Development of Small-size and Low-priced Speaker Detection Device Using Micro-controller with DSP Functions

Shinya Takahashi, Tsuyoshi Morimoto

***Abstract*—This paper report that the implementation of the speaker direction detector using a small and low-priced micro controller called ”dsPIC” and propose the speaker localization method integrating the results from the multiple detectors. To confirm a basic performance of the proposed method, an experiment for estimating 6 locations at intervals of 1m using a pair of the speaker direction detectors is conducted. The experimental results show a good prospect of the proposed method using the small and low-priced devices.**

***Index Terms*—Speaker Localization, Direction of Arrival, Cross-correlation, DSP, Embedded System.**

I. Introduction

IN conventional speech processing system, it is important to capture the clear and noiseless speech signals so that a close-talking microphone, which is equipped near the mouth of a speaker, is usually used. It is, however, inconvenient to equip such a head-set microphone and it has some constraints in respect of the number of the speakers and the cost of the instruments. To cope with this problem a hands-free speech processing technology with distant microphones has been studied in recent years. Using such a distant microphone, a natural communication can be realized without any user’s burden.

It is necessary to reduce the influence of environmental noises for estimating a speaker direction with the distant microphone. It is also important to locate a speaker (“who spoke where?”), which is called “speaker localization” problem, for the speaker detection in the remote conference system or the automatic meeting recording system with the distant microphones.

Generally, a microphone-array speech processing system is widely used for the estimation of speaker direction, so-called DOA (Direction Of Arrival) estimation in signal processing[1],[2].

Microphone-array can also be used to enhance target speech signals, remove noises and/or cancel echos. However, it is not easy to actually implement the microphone-array in respect of the cost and portability because it requires a large scale device with a number of microphones[3].

In this paper we propose a distributed speech processing system using small-size and low-priced micro controllers with DSP functions. The main topic of this paper is the realization of the speaker localization as the embedded system in trial. In the following section we firstly describe an overview of the distributed speech processing system and then report the implementation of the actual device and the experimental results.

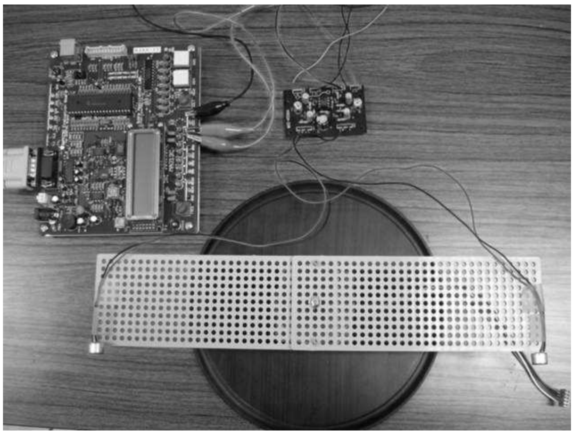

Fig. 1. Overview of Speaker Direction Detector

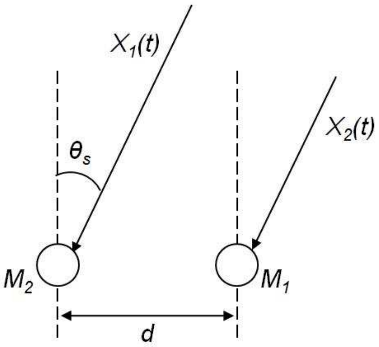

Fig. 2. Direction of Arrival

II. Speaker Localization

Fig.1 is the speaker direction detector device we used. This device receives speech signals from a stereo microphone amplifier as an input, executes digital signal processing on a small-size micro controller called “dsPIC”, and sends an estimation result of the speaker direction to a host computer. We explain the principle of the estimation of speaker direction and the speaker localization in this section and describe the specification of the implementation in the next section.

A. Estimation of Speaker Direction Using Cross Correlation

Let us consider the speech signal received by two microphones which are placed at a distance of d as shown in Fig. 2.

