2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

DSP Integration of Sound Source Localization and Multi-channel Wiener Filter

Byoung-gi Lee, Hyun-dong Kim, Jong-suk Choi, Seyun Kim, and Nam Ik Cho

Abstract—**This paper describes a DSP integration of sound source localization (SSL) and multi-channel Wiener filter (MWF). To develop a robot audition system, we integrated SSL module and MWF module into a DSP system. SSL is a module to perceive the direction of a human user's call. It measures time delay of arrival among microphones and estimates the direction of sound source. Also, it post-processes the resulted estimations of direction by histogram to perceive the direction robustly under noisy environment. MWF is a module to reduce background noises from raw voice signal to enhance the performance of robot's speech recognition. It gathers information of background noises during noise-period and then reduces noises during voice-period. This SSL-MWF combination system will be a cheap, high-performing and convenient solution for robot audition.**

I. Introduction

ROBOT audition systems are commonly composed of microphone array, sound amplifier, A/D converter, and computer for various sound signal processing. But this PC-based system is not convenient and efficient in the robotics. A heavy and power-consuming system deteriorates the portability and battery duration of robots. Instead of the PC-based system, we designed a set of small and optimized boards for a robot audition system, which consists of Sound Localization Processing board (SLP) and Nonlinear Amplifying Board (NAB). It is convenient to attach to various robot systems because it is a completely independent audition module. As indicated by their names, the SLP is a DSP board for signal process performing sound source localization (SSL), and the NAB is a data acquisition board equipped with nonlinear amplifier. Fig. 1 shows the whole SLP & NAB system using a 3-channel microphone-array. SLP has a high performance DSP chip and convenient developmental environment serviced by Analog Device Company. NAB has synchronized A/D converters and gain-variable amplifier. This optimized hardware system can reduce the burden on software, and minimize computational overhead and delay by simultaneously receiving real-time speech signals from multi-channel microphone array.

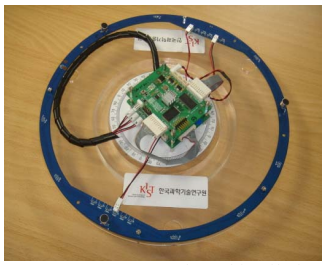


Fig. 1. SLP & NAB system. SLP board is piled up above NAB board. Triangular microphone array is connected to NAB board.

We have developed SSL algorithm on a set of SLP & NAB [1-3]. SSL is an algorithm estimating the direction of sound source, mainly speaking person, using the time delay of arrival (TDOA) among microphones. However, TDOA is vulnerable to background noise, interfering sound, and reverberant sound. There exist numerous approaches to establish a robust SSL algorithm. To relieve this problem, Kalman filters [4], Bayesian networks [5], Sequential Monte Carlo method [6], and particle filters [7] have been proposed. We introduced a reliable detection method by transforming cross-correlation into a spatial function [1]. A post-process using histogram is applied to our method to achieve robust localization results.

In this paper, we add a speech enhancing process, multi-channel Wiener filter (MWF) to our DSP system. Noise reduction is very important in the case of distant or hands-free speech acquisition system such as robot audition. Hence, there have been many researches on noise suppression for several decades. Among them, beam-former and Wiener filter are regarded as typical noise reduction techniques. Multi-channel Wiener filter has been shown to provide better performance than the standard beam-former [8]-[10]. More recently, GSVD (Generalized Singular Value Decomposition)-based subspace approach is developed for multi-channel Wiener filter [9]. The underlying principle of subspace approach is based on the low-rank subspace model of speech signal. The vector space of noisy input signal is decomposed into signal and noise subspace and the noise reduction is achieved by removing the noise subspace. In this paper, we present an embedded multi-channel Wiener filter (e-MWF) which is also based on the subspace decomposition using generalized eigen-value problem in the frequency domain.

By integration of SSL and e-MWF, we could achieve a small and efficient robot audition system which can notify the robot's main PC of the direction of detected sound source,

Manuscript received September 15, 2009. This work was supported by the Korea Ministry of Knowledge Economy under Grant of the 21st century frontier project.

Byoung-gi Lee is with Korea Institute of Science and Technology, Seoul, Republic of Korea (e-mail:leebg03@ kist.re.kr).

