through experiments with a mobile platform and results are shown in Section IV. Then conclusion follows(Sec. V).

II. Simultaneous Perturbation Stochastic Approximation

Simultaneous Perturbation Stochastic Approximation(SPSA) has been studied as an efficient optimization method[4], [5]. This approach requires the evaluated value of the target function to optimize at a few points(usually two points) for each optimization step in order to approximate the gradient of the function. This method is useful when 1) an explicit mathematical model of the target function is not

M. Kumon is with Department of Intelligent Mechanical Systems, Graduate School of Science and Technology, Kumamoto University, 2-39-1, Kurokami, Kumamoto, 860-8555, Japan kumon@gpo.kumamoto-u.ac.jp

available to compute its gradient, or 2) the evaluation of the target function requires extensive computational cost, which fits the case considered in the present paper.

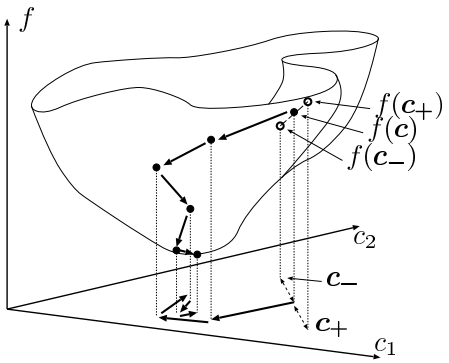


Fig. 1. SPSA optimization : a sketch

A schematic diagram of the optimization steps is shown in Fig.1. Consider that the optimization of a function that is denoted as f is optimized by tuning the parameter vector $\boldsymbol{c}$. Instead of computing the gradient of f as $\frac{\partial}{\partial \boldsymbol{c}} f$, the difference of f at $\boldsymbol{c}_+$ and $\boldsymbol{c}_-$ are used to approximate the gradient statistically. Here, $\boldsymbol{c}_\pm$ are perturbed parameter vectors and every element of the vector is perturbed, that is, $c^i_\pm \neq c^i$ for any i where $c^i_\pm$ and c^i denote the i-th element $\boldsymbol{c}_\pm$ and $\boldsymbol{c}$ respectively.

Given the variable vector $\boldsymbol{c}_k$ at the k-th step of the optimization, the i-th element of the vector at $k+1$-th step, which is denoted as $\boldsymbol{c}^i_{k+1}$, is given as

$$\boldsymbol{c}^i_{k+1} = \boldsymbol{c}^i_k - \alpha \frac{f(\boldsymbol{c}_{+k}) - f(\boldsymbol{c}_{-k})}{c^i_{+k} - c^i_{-k}} \quad \text{for } \forall i, \qquad (1)$$

where α is a positive update gain. It is worth repeating that f is evaluated at $\boldsymbol{c}_+$ and $\boldsymbol{c}_-$ only, no matter how large the dimension of the optimization space is.

The perturbation affected to $\boldsymbol{c}$ is given by random variables of appropriate distributions. For example, ± 1 of Bernoulli distribution can be used as a perturbation generator such as

$$c^{ik}_\pm = c^{ik} \pm_{ik} \beta,$$

where β represents a positive perturbation parameter and $\pm_{ik}$ shows that the sign is decided for each i and k. In the present paper, the perturbation is generated using the above model because of its simplicity that leads computational efficiency.

α and β can be given as a function of the step number k to realize simulated annealing as shown in [4], that is, α and β tend to 0 as k increases.

III. Motion Planning for Auditory Robots

In this section, a motion planning method for a mobile auditory robot based on SPSA is proposed. The target robot system, and the objective of the approach is given first, then the quantity to optimize in order to realize the objective is introduced.

