The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Sound Source Localization in Reverberant Environment using Visual information

Byoung-gi Lee, JongSuk Choi, Daijin Kim, and Munsang Kim

Abstract—**Recently, many researchers have carried out works on audio-video integration. It is worth exploring because service robots are supposed to interact with human beings using both visual and auditory sensors. In this paper, we propose an audio-video method for sound source localization in reverberant environment. Using visual information from a vision camera, we could train our audio localizer to distinguish a real source from fake sources and improved the performance of audio localizer in reverberant environment.**

I. Introduction

HUMAN beings have several sensors to detect and understand real world where they live. They look by their eyes, hear by their ears, feel by their skin, taste by their tongues and smell by their noses. All these sensors are working together for our brain to imagine our surroundings vividly. Since each sensor has its advantages and also disadvantages, a combination of two or more sensors performs much more efficiently. Since eyes and ears are the most important sensors of human sensors, many researchers have tried to design a system where audition and vision are working together. Lathoud et al. provided a corpus of audio-visual data, called "AV16.3" [1]. It was recorded in a meeting room where 3 cameras and two 8-microphone arrays are equipped. It targeted researches on audio-visual speaker tracking. Busso et al. developed a smart room which can identify the active speaker and participants in a casual meeting situation [2]. They used 4 CCD cameras, an omni-directional camera and 16 microphones distributed in the room. They showed that complementary modalities could increase the smart room's performance of identification and localization. With intelligent meeting room, mobile service robot is also a prospective research area of audio-video fusion. Lim et al. developed a mobile robot which can track multiple people and select the current speaker of them by sound source localization and face detection [3]. Their robot could associate sound event with vision event and make audio-video information fusion using particle filter. Nakadai et al. designed a robot audition system for humanoid SIG [4]. SIG also associated auditory stream and visual stream to tracking people when they are speaking and moving.

In this paper, we give another example of audio-video complementary system which is a little different from previous audio-video system in that it is not simply fusing two modalities but focusing on improving auditory performance with a help of vision. One of the most difficult problems of sound source localization is that the performance is easily messed up in the echoic environments. In a closed room, each wall, ceiling and floor cause to reflect sound waves. They make many fake sound sources and impede proper sound source localization. As you know, the reflected sound is almost the same as the original sound contrary to the other interfering noises. It is why reverberant condition is worse than noisy condition.

In this paper, we propose a method of sound source localization in a reverberant environment using visual information. Our motivation is simple and natural. If we see some sound sources by our eyes, we can learn how to distinguish real sound sources from virtual sound sources, and finally adapt our ears to an echoic room. In the proposed method, we train a neural network as a verifier which would validate the result of the sound source localization in each frame. When a person is captured by a camera, this verifier is learning and when he speaks out of vision's view, it would improve the performance of sound source localization.

In the next section, we present our basic algorithm of sound source localization system. In the section III, we propose features and talk about how to verify them and how to train a neural network using visual information. In the section IV, we provide experimental results of the proposed method and in the final section, we conclude our method and mention about further work.

II. Sound Source Localization

A. Microphone Array

We've used a 3-microphone array system for sound source localization. We pursue a small and light system with smart and strong performance. Our microphone array is within 7.5cm radius circle. We put 3 microphones on the vertices of equilateral triangle in the free field. We assume no obstacle from a sound source to each microphone, which means no HRTF (head related transfer function) is required and makes the localization very simple and its performance very even with no angle dependency. But its disadvantage is that the

Manuscript received February 28, 2010. This work was supported in part by the Korea Ministry of Knowledge Economy under the 21st century Frontier project.

Byoung-gi Lee is with the Center for Cognitive Robot Research, Korea Institute of Science and Technology, Seoul, Korea (e-mail: leebg03@kist.re.kr).

JongSuk Choi is with the Center for Cognitive Robot Research, Korea Institute of Science and Technology, Seoul, Korea (phone: +82-2-958-5618; fax: +82-2-958-5629; e-mail: cjs@kist.re.kr).

Daijin Kim is with the Dept. Computer Science and Engineering, Pohang University of Science and Technology, Korea (e-mail: dkim@postech.ac.kr).

