INTECH
open science | open minds

OPEN ACCESS
ARTICLE

International Journal of Advanced Robotic Systems

Simultaneous Blind Separation and Recognition of Speech Mixtures Using Two Microphones to Control a Robot Cleaner

Regular Paper

Heungkyu Lee[1,*]

1 Speech Group, Future IT R&D Lab, LG Electronics Advanced Research Institute, Seoul, Republic of Korea
* Corresponding author E-mail: heungkyu.lee@lge.com

Received 2 Jul 2012; Accepted 5 Dec 2012

DOI: 10.5772/55408

Abstract This paper proposes a method for the simultaneous separation and recognition of speech mixtures in noisy environments using two-channel based independent vector analysis (IVA) on a home-robot cleaner. The issues to be considered in our target application are speech recognition at a distance and noise removal to cope with a variety of noises, including TV sounds, air conditioners, babble, and so on, that can occur in a house, where people can utter a voice command to control a robot cleaner at any time and at any location, even while a robot cleaner is moving. Thus, the system should always be in a recognition-ready state to promptly recognize a spoken word at any time, and the false acceptance rate should be lower. To cope with these issues, the keyword spotting technique is applied. In addition, a microphone alignment method and a model-based real-time IVA approach are proposed to effectively and simultaneously process the speech and noise sources, as well as to cover 360-degree directions irrespective of distance. From the experimental evaluations, we show that the proposed method is robust in terms of speech recognition accuracy, even when the speaker location is unfixed and changes all the time. In addition, the proposed method shows good performance in severely noisy environments.

Keywords Blind Source Separation, Independent Vector Analysis, Noise Reduction, Distant Speech Recognition

1. Introduction

As a human-robot interaction interface, speech recognition has recently attracted considerable interest because its accuracy has been increased, even in severely noisy environments, through a lot of research in recent years. However, various issues to be considered on speech (or voice) user interfaces have newly emerged from autonomous mobile robot research fields. The first issue is derived from robot application characteristics. For example, a robot can move. So, microphones as speech input devices move with it. The distance from a speaker

to a microphone on a robot can increase. This issue is based on the assumption that people do not want to push a button to utter a voice command as well as use a remote controller. From this fact, distant speech recognition has become a fundamental function; also, a robot should be always listening to a speaker's voice command. Second, while an automated system including the speech user interface is applied to a home environment, the suppression or removal of a variety of noise sources such as TV sounds (speech, music, and other sounds), wind, and noise from air conditioners, as well as other noises from home appliances and babble noise, are still challenging issues to enhance speech recognition accuracy. In addition, a robot can generate noise signals by itself while it is moving. Thirdly, there is a real time issue. The speech recognition algorithm must be executed in real time on an embedded system such as a DSP (digital signal processing) board that has low memory and CPU resources. In the worst case, the speech recognition algorithm should share the CPU and memory resource simultaneously on the same DSP board with another control and service program such as simultaneous localization and map-building (SLAM). As a result, the optimized noise removal and recognition algorithms basically have to be executed in real time on an embedded system for intelligent human and robot interaction.

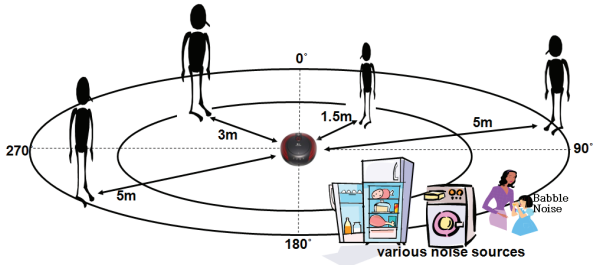

Figure 1. System configuration and conditions for distant speech recognition in a mobile home-robot cleaner in an indoor noisy environment.

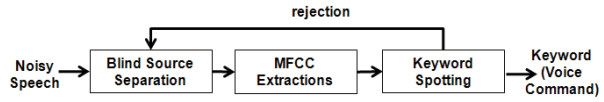

Figure 2. System flow diagram for simultaneous separation and recognition of spoken words.

In this paper, our target application is a home-robot cleaner encountering the various issues mentioned above. People can utter a voice command at a distance at any time, as shown in Figure 1. There exist a lot of noise sources such as TV sound, babble noise, refrigerator sound, and so on. To cope with the above issues, we designed the noise removal and speech recognition algorithms on a home-robot cleaner, as shown in Figure 2. As a noise removal method, we adopt the blind source separation technique, based on the independent vector analysis (IVA) using two microphones, because this algorithm is very effective to cope with a variety of noise sources. In addition, this method is very robust to white and non-stationary noises. In our lab, a real-time independent vector analysis algorithm for improving speech intelligibility in a mobile phone is developed and proposed in [14]. As one of the conventional frequency domains ICA [7][9], independent vector analysis (IVA) can solve some weaknesses of the traditional independent component analysis (ICA) approach [1][4][5][6], such as intensive computation and slow convergence of the time-domain approach, permutation problems [12], and scaling problems with the output [2]. In addition, it can adapt to a moving target signal promptly and improve the separation performance, which separates source signals by estimating an instantaneous demixing matrix on each frequency bin [3]. It solves the frequency domain blind source separation (BSS) effectively without suffering from the permutation problem between the frequencies by utilizing dependencies of frequency bins. The scales of the outputs may be different from the original ones, which can cause frequency distortion when the signal is reconstructed. However, this problem can be solved by adjusting the learned separation filter matrix, which uses the minimal distortion principle [13]. From these advantages, we use the IVA as a noise removal method. In addition, when the robot cleaner moves, it generates significant noise by itself – brush rotation sound, motor sound, and so on. Here, speech is mixed with the mechanical noises and the low sound energy disappears. From this fact, the noise suppression technique, e.g., an adaptive filter, is more suitable than the noise removal technique [8]. Here, the IVA algorithm plays a role of an adaptive filtering function. This is why we choose the IVA algorithm as a noise removal method.

