Sound Source Localization with Non-Calibrated Microphones

Tomoyuki Kobayashi, Yoshinari Kameda, and Yuichi Ohta

Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573, Japan
kobayashi@image.esys.tsukuba.ac.jp
{kameda,ohta}@iit.tsukuba.ac.jp
http://www.image.iit.tsukuba.ac.jp

Abstract. We propose a new method for localizing a sound source in a known space with non-calibrated microphones. Our method does not need the accurate positions of the microphones that are required by traditional sound source localization. Our method can make use of wide variety of microphone layout in a large space because it does not need calibration step on installing microphones. After a number of sampling points have been stored in a database, our system can estimate the nearest sampling point of a sound by utilizing the set of time delays of microphone pairs. We conducted a simulation experiment to determine the best microphone layout in order to maximize the accuracy of the localization. We also conducted a preliminary experiment in real environment and obtained promising results.

Key words: non-calibrated microphones, sound source localization, time-delay

1 Introduction

Sound source localization can play an important role in surveillance and monitoring purposes, especially in intelligent support of safe and secured life in daily situation. Once the location of a sound source is estimated, it can be a very powerful clue for many applications such as event classification, intruder detection at night, classification of the behaviors, and so on.

Traditional sound source localization methods need precise geometry of the microphones of a microphone array. Microphones are linearly, squarely, or roundly set in popular array layout and their position intervals are given in advance. Or, their positions needs to be precisely measured beforehand when they are sparsely scattered in a room. Some of the advanced researches have reported that sound source localization worked very well with these calibrated microphones [1][2][3][4][5][6][7]. Their performance, however, mainly relies on the accuracy of the microphone positions, which are not always easy to obtain in some situations.

In the literature of intelligent life support[8] and sensing[9][10] of daily life, we can exploit pre-recorded sound samples as a clue to localize a newly observed

sound source because microphones are usually fixed for a long term in such a space and they can record a large number of samples during the term. In addition, for some monitoring purposes, they can be satisfied by finding the geometrically nearest sound sample for the newly observed sound in the space.

We propose a new method for localizing a sound source in a known space with non-calibrated microphones. It does not need to know the positions of the microphones.

We assume that the microphones are permanently fixed in the space. Our method can work with wide variety of microphone layout in a large space because it does not need calibration step on installing the microphones. After a number of sampling points have been stored in the database, our system can estimate the nearest sampling point for a new sound by utilizing the set of time delays of microphone pairs.

The rest of the paper is organized as follows. Section 2 briefly describes the overview of our sound localization system, and section 3 explains the similarity scale that estimates the "distance" between two sound locations. Then, we examine the layout of microphones so that localization error is minimized in section 4 and a preliminary experiment in real environment is shown in section 5. We conclude the paper in section 6.

2 Sound Localization System

As for monitoring purposes, we should consider various kinds of sound in a space. For example, people in a room may make many kinds of sounds such as voice, cough, sneezing, footsteps, opening and closing sound of door, etc. Therefore, we cannot exploit the features that are only effective for human voice. In this paper, we use a relatively simple sound feature in our sound localization system. We exploit time-delays of all the possible pairs of microphones in our system.

If there are N ($N \geq 3$) microphones, the total number of microphone pairs becomes ${}_NC_2$. Therefore, a time-delay vector is expressed by $M = {}_NC_2$ elements. It holds rich information enough to estimate the spatial location of the sound source if the geometry of microphone layout is given to the system. Note that the time-delay vector only depends on the location of the sound source. It is uniquely given if a sound source is set in a certain place and all the microphones are fixed in the space. Therefore, if two time-delay vectors are same, it means the corresponding sound sources are placed in the same place.

We here assume that the temperature of the space is constant so that the sound speed is constant.

3 Query for Sound Sample

In our approach, a sound is localized by finding a sound sample that has the same time-delay vector. Sound samples had been recorded and their time-delay vectors were stored in a database in the system in advance.

As for applications, suppose the system has a calibrated camera and it found a new sound. The system will estimate the closest sound sample and point out the place of the sound source in the video image that has taken at the time when the sound was recorded.

