Hardware Implementation for Real-Time Speech Processing with Multiple Microphones

Cheong-Gyu Seok*, Jong-Suk Choi*, Munsang Kim* and Gwitea Park**
* Intelligent Robotics Research Center, KIST, Seoul, Korea
(Tel: +82-2-958-5630; E-mail: {robotics36,cjs, musang}@kist.re.kr.)
** Department of Electrical Engineering , Korea University, Seoul, Korea
(Tel : +82-2-3290-3218; E-mail: gtpark@elec.korea.ac.kr)

Abstract: Nowadays, various speech processing systems are being introduced in the fields of robotics. However, real-time processing and high performances are required to properly implement speech processing system for the autonomous robots. Achieving these goals requires advanced hardware techniques including intelligent software algorithms. For example, we need nonlinear amplifier boards which are able to adjust the compression radio (CR) via computer programming. And the necessity for noise reduction, double-buffering on EPLD (Erasable programmable logic device), simultaneous multi-channel AD conversion, distant sound localization will be explained in this paper. These ideas can be used to improve distant and omni-directional speech recognition. This speech processing system, based on embedded Linux system, is supposed to be mounted on the new home service robot, which is being developed at KIST (Korea Institute of Science and Technology)

Keywords: nonlinear amplification, embedded linux system, speech processing, EPLD, sound localization

1. INTRODUCTION

Recently, real-time sound localization also becomes technology of interest in the human-robot interaction area in order to recognize speech with high confidence [1][2]. And so, the real-time sound localization system is required for applications such as robust speech recognition and automatic teleconferencing, where, several audio signals obtained from an array of microphones are processed concurrently [3][4][5].

But the real-time processing of multiple audio signals is often expensive as it requires several processors for high performance and it often requires a large amount of power and complex system. To solve these problems, we propose an RTMTB (Real-Time Multi-channel Transfer Board) designed with EPLD for various controllers and algorithms, and implement it in an embedded linux system whose processor is PXA255. The RTMTB is composed of two controllers: one is for 4-channel A/D converters and the other is for memory management such as double buffering.

To efficiently amplify sound signals we use a nonlinear amplifier board. It used to control both compression ratio (CR) and amplification ratio of sound signals as well as linear pre-amplification. Also, the nonlinear amplifier board's parameters will be controlled by software programming to embody an intelligent sound system.

The detailed procedures of the sound localization are that the original speech signals are acquired by multiple microphones and amplified with the nonlinear amplifier which is useful for amplifying both distant and close speech signal without saturation in noisy environment. The nonlinear amplifier board is able to adjust the compression ratio via computer programming. By adjusting a resistance value by programming; we can change not only the compression radio but also amplification ratio.

After that, the signals are converted to digital signals by A/D converters. The converted data are stored in the memory, which is divided into two blocks, using a memory controller (EPLD). This memory controller enables the hardware to make double-buffering method.

Using this hardware of double-buffering, the real-time speech signal can be continuously buffered without interruption and delay, while the main processor keeps supporting the speech processing software algorithm.

By the double-buffering hardware, we can estimate the time-delay of arrivals (TDOA) which means time-delay between time-of-flights from the source of sounds to each microphone for the sound's direction detection in real-time and with high performance on the embedded system.

Also, a linux device driver (RTMTB device driver) and an application program are implemented on the embedded system.

The application program includes VAD (voice activity detection) and localization algorithm for seeking the source localization of speech. System platform is installed in the prototype robot, called IROBAA (Intelligent Robot for Active Audition)

Figure 1 shows the whole embedded system for the sound processing. It comprises PXA255 main board system, RTMTB, nonlinear amplifier board, and the secondary processor board.

Using the embedded system, the nonlinear amplification and the double-buffering method identified above can increase the real-time efficiency and performances in the processing of speech. Our proposed hardware system reduces the burden on software and prevents interruption and delay by simultaneously receiving real-time speech signals from multiple channels. To show its efficiency, we improved TDOA algorithm for the sound's direction detection.