Manuscript received December 29, 2010; revised January 17, 2011.

Dept. EECS, Fukuoka University, 8-19-1 Nanakuma, Jonan-ku, Fukuoka-shi, 814-0180 Japan (Tel: +81-92-871-6631(ext.6572); Fax: +81-92-871-6031; Email: takahasi@tl.fukuoka-u.ac.jp).

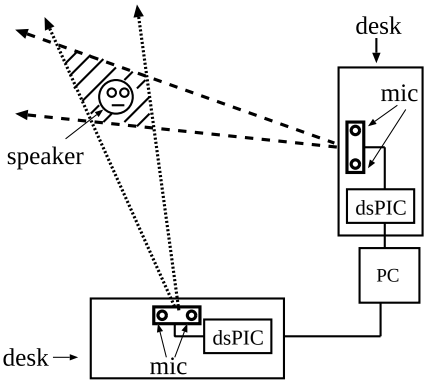

Fig. 3. Estimation method of speaker location

The signal from the direction θ_s arrives at the microphone M_1, then travels a distance ξ and arrives at the microphone M_2. Since an equation $\xi = d \sin \theta_s$ is hold, the DOA θ_s can be calculated by the following equation.

$$\theta_s = \sin^{-1}\left(\frac{v\tau_s}{d}\right), \tag{1}$$

where v is the velocity of sound and τ_s is a time that the signal requires to travel the distance ξ.

Therefore, if a correct τ_s (= $\hat{\tau}$) can be obtained, then DOA can be estimated. Many methods for calculating this $\hat{\tau}$ have been proposed so far. We use TDOA (Time Difference of Arrival) approach, which is one of the most basic and simplest approach, in this paper.

In TDOA, $\hat{\tau}$ can be obtained by maximizing the cross-correlation between $X_1(t)$ and $X_2(t)$ as follows,

$$\Phi(\tau) = \frac{1}{N} \sum_{t=0}^{N-1} X_1(t) X_2(t+\tau), \tag{2}$$

provided that the range of $-T \leq \tau \leq T$ [1].

B. Speaker localization by combining two speaker direction detectors

Although the DOA can be estimated by using the above principle, it is not easy to estimate the speaker location with only one speaker direction detector. So we consider utilizing a pair of the speaker direction detectors.

That is, the location to be estimated is the region enclosed by the range of the DOA result plus a permissible error as shown in Fig. 3.

III. Implementation on dsPIC

A. Target hardware

The DOA estimation program is implemented using a micro controller chip called dsPIC30F4011 made by Microchips. In this paper, we used an evaluation board of MA213 made by MicroApplication, which has the above dsPIC on board, to shorten the time to prepare a circuit. The block diagram of MA213 is shown in Fig.4. Table I shows a data sheet for dsPIC and the microphone amplifier.

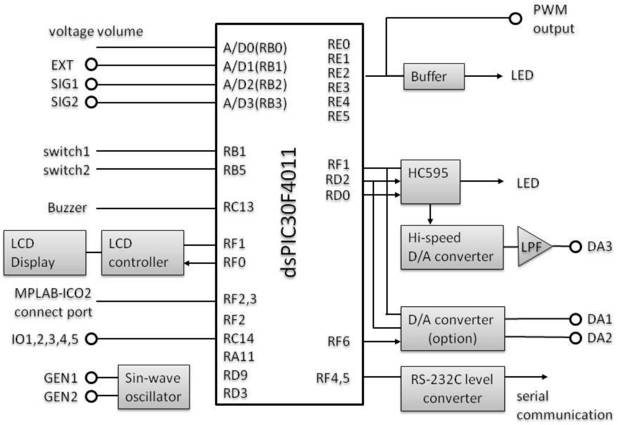

Fig. 4. Block diagram of MA213

TABLE I
HARDWARE DATA SHEET

dsPIC30F4011	
CPU speed	Up to 30 MIPs
Program memory size	48kbytes
Data memory size	2048 bytes
AD converter	10bit, 4channel