The system needs a large number of sound samples in the database to cover the whole space. However, as the similarity estimation of our method is rather simple (explained in 3.2), the computation cost to find the closest sound sample in the database is within the practical range for various on-line applications.

3.1 Time-delay Vector

By recording a sound with two microphones i and j, we can obtain a time-delay $\tau_{i,j}$ between i and j.With N microphones, we will have $M = {}_NC_2$ time-delay values for one sound. Time-delay vector T is denoted as $T = (\tau_{1,1}, \tau_{1,2}, \cdots, \tau_{N-1,N})$.

We basically estimate the time-delays of a microphone pair by simply searching the start time of sound wave at each microphone for an isolated single sound (such as closing sound of a door). We also plan to exploit Cross-Power Spectrum Phase (CSP)[11] and its improved approach to estimate more accurate time-delay τ because we need to handle various kinds of environmental noise such as noise of fans and motors of electric devices.

3.2 Similarity Scale for Time-Delay Vector

We define a similarity scale for evaluating the similarity between the time-delay vectors of two sound sources.

Obviously, the distance should be zero and the similarity should be high if the two time-delay vectors are same because it implies that the two relevant sound sources are at the same place.

Suppose there are two sound sources α and β, and the corresponding time-delay vectors T_α and T_β. We define the similarity scale $ss(T_\alpha, T_\beta)$ for two time-delay vectors based on the definition of the Euclidian distance.

$$ss(T_\alpha, T_\beta) = \Big(\sum_{1 \le k \le M} (\tau_{\alpha_k} - \tau_{\beta_k})^2 \Big)^{\frac{1}{2}} \tag{1}$$

$\tau_{\alpha k}$ indicates the kth element of the time-delay vector T_α.

Since time-delay vector space is a none-linear projection of the real space where sound-α and sound-β exists, we need to examine the behavior of the similarity scale when the two sound sources are in different positions. For example, $ss(T_\alpha, T_\beta) = ss(T_\beta, T_\gamma)$ does not imply $ed(\alpha, \gamma) = 0$, where the function $ed(\ )$ shows the Euclidian distance betweem the two sound sources in the real space.

This behavior is affected by the layout of the microphones in a space and the location of sound source.

Therefore, if we find a good microphone layout in which the similarity scale has the strong correlation with the actual distance in the real space, we can use

the scale to find the closest sound sample in the database for a newly observed sound.

As for a comparison of the proposed similarity scale, we also prepare the normalized inner product $ip()$ in this paper. It is often used as a similarity scale for two multidimensional vectors. Suppose there are two sound sources α and β, and the corresponding time-delay vectors T_α and T_β, the normalized inner product is defined as:

$$ip(T_\alpha, T_\beta) = \frac{(T_\alpha, T_\beta)}{\|T_\alpha\|\|T_\beta\|} \tag{2}$$

4 Simulation Experiment

We have examined five kinds of popular microphone layout in a room in simulation experiment. We also compare the proposed similarity scale with a normalized inner product to evaluate the accuracy of the proposed method.

In the simulation, we set the size of experimental space as 6.201 by 7.288 by 3.094 (depth/width/height)[m], which is the size of our real experiment room. We prepare the pre-recorded sound samples at the interval of 0.3, 0.6, and 1.0 meter for depth, width, and height direction. The numbers of the pre-recorded sound samples are 223, 857, and 5,774 respectively. The corresponding time-delay vectors are calculated in advance. We assume ominidirectional and undamped sound sources and sound samples. And we do not consider reverberation during the simulation process.

