Human-Robot Interaction in Real Environments by Audio-Visual Integration

Hyun-Don Kim, Jong-Suk Choi*, and Munsang Kim

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

Keywords: Audio-visual integration, face tracking, human-robot interaction, sound source localization, voice activity detection.

1. INTRODUCTION

In the near future, we expect participation of intelligent robots to grow rapidly in human society. Therefore, since effective interaction between robots and average people will be essential, robots need to be able to identify a speaker among a group of people and recognize speech signals in a real environment. For example, in order to recognize speech with high confidence, the techniques that 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 behavior need to integrate with visual and auditory information in order that they become friendly toward human beings. One of the reasons for integrating with visual and auditory information is to locate a speaker who wants to talk with a robot effectively. This is because robots need to locate a speaker so as to perform speech recognition and sound source separation. If they succeed in locating a desired speaker, that can help them to improve those performances. Therefore, many robot experts have a growing concern as to how they can integrate effectively with visual and auditory information as well as data from various sensors.

The objective of this research is to develop techniques that enable speaker localization by audio-visual integration. In detail, detecting intervals of speech signal, finding its direction and turning a robot's head to the direction of a speaker's face can help average people to interact with robots naturally [1-6]. Besides, it is necessary to use the visual processing technology that can support robots to detect and track a specific speaker's face. 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 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].

In conventional systems to locate a speaker, auditory systems such as voice activity detection (VAD) and sound source localization have mainly been used in robots. However, finding a speaker only

Manuscript received October 3, 2006; revised June 12, 2006; accepted September 8, 2006. Recommended by Editorial Board member Sooyong Lee under the direction of Editor Jae-Bok Song. This research was supported by Development of Active Audition System Technology for Intelligent Robots through Center for Intelligent Robotics.

Hyun-Don Kim had been with Intelligent Robotics Research Center at KIST and now moved to Speech Media Processing Group in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan (e-mail: hyundon@kuis.kyoto-u.ac.jp).

Jong-Suk Choi is with the Intelligent Robotics Research Center at KIST, 39-1, Hawolgok-dong, Seongbuk-gu, Seoul 136-791, Korea (e-mail: cjs@kist.re.kr).

Munsang Kim is with the Center for Intelligent Robotics, Frontier 21 Program at KIST, 39-1, Hawolgok-dong, Seong buk-gu, Seoul 136-791, Korea (e-mail: munsang@kist. re.kr).

* Corresponding author.

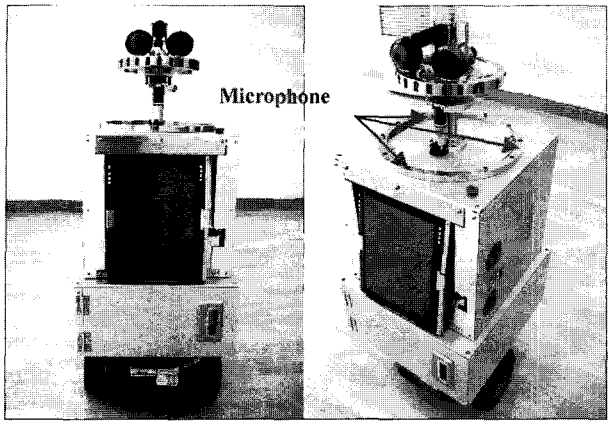


Fig. 1. IROBAA (Intelligent ROBot for Active Audition).

