2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

Two-Channel-Based Voice Activity Detection for Humanoid Robots in Noisy Home Environments

Hyun-Don Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno

Abstract—**The purpose of this research is to accurately classify the speech signals originating from the front even in noisy home environments. This ability can help robots to improve speech recognition and to spot keywords. We therefore developed a new voice activity detection (VAD) based on the complex spectrum circle centroid (CSCC) method. It can classify the speech signals that are received at the front of two microphones by comparing the spectral energy of observed signals with that of target signals estimated by CSCC. Also, it can work in real time without training filter coefficients beforehand even in noisy environments (SNR > 0 dB) and can cope with speech noises generated by audio-visual equipments such as televisions and audio devices. Since the CSCC method requires the directions of the noise signals, we also developed a sound source localization system integrated with cross-power spectrum phase (CSP) analysis and an expectation-maximization (EM) algorithm. This system was demonstrated to enable a robot to cope with multiple sound sources using two microphones.**

I. Introduction

Since we expect intelligent robots to participate widely in the near future society, effective interaction between them and us will be essential. For the purposes of natural human-robot interactions, they should firstly localize voices and faces in social and home environments to find and track their communication partners because people usually talk while looking at robots. Therefore, localization and tracking systems for voices and faces have been extensively studied and developed [1-3].

Robots then need a Voice Activity Detection (VAD) system that helps them to recognize speech well and correctly [4-8]. Although various voice activity detection (VAD) algorithms have been applied to such applications as speech recognition, speech enhancement, and speech coding, conventional VAD algorithms work poorly in extremely noisy environments and are unreliable in the presence of non-stationary or broad band speech-like noise [4-6]. Therefore, researchers have introduced multi-channel algorithms to improve VAD performance by exploiting the spatial selectivity [7,8]. Specifically, Le Bouquin et al. assumed that the spatial correlation between the disturbing noises was weak for all frequencies of interest while the speech signals were highly correlated [7]. However, this technique based on coherence function is usually difficult to cope with vocal noises generated by television sets or audio devices. Recently, although Hoffman et al. estimated the target-to-jammer ratio (TJR) using the generalized sidelobe canceller (GSC) as a measure for VAD [8], this way requires relatively many microphones and the training of adaptive filter coefficients to accurately estimate TJR.

In this paper, using two microphones, we developed a method that can accurately classify the speech signals originating from the front even in noisy home environments. It is realized by comparing the spectral energy of observed signals with that of target signals separated by complex spectrum circle centroid (CSCC) [9] method. The CSCC method which has recently been proposed utilizes geometric information of the target signal that should be received at the front of microphones and the observed signal obtained by microphones in a complex spectrum plane. It actually requires at least three microphones which are disposed on a straight line. However, since the form of a microphone array is difficult to be equipped with systems of various shapes such as robots, we used a new way that makes the CSCC method estimate the target signals using only two microphones. This method can reduce noise in real time without training beforehand and also achieve high performance. Although our VAD based on the CSCC method can only classify front target signals, this system may be suitable to communicate with someone because people usually talk while facing the communication target. The allowable range of target signals for our VAD is within about ±8°where 0°is the front of two microphones, the sampling rate is 16 kHz, and the distance between two microphones is 0.15 m (refer to Equation 3). This is because the target signals are available as long as the delay of arrival (DOA) between two microphones does not occur.

In addition, to use the CSCC method, we need two sound directions for noise and target signals. However, localizing several sound sources usually requires an array microphone and some methods require impulse response data. Thus, using two microphones, we developed a method based on probability for estimating the number and localization of sound sources. For our method, we first need to accumulate cross-power spectrum phase (CSP) analysis [10] results for three frames (shifting every half a frame). Then, the expectation-maximization (EM) algorithm [11] is used to estimate the distribution of the accumulated data. It can localize two sound sources using only two microphones, and

Hyund-Don Kim, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno are with Speech Media Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan (e-mail: {hyundon, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp).

it does not need impulse response data.

The rest of this paper is organized as follows. Section II describes the sound source localization that we developed. Section III describes sound classification using Gaussian Mixture Model (GMM) and also the VAD system based on the CSCC method. In Section IV, we applied our VAD to a humanoid robot and did experiments to detect the intervals of specific keywords in noisy environments. Section V concludes this paper.

II. Sound Source Localization

For sound source localization, the latest systems for robots mostly use one of three methods: head-related transfer function (HRTF) [1,12,13], multiple signal classification (MUSIC) [2,14], and CSP [10,15]. HRTF and MUSIC typically need impulse response data and an array of microphones in order to localize several sound sources. Impulse response data must thus be measured for every discrete azimuth and/or elevation before these methods can be applied to robots. Even though a lot of microphones and impulse response data would improve localization performance, they would also increase the calculation time. Furthermore, configuring the microphones in the robot would be problematic.

