Human-Robot Interface Using System Request Utterance Detection Based on Acoustic Features

Tetsuya Takiguchi
Organization of Advanced Science and Technology
Kobe University, Japan
takigu@kobe-u.ac.jp

Tomoyuki Yamagata
Graduate School of Science and Technology
Kobe University, Japan
yamagata@me.cs.scitec.kobe-u.ac.jp

Atsushi Sako
Graduate School of Science and Technology
Kobe University, Japan

Nobuyuki Miyake
Graduate School of Science and Technology
Kobe University, Japan

Jerome Revaud
Institut National des Sciences Appliquées de Lyon
69621 Villeurbanne Cedex, France

Yasuo Ariki
Organization of Advanced Science and Tech.
Kobe University, Japan

Abstract

For a mobile robot to serve people in actual environments, such as a living room or a party room, it must be easy to control because some users might not even be capable of operating a computer keyboard. For non-expert users, speech recognition is one of the most effective communication tools when it comes to a hands-free (human-robot) interface. This paper describes a new mobile robot with hands-free speech recognition. For a hands-free speech interface, it is important to detect commands for a robot in spontaneous utterances. Our system can understand whether user's utterances are commands for the robot or not, where commands are discriminated from human-human conversations by acoustic features. Then the robot can move according to the user's voice (command). In order to capture the user's voice only, a robust voice detection system with AdaBoost is also described.

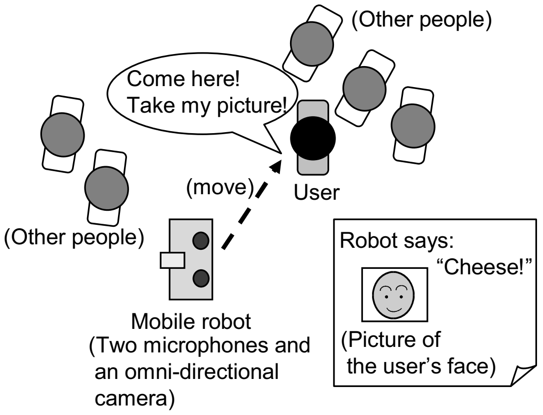

Figure 1. Scenario of mobile picture-taking robot.

1: Introduction

Robots are now being designed to become a part of the lives of ordinary people in social and home environments, such as a service robot at the office, or a robot serving people at a party [1][2]. One of the key issues for practical use is the development of technologies that allow for user-friendly interfaces. This is because many robots that will be designed to serve people in living rooms or party rooms will be operated by non-expert users, who might not even be capable of operating a computer keyboard. Much research has also been done on the issues of human-robot interaction. For example, in [3], the gesture interface has been described for the control of a mobile robot, where a camera is used to track a person, and gestures involving arm motions are recognized and used in operating the mobile robot.

Speech recognition is one of our most effective communication tools when it comes to a hands-free (human-robot) interface. Most current speech recognition systems are capable of achieving good performance in clean acoustic environments. However, these systems require the user to turn the microphone on/off to capture voices only. Also, in hands-free environments, degradation in speech recognition performance increases significantly because the speech signal may be corrupted by a wide variety of sources, including background noise and reverberation. In order to achieve highly effective speech recognition, in [4], a spoken dialog interface of a mobile robot was introduced, where a microphone array system is used.

In actual noisy environments, a robust voice detection algorithm plays an especially important role in speech recognition, and so on because there is a wide variety of sound sources in our daily life, and because the mobile robot is requested to extract only the object signal from all kinds of sounds, including background noise. Most conventional systems use an energy- and zero-crossing-based voice detection system [5]. However, the noise-power-based method causes degradation of the detection performance in actual noisy environments.

Also, for a hands-free speech interface, it is important to detect commands in spontaneous utterances. Most current speech recognition systems are not capable of discriminating system requests - utterances that users talk to a system - from human-human conversations. Therefore, a speech interface today requires a physical button which on and off the microphone input. If there is no button for a speech interface, all conversations are recognized as commands for the system. The button spoils the merit of speech interfaces that users do not need to operate by the hand. Concerning this issue, there are researches on discriminating system requests from human-human conversation by