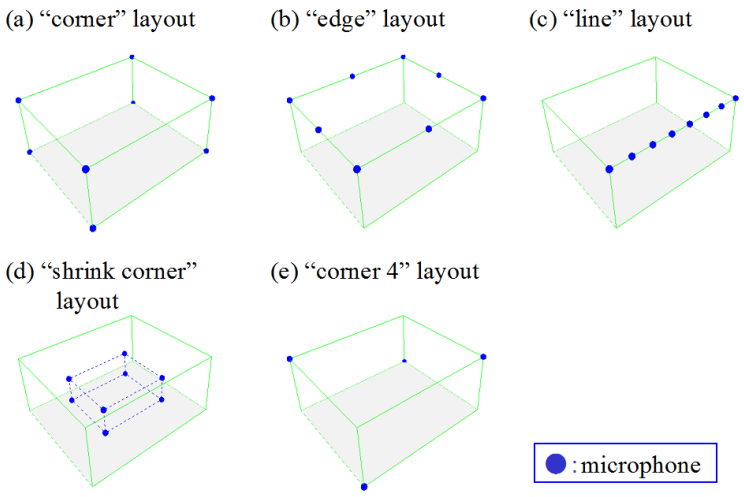

Fig. 1. Microphone layouts

Table 1. Simulation result

layout	method	interval	ref	success	neighbor	failure	average [m]	variance	deviation
corner	ss	1.0 [m]	a	4675	325	0	0.483	0.020	0.143
		0.6 [m]	b	4713	287	0	0.286	0.007	0.085
		0.3 [m]	c	4724	276	0	0.144	0.002	0.043
	ip	1.0 [m]	d	2009	1665	1326	0.904	0.308	0.555
		0.6 [m]	e	1846	1428	1726	0.647	0.233	0.483
		0.3 [m]	f	1728	1392	1880	0.380	0.123	0.351
edge	ss	1.0 [m]	g	1881	1586	1533	0.930	0.336	0.580
		0.6 [m]	h	1700	1423	1877	0.618	0.166	0.408
		0.3 [m]	i	1630	1346	2024	0.331	0.052	0.229
	ip	1.0 [m]	j	1831	1797	1372	0.920	0.319	0.565
		0.6 [m]	k	1869	1510	1621	0.609	0.200	0.448
		0.3 [m]	l	1787	1496	1717	0.337	0.086	0.294
linear	ss	1.0 [m]	m	714	1528	2758	1.425	0.639	0.799
		0.6 [m]	n	482	998	3520	1.256	0.659	0.811
		0.3 [m]	o	261	533	4206	1.134	0.612	0.782
	ip	1.0 [m]	p	427	1050	3523	1.939	1.374	1.172
		0.6 [m]	q	299	696	4005	1.532	0.865	0.930
		0.3 [m]	r	169	328	4503	1.280	0.676	0.822
shrink corner	ss	0.6 [m]	s	3240	1395	365	0.378	0.045	0.213
corner 4	ss	0.6 [m]	t	3454	1478	68	0.332	0.019	0.139

The performance of sound source localization is evaluated as follows. A new sound sample is randomly placed in the space, and the corresponding time-delay vector is calculated. Then, the system finds the most similar time-delay vector by applying the proposed similarity scale (and the normalized inner product for comparison). If the answer is also the closest sample to the new sound sample in the real space, it is marked as success. If the answer is not the closest one but is one of the 7 second-best pre-recorded samples (8 corners of a cube that involves the new sound source, excluding the closest corner itself), it is marked as neighbor. Otherwise, it is marked as failure. Each experiment is conducted with 5,000 randomly placed samples.

We have first examined the three kinds of microphone layouts to compare the proposed method ("ss") with the normalized inner product ("ip").

The first layout has 8 microphones at each corner of the space ("corner"). The second one has 4 microphones at the corners on the ceiling, and 4 microphones in the mid point of the 4 edges of the ceiling ("edge"). In the third layout, 8 microphones are linearly arranged and their intervals are equally set. This linear microphone array is set on one side of the wall at the ceiling height ("line"). Fig 1 (a) - (c) shows these three layouts.