Munsang Kim is with the Center for Intelligent Robotics, Korea Institute of Science and Technology, Seoul, Korea (e-mail: munsang@kist.re.kr).

smallest number of microphones which doesn't suffer from the front-back confusion is three, while a system using HRTF needs just two. Fig. 1 shows our triangular microphone array.

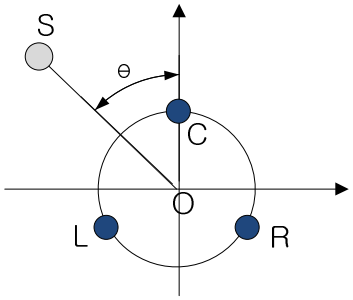


Fig. 1. Arrangement of 3-microphone array

B. Angle-TDOA Map

From our assumption of no HRTF, we can easily calculate TDOAs (time delay of arrival) between microphones by geometric relations. TDOA is determined by the position of sound source and actually it depends on almost only the direction of sound source [5].

We can survey the relation between the azimuth angle of sound source and TDOAs which is given by (1).

$$\begin{cases} TD_{LC} = \dfrac{\overline{SL} - \overline{SC}}{v_{sound}} \\ TD_{CR} = \dfrac{\overline{SC} - \overline{SR}}{v_{sound}} \\ TD_{RL} = \dfrac{\overline{SR} - \overline{SL}}{v_{sound}} \end{cases} \quad (1)$$

, where v_{sound} is the speed of sound in the air.

After surveying, we can get a TDOA map of source angle. We call it Angle-TDOA Map and denote it as (2).

$$\begin{cases} TD_{LC} = \tau_{LC}(\theta) \\ TD_{CR} = \tau_{CR}(\theta) \\ TD_{RL} = \tau_{RL}(\theta) \end{cases} \quad (2)$$

Angle-TDOA Map is the essential part of TDOA-based sound source localization method. Its inverse map tells us where the sound source from measured TDOAs.

C. Cross-Angle-Correlation function

Generally, TDOAs are measured by cross-correlation or its variations such as GCC (generalized cross-correlation) and CPSP (cross-power spectrum phase) [6]. In our localization system, we use cross-correlation in a unique way. We intermingle cross-correlation with Angle-TDOA Map. We call the intermingled result Cross-Angle-Correlation function.

Cross-correlation is to compare two signals crossing all possible time delays. By Cross-Angle-Correlation, we want to compare two signals crossing all possible source angles. It is possible by composite function of cross-correlation and Angle-TDOA Map.

$$\begin{cases} R_{LC}(\theta) = r_{LC}\left(\tau_{LC}(\theta)\right) \\ R_{CR}(\theta) = r_{CR}\left(\tau_{CR}(\theta)\right) \\ R_{RL}(\theta) = r_{RL}\left(\tau_{RL}(\theta)\right) \end{cases} \quad (3)$$

, where r_{LC}, r_{CR}, and r_{RL} are cross-correlation functions.

We integrate these functions of (3) in the way of (4) and call the integrated result Cross-Angle-Correlation function.

$$R(\theta) = \left[\bar{R}_{LC}(\theta)\bar{R}_{CR}(\theta) + \bar{R}_{CR}(\theta)\bar{R}_{RL}(\theta) + \bar{R}_{RL}(\theta)\bar{R}_{LC}(\theta)\right]/3 \quad (4)$$
$$, where \quad \bar{R}_{AB}(\theta) = \max\left(0, \quad R_{AB}(\theta)\right)$$

Fig. 2 shows an example of Cross-Angle-Correlation function. While Cross-Correlation gives us time information of the detected sound, Cross-Angle-Correlation gives us spatial information of the detected sound.

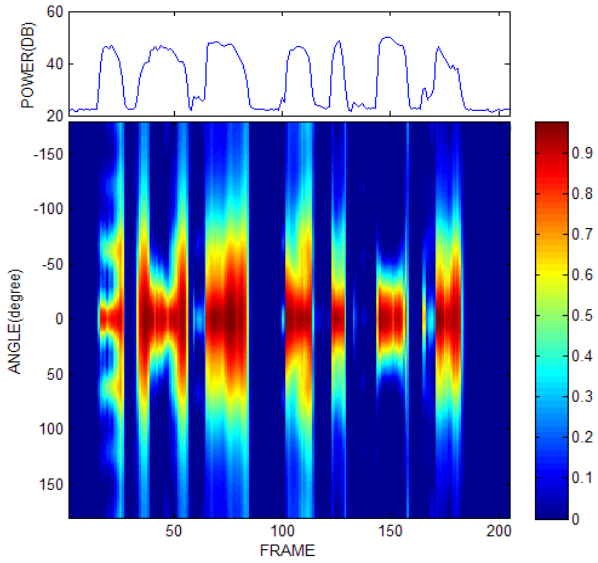


Fig. 2. An example of Cross-Angle-Correlation (bottom) and the power of signal (upper)