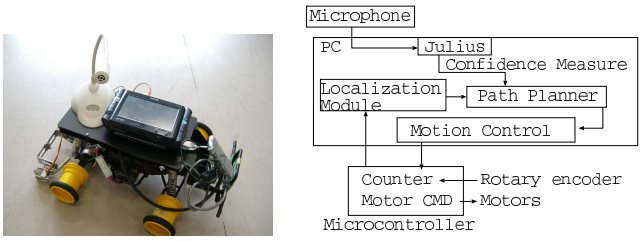


Fig. 2. Mobile Auditory Robot and its system structure

A. Mobile Auditory Robot

The target device to implement the proposed method is a mobile platform with a microphone. Fig.2 shows the robot and its system structure. The robot has two steering wheels and two drive wheels, and onboard computer is installed. Two rotary encoders are equipped for odometry. Steering wheels and drive wheels are actuated by electric motors that are controlled by an embedded microprocessor communicating with the onboard computer. The onboard computer has a microphone interface and it can run a speech recognition software on realtime.

B. Objective

In order to achieve the final goal, the paper proposes a method to compute the move to the location where the robot can recognize the speech stably.

It is impossible for robots to measure the recognition accuracy directly because it needs the exact sentence given to the robot to compare the recognized sentence. On the other hand, the speech recognition engine computes various quantities to evaluate the recognized sentences. Some of those quantities or combinations of them are known to have positive correlation with the recognition accuracy. For example, the ratio of likelihood for the best two recognized candidates can be used to monitor the performance. The algorithm can reject the results when the ratio is close to 1 since the best candidate is hard to be distinguished from the second one.

Following this approach, this paper proposes a motion planning method using one of those quantities in speech recognition without using the information of the recognition accuracy directly.

C. Word Confidence Measure in Speech Recognition

Julius[6] is used as a speech recognition engine in this research. Julius computes likelihood of each candidate and selects the most likely, or those satisfy a given criteria, as the recognized sentence(s). The confidence measure[7] is one of those quantities that evaluate how reliable each recognized word is and the measure is within the range from 0 to 1. Although the measure is not equal to the recognition accuracy rate itself, it can be used as an index to estimate how much the user can rely on the recognized result. This implies that the measure can be used to compute the robot's motion in order to achieve the objective.

Example of confidence measure (CM) for N sentences ($S_1 \cdots S_N$)

S_1	Kumamoto	no	Raishu	no	Tenki	ha?
CM	0.954	0.999	0.768	0.999	0.911	0.999
S_2	Kumamoto	no	Raishu	no	Tenki	ha?
CM	0.924	1.000	0.858	0.999	0.950	0.995
⋮			⋮			
S_N	Kumamoto	no	Raishu	no	Tenki	ha?
CM	0.854	0.999	0.968	0.950	0.899	1.000

underline shows the minimum confidence measure of the sentence.

$$\Rightarrow \underset{1\cdots N}{\text{median}}(0.768, 0.858, \cdots, 0.854) = 0.858$$

Fig. 3. Word Confidence Measure

One recognized sentence provides those scores as many as the number of words that the sentence has. Furthermore, the robot may recognize plural sentences at one spot. In order to compute the robot's motion by a simple way, one of the scores is selected as follows.

Denote the number of recognized sentences at k-th trial as $N(k)$, the number of words in the i-th sentence as $N^i(k)$, and the measure of the j-th word in the i-th sentence as $s^i_j(k)$. Then the representative value of those measures, which is denoted as $s(k)$, is defined as

$$s(k) = \log\left\{\underset{i=1\cdots N(k)}{\text{median}}\left(\min_{j=1\cdots N_i(k)}\left(s^i_j(k)\right)\right)\right\}.$$

Since the computed measure may be varied and disturbed rather largely, the median of the worst scores is utilized in order to reject the effect of outliers. The logarithmic function enlarges the range of s that makes the numerical optimization process easier. In what follows, a set of 17 words and 2 types of grammar is utilized as an example. Although this might be thought rather simple and small set for normal conversation, it is a good size for the robots that accepts simple voice commands.

D. Speech Recognition in a Room

Because of reflection and diffraction by objects or walls in a room, the speech signal is distorted when it is transmitted from the speaker to the listener. This implies that the distortion is a function of position and configuration of the speaker and the listener. It is not surprising to assume that the function is smooth, at least locally. It is also expected that the word recognition confidence parameter depends on position and configuration of the speaker and the listener smoothly. In order to validate this assumption, preliminary experiments to measure the score in a room were conducted. Since this paper considers the case when the listener, or the robot, moves, the listening location is varied.

