2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

# Target Speech Detection and Separation for Humanoid Robots in Sparse Dialogue with Noisy Home Environments

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

*Abstract*— **In normal human communication, people face the speaker when listening and usually pay attention to the speaker' face. Therefore, in robot audition, the recognition of the front talker is critical for smooth interactions. This paper presents an enhanced speech detection method for a humanoid robot that can separate and recognize speech signals originating from the front even in noisy home environments. The robot audition system consists of a new type of voice activity detection (VAD) based on the complex spectrum circle centroid (CSCC) method and a maximum signal-to-noise (Max-SNR) beamformer. This VAD based on CSCC can classify speech signals that are retrieved at the frontal region of two microphones embedded on the robot. The system works in real-time without needing training filter coefficients given in advance even in a noisy environment (SNR > 0 dB). It can cope with speech noise generated from televisions and audio devices that does not originate from the center. Experiments using a humanoid robot, SIG2, with two microphones showed that our system enhanced extracted target speech signals more than 12 dB (SNR) and the success rate of automatic speech recognition for Japanese words was increased about 17 points.**

## I. Introduction

Up to a few years ago, most robot applications were related to industries and manufacturing, and most robots in general use were industrial robots. Today, robots work in almost all fields of service, ranging from housekeeping to high technology space exploration, and robot technology has had a significant impact on daily life. Recently, service provided by humanoid robots has received an increasing amount of attention. But, although humanoid robots are increasingly expected to possess perceptual capabilities similar to those of humans due to an increasing demand for symbiotic interaction between humans and robots, the ability of robots in this respect is still very lacking. Since we expect intelligent robots to participate widely in the society of the near future, effective interaction between them and humans will be essential. To facilitate natural human-robot interaction, robots should firstly be able to localize voices and faces in social and home environments so that they can find and track their communication partner, this is important because people usually look directly at robots while addressing them. Therefore, localization and tracking systems for voices and faces have been extensively studied and developed [1-3]. Regarding, the human-robot interaction for speech, robots have to be able to separate the voices of participants in a conversation or an actual meeting where sources are sometimes active but silent. In practice, humanoid robots will often be confronted with sparse dialogue in noisy home environments.

In this paper, we consider the detection and separation of speech signals for the purpose of improving speech recognition or spotting keywords from the point of view of humanoid robots with two microphones. To realize this, voice activity detection (VAD) and sound source separation are essential for robots to communicate with people in real environments. Therefore, we developed a system which has some primary capabilities:

1) Our VAD can accurately classify a target speech originating from the front in real time even in noisy environments (SNR > 0 dB).
2) It can cope with vocal noise generated by television sets or audio devices in home environments.
3) The two microphones of our system can enhance detected target speech even when interference noise occurs or varies.

First, using two microphones, we developed a method that can accurately classify speech signals originating from the front even in noisy home environments. This was realized by comparing the spectral energy of observed signals with that of target signals separated by the complex spectrum circle centroid (CSCC) [9] method. Although our VAD based on the CSCC method can only classify frontal target signals, this system may be suitable for communicating with a person because people usually face the communication target while talking. The allowable range of target signals for our VAD is about ±8°, where 0° denotes a position directly in front of the two microphones. The sampling rate is 16 kHz, and the distance between the two microphones is 0.15 m. This separation distance is important because the target signals can be obtained as long as no delay of arrival (DOA) occurs between the two microphones.

Second, robots need also to recognize speech and/or keywords for communication with a person. However, since there are various sources of background noise in real home environments, such as music, TV audio, and unrelated dialogue, robots have to be able to separate and recognize detected target speech. Therefore, we used a max-SNR beamformer [10,11] to reduce noise because it does not need

* 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).

** Jinsung Kim is with the Center for Cognitive Robotics Research, Korea Institute of Science and Technology, Seoul, Korea, Republic (e-mail: jinsung@kist.re.kr).

information on the target location while most widely used beamformers [11,12] need that. It is also an effective means to separate the target speech when the target and noise signals are sometimes both active, as is sometimes the case in sparse dialogue. Since our system with two microphones can enhance the target speech detected by our VAD, it improves robot speech recognition. All of the above methods run on a PC equipped with a Celeron 2.4 GHz CPU and 512 RAM.