Hyun-dong Kim is with Korea Institute of Science and Technology, Seoul, Republic of Korea (e-mail:hdkim@ kist.re.kr).

Jong-suk Choi is with Korea Institute of Science and Technology, Seoul, Republic of Korea (phone: +82-2-958-5618; fax: +82-2-958-5629; e-mail: cjs@kist.re.kr).

Seyun Kim is with Department of Electrical Engineering, Seoul National University, Seoul, Republic of Korea (e-mail: light4u@ispl.snu.ac.kr).

Nam Ik Cho is with Department of Electrical Engineering, Seoul National University, Seoul, Republic of Korea (e-mail: nicho@snu.ac.kr).

and simultaneously provide enhanced speech signal for robot's speech recognition engine. In the rest of this paper, we explain more details about our DSP hardware system, SSL algorithm and e-MWF algorithm, and present experimental results of our robot audition system.

II. Hardware Architecture

Our hardware is composed of two parts. One is NAB (Nonlinear Amplifying Board) that amplifies sound signals and digitizes amplified sound signals. The other is SLP (Sound Localization Processing board) that processes digitized sound signals. Fig. 2 shows the system architecture.

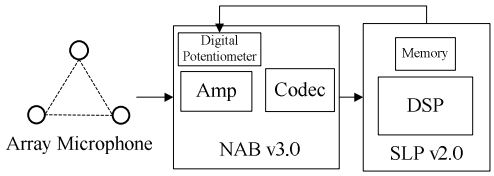


Fig. 2. System Architecture

A. Microphone Arrangement

We have pursued small and efficient robot audition system. We use an array of three microphones because a 3-microphone-array is the smallest structure which doesn't make an ambiguity in determining the sound source direction. When 2-microphones is used, there are two cases having the same TDOA (Time delay of arrival) which make front-back confusion. We placed microphones L, C and R at equi-spaced points as shown Fig. 3. Two of them make an angle of 120 degrees. The symmetric positions of microphones make localization performance less partial over all directions.

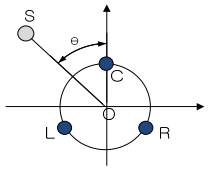


Fig. 3. Microphone arrangement

B. NAB (Nonlinear Amplifying Board)

NAB is composed of two parts, that is, nonlinear amplifier and codec as shown in Fig. 4. It can support 4-channel signal.

1) *Amplifier:* NAB has a non-linear amplifier, SSM2166 chip. We can change the compression ratio to capture low-dB voice or avoid saturation of loud voice. But in this paper, we set the compression ratio as linear amplification because of speech enhancement process, MWF. The SSM2166 has two resistor values for compression ratio and amplification ratio. We can change them in a run-time using digital potentiometer which is controlled by SLP.

2) *Codec:* Amplified analog signals pass through a codec, AD1836. The codec converts analog signal into 16 kHz digital signal. After conversion, it sends the outputs to the SLP module through TDM port.

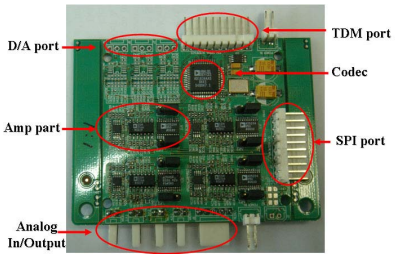


Fig. 4. NAB (Nonlinear Amplifying Board)

C. SLP (Sound Localization Processing board)

We have developed SLP module using Blackfin processor (BF533) of Analog Device Company. The overview of SLP is given by Fig. 5. It has ADSP-BF533 for main processor, flash memory for local memory, SPI port for auto gain control of NAB, UART port for PC interface, and TDM port for codec (AD1836). SLP has features of low power, high performance, adaptively embedded system of small-size and convenient developmental environment serviced by Analog Device Company. It is a suitable system for implementing not only SSL algorithm but also any other process on robot audition.

