AN INTEGRATED AUDIO-VIDEO TRACKING SYSTEM WITH A PTZ VIDEO CAMERA AND A 3-D MICROPHONE ARRAY

Kuo-Lun Huang, Yi-Ta Cheng, and Pao-Chi Chang[*]
Department of Communication Engineering
National Central University, Jhongli, Taoyuan 32001, Taiwan
E-mail: [*]pcchang @ vaplab.ce.ncu.edu.tw

ABSTRACT

Localization of an audio source or a video object is critical in many audio-video applications such as video surveillance and video conferencing. A microphone array with three-dimensional (3-D) space layout is able to locate an audio source in a 3-D space, which improves the accuracy of localization in different environments. A Pan-Tilt-Zoom (PTZ) video camera is able to track and zoom a video object with proper control. Conventionally, a PTZ camera can be controlled manually or by video segmentation. Audio source localization can nicely complement vision to help localize a person or a video object in an environment.

In this work, we present an integrated audio-video tracking system that includes a 3-D microphone array as well as a PTZ video camera. The audio localization that is based on time delay of arrival estimation is used to control the PTZ camera.

The experiment results show that a 3-D microphone array with 4 microphones is able to localize human voice with the error less than 7°, which provides sufficient accuracy to control a PTZ camera in a tracking system.

Index Terms — *3-D microphone array, PTZ, audio localization*

1. INTRODUCTION

Pan-Tilt-Zoom camera (PTZ) video camera that is able to track and zoom in/out a video object is commonly used in many surveillance systems. By using RS-232 to connect the computational server and PTZ, we can control PTZ lens to capture the video and track the objects. Although PTZ is powerful, it still cannot track objects which are not in the PTZ viewing range. In this work, we integrate a PTZ camera with a 3-D microphone array to localize objects. Localization by microphone arrays has been studied extensively. For examples, Valin *et. al* [2] use 8 microphones to build a rectangular prism and integrate a mobile robot to track objects. Tamai and Kagami *et. al* [5][7] layout a 128-channel huge microphone array in a room to localize audio source.

The most common way for audio source localization from 2-D to 3-D is to build an array by using more than 8 microphones. Because of their high computational complexity and device cost, we reduce number of microphones and reshape the layout to save the complexity and cost.

In this work, we use a 4-microphone array to localize an audio source in 3-D space and use the localization results to control the PTZ lens to track the source. In our study, the layout of the 4 microphones is shown in Fig.1. The audio signal is captured by the microphone array and then sent to the server. By finding the time difference from two microphones, the direction estimation method can determine the direction of voice in the 2-D space. Combining the direction of another 2-D space determined by two other microphones, we can build a 3-D localization system and get the source position. The next step is to feed the horizontal and vertical degrees into PTZ control instructions. Then the PTZ will focus on the object and capture the video frames.

This paper mainly studies a 3-D audio source localization by utilizing the Time Difference Of Arrival (TDOA) from the audio source to four microphones. There exist many TDOA algorithms in time or frequency domains [3]. Although frequency-domain methods may yield better results, it is complex and time-consuming. In order to shorten the processing time to support real-time applications, we choose the time-domain algorithms.

The time-domain methods mainly include Average Magnitude Difference Function (AMDF), ratioAMDF, least squares, and cross-correlation methods. According to [1], AMDF is the most accurate method in computing TDOA. However, the cross-correlation method is more convenient in implementation and faster than AMDF. Although cross-correlation method has the restriction that the distance between the object and the microphones needs to be at least 1 meter for accurate results, it is an appropriate method to implement our system.

The rest of this paper is organized as follows. Section

2 illustrates how to compute TDOA by using cross-correlation. Section 3 explains how to compute the audio source direction based on TDOA. Section 4 shows the experimental results. Section 5 draws the conclusion.

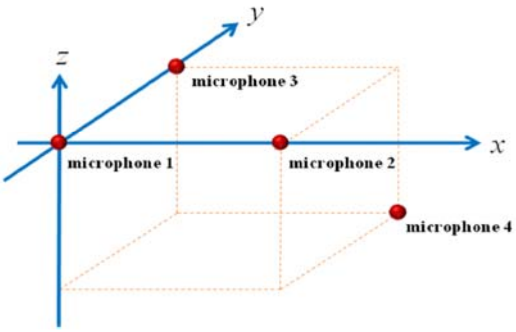


Fig. 1. 4-microphone array structure

2. TDOA ESTIMATION

The procedure of TDOA estimation is described in this section. The received audio signals are first pre-processed to eliminate the noise by dividing the audio signals of four microphones into several frames and computing the power of each frame. The signals in a frame which power is below a certain threshold are set to zero. These pre-processed signals are fed into TDOA algorithm to estimate the time difference.