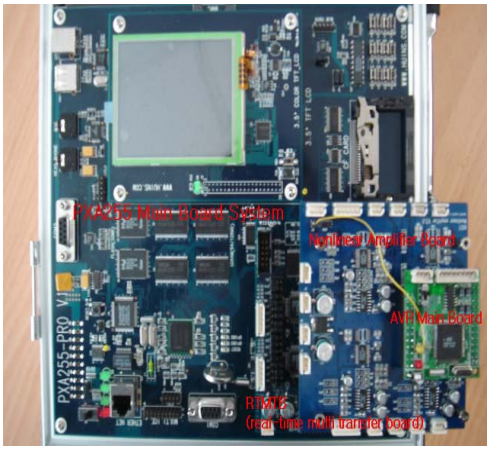

Fig.1 Embedded system for sound localization and processing

As for the further works, we will consider advanced hardware implementation for speech processing which utilizes DSP (Digital signal processor)-based or FPGA (Field programmable gate array)-based algorithms for higher performances and more modularization.

2. NONLINEAR AMPLIFICATION FOR LOCALIZATION AND RECOGNITION

In this section, we introduce both voice activity detection and localization method using the nonlinear amplifier board. We used the nonlinear amplification board which is updated such that it is able to adjust both compression ratio (CR) and the amplification ratio via computer programming in the AVR board. Fig. 2 shows the schematic diagram of the nonlinear amplifier board based on the SSM 2166 [6], and the way how to adjust the CR value by software set-up. By adjusting resistance value by programming, we can change not only the compression ratio but also the amplifying ratio.

Nonlinear amplification which is able to make dynamically variable amplification rate according to the signal's magnitude is required to increase the range of detectable distance. If the ratio of amplification is fixed to small one, the signal of speech occurring at the long distance can be hardly extracted from its received signal whose magnitude is too small to be separated from noises. To the contrary, with large ratio, the signal occurring nearby sometimes may be saturated in the A/D conversion. Here, we propose the nonlinear amplification where smaller signal can be amplified with larger amplification ratio to implement the nonlinear property. We use SSM2166, made by Analog Device Corporation.

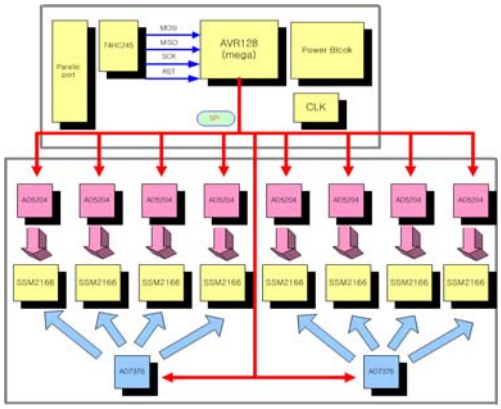

Fig.2 Nonlinear amplifier schematic block diagram

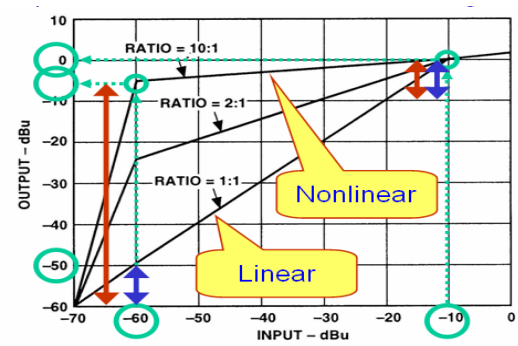

Fig. 3 Nonlinear amplification characteristics of SSM 2166 A/D board at CR is 5:1

Fig.4 Developed nonlinear amplifier board

Figure 3 shows the characteristics of the nonlinear amplification board with respect to several CRs. The x-axis represents the value of signal input and y-axis shows its amplified value in dB scale.

Think about two inputs: one is small input of 60dBu and the other is large input of -10dBu (as much of 50dBu difference in magnitude). When using 1:1 ratio, the difference in the magnitude of outputs is the same value of 50dBu as that of input, which shows the linear property. However, when using 10:1 ratio, the difference in the magnitude of outputs is just 5dBu smaller than that (50dBu) of inputs showing that the larger is the amplified ratio of the weaker signal and the smaller is the amplified ratio of stronger signal.