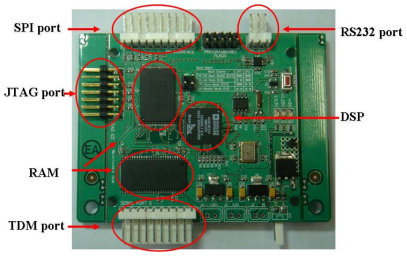


Fig. 5. SLP (Sound Localization Processing board)

III. Sound Source Localization

Background noises, many interfering noises from radio, TV, car, etc., and reverberant sounds cause critical errors in measuring TDOA (Time Delay of Arrival) among microphones. To localize a sound source robustly, we use a reliable detection method by transforming cross-correlation into a spatial function and histogram method.

To measure TDOA of a sound source, we calculate the cross-correlation between two microphones. It is a typical way of TDOA measurement to search the maximum point of the cross-correlation. But we use not only the maximum point but also all the other points of the cross-correlation to set up our spatial function which scores the count of existence value at every direction.

Let $x_L[n]$, $x_C[n]$ and $x_R[n]$ be signals from microphones L, C, and R, respectively. Cross-correlations are given by (1).

$$r_{LC}[k]=cc(L,C),\quad r_{CR}[k]=cc(C,R)\quad and\quad r_{RL}[k]=cc(R,L)$$
$$,where\quad cc(A,B)=\frac{\sum_n x_A[n]\cdot x_B[n-k]}{\sqrt{\sum_n x_A^2[n]\sum_n x_B^2[n]}} \tag{1}$$
$$and\quad -f_s\cdot\frac{dist(A,B)}{v_{sound}}\le k\le f_s\cdot\frac{dist(A,B)}{v_{sound}}$$

We need to calculate the (1) only over the possible index *k*'s which are determined by the distance between microphones, the sampling frequency f_s, and sound speed v_{sound} in the air.

Next, we are going to transform $r_{LC}[k]$, $r_{CR}[k]$, and $r_{RL}[k]$ into functions of θ. So, we need a mapping from *k* into θ such as (2).

$$k_{AB}(\theta)=f_s\cdot\frac{dist(S,A)-dist(S,B)}{v_{sound}} \tag{2}$$

,where *S* is the sound source placed at the azimuth angle of θ. Using (2), we define a spatial function which indicates where a sound source exists "strongly" such as (3).

$$R(\theta)=R_L(\theta)+R_C(\theta)+R_R(\theta)$$
$$,where\quad \begin{cases} R_L(\theta)=r_{LC}[k_{LC}(\theta)]\cdot r_{RL}[k_{RL}(\theta)] \\ R_C(\theta)=r_{LC}[k_{LC}(\theta)]\cdot r_{CR}[k_{RL}(\theta)] \\ R_R(\theta)=r_{CR}[k_{LC}(\theta)]\cdot r_{RL}[k_{RL}(\theta)] \end{cases} \tag{3}$$

To get a more reliable spatial function, we modify (3) into (4) by a threshold R_{th}.

$$\tilde{R}(\theta)=\begin{cases}\frac{R(\theta)-R_{th}}{R_{\max}-R_{th}} & ,if\quad R(\theta)>R_{th}\\ 0 & ,otherwise\end{cases} \tag{4}$$
$$,where\quad R_{\max}=\max_\theta R(\theta)\quad and\quad R_{th}=0.9R_{\max}$$

Finally, we can estimate the direction of sound source by calculating a centroid of $\tilde{R}(\theta)$ as follows.

$$\hat{\theta}=\frac{\sum_\theta \tilde{R}(\theta)\cdot\theta}{\sum_\theta \tilde{R}(\theta)} \tag{5}$$

Our system processes 3-channel signal by 320-sample frame and the sampling frequency is 16 kHz, thus one frame is just 20ms long. It is too short to achieve a good localization result. Post-processing is required to get a better localization. For the post-processing, we first gather the estimated angles from each frame within a voice segment. In our old system, we took the average of the gathered angles. Now, we changed the post-processing from the average into a histogram method – building the histogram of the gathered angles and then taking the mode of the histogram. This new post-processing is more effective than the old one. It will be shown in the part of simulation and experiment.

IV. Multi-channel Wiener Filter

A. Multi-channel Wiener Filter

To get a clean speech signal from distant microphone-array, we used multi-channel Wiener filter (MWF) as a noise reduction method. MWF is a useful method for reducing stationary noise from noisy signal.