※1. Simulated signal : angle 0 / sampling rate 16kHz
2. Frame : shift 15msec / length 20msec

As you can see from Fig. 2, Cross-Angle-Correlation function has high values at directions from which sound is coming but it is somewhat blurred depending on the temporal characteristic of sound. Also, it is most likely that in a very short time interval, only one sound source among multiple sound sources is dominant to the other sources and can be detected by the original Cross-Angle-Correlation[7]. Therefore, instead of Cross-Angle-Correlation, we take a Gaussian function located on the maximum point of

Cross-Angle-Correlation function for each frame.

$$\hat{R}(\theta) = R_{\max} \cdot \exp\left(-\frac{(\theta-\theta_{\max})^2}{50}\right) \quad (5)$$

$$\text{, where } \begin{cases} R_{\max} = \max_{\theta} R(\theta) \\ \theta_{\max} = \arg\max_{\theta} R(\theta) \end{cases}$$

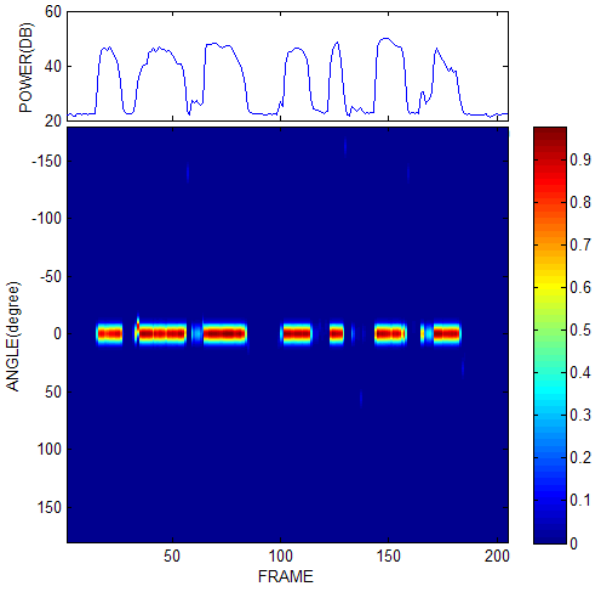


Fig. 3. Transformed image of Fig.2 by Gaussian function

III. Real Source Verification

A. Visual Information: Face Detection

We want our sound source localization system to learn how to distinguish real sources from fake sources. Vision camera can give us useful information. We assumed that we are interested in only human voice and determined to use face detection module to get visual information. It is a good approach because other sound from a dog, TV or a vacuum cleaner is considered as interfering noise in the situation of human-machine interaction.

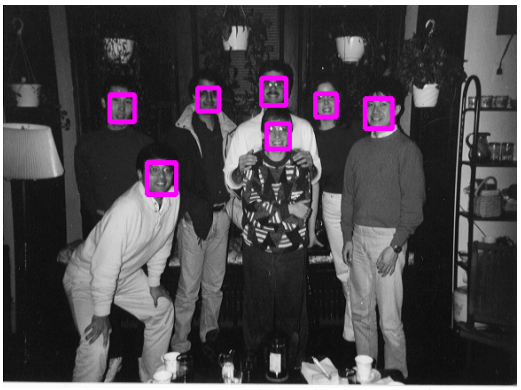


Fig. 4. An example of face detection result

Intelligent Media Lab, Postech provided us face detection module [8]. It can process about 23 frames per second and tell us the number of detected faces and their rectangular regions in the picture. From it, we can know the angles of which people are standing [9].