In contrast, CSP does not need impulse response data and can accurately determine the direction of a sound using only two microphones. Using CSP with two microphones can locate only one sound source each frame even if several sound sources are present. This is because CSP obtains the sound localization information from the spatial correlation between two signals. Besides, CSP is usually unreliable in noisy environments. To overcome these weaknesses, we developed a new method based on probability for estimating the number and location of sound sources. First, the CSP results for three frames (shifting every half frame) are collected. Then, an EM algorithm [11] is used to estimate the distribution of the data. In this way, our method can localize several sound sources using the distribution of CSP results and can reduce the error in sound source localization.

A. Cross-power Spectrum Phase analysis (CSP)

The direction of a sound source can be obtained by estimating the Time Delay Of Arrival (TDOA) between two microphones [3]. When there is a single sound source, the TDOA can be estimated by finding the maximum value of the cross-power spectrum phase (CSP) coefficients [10] derived by

$$csp_{ij}(k) = IFFT\left[\frac{FFT\left[s_i(n)\right]FFT\left[s_j(n)\right]^*}{\left|FFT\left[s_i(n)\right]\right|\left|FFT\left[s_j(n)\right]\right|}\right] \quad (1)$$

$$\tau = \arg\max\left(csp_{ij}(k)\right) \quad (2)$$

where k and n are the number of samplings for the delay of arrival between two microphones, $s_i(n)$ and $s_j(n)$ are signals entering into the microphone i and j respectively, FFT (or IFFT) is the fast Fourier transform (or inverse FFT), * is the complex conjugate, and τ is the estimated TDOA. The sound source direction is derived by

$$\theta = \cos^{-1}\left(\frac{v\cdot\tau}{d_{\max}\cdot F_s}\right) \quad (3)$$

where θ is the sound direction, v is the sound propagation speed, F_s is the sampling frequency, and d_{max} is the distance with the maximum time delay between two microphones. The sampling frequency of our system was 16 kHz.

B. Localization of multiple sound sources by EM

Figure 1 (A) shows the sound source localization events extracted by CSP according to time or frame lapses. We can see events that lasted 192 ms are used to train the EM algorithm to estimate the number and localization of sound sources. We experimentally decided that the appropriate interval for the EM algorithm was 192 ms [15]. Figure 1 (B) shows the training process for the EM algorithm to estimate the distribution of sound source localization events. Figure 1 (C) shows that the EM training results indicate the refined localizations of sound sources by iterating processes (A) and (B) in the same way. The interval for EM training is shifted every 32 ms.

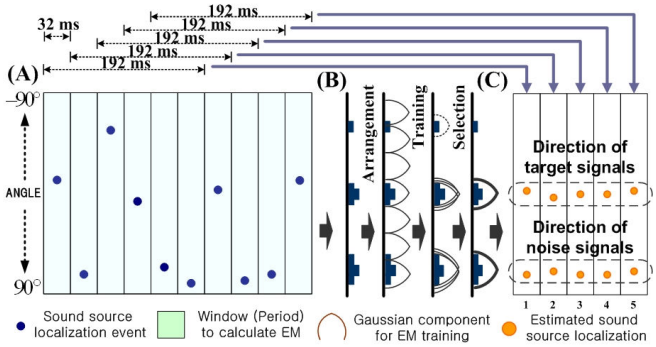

Fig. 1. Estimating localization of multiple sound sources.

Here, we explain the process of applying EM algorithm. Figure 2 describes the process in Figure 1 (B) in detail. In (A) of Figure 2, as the first step of EM training, sound source localization events were gathered for 192 ms. Next, Gaussian components defined by using equation (4) for training the EM algorithm were uniformly arranged on whole angles.

$$P\left(X_m\middle|\theta_k\right) = \frac{1}{\sqrt{2\pi\sigma_k^{\,2}}}e^{\frac{(X_m-\mu_k)^2}{2\sigma_k^{\,2}}} \quad (4)$$

where μ_k is the mean, $\sigma^2{}_k$ is the variance, θ_k is a parameter vector, m is the number of data, and k is the number of mixture components. At that time, in (A) of Figure 2, the μ and σ parameters in Gaussian components are the respective center and radius values of each component. Then, the sound localization events are applied to the arranged Gaussian components to find the parameter vector, θ_k,

describing each component density, $P(X_m|\theta_k)$, through iterations of the E and M steps. This EM step is described as follows:

1) E-step: The expectation step essentially computes the expected values of the indicators, $P(\theta_k|X_m)$, where each sound source localization event X_m is generated by component k. Given N is the number of mixture components, the current parameter estimates θ_k and weight w_k, using Bayes' Rule derived as

$$P(\theta_k|X_m) = \frac{P(X_m|\theta_k)\cdot w_k}{\sum_{k=1}^{N} P(X_m|\theta_k)\cdot w_k} \tag{5}$$