For convenience, we change the notation of L, C, and R microphones into 1st, 2nd, and 3rd microphones. A three-channel signal can be expressed such as (6).

$$y_i(t)=h_i(t)*s(t)+n_i(t)=x_i(t)+n_i(t),\quad i=1,2,3 \tag{6}$$

, where $y_i(t)$ denotes the observed signal at the i^{th} microphone at time *t*, $x_i(t)$ and $n_i(t)$ are speech and additive stationary noise components, respectively, $s(t)$ is the desired speech source and $h_i(t)$ is the acoustic transfer function from the speech source to the i^{th} microphone. The frequency domain vector representation of (6) is

$$\mathbf{Y}(f)=\begin{bmatrix}Y_1(f)\\Y_2(f)\\Y_3(f)\end{bmatrix}=S(f)\begin{bmatrix}H_1(f)\\H_2(f)\\H_3(f)\end{bmatrix}+\begin{bmatrix}N_1(f)\\N_2(f)\\N_3(f)\end{bmatrix}$$
$$=S(f)\mathbf{H}(f)+\mathbf{N}(f)=\mathbf{X}(f)+\mathbf{N}(f) \tag{7}$$

, where $Y_i(f)$, $H_i(f)$, $S(f)$, $N_i(f)$, and $X_i(f)$ are the Fourier transforms of $y_i(t)$, $h_i(t)$, $s(t)$, $n_i(t)$, and $x_i(t)$ respectively.

The MWF (Multi-channel Wiener Filter) is optimal in the sense of minimum mean square error (MMSE) between the estimated signal and the desired source. If we estimate the speech component at the first microphone and assume that the speech and noise signals are uncorrelated, the MWF in the frequency domain is expressed as follows.

$$\mathbf{W}(f)=\mathbf{R}_{\mathbf{YY}}^{-1}(f)\mathbf{R}_{\mathbf{XX}}(f)\mathbf{e}_1$$
$$=\mathbf{R}_{\mathbf{YY}}^{-1}(f)[\mathbf{R}_{\mathbf{YY}}(f)-\mathbf{R}_{\mathbf{NN}}(f)]\mathbf{e}_1 \tag{8}$$
$$,where\quad \begin{cases}\mathbf{e}_1=[1\quad 0\quad 0]^{\mathrm{T}}\\ \mathbf{R}_{\mathbf{YY}}(f)=E\{\mathbf{Y}(f)\mathbf{Y}^H(f)\}\\ \mathbf{R}_{\mathbf{XX}}(f)=E\{\mathbf{X}(f)\mathbf{X}^H(f)\}\\ \mathbf{R}_{\mathbf{NN}}(f)=E\{\mathbf{N}(f)\mathbf{N}^H(f)\}\end{cases}$$

B. Subspace-based MWF

The rank of the narrow-band spatial covariance matrix (NSCM) of $\mathbf{X}(f)$ is theoretically equal to 1 because of (9).

$$\mathbf{R}_{\mathbf{XX}}(f)=E\{\mathbf{X}(f)\mathbf{X}^H(f)\}$$
$$=E\{S(f)S^*(f)\}\mathbf{H}^H(f)\mathbf{H}(f) \tag{9}$$

Hence, we are going to make the estimated NSCM of $\mathbf{X}(f)$, i.e., $\mathbf{R}_{\mathbf{YY}}(f)-\mathbf{R}_{\mathbf{NN}}(f)$ to be a rank-1 matrix. If we assume that $\mathbf{R}_{\mathbf{NN}}(f)$ is not rank-deficient, $\mathbf{R}_{\mathbf{YY}}(f)$ and $\mathbf{R}_{\mathbf{NN}}(f)$ can be jointly diagonalized by solving the generalized eigenvalue problem such as

$$\begin{aligned}&\mathbf{R}_{\mathbf{YY}}(f)\mathbf{Q}(f)=\mathbf{R}_{\mathbf{NN}}(f)\mathbf{Q}(f)\mathbf{\Lambda}(f)\\&\begin{cases}\mathbf{Q}^{H}(f)\mathbf{R}_{\mathbf{YY}}(f)\mathbf{Q}(f)=\mathbf{\Lambda}_{\mathbf{Y}}(f)\\\mathbf{Q}^{H}(f)\mathbf{R}_{\mathbf{NN}}(f)\mathbf{Q}(f)=\mathbf{\Lambda}_{\mathbf{N}}(f)\end{cases}\end{aligned}\tag{10}$$

