

Sound Source Localization Based on Audio-visual Information for Intelligent Service Robots

Beom-Cheol Park
and Kyu-Dae Ban
Computer software and engineering
University of Science and Technology
Daejeon, Korea
Email: {parkbc,kdban}@etri.re.kr

Keun-Chang Kwak
and Ho-Sub Yoon
Human Robot Interaction Team
Intelligent Robot Research division
Electronics and Telecommunications-
Research Institute
Daejeon, Korea
Email: {kwak,hsyoon}@etri.re.kr

Abstract— In this paper, we present an Sound Source Localization (SSL) based on audio-visual information with robot auditory system for a network-based intelligent service robot. The main goal of this paper is to combine audiovisual-based Human-Robot Interaction (HRI) components that can naturally interact between human and robot for SSL. The proposed approach includes two main steps. The first step performs speech recognition and sound localization to know whether the user calls the robot or not as well as the direction of the caller respectively, when someone calls robot's name. Here sound localization is based on GCC(Generalized Cross-Correlation)-PHAT(Phase Transform) by frequency characteristics. In the second step, a robot moves forward to the caller based on face detection. The robot platform used in this work is wever-R2, which is a network-based intelligent service robot developed at Intelligent Robot Research Division in ETRI. The effectiveness of the proposed approach is compared with audio-based SSL itself.

Index Terms — sound source localization, intelligent robot, GCC-PHAT, face detection, human-robot interaction

I. INTRODUCTION

Today, the interest in Intelligent Service Robots has been brought to public attention. There are a lot of needed techniques for developing intelligence robot, especially Human-Robot Interaction (HRI) technique that naturally interact between human and robot using images and voice information from cameras microphones has been researched in places [3]. Among these techniques, sound source localization finds the direction of sound source. From this technique, the robot can move and help for giving aid to a person by recognizing and judging a situation in public places and home [1]. Currently, many sound source localization techniques have been researched generally through analysis in time domain or frequency domain. The most representative techniques frequently used in conjunction with sound source localization are intensity difference between microphones [2], Time Delay of Arrival (TDOA) method [6] and Generalized

Cross Correlation-Phase Transform (GCC-PHAT) method[7], Beam-forming method[4], and so on. The TDOA are widely used due to accuracy and simple computation [5]. On the other hand, the GCC-PHAT has a good performance under noise or reverberation environments.

This paper is concerned with audio-visual sound localization that combines GCC-PHAT and face detection based on adaboosting. Sound source location is performed by estimating localization angle from several candidate angles without section selection method. For performance analysis, database used in this paper is constructed through the wever-R2 in test bed environments. This database is built by sound source distance(1m~5m), angle(2 channel : $0^\circ \sim 180^\circ$, 3 channel : $0^\circ \sim 360^\circ$, 45° interval), number of microphone channel(2, 3). The localization performance is compared by Localization Success Rate (LSR) and Average Localization Error (ALE) from FOV(Field of View) range.

This paper is organized in the following manner. Section 2 describes GCC-PHAT method and approach to obtain reliable localization angle. Section 3 presents multimodal sound localization including face detection under network-based environments. Section 4 covers the experimental results concerning sound localization. Finally concluding comments are covered in Section 5.

II. SOUND LOCALIZATION METHOD BASED ON GCC-PHAT

A. GCC-PHAT method

GCC(Generalized Cross Correlation) is a correlation method in frequency domain. Sound localization based on GCC-PHAT has a lot of merits in noise environments and reverberation environments. The GCC-PHAT sound localization method can be described as follows

When the signals $x_1(n)$ and $x_2(n)$ are obtained by each of two microphones, the generalized cross-correlation

between $x_1(n)$ and $x_2(n)$ can be obtained by the following equation.

