Reliable Sound Source Localization for Human Robot Interaction

Hyun-Don Kim*, Jong-Suk Choi*, Chang-Hoon Lee**, and Mun-Sang Kim*

* Korea Institute of Science and Technology, Intelligent Robotics Research Center, Seoul, Korea
(Tel : +82-2-958-5618; E-mail: reynolds,cjs,munsang@kist.re.kr)
** Department of Information Communication Engineering, Paichai University, Teajon, Korea
(Tel : +82-42-520-5702; E-mail: naviro@pcu.ac.kr)

Abstract: In this paper, we propose a humanoid active audition system which detects the direction of sound and performs speech recognition using just three microphones. Compared with previous researches, this system comprises simpler algorithm and better amplifier system having advantages to increase a detectible distance of sound signal in spite of simple circuit. In order to verify our system's performance, we install the proposed active audition system to the home service robot, called Hombot II, which has been developed at the KIST (Korea Institute of Science and Technology), thus we confirm excellent performance by experimental results

Keywords: Sound's direction detection, Speech recognition system, Humanoid active audition

1. INTRODUCTION

The speech recognition has been applied into various systems and its performance has been improved greatly. In addition to the recognition, sound localization becomes a technology of much interest in the research field of human-robot interaction. In order to recognize speech with high confidence, the techniques which separate speech from various sound and remove noise from the signal of speech have been received a great deal of attention. Also, humanoid robots integrated with computer vision and various sensors have been developed for similar behaviors of human [1]-[4].

The objective of this research is to propose reliable sound source localization for human robot interaction. Compared with previous researches, our system comprises simpler algorithm and nonlinear amplifiers which have advantage to increase a detectible distance of sound signal in spite of simple circuits. Furthermore, to make reliable detection of sound's direction, we propose a new performance index using differences of cross-correlation. Also, since this system includes a function of VAD (Voice Activity Detection), the robot can know when to start finding the direction and performing speech recognition automatically.

To verify our system's feasibility, the proposed audition system is installed in the home service robot, called Hombot II, which has been developed at the KIST (Korea Institute of Science and Technology). Fig. 1 shows the audition system installed in Hombot II.

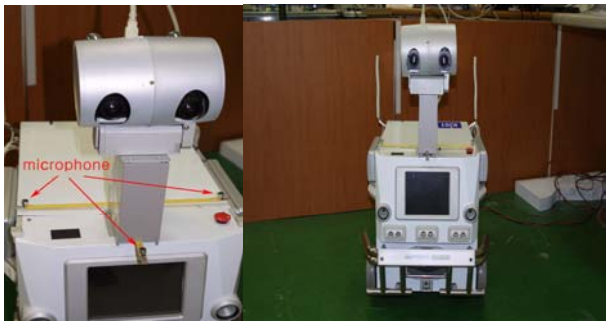

Fig. 1 Active audition system installed in Hombot II.

2. SYSTEM HARDWARES

The audition system is composed of pre-amplifier board, mic-mounted triangular rod, commercial AD converter, and a single board computer to execute our program. The AD converter samples data from three microphones to the rate of 11 kHz for each [6]-[7].

2.1 Nonlinear amplifier system

Nonlinear amplification which is able to make dynamically variable amplification according to the signal magnitude is required to increase the range of detectable distance. If the ratio of amplification is fixed to small one, the signal of speech occurring at the long distance can be hardly extracted from its received signal whose magnitude is small enough for the contents of speech to be canceled by noise. To the contrary, with large ratio, the signal occurring nearby sometimes may be saturated in the AD conversion. For this reason, the speech recognition system which is less affected by the distance to sound's source is necessary. To resolve this problem, we propose the use of SSM2166, made by Analog Device Corporation, which enables the nonlinear amplification. Our circuit, as shown in Fig. 2, is adjusted to compression ratio of 5:1.