The simulation results are shown in Table 1. The average indicates the average error distance between a random sample and the corresponding "closest"

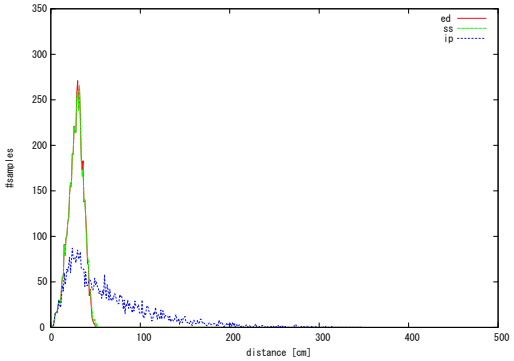

Fig. 2. Error distribution of “corner”

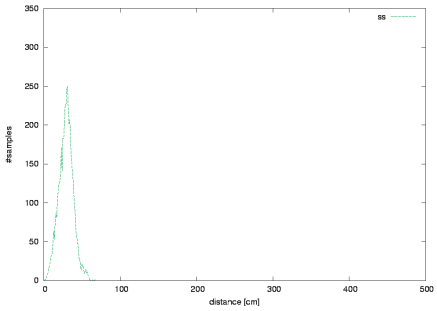

Fig. 3. Error distribution of “ss” on “corner” layout

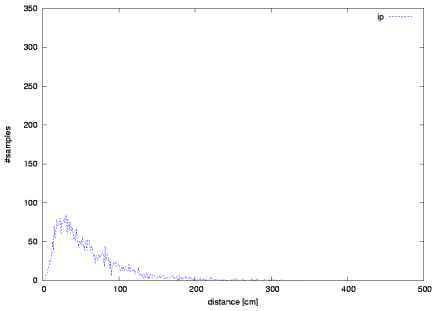

Fig. 4. Error distribution of “ip” on “corner” layout

pre-recorded sample in the real space. The variance and the deviation show the distribution of the error.

The (a) - (c) rows in Table 1 clearly show that our proposed method achieved the best score with “corner” layout. The proposed method shows the better results than those of the comparison method “ip” in “corner” and “line” layouts at any resolution.

Fig 2 - 4 shows the distance distribution of the samples at 0.6 meter interval. The line “ed” in Fig 2 indicates the Euclidian distance between the random samples and their nearest pre-recorded samples. The line “ss” plots the Euclidian distance between the random samples and their answers of the proposed method, and “ip” plots the Euclidian distance between the samples and the answers given by the normalized inner product method. Since “ss” is very similar to “ed”, we can say that “ss” can be treated as the real Euclidian distance in this case. In Fig 2, “ss” and “ed” have a very similar distribution and they are almost overlapped each other.

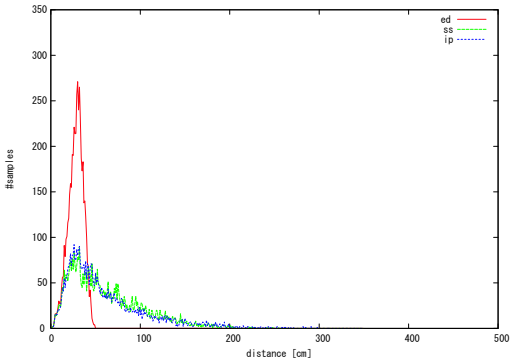

Fig. 5. Error distribution of “edge”

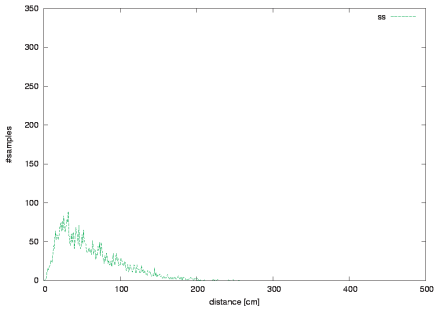

Fig. 6. Error distribution of “ss” on “edge” layout

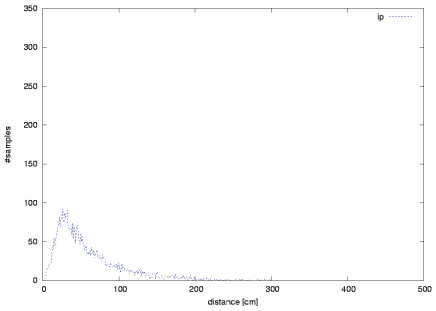

Fig. 7. Error distribution of “ip” on “edge” layout

However, even with the “ss”, the performance becomes worse if the microphone layout loses the cubic expansion of its baselines(see Table.1 (g-i)(m-o)). Fig 5 - 7 and Fig 8 - 10 also shows this fact because “ss” is different from “ed” in these cases. Therefore, we can say that the cubic layout is preferred to obtain good performance on sound source localization by our method.