Our nonlinear amplifier board, as shown in Fig. 4, adjusts compression ratio of 5:1 or 1:1 and amplification ratio from 0dB to 30dB and drives up to 8 channels.

3. REAL-TIME MULTI TRANSFER BOARD

In this section, we introduce the RTMTB for pre-processing of input signal. It takes charge of sound raw data processing using an A/D converter controller for sound signal conversion and a memory controller for double buffering.

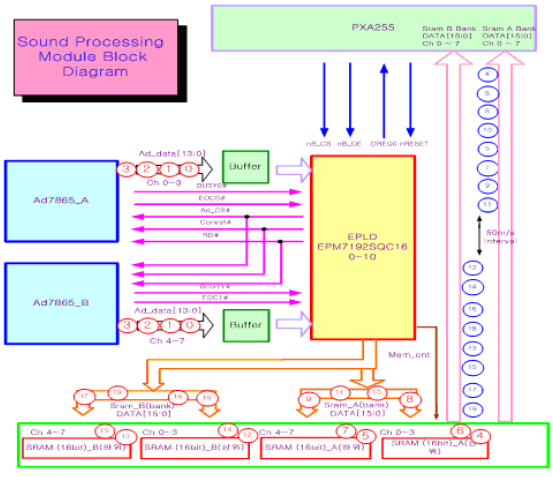

Fig.5 RTMTB block diagram for sound processing

Figure 5 shows RTMTB's block diagram. The signals are converted to digital signals by 2 A/D converters which have

14-bit resolution, and capability to handle 4-channels simultaneously. The converted data are stored in the memory which is divided into two blocks, using a memory controller (EPLD). This memory controller enables the hardware to make double-buffering method.

As for the detail of the double-buffering, the first stream set of the converted digital data is written into the first block of memory, and the next stream set is written to the second block of memory. While the secondly converted data are being written, the PXA255-based embedded system brings all the first block's data immediately and performs signal processing algorithms with the first converted data. By this way, while the main system (PXA 255) is communicating and processing with one block, new data are written to the other block without loss of data.

Figures 5 and 6 show simulation results about the A/D converter controller and the memory controller. A/D converters that we used have simultaneous sampling of 4 channel and the resolution of 14bits. And the RTMTB have 4 memory chips (16bit, SRAM).

The sampling frequency of the A/D converter is 16KHz (the next conversion time have 62.5usec delay time) and the whole channels have 9.6usec time during once conversion. The A/D converter controller is implemented using state machine blocks of EPLD while the memory controller using counter blocks.

The memory controller is designed for the double buffering structure which has the buffering time of 0.05 sec that makes the first memory block full. Figure 8 shows our RTMTB (real-time multi transfer board).

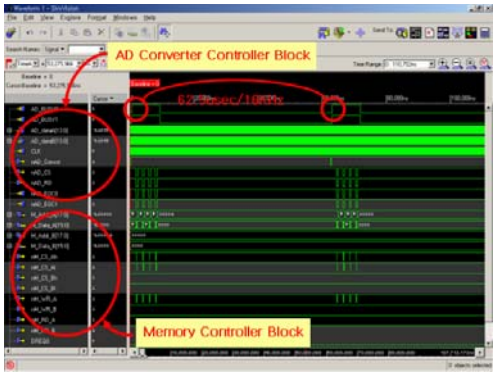

Fig.6 Simulation Result
(A/D converter controller & memory controller)

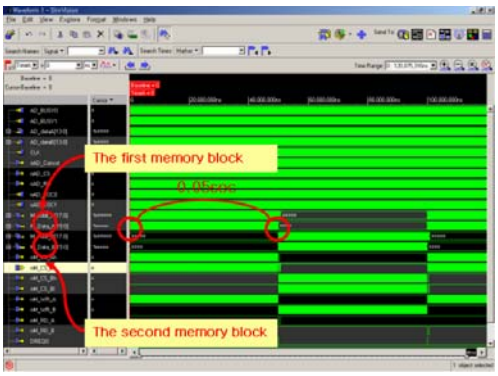

Fig.7 Simulation Result
(Double buffering timing simulation)

Fig.8 RTMTB (real-time multi transfer board)

4. EMBEDDED SYSTEM FOR SOUND LOCALIZATION

Our embedded system for sound processing has the main processor PXA255, based on the Intel XScale 32bit processor with the clock frequency of 400MHz. It comprises 32MB Flash Memory (local rom), 128M SDRAM (local memory) and support JTAG, UART, Ethernet, and etc.