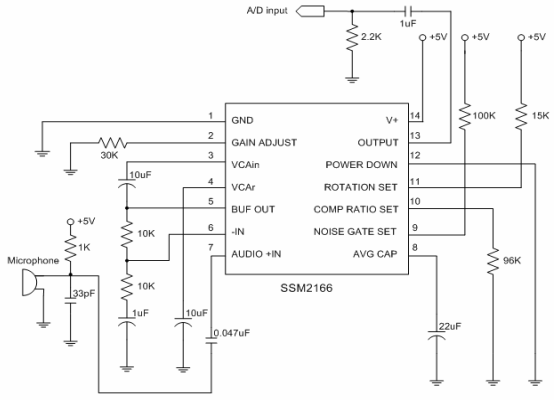

Fig. 2 Circuit of Nonlinear Pre-Amp

2.2 Comparing nonlinear amplifier with linear amplifier

In order to verify our nonlinear amplifier's performance, we should perform experiments to compare with normal l

inear amplifier. In the Fig. 3 and Fig. 4, the left hand shows sampled signals amplified by linear amplifier and the right hand shows sampled signals amplified by proposed nonlinear amplifier. Every source of the speech signals is the same one. Besides, main computer performs normalization within a range of ±0.5V so that the signals are suitable to speech recognition. Fig. 3 shows speech signals outputted at a distance of 0.5m. This figure shows that linear signals are more noise and unclear shape than nonlinear signals.

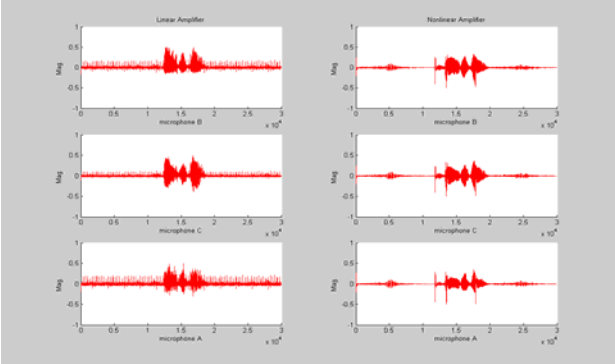

Fig. 3 Comparison between two kinds of data at a distance of 0.5m

Fig. 4 shows speech signals outputted at a distance of 1.5m. This figure shows that the linearly amplified signals (left) may be canceled with noise signal since they are far smaller than nonlinearly amplified signals (right). Ultimately, as a pre-amplifier board is made with the proposed nonlinear circuits, we can get advantages to increase detectable distance of sound signal and to reduce calculation time so as to execute various filtering algorithms such as low or high pass filtering.

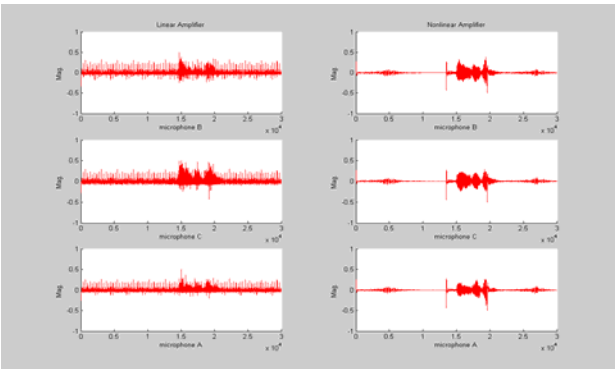

Fig. 4 Comparison between two kinds of data at a distance of 1.5m

3. SOUND SOURCE LOCALIZATION

3.1 Configuration of microphones and their cross-correlations

This paper uses DOA (Delay Of Arrival) for tracking the direction of sound [8][10]. DOA is the method that uses a time-delay from the source of sound to each microphone. Even though the time delay is short, the difference of arrival time occurs between array-shaped microphones. In Fig. 5, three microphones are arranged such that their distances from the center of triangular rod are the same. Two couples of A vs. C and B vs. C are selected in the view point of C. Note that the sampling data has maximum delay of time when a sound enters straightly through both A and C, or B and C.

