Proceedings of the 2nd World Congress on Electrical Engineering and Computer Systems and Science (EECSS'16)
Budapest, Hungary – August 16 – 17, 2016
Paper No. MHCI 110
DOI: 10.11159/mhci16.110

Implementation of Speaker Identification Using Speaker Localization for Conference System

Sung-Woo Byun, Seok-Pil Lee
Dept. Computer Science, SangMyung University
20, Hongjimun 2-gil, Jongno-gu, Seoul, Korea
123234566@naver.com, esprit@smu.ac.kr

Abstract - As signal processing and computing environment has developed, researches on speaker analysis technologies have been increasing. A speaker localization has become an active area of research with widespread applications in many speaker analysis fields. Many researches on a speaker localization have focused on steering camera and tracking active speakers. We also focus on tracking active speakers precisely. In this paper, we estimate 3-dimensional coordinates of the speaker using a time delay estimation and implement speaker identification for the conference system. For this, the 3-microphones array is used. To evaluate the performance of the proposed system, precision rate and recall rate are used.

***Keywords*:** Time Delay Estimation, TDOA, Speaker Localization, Speaker Identification, Conference System

1. Introduction

As signal processing and computing environment has developed, researches on speaker analysis technologies, like speaker recognition, speaker emotion analysis and speaker localization, etc. have been increasing. Speaker localization is defined as the determination of the coordinate of speaker in 3-dimensional spaces. Speaker localization has become an active area of research with widespread applications in many speaker analysis fields like an automatic steering, zooming cameras and a gesture recognition during the video teleconferencing. The accuracy of information of the speaker'. placement is also useful for various applications and other multimodal service [1].

In speaker localization, the location is estimated using time delay estimation (TDE) according to the difference in position of microphones. There are two ways the conventional strategies of time delay estimation. One is cross spectral function based and the other is generalized cross correlation (GCC) function [2]. For this research, we estimate the time difference of arrival (TDOA) using generalized cross correlation function.

Previous researches on the conference system using speaker localization have focused on steering camera and tracking active speakers [3][4][5]. We also focus on tracking active speakers precisely. In this paper, we estimate 3-dimensional coordinates of the speaker using a time delay estimation and implement speaker identification for the conference system. To evaluate the performance of the proposed system, precision rate and recall rate are used. The experimental result shows the performance of our system is very promising.

The rest of this paper is organized as follows. Section 2 explains the speaker localization. Section 3 shows the experiment configuration of our conference system based on Speaker identification and the result. Section 4 presents summaries and conclusions.

2. Speaker Localization

2.1. TDOA

The TDOA is defined as the time interval between each microphone from a speaker. For this research, we use 3 microphones to estimate the time difference of arrival to each microphone. First, in order to calculate the time interval, we compute the cross correlation function of the two signals. The lag at which the cross-correlation function has its maximum is taken as the time delay between the two signals [2]. The distance between the set of microphones is estimated by multiplying the time interval by the speed of sound in the acoustical medium (air, 330m/s).

$$T_{delay} = \text{argmax}(\rho_{s1s2}(\tau)) \quad (1)$$

Following the equation (1), $\rho_{s1s2}(\tau)$ is the cross-correlation function of the two signals.

2.2. 3-Dimensional coordinate estimation

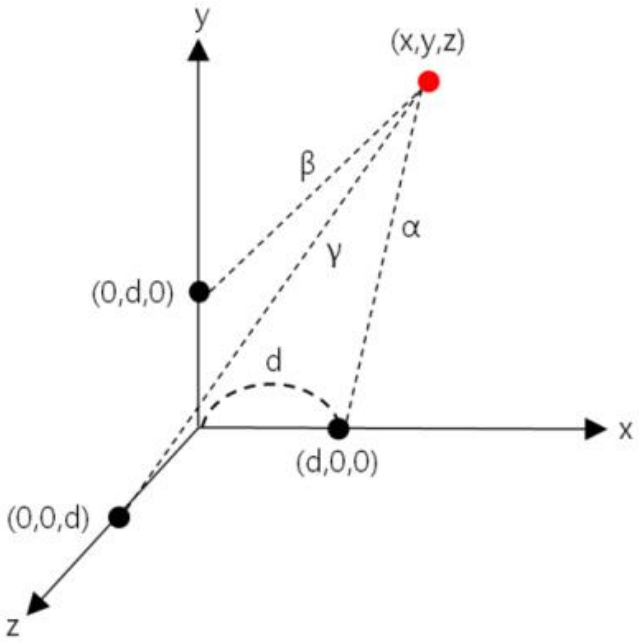

Fig. 1: The example of 3-dimensional coordinate estimation.