2) M-step: At the maximization step, we can compute the cluster parameters that maximize the likelihood of the data assuming that the current data distribution is correct. As a result, we can obtain the recomputed mean using Equation (6), the recomputed variance using Equation (7), and the recomputed mixture proportions (weight) using Equation (8). The total number of data is indicated by M.

$$\mu_k = \frac{\sum_{m=1}^{M} P(\theta_k|X_m) X_m}{\sum_{m=1}^{M} P(\theta_k|X_m)} \tag{6}$$

$$\sigma_k^2 = \frac{\sum_{m=1}^{M} P(\theta_k|X_m)\cdot (X_m-\mu_k)^2}{\sum_{m=1}^{M} P(\theta_k|X_m)} \tag{7}$$

$$w_k = \frac{1}{N}\sum_{m=1}^{M} P(\theta_k|X_m) \tag{8}$$

After the E and M steps are iterated an adequate number of times, the estimated mean, variance, and weight based on the current data distribution can be obtained.

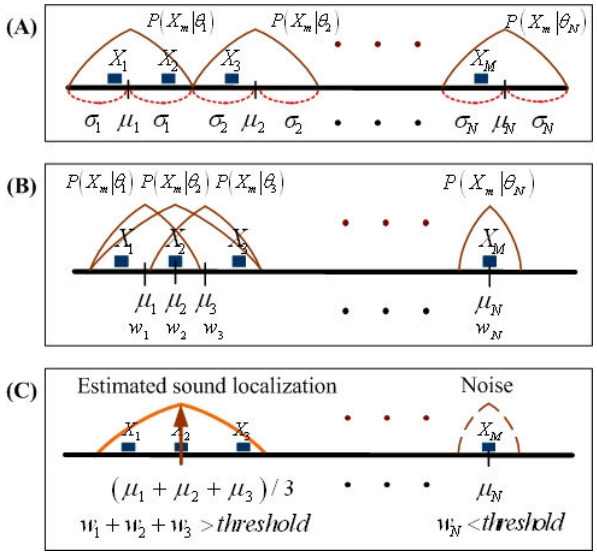

Fig. 2. Process of EM algorithm for estimating sound sources.

Then, in (B) of Figure 2, the weight and mean of Gaussian components are reallocated based on the density and distribution of the histogram data. Finally, in (C) of Figure 2, if the components overlap, each weight value of overlapping Gaussian components will be added. After that, if the weight value is higher than a threshold value, the system can determine the localization of the sound source by computing the average mean of the overlapping Gaussian components. In contrast, components with small weights are regarded as noise and will be removed.

C. Experiments and Results

To evaluate localization, we did an experiment observing conditions where two sound sources were 1.5 m from the head of a robot, and recorded female and male speech was simultaneously emitted from speakers for 7 sec at a magnitude of 85 dB. The symmetrical intervals between the two speakers were 60° (Experiment 1), 120° (Experiment 2), and 180° (Experiment 3) in Figure 3. The graphs show the results of sound source localization when there were two sound sources. The top graph plots the success rate, when the difference between the angle of speaker and observed angle was within 30°, for CSP with EM and HRTF and the bottom graph plots their average error. Our method, combining CSP and the EM algorithm, outperformed HRTF [1].

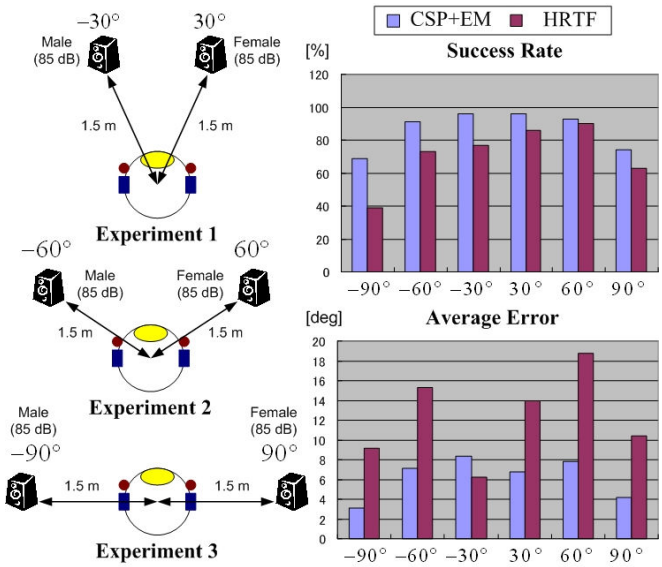

Fig. 3. Experimental conditions and results.