However, there are some issues to be considered in our application when we use the IVA. The problem is demixing convergence time for robust speech recognition. From experimental evaluation, we can know that it takes about two or three seconds to demix filter convergence on a dominant speech sound. However, two or three seconds is too long for speech recognition applications. The late convergence can generate speech distortion, resulting in a speech recognition failure. Thus, we propose the model-based approach where the IVA begins with the trained model off-line and online adaptation is continued. This makes the adaption speed faster. However, when the speaker location is different, the demixing filter coefficient is changed. To cope with this issue, we designed the two-microphone configuration in the vertical alignment direction. This configuration makes the previously trained demixing model similar to the online demixing environment, irrespective of speaker location. So, the demixing filter coefficient is not changed severely even

when the target speech (voice command) appears abruptly in the opposite direction. That is, the preprocessing issue for distant talking is a little bit simplified. Only when a distance is different are the demixing filter coefficients changed and adapted to a target speech.

We adopt the keyword spotting technique for speech recognition because a home-robot cleaner should be always ready to recognize a spoken word at any time, even while people talk with other people and a TV is turned on. So, if a captured sound is not a voice command, it should reject it. The keyword-spotting engine only responds to a predefined voice command (keyword). Talking sounds and sounds generated from a TV are not noise signals but are to be rejected in terms of speech recognition. In our application, the false alarm rate should be very low – unintentional operations of a robot cleaner due to false acceptance pose a reliability issue and cause the customer dissatisfaction.

This paper is organized as follows. In Section 2, we describe the proposed method for simultaneous separation of speech mixture and recognition on a home-robot cleaner. Here we describe the microphone alignment method for two-channel blind source separation in order to cope with spoken words and various noise sources covering 360-degree directions. In addition, the keyword-spotting technique is described to accept only a keyword (voice command) uttered at a distance and reject other signals around the clock. Then, we conduct the representative experiments and discuss the experimental results and performance issues in Section 3. Finally, the concluding remarks are presented in Section 4.

2. Proposed Simultaneous Separation and Recognition Method

2.1 Two-Channel Microphone Alignment

To obtain clean speech signals from noisy speech signals that can be randomly generated from 360-degree directions, we use two microphones. The performance of the multi-channel approach for noise reduction has been proven in many previous research works [17][28]. It can achieve high signal to noise ratio (SNR) from noisy speech signals by effectively removing the background noises. However, traditional speech recognition systems only handle the front 180-degree ranges and unmoving speech signals; utterance distance also has to be small. In this paper, we handle speech recognition [27] from near to 5 m distance, and our target system can move. Thus, a home-robot cleaner and microphones can move together even while a speaker is saying a voice command and spoken sound signals gradually die away from the home-robot cleaner. If we align two microphones by laying out the grounds horizontally on the upper side, we can obtain an opposite direction of arrival angle (DOA) when a speaker utters a voice command to the front side or rear of a robot. Especially in the case of the blind source separation method making use of a base microphone, the demixing filter coefficients should be adapted to a reversed direction of arrival angle as fast as possible. However, as the convergence rate is not fast, we can obtain slightly distorted speech signals. This result is serious in terms of speech recognition application because it results in a recognition failure.

To cope with the above issue, we design the two-microphone configuration by laying out the grounds vertically on the upper and lower sides respectively, as shown in Figure 3 (b). We can only see one microphone on the upper side – the basic input microphone, as shown in Figure 3 (a). The other microphone is below the upper cover and is used for capturing a reference noise signal. The advantage of this microphone configuration is that we can obtain a similar direction of arrival angle irrespective of speaker location, and distance covering 360-degree directions.

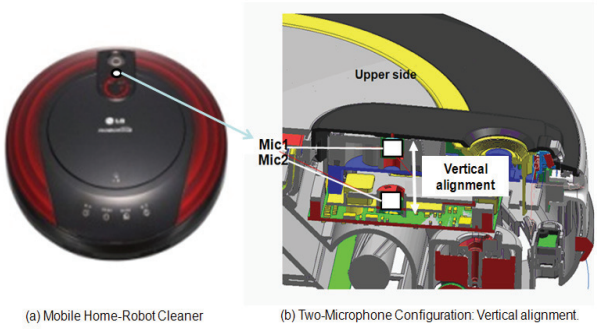

Figure 3. The mobile home-robot cleaner and its two-microphone configuration. Only the base microphone is visible; the other microphone is hidden within the robot.