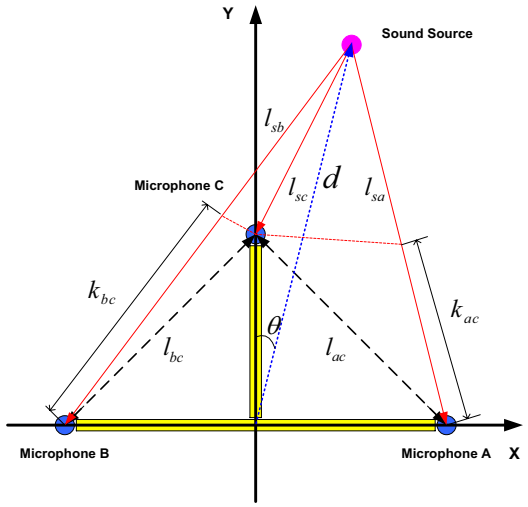

Fig. 5 Location of three microphones

In this case, the relative distance corresponding to the maximum delay is defined as l_{ac} (or l_{bc}). Also, the distance between sound's source and mic. A (mic. C) is defined as l_{sa} (or l_{sc}). The velocity of sound and sampling frequency are defined as v and F_s respectively. The number of sampling about the maximum delay is defined by (1) and (2) where n_{ac} is the number of sampling of maximum delay between A vs. C microphone and n_{bc} is the other one between B vs. C microphone.

$$n_{ac} = \frac{l_{ac}}{v} F_s \tag{1}$$

$$n_{bc} = \frac{l_{bc}}{v} F_s \tag{2}$$

The relation coefficient between mic. C and mic. A is defined by (3). Also, The relation coefficient between mic. C and mic. B is defined by (4). The variable t_g is a target number of delay in the g^{th} sampling period. Equation (3) and (4) is considered by sampling data from g=0 to g=∞. However, the real application of infinite period is impossible. Therefore, variable t_g is determined by suitable sampling data. We should decide the optimal sampling period consisted of 600 samples through experiments.

$$R_{ac}(k) = \frac{\sum_{g=0}^{\infty} \{A(t_g - k)C(t_g)\}}{\sqrt{\sum_{g=0}^{\infty} A(t_g - k)^2} \sqrt{\sum_{g=0}^{\infty} C(t_g)^2}} \tag{3}$$

$$R_{bc}(k) = \frac{\sum_{g=0}^{\infty} \{B(t_g - k)C(t_g)\}}{\sqrt{\sum_{g=0}^{\infty} B(t_g - k)^2} \sqrt{\sum_{g=0}^{\infty} C(t_g)^2}} \tag{4}$$

The variable k represents the number of actual delay sam-

ples. The number of delay k, in our configuration, spans to the range of $-n_{ac} \sim n_{ac}$ in this (3) and $-n_{bc} \sim n_{bc}$ in this (4) where its positive/negative value means that the sound enters microphone A and B earlier/later than microphone C.

Now, sound's direction should be calculated using relation coefficient R_{ac} and R_{bc} for all possible k_{ac} and k_{bc}. Fig. 6 illustrates the number of delay samples and the actual angle of sound's direction. An actual delay of sound's direction is expressed as (5) and (6).

$$k_{ac} = \frac{(l_{sc} - l_{sa})}{v} F_s \tag{5}$$

$$k_{bc} = \frac{(l_{sc} - l_{sb})}{v} F_s \tag{6}$$

However, we can't know the location of sound source (θ,d) yet. Therefore, the following method is proposed to estimate the sound source location. Matrix r presents the cross correlation of R_{ac} and R_{bc} for all possible k_{ac} and k_{bc}. All values of matrix r are calculated by (7).

$$r(\theta) = R_{ac}\left[k_{ac}(\theta)\right] R_{bc}\left[k_{bc}(\theta)\right] \quad \text{where } 1° \le \theta \le 360° \text{ i.e. } \theta = 1°, 2°, \cdots, 360° \tag{7}$$

Next, because we want to find the angle of sound's direction, we should first know the maximum value in the matrix r. After we fix threshold value in the r by using (8), we perform normalization to the r by using (9).

$$r_{thr} = 0.99 \max\left\{r(\theta)\right\}, \text{ where } \theta = 1°, 2°, \cdots, 360° \tag{8}$$

$$\begin{cases} r(\theta) = 0, & \text{if } r(\theta) < r_{thr} \\ \dfrac{(r(\theta) - r_{thr})}{(r_{\max} - r_{thr})}, & \text{if } r(\theta) \ge r_{thr} \end{cases}, \text{ where } \theta = 1°, 2°, \cdots, 360° \tag{9}$$

And, if we perform a weighted average to the r by using (10), we will find the angle of sound's direction.

$$\frac{\sum_{\theta=1}^{360} (r(\theta) \times \theta)}{\sum_{\theta=1}^{360} r(\theta)} = \theta_{sd} \tag{10}$$

3.1 Reliable detection of sound's direction

In a real speech signal, as there are reverberations, noise signals and consonants which have weakly periodic signals, wrong detections of sound's directions are calculated by computer frequently. Therefore, in order to find accurate directions of speech signal, we should detect sound's direction at the frame which has maximum energy within a period of speech signal.

