Proceedings of the 2006 IEEE/RSJ
International Conference on Intelligent Robots and Systems
October 9 - 15, 2006, Beijing, China

Probabilistic Speaker Localization in Noisy Environments by Audio-Visual Integration

Jong-Suk Choi and Munsang Kim
Intelligent Robotics Research Center
Korea Institute of Science and Technology
Seoul, Republic of Korea
{cjs, munsang}@kist.re.kr

Hyun-Don Kim
Dept. of Intelligence Science and Technology
Kyoto University
Kyoto, Japan
hyundon@kuis.kyoto-u.ac.jp

***Abstract* - In this paper, we have developed not only a probabilistic sound localization system including VAD (Voice Activity Detection) component using three microphones but also a face tracking system using a vision camera. Moreover, we have proposed a way to integrate these systems to compensate the errors in the localization of a speaker and to reject unnecessary speech or noise signals entering from the undesired directions effectively. For the purpose of verifying our system's performances, we have installed the proposed audition and vision system to the prototype robot, called IROBAA (Intelligent ROBot for Active Audition), and showed how to integrate an audio-visual system.**

Index Terms - sound localization, face tracking, voice activity detection, human robot interaction, audio-visual integration.

I. Introduction

In the near future, the importance and participation of intelligent robots will grow rapidly in the human society. Also, the interaction between robots and normal people without help from robot experts will be essential. Therefore, the technology related to robots has been applied into various areas and its performance has been improved greatly. Especially, speech signal processing and image processing become the technology of much interest in the research field of human-robot interaction. For example, in an audition system, robots require sound localization as well as speech recognition to seek out or to talk with human beings naturally. Nowadays, in order to recognize speech with high confidence, the techniques which separate speech signals from various non-speech signals and remove noises from the speech signals have received a great deal of attention. Besides, a vision system has been helping robots recognize specific objects such as human faces and find the location of the recognized targets correctly. Ultimately, humanoid robots developed for implementing human-like behaviour need to integrate with visual and auditory technologies resulting that they become friendly toward human beings. Recently, many robot experts have a growing concern how they can integrate effectively with visual and auditory information as well as data from various sensors.

The objective of our research is to develop a reliable and feasible system for human-robot interaction which is performed in real environments. First, detecting intervals of speech signal, finding its direction and turning a robot's head to the direction of a speaker's face can help normal people interact with robots naturally [1-6]. Second, it is necessary to use the visual processing technology which can support robots to detect and track specific speaker's face individually. Moreover, collaborating with vision systems will make robots not only compensate the errors in the sound localization of a speaker but also effectively reject unnecessary speech or noise signals entering from the undesired directions and will be able to improve the performance of speech recognition consequently. Finally, by integrating visual and auditory processing technology, we can extend this research to human-robot interaction technologies including multiple speech localization and speaker's face recognition [7-8].

To verify our system's feasibility, the proposed audition system is installed in a 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. IROBAA involved a pre-amplifier board, a mic-mounted circle pad, a commercial AD converter, a normal web camera and a single board computer to execute our programs. All the codes have been implemented by using GNU C and C++ language on Linux.

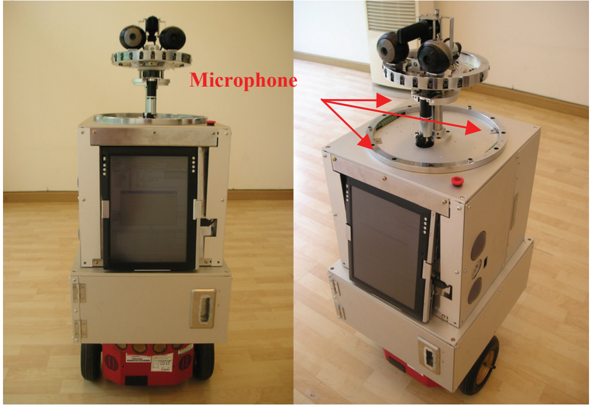


Fig. 1 IROBAA (Intelligent ROBot for Active Audition)