The embedded system supports a peripheral 100PIN GPIO (General peripheral I/O) and can interface the RTMTB by the GPIO. It uses Linux kernel 2.4.19 and the developmental environment is GNU C Compiler, arm-cross-compiler.

We have developed a device driver in order to process 'sound raw data' (sound input data) coming from the RTMTB. Sound raw data are used in the application program which performs the localization algorithm.

The raw data comprise 800 hexa values (or float type) per one frame. Then we process at once ten frames of total 32000 hexa values among which 8000 hexa values are used VAD and the other values are used localization algorithm.

5. LOCALIZATION ALGORITHM

4.1. Localization

This paper uses DOA (Delay Of Arrival) for tracking the direction of sound. DOA is the method that uses a time-delay from the source of sound to each microphone. Even though the time delay is short, the difference of arrival time occurs between array-shaped microphones.

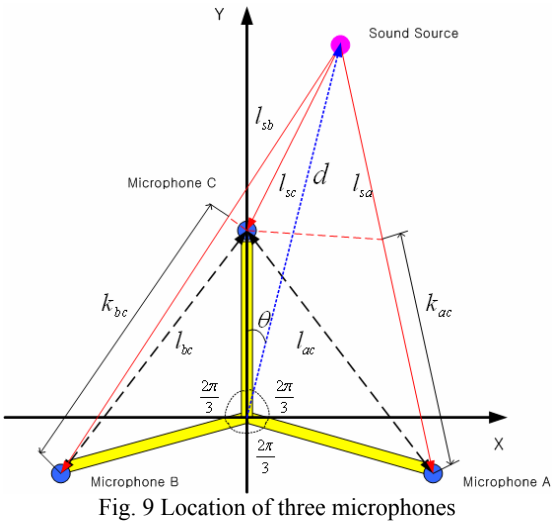

Fig. 9 Location of three microphones

In Fig. 9 three microphones are arranged such that their distances from the centre of triangular rod are the same. Two couples of A vs. C and B vs. C are selected in the view point of C. Note that the sampling data has maximum delay of time when a sound enters straight through both A and C, or B and C.

In this case, the relative distance corresponding to the maximum delay is defined as l_{ac} (or l_{bc}). Also, the distance between sound's source and microphone A (mic C) is defined as l_{sa} (or l_{sc}). The velocity of sound and sampling frequency are defined as v and F_s respectively. The number of sampling about the maximum delay is defined by (1) and (2) where n_{ac} is the number of sampling of maximum delay between A vs. C microphone and n_{bc} is the other one between B vs. C microphones.

$$n_{ac} = \frac{l_{ac}}{v} F_s \tag{1}$$

$$n_{bc} = \frac{l_{bc}}{v} F_s \tag{2}$$

The relation coefficient between mic C and mic A is defined by (3). Also, the coefficient of relation between mic C and mic B is defined by (4). The variable t_g is a target number of delay in the g^{th} sampling period. Equation (3) and (4) is considered by sampling data from g=0 to g=∞. However, the real application of infinite period is impossible. Therefore, variable t_g is determined by suitable sampling data. We should decide the optimal sampling period consisted of 552 samples through experiments.

$$R_{ac} = \frac{\sum_{g=0}^{\infty} \{A(t_g - k)C(t_g)\}}{\sqrt{\sum_{g=0}^{\infty} A(t_g - k)^2} \sqrt{\sum_{g=0}^{\infty} C(t_g)^2}} \tag{3}$$

$$R_{bc} = \frac{\sum_{g=0}^{\infty} \{B(t_g - k)C(t_g)\}}{\sqrt{\sum_{g=0}^{\infty} B(t_g - k)^2} \sqrt{\sum_{g=0}^{\infty} C(t_g)^2}} \tag{4}$$

The variable k represents the number of actual delay samples. The number of delay k, in our configuration, spans to the range of $-n_{ac} \sim n_{ac}$ in this (3) and $-n_{bc} \sim n_{bc}$ in this (4) where its positive/negative value means that the sound enters microphone A and B earlier/later than microphone C.