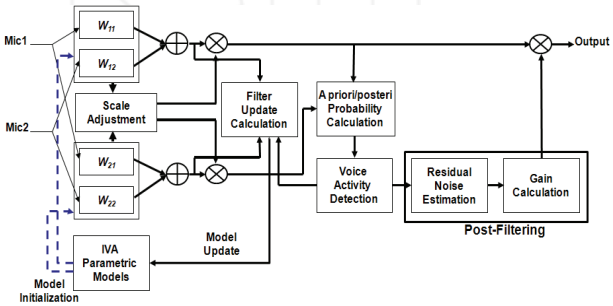

Figure 4. The proposed online two-channel based independent vector analysis and post filtering using a voice activity detection method to effectively remove residual noise from a separated speech signal part.

2.2 Model-based Independent Vector Analysis and Online Adaptation

We employ a real-time independent vector analysis (IVA) method for noise removal. By using two-channel IVA, we classify captured signals into speech and other noise signals because the number of separated signals cannot be greater than the number of microphones used. The base microphone is Microphone 1 and the reference source is the signal obtained from Microphone 2. The proposed system starts with the IVA parametric models previously trained for demixing the matrix of a defined target location. The IVA parametric model is not dependent on the direction. Then, the real-time adaptation method is combined to cope with other speech locations. The overall structure for the IVA is as shown in Figure 4, consisting of a mixing step, a separation step, and an online learning step. Additionally, the post-filtering method is employed to remove residual noise; the voice activity detection is also applied to enhance the performance of the post-filtering.

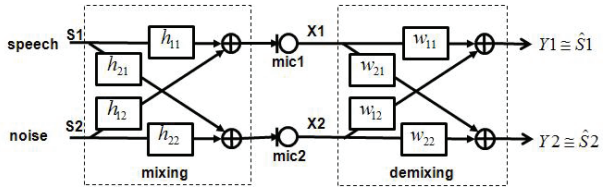

Figure 5. Two-channel based basic blind source separation architecture: mixing and demixing relation.

A. Mixing Step

The traditional concept of blind source separation is given in Figure 5, composed of mixing and demixing parts. We can assume that speech and noise are inputted to the microphones through the mixing process in a convolutive environment [10][11] as shown in the left part of Figure 5. That is, the input signals, $x_1(t)$ and $x_2(t)$ are computed as follows:

$$x_1(t) = \sum_{\tau=0}^{T-1} h_{11}(\tau)s_1(t-\tau) + \sum_{\tau=0}^{T-1} h_{12}(\tau)s_2(t-\tau) \quad (3)$$

$$x_2(t) = \sum_{\tau=0}^{T-1} h_{21}(\tau)s_1(t-\tau) + \sum_{\tau=0}^{T-1} h_{22}(\tau)s_2(t-\tau) \quad (4)$$

From the above assumption, speech and noise can be separated by estimating the demixing process as shown in the right part of Figure 5. It is blind separation because we cannot know the mixing process. First, the Hanning window is applied to the input signals, $x_1(t)$ and $x_2(t)$ respectively, as follows:

$$x_1' = w(t)\cdot x_1(t) \cong x_1'(n,k) = \sum_{t=0}^{K-1} w(t)x_1(n,k) \quad (5)$$

$$x_2' = w(t)\cdot x_2(t) \cong x_2'(n,k) = \sum_{t=0}^{K-1} w(t)x_2(n,k) \quad (6)$$

Here, the window length should be sufficiently longer than the length of the mixing filter $h_{ij}(t)$. Then, the fast Fourier transforms (FFT) are applied to equations (5) and (6) as follows:

$$X_1(n,k) = \sum_{t=0}^{K-1} x_1'(nJ+t)e^{-jW_k t} \quad (7)$$

$$X_2(n,k) = \sum_{t=0}^{K-1} x_2'(nJ+t)e^{-jW_k t} \quad (8)$$

Equations (7) and (8) are the same as the following mixed form:

$$X_1(n,k) \cong H_{11}(k)S_1(n,k) + H_{12}(k)S_2(n,k) \quad (9)$$

$$X_2(n,k) \cong H_{21}(k)S_1(n,k) + H_{22}(k)S_2(n,k) \quad (10)$$

B. Separation Step

Next, we can estimate the original signals S_1 and S_2 through estimating the inverse matrix, W of the mixing matrix, H. Let the original signals $S_1=Y_1$ and $S_2=Y_2$ in the frequency domain; then, the separated source signals, $Y=WX$, are given as:

$$Y_1(n,k) = W_{11}(k)X_1(n,k) + W_{12}(k)X_2(n,k) \cong S_1(n,k) \quad (11)$$

$$Y_2(n,k) = W_{21}(k)X_1(n,k) + W_{22}(k)X_2(n,k) \cong S_2(n,k) \quad (12)$$

where the demixing matrix W is same as H^{-1}. To resolve the scale problem, we apply the minimal distortion principle as follows:

$$\bar{W} = P\cdot D\cdot H^{-1} \quad (13)$$

where P is permutation matrix. If we assume that P is identity matrix I, equation (13) is rewritten as

$$\bar{W} = D\cdot H^{-1} \quad (14)$$

where D is *diag(H)*. Thus , equation (14) can be written as

$$\bar{W} = \mathrm{diag}(H)\cdot H^{-1} \quad (15)$$

Equation (15) should be satisfied with the following condition with respect to any diagonal matrix, E.

$$\mathrm{diag}(H\cdot E)(H\cdot E)^{-1} = \mathrm{diag}(H)\cdot H^{-1} \quad (16)$$

Here, $W=EH^{-1}$. Thus, W can be written as equation (17) and *diag(H)* can also be replaced as equation (18).