B. Sound Feature Extraction

We want to make a feature that could characterize the direct-path sound and reflected sound. We took notice of Precedence effect [10]. It is a well-known phenomenon which explains how human being improves his sound source localization in a reverberant environment. According to Precedence effect, in the human auditory system, lagging spatial cues (such as interaural time/level difference) are suppressed if its leading signal arrived 25-35msec earlier than it and its signal is not 10dB stronger than its leading signal. It is a simple but effective solution.

There are two criteria of Precedence effect relevant with the time and power. It says that a reverberant condition can be handled enough well using just a rule relevant with time and power. For this reason, we made a delta-power filter which has a time parameter γ and a power parameter δ.

$$f_{\gamma,\delta}(n,\theta) = \gamma \cdot f_{\gamma,\delta}(n-1,\theta) + \mu_{\delta}(\Delta p) \cdot \hat{R}(n,\theta)$$
$$\mu_{\delta}(\Delta p) = \frac{1}{(1+\exp(-2(\Delta p-\delta)))} \quad (6)$$

, where Δp is a power increment, and $\hat{R}(n,\theta)$ is a transformed Cross-Angle-Correlation by Gaussian function at the n^{th} frame.

A delta-power filter plays a role of a temporal memory for $\hat{R}(n,\theta)$ at increasing-power frames. If current power increment is larger than power parameter δ, $\hat{R}(n,\theta)$ is recorded on our filter and it fades out with γ-rate as frame goes on. With our delta-power filter, we can extract a feature in the way of (7).

$$\left\{\zeta_{\gamma,\delta}(n) = \sum_{\theta} f_{\gamma,\delta}(n,\theta) \cdot \hat{R}(n,\theta)\right\}_{\gamma,\delta} \quad (7)$$

We constituted a feature vector using (7) with various (γ,δ) combinations. Its dimension is about 10-20 depending on the experimental environment. This feature can indicate how much the spatial cues of current frame conform to the previous spatial cues of increasing-power frames. The spatial cues not conforming will be suppressed similarly as Precedence effect.

The reason we tried to watch the increasing-power frames is that it is likely to come from the direct-path sound because reflected sound might lose its power and be difficult to make a striking power increment.

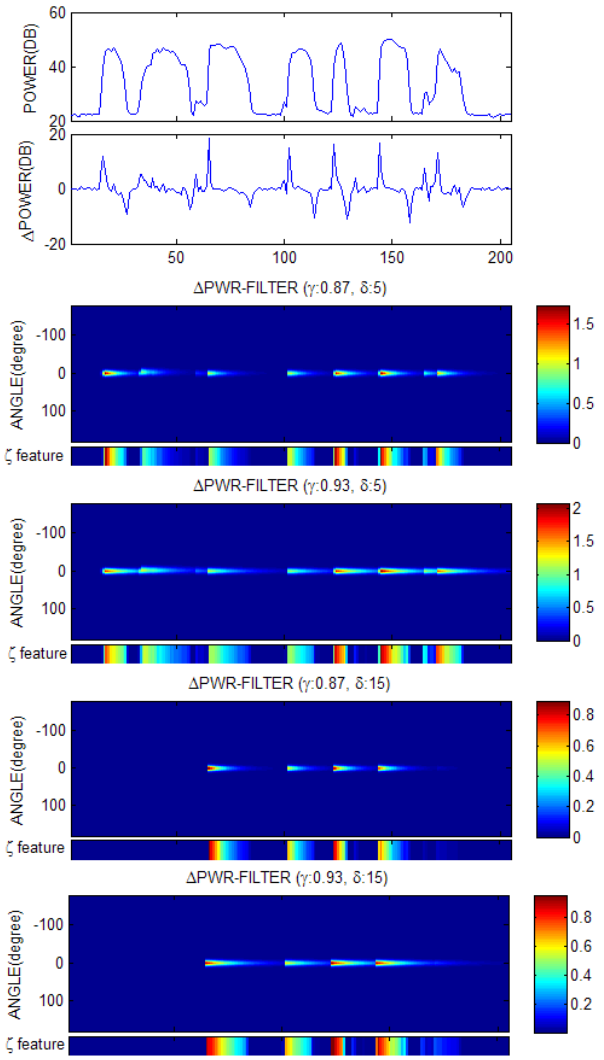


Fig. 5. An example of delta-power filters and extracted features of Fig.3

C. *Verifier and its Training*

We took a neural network classifier as our verifier. Our training space is very simple – accept or reject. Therefore we minimized the structure of our network as one hidden layer of one node. For its training, we could get target values from the detected face position through vision camera. If the estimated source angle from audio conforms to the face position from video, the feature of that frame is trained valid and otherwise, invalid. The training procedure is given as follows.