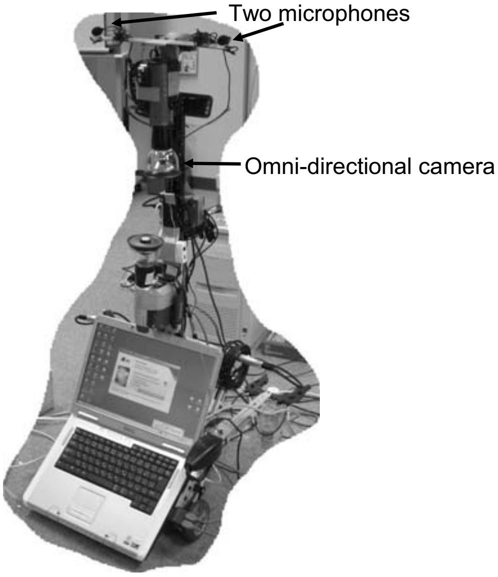

Figure 2. Picture of mobile robot built in this work.

acoustic features calculated from each utterance [6]. And also, there are discrimination techniques using linguistic features. Keyword or key-phrase spotting based methods [7, 8] have been proposed. However, using keyword spotting based method, it is difficult to distinguish system requests from explanations of system usage. It becomes a problem when both utterances contain a same "key-words." For example, the request speech is "come here" and the explanation speech is "if you say come here, the robot will come here." In addition, it costs to construct a network grammar to accept flexible expressions.

In this paper, an advanced method of discrimination using only acoustic features is described. The difference of system requests and spontaneous utterances usually appears on the head and the tail of the utterance [9]. By separating the utterance section and calculating acoustic features from each section, the accuracy of discrimination was improved. In adition, a robust voice/non-voice detection algorithm using AdaBoost, which can achieve extremely high detection rates in noisy environments, is described in this paper [10].

Also, the user's direction estimation by CSP (Crosspower-Spectrum Phase) is implemented on the mobile robot. That enables the mobile robot to serve the user who calls to it from among other people. The two-channel noise reduction method is also implemented in order to improve the speech recognition performance. Using the user's direction estimated by the CSP method, the robot can move freely from its position to the user's position. After the mobile robot moves to the target position, it detects the user's face using the OpenCV library and takes the picture, which is integrated into the mobile robot's operating program.

2: Mobile Picture-Taking Robot

Figure 1 shows a scenario of the multi-modal robot that can take the user's picture (mobile picture-taking robot), and Figure 2 shows a picture of the mobile robot built in this work. The mobile robot can move intelligently in the user's direction by listening to the user's voice, and recognize what the user asks it. As shown in Figure 3, the robust speech recognition system on the mobile robot is composed of four steps. The first step is voice detection with AdaBoost, where

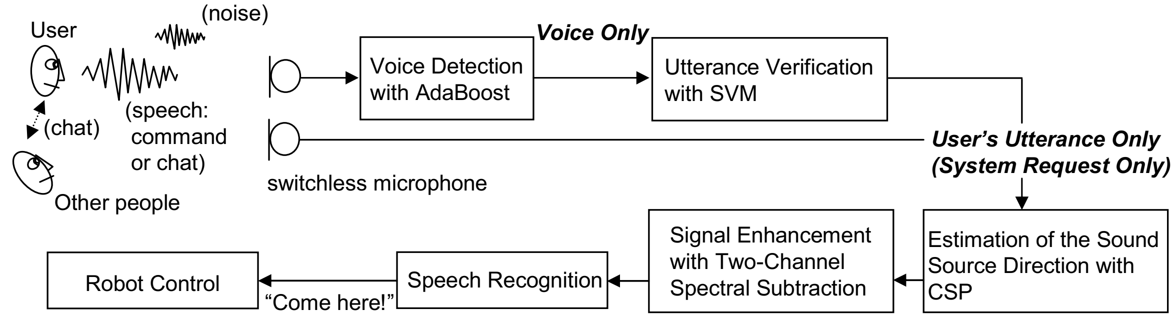

Figure 3. System overview of mobile robot.

the system identifies whether the observed signal is a voice or not. When the signal is a voice, the system performs the second step. The second step is system request detection, where utterances (commands) for a robot only are extracted based on SVM (Support Vector Machine). The third step is estimation of the sound source direction using the CSP (Crosspower-Spectrum Phase) method, where two microphones are used. The fourth step is the signal enhancement for the estimated direction using two-channel SS (two-channel Spectral Subtraction), after which the system carries out speech recognition, and controls the robot according to the speech recognition results. In these experiments, the total number of robot actions is set at 18. For example, the user can say "*Tomatte.* (Stop.)," "*Kocchi ni kite.* (Come here.)," "*Shashin wo totte.* (Take a picture.)," "*Muko wo muite.* (Look over there.)," and so on.