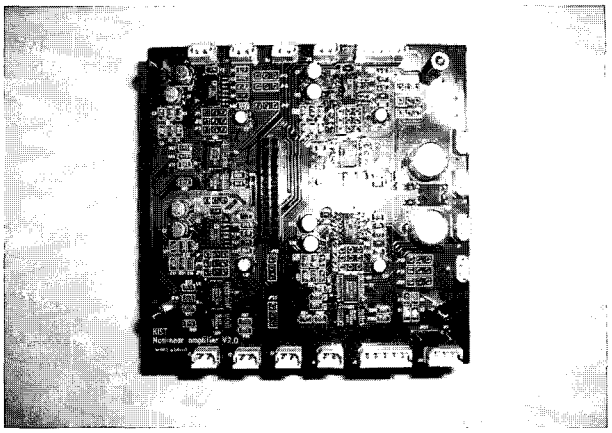


Fig. 2. Developed nonlinear pre-amp. board.

by auditory systems has frequently failed because speech signals are hard to extract in a noisy environment. For this reason, we propose a way to integrate audio-visual information for classifying speech signals and locating a speaker even if a new person appears for tracking a desired person. Furthermore, in comparison to the similar studies related to a human-robot interaction by audio-visual integration, our system has some advantages to deal with locating a speaker at whole direction (0°~360°) using just three microphones. Other systems can deal with the front area (0°~180°) in the case of using two or three microphones. Also, while there exists sound source localization, voice activity detection and a face tracking system, our system has a computer for calculation because our proposed algorithms are simple and compact. Other systems which have similar abilities usually have more than two computers for audio and visual processing.

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 involves 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.

2. 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 be small, the signal of speech occurring at the long distance can be hardly extracted from its received signal whose magnitude is so small that the contents of speech are cancelled by noise. On the contrary, with a large ratio, the signal occurring nearby may be saturated in the AD conversion. To solve this problem, we propose the nonlinear amplification where smaller signals can be amplified with larger amplification ratio. To implement the nonlinear property, we used SSM2166, made by the 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.

3. SOUND LOCALIZATION

3.1. Tracking of sound's direction

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

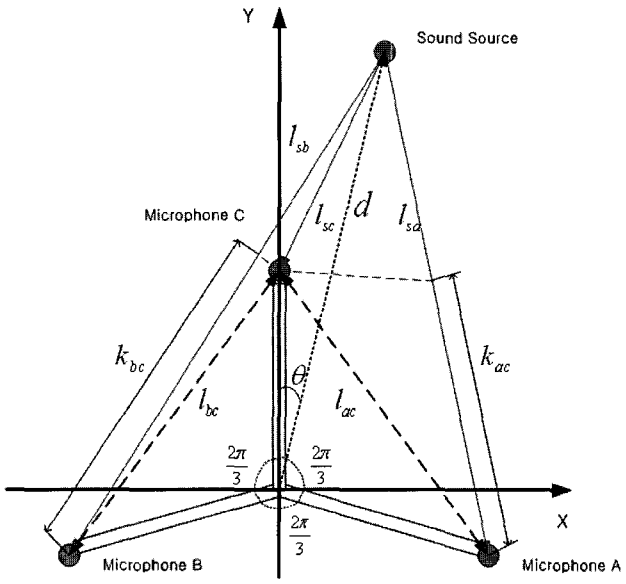


Fig. 3. Location of three microphones.

between array-shaped microphones.

In Fig. 3, three microphones are arranged such that their distances from the center of the triangular rod are the same. Two couples of A vs. C and B vs. C are selected in the viewpoint 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. In this case, the relative distance corresponding to the maximum delay is defined as l_{ac} (or l_{bc}). Also, the distance between the 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. Equations (3) and (4) are 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 consisting of 800 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 $-n_{ac} \sim n_{ac}$ in (3) and $-n_{bc} \sim n_{bc}$ in (4) where its positive/negative value means that the sound enters microphones A and B earlier/later than microphone C. Now, the sound's direction should be calculated using relation coefficient R_{ac} and R_{bc} for all possible k_{ac} and k_{bc}. Fig. 3 illustrates the number of delay samples and the actual angle of the sound's direction. An actual delay of the 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).

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

Next, because we want to find the angle of the 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).