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} = \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^{-1} \quad (17)$$

$$\mathrm{diag}(H) = \mathrm{diag}\left(W^{-1}\right)$$

$$= \mathrm{diag}\left(\left[\begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^{-1} \right]^{-1} \right)$$

$$= \mathrm{diag}\left(\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} 1/E_1 & 0 \\ 0 & 1/E_2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} \begin{bmatrix} 1/E_1 & 0 \\ 0 & 1/E_2 \end{bmatrix} \quad (18)$$

Thus, equation (15) can be arranged by using the equation (18) as follows:

$$\bar{W} = \mathrm{diag}(H) \cdot H^{-1}$$

$$= \begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} \begin{bmatrix} 1/E_1 & 0 \\ 0 & 1/E_2 \end{bmatrix} \begin{bmatrix} E_1 & 0 \\ 0 & E_2 \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^{-1} \quad (19)$$

From equation (19), the separation model, *Y=WX* become $\bar{Y} = \bar{W} \cdot X$ as in equation (20) or equation (21) and (22):

$$\begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix} = \begin{bmatrix} H_{11} & 0 \\ 0 & H_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}^{-1} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (20)$$

$$\begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix} = \frac{1}{W_{11}W_{22} - W_{12}W_{21}} \begin{bmatrix} W_{22} & 0 \\ 0 & W_{11} \end{bmatrix} \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (21)$$

where *diag(W-1)* is $\frac{1}{W_{11}W_{22} - W_{12}W_{21}} \begin{bmatrix} W_{22} & 0 \\ 0 & W_{11} \end{bmatrix}$, *W* is *H1*, and *Y* is $\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. Therefore, equation (21) can be written as follows:

$$\begin{bmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{bmatrix} = \frac{1}{W_{11}W_{22} - W_{12}W_{21}} \begin{bmatrix} W_{22} & 0 \\ 0 & W_{11} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \quad (22)$$

where $\frac{1}{W_{11}W_{22} - W_{12}W_{21}} \begin{bmatrix} W_{22} & 0 \\ 0 & W_{11} \end{bmatrix}$ is for the scale adjustment term.

C. Online Learning Step

For real-time blind source separation, it is necessary to extract outputs immediately. Thus, the learning process must be a fully online algorithm that is appropriate for practical embedded systems. Let the demixed signal *y* be the *wx*. Then, the coefficients of the separation filter matrices are updated at every frame as follows:

$$w(n+1) = w(n) + \eta \Delta W(n) \quad (23)$$

$$\Delta W(n) = \sum_{l=1}^{L} \left(I_{il} - R_{il}(n) \right) W(n) \quad (24)$$

where $R_{il}(n)$ is the online version of the scored correlation at the current frame. We apply the natural gradient in order to compute the demixing filter coefficients. The equation (24) is the online natural gradient learning rule. In addition, we apply the nonholonomic constraint [15] as given in equation (25) in order to solve the stability problem that can arise in the case of online learning. We can obtain the following gradient with the constraint by simply replacing the identity matrix I_{il} with $\Lambda_{ii}(n)$.

$$\Delta W_{ij}(n) = \sum_{l=1}^{L} \left(\Lambda_{il}(n) - R_{il}(n) \right) W_{ij}(n) \quad (25)$$

where $\Lambda_{ii}(n)$ is equal to $R_{ii}(n)$, and $\Lambda_{ii}(n)$ is 0 when *i* is not equal to *l*. We can adjust the learning rate with a normalization factor $\xi^{-1}(n)$ as given in equation (26). Furthermore, equation (26) is normalized with respect to the input level in order to improve the convergence property as given in equation (27).

$$W(n+1) = W(n) + \eta \cdot \sqrt{\xi^{-1}(n)} \cdot \Delta W(n) \quad (26)$$

$$\xi(n) = \beta \xi(n-1) + (1-\beta) \sum_{i=0}^{L} \frac{|x(n)|^2}{L} \quad (27)$$

D. Voice Activity Detection and Post-Filtering

The condition of a blind source separation is that the number of sources *L* is less than or equal to the number of observed signals *M*. However, we use only two microphones although there are a lot of noise sources in a home environment. Thus, it is not possible to separate all of the sources within a home environment. We therefore only classify the captured signals into speech signals and noise signals according to the decision of the voice activity detector [19] while the blind separation process is going on. Furthermore, even when there are only two observations, the real-time separated outcome is not satisfactory for speech recognition because of residual noises. Thus, we apply the post filtering method [17][18] to the separated output signal. The post-filtering method is based on the minimum mean square estimation (MMSE). To enhance the performance of the post filtering for noise reduction, the voice activity detection algorithm is also utilized because accurate noise power estimation

can increase the performance of the speech enhancement. The voice activity detection and post-filtering module is combined with the IVA as shown in the bottom right of Figure 4. These algorithms are well known and proved in many research papers.