The rest of this paper is organized as follows. Section II describes the VAD system based on the CSCC method and sound classification using the Gaussian Mixture Model (GMM). Section III describes the max-SNR beamformer. Section IV, looks at the experiments, we did on the recognition of specific keywords after the detection and separation of their intervals in the presence of noise using the VAD system and max-SNR beamformer. Section V concludes the paper.

## II. Voice Activity Detection based on CSCC

Although various VAD algorithms have been applied for applications such as speech recognition, speech enhancement, and speech coding [4-8], 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, multi-channel algorithms have been introduced to improve VAD performance by exploiting spatial selectivity [7,8]. Specifically, Le Bouquin et al. assumed that the spatial correlation between disturbing noises was weak for all frequencies of interest while speech signals were highly correlated [7]. However, this technique based on a coherence function usually has difficulty coping with vocal noise generated by television sets or audio devices. Although Hoffman et al. recently estimated the target-to-jammer ratio (TJR) using the generalized sidelobe canceller (GSC) as a measure for VAD [8], this requires relatively many microphones and training of adaptive filter coefficients to accurately estimate TJR.

To overcome these problems, we applied a complex spectrum circle centroid (CSCC) method to our VAD. The CSCC method uses geometric information regarding the target signal that should be received from in front of the microphones and the observed signal obtained by the microphones in a complex spectrum plane. It typically requires at least three microphones disposed in a straight line. However, since this form of a microphone array is difficult to install in systems of various shapes, such as robots, we developed a way to enable the CSCC method to estimate target signals using only two microphones. This method can reduce noise in real time without training beforehand while still enabling achieve high performance. In addition, to use the CSCC method, we need two sound directions for noise and target signals. Thus, using two microphones, we developed a method based on probability to estimate the number and localization of sound sources. We do not describe this method here, but a full description is available elsewhere [3].

### A. Complex Spectrum Circle Centroid (CSCC)

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

$$M_1(\omega) = S(\omega) + N(\omega) \tag{1}$$

$$M_2(\omega) = S(\omega) + N(\omega)e^{-j\omega\tau} \tag{2}$$

where $M_1(\omega)$ and $M_2(\omega)$ are the spectrums of the observed signals, and $S(\omega)$ and $N(\omega)$ denote the respective spectrums of the target signals and the noise signals. The value $\tau$ denotes the time delay between the two microphones with respect to the noise signal. As shown in Figure 1, $S(\omega)$ is located at an equal distance from $M_1(\omega)$ and $M_2(\omega)$, and the distance is $N(\omega)$. Subtracting Equation (2) from Equation (1) gives the value of $N(\omega)$ as

$$\left\|N(\omega)\right\| = \frac{\left\|M_1(\omega) - M_2(\omega)\right\|}{\left\|1 - e^{-j\omega\tau}\right\|} \tag{3}$$

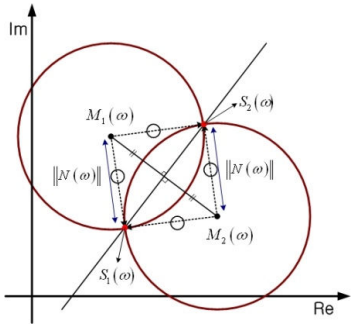


**Fig. 1.** Estimating target signal spectrum using two channels.

Figure 1 outlines the process used to estimate $S(\omega)$ using two microphones. First, we draw a perpendicular bisector toward a straight line connecting $M_1(\omega)$ and $M_2(\omega)$ in a complex spectrum plane. Next, we draw a circle with the radius of $N(\omega)$ shown in Equation (3) and its center at $M_1(\omega)$. The coordinates of each spectrum in Figure 1 are defined as