$$R_{x_1x_2}(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} W(\omega) X_1(\omega) X_2^*(\omega) e^{j\omega n} d\omega \quad (1)$$

where $W(\omega)$ is a frequency weighting function also called as PHAT(Phase Transform) [8] and this weighting function is the reciprocal of $X_1(\omega) X_2^*(\omega)$. PHAT is a weighting function subordinated to frequency that determines the relative importance of each frequency, equation can be represented as follows

$$W(\omega) = \frac{1}{|X_1(\omega) X_2^*(\omega)|} \quad (2)$$

The delay time between $x_1(n)$ and $x_2(n)$ can be obtained by the following expression

$$\tau = \arg \max R_{x_1x_2}(n) \quad (3)$$

B. Estimation of reliable localization angle

In what follows, we propose the method to estimate reliable localization angle from several candidate angles obtained by GCC-PHAT method. In this point, when we suppose that wavelength of a sound source is plane wave, three angles are obtained based on each time delay obtained by GCC-PHAT as the following equations

$$\alpha = \cos^{-1} \left(\frac{v\tau_{12}}{d} \right) \quad (4)$$

$$\beta = \cos^{-1} \left(\frac{v\tau_{23}}{d} \right) \quad (5)$$

$$\gamma = \cos^{-1} \left(\frac{v\tau_{13}}{d} \right) \quad (6)$$

where d is a distance between each microphone and v is the velocity of sound, τ_{12} is time delay between channel 1 and channel 2, τ_{23} is time delay between channel 2 and channel 3, τ_{13} is time delay between channel 1 and channel 3. From these angles, we obtain six candidate angles as follows

$$\begin{aligned} \Phi_1 &= \alpha - 30, \Phi_2 = -\alpha - 30, \Phi_3 = \beta + 90 \\ \Phi_4 &= -\beta + 90, \Phi_5 = \gamma + 30, \Phi_6 = -\gamma + 30 \end{aligned} \quad (7)$$

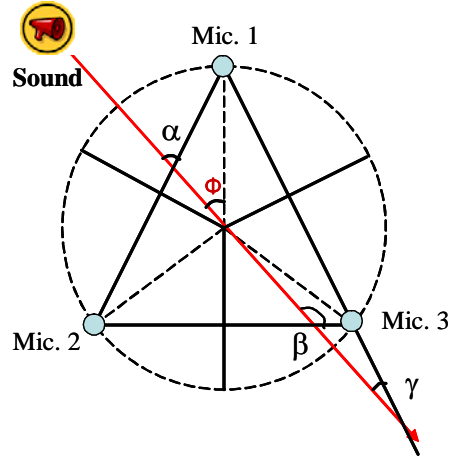


Fig. 1. six angles calculation method (3 channel case)

Fig.1 shows the method to obtain the reliable angle from six angles in the case with three microphones. Here because it is difficult to obtain ideal time delay, we have to select the two closest Φ . And then we can estimate the reliable angle by averaging these two angles.

III. MULTIMODAL SOUND SOURCE LOCALIZATION

A. Face detection

After performing sound localization, we use face detection/recognition to compensate localization error and realize humanlike visual system. Here face detection is comprised of three steps including preprocessing, detection, and postprocessing. In the first stage, we revise the modified census transform to compensate the sensitivity to the change of pixel values. The second stage performs Adaboost that constructs the weak classifier which classifies the face and nonface patterns and the strong classifier which is the linear combination of weak classifiers. The last stage performs face certainty map based on facial information such as facial size, location, rotation, and confidence value to reduce False Acceptance Rate (FAR) with constant detection performance.

B. Face recognition

On the other hand, we use a MPCA (Multiple PCA) and SVM (Support Vector Machine) for face recognition. The most representative recognition technique frequently used in conjunction with face recognition is PCA. The PCA approach, also known as eigenface method, is a popular unsupervised statistical technique that supports finding useful image representations. It also exhibits optimality when it comes to dimensionality reduction. The use of the MPCA in this setting is motivated by its insensitivity to variation in comparison to PCA itself. This method consists of preprocessing, feature extraction, and identification. In the preprocessing, we select the face region such as eigenface, eigenUpper, and eigenTzone for multiple PCA. Furthermore, a geometric and photometric normalization is used to adjust the location of facial features and improve the quality of the face image,

respectively. In the feature extraction, the weight and edge distribution vectors are obtained. Finally we perform face recognition using SVM as a nonlinear classifier from feature vectors obtained by MPCA and edge distribution.