We also conducted two additional experiments to examine the performance of the microphone layout that saves space and number. The row (s) in Table 1 is the result of the 8 microphones that are similar to “corner”, but it is shrunk in half to the center of the space (“shrink corner”). This layout marked the better result than “edge” and “line”. If we are allowed to place microphones at corners, we can get very good performance even if we use only 4 microphones (2 at the opposite corners on the floor and 2 at the other opposite corners on the ceiling) as shown in row (t), “corner 4”.

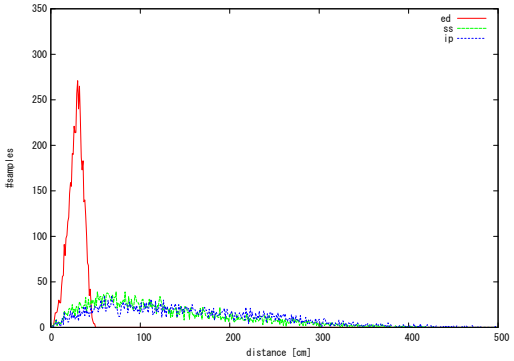

Fig. 8. Error distribution of "line"

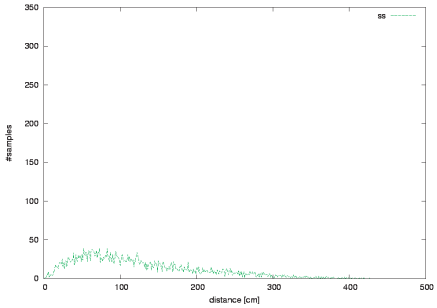

Fig. 9. Error distribution of "ss" on "line" layout

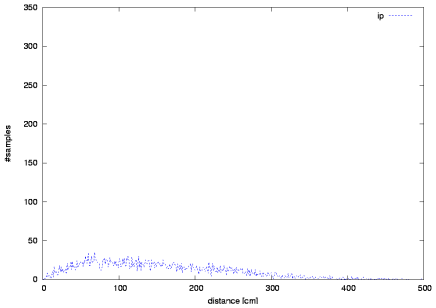

Fig. 10. Error distribution of "ip" on "line" layout

5 Experiment in Real Environment

We also conducted a preliminary experiment in a real room. We placed 4 microphones so as to expand them three dimensionally. Fig 11 and Table 2 show the layout of microphones and the positions of sound sources. Fig 12 shows a snapshot of the room.

Table 2. Positions of microphones and sound sources

	mic1	mic2	mic3	mic4
X	5.090	5.643	2.210	3.059
Y	5.314	0.958	5.824	0.4130
Z	0.159	1.229	0.812	2.599

	place1	place2	place3	place4
X	2.182	4.814	4.771	2.723
Y	5.306	5.054	3.095	3.073
Z	0.686	0.002	0.757	0.001