*1)* The spectrum of the observed signal:

$$M_1(\omega) = (M_{1x}, M_{1y}),\ M_2(\omega) = (M_{2x}, M_{2y}) \tag{4}$$

*2)* The candidate for the target signal spectrum:

$$\tilde{S}(\omega) = \{S_1(\omega), S_2(\omega)\} = \{(S_{1x}, S_{1y}), (S_{2x}, S_{2y})\} \tag{5}$$

*3)* The midpoint:

$$C(\omega) = (C_x, C_y) = \left(\frac{M_{1x} + M_{2x}}{2}, \frac{M_{1y} + M_{2y}}{2}\right) \tag{6}$$

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

$$\tilde{S}_y(\omega) - C_y(\omega) = \frac{M_{1x}(\omega) - M_{2x}(\omega)}{M_{2y}(\omega) - M_{1y}(\omega)} \cdot \left(\tilde{S}_x(\omega) - C_x(\omega)\right) \tag{7}$$

$$\left(\tilde{S}_x(\omega)-M_{1x}(\omega)\right)^2+\left(\tilde{S}_y(\omega)-M_{1y}(\omega)\right)^2=\left\|N(\omega)\right\|^2 \quad (8)$$

The spectrum of the target signal, $S(\omega)$, is located at the intersection of the perpendicular bisector and the circle. Hence, $S_1(\omega)$ and $S_2(\omega)$ are obtained by solving the simultaneous formulae between Equation (7) and Equation (8). Actually, the CSCC method needs at least three microphones to accurately estimate the 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 shown in Figure 1, $S_1(\omega)$ was chosen as the target signal spectrum.

### B. Sound Source Classification by GMM

The Gaussian Mixture Model (GMM) is a powerful statistical method widely used for speech classification [5]. Here, we applied the 0 to 12th coefficients (a total of 13 values) and the Δ1 to Δ12th coefficients (a total of 12 values) of Mel Frequency Cepstral Coefficients (MFCCs) to the GMM defined by Equation (9) and the weights as denoted by Equation (10).

$$P_{mixture}\left(X_{1\sim25}\middle|\theta_{1\sim25}\right)=\sum_{L=1}^{25}P_L\left(X_L\middle|\theta_L\right)w(L) \quad (9)$$

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

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 Δ1 to Δ12th coefficients, and $\theta$ 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

$$f=\log\left(P_s\left(X_s\middle|\theta_s\right)\right)-\log\left(P_n\left(X_n\middle|\theta_n\right)\right) \quad (11)$$

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 parameter, $\theta_s$. 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 parameter, $\theta_n$. If the final value, $f$, denoted as Equation (11), is higher than the value of the threshold to discriminate the speech signal from the GMM, signals in the $t$-th frame will be regarded as speech signals.

$$\begin{aligned}&\text{If } f(t)>threshold \text{ then } f(t)=1\ (speech)\\ &\qquad\text{else } f(t)=0\ (noise)\end{aligned} \quad (12)$$

We used 30 speech samples (from 15 males and 15 females) as speech parameters to train the GMM parameters, and 77 noise samples 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 quality of the GMM parameter training, we classified the sound sources using speech and noise data for training. We obtained a success rate for speech classification of 95.5% and a success rate for noise classification of 72.8%.

### C. Design of Voice Activity Detection

To classify the speech signals of a communication partner who is directly in front of a robot, we classified the signals after CSCC had reduced the noise signals which arrived from directions other than directly ahead of the robot. Since the two microphones were installed in the robot's head, rather than in a free space, CSCC had to take into account the effect of diffraction by the robot head's cover. We assumed that there was no diffraction effect for the center direction because speech signals arrived at the two channels at the same time without delay. To make allowance for errors generated from diffraction of noise signals arriving from the side directions, we considered the frame energy greater than the thresholds we experimentally determined as shown in Equation (15). 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 observed from microphones 1 and 2:

$$E_{m1}=\frac{1}{N}\sum_{\omega=0}^{N}\left|M_1(\omega)\right|,\quad E_{m2}=\frac{1}{N}\sum_{\omega=0}^{N}\left|M_2(\omega)\right| \quad (13)$$