When the source is occurred at red point (x, y, z), the equation can be expressed as,

$$(x-d)^2 + y^2 + z^2 = \alpha^2 \quad (2)$$
$$x^2 + (y-d)^2 + z^2 = \beta^2 \quad (3)$$
$$x^2 + y^2 + (z-d)^2 = \gamma^2 \quad (4)$$

Solve for (x, y, z) in the following equation,

$$x = \frac{\beta^2 - \alpha^2}{2d} + y \quad (5)$$
$$y = \frac{\gamma^2 - \beta^2}{2d} + z \quad (6)$$
$$z = \frac{\alpha^2 - \gamma^2}{2d} + x \quad (7)$$

And then equations (2), (3), (4) are substituted by equation (5), (6), (7) as following:

$$3x^2 + 2(i-d-k)x + d^2 + k^2 + i^2 - \alpha^2 = 0 \quad (8)$$
$$3y^2 + 2(k-d-j)y + k^2 + d^2 + j^2 - \beta^2 = 0 \quad (9)$$
$$3z^2 + 2(j-i-d)z + i^2 + j^2 + d^2 - \gamma^2 = 0 \quad (10)$$

From this, k is $(\beta^2 - \alpha^2)/2d$.i is $(\alpha^2 - \gamma^2)/2d$ and j is $(\gamma^2 - \beta^2)/2d$. Therefore, the 3-dimensional coordinate of source can be estimated by solving the quadratic equation from x, y, z.

2.3. Speaker localization experiment

Fig. 2: The experimental environment.

We performed an experiment of speaker localization while increasing the distance between the speaker and the microphone. The speaker was randomly placed. The distance between microphones and the origin were set to d = 30cm, and the distance of 1 cm was set to 1 point at 3-dimensional space. The experiment was done 10 times for each case to minimize the human error. The experimental environment was done in the figure 2 diagram.

Table 1: The result of speaker localization experiment.

	Mean for absolute error of the position		
	x	y	z
100Cm	5.91	5.6	2.05
150Cm	6.59	7.02	2.6
200Cm	9.26	9.45	3.8

Table.1 shows the mean for absolute error of the position. According to this result, in the case of 100Cm, the mean for absolute error is about 4.52. As for the distance increasing, the mean for absolute error also increases.

3. Speaker Identification for the Conference System

3.1. Experiment configuration

For the experiment configuration, we performed an experiment in a regular room with a size of 500cm in width, 630cm in length and 250cm in height. The speakers were placed about 150cm apart from the 3-microphones array, and the experiment was performed 3 times. Within each experiment, we placed 2, 4, and 6 speakers randomly. Each experiment was performed for 10 minutes, and each speaker spoke in free debate for 30 seconds at once.

After the experiments were finished, we extracted the voice segments, and mapped it with the correspondence according to the speaker localization. The signals are recorded at sampling rate of 16 kHz and 16 bit resolution.

3.2. Evaluation Criteria

In this research, we used the precision rate and the recall rate which are commonly used for basic measures to evaluate the experiment.

The precision rate is the ratio of the number of correctly labelled segments to the total number of extracted segments. It is shown in (11)

$$\text{Precision rate} = \frac{|T \cap R|}{R} \tag{11}$$

The recall rate is the ratio of the number of correctly labelled segments to the total number of correct segments. It is shown in (12)

$$\text{Recall rate} = \frac{|T \cap R|}{T} \tag{12}$$

From this, R is the total number of extracted segments. T is the total number of correct segments.