Verifier Training Procedure

For each audio frame,

1. Gather the information from audio and video
   - A. Localize sound source from audio signal
   - B. Read current face positions from the face detection module
2. Make a feature vector
   - A. Calculate a set of delta-power filters for various time and power parameters
   - B. Make a feature vector from delta-power filters
3. If no face is detected, no training
   Otherwise, do on-line training
   - A. Decide the target value
     If audio conforms video, set valid
     Otherwise, set invalid
   - B. Save the feature vector and target value
   - C. Train the verifier with recent M-frame training data
4. Verify the validity of the audio result of current frame

IV. Simulation and Experiment

A. *Simulation*

To test our proposed method, we simulated three reverberant environments by Roomsim program in MATLAB [11]. The selected rooms and its conditions are listed in Table I and Fig 6 shows the virtual room configuration used in Roomsim.

TABLE I
SIMULATED ROOM CONDITIONS

Room	RT60 (sec)	Absorption Rate of Wall					
		125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz
Quiet-room	0.07	0.9	0.9	0.9	0.9	0.9	0.9
Acoustic-plaster	0.62	0.10	0.20	0.50	0.60	0.70	0.70
Plywood	1.12	0.60	0.30	0.10	0.10	0.10	0.10

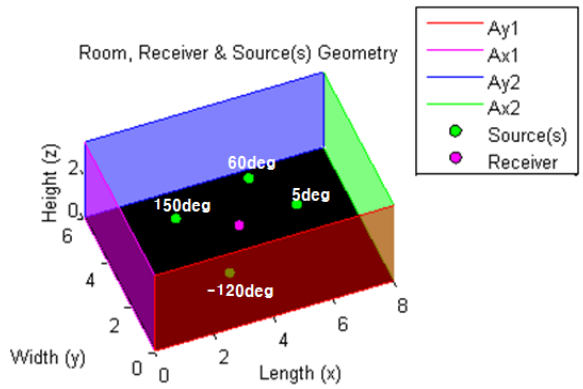


Fig. 6. Configuration of virtual room in Roomsim

Actually, Roomsim generates impulse responses for one-microphone or two-microphone arrays but our microphone-array has 3-microphones. Therefore, we generated an impulse response for each microphone and bound them together as an impulse response for a 3-microphone array.

The simulation scenario is shown in Fig. 7. Our vision system has its coverage of about ±20 degrees in its FOV (Field Of View). At the beginning, a source is detected at 5 degrees by both audio and video sensors. At this time, our verifier is trained. Next, sources at 60, 150, and -120 degrees are sequentially detected by only audio sensor. At this time,

our verifier is tested.

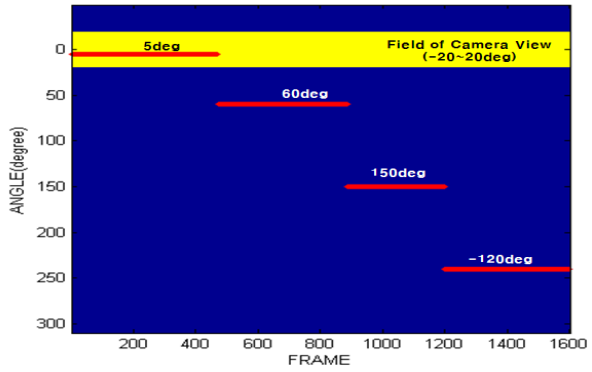
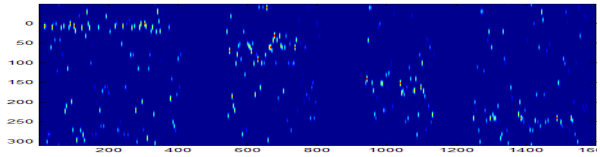
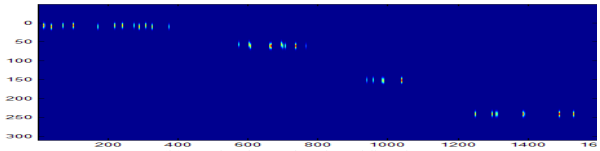