*2)* The spectral frame energies of the target and the average frame energy between $E_{m1}$ and $E_{m2}$:

$$E_{\text{target}}=\frac{1}{N}\sum_{\omega=0}^{N}\left|S_{\text{target}}(\omega)\right|,\quad E_{ave}=\frac{E_{m1}+E_{m2}}{2} \quad (14)$$

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

Next, we can detect the interval of target signals coming from the front as follows. First, if we assume the robot knows the direction of noise signals coming from the side, the frame energy of the separated target signals will be less than that of the observed signals as defined in Equation (15). Second, from the definition of Equation (16), we can determine that noise signals are not coming from the side, and that there are target signals coming from the front if the difference of frame energy between both microphones is almost the same.

$$E_{ave}/E_{\text{target}}>threshold \quad (15)$$

$$thr_{Low}<\left(E_{m1}/E_{ave}-E_{m2}/E_{ave}\right)<thr_{High} \quad (16)$$

Here, if equations (15) and (16) are not satisfied, even where the $t$-th frame includes speech signals, $f(t)$=1, the robot will change the value of $f(t)$ into 0. This is because speech or noise signals in the $t$-th frame were generated from the side. We can then detect the period of the target speech by using

$$\sum_{i=i-n_a}^{i+n_b} f(i) \geq \text{threshold} \tag{17}$$

where $f(i)$ is the $i$-th speech frame classified by equation (12). Finally, if some speech frames exist within the interval of designated frames from the $n_a$-th frame to the $n_b$-frame, we can determine that the $i$-th frame is within the interval of the target speech.

## III. Enhanced Target Speech

To separate target speech signals after VAD is used to detect signals, including noise, we applied a max-SNR beamformer [10,11] to a humanoid robot called SIG2. Two methods are commonly used for sound source separation (SSS). One is geometric source separation (GSS) and one of its well-known methods is as an adaptive beamformer [12]. This requires many microphones and prior information on the target location. The other method is blind source separation (BSS), which is widely used in independent component analysis (ICA) [13]. ICA is normally unsuitable for environments where the number of sound sources dynamically changes because in principle the required number of microphones is equal to the number of sound sources. In addition, to achieve high performance, ICA usually requires a large quantity of sampling data and assumes that most of the data includes mixing with noise. In contrast, the max-SNR beamformer can effectively reduce noise even if there are only a few microphones and where speech and/or noise signals sometimes occur, but training of the max-SNR beamformer weights (refer to Equation 28) is required beforehand.

### A. Maximum SNR Beamformer

We took a time-frequency domain approach. Suppose that speech sources are convolutedly mixed and observed at microphones with a short-time Fourier transform (STFT):

$$x_j(f,\tau) = \sum_{k=1}^{N} h_{jk}(f) s_k(f,\tau) + n_j(f,\tau) \tag{18}$$

where $h_{jk}(f)$ is the frequency response from source $k$ to sensor $j$, $s_k(f,\tau)$ and $n_j(f,\tau)$ are the STFTs of a source $s_k$ and noise $n_j$, respectively. $f \in \{0,(1/T)f_s,\cdots,\{(T-1)/T\}f_s\}$ is a frequency ($f_s$ is the sampling frequency) and $\tau$ (=1, … , $K$) is a time-frame index. The vectors are $X=[x_1, \ldots, x_M]^T$, $\mathbf{h}_k=[h_{1k}, \ldots, h_{Mk}]^T$ and $\mathbf{n}=[n_1, \ldots, n_M]^T$.

The max-SNR beamformer maximizes the ratio between the output powers for the target-active and the target-silent periods. When such a beamformer $W_k(f)$ is obtained for source $k$, the $k$-th output signal can be obtained by

$$y_k(f,\tau) = W_k^H(f) X(f,\tau) \tag{19}$$

Let $P=\{1, \ldots, K\}$ be the whole period of $K$ observations $X(f,1),\ldots,X(f,K)$ at each frequency, $P_T^k \subset P$ be the target-active period when the target source $s_k$ is active, and $P_I^k \subset P$ be the target-silent period when the target $s_k$ is NOT active but interference and noise may be active. In this paper, we assume $P_T^k \cup P_I^k = P$.