The energy of a frame is expressed as (11).

$$E_{frame} = \frac{1}{k} \sum_{i=0}^{k} x^2(i) \tag{11}$$

The $x(i)$ is a sampling data of i-th in speech signals. However, a method using frame energy has several problems. First, if much noise is included in a speech signal, it will be able to select a frame which is not a period of speech signal. Second, because the frame having a maximum energy has not always good data to find an accurate direction of sound, accuracy related to detecting sound's direction can be reduced.

To fix these problems, we propose a new performance index rather than the frame energy. Given each frame, the performance index is expressed as (12).

$$P = r_{\max} - r_{\min} \tag{12}$$

We've found a notable feature through lots of experim ental investigation: it is true that when we spread values calculated by using (7) on the range of all angles, the di fference between magnitudes of the cross-correlation is ve ry informative to find reliable detection of sound's directi on. After selecting the reference frame having the maxim um value of our performance index P in a sample period, we decide direction whose cross-correlation value is the maximum at the selected frame as the final result.

To compare a frame energy method with a cross-correlation method, we used three commands such as "look at me," "go to a big room," and "patrol my home." Generating spots of each command were total 13 points at a distance of 1 meter. The azimuth which ranges from -90° to 90° was divided by every 15°. Table 1 is the average of experimental results.

Table 1 Comparison of performances

	Successful detection of sound's direction		**Angle error of sound's direction**	
Method	Frame Energy	Proposed	Frame Energy	Proposed
Average	82%	97%	7.2°	5.9°

As you see Table I, the cross-correlation method is better in the percentage of successful detection and average of angle error than the frame energy method.

Fig. 6 illustrates a 3 dimension graph which consists of numerical values calculated by using (7). At this time, used speech command is "patrol my home" coming at a distance of 1 meter and 30°. Where a frame has the proposed performance index which have the largest value throughout all the frames (See the inside of blue circle in Fig. 6), we can find an accurate direction of sound.

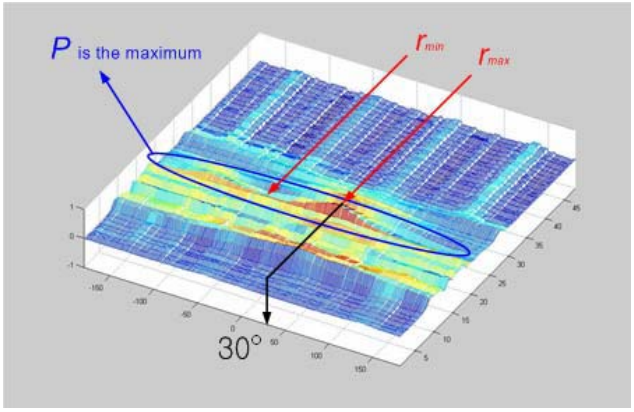

Fig. 6 The 3D graph of cross-correlation

4. VOICE ACTIVITY DETECTION

For the purpose of effective interaction between human being and a robot, it is necessary to extract the period in which only voice signals are included: Non-voiced or silent periods are unnecessary or undesirable. Therefore, we propose a function of VAD (Voice Activity Detection) using autocorrelation method to find pitch information. Hombot II executes a reliable detection of sound's direction and speech recognition when the robot has detected any signals of voice by VAD method. Pitch information rather than energy is applied to the VAD since the former has the advantage of a robust feature against noises [10]. Beside, In order to detect a pitch, we use an autocorrelation method which is composed of simpler algorithm instead of FFT.