, where $\mathbf{Q}(f)$ is an invertible, but not necessarily orthogonal matrix and, $\mathbf{\Lambda}(f)$, $\mathbf{\Lambda}_{\mathbf{Y}}(f)$, and $\mathbf{\Lambda}_{\mathbf{N}}(f)$ are diagonal matrices. Using (10), we can obtain the subspace-based MWF equation of (11) instead of the original MWF equation of (8).

$$\mathbf{W}(f)=\mathbf{Q}(f)[\mathbf{I}-\mathbf{\Lambda}_{\mathbf{Y}}^{-1}(f)\mathbf{\Lambda}_{\mathbf{N}}(f)]\mathbf{Q}^{-1}(f)\mathbf{e}_1\tag{11}$$

From (11), we just need to solve (10) to get MWF's filter coefficients $\mathbf{W}(f)$. $\mathbf{R}_{\mathbf{YY}}(f)$ and $\mathbf{R}_{\mathbf{NN}}(f)$ can be easily measured if a voice activity detector (VAD) is equipped. Within background noise segment, $\mathbf{R}_{\mathbf{NN}}(f)$ is measured, while within voice segment, $\mathbf{R}_{\mathbf{YY}}(f)$ is measured. After measuring $\mathbf{R}_{\mathbf{YY}}(f)$ and $\mathbf{R}_{\mathbf{NN}}(f)$, we can solve (10) by a conventional iterative method.

V. Embedded Robot Audition

We have developed an embedded robot audition system based on DSP system. Once, SSL (Sound Source Localization) is implemented [3], now MWF (Multi-channel Wiener Filter) for noise reduction is added to the same system. Hence, we are going to focus more on the implementation of e-MWF (embedded MWF) than embedded SSL.

A. Overview of our Robot Audition System

Our integrated system is shown in Fig. 6. The integrated system can notify the robot's main system of the angle of detected sound source, and provide enhanced speech signal for the robot's speech recognition engine at the same time. As mentioned before, its sampling rate is 16kHz and its frame size is 320 samples (frame time is 20ms). Frame sliding size is 240 samples and successive frames overlap each other by 80 samples. In this frame strategy, our DSP system should process one frame in less than frame sliding time (15ms).

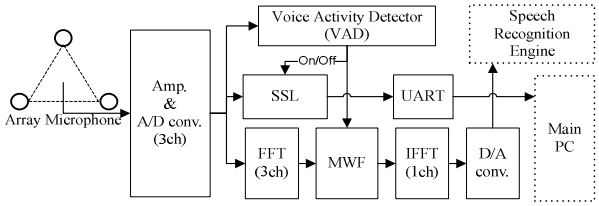


Fig. 6. Integrated System

B. Two problems of DSP

There are two problems about DSP implementation of robot audition system in Fig. 6. One is an inaccurate FFT function of DSP and the other is lack of float point unit (FPU) in DSP. Inaccuracy of FFT results from the lack of FPU. DSP library provides only FFT functions using 16-bit fixed point operation. These FFT functions do one-bit right shift after every multiplication of FFT algorithm to avoid overflow. That means loss of lower bits proportional to the logarithm of frame size because FFT algorithm has $\log_2 N$ multiplications if frame size is *N*. For this reason, its output is very poor when N is larger than 64 as shown Fig. 7.

It doesn't matter for SSL module because our SSL doesn't need FFT and can be performed enough with fixed point operations. However, these problems are very severe for MWF module because it is operated in frequency domain and its iterative method is too delicate to be operated by fixed point. As iteration goes on, the error caused by fixed point operation grows and iteration diverges.