Component of Microphone input	
Amp	OP-AMP4588
Voltage gain	220
Mic	condenser microphone × 2
Distance between mics	0.3m

B. Implementation

Sampling signals from the microphones are executed with 625 cycles of interrupt timing for A/D converter. The sampling rate is 32 kHz. After A/D converting 256 samples per 1 frame, which equals 8ms, as fixed-point numbers (Q15 format) from the left and right microphones respectively, the sampling data are stored in 10-bit accuracy in two working memories (X Data RAM/Y Data RAM) which support high-speed addressing[4]. At first, the power of the speech signal is calculated by a DSP function VectorPower() of dsPIC to decide whether the input signal should be target or not. Next, the cross-correlation is calculated by a DSP function VectorDotProduct() after shifting sampling data in each frame with the range of $-30 < \tau < 30$. Here, the dsPIC has a DSP function VectorCorrelate() which can calculate a cross-correlation, but it requires more memories so that we used VectorDotProduct() which simply multiplies two arrays for saving memory. Then, an angle of DOA is estimated by calculating the maximum value of the accumulative cross-correlation for a block of 32ms which consists of 4 successive frames. In this paper, we set the step size of the DOA estimation to 5° because the average error of the delay for 1 sample is expected to be 3.2° from the fact that the distance of the microphones is 30cm.

The reason why we calculate the accumulative values of the cross-correlation for 1 block is due to saving the

[1] T is the delay range restricted by a sampling rate f_s and the distance between microphones, d. It can be calculated as

$$T = \frac{df_s}{v} + 1$$

.

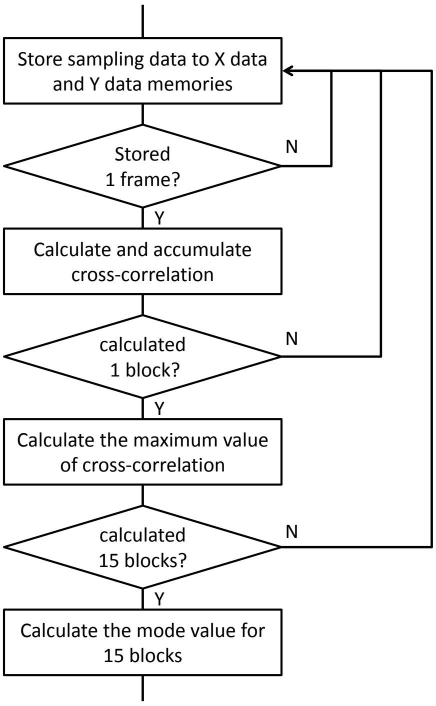
Store sampling data to X data and Y data memories
Stored 1 frame?
N
Y
Calculate and accumulate cross-correlation
calculated 1 block?
N
Y
Calculate the maximum value of cross-correlation
calculated 15 blocks?
N
Y
Calculate the mode value for 15 blocks

Fig. 5. Flowchart of DOA calculation

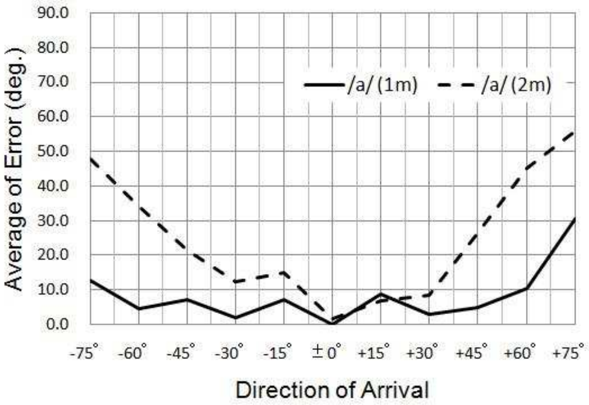

Fig. 6. Experimental results

working memory. It is equivalent to calculating the whole cross-correlation for 1 block because the calculation of the cross-correlation is simply accumulating the multiplication of values in two vectors. Therefore the equation 2 can be replaced the following form,

$$\Phi(\tau) = \sum_{n=1}^{M} \phi_n(\tau),$$

where ϕ_n is a value of the cross-correlation in a frame and M is the number of frames in 1 block. For implementation, the overlapping samples at boundaries of adjacent frames are processed appropriately.