II. Nonlinear Amplification Board

Nonlinear amplification which is able to make dynamically variable amplification according to the signal magnitude is required to increase the range of detectable distance in the acquisition of sound signals. 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

1-4244-0259-X/06/$20.00 ©2006 IEEE

whose magnitude is so small that the contents of speech are cancelled by noise. To the contrary, with large ratio, the signal occurring nearby may be saturated in the AD conversion. To solve this problem, we propose the nonlinear amplification where smaller signal can be amplified with larger amplification ratio. To implement the nonlinear property, we use SSM2166, made by Analog Device Corporation. Our amplifier board, as shown in Fig. 2, is adjusted to compression ratio of 5:1 and is made up of 4 channels.

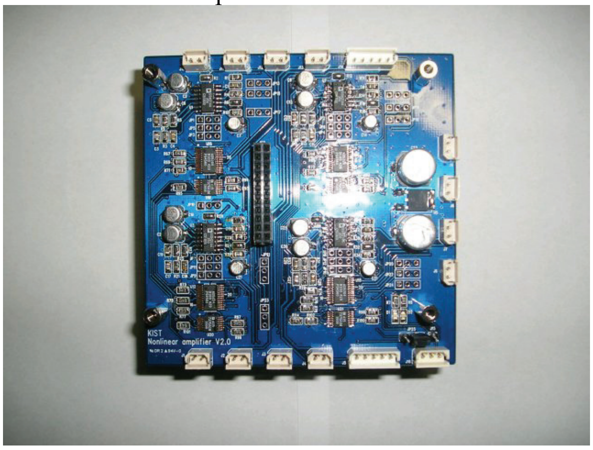


Fig. 2 Developed nonlinear pre-amp. board

III. Sound Localization

A. Probabilistic Tracking of Sound's Direction

This paper uses DOA (Delay Of Arrival) for tracking the direction of sound [6]. 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. 3, three microphones are arranged such that their distances from the center of triangular rod are the same.

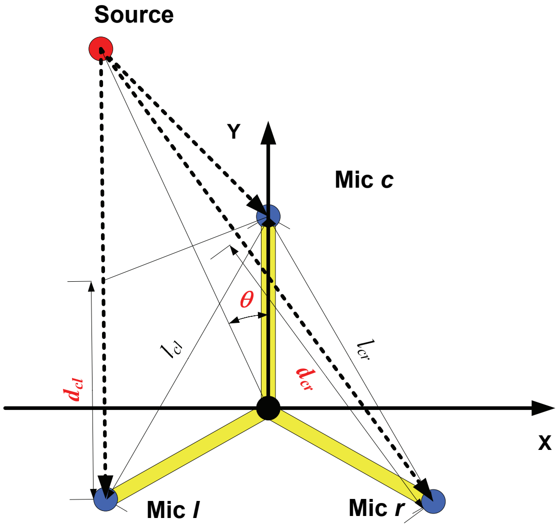


Fig. 3 Location of three mics

Two couples of mic *l* vs. mic *c* and mic *r* vs. mic *c* are selected in the view point of mic *c*. Note that the sampling data has maximum delay of time when a sound enters straight through both mic *l* and mic *c*, or mic *r* and mic *c*. Also, the distance between mic *l* (or mic *r*) and mic *c* is defined as l_{cl} (or l_{cr}). The sound velocity and sampling frequency are defined as v and F_s respectively. If a person speak from a direction θ defined as shown in Fig. 3, the DOAs between mic *c* and mics *l* and *r* are described approximately as

$$\begin{aligned} d_{cl} &\cong l_{cl}\cos\left(\theta+\frac{\pi}{6}\right) \\ d_{cr} &\cong l_{cr}\sin\left(\theta+\frac{\pi}{3}\right) \end{aligned} \tag{1}$$