C. Modifying FFT of DSP

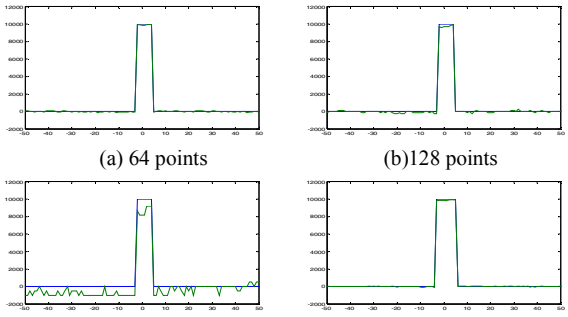
There are two choices for FFT function problem. One is to utilize the inaccurate FFT function of DSP library. The other is to make a more accurate user function of FFT. The latter seemed more reasonable and we could make a good FFT function using 32-bit fixed point operation. But it has another problem, that is, time. It was much slower than FFT of DSP library. It took almost 2ms to FFT 512-sample signal. (The frame size is 320 samples but generally FFT requires input signal length to be a power of 2. 512 is the smallest larger than 320.) Our system has 3 channels and we need three times of FFT. It is not acceptable to spend 6ms of 15ms for just FFT.

Hence, we went back to the FFT of DSP library. Anyway, it is fast and its inaccuracy is tolerable if the signal length is less than 64. We tried to make a desired FFT function based on 64-point FFT of DSP library. It could be realized by (12).

$$\begin{aligned}X_k&=\sum_{n=0}^{5N-1}x_n e^{-i\frac{2\pi kn}{5N}}=\sum_{n=0}^{N-1}\sum_{m=0}^{4}x_{5n+m}e^{-i\frac{2\pi k(5n+m)}{5N}}\\&=\sum_{m=0}^{4}\sum_{n=0}^{N-1}x_{5n+m}e^{-i\frac{2\pi kn}{N}}e^{-i\frac{2\pi km}{5N}}=\sum_{m=0}^{4}X_k^{(m)}e^{-i\frac{2\pi km}{5N}}\\&,where\quad N=64\quad and\quad X_k^{(m)}=\sum_{n=0}^{N-1}x_{5n+m}e^{-i\frac{2\pi kn}{N}}\end{aligned}\tag{12}$$

We made a 320-point FFT function by combining five 64-point FFTs and it took 1ms to do 320-point FFT for all 3

channels. Its FFT example is shown in Fig. 7-(d).



(a) 64 points (b)128 points

(c)512 points (d)320 points

Fig. 7. FFT-IFFT results of rectangular pulse: (a)-(c) use the original FFT function, and (d) uses a combination of five 64pt FFT functions.

D. *Filter sharing of neighboring frequencies.*

For the second problem of DSP, there is a bad solution, to use float point operations emulated by fixed point operations. It is provided by DSP compiler implicitly but it is very slow because of lack of float point unit (FPU). It is why we express it a bad solution. However, MWF is so complicated that float point operation is indispensable.

Now, the problem is changed into how to meet the limit of processing time, 15 ms. Fortunately, the measured processing time of e-MWF (embedded MWF) is 11.44 ms per frame. Since the processing time of e-SSL (embedded SSL) is just 2 ms, it is no problem to meet the time limit. However, we have a plan beyond the system shown in Fig. 6. Actually, our next goal is to integrate one more module, a simple speech recognizer, on DSP. Therefore, we need a way to reduce the amount of MWF computation. We propose a filter sharing method to reduce the number of frequency bin in which the MWF filter coefficients, $\mathbf{W}(f)$ are calculated. Since the neighboring frequencies are very similar to each other, it is reasonable and effective to share neighboring filters. When choosing frequencies of which filter is just copied from neighboring frequency, mel-scaled frequency was regarded since our robot speech recognizer uses MFCC (Mel-frequency Cepstral Coefficient). Hz-frequency is transformed to mel-frequency by (13).

$$m = 1127.01048 \ln\left(1 + f / 700\right) \tag{13}$$

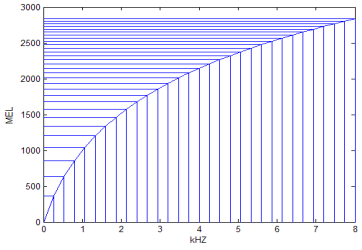


Fig. 8. Hz-frequency vs. Mel-frequency

As shown Fig. 8, equi-spaced Hz-frequencies are mapped unevenly spaced mel-frequencies. Higher mel-frequencies are denser than lower mel-frequencies. Therefore, more sharing filters in higher frequencies would minimize the loss of MFCC features. Since the frame size is 320 samples, there are 161 frequency bins. (we need to process just half of frequencies because of its symmetry.) When sharing filters of 40 frequencies (25% sharing), the processing time of MWF is reduced to 8.98 ms per frame.