III. Voice Activity Detection

A. Sound Source Classification by GMM

Gaussian Mixture Model (GMM) is a powerful statistical method widely used for speech classification [5]. Here, we applied the 0 to 12th coefficients (total 13 values) and the Δ 1 to Δ 12th coefficients (total 12 values) of Mel Frequency Cepstral Coefficients (MFCCs) to GMM defined by Equation (9) and the weight as denoted by Equation (10).

$$P_{mixture}(X_{1\sim 25}|\theta_{1\sim 25}) = \sum_{L=1}^{25} P_L(X_L|\theta_L)\, w(L) \tag{9}$$

$$\sum_{L=1}^{25} w(L) = 1, \quad 0 \le w(L) \le 1 \tag{10}$$

where P is the component density function, L is the number of MFCC parameters, X is the value of the MFCC data of the 0 to 12th and the Δ 1 to Δ 12th coefficients, and θ is the parameter vector concerning each MFCC value. Moreover, to classify speech signals robustly, we designed two GMM models for speech and noise derived as

$$f = \log\left(P_s\left(X_s|\theta_s\right)\right) - \log\left(P_n\left(X_n|\theta_n\right)\right) \tag{11}$$

where P_s is the GMM related to speech, and X_s is the MFCC data set at the t-th frame belonging to the speech parameters, θ_s. On the other hand, P_n is the GMM related to noise and X_n is the MFCC data set at the t-th frame belonging to the noise parameters, θ_n. Finally, if the final value, f, denoted as Equation (12), is higher than the value of the threshold to discriminate the speech signal from GMM, signals at the t-th frame will be regarded as speech signals.

$$f_{\text{speech}} > threshold > f_{\text{noise}} \tag{12}$$

We used 30 speech data (15 males and 15 females) for the speech parameters to train the GMM parameters, and 77 noise data generated in home environments such as the sounds of a door opening or shutting and those of electrical home appliances (e.g., a vacuum cleaner, a hair drier, and a washing machine) for the noise parameters. To verify the performance of GMM parameter training, we classified the sound sources using speech and noise data for training. As a result, we obtained a success rate for speech classification of 95.5% and a success rate for noise classification of 72.8%.

B. Complex Spectrum Circle Centroid (CSCC)

To cope with vocal noises originating from the sides, we applied sound source separation (SSS) to our VAD. Two methods are commonly used for SSS. One is geometric source separation (GSS) and one of its well-known methods is as an adaptive beamformer [16]. This requires many microphones and prior training of the post-filter coefficients. The other is blind source separation (BSS) and it is well-known in independent component analysis (ICA) [17]. ICA is normally unsuitable in environments where the number of sound sources is dynamically changed because it is needs the same number of microphones as that of sound sources in principle. Also, to achieve high performance, ICA usually requires a large number of sampling data and much executing time. Therefore, we used the CSCC method because it can reduce noise in real time without training beforehand and also achieve high performance.

As seen in Figure 4, if the signals propagate as a plane wave, the spectrums of the signals observed using a 2-channel microphone are given as

$$M_1(\omega) = S(\omega) + N(\omega) \tag{13}$$

$$M_2(\omega) = S(\omega) + N(\omega)e^{-j\omega\tau} \tag{14}$$

where $M_1(w)$ and $M_2(w)$ are the spectrums of the observed signal, and $S(w)$ and $N(w)$ denote the respective spectrums of the target signal and the noise signal. The value τ denotes the time delay between the two microphones in respect to the noise signal.

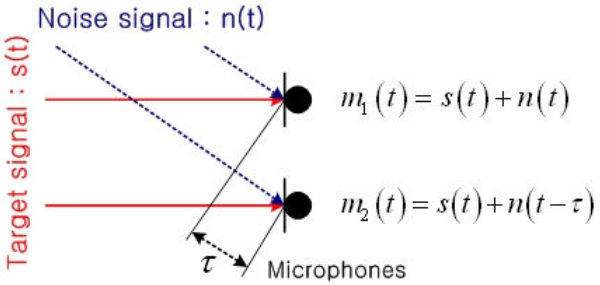

Fig. 4. Signal propagating toward two microphones.

As seen in Figure 5, $S(w)$ is located at an equal distance from $M_1(w)$ and $M_2(w)$, and the distance is $N(w)$. Subtracting Equation (14) from Equation (13) gives the value of $N(w)$ as

$$\left\|N(\omega)\right\| = \frac{\left\|M_1(\omega) - M_2(\omega)\right\|}{\left\|1 - e^{-j\omega\tau}\right\|} \tag{15}$$

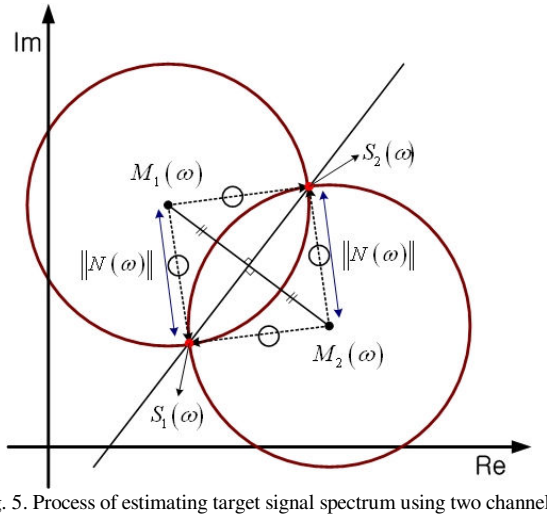

Fig. 5. Process of estimating target signal spectrum using two channels.