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

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

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

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

3.2. Reliable detection of sound's direction

In a real speech signal, as there are reverberations, noise signals and consonants that have weakly periodic signals, incorrect detections of sound's directions are calculated by computer frequently. Therefore, in order to find accurate directions of speech signal, we should detect the sound's direction at the frame that has the maximum energy within a period of speech signal. 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 that is not a period of speech signal. Second, because the frame having a maximum energy does not always have good data to find an accurate direction of sound, accuracy related to detecting the sound's direction can be reduced. To fix these problems, we propose a new performance index rather

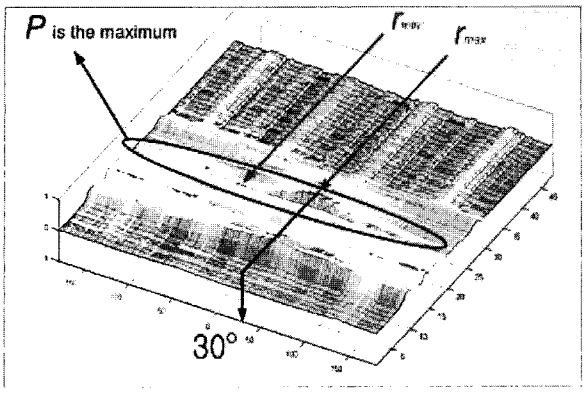


Fig. 4. The 3D graph of cross-correlation.

Table 1. Compare frame energy with proposed index at 1m distance.

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

than the frame energy. Given each frame, the performance index is expressed as (11).

$$P = r_{\max} - r_{\min} \tag{11}$$

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 assist in finding reliable detection of the 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. Fig. 4 illustrates a 3-dimension graph that 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°. When a frame has the proposed performance index with the largest value throughout all the frames (see the inside of blue circle in Fig. 4), we can find an accurate direction of sound.

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." The spots of generating each command were total 13 points at a distance of 1 meter. The azimuth, which ranges from -90° to 90°, was divided at intervals of 15. Table 1 is the average of experimental results. As a result, the cross-correlation method is better in the percentage of successful detection and in the average of angle error than the frame energy method as you see in Table 1.

4. VOICE ACTIVITY DETECTION

For the purpose of effective interaction between a person 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 the cepstrum to find pitch information [9]. The word 'cepstrum' is used to indicate 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 the cepstrum is that if the signal is periodic, the signal made by the cepstrum will also present peak signals at intervals of each period. Furthermore, compared to pitch detection method using autocorrelation at time domain, the cepstrum has distinct peaks at intervals of each period and the first peak is always bigger than the second or the third one. Consequently, the cepstrum can reliably extract the pitch of a speech signal. Given a signal $x(t)$, the equation of the cepstrum is expressed as (12).

$$c_c(\tau) = F^{-1}\{\log X(f)\} = F^{-1}\{\log|X(f)| + j\phi_x(f)\} \tag{12}$$

Fig. 5 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

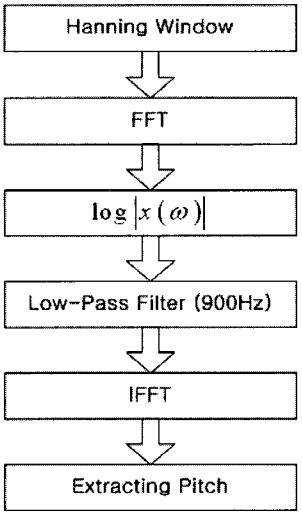


Fig. 5. Procedure of the method extracting pitch.

voice.

Therefore, to minimize the disturbance of noises when a robot tries to extract pitches, we apply a low pass filter that has the range between 0 and 900Hz to the pitch-detection algorithm.