2.3 Keyword Spotting in Stop and Running Modes

In our system, a home-robot cleaner should be always in a recognition-ready state to respond to a voice command in a home environment. There are a lot of noise sources here, such as talking, children, TV sounds, mechanical noises from refrigerators, air-conditioners, and so on. Thus, false acceptance is a critical issue because the proposed system can behave abnormally. To resolve this issue, the keyword spotting technique is applied. Keyword spotting refers to the detection of all occurrences of any given word in a speech signal [17]. Most previous work on keyword spotting and our system are based on hidden Markov models (HMMs) as in [23][24][25][26]. The keyword spotting engine has filler models. The filler models can compete with the keyword models in terms of log likelihood in each state sequence. If a final output is a keyword, the accumulated log likelihood values are compared to the predefined threshold. If it is greater than the predefined threshold, it is only accepted. The threshold is defined from the off-line experiments that evaluated the false acceptance rate (FAR) and the false rejection rate (FRR). Our main interest here is the false acceptance rate. Thus, we define the threshold when the FAR is under 5% even though the FRR is high.

In our system, the keyword spotting method has two main functions. The first function is to play the role of an activator to start a speech recognition system like an end point detector (EPD). First, we give the system a name, such as "Robo-king". After that, a real voice command is spoken such as "start cleaning". The other function is the main speech recognition function. Thus, we have to utter a voice command to control our system such as "Robo-king, start cleaning". We classify the recognition modes into the stop mode and the running mode. When a robot cleaner is cleaning a room in the running mode, severe noises occur because of the operation of brushes and motors. In addition, the simultaneous localization and map-building (SLAM) algorithm is run for autonomous cleaning. Thus, the speech recognition engine should share the CPU and memory resources with the SLAM engine. From this condition, we use only one keyword recognition mechanism in the running mode in order to use the CPU and memory resources as little as possible. The one keyword is the name of the robot cleaner. If a speaker says the name, "Robo-king" in the running mode, the robot cleaner is stop to recognize the next voice command. At this time, the mode of the robot cleaner is changed into the stop mode. If there is not any voice command for some periods of time, the cleaner start to clean a room continually. In stop mode, 25 keywords are used to control the robot and provide information for a user.

3. Experimental Simulations

3.1 Experimental Setup

To implement the proposed method, we used an ARM-11 DSP board and an RVDS 4.0 compiler. The operating system is the embedded Linux. The speech input is sampled to 16 khz PCM, and the 39th mel-frequency cepstral coefficients (MFCC) feature vectors are used for robust speech recognition. The frame length is 25 msec (400 samples), and the frame shift interval is 10 msec (160 samples). The number of keywords is 25. In addition, the TTS (Text-To-Speech) engine is applied as well to respond to a voice commend of a speaker where the output speech sampling rate is also 16 khz. The robot control engine and speech recognition engine are run simultaneously as an independent thread on the same embedded operating system. They use a message-passing mechanism for communicating between them. All kinds of test speech databases are generated in a reverberation chamber. The room size is 7 m *x* 5 m, and the height is 2.75 m.

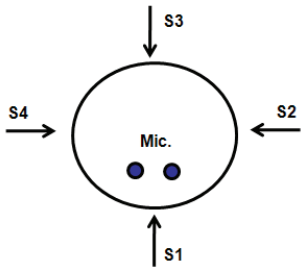

Figure 6. The horizontal two-microphone alignment setup.

3.2 Experimental Evaluations

3.2.1 Evaluation of Speech Recognition Accuracy according to a Microphone Alignment

Prior to showing the robustness of the proposed vertical microphone alignment method, the defect of the traditional horizontal microphone alignment is proved to evaluate it in each direction with same test data. Two microphones are attached on the upper side panel of a home-robot cleaner as shown in Figure 6. The distance between microphones is 12 cm. Then, speech recognition tests are done at the four different locations, S1(0 degree), S2(90 degree), S3(180 degree), and S4(270 degree). The two-channel input data are computed by the IVA method and then passed to the isolated word recognizer. To generate the test speech database, we recorded the test files with 20 men and 20 women speaking the 10 voice commands three times at 3 m and 5 m distance, respectively. Thus, the total 2400 speech utterances are used for each direction respectively. To verify the

recognition accuracy at all directions and in the same conditions, 2400 files from the recorded clean speech database are played by the mouth simulator (loudspeaker) and then evaluated. The experimental results are shown in Table 1, where the isolated word recognizer is used in order to verify just the speech recognition accuracy in each direction. As a starting point, we used the demixing filter coefficients that are trained in off-line for the S1 direction. So, the recognition accuracy showed good results in the direction S1. Meanwhile, we obtained the degraded accuracy at other directions because the demixing filter coefficients could not be adapted promptly from the S1 direction to other spoken directions (S2, S3, and S4), where a distorted speech output may be passed promptly to the feature extractor and recognizer in order to meet the real-time constraint. The worst recognition accuracy was obtained at the opposite side, S3, because the direction of arrival angle is abruptly changed into the opposite direction.

	S1	S2	S3	S4
Recognition Rate (%)	90.51%	89.35%	83.10%	89.58%

Table 1. Respective average speech recognition rate and comparison results in four different directions. The distance from a mouth simulator playing recorded speech files to the home-robot cleaner was 3 m and 5 m.