3.3. Results

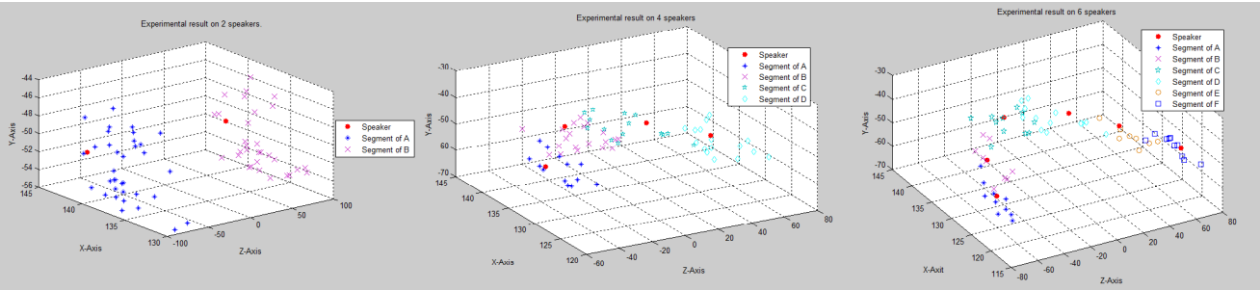

Fig. 3: the result to map the voice segments onto each speaker.

As a result, figure 3 shows the result to map the voice segments onto each speaker during the 10 minute conversation. The location of the voice segments and the speakers are marked in 3-dimensional space.

Table 2: The result of the conference system experiment.

Number of speaker	Average precision rate	Average recall rate
2	100%	100%
4	90.5%	91%
6	86.5%	86.7%

Table 2 shows the result of evaluating performances. In this result, the average of precision rate and recall rate was at maximum 100% for the case which contained 2 speakers, and it has decreased while increasing the number of speakers. In the case of 4 speakers, the average precision rate was 90.5% and the average recall rate was 91%. For the 6 speakers, the average precision rate was 86.5% and the average recall rate was 86.7%.

In a previous research, Rafal Samborski performed the conference system based on 2-dimensional information such as the phase feature. He tested on five male by setting them randomly around the table during the 28 minutes [3]. We compared Rafal's research method with our 3-dimensional coordinate estimation method.

Table 3: a comparison of 3-dimensional coordinate estimation with 2-dimensional information method.

	3-dimensional coordinate estimation	2-dimensional information method[3]
Accuracy	86.2%	80%
Precision rate	87.7%	43%
Recall rate	86.1%	50%

As shown on table 3, Accuracy, precision rate and recall rate of 3-dimensional coordinate estimation were better than 2-dimensional information method.

4. Conclusion

Speaker localization has become an active area of research with widespread applications in many speaker analysis fields. Previous researches on the conference system using speaker localization have focused on steering camera and tracking active speakers. We also focus on tracking active speakers precisely. In this paper, we estimate 3-dimensional coordinates of the speaker using a time delay estimation and implement speaker identification for the conference system. For this, the experiment is done 3 times with the 3-microphones array. To evaluate the performance of the proposed system, precision

rate and recall rate are used. The experimental result shows the performance of our system is very promising. Additional tests for estimating 3-dimensional coordinates in noisy environment has been left for future works.

Acknowledgements

"This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-R0992-16-1014) supervised by the IIT((Ittt itut frr IIff rr mtt i&&&&mmmmiii aati hhhhlll ggymmmmmm)) "

References

[1] M. Hesam and H. Marvi, II mrr ovemttt ff eett or aaee mll tipl ppaakrr loaaliztt i i mmrr t rmmm in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, 2010.

[2] A. K. Tellakula, ccc sss ti ccccc ccaaaa liztt i Usigg Tim Dll yy ttt imtt inn"" Degree Thesis. Bangalore, India: Supercomputer Education and Research Centre Indian Institute of Science

[3] R. Samborski and M. Ziolko, aaaaa arr lcc ll iztt io i ffff ernnii gg yytt mm mmll yyi ssss features and wavelet trnnfformiiii n *Signal Processing and Information Technology (ISSPIT)*, 2013.

[4] H. Sayoud, S. Ouamour, and S. Khennouf, n mtt ff eee.. rr loaalization using the filtered correlation in *Industrial Mechatronics and Automation (ICIMA), 2010 IEEE 2nd International Conference on*, 2013.

[5] S. Ouamour and H. Sayoud, mmmmmmm ppaakrr laaaliztt i aaee ppaak.. innntification - A smart room llll iaati""""" in *Information and Communication Technology and Accessibility (ICTA), 2013 IEEE Fourth International Conference on.*