Finally, with the number of samples between two peak signals found, the pitch can be detected by (13).

$$Pitch = \frac{\text{Sampling Frequency}}{\text{A number of samlpes between the two peaks}} \tag{13}$$

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 (14).

$$E_{frame} = \frac{1}{k}\sum_{i=0}^{k} x^2(i), \tag{14}$$

where $x(i)$ means the sampling data of i-th step and k is the number of steps. The ZCR indicates how many times the sign of signals are changed at the period of a frame. The ZCR is expressed as (15).

$$ZCR = \sum_{i=0}^{N-1} \left|\text{sgn}\left[x(i)\right] - \text{sgn}\left[x(i+1)\right]\right| \times \frac{1}{2} \tag{15}$$

In the interval of noise signals or consonants that 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 in 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]:

$$R_C = \left\{C_i \middle| \min(F) < C_i < \max(F)\right\}, \tag{16}$$

$$R_Z = \left\{ZCR_i \middle| \min(Z) < ZCR_i < \max(Z)\right\}, \tag{17}$$

$$R_E = \left\{E_i \middle| \min(E) < E_i < \max(E)\right\}, \tag{18}$$

where 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:

$$\left[(C_i \in R_C) \Lambda (ZCR_i \in R_Z) \Lambda (E_i \in R_E)\right] \Rightarrow (i \in Voice), \tag{19}$$

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

Besides, since the A/D converter that 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 the microphones.

5. 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, as can be seen in Fig. 6, we should calculate the distance and angle between the detected face and the center of a camera lens at the captured picture.

Firstly, we can get an estimated distance between the center of a camera lens and an original point by (20).

$$D_{est} = \frac{P_{ref}}{P_{obs}} \cdot D_{ref}, \tag{20}$$

where D_{ref} is a reference distance, P_{ref} is the

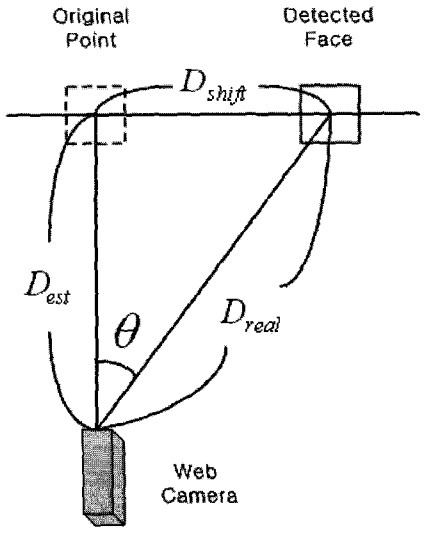


Fig. 6. The illustration of an estimated distance and angle.

number of reference pixels corresponding to the reference distance, and P_{obs} is the number of observed pixels corresponding to a detected face.

Second, we can calculate the distance between the center of a detected face and an original point by (21).

$$D_{shift} = \frac{D_{est}}{D_{ref}} \cdot P_{shift} \cdot \alpha, \tag{21}$$

where P_{shift} is the number of pixels between the detected face and the original point and α is the gap between pixels at the reference distance. Then we can get the angle between the center of a detected face and the original point by (22).

$$\theta = \tan^{-1}\left(\frac{D_{shift}}{D_{est}}\right) \tag{22}$$

Finally, we can get the real distance between the center of detected face and the original point by (23).

$$D_{real} = \sqrt{D^2_{est} + D^2_{shift}} \tag{23}$$

As a result, we developed a simple face tracking system. In order to track only a particular face among multi-faces detected by OpenCV, we used the information of a color histogram that is caught from the clothing 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 a 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 a stereo camera in order to obtain an accurate distance and the angle coordinates of detected faces.

6. FACE TRACKING SYSTEM

6.1. Bayes model for IROBAA

We applied a modified Bayes model (24) to a robot in order to integrate audio-visual information [12].

$$P(\overline{F}_i|T) = \frac{P(\overline{F}) \cdot P(T|\overline{F}_i)}{P(T)} = \frac{P(\overline{F}) \cdot P(T|\overline{F}_i)}{\sum_{j=1}^{k} P(\overline{F}) \cdot P(T|\overline{F}_i)}, \tag{24}$$

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 '$P(T|F_i)$' signifies 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 (24), we will be able to find the target face among the detected faces ultimately like shown in (25).

$$\text{Target Face} = \arg\max_i \left\{P(\overline{F_i}|T)\right\} \tag{25}$$

6.2. 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 the screen, we applied the Bivariate Gaussian (normal) Density (26), which has the maximum value on a center of the screen, to our Bayes model.

$$P(\overline{F_i}) = P(x_i, y_i) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x_i-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y_i-\mu_y}{\sigma_y}\right)^2\right]} \tag{26}$$

In (26), μ is the mean value correspond-ing to the coordination of the center of the screen and σ is the variance that can be set up by the experiments.