In our work, cross-correlation method is used to implement TDOA algorithm. The cross-correlation method, using two microphones as an example for convenience, is expressed by

$$R_{12\,i}(\lambda) = \sum_{n=0}^{N-1} x_{1i}[n]x_{2i}[n+\lambda], i = 1,2,3..., M \tag{1}$$

$$\lambda'_i = \arg\max_{\lambda}\{R_{12i}(\lambda)\}, i = 1,2,3..., M \tag{2}$$

$$\hat{\lambda} = \text{mode}\{\boldsymbol{\lambda}'\} \tag{3}$$

where $x_{1i}[n]$, $x_{2i}[n]$ are signals of the frame number i received by microphone *1* and microphone *2*, respectively, $R_{12i}(\lambda)$ is cross-correlation function of these two audio channel signal, M and N are the number of frame and number of sample in each frame, respectively. λ'_i denotes the sample difference of the frame *i*, $\hat{\lambda}$ is most frequent number in the set $\boldsymbol{\lambda}'$. Namely we select $\hat{\lambda}$ which results in maximum correlation value to represent the sample differnce of two microphones. Finally, we can get the time difference of arrival ΔT_{12} for two microphones with sampling rate f_s as

$$\Delta T_{12} = \frac{\hat{\lambda}}{f_s} \, . \tag{4}$$

3. AUDIO SOURCE DIRECTION ESTIMATION

As the analysis in [2], we also assume that the distance from the microphone array to the audio source is almost infinite. From this assumption, the paths from sound source to microphones can be regarded as parallel.

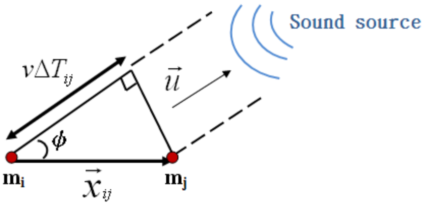


Fig. 2. Computing source direction from TDOA

Fig. 2 illustrates the case of a 2-microphone array with an audio source in the far-field. Using the cosine law, we can state

$$\cos\phi = \frac{\vec{u}\cdot\vec{x}_{ij}}{\|\vec{u}\|\|\vec{x}_{ij}\|} = \frac{\vec{u}\cdot\vec{x}_{ij}}{\|\vec{x}_{ij}\|} \tag{6}$$

where $\vec{x}_{ij}$ is the vector that starts from microphone i to microphone j and $\vec{u}$ is a unit vector pointing in the direction of the source. From fig. 2, we can also obtain

$$\vec{u}\cdot\vec{x}_{ij} = v\Delta T_{ij} \tag{7}$$

where v is the speed of sound, ΔT_{ij} is the TDOA of two corresponding microphone i and j. Equation (7) can also be written as

$$a(x_j - x_i) + b(y_j - y_i) + c(z_j - z_i) = v\Delta T_{ij} \tag{8}$$

where $\vec{u} = (a,b,c)$ and $\vec{x}_{ij} = (x_j - x_i, y_j - y_i, z_j - z_i)$, the position of microphone i being (x_i, y_i, z_i).

Using an array of K microphones, there are *K(K-1)/2* different cross-correlations, but the only *K-1* are independent. We choose only ΔT_{1j} values for calculation. Then we obtain a system of *K-1* equations:

$$\begin{bmatrix} (x_2 - x_1) & (y_2 - y_1) & (z_2 - z_1) \\ (x_3 - x_1) & (y_3 - y_1) & (z_3 - z_1) \\ . & . & . \\ . & . & . \\ (x_K - x_1) & (y_K - y_1) & (z_K - z_1) \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} v\Delta T_{12} \\ v\Delta T_{13} \\ . \\ . \\ v\Delta T_{1K} \end{bmatrix} \tag{9}$$

Consequently, the direction of voice source can be determined from the position vector $\vec{u} = (a,b,c)$ in (9).