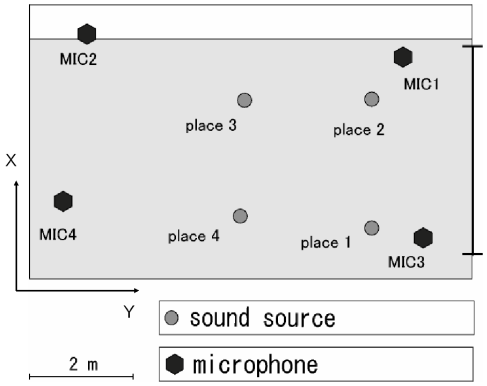

Fig. 11. Layout of microphones and positions of sound sources

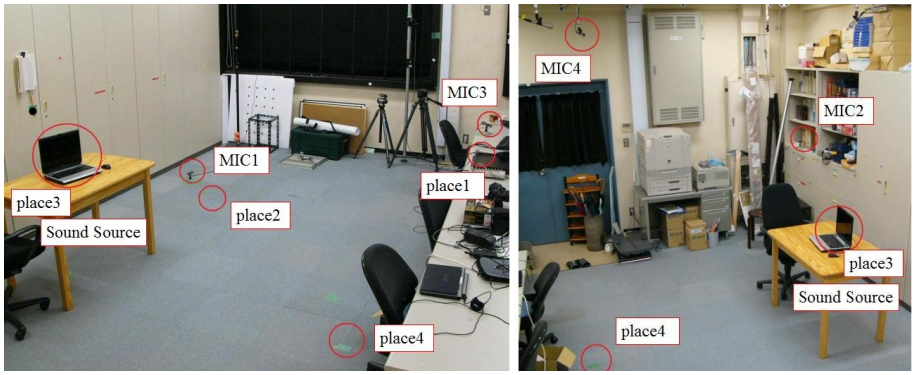

Fig. 12. Positions of microphones and sound sources

We prepared two kinds of recorded sounds. One is "voice" and the other is door closing sound of a locker ("door"). They were played on a speaker at 4 different places. Table 3 shows the values of the similarity scale between "voice" and "door". Note that the values are always low when the both sounds are observed at the same place.

6 Conclusion

We presented a method to localize sound sources by using a number of non-calibrated microphones. Our method exploits time-delay vectors and it does not need calibration step of the microphones. Since the performance of the proposed method is affected by microphone layout, we conducted a simulation experiment and concluded that spatially expanded layout is the best. We also conducted a preliminary experiment in real environment and got a promising result.

Table 3. The value of the similarity scale in real environment

		door			
		place 1	place 2	place 3	place 4
voice	place 1	47.0	868.9	789.7	460.0
	place 2	683.2	284.6	558.9	448.9
	place 3	848.1	894.5	75.1	433.5
	place 4	471.1	698.5	417.9	22.9

Since the experiments are at preliminary level, we need to apply the proposed method in various real situations to validate it. As for future works, we plan to improve the method to cope with multiple sound sources and reverberation that sometimes make influence on the performance of the sound source localization in real situations.

References

1. M. Brandstein and D. Ward, Eds.: Microphone Arrays: Signal Processing Techniques and Applications. Springer, Reading, Massachusetts, 2001.
2. M. Brandstein, J. Adcock and H. Silverman.: A closed-form method for finding source locations from microphone-array time-delay estimates. ICASSP95, 3019–3022, 1995.
3. M. Omologo and P. Svaizer.: Acoustic source location in noisy and reverberant environment using csp analysis. ICASSP96, 921–924, 1996.
4. J. M. Peterson and C. Kyriakakis.: Hybrid algorithm for robust, real-time source localization in reverberant environments. ICASSP05, Vol.4, 1053–1056, 2005.
5. V. C. Raykar, B. Yegnanarayana, S. R. Parasanna, and R. Duraiswami.: Speaker localization using excitation source information in speech. IEEE Trans. Speech and Audio Processing, Vol.13, no.5, 751–761, 2005.
6. A. O' Donovan, R. Duraiswami and J. Neumann.: Microphone arrays as generalized cameras for integrated audio visual processing. CVPR07, 1–8, 2007.
7. J. Scott and B. Dragovic.: Audio location: Accurate low-cost location Sencing. Pervasive05, LNCS 3468, 1–18, 2005.
8. M. H. Coen.: Design principles for intelligent environments. Proceedings of AAAI, 547–554, 1998.
9. T. Mori, H. Noguchi, A. Takada, and T. Sato.: Informational support in distributed sensor environment sensing room. RO-MAN04, 353–358, 2004.
10. S. Nishiguchi, Y. Kameda, K. Kakusho, and M. Minoh.: Automatic video recording of lecture's audience with activity analysis and equalization of scale for students observation. JACIII, Vol.8, No.2, 181–189, 2004.
11. M. Omologo. et al.: Acoustic event location using a crosspower-spectrum phase based. technique. ICASSP94, Vol.2, 273–276, 1994.