6.3. Target candidate model

Finally, we need to define the target candidate model (27) 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 clothing color under each detected face. This is because the color of the face depends on the illumination condition and also the difference between each face is small.

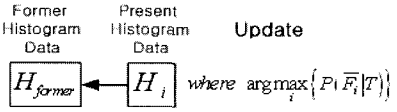
$$P(T|\overline{F_i}) = \{R_i(red) + R_i(blue) + R_i(green)\}/3 \tag{27}$$

Equation (27) indicates the probability calculated using histogram data from three colors (red, blue, green) of the upper clothing concerning each detected face. Here, each R_i expresses the correlation results between histogram data of the 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 (28).

$$R_i(color) = \frac{\sum_{d=1}^{256} \{H_i(d) \cdot H_{former}(d)\}}{\sqrt{\sum_{d=1}^{256} H_i(d)^2} \sqrt{\sum_{d=1}^{256} H_{former}(d)^2}} \tag{28}$$

6.4. Update

Finally, after a robot obtains the information of a target face by using (25), it has to update the histogram data pertaining to the target face so as to compare with all the faces at the next frame. That is expressed as (29).



(29)

7. AUDIO-VISUAL INTEGRATION

Two merits have been revealed as a result of this research. First of all, collaborating with vision systems can help a robot compensate the errors in sound source localization. According to the results of our previous experiments [6], we could confirm excellent performance using only audio information at a short distance (1m) as shown in Table 2 the percentage of successful detection of the sound's direction is 90.3%. Then the average of errors and standard deviation concerning the estimated sound's direction are 5.1° and 4.6°, respectively. Moreover, once a robot locates a face after it has turned its head towards the sound's direction, it can compensate the angle error and start tracking the face by visual information. After that, even if other speakers appear in the screen, a robot can distinguish the tracking face using histogram data from the upper clothing regardless of the distance. For this reason, the angle error was deceased ($1° \pm 1°$) in case of integrating with audio and visual information. However, we obtained the same success rate (90.3%) in regard to successful detection of the sound's direction irrespective of integrating visual information. This is because the angle errors at 1m distance are almost out of the field of view ($\pm 18°$) for our camera whenever a robot fails to find the direction of sound at a short distance.

On the other hand, results at the 2m distance show poor performance only using audio information. Therefore, to alleviate this problem, we integrated with audio and visual information. Consequently, we acquired good results as shown in Table 2. Especially, we cannot consider doing an experiment at 3m distance because of certain factors. One is that human-robot interaction is normally carried out within a 2m distance. The other is that our system cannot determine a face over 2m away and the performance of sound source localization at long distance is also not good.

Second, collaborating with vision systems can help a robot effectively reject unnecessary speech or noise signals entering from undesired directions. That will make the performance of speech recognition improved. Therefore, IROBAA can perform the following scenario or sequence. (1): 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. (2): After that, it will track the face in order to communicate with the recognized person. Also, a robot can track only the selected speaker even if other faces are detected randomly. (3): 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 a noisy environment. (4): Finally, if a particular speaker is disappeared, it will try finding the target again within two steps because OpenCV isn't always able to detect a particular face perfectly. However, once losing the target face (that is to say, when IROBAA can't detect the target face over three frames), the robot will stand by until it finds a new voice command and the corresponding target face.