By the way, the DOA measurements are noisy in the real world. Hence, we need to take a probabilistic approach. Let us assign, to the sound's direction θ, a normally distributed random variable whose mean and variance are assumed by $\bar{\theta}$ and Λ_θ respectively. Then, the DOAs can also be regarded as normally distributed random variables whose means and variances are

$$\begin{aligned} D_{cl} &\sim N\left(\bar{d}_{cl},\Lambda_{d_{cl}}\right) \\ D_{cr} &\sim N\left(\bar{d}_{cl},\Lambda_{d_{cl}}\right) \end{aligned} \tag{2}$$

where

$$\begin{pmatrix}\bar{d}_{cl}\\ \bar{d}_{cr}\end{pmatrix}=\begin{pmatrix} l_{cl}\cos\left(\bar{\theta}+\frac{\pi}{6}\right)\\ l_{cr}\sin\left(\bar{\theta}+\frac{\pi}{3}\right)\end{pmatrix} \tag{3}$$

$$\begin{pmatrix}\Lambda_{d_{cl}}\\ \Lambda_{d_{cr}}\end{pmatrix}=\begin{pmatrix}\left(\frac{\partial d_{cl}}{\partial\theta}\right)^2\\ \left(\frac{\partial d_{cr}}{\partial\theta}\right)^2\end{pmatrix}\Lambda_\theta=\begin{pmatrix} l_{cl}^2\sin^2\left(\theta+\frac{\pi}{6}\right)\\ l_{cr}^2\cos^2\left(\theta+\frac{\pi}{3}\right)\end{pmatrix} \tag{4}$$

Given the possible discrete angle θ_j between [0, 360), find the probability that instantaneous DOAs d_{cl} and d_{cr} are measured like

$$\begin{aligned} p\left(d_{cl}\mid\bar{\theta}_j\right) &= N\left(d_{cl},\bar{d}_{cl,j},\Lambda_{d_{cl},j}\right) \\ p\left(d_{cr}\mid\bar{\theta}_j\right) &= N\left(d_{cr},\bar{d}_{cr,j},\Lambda_{d_{cr},j}\right) \end{aligned} \tag{5}$$

Here, define a joint probability

$$p\left(d_{cl},d_{cr}\mid\bar{\theta}_j\right)=p\left(d_{cl}\mid\bar{\theta}_j\right)\cdot p\left(d_{cr}\mid\bar{\theta}_j\right) \tag{6}$$

Then we propose a probabilistic estimation about the sound's direction and its reliability as Eq. (7) at k^{th} frame.

$$\begin{aligned} \theta_{est,k} &= \arg\max_{\bar{\theta}} p\left(d_{cl},d_{cr}\mid\bar{\theta}_k\right)\ \text{ at } k^{\text{th}} \text{ frame} \\ R_{est,k} &= E_k\cdot\left(1-e^{-\alpha p_{est,k}}\right) \end{aligned} \tag{7}$$

where

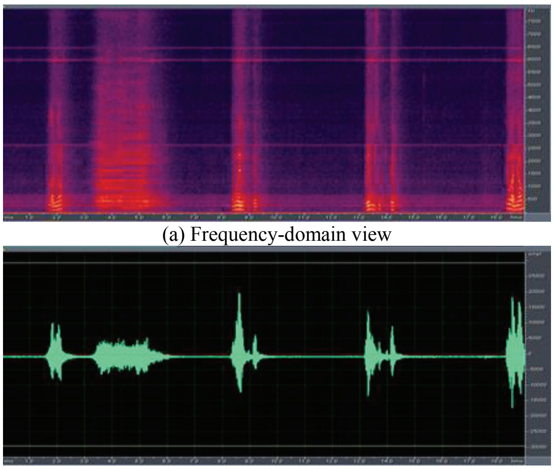
$$\begin{aligned} &E_k:\ \text{short-time energy at } k^{\text{th}} \text{ frame} \\ &p_{est,k}=p\left(d_{cl},d_{cr}\mid\theta_{est,k}\right) \end{aligned} \tag{8}$$