Figure 5 outlines the process used to estimate $S(w)$ using two microphones. First, we draw a perpendicular bisector toward a straight line connecting $M_1(w)$ and $M_2(w)$ in a complex spectrum plane. Next, we draw a circle with the radius of $N(w)$ shown in Equation (15) and its center at $M_1(w)$. The coordinates of each spectrum in Figure 5 are defined as

1) The spectrum of the observed signal:

$$M_1(\omega) = \left(M_{1x}, M_{1y}\right),\ M_2(\omega) = \left(M_{2x}, M_{2y}\right) \tag{16}$$

2) The candidate for the target signal spectrum:

$$\tilde{S}(\omega) = \left\{S_1(\omega), S_2(\omega)\right\} = \left\{\left(S_{1x}, S_{1y}\right), \left(S_{2x}, S_{2y}\right)\right\} \tag{17}$$

3) The midpoint:

$$C(\omega) = \left(C_x, C_y\right) = \left(\frac{M_{1x} + M_{2x}}{2}, \frac{M_{1y} + M_{2y}}{2}\right) \tag{18}$$

where subscript x and y correspond to the coordinates of the real and imaginary parts respectively.

The perpendicular bisector and the circle are given as

$$\tilde{S}_y(\omega) - C_y(\omega) = \frac{M_{1x}(\omega) - M_{2x}(\omega)}{M_{2y}(\omega) - M_{1y}(\omega)} \cdot \left(\tilde{S}_x(\omega) - C_x(\omega)\right) \quad (19)$$

$$\left(\tilde{S}_x(\omega) - M_{1x}(\omega)\right)^2 + \left(\tilde{S}_y(\omega) - M_{1y}(\omega)\right)^2 = \left\|N(\omega)\right\|^2 \quad (20)$$

The spectrum of the target signal, $S(w)$, is located at the intersection of the perpendicular bisector and the circle. Hence, $S_1(w)$ and $S_2(w)$ are obtained by solving the simultaneous formulae between Equation (19) and Equation (20). Actually, the CSCC method needs at least three microphones to estimate the accurate target signal. However, since we used only two microphones, we must choose the most appropriate spectrum from the two candidates for the target signal. Here, we chose the candidate whose spectrum power was smaller, since we considered that the power of the estimated clean signal would be smaller than that of the observed noisy signal. In the case in Figure 5, $S_1(w)$ was chosen as the target signal spectrum.

C. Speech Classification based on CSCC

To classify the speech signals of a communication partner who is in front of a robot's face (i.e., speech signals arriving at two channels simultaneously without delay), we classified them after CSCC had reduced the noise signals that had arrived from the side of the robot's face. In particular, to classify the interval of target signals using CSCC, we first had to obtain the various types of frame energies in the frequency domain. The frame energies in the frequency domain of all types are defined as

1) The spectral frame energies of target and observed signals:

$$E_{\text{target}} = \frac{1}{N}\sum_{\omega=0}^{N}\left|S_{\text{target}}(\omega)\right|, \quad E_c = \frac{1}{N}\sum_{\omega=0}^{N}\left|C(\omega)\right| \quad (21)$$

2) The spectral frame energies observed from microphone 1 and 2:

$$E_{m1} = \frac{1}{N}\sum_{\omega=0}^{N}\left|M_1(\omega)\right|, \quad E_{m2} = \frac{1}{N}\sum_{\omega=0}^{N}\left|M_2(\omega)\right| \quad (22)$$

where w is the frequency value of FFT, N is the order of FFT, and $S_{\text{target}}(w)$ is the target signal spectrum separated by CSCC. Here, $M_1(w)$ is the signal spectrum observed from microphone 1, $M_2(w)$ is the signal spectrum observed from microphone 2, and $C(w)$ is the observed signal spectrum calculated by Equation (18).

Next, we can detect the interval of target signals coming from the front as follows. First, if there are noise signals coming from the side, the frame energy of the separated target signals will be less than that of the observed signals. This condition is defined in Equation (23). Second, as the definition of Equation (24), we can determine whether noise signals are coming from the side if the frame energy observed from both microphones is more than that of the observed signals.

$$E_c / E_{\text{target}} > threshold \quad (23)$$

$$thr_{Low} < \left(E_{m1}/E_c - E_{m2}/E_c\right) < thr_{High} \quad (24)$$

Finally, we have to classify whether the target signals are speech or not using Equation (12).

D. Experiments and Results

We used two metrics to evaluate our VAD in noisy environments. These were the speech hit rate (SHR) and non-speech hit rate (NSHR) defined as

$$SHR = \frac{N_S}{N_{Sref}}, \quad NSHR = \frac{N_N}{N_{Nref}} \quad (25)$$

where N_S and N_{Rref} are the numbers of all speech samples correctly detected and real speech in the whole database, and N_N and N_{Nref} are the numbers of all non-speech samples correctly detected and real non-speech in the whole database.

