RELIABLE DETECTION OF SOUND'S DIRECTION FOR ACTIVE AUDITION

Hyun-Don Kim, Jong-Suk Choi, *Chang-Hoon Lee and Munsang Kim
Intelligent Robotics Research Center Korea Institute of Science and Technology
*Department of Information Communication Engineering, Paichai Univ., Taejon, Korea
E-mail : {reynolds, cjs, munsang}@kist.re.kr, *naviro@pcu.ac.kr

Abstract: In this paper, we propose reliable detection of sound's direction for human robot interaction. Compared with previous researches, this system comprises simpler algorithm and a proposed nonlinear amplifier which has advantages to increase a detectible distance of sound signal in spite of simple circuits. Moreover, we propose the new performance index using cross-correlation to make reliable detection of sound's direction. Also, since this system include a function to detect a period of voice signals, the robot can know when to start finding sound's direction and performing speech recognition automatically. In order to verify our system's performances, we install the proposed audition system to the prototype robot, called IROBAA (Intelligent ROBot for Active Audition), and describe the experimental results.

Keywords : sound localization, voice activity detection, human robot interaction, intelligent robot

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]-[10].

The objective of this research is to develop the techniques which enable the following scenario. Our robot recognizes its name and the direction of sound as well when a man calls. Then it turns its face to the direction, and can possibly recognize the person through face recognition technique to communicate with him.

We use nonlinear amplification as a preprocessor in order that speech recognition system can recognize speech at a long distance robustly against noises in the environment. 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 prototype robot, called IROBAA (Intelligent ROBot for Active Audition), which has been developed at the KIST (Korea Institute of Science and Technology).

Fig. 1 shows the audition system installed in IROBAA. The audition system is composed of pre-amplifier board, mic-mounted circle pad, commercial AD converter, and a notebook computer to execute our program. The AD converter samples data from three microphones to the rate of 11 kHz for each.

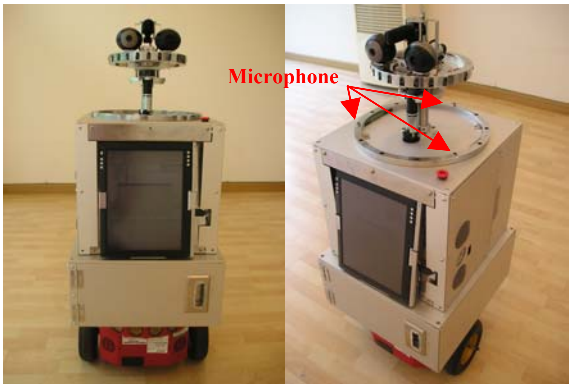

Figure 1. Active audition system installed in IROBAA.

2. Audition System of IROBAA

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.

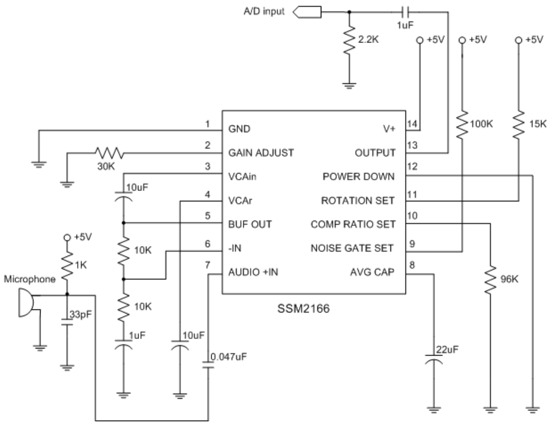

Figure 2. Circuit of Nonlinear Pre-Amp

In order to verify our nonlinear amplifier's performance, we should perform experiments to compare with normal linear amplifier. In the Figs. 3 and 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 speech signals is the same. Besides, main computer performs normalization within a range of ±0.5V so that they are suited to speech recognition.