3: Hands-Free Speech Recognition

Speech recognition is one of our most effective communication tools when it comes to a hands-free (human-robot) interface. In the new mobile robot, the robust speech recognition system is composed of four steps (see Figure 3). We describe each step in the following subsection.

3.1: Voice Detection with AdaBoost

In hands-free environments, a speech detection algorithm plays an especially important role in noise reduction or speech recognition, because the user is not able to push a button for recording. In this subsection, a speech/non-speech detection algorithm using AdaBoost, which can achieve extremely high detection rates, is described.

Figure 4 shows the overview of the voice detection system based on AdaBoost. The AdaBoost algorithm [11] uses a set of training data, $\{(X(1), Y(1)), \ldots, (X(N), Y(N))\}$, where $X(n)$ is the n-th feature vector of the observed signal and Y is a set of possible labels. For the speech detection, we consider just two possible labels, Y = {-1, 1}, where the label, 1, means voice, and the label, -1, means noise. Next, the initial weight for the n-th training data is set to

$$w_1(n) = \begin{cases} \frac{1}{2m}, & Y(n) = 1 \text{ (voice)} \\ \frac{1}{2l}, & Y(n) = -1 \text{ (noise)} \end{cases}$$

where m is the total voice frame number and n is the total noise frame number.

As shown in Figure 4, the weak learner generates a hypothesis h_t: $X \rightarrow$ {-1, 1} that has a small error. In this paper, single-level decision trees (also known as decision stumps) are used as the base

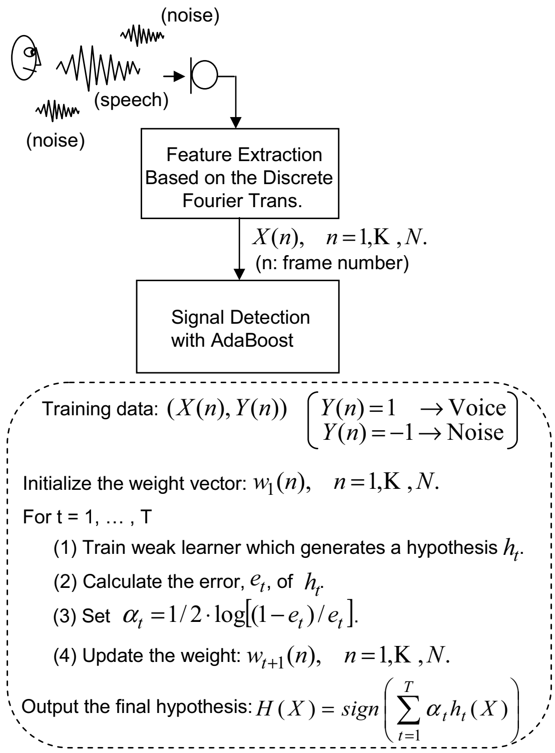
(noise)
(speech)
(noise)
Feature Extraction Based on the Discrete Fourier Trans.
$X(n), \quad n=1,K,N.$
(n: frame number)
Signal Detection with AdaBoost
Training data: $(X(n), Y(n))$ $\begin{bmatrix} Y(n)=1 \rightarrow \text{Voice} \\ Y(n)=-1 \rightarrow \text{Noise} \end{bmatrix}$
Initialize the weight vector: $w_1(n), \quad n=1,K,N.$
For t = 1, … , T
(1) Train weak learner which generates a hypothesis h_t.
(2) Calculate the error, e_t, of h_t.
(3) Set $\alpha_t = 1/2 \cdot \log[(1-e_t)/e_t]$.
(4) Update the weight: $w_{t+1}(n), \quad n=1,K,N.$
Output the final hypothesis: $H(X) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(X)\right)$

Figure 4. Voice detection with AdaBoost.

classifiers. After training the weak learner on t-th iteration, the error of h_t is calculated by

$$e_t = \sum_{n:h_t(X(n)) \neq Y(n)} w_t(n) \tag{1}$$

Next, AdaBoost sets a parameter α_t. Intuitively, α_t measures the importance that is assigned to h_t. Then the weight w_t is updated.

$$w_{t+1}(n) = \frac{w_t(n)\exp\{-\alpha_t \cdot Y(n) \cdot h_t(X(n))\}}{\sum_{n=1}^{N} w_t(n)\exp\{-\alpha_t \cdot Y(n) \cdot h_t(X(n))\}} \tag{2}$$

The equation (2) leads to the increase of the weight for the data misclassified by h_t. Therefore, the weight tends to concentrate on "hard" data. After T-th iteration, the final hypothesis, $H(X)$, combines the outputs of the T weak hypotheses using a weighted majority vote.