The design criterion for the beamformer $W_k(f)$ is to maximize the ratio $\lambda(f)$ of the output power between the target-only period $P_T^k$ and the interference-and-noise-only period $P_I^k$:

$$\lambda(f) = \frac{\varepsilon\{|Y_k(f,\tau)|^2\}_{P_T^k}}{\varepsilon\{|Y_k(f,\tau)|^2\}_{P_I^k}} = \frac{W_k^H(f) R_T^k(f) W_k(f)}{W_k^H(f) R_I^k(f) W_k(f)} \tag{20}$$

where $R_T^k(f)$ and $R_I^k(f)$ are the correlation matrices of observations

$$R_T^k = \varepsilon\{X(f,\tau)X(f,\tau)\}_{P_T^k} = \frac{1}{|P_T^k|}\sum_{\tau\in P_T^k} X(f,\tau)X^H(f,\tau) \tag{21}$$

$$R_I^k = \varepsilon\{X(f,\tau)X(f,\tau)\}_{P_I^k} = \frac{1}{|P_I^k|}\sum_{\tau\in P_I^k} X(f,\tau)X^H(f,\tau) \tag{22}$$

where $|P|$ denotes the number of elements of the set $|P|$.

By differentiating $\lambda(f)$ with $W_k(f)$ and setting it to 0, we have

$$R_T^k(f) W_k(f) = \lambda(f) R_I^k(f) W_k(f) \tag{23}$$

Obtaining the maximum $\lambda(f)$ corresponds to calculating the largest eigenvalue of the generalized eigenvalue problem (23), and the corresponding eigenvector $e(f)$ gives the solution for the max-SNR beamformer

$$W_k(f) = e(f) \tag{24}$$

Equation (23) is simplified to an eigenvalue problem by multiplying both sides by $[R_T^k(f)]^{-1}$.

The max-SNR beamformer does not have any constraint for its gain, so the beamformer gain provided by equation (24) has scaling ambiguity. This characteristic should be compensated for if the MaxSNR beamformer is applied to wide-band signals such as speech. Inspired by the deflation-based blind source separation algorithm [14], we propose compensating $W_k(f)$ so that the output $y_k(f,\tau)$ becomes as close as observations:

$$X(f,\tau) \cong a(f) Y_k(f,\tau) = a(f) W_k^H(f) X(f,\tau) \tag{25}$$

That is, we calculate $a(f)$, which minimizes the following cost function:

$$\varsigma(a(f)) = \varepsilon\{\|X(f,\tau) - a(f)Y_k(f,\tau)\|^2\} \tag{26}$$

This is a linear least-mean-squares estimation problem [10]. Therefore, an optimal $a(f)$ can be obtained by setting the differentiation $\frac{\partial\varsigma(a(f))}{\partial a(f)}$ to zero:

$$a(f) = \frac{\varepsilon\{Y_k^*(f,\tau)X(f,\tau)\}}{\varepsilon\{|Y_k(f,\tau)|^2\}} = \frac{R_X(f)W_k(f)}{W_k^H(f)R_X(f)W_k(f)} \tag{27}$$

where $R_X(f)=\varepsilon\{X(f,\tau)X^H(f,\tau)\}$ is the observation correlation matrix. The scale compensated beamformer is given by a selecting the $J$-th component $a_J$,

$$W_k(f) \leftarrow a_J W_k(f) \tag{28}$$

### B. Evaluation of the max-SNR beamformer

An evaluated an experimental system for target speech

separation on a humanoid robot called SIG2, equipped with four omni-directional microphones on the left, front, right and back side of its head (Figure 2, right side). Experiments were done using two (Mic. #1,3), three (Mic. #1,2,3), or four (Mic. #1,2,3,4) channels. Three speakers were prepared as sound sources and, located as shown in the left part of Figure 2 (one in front with, side speakers set at ±30°, ±60° and ±90°, all at a distance of 1.5 m). The test set included 200 different phonetically balanced isolated Japanese words for each of the three speakers and 200 Japanese words were simultaneously emitted by the three speakers (each word spoken once). At that time, each speaker emitted a different word. To obtain the weight of the max-SNR beamformer (refer to equation 28), we used five words beforehand taken from among the 200 words emitted by each speaker.