Fig. 7. Simulation Scenario

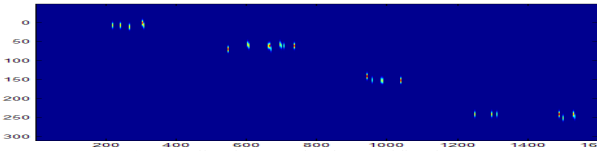
Fig. 8 shows an example of our simulation, that is, the simulation result in the Plywood room environment. Fig. 8-(a) shows how sound source localization in a reverberant condition is confused. Although a large number of results are still distributed around the directions of real sources, the results from fake sources are too many for us to make decisions clearly on where is the sound source. Fig. 8-(b) shows a desired result of verification. Frames with error less than 5 degrees are passed and others are blocked. Fig. 8-(c) shows the result of our verification method. It shows a good performance comparing to the desired result. Only from 0 to 200 frames, it blocks almost frames, but it is because the verifier went through an adaptation time at the beginning.



(a)Localization from audio



(b)Desired result of verification



(c)Localization result after verification

Fig. 8. Simulation Result in Plywood room

All simulation results are listed in Table II. "Hit" means verification accords with the desired and "Miss" means verification discords with the desired at a frame. In detail, there are two kinds of Miss, the one is when an invalid frame is passed and the other is when a valid frame is blocked by our verifier. According to the simulation results, our method shows a good performance. Its hit rate is higher than 85% and up to 92.44%. An interesting point is that its performance doesn't depend on the acoustic conditions. This upholds that our approach is reasonable and successful.

TABLE II
SIMULATION & EXPERIMENT RESULTS

Room	Hit [frames]	Miss [frames]	
		Pass invalid	Block valid
Quiet-room	1313 (88.48%)	89 (6.00%)	82 (5.53%)
Acoustic-plaster	1385 (86.51%)	88 (5.50%)	128 (8.00%)
Plywood	1480 (92.44%)	28 (1.75%)	93 (5.81%)
Real-Hall	2197 (87.77%)	195 (7.79%)	111 (4.43%)

B. *Experiments*

Our algorithm was implemented on a robot system which consists of a robot head we made and a Peoplebot platform of MobileRobots Inc. Its head has 2 vision cameras (but we used just one camera) and 3 microphones positioned on the vertices of a triangle within a circle of 7.5cm radius.

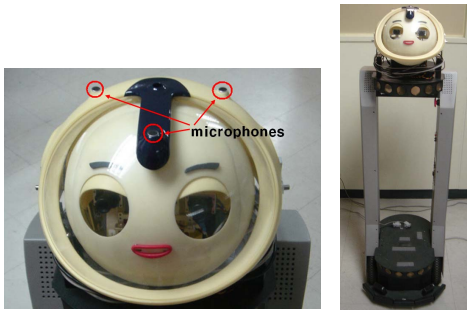


Fig. 9. Robot platform

In addition to simulations, we performed a real experiment. The scenario of our experiment is similar to the simulation except the difference in the source angles. At first, a person speaks at 0 degree. At this time, the vision camera can detect him and our verifier is trained. Next, he moves to 90, 180, and -90 degrees sequentially and says words. While he moves, he is out of the field of camera view and the verifier refines the result from the audio sensor. This experiment was done in a large hall of 19.5x9.1m^2 where RT60 was measured about 0.6sec.