B. Reliable Detection of Sound's Direction

In a real world, as there are motor noises, reverberations, and consonants which have weakly periodic signals, wrong detections of sound's directions are calculated at some instantaneous frame. However, in our probabilistic method, the reliability measure $R_{est,k}$ shows lower value since the $p_{est,k}$ gets lower for those irregular cases. Hence, more reliable results can be acquired continuously if we do post-filtering with regard to both $p_{est,k}$ and $R_{est,k}$ like

$$\theta_{fil,k} = \frac{\theta_{fil,k-1}\cdot R_{fil,k-1}\cdot(k-1)}{R_{fil,k-1}\cdot(k-1)+R_{est,k}} + \frac{\theta_{est,k}\cdot R_{est,k}}{R_{fil,k-1}\cdot(k-1)+R_{est,k}}$$
$$R_{fil,k} = \frac{R_{fil,k-1}\cdot(k-1)}{k} + \frac{R_{est,k}}{k} \quad (9)$$

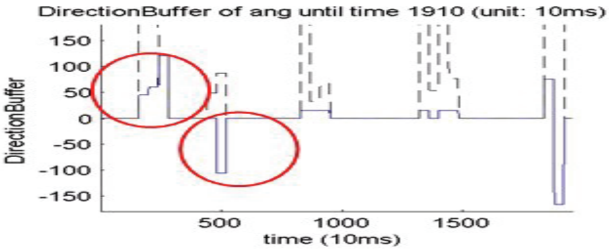
Fig. 4 shows both frequency-domain view and time-domain view of our speech signals which are generated from a direction of 50 degree and 2 meters ahead while changing the direction within 30 degree continuously for long frames. The overall signals are divided into five groups where the second is composed of motor noises only and the last is composed of voices and motor noises in the sequential order. As shown in Fig. 5, while traditional method (a) for the sound localization gives wrong results at the second noise group, the proposed probabilistic method (b) removes them. Furthermore, the post-filtering (c) removes the instantaneously abrupt result, at the last group, which could not be eliminated in (b).



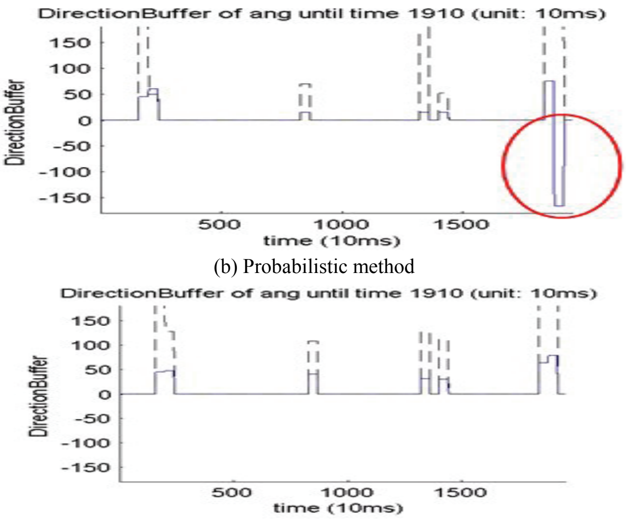
(a) Frequency-domain view

(b) Time-domain view

Fig. 4 Speech signals with motor noises



(a) Traditional method



(b) Probabilistic method

(c) Probabilistic method with post-filtering

Fig. 5 Comparison of results about sound localization for long frames

IV. 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 cepstrum to find pitch information [9]. The word 'cepstrum' was to mean the 'spectrum of a natural logarithmic (amplitude) spectrum'. That is to say, cepstrum means the signals made by inverse fourier transform of the logarithm of fourier transform of sampled signals. One of the most important features of cepstrum is that if the signal is periodic signal, the signal made by cepstrum will also present peaks signal at intervals of each period. Furthermore, compared to pitch detection method using autocorrelation at time domain, cepstrum has distinct peaks at intervals of each period and the first peak is always bigger than the second or the third one. Consequently, cepstrum can reliably extract the pitch of a speech signal. Given a signal $x(t)$, the equation of the cepstrum is expressed as Eq. (10).

$$c_c(\tau)=F^{-1}\{\log X(f)\}=F^{-1}\{\log|X(f)|+j\phi_x(f)\} \quad (10)$$

Fig. 6 shows the sequence of extracting pitch signals at IROBAA. First, to minimize frequency leakage effects, we apply hanning window to the sampled signals foremost. Then, after performing FFT (Fast Fourier Transform), the robot performs IFFT (Inverse Fast Fourier Transform) of the logarithm of these signals. At that time, since the frequency of a vocal cord concerning human beings exists in the range between 50 and 250Hz in case of a male and between 120 and 500Hz in case of a female, it has no problem even if we just search the pitch signals within the range of the fundamental frequency of human voice. Therefore, to minimize the disturbance of noises when a robot tries to extract pitches, we apply a low pass filter which has the range between 0 and 900Hz to the pitch-detection algorithm.