Finally, we obtain the estimation angle of DOA as the mode value of the estimated angles in 15 blocks, which equals 480ms, to avoid including large errors. Fig.5 shows the flowchart of the above process.

TABLE II
SPEAKER LOCALIZATION IN EXPERIMENT 2

speaker location	from dsPIC-1			from dsPIC-2		
	direction		distance	direction		distance
A	left	-45°	100cm	left	-19°	224cm
B	center	±0°	141cm	center	±0°	141cm
C	right	+19°	224cm	right	+45°	100cm
D	left	-45°	200cm	center	±0°	282cm
E	left	-19°	224cm	right	+19°	224cm
F	center	±0°	282cm	right	+45°	200cm

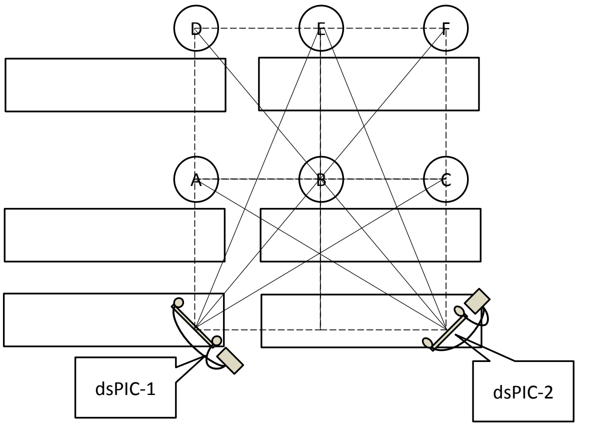

Fig. 7. Experimental environment

IV. EXPERIMENT

A. Exp.1: Estimation of speaker direction

We conducted an experiment for the DOA estimation from $-75°$ to $75°$ with a step of $15°$. The front of the microphone is $0°$. We obtained the average of the errors between the correct angle and the estimated value.

We used 2 seconds of a vowel /a/ repeating 30 times as the sound source from a computer speaker. The distance from the sound source to the microphone is 1m or 2m. Fig. 6 shows the experimental result.

As shown in Fig.6, DOA can be estimated stably in the case of 1m. However the estimation errors in the broadside angles becomes large in the case of 2m. This result suggests that the devices that can cover sound sources in the range of $\pm45°$ should be arranged if a permissible error is $20°$.

B. Exp.2: Speaker Localization

Next, we carried out the experiment for the speaker localization using a pair of the speaker direction detectors. Each detector is arranged as shown in Fig 7, based on the result of the experiment 1. In this experiment, let $-10° \leq \theta_s \leq 10°$ be the center, $-60° \leq \theta_s < -10°$ be the left and $10° < \theta_s \leq 60°$ is the right, combining the estimation result from each speaker direction detector, we estimated the speaker localization for 6 places from A to F shown in Fig. 7. We used the vowel sound /a/, which is same as the experiment 1, and estimated 50 times for each location.

Fig.8 and Table III show the experimental results. The DOA estimation results from each speaker direction detector are plotted in Fig.8.