In hands-free speech recognition, speech signals may be severely corrupted by noise because the user speaks far from the microphone. In such situations, the speech signal captured by the microphone will have a low SNR (signal-to-noise ratio) which leads to "hard" data. As the AdaBoost trains the weight, focusing on "hard" data, we can expect that it will achieve extremely high detection rates in low SNR situations. For example, in [10], the proposed method has been evaluated on car environments, and the experimental results show an improved voice detection rate, compared to that of conventional detectors based on the GMM (Gaussian Mixture Model) in a car moving at highway speed (the SNR of 2 dB).

3.2: Utterance Verification in Spontaneous Speeches Using Acoustic Features

We describe the system request detection based on SVM (Support Vector Machine) using acoustic features. The proposed method is able to detect system requests reasonably with acoustic fea-

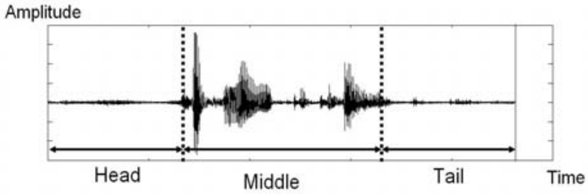

Figure 5. A sample of system request.

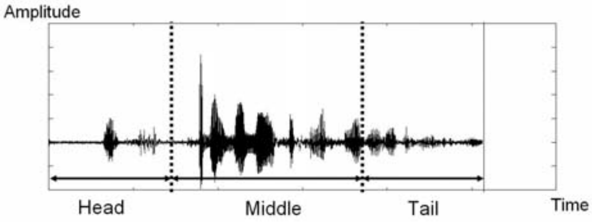

Figure 6. A sample of spontaneous utterance (chat).

tures, because it does not need to reconstruct the discriminator when the system requests are added or changed.

Even if we speak unconsciously, there are acoustic differences between utterances to equipments and those to humans under the condition the subject equipment is machinelike [6]. In our work, we focus on the different characteristics of commands and human-human conversations which usually appear on the head and the tail of the utterance.

The start point and the end point of the utterance are indistinct in chatters while there are no sounds before and after the utterance in commands. There are mainly two reasons that make the start and the end point unclear. One reason is there are usually fillers and falters in chatters while there are short pauses on the head and the tail of utterances in commands. We usually put a short pause before a command to clarify and keep quiet until the system responds something. The other reason is the following person often begins to talk while the current person does not finish talking yet. In this section, we deal with the former case. To put the former phenomenon to practical use, we calculate acoustic parameters not from the whole utterance section but from each three sections below.

To extract the head and the tail of the utterance, the power and zero-crossing are used in this paper. Figure 5 is the wave form of a command utterance, and Figure 6 is that of a spontaneous utterance (chat). The head and tail of the utterance are indistinct in chatters while there are no sounds before and after the utterance in commands as described above. Therefore, as the head and tail of the utterance contain useful information written above, we do not join these margins to the detected utterance section, but calculate acoustic parameters (Table 1) also from each margin separately.

Calculated acoustic parameters are 8 dimensions shown in Table 1, but we calculate them from three sections described above. Thus, the acoustic features are 24 dimensions. The power is computed by Root Mean Square (RMS). The pitch is calculated by LPC residual correlation.

Table 1. Acoustic parameters.

Power	Ave.	S.D.	Max.	Max. - Min.
Pitch	Ave.	S.D.	Max.	Max. - Min.

3.3: Estimation of User's Direction with CSP

The mobile robot is requested to detect a person who calls to it from among a group of persons. This subsection describes the estimation of the user's direction from the user's voice. As the mobile robot may require a small computation resource due to its limitations in computing capability, the CSP (Crosspower-Spectrum Phase)-based technique [12] has been implemented on the mobile robot for a real-time location system.