IV. DATABASES FOR THE USE OF SOUND SOURCE LOCALIZATION

A database used in this paper has been constructed in test bed that is similar with home environment to evaluate the sound source localization algorithm. Fig. 2 shows the network based intelligent robot “wever-R2” developed by intelligent robot research division in ETRI. Fig. 3 shows microphone arrangement with 120 degrees interval on wever-R2. The three microphones are equipped with multi-channel sound source board MIM (Multimodal Interface Module) that is developed by ETRI. We used “wever” speech as sound source localization database. The two person speak three times at 45° interval from 0° to 360° . The first data set consists of 120 speeches at each meter from 1meter to 5 meter (CH3_M1, CH3_M2) when the number of microphone is three. On the other hand, in the case of two channels, we used 105 set at each meter from 1 meter to 5 meter with 30° interval from 0° to 180° (CH2_M1, CH2_M2) as the second data set. The recording was done in an office environment. The audio is stored as a mono, 16bit, 16kHz. Fig. 4 shows example of voice samples obtained in the case of 0° , 1m.

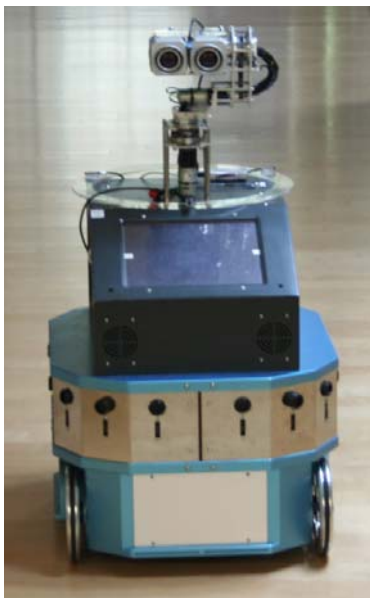


Fig. 2. Intelligent robot “wever-R2” of ETRI

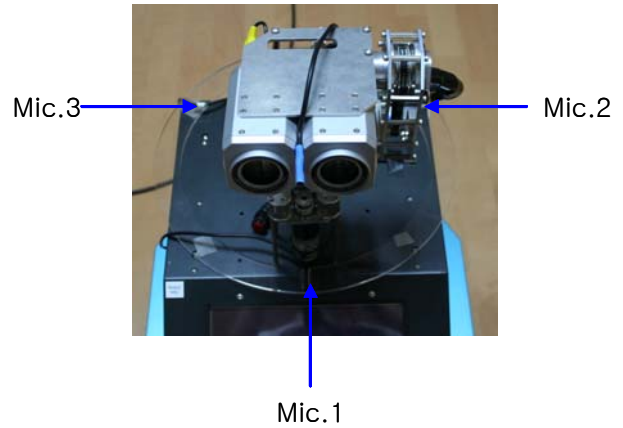


Fig.3. Microphone arrangement (3 microphones)

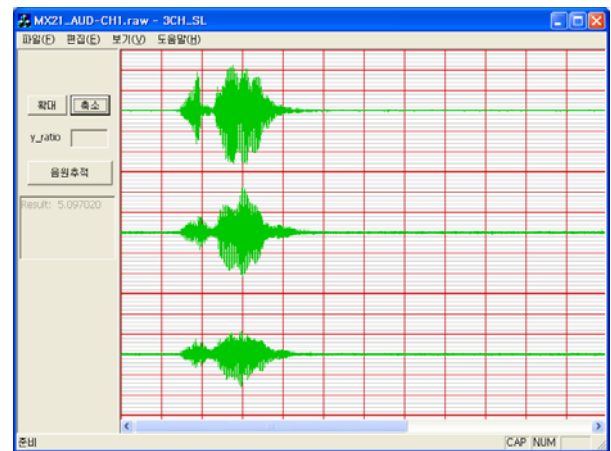


Fig. 4. Microphone arrangement (3 channel case)

V. EXPERIMENTS AND RESULT

In this section, the GCC-PHAT sound source localization method of frequency domain is compared with TDOA sound source localization method of time domain with using databases for the use of sound source localization. TDOA is the variation of TDOA and partitioned by angle measurement method. Type1 is the experiential section selection measurement method and Type2 is the proposed angle measurement method. At this point, the variation of TDOA uses new voice signal that is made by subtracting previous sample from present sample. Localization success rate (LSR) and Average localization error (ALE) are considered as performance indicator. Localization success rate considers FOV(± 10 , ± 15) because sound source localization is used with face detection when robot approach caller. FOV of Cameras of WEVER-R2 is ± 24 .