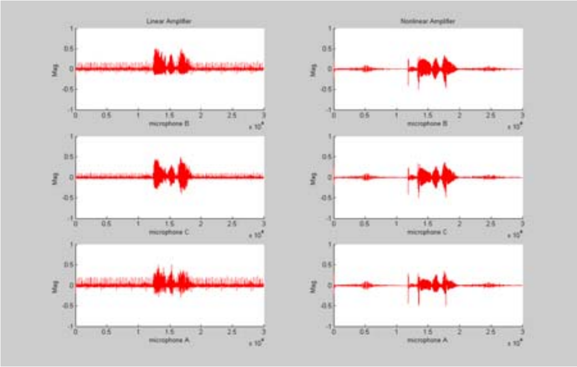

Figure 3. Compared two data of a distance of 0.5m

Fig. 3 shows speech signals outputted at 0.5m distance. This figure shows that linear signals are more noise and unclear shape than nonlinear signals.

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

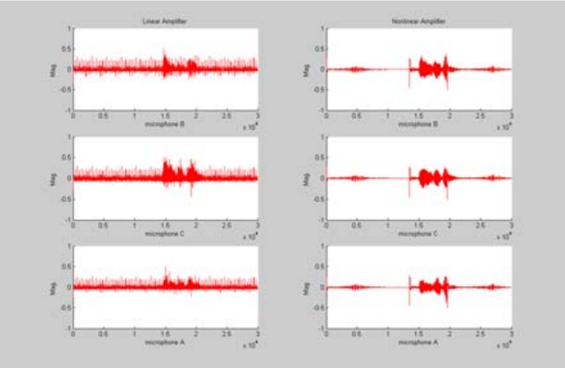

Figure 4. Compared two data of a distance of 1.5m

3. Tracking of Sound's Direction

This paper uses DOA (Delay Of Arrival) for tracking the direction of sound [6]-[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.

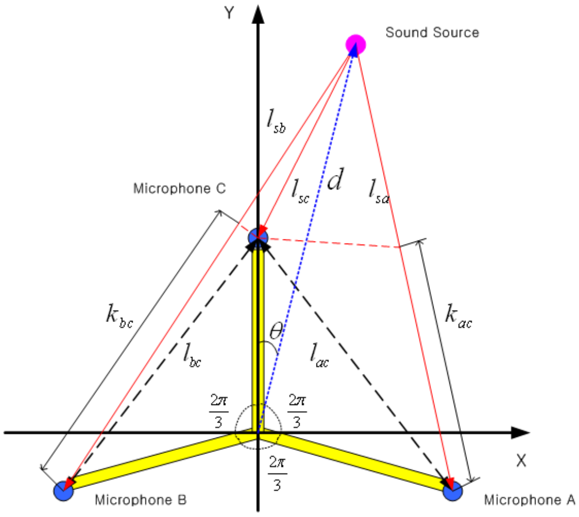

Figure 5. Location of three mirophones

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 552 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 samples. The number of delay k, in our configuration, spans to the range of -n_{ac} ~ n_{ac} in this (3) and -n_{bc} ~ 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. 5 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).

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

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).

$$r_{thr} = 0.99 \times \max\{r(\theta)\}$$
$$\text{where } 1° \le \theta \le 360° \text{ i.e. } \theta = 1°, 2°, ..., 360° \tag{8}$$

$$r(\theta) = 0 \quad if\ r(\theta) < r_{thr}$$
$$\frac{(r(\theta) - r_{thr})}{(r_{\max} - r_{thr})} \quad if\ r(\theta) \ge r_{thr}$$
$$\text{where } 1° \le \theta \le 360° \text{ i.e. } \theta = 1°, 2°, ..., 360° \tag{9}$$

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

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

4. 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 experimental investigation: it is true that when we spread values calculated by using (7) on the range of all angles, the difference between magnitudes of the cross-correlation is very informative to find reliable detection of sound's direction. After selecting the reference frame having the maximum 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 "go to a living room," "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 I is the average of experimental results.

Table I. Compare frame energy with proposed index

	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. Figure 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.