If we use the horizontal microphone alignment method shown in Figure 6, the four kinds of demixing filter coefficients are required. In addition, we would have to choose the maximum likelihood value among the output values estimated in all directions. These require high computational and memory resources that are not pertinent to an embedded system. Thus, we applied the proposed vertical microphone alignment method shown in Figure 3. This alignment was not dependent on the 360-degree directions. Through the experimental evaluations under the same conditions and with the same data used in Table 1, we proved the robustness of the proposed method. We obtained a similar result in all directions when we used the vertical microphone alignment method. The average speech recognition rate was 91% and the difference in the speech recognition rates between test locations was smaller than 0.3%. When we evaluated the baseline speech recognition rate by using one-channel PCM data without a noise reduction method, the speech recognition rate was 90.7%. From this experiment, we can see that the IVA demixing filter coefficients are trained and adapted very well to a target speech.

3.2.2 Performance Evaluation of the Keyword spotting

First, we evaluated the performance of the traditional end-point detection based isolated word recognizer to prove the effectiveness of the keyword spotting technique. Then, we compared it to the keyword spotting engine. Actually, the traditional EPD-based speech recognition system has severe weaknesses because the EPD fails to find the start point of a speech in severe non-stationary noisy environments; furthermore, in some applications there is no PTT (Push To Talk) button. Thus, an EPD-based speech recognition interface is not an appropriate method in noisy service environments. To prove this, we evaluated EPD-based speech recognition accuracy using a noisy speech database recorded where a television was turned on. Our hypothesis was that the TV sounds would be a critical factor that could cause the degradation of detection rate in an indoor environment. The number of test utterances was 6,400. The speaking distance was 3 m, and the SNR (Signal to Noise Ratio) was between 5 dB and 10 dB on average. The experimental results are shown in Table 2. The baseline test result of the EPD-based isolated word recognizer is very poor because the speech sounds generated from the TV indeed prevented the EPD from finding the start point of a voice command. In addition, we can see that the two-channel based IVA method did not discriminate well between a voice command and TV sounds even though the speech recognition accuracy increased 17.21% after applying the IVA when compared to the baseline recognition rate. The speech-like residual noises in the separated speech output caused the degradation of the speech recognition accuracy. Therefore, we evaluated the performance of the keyword spotting engine, and obtained an average 19.23% improvement in detection rate when compared to the EPD-based isolated word recognizer with two-channel based IVA.

Noise Type	EPD based ASR; Baseline	EPD based ASR with 2Ch IVA	Keyword Spotting with 2Ch IVA
Drama	46.51%	61.02%	80.5%
Music	58.86%	80.15%	86.3%
Music + Speech	50.99%	70.90%	85.4%
News	32.47%	45.59%	82.4%
Average	47.21%	64.42%	83.65%

Table 2. Speech recognition experimental results for TV sounds: drama, music, music and speech, and news. The SNR is between 5 dB and 10 dB on average, and the distance from the TV to the home-robot cleaner is 1 m. The speaking distance is 3 m.

From the experimental evaluations in Table 2, we employed the keyword spotting based speech recognition approach. Speech recognition accuracy is important; however, the false acceptance rate is a critical problem in our system because the robot cleaner can execute operations at any time. Thus, the out-of-vocabulary (OOV) rejection method is a crucial factor in. To do this, we applied the 25 filler models to our keyword-spotting engine. Then, we defined the threshold so that the false acceptance rate could be 5% lower. In our system, we

decided that the threshold was 8 where the detection rate was 97.1% and the false acceptance rate was 4.8% as shown in Table 3. Using this configuration, the keyword spotting test evaluation for TV sound noises as shown in Table 2 is performed. The test result in Table 3 is evaluated using a total of 6,400 clean speech files from a database, recorded at distances of 3 m and 5 m. In addition, the number of the OOV test files is 70,000.

Threshold	0	8	10	20	30	40
Detection Rate	95.4%	97.1%	97.9%	98%	97.1%	95.8%
False Acceptance Rate	3.9%	4.8%	5.4%	6.9%	8.8%	11.5%

Table 3. Experimental results to decide the threshold for rejecting an out-of-vocabulary command.

3.2.3 Performance Evaluation of Two-channel based IVA

As a speech enhancement technique, independent vector analysis is well known as a method that does not distort the inputted speech signals. We have already shown the improvements in terms of speech enhancement by using the signal-to-interference ratio (SIR) measure in previous work [14]. To show the robustness of two-channel based IVA in speech recognition application, experimental evaluations are done using a total of 6,400 files recorded under different noisy conditions. These data are generated using the speech database set in which 64 men and 64 women articulate 25 keywords two times at a distance of 50 cm. To generate the noisy speech database, we recorded the test files in each noisy environment respectively after the 6,400 original speech files were played at a 3 m distance and the noise signal played simultaneously at a 3 m distance and at a 45 degree angle. The test data used two types of noise – babble and pub noises. The sound level of babble and pub noises is adjusted in order to make the SNR 0, 10, 20, and 30 dB.