Fig.4 shows a sketch of the room and the hatched 2m×2m region was considered as the space to search. The confidence measure was computed at every 0.5m grid points, hence there were 25 points in the target region. The microphone was installed at the grid point and it was pointed toward the sound source. Fig.5 shows the measured result in bird eye's view. Roughly speaking, the closer the listener was to the speaker,

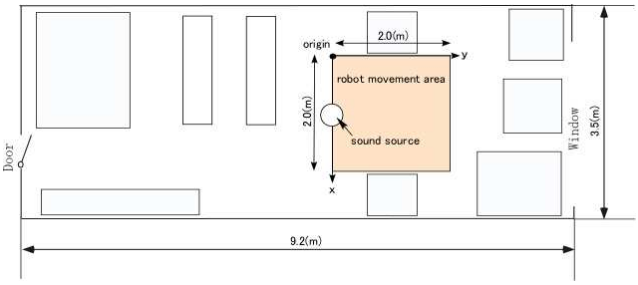


Fig. 4. Sketch of the room for experiments

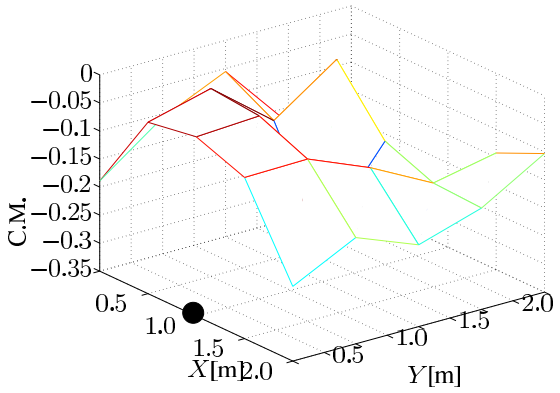


Fig. 5. Confidence Measure in a room: pointing to the source

the better the confidence was, as it was expected. But the peak, or the most confident location, was slightly left from the speaker. This might be caused by acoustic effects of the environment.

Fig.6 shows the result when the direction of the microphone was also considered as a parameter. Since it is impossible to draw the result as in Fig.5, confidence is shown in radar charts. The center of each radar shows the position of the listener, and the direction of the microphone is shown as the angle. The confidence is shown as the radius from the center. In this case, 3 by 3 points were selected to measure.

As shown in Fig.5, the same tendency can be seen that the closer listening point shows the better confidence. It might be surprising that there was no directional effect seen when the listener was close to the speaker, but this was because the test signal was loud and the microphone could get a good signal for any direction. It is hard to find a systematic trend with respect to the direction, but better confidence was obtained when the microphone was not pointing the opposite direction to the sound source.

E. Proposed method

As shown in the previous subsection, the confidence measure, that is expected to represent the recognition accuracy, depends on the listener's position. Then, it is natural to command the robot to move to a certain point, or a region, where the best confidence measure can be obtained.

The logarithmic confidence measure is used as the index to be maximized, and the score can be computed by listening

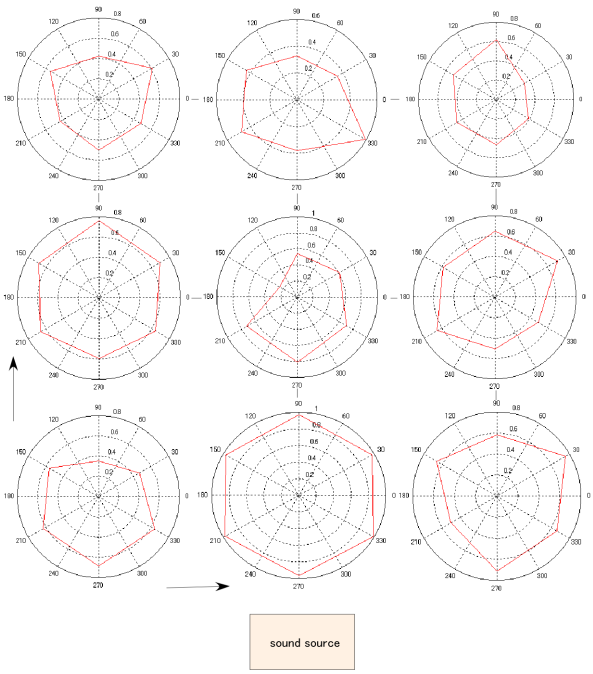


Fig. 6. Logarithmic confidence in a room: with direction.