Fig. 7 shows the algorithm sequence of IROBAA corresponding to the scenario and Fig. 8 shows the GUI of the application program which is developed by gcc on Linux. The application program for IROBAA consists of three windows. The left-up window shows the captured picture by a web camera and detected faces as well. Especially, the black box represents the target face. On the contrary, red boxes represent the detected faces. Also, all blue boxes represent the area of the clothes to catch the histogram data. The right-up window shows not only a distance and angle from the camera to the detected faces but also audio information such as pitch frequency, voice's direction and frame energy. The bottom window reveals sampled signals entered from three microphones and speech signals extracted by VAD. This program has all algorithms run at intervals of 0.5 seconds. As soon as we run this program, IROBAA performs a programmed scenario.

Table 2. Experiment results at 1m and 2m distances.

		Successful detection of sound's direction		Angle error of sound's direction	
Method		Audio information	Audio-Visual integration	Audio information	Audio-Visual integration
Average	1m	90.3%	90.3%	$5.1° \pm 4.6°$	$1° \pm 1°$
	2m	63.9%	80%	$13.1° \pm 13.1°$	$1.7° \pm 2°$

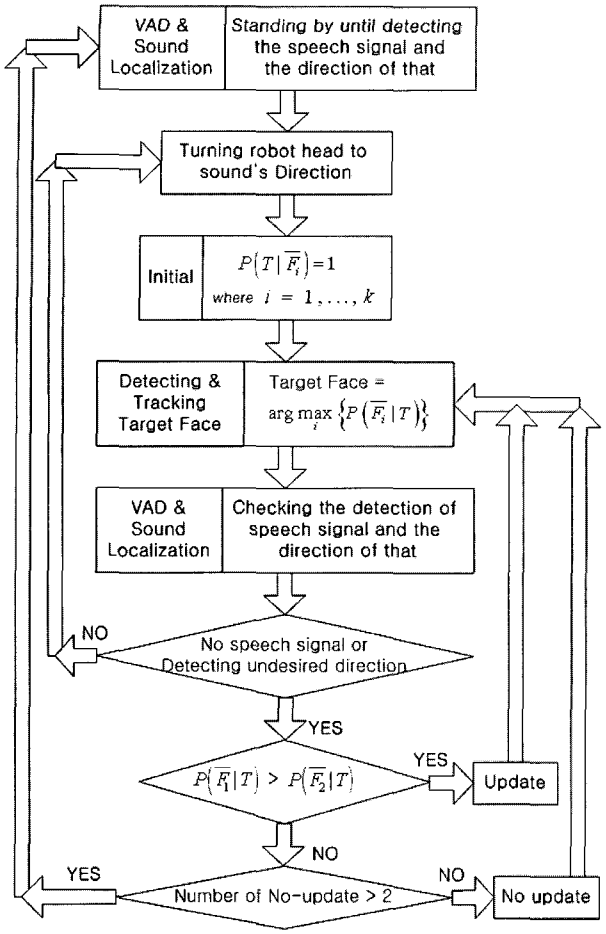


Fig. 7. Sequence of algorithm of IROBAA.

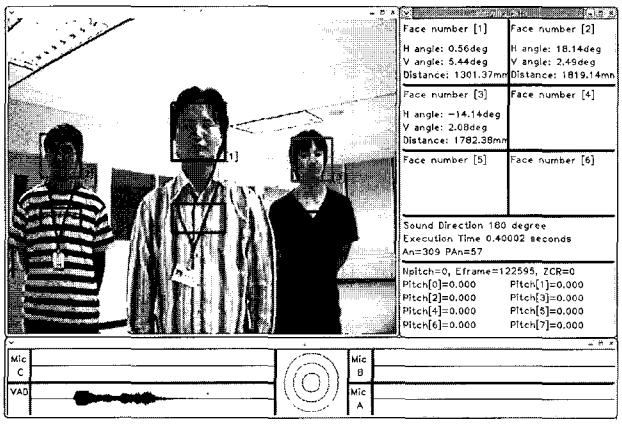


Fig. 8. GUI of the application program for IROBAA.