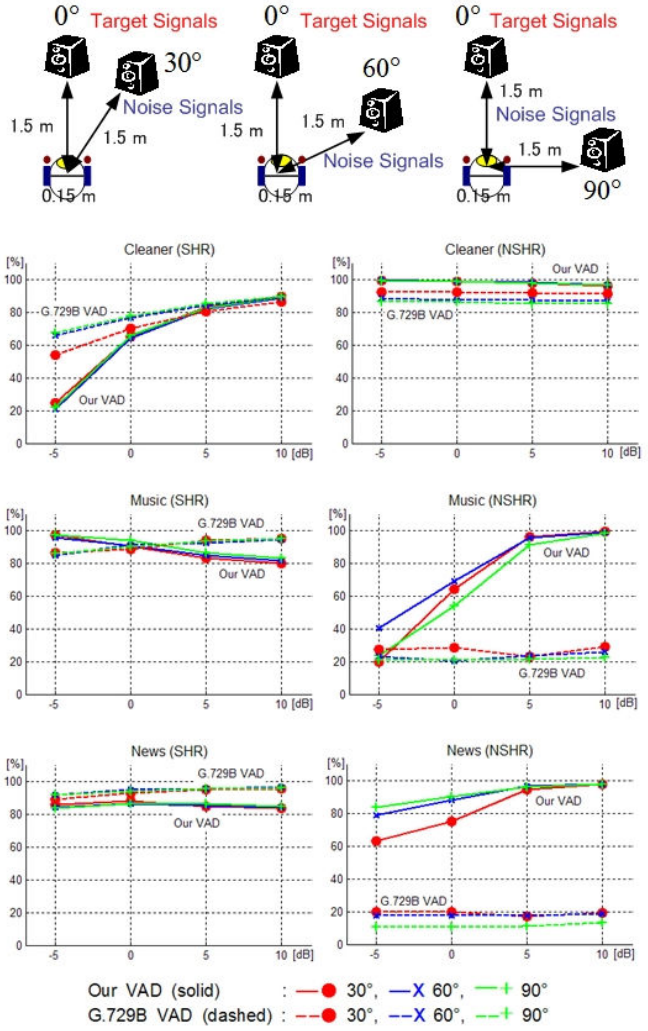

Fig. 6. Experiments and results of VAD based on CSCC.

We conducted experiments under the following conditions. We used two omnidirectional microphones installed at the left and right ear positions of the humanoid robot SIG2 [15]. The distance between two microphones was 0.15 m. The sampling rate is 16 kHz and 1024-point FFT is applied to the windowed data with 512 sample overlap. As shown at the top of Figure 6, the target signals and noise signals were 1.5 m

from two microphones. The target signals were in front of the microphones, and the noise signals were at 30°, 60°, or 90° from the side. Two loud sounds were simultaneously emitted from two speakers for 30 sec. We used 10 speech data (for 5 men and 5 women) for target signals, and 3 noise data (vacuum cleaner, television news, and contemporary pop music including vocals). The words of a numeral one to a numeral ten in Japanese were randomly recorded for each target signal data for 30 sec. The signal to noise ratios (SNRs) were -5, 0, 5, or 10 dB.

Figure 6 shows the performance results for our VAD algorithm compared to G.729 Annex B VAD [6], which the International Telecommunication Union (ITU-T) adopted. The standard G.729B VAD makes a voice activity decision every 10 ms, and its parameters are the full band energy, the low band energy, the zero-crossing rate and the spectral measure. Here, since G.729B is the one-channel-based VAD, we obtained the performance results for the G.729B VAD after averaging the results obtained by the left and right microphones.

At vacuum cleaner noise in Figure 6, SHR of our VAD was similar to that of G.729B VAD and NSHR of our VAD was better than that of G.729B VAD. Especially, the G.729B VAD performed poorly non-speech detection accuracy (NSHR) with the vocal noise (music and TV news) while speech detection accuracy (SHR) was good (higher than 90%). This is because the G.729B VAD regarded noises containing vocal signals as speech signals. On the other hand, at noise containing vocal signals, SHR of our VAD was better than about 85% for all SNRs, and NSHR of our VAD was considerably better than that of the G.729B VAD. NSHR was better than 80% except for at -5 and 0 dB SNR for music noise and for at 30° at -5 and 0 dB SNR for TV news noise. Our system can thus usually be used at SNRs larger than 0 dB regardless of the kinds of noise signals.