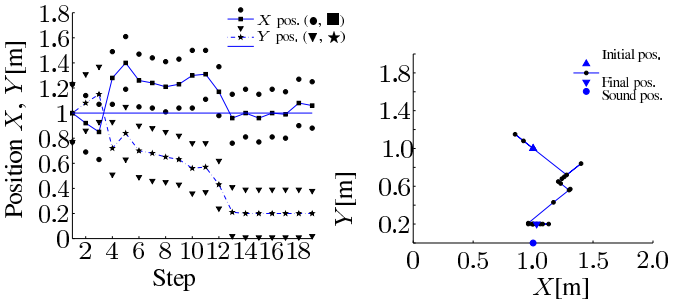
the speech at two different locations.

Let denote the location and configuration of the robot at time k as $\boldsymbol{X}_k$, and the confidence measure obtained at $\boldsymbol{X}_k$ can be expressed as $s(\boldsymbol{X}_k)$. It can be assumed that the robot can achieve the command to move its state $\boldsymbol{X}_k$ to the given desired state owing to the low-level controller.

The method is summarized as follows:

1) Generate the perturbation vectors $\boldsymbol{c}_{\pm k}$, whose i-th element is given as
$$c^i_{\pm k} = \boldsymbol{X}^i_k \pm_{ik} \beta_k.$$
2) Move the robot to $\boldsymbol{c}_{+k}$ and listen the speech
3) Compute the confidence measure $s(\boldsymbol{c}_{+k})$
4) Move the robot to $\boldsymbol{c}_{-k}$ and listen again
5) Compute the confidence measure $s(\boldsymbol{c}_{-k})$
6) Compute the difference of the measure, denoted as $\Delta\boldsymbol{X}_k$ as an approximated gradient:
$$\Delta\boldsymbol{X}_{k_i} = \frac{s(\boldsymbol{c}_{+k}) - s(\boldsymbol{c}_{-k})}{c^i_{+k} - c^i_{-k}},$$
where a subscript $_i$ represents the i-th element of a vector.
7) Move the robot to $\boldsymbol{X}_k + \alpha_k \Delta\boldsymbol{X}_k$.
8) Let $k \to k+1$ and back to Step 1

$\boldsymbol{c}_k$ is defined by a random variable governed by Bernoulli distribution of amplitude β_k. According to SPSA



(a) Motion of the robot with respect to time. (b) Motion of the robot (horizontal view)

Fig. 7. Experiment 1(X,Y case)

formulation[4], α_k and β_k can be given as

$$\alpha_k = \alpha_0 \frac{1}{(k+1+A)^\gamma}$$
$$\beta_k = \beta_0 \frac{1}{(k+1)^\epsilon},$$

where α_0, β_0, A, γ and ϵ show appropriate constants. γ and ϵ should be selected carefully since they play the role to control how the algorithm converges to a solution. When γ and ϵ are large, the algorithm stops updating the parameter quickly, and it might stack at a pseudo maximum.

IV. Experiment

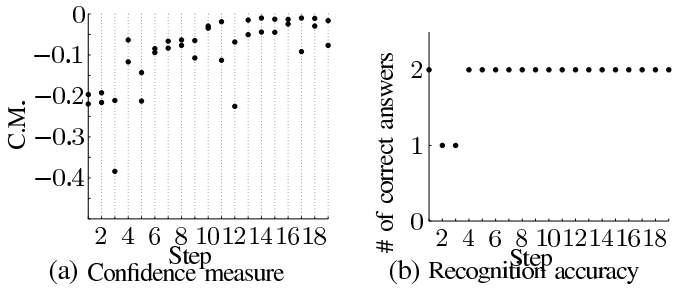
The proposed algorithm was evaluated by experiments in the room shown in Fig.4. Corresponding to Fig.5 and 6, the algorithm was tested in a horizontal displacement case and a horizontal and angular displacement case. In the first case, the microphone was pointed to the sound source and only the horizontal displacement was considered. This required the information about the sound source direction from the robot, which might be not always available. In order to consider the case when no directional information is available, the second case considered the directional displacement of the microphone as well. Adding to those two experiments, the method was also tested in a different room that was larger and noisy in order to verify the robustness.