In this method, we need at least four microphones to estimate the direction of the sound in a three-dimensional space. Hence, we utilize four microphones, *i.e., K*=4, to implement our tracking system.

4. EXPERIMENTAL RESULTS

In this section, we demonstrate the experimental results with various positions of human voice sources. Fig. 3 shows the picture of the entire tracking system, in which corresponding equipments are specified in the Table 1. The audio signals which are received from the microphones are acquired by a Data AcQuisition (DAQ) board. The server computes the TDOA of digitized audio signals and estimates the direction of audio source. Following the PTZ technical specification [9], we feed the estimated horizontal and vertical degrees in the specific command format into PTZ to control the lens rotation.

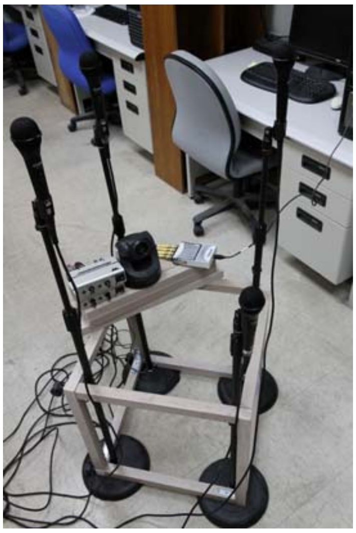


Fig. 3. The 4-microphone tracking system

Table. 1. Equipments for implement

Software	Operating system	Microsoft Windows XP
	Development Code	Microsoft Visual C++
Hardware	Computational Sever	Intel Core i7 CPU 920@ 2.67Hz, 2.99GB RAM
	Audio interface	Dynamic Signal Acquisition NI DAQ USB-9233
	Microphones	AKG D 660 S
	Microphone pre-amplifier	M-Audio Audio Buddy
	PTZ camera	Sony EVI-D70/70P
	Video interface	(TV box)Aver media AVerTV USB820

We test the tracking system in various voice directions. Fig. 4 illustrates the relative positions of the PTZ and the sound source. The operating angle of this system ranges from 0 to 270 degrees on horizontal and 0 to 55 degrees on vertical, which can usually cover a person head direction.

Table 2 shows the actual and estimated degrees in comparing the accuracy in each direction. From Table 2, the results from 0 to 45 degrees on horizontal are more accurate than other degrees. Although other angles have error with 6 to 8 degrees, the PTZ camera can still focus on the target. The captured images of PTZ are shown in Fig. 5.

In general, the position estimation will be affected by the echo and noise if the experiments are not performed in an empty space. However, in our experiments, PTZ tracking system by using microphone array is still reliable in a non-empty space such as in a classroom with tables and chairs.

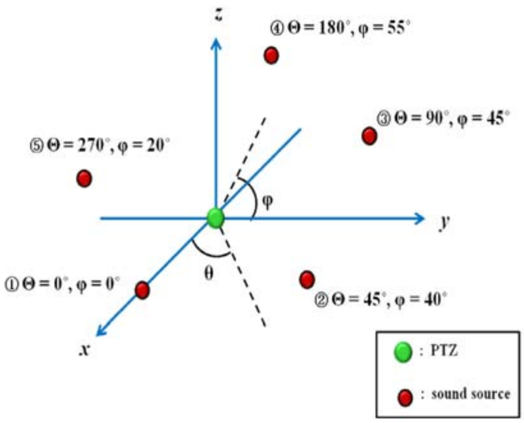
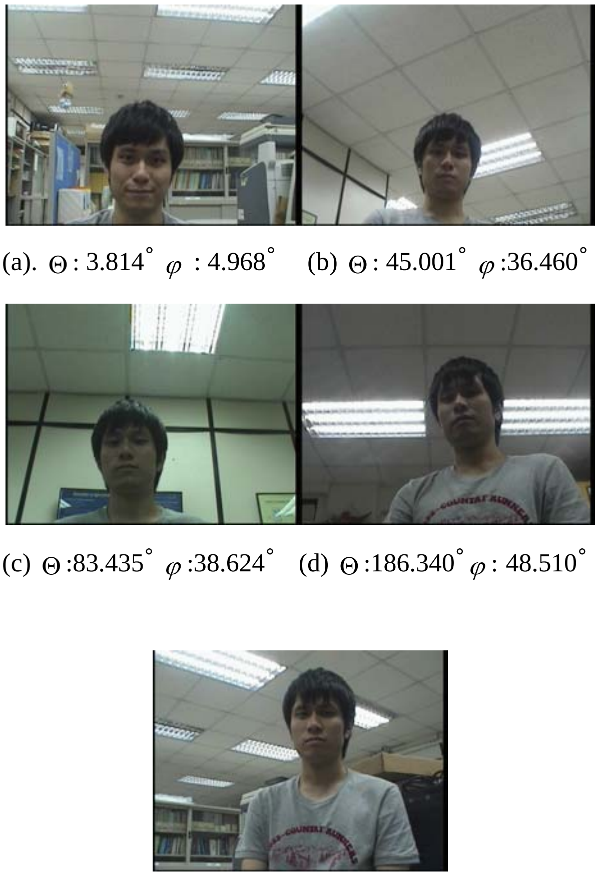