IV. Voice Activity Detection for Humanoid Robots

A. System Overview

Figure 7 shows the overview of structure of our VAD system based on the CSCC method and the photograph of a humanoid robot called SIG2. The robot has two omni-directional microphones inside humanoid ears at the left and right ear positions. First, to use the CSCC method, the robot needs the direction of noise signals. Therefore, we localized sound sources by combining the CSP method with the EM algorithm as discussed in Section II. Then, after finding the direction of noise signals, the CSCC method can reduce the noise signals from the target signals. Also, as discussed in Section III, the robot is able to determine whether target signals exist or not and whether the target signals are voice or not through CSCC and GMM, respectively. Finally, after VAD has counted the voice frames for 192 ms, it can determine the appropriate interval for speech spoken by the communication partner. This process for VAD iterates every 32 ms. The computer we used followed this specification, Celeron 2.4 GHz, 512 M ram.

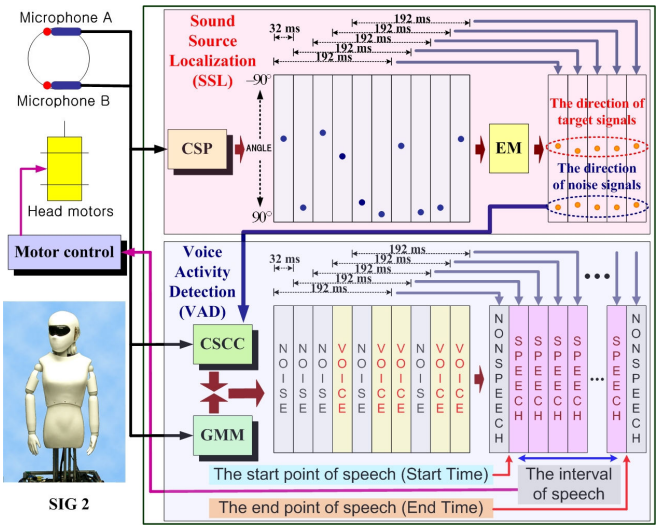

Fig. 7. System overview for the keyword length detection.

B. Experiments and Results

The goal of this paper was to accurately detect the intervals of specific keywords generated from the front of the robot even in noisy home environments. This is because people naturally look at robot's faces in order to communicate with them. If the robot is also able to classify the length of keywords that the communication partner spoke even in a noisy environment, this ability will help robots to improve its speech recognition and to spot the specific command for a keyword. To verify our system's feasibility, we applied the VAD we developed to a humanoid robot, SIG2, and we recorded two commands, "sig" and "ohayogozaimas", as specific keywords. The Japanese command for "ohayogozaimas" means "Good morning" in English. For the experiment, three sounds (vacuum cleaner, TV news, and pop music) were generated by the side speaker at 30°, 60°, and 90°. The target and noise signals were simultaneously emitted ten times at a magnitude of 90 dB every item on the Table I. Table I lists the experimental results that show the good performance of the robot in detecting the interval of two commands emitted by the front speaker. Detecting two commands was almost perfect except for the item at 30° and Cleaner. This is because GMM could not classify the speech signals well due to the close gap between speech and noise signals. In addition, the average intervals of detected commands were similar to original intervals for "sig" and "ohayogozaimas" whose lengths were about 1.5 and 1.8 sec, respectively. Also, the standard deviations of detected command intervals were usually within 0.1 sec. Figure 8 shows snap-shots of the robot detecting intervals of specific keywords. A in Figure 8 shows that the robot has neglected noise signals generated from its side, and B and C in Figure 8 show that the robot nodded when detecting the keywords with

the length for about 1.5 sec concerning "sig" (C shows when the robot detected the keyword length where noise signals have occurred). D in Figure 8, the robot tilted its head when detecting the keywords with the length for about 1.8 sec concerning "ohayogozaimas".

TABLE I
THE RESULTS OF DETECTING COMMAND INTERVALS

CMD	"sig" (1.5 sec)			"ohayogozaimas" (1.8 sec)		
Angle	**30°**	**60°**	**90°**	**30°**	**60°**	**90°**
The success rate of VAD [%]						
Cleaner	50	90	100	50	100	100
News	90	100	100	100	100	100
Music	100	100	100	100	100	100
The average intervals of commands detected by VAD [sec]						
Cleaner	1.38	1.52	1.55	1.68	1.71	1.89
News	1.55	1.52	1.59	1.84	1.85	1.88
Music	1.54	1.55	1.58	1.85	1.88	1.90
The standard deviation of detected intervals [sec]						
Cleaner	0.054	0.06	0.138	0.094	0.221	0.06
News	0.036	0.067	0.059	0.093	0.082	0.039
Music	0.086	0.06	0.04	0.045	0.041	0.034

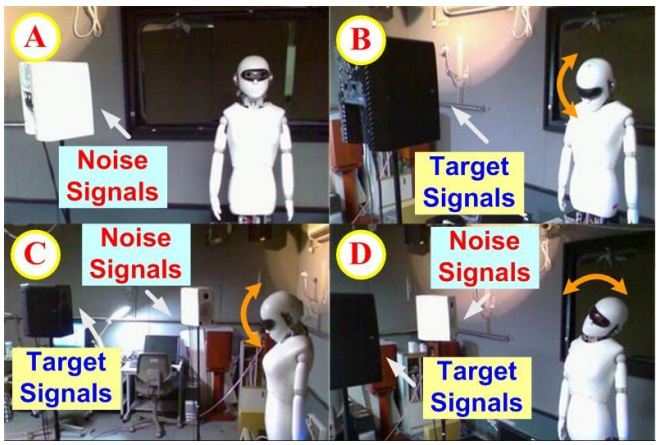

Fig. 8. Snap-shots when the robot detects specific intervals of speech.