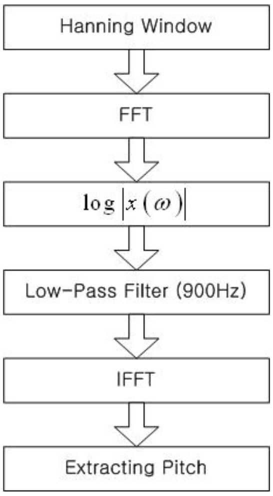


Fig. 6 Procedure of the method extracting pitch

Finally, with the number of samples between two peak signals found, the pitch can be detected by Eq. (11).

$$\text{Pitch} = \frac{\text{Sampling Frequency}}{\text{A number of samlpes between the two peaks}} \quad (11)$$

Here, we need to consider adding supplementary methods to VAD so as to reduce the effects of noises or improve the successful rate of VAD. As the supplementary methods, there used to be the short-time energy and ZCR (Zero Crossing Rate) [10] which are very simple but able to help our VAD to improve its efficiency. The short-time energy is used to know whether there are signals or not according to the magnitude. However, it is impossible to know whether the signals are real speech signals or noise signals. The short-time energy of a frame is expressed as Eq. (12).

$$E_{frame} = \frac{1}{k}\sum_{i=0}^{k} x^2(i) \quad (12)$$

where $x(i)$ means the sampling data of i-th step and k is the number of steps. The ZCR means that how many times the sign of signals are changed at the period of a frame. The ZCR is expressed as Eq. (13)

$$ZCR = \sum_{i=0}^{N-1} \left|\operatorname{sgn}\left[x(i)\right] - \operatorname{sgn}\left[x(i+1)\right]\right| \times \frac{1}{2} \quad (13)$$

In the interval of noise signals or consonants which have weakly periodic signals, the number of ZCR is increased in comparison with the interval of a vowel. Therefore, we can find the interval of speech signals roughly.

Now, we should develop a VAD algorithm which the three items - pitch, ZCR and short-time energy - are combined properly. Consequently, we need to set up the condition to select voiced regions [10]:

$$\begin{aligned} R_C &= \left\{C_i \,\middle|\, \min(F) < C_i < \max(F)\right\} \\ R_Z &= \left\{ZCR_i \,\middle|\, \min(Z) < ZCR_i < \max(Z)\right\} \quad (14) \\ R_E &= \left\{E_i \,\middle|\, \min(E) < E_i < \max(E)\right\} \end{aligned}$$

F, Z and E denote the frequency of pitch, the number of zero-crossing rate and the magnitude of frame energy respectively corresponding to the i-th frame of speech signals. Based on the above condition, the i-th frame is roughly declared *voiced* if the following logical expression is satisfied:

$$\left[\left(C_i \in R_C\right) \Lambda \left(ZCR_i \in R_Z\right) \Lambda \left(E_i \in R_E\right)\right] \Rightarrow \left(i \in Voice\right) \quad (15)$$

where 'Λ' denotes the logical 'and' operation, and *Voice* is the set of voiced indices.

Besides, 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 0.5 second intervals without loss of raw data. Therefore, it can automatically and continuously perform finding direction of voice and classify the interval of speech signals whenever speech commands enter to microphones.