The sound signal given to the robot was a recorded sentence by a male subject and it was repeated two times at each listening position. In order to avoid the disturbance by ego noise, the sound signal was measured while the robot located at the listening spot quietly.

A. Horizontal Displacement Case (X and Y)

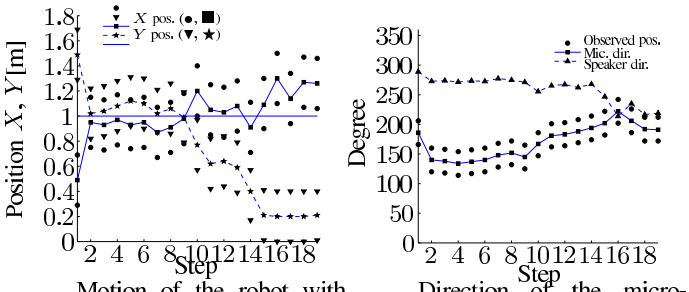
As the microphone was controlled to point the sound source in this experiment, the quantities to be optimized were X and Y location of the robot. α_0 and β_0 were selected to 0.6 and 0.2 respectively. According to [4], $\gamma = 0.101$ and $\epsilon = 0.602$. Because of the limit of the experiment area, Y coordinate was limited to $Y \geq 0.2$. This was realized by clipping the generated motion with respect to Y axis.

The generated motion is shown in Fig.7. Fig.7(a) shows the time history of X and Y location. Markers without lines show the location of perturbed measurement points and

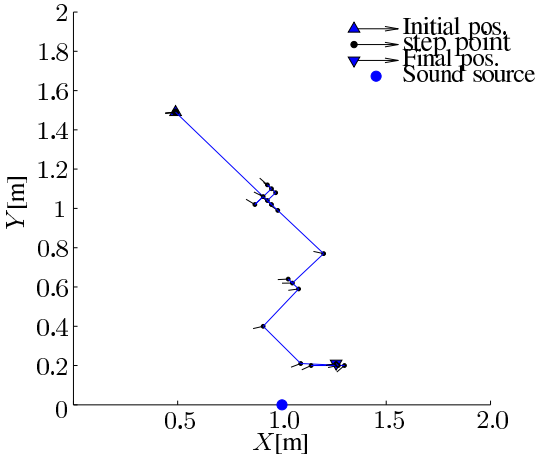


(a) Confidence measure (b) Recognition accuracy

Fig. 8. Confidence measure in speech recognition and its performance



(a) Motion of the robot with respect to time. (b) Direction of the microphone with respect to time



(c) Motion of the robot (horizontal view)

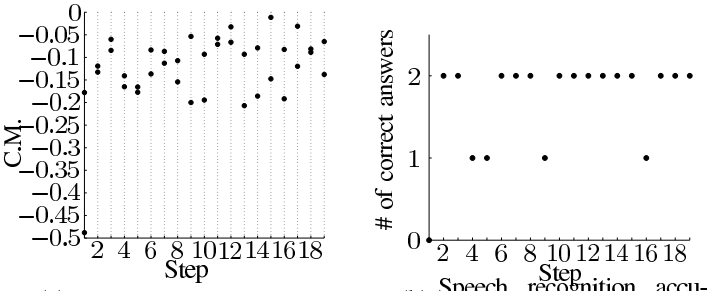
Fig. 9. Experiment 2(X,Y, θ case)

markers with line segments show the location of the robot. As time passed, Y component converged to 0.2 that was the limit of Y and X converged to almost 1.0 where the sound source located $(1.0, 0.0)$. The confidence measure and the recognition accuracy were improved as shown in Fig.8. At each step, two values of confidence measure and recognition accuracy were evaluated corresponding to the measurement points $\boldsymbol{c}_{+k}$ and $\boldsymbol{c}_{-k}$. Fig.8(a) shows the history of the logarithmic confidence measure defined in Sec.III. From the figure, it can be concluded that the confidence measure was improved and got close to the best value 0 as k increased. The recognition accuracy in Fig.8(b) is expressed by the number of correctly recognized sentences. Since the robot obtained two sentences at every measurement, 2 was the best (prefect) recognition. When the logarithmic confidence got higher than -0.1, the recognition accuracy was also 2 (100% accurate).