V. CONCLUSION

We developed the VAD system that enables robots to accurately detect the intervals of specific keywords or commands generated in front of them even in noisy home environments and confirmed that it performed well. Our system has some principle capabilities. First, the VAD we developed can classify the intervals of speech arriving from the front in real-time even where there is speech competing. Also, our results indicated that our system can reliably classify the intervals of speech in noisy environments larger than SNR 0 dB. Second, since it can work using only two channels and a normal sound card device, it can be used in various kinds of robots and systems. Our system combining the CSP method and the EM algorithm can localize several sound sources despite only having two microphones and does not use impulse response data. Finally, in the next step, we are considering adding a speech recognition engine to our VAD system because robots must also be able to recognize the meaning of keywords or commands.

ACKNOWLEDGMENT

This research was partially supported by MEXT, Grant-in-Aid for Scientific Research, and Global COE program of MEXT, Japan.

REFERENCES

[1] Kazuhiro Nakadai, Ken-ichi Hidai, Hiroshi Mizoguchi, Hiroshi G. Okuno, and Hiroaki Kitano, "Real-Time Auditory and Visual Multiple-Object Tracking for Humanoids," in Proc. of 17th International Joint Conference on Artificial Intelligence (IJCAI-01), Seattle, Aug. (2001) pp. 1425-1432.

[2] I. Hara, F. Asano, Y. Kawai, F. Kanehiro, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004), Oct. (2004) pp. 2404-2410.

[3] H-D. Kim, J. S. Choi, and M. S. Kim, "Speaker localization among multi-faces in noisy environment by audio-visual integration", in Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA2006), May (2006) pp. 1305-1310.

[4] L. Lu, H. J. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation," IEEE Trans. on Speech and Audio Processing, vol. 10, no 7, pp. 504-516, 2002.

[5] M. Bahoura and C. Pelletier, "Respiratory Sound Classification using Cepstral Analysis and Gaussian Mixture Models," IEEE/EMBS, pp. 9-12, Sep. 2004.

[6] ITU-T, "A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70," ITU-T Rec. G.792, Annex B, 1996.

[7] R. Le Bouquin and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," Speech communication vol. 16, pp. 245-254, 1995.

[8] M. Hoffman, Z. Li, and D. Khataniar, "GSC-based spartial voice activity detection for enhanced speech coding in the presence of competing speech," IEEE Trans. on Speech and Audio Processing, vol. 9, no. 2, pp. 175-179, March 2001.

[9] T. Ohkubo, T. Takiguchi, and Y. Ariki, "Two-Channel-Based Noise Reduction in a Complex Spectrum Plane for Hands-Free Communication System," Journal of VLSI Signal Processing Systems 2007, Springer, Vol. 46, Issue 2-3, pp. 123-131, March 2007.

[10] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of multiple sound sources based on a CSP analysis with a microphone array," IEEE/ICASSP Int. Conf. Acoustics, Speech, and Signal Processing, June (2000) pp 1053-1056.

[11] T. K. Moon. "The Expectation-Maximization algorithm," IEEE Signal Processing Magazine, Nov. (1996) 13(6) pp. 47-60.

[12] C. I. Cheng & G. H. Wakefield, "Introduction to Head-Related transfer Functions (HRTFs): Space," Journal of the Audio Engineering Society, vol. 49, no. 4, pp.231-248, 2001.

[13] S. Hwang, Y. Park, and Y. Park, "Sound Source Localization using HRTF database," in Proc. Int. Conf. on Control, Automation, and Systems (ICCAS2005), June, 2005, pp.751-755.

[14] R. O. Schmidt, "Multiple Emitter Location and Signals Parameter Estimation," IEEE Trans. Antennas Propag., AP-34, 1986, 276-280.

[15] H. D. Kim, K. Komatani, T. Ogata, and H. G. Okuno, "Auditory and Visual Integration based Localization and Tracking of Multiple Moving Sounds in Daily-life Environments," Proc. of IEEE/ROMAN Aug. (2007), pp. 399-404.

[16] J-M. Valin, J. Rouat, and F. Michaud, "Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2004), Sep. (2004) pp. 2123-2128.

[17] 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 Eras," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2006), Sep. (2006) pp. 878-885.