VI. Simulation and Experiment

A. *Performance of e-SSL*

We tested our embedded SSL (e-SSL) system in noisy environment. In a demo room, we prepared a TV and an air-conditioner which were noise sources. We made various noise conditions by changing the volume of TV or turning on the air-conditioner. In each case, the utterance of "Hello, H-robot, come here" spoken by an audio system is kept louder than the noise signal by 15dB. This is from an assumption that most users might call their robot loudly enough if they are in noisy environment. Success rates are measured with 15 deg.-error and 30 deg.-error. The error tolerance is determined regarding the robot's field of view in its vision sensor. As long as SSL error is less than 30 deg., robot is likely to find out its user using the vision system. The result is given by Table I. Two different post-processing methods are compared. The one is our old method which is to take the average of gathered angles and the other is new method which is to take the mode value of gathered angles using histogram.

The result shows the histogram method is much better than the average method. The histogram method didn't fail, while the average method was very subjective to TV noise. It seems that the average method compounds desired angles and undesired angles and concludes an irrelevant angle, while the histogram method preserves desired or undesired angle values and concludes the predominant angle which might come from the real source.

TABLE I
Experiment : Success Rate of SSL

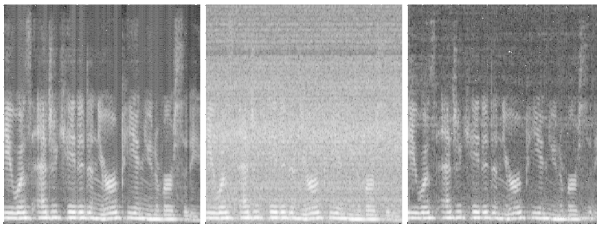
Condition	Average method	Histogram method
Quiet 47dB	100% (100%)	100% (100%)
TV 53dB	27.3% (61.4%)	100% (100%)
TV 57dB	14.0% (41.9%)	100% (100%)
TV 60dB	7.0% (41.9%)	100% (100%)
Air-conditioner 54dB	100% (100%)	100% (100%)
Total	50% (69.3%)	100% (100%)

Each case has 43~44 utterances of "Hello, H-robot, come here." and total 218 utterances are tested. The utterances are spoken 15dB louder than noise at the azimuth angles of 0, 90, 180 or 270 deg. in the 3m distance. SSL is regarded successful if the error of SSL is less than or equal to 15 deg. or 30 deg. The success rate of 15 deg. error is written without parentheses, while that of 30 deg. error is written within parentheses.

B. *Performance of e-MWF*

We had a simulation to measure the performance of our embedded MWF (e-MWF). Its performance is measured by noise reduction ratio. Since e-MWF was added to our robot

audition system to enhance the performance of robot's speech recognition which uses MFCC features, we measured how much MFCC distortion was reduced by MWF filters. To measure MFCC distortion, we need the original clean signals. However, in real situation, it is difficult to obtain the clean signal and to measure the performance of e-MWF objectively. It is the reason we had a simulation instead of a real experiment. The result is given by Table II. We compared the performances of PC version of MWF, e-MWF and filter-sharing e-MWF. In Table II, the performance of PC MWF is the best as expected. But the performance of e-MWF is very close to that of PC MWF. The difference is just 1.14% points. Also, the filter-sharing e-MWF is close to that of PC MWF and the difference is 4.13% points. This result shows that filter-sharing is an effective way to reduce the processing time of MWF. We could save 2.46ms per frame at the cost of 2.99% points of noise reduction ratio. We provide an example of e-MWF result in Fig. 9. It shows visually how effective e-MWF is. The spectrogram of source signal has a clear sound pattern and it could be almost restored by e-MWF.