B. Horizontal and Angular Displacement Case (X, Y and θ)

The second experiment was conducted in the same room. In this case, γ and ϵ were set to 0 in order to search larger parameter space than the previous one because θ was included. Fig.9 shows the motion of the robot and the direction of the microphone. As seen in the previous experiment, Fig.9(a) shows that the robot got close to the sound source. Adding to this, the direction of the microphone also converged to the direction pointing to the sound source (Fig.9(b)).

Confidence measure and the recognition accuracy are shown in Fig.10. Although the robot could reach the sound source with pointing its microphone toward the sound source, the performance index was not as good as the previous case. This was because the direction of the microphone that effected the performance was varied and its motion was kept active until the end of the experiment($\gamma = \epsilon = 0$).



(a) Confidence measure (b) Speech recognition accuracy

Fig. 10. Confidence measure in speech recognition and its performance

C. Robustness

The method was also tested in a different room shown in Fig.11. The room was larger (18.5×10[m] size) than the previous one, and there was the second sound source which generated white noise during the trial. As the first case, $\gamma = 0.101$ and $\epsilon = 0.602$, and the robot was initially located about 1[m] from the sound source and 2[m] from the noise source.

Fig.12 shows the result of the experiment. From Fig.12(a), X decreased gradually and converged around 0, and Y increased at the beginning and then decreased. This is plotted in Fig.12(b), and the motion was indicated by a counter clockwise arrow. The robot escaped from the noise at the beginning and then reached the sound source from the far side of the noise, which was reasonable. Owing to the motion, the performance shown in Fig.12(c) and (d) was improved and succeeded in recognizing the sentence perfectly.

D. Discussion

As shown in Fig.5, the optimal listening spot with respect to the confidence measure located close to the speaker. In this sense, it is acceptable that the robot got close to the sound

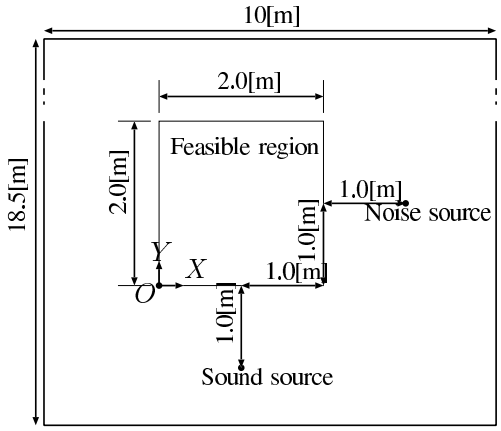
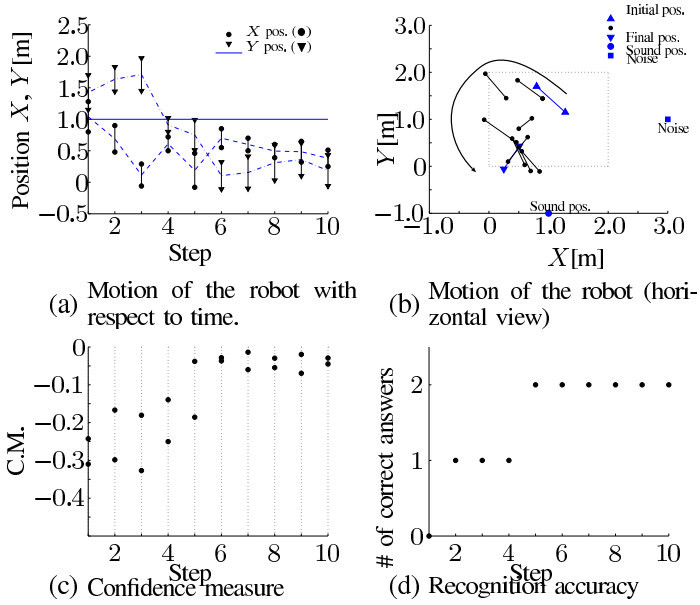


Fig. 11. Large room case, with noise source



(a) Motion of the robot with respect to time.