V. Vision System of IROBAA

For the purpose of the detection of human faces, we used OpenCV (Open Computer Vision), the open source vision library made by Intel company. This vision library supplies the function concerning human face detection to users. Thus, it is able to track a human face using just one of two web cameras installed in the head of IROBAA. Based on OpenCV, we can just know the information concerning the number and the coordination of the detected faces. Therefore, to obtain the distance and angle between the detected face and an original point at the captured picture, we should calculate the distance and the angle of faces using the given information. As a result, we developed a simple face tracking system. Beside, to track only particular face among multi-faces detected by OpenCV, we used the information of color histogram which is caught from the cloths of people whose faces are detected. However, since we use only one of two web cameras, it has a disadvantage that the calculated distance and angle are less accurate than the results calculated by a method using stereo camera in spite of the advantages that it has a simple algorithm and a short execution time [11]. Therefore, we need to develop an algorithm using stereo camera in order to obtain an accurate distance and angle coordinates of detected faces.

VI. Face Tracking System

A. *Bayes Model for IROBAA*

We applied a modified Bayes model (Eq. (16)) to a robot in order to integrate audio-visual information [12].

$$P\left(\overline{F}_i \middle| T\right) = \frac{P\left(\overline{F}_i\right)\cdot P\left(T\middle|\overline{F}_i\right)}{P(T)} = \frac{P\left(\overline{F}_i\right)\cdot P\left(T\middle|\overline{F}_i\right)}{\sum_{j=1}^{k} P\left(\overline{F}_j\right)\cdot P\left(T\middle|\overline{F}_j\right)} \quad (16)$$

where '$P(F_i|T)$' means the probability that a target face 'T' is to be a detected face 'F_i', '$P(F_i)$' means the probability responding to the coordination of the detected face 'F_i' and '$P(T|\ F_i)$' means the conditional probability that each detected face 'F_i' is to be the target face 'T'. Also, 'k' denotes the total number of detected faces. That is to say, by using (20), we will be able to find the target face among the detected faces ultimately like Eq. (17).

$$\text{Target Face}=\arg\max_{i}\left\{P\left(\overline{F_i}\mid T\right)\right\} \tag{17}$$

B. Target Probability Model

Here, we can define the target probability model in order to select the target face among multi-faces effectively after a robot turns its head to the direction of the detected speech through an audition system. Since the head of the robot is tracking the target face in order to have the face located in the center of screen, we applied the Bivariate Gaussian (normal) Density Eq. (18) which has the maximum value on a center of screen to our Bayes model.

$$P\left(\overline{F_i}\right)=P\left(x_i,y_i\right)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}\left[\left(\frac{x_i-\mu_x}{\sigma_x}\right)^2+\left(\frac{y_i-\mu_y}{\sigma_y}\right)^2\right]} \tag{18}$$

In equation (22), μ is the mean value corresponding to the coordination of the center of screen and σ is the variance which can be set up by experiments.

C. Target Candidate Model

Finally, we need to define the target candidate model Eq. (19) in order to maintain classifying the target face even if new faces are detected unexpectedly. Therefore, for obtaining reliable performance with simple algorithms to reduce the execution time on computer, we used color information (histogram) corresponding to the color of clothes under each detected face. This is because the color of face depends on illumination condition and also the difference between each face is small.

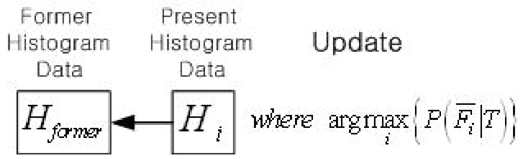
$$P\left(T\mid\overline{F_i}\right)=\left\{R_i(red)+R_i(blue)+R_i(green)\right\}/3 \tag{19}$$

The Eq. (19) means the probability calculated using histogram data from three colors (red, blue, green) of the upper clothes concerning each detected face. Here, each R_i is the correlation results between histogram data of present detected faces, $H_i(d)$, and that of the former selected target face, $H_{former}(d)$, with regard to the corresponding color by using Eq. (20).

$$R_i(color)=\frac{\sum_{d=1}^{256}\{H_i(d)\cdot H_{former}(d)\}}{\sqrt{\sum_{d=1}^{256}H_i(d)^2}\sqrt{\sum_{d=1}^{256}H_{former}(d)^2}} \tag{20}$$

D. Update

Finally, after a robot obtains the information of a target face by using Eq. (16), it has to update the histogram data about the target face so as to compare with all the faces at the next frame. That is expressed as Eq. (21).