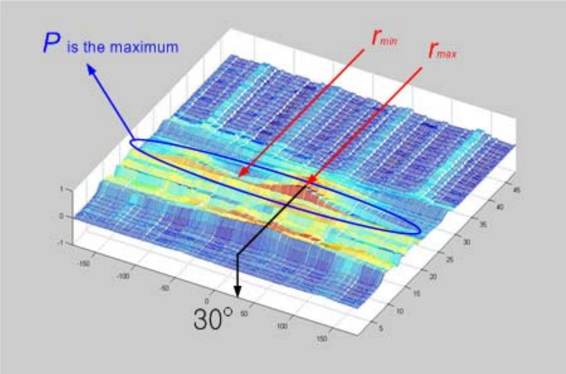

Figure 6. The 3D graph of cross-correlation

5. 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-voice or silent periods are unnecessary or harmful. Therefore, we propose a function of VAD (Voice Activity Detection) using autocorrelation method to find pitch information. IROBAA executes a reliable detection of sound's direction and speech recognition when the robot is decided to detect 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 [11][12]. 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 552 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).

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

Then, after we perform a medium filter which has excellent features in removing impulse noise, edge signal preservation and smoothing, we can calculate differential values about autocorrelation results. In case of real voice signals, since magnitude of a periodic form and differential values between sampled signals is large, the peak values can be calculated by applying threshold value to differential values.

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.

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

Now, after making weighted sum of the calculated pitches of 20 frames, we can infer extracting the period of voice signal.

Since the A/D converter which is installed in IROBAA has the function of double buffering, the robot can continuously execute the VAD algorithm at a second intervals without loss of raw data. Consequently, it can automatically and continuously perform finding direction of sound and speech recognition whenever speech commands enter to microphones.

6. 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. IROBAA 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. 7 shows the block diagram of the speech recognition system.

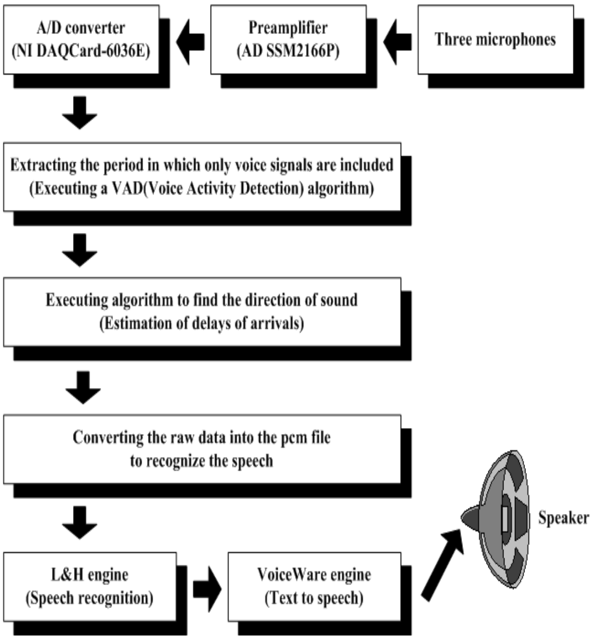
A/D converter
(NI DAQCard-6036E)
Preamplifier
(AD SSM2166P)
Three microphones
Extracting the period in which only voice signals are included
(Executing a VAD(Voice Activity Detection) algorithm)
Executing algorithm to find the direction of sound
(Estimation of delays of arrivals)
Converting the raw data into the pcm file
to recognize the speech
Speaker
L&H engine
(Speech recognition)
VoiceWare engine
(Text to speech)

Figure 7. Block diagram of the speech recognition system

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

7. Experimental Results

Fig. 8 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 24 points (See the red points in Fig. 8). The angle is divided by every 15° azimuth at a distance of 1 meter. The computer's speaker emits commands which were previously recorded in about 90dB volume. If the error of detected sound's direction is up to 30°, the result will be regarded as failure.