As can be seen from this plotted graph, the results for the area A, B, C, and E are separated well. However, the results for the area D and F are mixed with the other areas. This

TABLE III
RESULTS FOR SPEAKER LOCALIZATION

Speaker location	Correct rate	Ave. Error(°)	
		dsPIC-1	dsPIC-2
A	38%	8.3	11.2
B	96%	2.2	3.9
C	28%	9.1	9.4
D	76%	19.1	11.0
E	32%	5.5	8.0
F	64%	11.9	22.6

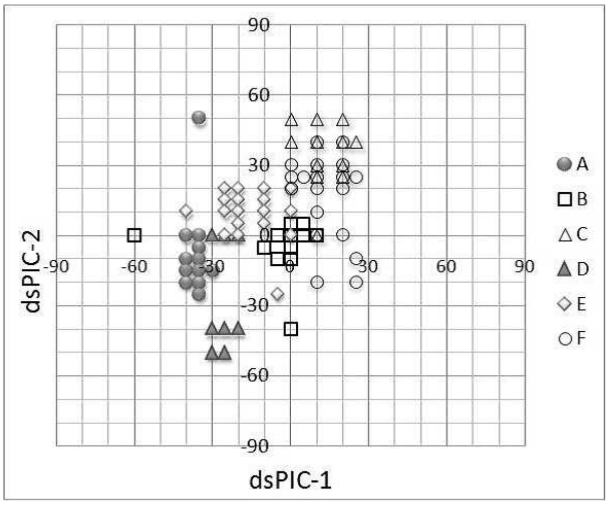

Fig. 8. Plot of experimental results

result means that it is difficult to detect accurately for the area in a far distance.

As the table shows, the correct rates are symmetrical about B-E line. The reason why the correct rates of A and C are low is because the speaker direction detector in far estimated the center by mistake. In the same way, the estimation result for E tends to be mistaken as B because the area E is located in the center of two speaker direction detector devices. As shown in the estimation error results, the accuracy of DOA for E is not low. It would be expected that E can be classified to B if it is estimated from the other side.

V. DISCUSSION & CONCLUSION

In this paper, we investigated the implementation of the speaker direction detector using a small and low-priced micro controller called ”dsPIC” and proposed the speaker localization method integrating the results from the multiple detectors. To confirm a basic performance of the proposed method, we conducted an experiment for estimating 6 locations at intervals of 1m using a pair of the speaker direction detectors. From the experimental results, although the accuracy of each detector is not so high, for a far distance especially, it can be covered the far area by appropriately arranging more detectors in the room. It is easy to increase the number of the devices because of the portability and cost performance. In future, It is necessary to investigate how to arrange more devices for the speaker localization at any places in wider room.

With further research, we intend to embed the another method for DOA such as a spectral-based algorithm into a more powerful dsPIC micro-controller for more accuracy. In addition we also intend to connect each speaker direction detector with network using Ethernet functions in the higher class dsPIC.

As other approach for increasing accuracy, we can use not only audio information but also visual information. Although there are many researches for such an audio-visual approach[5], almost of them used a special hardware or instrument for microphone arrays and video cameras. We have proposed the speaker detection system with low cost camera controlled by PIC[6]. We intend to investigate integrating the visual approach and the proposed method in this paper.

REFERENCES

[1] G.W.Eiko, “Microphone array systems for hands-free telecommunication,” Speech Communication, Vol. 20, pp.229–240 (1996).

[2] T.Yamada, *et.al.*, ”Hands-free Speech Recognition with Talker Localization by a Microphone Array (in Japanese),” Journal of Information Processing, Vol.39, No.5, pp.1275–1284 (1998).

[3] G. Lathoud and J. Odobez, “Short-Term Spatio-Temporal Clustering Applied to Multiple Moving Speakers,” IEEE Trans. on Audio, Speech & Language Processing, pp.1696-1710 (2007)

[4] Microchip, “dsPIC30F/33F Programmer ’s Reference Manual, High-Performance Digital Signal Controllers,” http://www.microchip.com/downloads/en/DeviceDoc/

[5] G. Friedland , C. Yeo , H. Hung, “Visual speaker localization aided by acoustic models,” Proc. of the 17th ACM international conference on Multimedia, pp.195–202 (2009)

[6] D. Inoue, S. Takahashi, T. Morimoto, “Speaker Detection by Sound Source Direction and Face Image (in Japanese),” Fukuoka Univ. Review of Technological Sciences, Vol. 76, pp. 23-29 (2006)