Former Histogram Data
Present Histogram Data
Update
H_{former}
H_i
where $\arg\max_i\left\{P\left(\overline{F_i}\middle|T\right)\right\}$

(21)

VII. Audio-Visual integration

As the results of this research, we can mainly get the two merits. First of all, collaborating with vision systems can help a robot compensate the errors in the sound localization. According to the results of our previous experiments [6], we could confirm excellent performance at the short distance (1m): the percentage of successful detection of sound's direction is 90.3% and the average of errors about the estimated sound's direction is 5.1°. However, results at 2m distance show poor performance. Therefore, to alleviate this problem, we improved the performance of VAD and integrated with audio and visual information and, consequently, we have acquired the good results as shown in Table I. Especially, if a detected face exists in the screen, after a robot turns its head to the sound's direction, face tracking system can compensate the angle error, resulted from the sound localization, within the range of the file of view (± 18°). In this reason, we could get the excellent angle accuracy.

TABLE I The experimental results at the distance of 2m.

	Successful detection of sound's direction		Angle error of sound's direction	
Method	Previous	Proposed	Previous	Proposed
Average	63.9%	82%	13.1°	1.7°

Second, collaborating with vision systems can help a robot effectively reject unnecessary speech or noise signals entering from the undesired directions. That will make the performance of speech recognition improved. Therefore, IROBAA can perform the following scenario or sequence in Figure 6. Firstly IROBAA recognizes the voice command and the direction of the voice as well when someone calls. Then it turns its face to the direction, and can recognize someone's face through the vision system. After that, it will track the face in order to communicate with the recognized person. At that time, if the robot catches a new voice command or noise signal entering from other directions except for the direction of a selected speaker, the robot will reject the voice or the signal so as to talk with a particular speaker efficiently in noisy environment. Also, a robot can track only the selected speaker even if other faces are detected randomly. Finally, if a particular speaker is disappeared, the robot will stand by until it finds a new voice command and the corresponding target face. However, if a particular speaker isn't detected, it will try finding the target again within two steps because OpenCV isn't always able to detect a particular face perfectly.