The frequency of a vocal cord concerning human being exists in the range between 50 and 250Hz in case of a male and between 120 and 500Hz in case of a female. Therefore, if we put 600 samples per one frame into the autocorrelation equation, the executed signal will show pitch having periodic form of human vocal cord. The equation of the autocorrelation is expressed as (13).

$$R_{cc}(k) = \frac{\sum_{g=0}^{600}\{C(t_g - k)C(t_g)\}}{\sqrt{\sum_{g=0}^{600} C(t_g - k)^2}\sqrt{\sum_{g=0}^{600} C(t_g)^2}} \quad (13)$$

Then, after we perform a medium filter which has excellent features in removing impulse noise, edge signal preservation and smoothing, we can know peak vertexes of the related pitch. Finally, as we can know the number of samples between two peak signals, the pitch can be detected by (14). To improve accuracy of VAD, we should also detect the second pitch in a frame.

$$Pitch = \frac{\text{Sampling Frequency}}{\text{A number of samlpes between the two peaks}} \quad (14)$$

Now, after making weighted sum of the calculated pitches of 9 frames, we can infer extracting the period of voice signal.

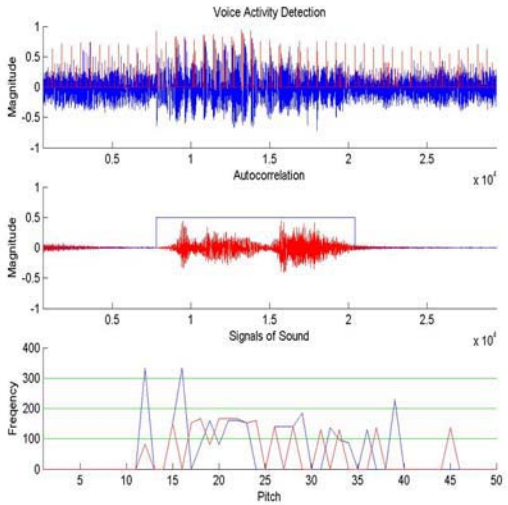

Fig. 7 Voice Activity Detection

Fig. 7 illustrates the experimental result which is performed by VAD. In the top figure, blue lines show the results of the autocorrelation and red lines show the detected periodic signal. In the middle figure, red lines show the original signal and a blue line shows the extracted speech signal. Also, in the bottom figure, red lines show the first pitches detected and blue lines show the second pitches finally.

5. SPEECH RECOGNITION SYSTEM

For the purpose of human-robot interaction, it is necessary for performing speech recognition and synthesis as well as detection of sound's direction. Hombot II performs speech recognition using a commercial engine, however, with no additional mic. for the recognition only. To recognize speech reliably, we input pre-processed signal of speech to L&H engine. Fig. 8 shows the block diagram of the speech recognition system.

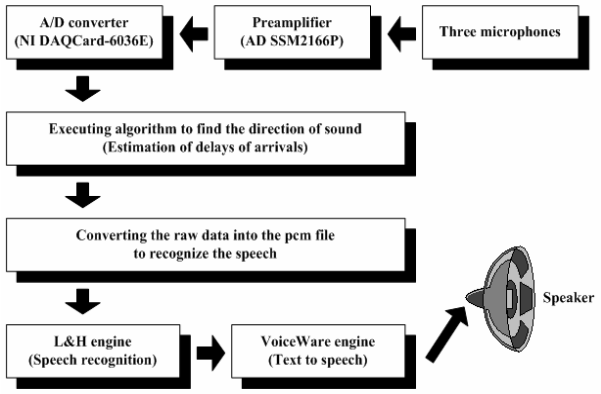

Fig. 8 Block diagram of the speech recognition system

First, in order to bring the speech signal to the L&H engine, speech signal with noise reduced by pre-amplifier board should be converted from analog signal into digital signal by AD converter. Next, digital signal should be made into PCM (Pulse Code Modulation) file of 11 kHz and 16bits after normalization. Fig. 9 shows the procedure of writing PCM file and the related C++ codes about it.

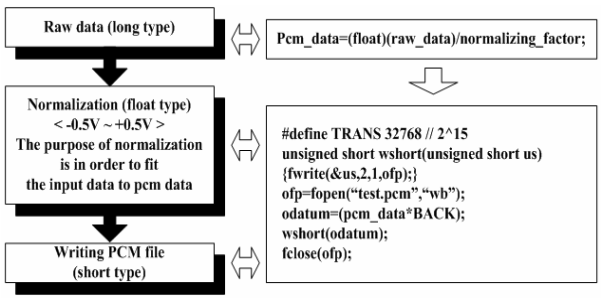

Fig. 9 Block diagram of writing PCM file

The speech synthesis of Hombot II uses commercial speech synthesis engine which outputs the most natural speech through a speaker.