3 channel experiment results that compare performance about localization success rate and average localization error are showed TABLE 1 and TABLE 2. As tables showed, both of CH3_M1 and CH3_M2 DB results show that GCC-PHAT (type2) based sound source localization method is better than TDOA based sound source localization method. Also, the more far distance between microphone and caller, the worse performance is showed in TDOA, but GCC-PHAT shows that

TABLE 1
Performance comparison for CH3_M1 DB
LSR : Localization Success Rate(%), ALE: Average
Localization Error (degree)

	FOV±10		FOV±15	
	LSR	ALE	LSR	ALE
TDOA (type 1)	42.5	3.3	52.5	5.0
TDOA (type 2)	66.6	4.2	75	5.15
GCC-PHAT (type 2)	88.3	2.9	98.3	3.8

TABLE 2
Performance comparison for CH3_M2 DB

	FOV±10		FOV±15	
	LSR	ALE	LSR	ALE
TDOA 변형 (type 1)	36.7	3.8	45	5.3
TDOA 변형 (type 2)	62.5	3.9	69.2	4.4
GCC-PHAT (type 2)	89.2	4.2	94.2	4.6

TABLE 3
Performance comparison for CH2_M1 DB

	FOV±10		FOV±15	
	LSR	ALE	LSR	ALE
TDOA 변형 (type 2)	30.5	3.4	53.3	7.4
GCC-PHAT (type 2)	81.9	8.4	86.7	8.0

there is no performance dissimilarity until 5m.

In 2 channel case, only proposed angle measurement methods are compared. Table 3 and 4 is the comparison of performance about localization success rate and average localization error. Tables show that GCC-PHAT based sound source localization represents better performance more 35% than the variation of TDOA method.

TABLE 4
Performance comparison for CH2_M2 DB

	FOV±10		FOV±15	
	LSR	ALE	LSR	ALE
TDOA 변형 (type 2)	25.7	3.5	46.7	7.5
GCC-PHAT (type 2)	81.0	8.2	84.8	8.0

VI. CONCLUSION

In this paper, performance of TDOA based and GCC-PHAT based sound source localization methods are compared with applying to the intelligent robot WEVER-R2 in test bed where is similar with real home environment. And multimodal sound source localization is experimented with face detection and face recognition. In real home environment, delay times that are obtained by transforming to frequency domain have more accuracy information than delay times that is obtained in time domain. Also, we can confirm availability of trusted localization method from several presumed angles. The sound localization method with multimodal that is introduced in this paper can improve of localization success rate.

ACKNOWLEDGMENT

This work was supported in part by IT R&D program of MIC & IITA [2005-S-033-03, Embedded Component Technology and Standardization for URC

REFERENCES

- [1] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model based sound localization system and its application to robot navigation," *Robotics and Autonomous Systems*, pp. 199-209, 1999.
- [2] Jiyeoun Lee, and Minsoo Hahn, "Sound Localization Technique for Intelligent Service Robot "WEVER", Proceedings of the KSPS conference, pp. 117-120, Nov. 2005.
- [3] O. Deniz, M. Castrillon, J. Lorenzo, C. Guerra, D.Hernandez, M.Hernandez, "CASIMIRO: A RobotHead for Human-Computer Interaction", Proceedings the 2002 IEEE. Int. WorkShop in Robot and Human Interactive Communication, 2002.
- [4] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, New York, 2001.
- [5] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput., Speech Lng.*, vol. 11, no. 2, pp. 91-126, 1997.
- [6] Jisung Choi, Jiyeoun Lee, Sangbae Jeong, Keunchang Kwak, Suyoung Chi, Minsoo Hahn "Multimodal Sound Source Localization for Intelligent Service Robot," International Conference on Ubiquitous Robots and Ambient Intelligence, 2006.
- [7] C.H Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustic. Speech Signal Processing.*, Vol. 24, No. 4, pp.320-327, 1976.
- [8] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform (SCOT)," *Proceedings of the IEEE*, vol. 61, pp.1497-1498, 1973.