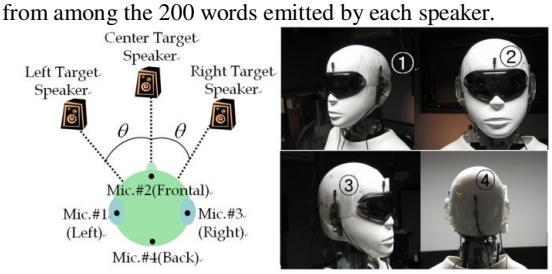


**Fig. 2.** Experiment condition and SIG2 with four microphones.

Figure 3 shows original mixing speech signals and three signals separated by the max-SNR beamformer with four microphones and the speakers 90° apart at a distance of 1.5 m. Table I shows the average SNR of the original mixing words and the separated words. These results show that more microphone channels and larger angles between the sources produce better separation results.

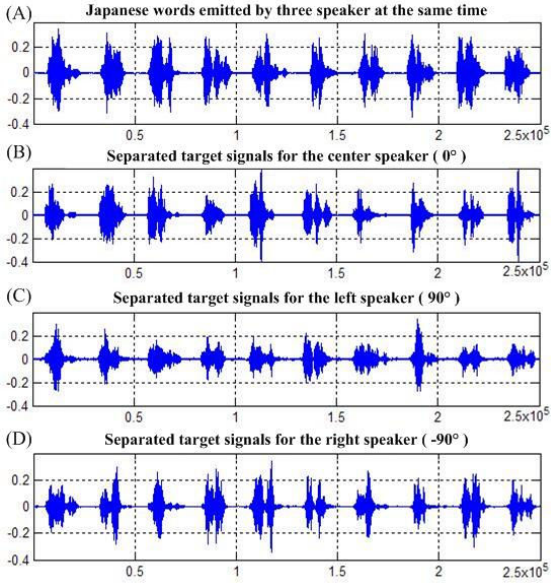


**Fig. 3.** Original and separated signals with four microphones and 90° separation.

TABLE I
SNR RESULTS OF SPEECH SEPARATION AT SIG2

| Mics | θ | SNR [dB] | | |
|---|---|---|---|---|
| | | Before MSNRBF | After MSNRBF | Improvement |
| 2 Ch. | ±30° | 10.8965 | 23.5050 | 12.6085 |
| | ±60° | 14.7353 | 29.8516 | 15.1163 |
| | ±90° | 15.8177 | 27.9715 | 12.1538 |
| 3 Ch. | ±30° | 11.6157 | 28.4096 | 16.7939 |
| | ±60° | 15.2691 | 23.3600 | 8.0909 |
| | ±90° | 19.1238 | 49.3091 | 30.1853 |
| 4 Ch. | ±30° | 12.5819 | 35.8380 | 23.2561 |
| | ±60° | 14.2327 | 51.7311 | 37.4984 |
| | ±90° | 20.4983 | 57.6360 | 37.1377 |