6. EXPERIMENTAL RESULTS

Fig. 11 shows experiment setup. This experiment was conducted in an ordinary room, where background noise of about 55dB is generated by computer fans, an air-conditioner and motor noise in through the robot. Spots for sound's sources are total 27 points (See the red points in Fig. 16). The angle is divided by every 15° azimuth at a distance of 1 meter and every 30° azimuth at a distance of 2 and 3 meter. The computer's speaker emits commands which were previously recorded in about 85dB volume. If the error of detected sound's direction is up to 30°, the result will be regarded as failure.

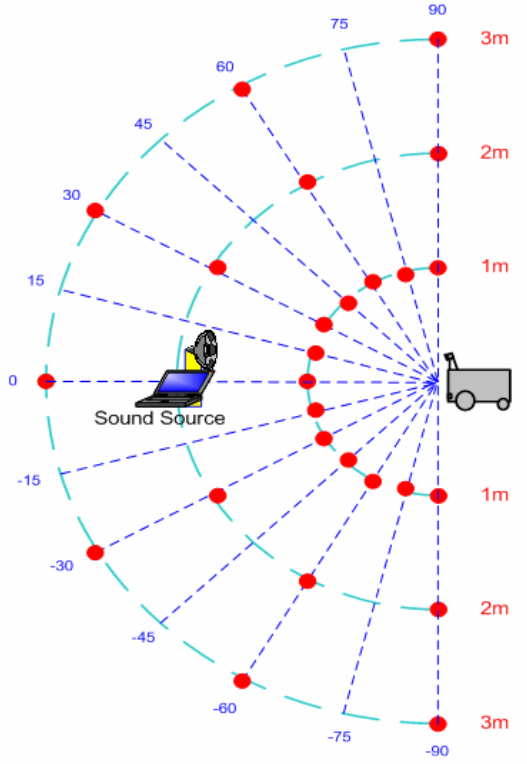

Fig. 10 Experiment setup

Table 2 is experimental results at 1m distance, at 2m distance, and at 3m distance. Three commands are mostly used commands in Hombot II: "Look at me," "Go to a big room," and "Patrol my home."

Table 2 The results of experiment at the distances of 2M and 3M

command		Look at me		Go to a big room		Patrol my home	
Distance	angle	recognition	direction	recognition	direction	recognition	Direction
1M	90°	OK	73°	OK	91°	OK	90°
	75°	OK	73°	OK	84°	OK	73°
	60°	OK	67°	OK	FAIL	OK	67°
	45°	OK	45°	OK	32°	OK	45°
	30°	OK	32°	OK	8°	OK	23°
	15°	OK	18°	OK	11°	OK	0°
	0°	OK	-12°	OK	16°	OK	0°
	-15°	OK	0°	OK	-32°	OK	-3°
	-30°	OK	-43°	OK	-42°	OK	-45°
	-45°	OK	-49°	OK	-45°	OK	-51°
	-60°	OK	-67°	OK	-62°	OK	-90°
	-75°	OK	-77°	OK	-78°	OK	-90°
	-90°	OK	-84°	OK	-95°	OK	-90°
average		**100%**	**6.9°**	**100%**	**8.7°**	**100%**	**8.4°**
2M	90°	FAIL	FAIL	OK	103°	OK	90°
	60°	FAIL	90°	OK	45°	OK	67°
	30°	FAIL	58°	OK	40°	OK	36°
	0°	FAIL	-5°	OK	-12°	OK	-4°
	-30°	OK	-58°	OK	-28°	OK	-58°
	-60°	OK	FAIL	OK	-69°	OK	-32°
	-90°	FAIL	-95°	FAIL	-90°	OK	-90
average		**29%**	**19.2°**	**86%**	**8.7°**	**100%**	**10.4°**
3M	90°	FAIL	95°	OK	FAIL	FAIL	FAIL
	60°	FAIL	FAIL	OK	FAIL	FAIL	FAIL
	30°	FAIL	FAIL	OK	36°	OK	35°
	0°	FAIL	FAIL	FAIL	FAIL	OK	FAIL
	-30°	FAIL	FAIL	OK	FAIL	FAIL	-45°
	-60°	FAIL	-62°	OK	FAIL	OK	-77°
	-90°	FAIL	-107°	OK	-89°	OK	-90°
average		**0%**	**6.2°**	**86%**	**3.5°**	**57%**	**9.3°**