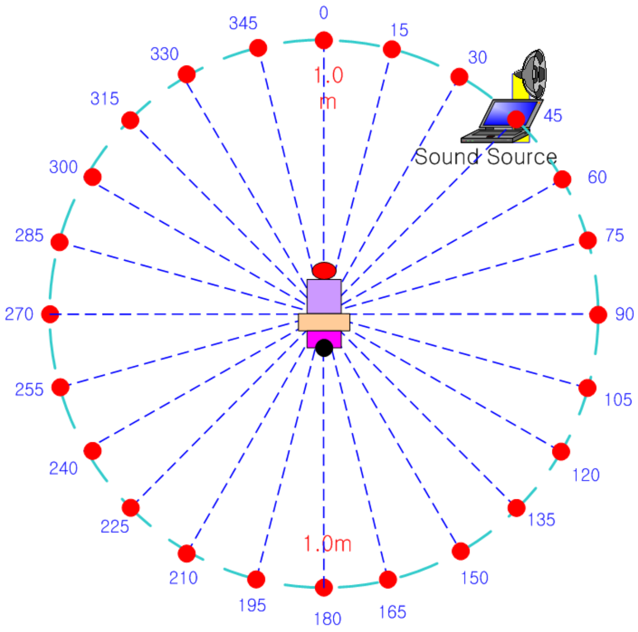

Figure 8. Experiment Setup

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

Table II. The results of experiment at 1m distance

Command	Go to a living room		Go to a big room		Patrol my home	
Angle	recognition	direction	recognition	direction	recognition	direction
0°	FAIL	352°	FAIL	352°	FAIL	FAIL
15°	OK	13°	OK	352°	FAIL	13°
30°	OK	21°	OK	20°	OK	29°
45°	OK	36°	FAIL	44°	FAIL	43°
60°	FAIL	62°	OK	72°	OK	62°
75°	FAIL	83°	OK	76°	OK	76°
90°	OK	90°	OK	95°	FAIL	98°
105°	OK	104°	OK	108°	FAIL	104°
120°	OK	FAIL	OK	FAIL	FAIL	121°
135°	OK	FAIL	OK	138°	OK	FAIL
150°	OK	FAIL	OK	155°	FAIL	155°
165°	FAIL	165°	FAIL	165°	FAIL	165°
180°	OK	204°	OK	194°	FAIL	198°
195°	OK	199°	OK	199°	OK	199°
210°	OK	205°	OK	205°	OK	208°
225°	OK	221°	FAIL	227°	OK	201°
240°	FAIL	239°	FAIL	239°	OK	227°
255°	FAIL	260°	FAIL	256°	OK	256°
270°	OK	268°	FAIL	276°	OK	270°
285°	OK	FAIL	OK	280°	OK	280
300°	OK	304°	OK	288°	OK	304°
315°	FAIL	316°	FAIL	331°	OK	319°
330°	FAIL	331°	FAIL	331°	FAIL	331°
345°	OK	347°	OK	342°	FAIL	347°
Successful Average	**Sound's Direction**		**Angle Error**		**Speech Recognition**	
	90.3%		**5.1°**		**59.7%**	

This results show excellent performance at the short distance (1m): percentage of successful sound's direction is 90.3%. Moreover, the average of errors about the estimated sound's direction is 5.1° corresponding to three commands. However, percentage of successful speech recognition is not so good. This may result from the limited performance of speech recognition engine and the incompleteness in extracting exact period of voice from VAD. Therefore, we should analyze and improve the VAD algorithm to increase successful rate of speech recognition in the integrated system.

8. Conclusions

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 IROBAA 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 pre-amplifier with simple circuits, we can get advantages to increase the detectible distance of sound's signal and to reduce noise. Second, tracking of sound source direction is often obstructed by echo sound due to surrounded obstacles. Moreover, as there are noise signals and consonants which have weakly periodic signals in the sampled speech signals, wrong detections of sound's directions may be calculated by computer frequently. In order to solve the problem, we proposed the new method to select the frame which has good information to find sounds direction (The fame is composed of 552 speech sampled data) based on cross-correlation. Third, as we apply VAD using autocorrelation, it can automatically and continuously perform finding direction of sound and speech recognition whenever speech commands enter to microphones. Finally, 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.

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, 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.

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

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

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

[5] 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.

[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] 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.

[11] 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.

[12] G. Monti, M. Sandler, “Monophonic Transcription with Autocorrelation”, GOST G-6 Conference, Dec. 2000.