Now, sound's direction should be calculated using relation coefficient R_{ac} and R_{bc} for all possible k_{ac} and k_{bc}. Fig. 9 illustrates the number of delay samples and the actual angle of sound's direction. An actual delay of sound's direction is expressed as (5) and (6).

$$k_{ac} = \frac{(l_{sc} - l_{sa})}{v} F_s \tag{5}$$

$$k_{bc} = \frac{(l_{sc} - l_{sb})}{v} F_s \tag{6}$$

However, we cannot know the location of sound source (θ_d) yet. Therefore, the following method is proposed to estimate the sound source location. Matrix r presents the cross correlation of R_{ac} and R_{bc} for all possible k_{ac} and k_{bc}. All values of matrix r are calculated by (7).

$$r(\theta) = R_{ac}[k_{ac}(\theta)] \cdot R_{bc}[k_{bc}(\theta)], \text{ where } 1^\circ \le \theta \le 360^\circ \tag{7}$$

Next, because we want to find the angle of sound's direction, we should first know the maximum value in the matrix r. After we fix threshold value in the r by using (8), we perform normalization to the r by using (9).

$$r_{th} = 0.99 \times \max\{r(\theta)\}, \text{ where } 1^\circ \le \theta \le 360^\circ \tag{8}$$

$$\begin{aligned} &r(\theta) = 0, && \text{if} \quad r(\theta) < r_{th}, \text{ where } 1^\circ \le \theta \le 360^\circ \\ &\frac{r(\theta) - r_{th}}{r_{\max} - r_{th}}, && \text{if} \quad r(\theta) > r_{th}, \text{ where } 1^\circ \le \theta \le 360^\circ \end{aligned} \tag{9}$$

And, if we perform a weighted average to the r by using (10), we will find the angle of sound's direction.

$$\frac{\sum_{\theta=1}^{360} \{r(\theta) \times \theta\}}{\sum_{\theta=1}^{360} r(\theta)} = \theta_{sd} \tag{10}$$

4.1. Voice activity detection (VAD)

For the purpose of effective interaction between human being and a robot, it is necessary to extract the period in which only voice signals are included: Non-voice or silent periods are unnecessary or harmful. Therefore, we propose a function of VAD (Voice Activity Detection) using autocorrelation method to find pitch information. IROBAA executes a reliable detection of sound's direction and speech recognition when the robot is decided to detect signals of voice by VAD method. Pitch information rather than energy is applied to the VAD since the former has the advantage of a robust feature against noises. Beside, In order to detect a pitch, we use an autocorrelation method, which is composed of simpler algorithm instead of FFT. The frequency of a vocal cord concerning human being exists in the range between 50 and 250Hz in case of a male and between 120 and 500Hz in case of a female. Therefore, if we put 800 samples per one frame into the autocorrelation equation, the executed signal will show pitch having periodic form of human vocal cord. The equation of the autocorrelation is expressed as (11).

$$R_{cc} = \frac{\sum_{g=0}^{\infty} \{C(t_g - k)C(t_g)\}}{\sqrt{\sum_{g=0}^{\infty} C(t_g - k)^2} \sqrt{\sum_{g=0}^{\infty} C(t_g)^2}} \tag{11}$$

Then, after we perform a median filter, which has excellent features in removing impulse noise, edge signal preservation and smoothing, we can calculate differential values about autocorrelation results. In case of real voice signals, since

magnitude of a periodic form and differential values between sampled signals is large, the peak values can be calculated by applying threshold value to differential values. Finally, as we can know the number of samples between two peak signals, the pitch can be detected by (12). To improve accuracy of VAD, we should also detect the second pitch in a frame.

$$Pitch = \frac{\text{Sampling Frequency}}{\text{A number of samples between the two peaks}} \quad (12)$$

Now, after making weighted sum of the calculated pitches of 10 frames, we can infer extracting the period of voice signal. Since the A/D converter, which is installed in IROBAA, has the function of double buffering, the robot can continuously execute the VAD algorithm at 0.05-second intervals without loss of raw data. Consequently, it can automatically and continuously perform finding direction of sound and speech recognition whenever speech commands enter to microphones.