(b) Motion of the robot (horizontal view)

(c) Confidence measure

(d) Recognition accuracy

Fig. 12. Figures corresponding to Fig.7 and Fig.8

source in order to get better confidence (Fig7, 9 and 12). Since the performance of speech recognition was improved, the assumption that the confidence improvement provides the better recognition accuracy was also validated.

Microphone direction of the second experiment reached to the relative direction of the sound source, which implied that the method succeeded in guiding the robot. However, the speech recognition performance could not be improved stably. This was because the direction of the microphone kept moving since parameters γ and ϵ were set to be 0. Although this parameter configuration is good for the case when the robot needs to explore in a large area, the result shows that an appropriate selection of the parameters is necessary.

From the third experiment, the method succeeded in guiding the robot with reducing the effect from the noise. It is not shown in this paper, but authors have examined that the method also succeeded under the noise of vocal signal or music as far as its intensity did not varied largely. It is expected that the vocal noise could be rejected because of the grammar, but further study is needed.

The algorithm does not require any *a priori* information about the sound field, but just compute the confidence measure at several points. This is practical and important for real active audition. It might be thought that the first experiment requires the information about the location of the sound source in order to keep the microphone pointing to the source. This can be overcome by estimating the location of the sound using microphone arrays, or dynamic robotic head controlled by audio servo[8].

V. Conclusion

A preliminary experiment in this paper revealed that the confidence measure could be modeled as a smooth function of the listener's position. Based on the assumption that the better confidence measure gives more accurate speech recognition, it is also proposed to utilize word confidence measure optimization as a method to plan the motion for a mobile auditory robot. The approach was validated through experiments that showed that the robot could get more accurate recognition after moving from the initial location. Even under noisy situation, the robot could reach the better listening spot in the sense of the speech recognition. However, the study also shows that an appropriate parameter selection is necessary, which will be studied as a future work.

References

[1] Y.Tamai, Y.Sasaki, S. Kagami and H. Mizoguchi, "Three Ring Microphone Array for 3D Sound Localization and Separation for Mobile Robot Audition", Proc. of 2005 IEEE/RSJ Int. Conf. on Intell. Robot. and Sys., 2005, pp.903–908

[2] R. Takeda, S. Yamamoto, K. Komatani, T. Ogata and H.G. Okuno, "Missing-Feature based Speech Recognition for Two Simultaneous Speech Signals Separated by ICA with a pair of Humanoid Ears", Proc. of 2006 IEEE/RSJ Int. Conf. on Intell. Robot. and Sys., 2006, pp. 878–885

[3] J. Even, H.Saruwatari and K. Shikano, "Real-time Implementation of Blind Spatial Subtraction Array for Hands-free Robot Spoken Dialogue System", Proc. of 2008 IEEE/RSJ Int. Conf. on Intell. Robot and Sys., 2008, pp. 2172–2177

[4] J.C. Spaal, An Overview of the Simultaneous Perturbation Method for Efficient Optimization, John Hopkins APL Technical Digest, 19, 1998, pp.482–492.

[5] Y. Maeda, Simultaneous Perturbation Optimization Methods and Their Applications, Systems, Control and Information, 52(2), 2008, pp.47–53. (in Japanese)

[6] A. Lee, T. Kawahara and K. Shikano, "Julius — an open source realtime large vocabulary recognition engine", Proc. European Conf. on Speech Comm. and Tech. 2001, pp. 1691–1694.

[7] A. Lee, K. Shikano and T. Kawahara, "Real-time Word Confidence Scoring using Local Posterior Probabilities on Tree Trellis Search", Proc. of IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2004, pp.793–796.

[8] M. Kumon, T. Shimoda, R. Kohzawa, I. Mizumoto and Z.Iwai, "Audio Servo for Robotic Systems with Pinnae", Proc. of IEEE/RSJ Int. Conf. on Intell. Robot. and Sys., 2005, pp. 885-890.