This results show excellent performance at the short distance (1m): percentage of successful speech recognition is 100% and failure of detecting sound's direction is just one. Moreover, error averages about searching sound's direction are 6.9°, 8.7° and 8.4° corresponding to each command. However, performances at long distance (3m) are not so good: percentage of successful speech recognition shows poor performance and number of failure of detecting sound's direction is increased.

7. CONCLUSION

In this paper, conventional form of array-typed microphones is avoided. Also, simple and reliable algorithm with new pre-processing hardware is developed such that we are able to find the direction of sound's source from entire azimuth by using three microphones. Furthermore, it makes possible to perform speech recognition without another specific microphone (i.e., wireless or unidirectional sensitive microphone).

The audition system of Hombot II is designed for the optimized performance in the interaction between a human being and a robot. Consequently, this system has some distinguished functions. First, using the proposed preamplifier with simple circuits, we can get advantages to increase the detectible distance of sound's signal and to reduce noise. Second, we have proposed the new method to select the frame, which has good information for finding sounds' direction and has robustness to noisy environments with echo signals and consonants. Third, for the purpose of effective interaction between a human being and a robot, we have integrated functions being able to perform speech recognition and synthesis. Finally, as we apply VAD using autocorrelation, our system can automatically and continuously perform finding direction of sound and speech recognition whenever speech commands enter to microphones.

For further application to the real life, the system should extract the desired signal when voices of several people are mixed. Also, it should eliminate the noises even though large ones are mixed.

REFERENCES

[1] J. Huang, K. Kume, A. Saji, "Robotics spatial sound localization and its 3D sound human interface" Cyber Worlds, pp. 191-197, Nov. 2002.

[2] J. Huang, N. Ohnishi, N. Sugie, "A biomimetic system for localization and separation of multiple sound sources", IEEE/IMTC, Vol. 2, pp. 967-970, May 1994.

[3] J. Huang, N. Ohnishi, N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect", IEEE Trans., Vol. 46, pp. 842-846, Aug. 1997.

[4] J. Huang, N. Ohnishi, N. Sugie, "Spatial localization of sound sources: azimuth and elevation estimation" IEEE/IMTC, Vol. 1, pp. 330-333, May 1998.

[5] J. Huang, T. Supaongprapa, I. Terakura, N. Ohnishi, N. Sugie, "Mobile robot and sound localization" IEEE/RSJ, Vol. 2, pp. 683-689, Sept. 1997.

[6] H. D. Kim, J. S. Choi, C. H. Lee, G. T. Park, M. S. Kim, "Sound's direction Detection and Speech Recognition System for Humanoid Active Audition", ICCAS2003 Int. conference, Oct. 2003.

[7] H. D. Kim, J. S. Choi, C. H. Lee, G. T. Park, M. S. Kim, "Humanoid Active Audition System Using the Fuzzy Logic System", Journal of Control, Auto., and Sys. Eng., Vol. 9, No. 5, May, 2003.

[8] K. Nakadai, T. Matsui, H. G. Okuno, H. Kitano, "Active Audition System and Humanoid Exterior Design", IEEE/RSJ vol. 2, pp. 1453-1461, 2000.

[9] K. Nakadai, H. G. Okuno, H. Kitano, "Epipolar Geometry Based Sound Localization and Extraction for Humanoid Audition", IEEE/RSJ vol. 3, pp. 1395-1401, 2001.

[10] R. V. Prasad, A. Sangwan, H. S. Jamadagni, Chiranth M. C, Rahul S, "Comparison of Voice Activity Detection Algorithms for VoIP", IEEE/ISCC'02, pp. 530-535, 2002.

[11] S. A. Sekmen, M., Wilkes, K. Kawamura, "An Application of Passive Human-Robot Interaction: Human Tracking Based on Attention Distraction", IEEE Trans. Sys., Man and Cybern., vol. 32, no. 2, pp. 248-259, 2002.