Fig. 4. Sound source composition

Table. 2. Experimental result and data

Actual value (degree)		Experimental value (degree)	
horizontal(Θ)	vertical(φ)	horizontal	vertical
0	0	3.814	4.968
45	40	45.001	36.46
90	45	83.435	38.624
180	55	186.34	48.51
270	20	276.34	11.132
Actual value (degree)		Error (degree)	
horizontal	vertical	horizontal	vertical
0	0	3.814	4.968
45	40	0.001	-3.54
90	45	-6.565	-6.376
180	55	6.34	-6.49
270	20	6.34	-8.868



(a). Θ : 3.814° φ : 4.968° (b) Θ : 45.001° φ :36.460°

(c) Θ :83.435° φ :38.624° (d) Θ :186.340° φ : 48.510°

(e) Θ :276.340° φ :11.132°

Fig. 5 Captured experimental images.

5. CONCLUSION AND FUTURE WORK

This paper has proposed a localization system that traces the voice source automatically. Instead of building a rectangular prism structure by using 8 microphones, we use only four microphones to localize the object in 3D space to reduce the cost. In all test results, estimated errors are not more than 8.8°, the localization program can help the PTZ put the target in the middle of the frame smoothly.

The accuracy of the position estimation is seriously affected by the environment seriously. In some cases, the echo can make the estimation program less accuracy. We suggest following methods that can make the PTZ localization system more accurately. 1) Use a better echo cancellation method, 2) Track objects according to the voice feature database, and 3) Combine it with the image recognition system to improve the accuracy of recognition.

6. REFERENCES

[1] S. L. Tu, and Master J. S. Roger Jang, *Two-dimensional Source Localization : Implementation and Discussions of Time Domain Methodologies*, Department of Computer Science, National Tsing Hua University (NTHU), Hsinchu, Taiwan, 2004.

[2] J. M. Valin, F. Michaud, J. Rouat, and D. L´etourneau, “Robust Sound Source Localization Using a Microphone Array on a Mobile Robot,” in *Proceedings IEEE/RSJ International Conference*, Vol 2, pp. 1228-1233, Oct. 2003.

[3] J. Benesty, J. Chen, Y. Huang, *Microphone Array Signal Processing,* Springer 2008.

[4] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach, Third Edition,* McGraw-Hill, 2006.

[5] S. Kagami, H. Mizoguchi, Y. Tamai, and T. Kanade, “Micophone Array for 2D Sound Localization and Capture,” in *Proceedings of the 2004 IEEE International Conference on Robotics & Automation,* New Orleans, LA , Apr. 2004

[6] C. H. Knapp, G. C. Carter, “The Generalized Correlation Method for Estimation of Time Delay”, *IEEE Trans. On Acoustics, Speech and Signal Processing*, Vol. ASSP-24, No.4, Aug. 1976.

[7] Y. Tamai, S. Kagami, H. Mizoguchi, Y. Amemiya, K. Nagashima, Takano, “Real-Time 2 Dimensional Sound Source Localization by 128-Channel Huge Microphone Array”, in *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, Kurashiki, Okayama Japan, Sep. 20-22, 2004.

[8] K. Obata, K. Noguchi, Y. Tadokoro, “A New Source Location Algorithm Based on Formant Frequency for Sound Image Localization”, in *Proceedings of the 2003 International Conference on Multimedia and Expo, Vol. 2*, 2003.

[9] *SONY color video camera EVI D-70 series technical manual.*