Noise Type	SNR	Baseline	2Ch IVA
Babble Noise	0dB	0%	55.88%
	10dB	77.31%	90.34%
	20dB	92.44%	95.8%
	30dB	98.32%	98.32%
Pub Noise	0dB	0%	44.54%
	10dB	61.34%	86.55%
	20dB	91.6%	92.44%
	30dB	98.32%	98.74%
Average		64.92%	82.83%

Table 4. Experimental results from noisy environments using babble and pub noises in SNR 0, 10, 20, and 30 dB. After applying the two-channel IVA method, we obtained an average 17.91% improvement in detection rate. The speaker distance was 3 m from a mouth simulator (loudspeaker) to a home-robot cleaner.

The experimental results are given in Table 4, where the false acceptance rate of the baseline keyword spotting engine is 4.8%. Table 4 describes the detection rate. Here, 1,000,000 speech files are adapted to the original HMM models using maximum posterior estimation. Those files from the speech database are uttered at a distance. Table 4 shows that the detection rate at SNR 30 dB was about 98%. This result showed about 7.5% improvement when compared to the recognition result of the isolated word recognizer in Table 1. From this result, we can see that the matched condition between acoustic model and test feature vectors could increase the recognition rate. In addition, we obtained the average 17.91% improvement compared to the average baseline detection rate when the two-channel based IVA is applied.

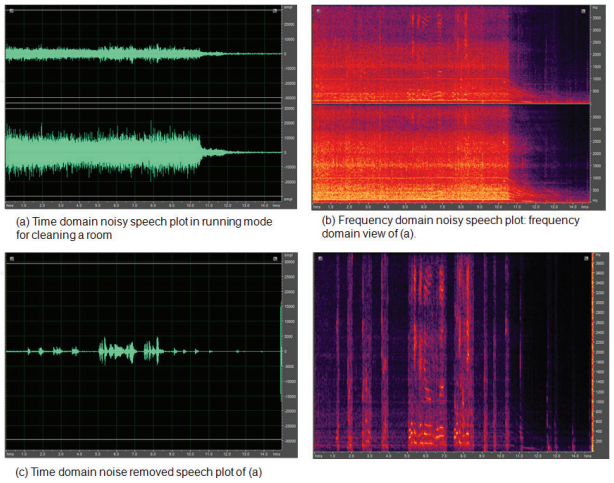

Figure 7. The captured original noisy speech data and its separated speech data for when a robot cleaner was moving and then stopped after recognizing a voice command, "Robo-king stop". While capturing data some people were talking quietly, and this talking is also obtained even though the SNR was under 0 dB, as shown in (c).

Noise Type	SNR	Baseline	2Ch IVA
Motor Noise	5dB ~10dB	58.74%	88.5%

Table 5. Detection rate in running mode. Only one keyword is applied to deactivate a home-robot cleaner. The motor noise signal is generated when a home-robot cleaner is moving, This caused the degradation of speech recognition rate.

The two-channel based IVA method showed the robustness even in the running mode of a home-robot cleaner. Figure 7 describes the time domain and frequency domain view of the captured sample file before and after the two-channel based IVA method is applied. While a home-robot cleaner is moving, the motor and brush noise sounds are generated. These noises caused the degradation of speech recognition rate. When we checked the SNR, it was approximately between 5 dB and 10 dB because the moving speed was different. To obtain the statistical information in the running mode, we performed the offline tests using 6,400 recorded speech files. To generate the noisy speech database in the running mode, we made a home-robot cleaner clean in a reverberation chamber, where the 6,400 original speech files were played at a 3 m distance. After that, we recorded the above status. The off-line experimental

result in the running mode is shown in Table 5. We obtained 29.76% improvement in detection rate after the two-channel based IVA method was applied.

4. Conclusions

In this paper, our main focus was to recognize a keyword uttered at a distance in noisy environments around the clock where the false acceptance rate should be lower. Our system can move and a user can utter a voice command at a long distance. The performance of a speech recognizer in such a situation is vulnerable to various noises. Thus, we employed the independent vector analysis based two-channel noise reduction method for robust speech recognition on a mobile home-robot cleaner. Additionally, we did not use a remote controller to activate a speech recognition function. A home-robot cleaner should be always listening to all kinds of sound signals generated in real life, and then promptly respond to a specific keyword. Meanwhile, it should reject other sounds and speech signals. To cope with the above issue, the keyword spotting technique is applied. Here, the real-time blind separation of noisy speech mixtures and recognition are performed on an ARM-11 digital signal processing board.

In our system, our goal is to provide reliable and stable speech recognition. We prefer a low false acceptance rate to a high recognition rate. So, we focused on preventing abnormal operation. We did not deal with the distance speech recognition issue in order to increase the accuracy of speech recognition. Speech feature enhancement, search problems to find the best word hypothesis, and hidden Markov model parameter estimation can be considered to enhance the performance of distant speech recognition [27], and these could be some ideas for future work. In addition, we failed to consider the reverberation issue within an indoor environment. The speech recognition rate is abruptly decreased by up to 40% when the reverberation time is greater than one second. The dereverberation method is still unsolved The performance of dereverberation methods is still incomplete because the room impulse response can be changed according to a variety of conditions and materials even though there a lot of work has been done in this area [20][21][22]. We think that this is an important research area and future work should aim to solve the speech recognition problem in reverberant environments.