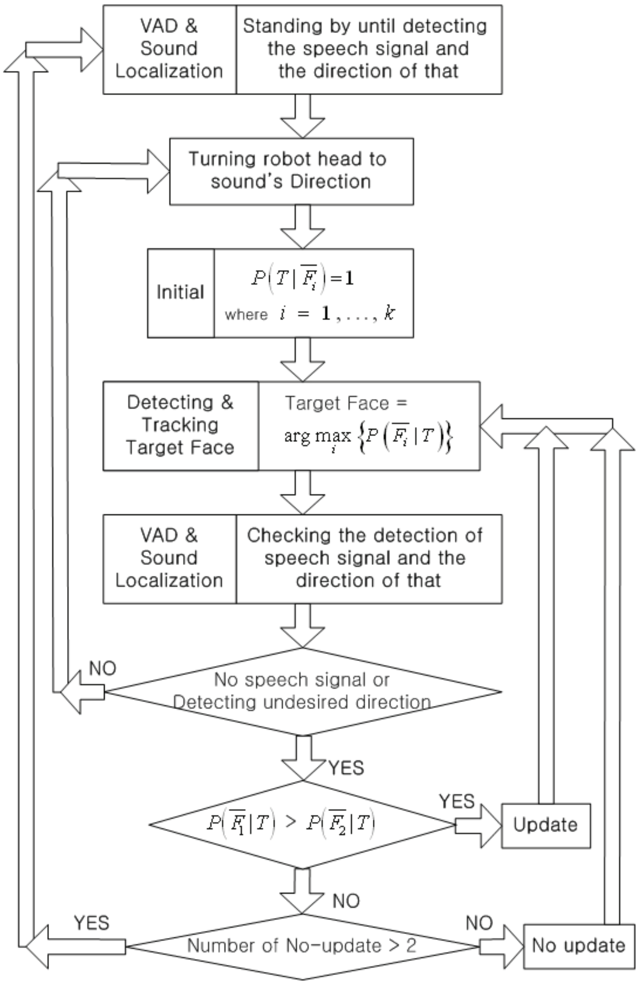

VAD & Sound Localization
Standing by until detecting the speech signal and the direction of that
Turning robot head to sound's Direction
Initial
$P(T|\overline{F}_i)=1$
where $i = 1, \ldots, k$
Detecting & Tracking Target Face
Target Face = $\arg\max_i \{P(\overline{F}_i|T)\}$
VAD & Sound Localization
Checking the detection of speech signal and the direction of that
NO
No speech signal or Detecting undesired direction
YES
$P(\overline{F}_1|T) > P(\overline{F}_2|T)$
YES
Update
NO
YES
Number of No-update > 2
NO
No update

Fig. 7 Sequence of algorithm of IROBAA

VIII. Conclusion

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, it can get advantages to increase the detectible distance of sound's signal and to reduce noise. Second, a probabilistic sound localization with post-filtering has been proposed, which shows more reliable performances. Finally, by integrating visual and auditory processing technology, we were able to extend this research to particular speaker localization among multiple faces in noisy environment for the purpose of effective interaction between a human being and a robot.

However, since our research is just first step on the aim to implant a perception into robots, we have a lot of problems which we should overcome. Especially, 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. Of course, needless to say, improving a vision system is surely necessary for human robot interaction. Consequently, we should well integrate diverse information generated by audio and visual systems in order to realize the human robot interaction which we are regarding as a difficult technology in real environment. In addition, for the fusion of visual-audio information, we should consider applying artificial intelligent methods to robots.

References

[1] J. Huang, N. Ohnishi, and N. Sugie, "A Biomimetic System for Localization and Separatioin of Multiple Sound Sources," *in Proc. IEEE/IMTC Int. Conf. Intstrumentation and Measurement Technology*, Hamamatsu Japan, May 1994, pp. 967-970.

[2] J. Huang, N. Ohnishi, and N. Sugie, "Sound Localization in Reverberant Environment Based on the Model of the Precedence Effect," *IEEE Trans. on Instrumentation and Measurement*, vol. 46, no 4, pp. 842-846, 1997.

[3] J. Huang, T. Supaongprapa, I. Terakura, N. Ohnishi, and N. Sugie, "Mobile Robot and Sound Localization," *in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Grenoble France, Sep. 1997, pp. 683-689.

[4] J. Huang, N. Ohnishi, and N. Sugie, "Spatial localization of sound sources: azimuth and elevation estimation," *in Proc. IEEE/IMTC Int. Conf. Instrumentation and Measurement Technology*, St. Paul, MN USA, May 1998, pp. 330-333.

[5] J. Huang, K. Kume, and A. Saji, "Robotics Spatial Sound Localization and its 3D Sound Human Interface," *in Proc. IEEE Int. Sym. Cyber Worlds*, pp. 191-197, 2002.

[6] H. D. Kim, J. S. Choi, C. H. Lee, and M. S. Kim, "Reliable Detection of Sound's Direction for Human Robot Interaction," *in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Sendai Japan, Sep. 2004, pp.2411-2416.

[7] H. G.. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano, "Human-Robot Interaction through Real-Time Auditory and Visual Multiple-Talker Tracking," *in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Hawaii USA, Oct. 2001, pp. 1402-1409.

[8] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-Time Speaker Localization and Speech Separation by Audio-Visual Integration," *in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Washington DC USA, May 2002, pp. 1043-1049.

[9] H. Kobayashi, and T. Shimamura, "A Modified Cepstrum Method for Pitch Extraction," *IEEE/APCCAS Int. Conf. Circuits and Systems*, Nov. 1988, pp. 299-302.

[10] S. Ahmadi, and A. S. Spanias, "Cepstrum-Based Detection Using a New Statistical V/UV Classification Algorithm," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no 3, pp. 333-338, 1999.

[11] R. Y. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," *IEEE Journal of Robotics and Automation*, vol. 3, no 4, pp. 323-344, 1987.

[12] 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," *in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Sendai, Japan, Sep 2004, pp. 2404-2410.