The crosspower-spectum is computed through the short-term Fourier transform applied to windowed segments of the signal $x_i[t]$ received by the i-th microphone at time t: $CS(n;\omega) = X_i(n;\omega)$ $X_j^*(n;\omega)$, where $*$ denotes the complex conjugate, n is the frame number, and ω is the spectral frequency. Then the normalized crosspower-spectrum is computed by

$$\phi(n;\omega) = \frac{X_i(n;\omega)X_j^*(n;\omega)}{|X_i(n;\omega)||X_j(n;\omega)|} \tag{3}$$

that preserves only information about phase differences between x_i and x_j. Finally, the inverse Fourier transform is computed to obtain the time lag (delay) corresponding to the source direction.

$$C(n;l) = \mathcal{F}^{-1}\phi(n;\omega) \tag{4}$$

Given the above representation, the source direction can be derived. If the sound source is non-moving, $C(n;l)$ should consist of a dominant straight line at the theoretical delay. In this paper, the source direction has been estimated averaging angles corresponding to these delays. Therefore, a lag is given as follows:

$$\hat{l} = \underset{l}{\operatorname{argmax}} \left\{ \textstyle\sum_{n=1}^{N} C(n;l) \right\} \tag{5}$$

3.4: Signal Enhancement with Two-Channel SS

In actual environments, such as a living room or a party room, degradation in speech recognition performance increases significantly because the speech signal from the target speaker (user) may be corrupted by a wide variety of sources, including background noise. In order to improve the performance, the mobile robot must have the noise reduction capability. In this paper, the signal enhancement with two-channel SS [13][14] is used for the mobile robot.

The main beamformer forms a directivity pattern focused on the target direction and the sub-beamformer forms a directional null on the target. The output of the sub-beamformer is assumed to be the noise power and it is subtracted from the output of the main beamformer in the power-spectral domain [14]. The subtraction weight is estimated using an LMS algorithm so as to minimize the output when the target sound is absent.

4: Experiments

In this paper, two experiments were performed to evaluate our system. First, the hands-free speech recognition was evaluated on all command utterances. Second, the utterance verification (system request detection) was evaluated, where two people talk to each other and sometimes make request to the system.

Table 2. Recognition rates for user's request.

	Single mic.	Proposed method
stable	93.05	93.76
moving	86.81	89.93

4.1: Evaluation of Hands-Free Speech Recognition

Experiments were performed to test the hands-free speech recognition system on the mobile robot in a large meeting room, where all utterances are commands for the mobile robot. For speech recognition, we used the grammar-based engine Julian [15] and the speaker-independent HMMs. The dictionary contains about 40 words. The total number of the robot actions is set at 18. The grammar used in these experiments can generate about 60 kinds of sentences.

Six males are used as the testing speakers, and the total number of utterances is 417. In these experiments, we considered both a moving robot and non-moving (stable) robot. The SNR of the moving and non-moving robot is about 15 dB and 20 dB on average, respectively.

The table 2 shows the recognition rates for user utterance. Compared with that of the single microphone, our methods improve the recognition rate from 86.81% to 89.93% for the moving robot. When the mobile robot is not moving, there is essentially no difference between the single microphone and our methods. This is because the SNR for the non-moving robot was high.

Basically, as the SNR was relatively high in the meeting room that was used for the experiment, there was also essentially no difference in the voice detection performance between the conventional method and the AdaBoost-based method. But, in [10], the experimental results clarify the effectiveness of the AdaBoost-based method in a low SNR environment. Therefore, we will research our methods implemented on the mobile robot in noisy real environments, such as a party room.

After the mobile robot moves to the target position, using the user's direction estimated by the CSP method, it detects the user's face using the OpenCV library and takes a picture, which is then integrated into the mobile robot operating program. These processes worked well in the meeting room that was used for these experiments.

4.2: Evaluation of System Request Detection

Experiments were performed to test the utterance verification using the proposed parameters. The corpus for evaluation is recorded under the situation where two people and a system in a same place. Two people talk to each other and sometimes make request to the system (mobile robot).

The length of the recording time is 30 minutes. We did not show them the list of commands that the robot can accept. One reason is to increase the variation of system request commands. The other reason is that we are going to develop speech interfaces which accept not only specified commands but also various expressions. Therefore, they could speak commands that might be acceptable to the robot. We labeled those utterances as system requests manually. Table 3 shows the result of cutting out utterances.