## IV. TARGET SPEECH RECOGNITION FOR ROBOTS

### A. *System overview*

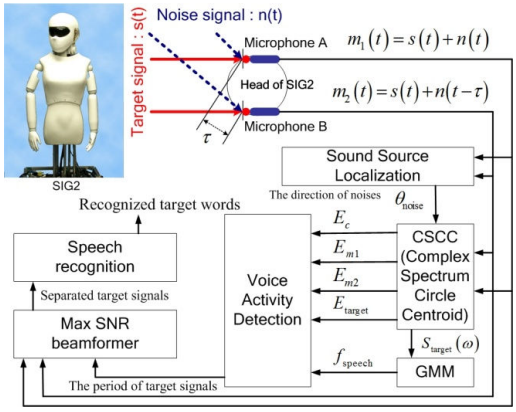


**Fig. 4.** System overview for recognizing target signals.

Figure 4 shows an overview of the structure of our system based on the CSCC and max-SNR beamformer methods and the SIG2 robot. This human-like auditory system has two omni-directional microphones, one at each of the left and right ear positions. First, to use the CSCC method, after finding the direction of noise signals, the robot is able to determine whether target signals exist and whether the target signals are voice or something different through CSCC and GMM, respectively, as discussed in Section II. The max-SNR beamformer can then separate target signals classified by VAD as discussed in Section III. In this section, after direct application of our VAD and max-SNR beamformer to the SIG2 auditory system, we discuss the results of VAD and the recognition of enhanced target signals.

### B. *Experiments and results for VAD*

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

$$SHR = \frac{N_S}{N_{Sref}}, \quad NSHR = \frac{N_N}{N_{Nref}} \tag{29}$$

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.

We conducted experiments under the following conditions. The distance between two microphones was 0.15 m. The sampling rate was 16 kHz and 1024-point FFT was applied to the windowed data with 512-sample overlap. As shown in Figure 5, 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 30°, 60°, or 90° to the side. Two loud sounds were simultaneously emitted from two speakers for 30 s. We used 10 speech samples (from five men and five women) for target signals, and three noise samples (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 s. The SNR values were -5, 0, 5, or 10 dB.

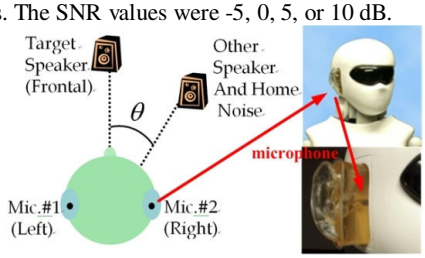


**Fig. 5.** Experiment conditions and SIG2 with two microphones.

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 a one-channel-based VAD, we obtained performance results for the G.729B VAD after averaging the results obtained by the left and right microphones.

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

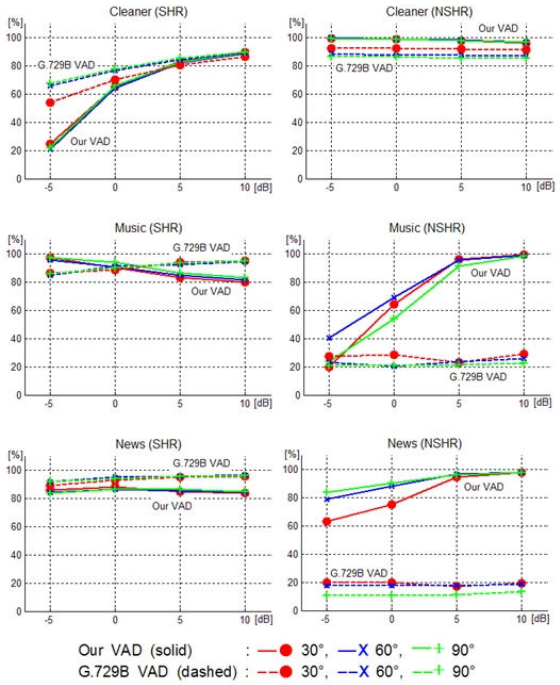


**Fig. 6.** Results of VAD based on CSCC.

### C. *Experiments and results for recognizing target signals*

The environment was set to be as similar to a typical room as possible. As shown in Figure 5, the target speaker stood in front of a humanoid robot while another speaker producing the noise was placed on the right at a 30°, 60° or 90° angle. All sound sources were at a distance of 1.5 m. The noise included classic music and the sound of a vacuum cleaner. The music was played throughout the experiment and the vacuum cleaner noise was produced occasionally. The interval during which speech was emitted from the center speaker sometimes overlapped the interference speech emitted from the side speaker; i.e., sparse dialogue. The SNR between target signals and interference signals was set to 5 dB and we obtained the weights for the max-SNR beamformer using five words and noise for about 10 s beforehand. Finally, to verify the speech recognition performance of the center speaker, we used target speech containing 50 words. These words were recognized by a speech recognition system and Table II shows those results. Here, in the case of recognizing words before separating the target speech, we averaged the results obtained by the left and right microphones.