6. EXPERIMENT AND RESULTS

6.1. Nonlinear Amplifier Board Experiment and result.

In order to verify our nonlinear amplifier's performance, we should perform experiments to compare with normal linear amplifier

First using nonlinear amplifier board at CR is 5:1 (nonlinear amplifier), we conduct experiment for localization and recognition, then in the same condition, we conduct the experiment again at CR is 1:1 (linear amplifier). Experiments took place rather ideally quiet living room. As shown in Fig. 10, a single male speaker locates in the corner of room and plays the same text, which is “go to the living room” for equal interval toward the robot. This experiment was performed several times, with distance of 0.5m, 1m, 2m, and 3m. L&H engine is commercial software and it accepts linear signal only for recognition in 11 kHz, then gives the recognition results and its confidence value, and the test room is rather ideal 130m^2 apartment living room.

Figure 11 and 12 shows experiment result of localization and recognition. In both figures, the x-axis is distance between sound and robot and the y-axis shows successful rate in percentage rate.

At fig. 11 the localization results at CR is 5:1 in solid line shows excellent till 3-meter range, whereas the success rate of recognition in dashed line performs poor after 2-meter range. Mainly, this is due to disturbance of signal by nonlinear amplification. However, since we are working on long-range localization and recognition over 2 meters, we have to develop the way to improve the recognition. Fig. 12 shows nonlinear amplification board SSM 2166 obtained from experiment as compression radio of 1:1 Experiment shows better results on sound recognition compare to 5:1, and especially around 1-meter, gives the best performance. However, in long range, especially over 2 meters, the localization drops its performance rapidly.

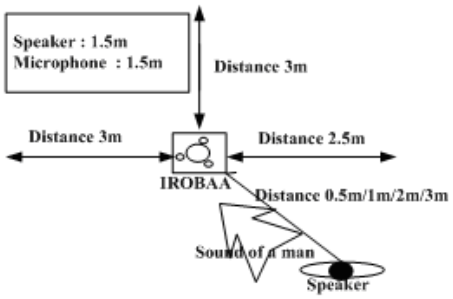

Fig.10 Developed nonlinear amplifier board

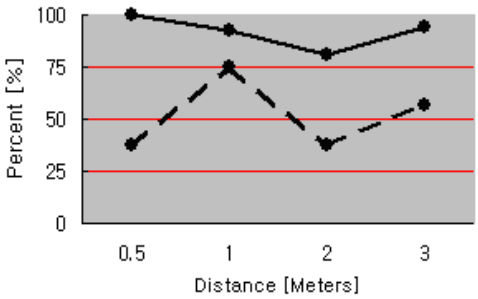

Fig.11 Experimental result at compression radio is 5:1

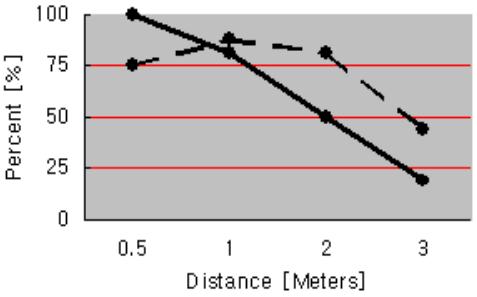

Fig.12 Experimental result of compression radio is 1:1

To improve recognition, using data sheet shown in Fig. 3 as reference, now we conduct curve fitting and interpolation to make the nonlinear signal as linear as possible. Noticing that the sound signal data is time discrete and the number of sampling data points are numerous, therefore to maintain the shape of original data is important during the curve fitting and interpolation.