We used SVM with RBF (Gaussian) kernel. Table 4 shows the results of utterance verification evaluated by leave-one-out cross-validation. The results are the cases F-measure became the maximum values. The F-measure became 0.86 where acoustic parameters (24 dim.) are calculated from proposed three utterance sections, while that was 0.66 where the feature values (8 dim.) are calculated from a whole utterance.

Table 3. The numbers of utterances and system requests.

Total utterance	System request
330	49

Table 4. Result of Utterance verification.

	Precision	Recall	F-measure
Acoustic (8 dim.)	0.71	0.61	0.66
Acoustic (24 dim.)	0.80	0.92	0.86

5: Conclusion

To facilitate natural interaction for a mobile robot, a hands-free speech recognition system with a new utterance and system request detection was employed in this paper. To discriminate commands from human-human conversations by acoustic features, it is efficient to consider the head and tail of an utterance. The different characteristics of system requests and spontaneous utterances appear on these parts of an utterance. Separating the head and the tail of an utterance, the accuracy of discrimination was improved.

Future work includes evaluation under the situation where the system accept many kinds of commands and enlarge the amount of corpus. The improvement of detecting utterance sections and the consideration of new kinds of features are also the assignments. We believe that our interface is applicable to a much larger range of up-and-coming service robots.

References

[1] H. G. Okuno, K. Nakadai, and H. Kitano. Social interaction of humanoid robot based on audio-visual tracking. In *Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, LNAI 2358, Springer-Verlag*, pages 725–735, 2002.

[2] J. Miura, Y. Shirai, N. Shimada, Y. Makihara, M. Takizawa, and Y. Yano. Development of a personal service robot with user-friendly interfaces. In *Proc. Int. Conf. on Field and Service Robotics*, pages 293–298, 2003.

[3] S. Waldherr, R. Romero, and S. Thrun. A gesture based interface for human-robot interaction. *Autonomous Robots*, 9(2):151–173, 2000.

[4] H. Asoh, T. Matsui, J. Fry, F. Asano, and S. Hayamizu. A spoken dialog system for a mobile robot. In *Proc. Eurospeech*, pages 1139–1142, 1999.

[5] R. Stiefelhagen, C. Fugen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech, head pose and gestures. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 2422–2427, 2004.

[6] S. Yamada, T. Itoh and K. Araki. Linguistic and Acoustic Features Depending on Different Situations - The Experiments Considering Speech Recognition Rate. In *Proc. of Interspeech 2005*, pages 3393–3396, 2005.

[7] T. Kawahara, K. Ishizuka, S. Doshita, and C.-H. Lee. Speaking-style Dependent Lexicalized Filler Model for Key-phrase Detection and Verification. In *Proc. of ICSLP98*, pages 3253–3259, 1998.

[8] P. Jeanrenaud, M. Siu, J. R. Rohlicek, M. Meteer, H. Gish. Spotting events in continuous speech. In *Proc. of ICASSP*, pages 381–384, 1994.

[9] T. Yamagata, A. Sako, T. Takiguchi, and Y. Ariki. System request detection in conversation based on acoustic and speaker alternation features. In *Proc. of Interspeech*, pages 2789–2792, 2007.

[10] T. Takiguchi, H. Matsuda, and Y. Ariki. Speech detection using real AdaBoost in car environments. In *Fourth Joint Meeting ASA and ASJ*, page 1pSC20, 2006.

[11] Y. Freund and R. E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[12] M. Omologo and P. Svaizer. Acoustic source location in noisy and reverberant environment using CSP analysis. In *Proc. ICASSP*, pages 921–924, 1996.

[13] H. Y. Kim, F. Asano, Y. Suzuki, and T. Sone. Speech enhancement based on short-time spectral amplitude estimation with two-channel beamformer. *IEICE Trans. Fundamentals*, E79-A(12):2151–2158, 1996.

[14] O. Ichikawa, T. Takiguchi, and M. Nishimura. Speech enhancement by profile fitting method. *IEICE Trans. Inf. & Syst.*, E86-D(3):514–521, 2003.

[15] A. Lee and et al. Continuous speech recognition consortium - an open repository for csr tools and models -. In *Int. Conf. on Language Resources and Evaluation*, pages 1438–1441, 2002.