5. References

[1] Y. Zhao, K.-C. Yen, S. Soli, S. Gao, and A. Vermiglio, (2002) On application of adaptive decorrelation filtering to assistive listening, *J. Acoust. Soc. Am.* Vol. 111, 1077–1085.

[2] K. Matsuoka and S. Nakashima, (2001) Minimal distortion principle for blind source separation, in Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation, pp. 722–727.

[3] T. Kim, H.T. Attias, S.-Y. Lee, T.-W. Lee, (2007) Blind source separation exploiting higher-order frequency dependencies, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 1, pp. 70–79.

[4] D. Yellin and E. Weinstein, (1996) Multichannel signal separation: methods and analysis, Vol. 44, No. 1, pp. 106–118.

[5] R. Lambert, (1996) Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures, Ph.D. dissertation, University of Southern California.

[6] K. Torkkola, (1996) Blind separation of convolved sources based on information maximization, in Proc. IEEE Int. Workshop on Neural Networks for Signal Processing, pp.423-432.

[7] T.-W. Lee, A. J. Bell, and R. Lambert, (1997) Blind separation of convolved and delayed sources, Proc. Advances in Neural Information Processing Systems, pp. 758–764.

[8] S. Weiß, (1997) On adaptive filtering on oversampled subbands, Ph.D. dissertation, Signal Processing Division, University of Strathclyde.

[9] P. Smaragdis, (1998) Blind separation of convolved mixtures in the frequency domain, Neurocomputing, Vol. 22, pp. 21–34.

[10] L. Parra and C. Spence, (2000) Convolutive blind separation of non-stationary sources, IEEE Trans. On Speech and Audio Processing, Vol. 8, No. 3, pp. 320–327.

[11] H. Buchner, R. Aichner, and W. Kellerman, (2005) A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics, IEEE Trans. Speech and Audio Processing, Vol. 13, No. 1, pp. 120–134.

[12] A. Hiroe, (2006) Solution of permutation problem in frequency domain ICA using multivariate probability density functions, in Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation, pp. 601–608.

[13] K. Matsuoka and S. Nakashima, (2001) Minimal distortion principle for blind source separation, in Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation, pp. 722–727.

[14] T. Kim, (2010) Real-time independent vector analysis for convolutive blind source separation, IEEE Transactions on Circuits and Systems I, Vol. 57, No. 7, pp.1431–1438.

[15] S.-I. Amari, T.-P. Chen, and A. Cichocki, (2000) Nonholonomic orthogonal learning algorithms for blind source separation, Neural Computation, Vol. 12, pp. 1463-1484.

[16] J. Keshet, D. Grangier and S. Bengio, (2009) Discriminative Keyword Spotting, Speech Communication, Vol. 51, No. 4, pp. 317–329.

[17] M. Brandstein, and D. Ward, (2001) Microphone Arrays: Signal Processing Techniques and Applications, Springer-Verlag Berlin Heidelberg New York, pp.39–57.

[18] C. Zheng, Y. Zhou, X. Hu, and X. Li, (2011) Two-Channel Post-filtering Based on Adaptive Smoothing and Noise Properties, ICASSP, May, pp.1745–1748.

[19] H.J. Kwon, S.H. Jin, and N.S. Kim, (2008) Voice Activity Detection Based on Conditional MAP Criterion, IEEE Signal Processing Letters, Vol.15, pp. 257–260.

[20] B. Yegnanarayana, P. Satyanarayana, (2000) Enhancement of reverberant speech using LP residual signal, IEEE Trans. On Speech and Audio Processing, Vol. 8, Issue. 3, pp. 267-281.

[21] T. Nakatani, (2007) Harmonicity-based blind dereverberation for single-channel speech signal (HERB), IEEE Trans. On Audio, Speech, and Language Processing, Vol. 15, Issue. 1, pp. 80-95

[22] E.A.P. Habets and S. Gannot, (2007) Dual-Microphone Speech Dereverberation Using a Reference Signal, IEEE Int. Conf. on Acoustics, Speech and Signal Processing(ICASSP), pp. IV-901-IV-904.

[23] Y. Benayed, D. Fohr, J.-P Haton, G. Chollet, (2004) Confidence measure for keyword spotting using support vector machines. Proc. of International Conference on Audio, Speech and Signal Processing, pp. 588–591.

[24] H. Ketabdar, J. Vepa, S. Bengio, H. Bourlard, (2005) Posterior based keyword spotting with a priori thresholds. Int. Conf. on Machine Learning for Multimodal Interaction (MLMI), pp.633-636.

[25] M.-C. Silaghi, H. Bourlard, (1999) Iterative posterior-based keyword spotting without filler models. Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop, Keystone, USA, pp. 213–216.

[26] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Fapso, M. Karafiat, J. Cernocky, (2005) Comparison of keyword spotting approaches for informal continuous speech. Proc. of INTERSPEECH-2005, Lisbon. Portugal, pp. 633-636..

[27] M. Wolfel, and J. McDonough, (2009) Distant Speech Recognition, John Wiley & Sons, Ltd. pp.135-229.

[28] P.C. Loizou, (2007) Speech Enhancement: Theory and Practice, CRC Press, Taylor & Francis Group. an informa business, pp. 213-286.