8. CONCLUSIONS

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 the sound's signal and to reduce noises. Second, detecting the interval and the direction of speech signal can help normal people to interact with robots naturally. Finally, by integrating visual and auditory processing technology, we were able to extend this research to particular speaker localization among multiple faces in noisy environments for the purpose of effective interaction between a human being and a robot.

However, since our research is just the first step toward implementing a kind of perception into robots, we have a lot of problems to overcome. Especially, for further application to real life, the system should extract the desired signal when voices of several people are mixed. Also, it should eliminate noises even though large ones are mixed with small ones. Of course, needless to say, improving the 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 the real environment. In addition, for the advanced fusion of audio-visual information, we should consider applying artificial intelligence to robots.

REFERENCES

[1] J. Huang, N. Ohnishi, and N. Sugie, "A biomimetic system for localization and separation of multiple sound sources," *Proc. of IEEE/IMTC Int. Conf. Instrumentation and Measurement Technology*, Hamamatsu Japan, pp. 967-970, May 1994.

[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," *Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Grenoble France, pp. 683-689, Sep. 1997.

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

[5] J. Huang, K. Kume, and A. Saji, "Robotics spatial sound localization and its 3d sound human interface," *Proc. of 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," *Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Sendai Japan, pp. 2411-2416, Sep. 2004.

[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," *Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Hawaii, USA, pp. 1402-1409, Oct. 2001.

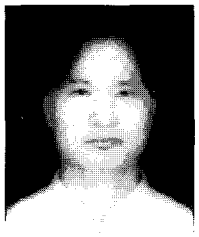
[8] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-time speaker localization and speech separation by audio-visual integration," *Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Washington, DC, USA, pp. 1043-1049, May 2002.

[9] H. Kobayashi and T. Shimamura, "A modified cepstrum method for pitch extraction," *Proc. of IEEE/APCCAS Int. Conf. Circuits and Systems*, pp. 299-302, Nov. 1988.

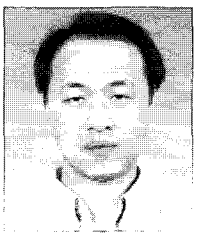
[10] S. Ahmadi and A. S. Spanias, "Cepstrum-based detection using a new statistical V/UV classification algorithm," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 3, pp. 333-338, 1999.

[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.

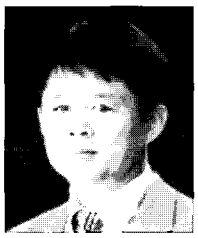
[12] I. Hara, F. Asano, Y. Kawai, F. Kanehiro, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," *Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Sendai, Japan, pp. 2404-2410, Sep. 2004.



Hyun-Don Kim received the B.S. degree in Control and Instrumentation Engineering from Korea University in 1997 and the M.S. degree in Electrical Engineering from Korea University in 2004. As of 2005, he has been a Ph.D. student with the Speech Media Processing Group, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto Japan. His research interests include sound signal processing, humanoid robot, vision system and artificial intelligence.



Jong-Suk Choi received the B.S., M.S., and Ph.D. in Electrical Engineering from the Korea Advanced Institute of Science and Technology in 1994, 1996, and 2001. In 2001, he joined the Intelligent Robotics Research Center, Korea Institute of Science and Technology (KIST), Seoul Korea as a Research Scientist, and now is a Senior Research Scientist at KIST. His research interests include signal processing, mobile robot navigation and localization.



Munsang Kim received the B.S. and M.S. degrees in Mechanical Engineering from Seoul National University in 1980 and 1982 respectively, and the Ph.D. in Robotics from the Technical University of Berlin, Germany in 1987. Since 1987, he has been working as a Research Scientist at the Korea Institute of Science and Technology (KIST), Korea where he is now a Principal Research Scientist. Also, he has been a Director at the 'Center for Intelligent Robots – The Frontier 21C Program' since Oct. 2003. His research interests include design and control of novel mobile manipulation systems, haptic device design and control, and sensor application to intelligent robots.