We evaluated our VAD after we manually listened to all detected target speech. The result graph after VAD is shown in Figure 7 (D). The max-SNR beamformer was then used to separate the detected target signals. Since only two microphones were used, so that the auditory system would closely resemble human hearing, interference signals were not perfectly removed but we confirmed that the

performance was improved. For example, while a larger gap between speech signals and noise usually enables better performance, the improvement at 90° in Table II was less than at 30° or 60°. This was because noise emitted by the speaker at 90° directly entered the left microphone; i.e., the noise magnitude was much larger than that of speech signals at the left microphone and using only a pair of microphones was insufficient to remove that noise. These results are shown in Table II and the result graph after noise was reduced is shown in Figure 7 (E).

TABLE II
TARGET SPEECH RECOGNITION RESULTS

| Mics. | $\theta$ | Success rate of VAD | Recognition of Target Speech | | Improvement |
|---|---|---|---|---|---|
| | | | Before max SNR beamformer | After max SNR beamformer | |
| 2 Ch. | 30° | 96 % | 55 % | 72 % | 17 % |
| | 60° | 100 % | 50 % | 64 % | 14 % |
| | 90° | 100 % | 38 % | 44 % | 6 % |

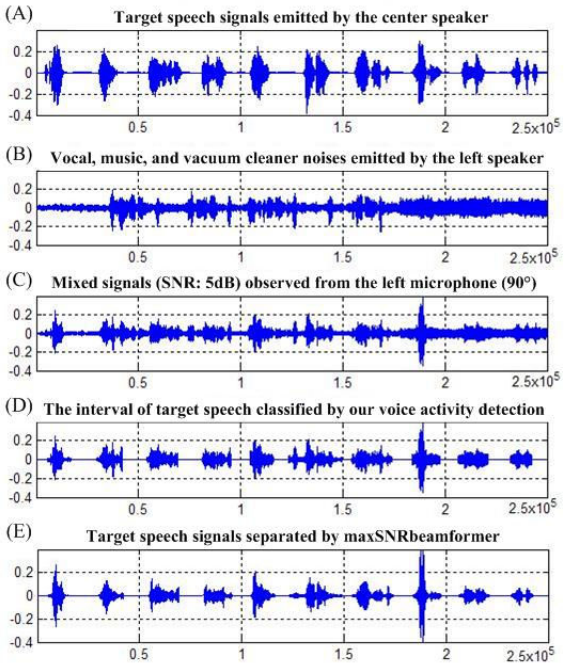


**Fig. 7.** Target Speech Detection and Separation with two microphones at 90°

## V. CONCLUSION

Our VAD system enables humanoid robots with two microphones to accurately detect the intervals of speech words or keywords generated in front of them even in noisy home environments, as was confirmed experimentally. In addition, the max-SNR beamformer helped improve the speech recognition performance for detected target signals. In the next step, to avoid the need for manual calculation, we are considering ways in which robots can automatically calculate the weights of the max-SNR beamformer after they have classified the intervals of target and interference signals. We are also considering the use of additional microphones, though only when separating target signals in order to reduce the execution time.

## 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, K. Komatani, T. Ogata, H. G. Okuno, "Auditory and visual integration based localization and tracking of humans in daily-life environments", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2007), Oct. (2007) pp. 2021-2027.

[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] Shoko Araki, Hiroshi Sawada and Shoji Makino, "Blind Speech Separation in a Meeting Situation with Maximun SNR Beamformers," IEEE/ICASSP Int. Conf. Acoustics, Speech, and Signal Processing, 2007, pp. I-41-44.

[11] H.L. Van Trees, Ed., *Optimum arrays processing*, John Wiley & Sons, 2002.

[12] 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.

[13] 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.

[14] A. Cichochi, R. Thawonmas, and S. Amari, "Sequential blind signal extraction in order specified by stochastic properties," Electronics Letters, vol. 33, no. 1, pp. 64-65, 1997.