6.2. Embedded system for sound Processing Experiment and result.

Figure 13 show whole experiment environment. Embedded system process sound raw data that receives speech signal from array microphone and so, using this sound raw data implement localization algorithm. Sound raw data comprise 800 hexa values (or float type) per one frame, accept total 32000 hexa values. To compare experiment result, we tested embedded system and PC in the same environment.

Test result, sound localization show similar values

embedded system or PC [1][2], but different performance. For example, embedded system reduce 10sec and PC reduce 0.1usec in VAD algorithm, embedded system reduce 19sec, but PC reduce .02usec in whole algorithms are included VAD and localization algorithm.

We can identify performance of embedded system poor through experiment result.

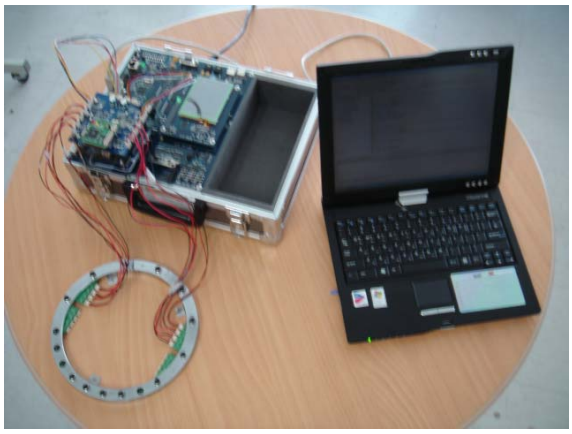

Fig.13 Embedded system for sound processing

7. CONCLUSIONS AND FUTURE WORKS

We conduct localization experiment using nonlinear amplifier board, RTMTB, embedded system is loaded linux OS, and also conduct both simulation and experiment on EPLD and embedded linux application program. And based on those results, we made conclusions as follows:

7.1. Conclusions

1) The nonlinear amplifier board shows satisfactorily performance for the programmable amplification.
2) Using nonlinear amplifier board, short distance was excellent about sound recognition and long distance was excellent about sound localization.
3) Through the simulation and real experiment results in RTMTB, we experienced high performance hardware implementation.
4) In experiment, we found whole performance poor in the embedded system, similar PC.
5) Our ultimate goal is to find suitable general algorithm that makes perfect hardware algorithm and optimized hardware platform for sound processing system, so that we can raise the performance, similar PC

7.2. Future Works

We achieved acceptable performance in hardware platform through the simulation as well as experiment. we will consider advanced hardware implementation for speech processing which utilizes DSP-based or FPGA-based algorithms for higher performances and more modularization.

REFERENCE

[1] Kim, H., Choi, J., Kim, M., and Lee, C., “Reliable Detection of Sound’s Direction for Human Robot Interaction” Proc. of 2004 IEEE/RJS Int. Conf. on Intelligent Robots and Systems, Sep. 28 ~ Oct. 2, 2004, Sendai, Japan.

[2] Kim, H., Choi, J., Kim, M., and Lee, C., “Sound’s Direction Detection and Speech Recognition System for Humanoid Active Audition” in Proc. Int. Conf. On Control, Automation, and Systems, Gyeongju, Korea, Oct. 2003, pp. 633 ~ 638.

[3] P. Aarabi and A. Mahdavi. The relation between speech segment selectivity and time-delay estimation accuracy. In Proceedings of ICASSP, May 2002

[4] P. Aarabi and S. Zaky. Robust sound localization using multi-source audiovisual information fusion. Information Fusion, 3:2:2009 ~ 223, September 2001

[5] M.S Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In Proc. of ICASSP, May 1997.

[6] SSM2166 Technical Manual”, Analog Device, 1999.

[7] Okuno, H., et al, “Assessment of General Applicability of Robot Audition System by Recognizing Three simultaneous Speeches”, Proc. of 2004 IEEE/RJS Int. Conf. on Intelligent Robots and Systems, Sep. 28 ~ Oct. 2, 2004, Sendai, Japan.