(a)Source signal (b)noisy signal (c) e-MWF output

Fig. 9. Spectrograms of sentence 3 (SNR 10dB)

VII. CONCLUSION

We integrated SSL (Sound Source Localization) and MWF (Multi-channel Wiener Filter) into an embedded system. Not only it could operate in real-time, but also it showed high performance. The e-SSL module using histogram post-processing could cope with the difficulties in the noisy environment and the e-MWF module suppressed the noise level effectively in the noisy speech signal and was expected to enhance the performance of robot's speech recognition engine. Our embedded robot audition is suitable for various portable robot systems because of its compactness and efficiency. Our next plan is to add a simple word recognition module into this embedded system.

VIII. ACKNOWLEDGMENTS

This work was conducted within the 21st century frontier project (http://irobotics.re.kr/) and funded by the Ministry of Knowledge Economy of Korea.

TABLE II
SIMULATION : NOISE REDUCTION RATIO OF MWF

SNR	Sentence	MWF (PC)	e-MWF	e-MWF (25% sharing)
20dB	1	55.73 %	38.75 %	37.15 %
	2	29.58 %	29.68 %	28.96 %
	3	47.36 %	41.68 %	39.90 %
	4	44.70 %	38.76 %	36.88 %
	5	28.49 %	25.17 %	25.27 %
10dB	1	51.98 %	47.11 %	44.80 %
	2	44.68 %	48.08 %	44.63 %
	3	54.01 %	53.61 %	48.72 %
	4	42.62 %	43.94 %	39.41 %
	5	44.90 %	42.78 %	40.36 %
0dB	1	37.58 %	38.69 %	36.69 %
	2	36.35 %	37.04 %	33.79 %
	3	41.41 %	41.82 %	38.18 %
	4	29.35 %	31.64 %	27.32 %
	5	37.45 %	39.06 %	35.32 %
Total		41.47 %	40.33 %	37.34 %

By simulation, 3-channel signals are generated from 5 mono-signals and white Gaussian noise is added to them in various SNR conditions.

The performance of MWF is measured by reduction ratio of MFCC feature distortion. When s, x, and z are the MFCC of source signal, noisy signal, and MWF output, respectively, the noise reduction ratio is given by $\frac{dist(s,x)-dist(s,z)}{dist(s,x)}*100(\%)$.

REFERENCES

[1] Hyun-Don Kim, Jong-Suk Choi, and Munsang Kim, "Reliable Detection of Sound's Direction for Human Robot Interaction," *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, Sendai, Japan, pp.2411-2416.

[2] Yoon Seob Lim, Jong Suk Choi, and Mun-Sang Kim, "Probabilistic sound localization," *International Conference on Control, Automation and Systems* 2007, Oct. 2007, pp.1925-1928.

[3] Seung Seob Yeom, Jong Suk Choi, Yoon Seob Lim, and Mignon Park, "DSP Implementation of Probabilistic Sound Source Localization," *IEEE Workshop on Signal Processing Systems*, Oct. 8-10, 2008, Washington, D.C. Metro Area, U.S.A., pp. 204-209.

[4] D. E. Sturim, M. S. Brandstein, and H. F. Silverman, "Tracking multiple talkers using microphone-array measurements," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[5] V. Pavlovic, A. Garg, and J. Rehg, "Multimodal Speaker detection using error feedback dynamic Bayesian networks," *Proceedings of the IEEE CVPR*, Hilton Head Island, SC, 2000.

[6] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," *Proceedings of IEEE ICCV*, Vancouver, July 2001.

[7] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, 2003, pp. 826-836.

[8] D. Florencio and H. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays", *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, May, 2001, Salt Lake City UT, USA, pp. 197-200.

[9] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process*, vol. 50 (9), 2002, pp. 2230-2244.

[10] G. Rombouts and M. Moonen, "QRD-based unconstrained optimal filtering for acoustic noise reduction", *Signal Process*, vol. 83 (9), 2003, pp. 1889-1904.

[11] S. Kurotaki et al., "Implementation of Active Direction-Pass Filter on Dynamically Reconfigurable Processor", *Proc. IROS 2005*, pp.3175-3180.

[12] Bonnal et al., "Speaker Localization and Speech Extraction with the EAR sensor", *Proc. IROS 2009*, pp.670-675.