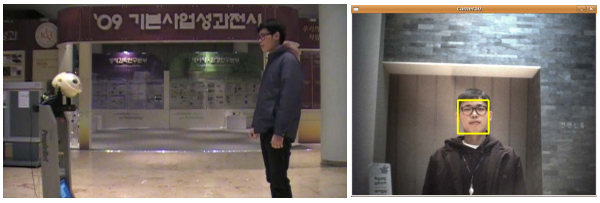
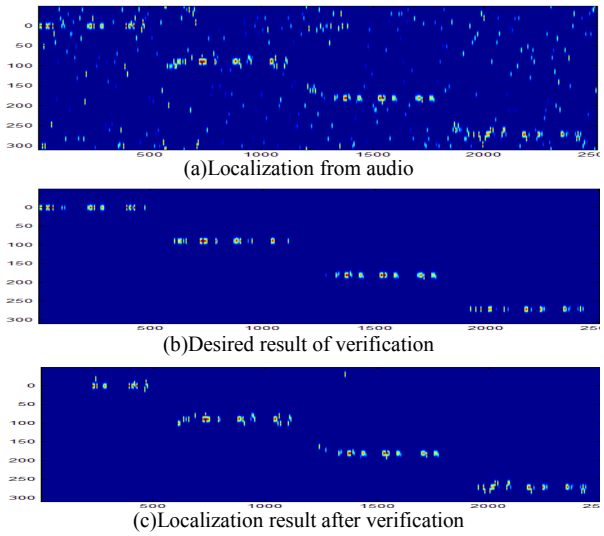


Fig. 10. Real Experiment in a Large Hall

Its result is given by Fig. 11 and Table II. Fig. 11-(a) shows how rough the acoustic condition is in the hall and Fig. 11-(c) shows that the proposed method can effectively handle the fake sources in a reverberant environment. According to the Table II, its hit rate in a real hall is 87.77% as good as

those of simulation results.



(a)Localization from audio

(b)Desired result of verification

(c)Localization result after verification

Fig. 11. Real Experiment Result in a Hall

V. CONCLUSION

By this work, we tried to develop a multi-modal system in which audio sensors and video sensors cooperate with each other. Especially, we want audio sensors to perform better using the information from video sensors. We designed a verifying algorithm which can adapt audio sensors to the reverberant environments by a visual learning procedure. We showed its effectiveness through simple simulations and a real experiment.

For a future work, we are going to merge the proposed method into an audio-video speaker tracking algorithm and implement it on our robot platform.

ACKNOWLEDGMENT

We really appreciate prof. Daijin Kim's IMLab members providing us their vision program. Also, we thank our lab members, Dohyeong Hwang and Dongjoo Kim. They spared no efforts for our implementation and experiments.

REFERENCES

[1] G. Lathoud, J.-M. Odobez, D. Gatica-Perez, "AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking," *Lecture Notes in Computer Science*, issu. 3361, pp.182-195, 2005.

[2] Carlos Busso et al., "Smart Room: Participant and Speaker Localization and Identification," in *Proc. IEEE ICASSP*, March, 2005, vol. 2, pp. ii/1117-ii/1120.

[3] Yoonseob Lim, Jongsuk Choi, "Speaker selection and tracking in a cluttered environment with audio and visual information," *IEEE Trans. Consumer Electronics*, vol. 55(3), pp.1581-1589, 2009.

[4] K. Nakadai, K. Hidai, H. G. Okuno, H. Kitano, "Real-Time Multiple Speaker Tracking by Multi-Modal Integration for Mobile Robots," in *Proc. Eurospeech 2001, Scandinavia*, pp.1193-1196.

[5] Byoung-gi Lee, Jongsuk Choi, "Analytic Sound Source Localization with Triangular Microphone Array," in *Proc. URAI 2009*, pp.29-32.

[6] P. Svaizer, M. Matassoni, M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE ICASSP*, April, 1997, vol. 1, pp.231-234.

[7] Byoung-gi Lee, JongSuk Choi, "Multi-source Sound Localization using the Competitive K-means Clustering," in Proc. IEEE Intl. Conf. Emerging Technologies and Factory Automation, September, 2010. (to be public)

[8] Intelligent Media Lab., Postech, hompage: http://imlab.postech.ac.kr/

[9] Bongjin Jun, Daijin Kim, "Robust Real-Time Face Detection Using Face Certainty Map," *Lecture Notes in Computer Science*, vol. 4642, pp.29-38, 2007.

[10] H. Haas, "The influence of a single echo on the audibility of speech," *Journal of the Audio Engineering Society*, vol.20, pp. 146-159, 1972.

[11] D. R. Campbell, *Roomsim User Guide (V3.4)*, 2007.

[12] Vermaak, J. and Blake, A., "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE ICASSP* 2001.

[13] Vermaak, J. and Gangnet, M. and Blake, A. and Perez, P., "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. IEEE Intl. Conf. on Computer Vision*, 2001.