國立交通大學

機械工程學系

碩士論文

具聽音辨位和語音增強及辨識的唱歌機器人

Source Localization, Speech Enhancement and Recognition of a Singing Robot

研 究 生：桂振益

指導教授：白明憲

中華民國九十九年六月

具聽音辨位和語音增強及辨識的唱歌機器人

Source Localization, Speech Enhancement and Recognition of a Singing Robot

研　究　生：桂振益　　Student：Chen-Yi Kuei

指導教授：白明憲　　Advisor：Ming-Sian Bai

國立交通大學

機械工程學系

碩士論文

A thesis
Submitted to Department of Mechanical Engineering
Collage of Engineering
National Chiao Tung University
in Partial Fulfillment of Requirements
for the Degree of
Master
in

Mechanical Engineering

June 2010

HsinChu, Taiwan, Republic of China

中華民國九十九年六月

具聽音辨位和語音增強及辨識的唱歌機器人

研究生：桂振益　　　　　　　　　　　　指導教授：白明憲 教授

國立交通大學 機械工程學系 碩士班

摘要

現今的機器人工業如雨後春筍般蓬勃發展，技術更是日新月異，各種功能的機器人舉凡保全機器人、軍事機器人、居家看護機器人、娛樂機器人等琳瑯滿目，而隨著社會水準以及人們對於生活品質要求的提高，娛樂機器人今日佔有相當重要的地位。本論文提出了一種點唱機器人，會追蹤且同時轉到使用者的方向，所以此點唱機器人必須具備聽音辨位及語音辨識的能力。其中聽音辨位的方法包括以物體轉移函數為基礎的辨位方法和交互相關及廣義交互相關；語音辨識則在萃取出特徵參數之後採用動態時軸校正的方法比對並且辨識。而為了要讓使用者命令的聲音純化以提高辨識率，我們採用語音增強的技術，包括陣列訊號處理和以相位差為基礎的語音增強方法。以上提及的演算法我們將擇其優者整合在樂高NXT機器人，而其操作平台為以windows為介面的資料擷取系統。

Source Localization, Speech Enhancement and Recognition of a Singing Robot

Student: Chen-Yi Kuei Advisor: Ming-Sian Bai

Department of Mechanical Engineering

National Chiao-Tung University

ABSTRACT

Nowadays, there are a variety of functional robots included security robot, military robot, household robot and recreational robot, etc. With social progress and the attention of quality life, entertainment undertakings play an important role recently. In this thesis, we present a nickelodeon robot with simultaneous human-tracking. Therefore, the robot contains source localization and speech recognition techniques. The methods of source localization include object-related transfer function (ORTF) based, cross correlation (CC) and generalized cross correlation (CC) method. We recognize words by employing dynamic time warping (DTW) to do dynamic matching after feature extraction. For the purpose of increasing the purity of the command voice, we adopt speech enhancement which contains array signal processing and phase difference (PD) method. All algorithms we take the best one of each purpose to implement on the LEGO NXT robot controlled by windows-based NI DAQ system.

誌謝

時光飛逝，兩年碩士班研究生涯轉眼就過去了，首先感謝指導教授白明憲博士的指導與教誨，使我順利完生學業與論文，在此致上最誠摯的謝意。而教授指導學生時豐富的知識，嚴謹的治學態度以及追求學問的熱忱亦是學生的學習效法的典範。

在論文寫作方面，感謝電機系冀泰石教授和電子系桑梓賢教授在百忙之中撥冗並提出相當寶貴的意見，使得本論文的內容更趨於完善與充實，在此致上無限的感激。

回顧這兩年的歲月，承蒙同實驗室的博士班林家鴻學長、李雨容學姐、陳勁誠學長、劉志傑學長與在職博士班蔡耀坤學長、曾瑞宏學長及碩士班何克男學長、艾學安學長、郭育志學長、王俊仁學長及劉冠良學長在研究與學業上的適時指點，並有幸與同學廖國志、曾智文、張濬閣、陳俊宏、劉嬡婷、廖士涵在研究上互相討論，也感謝這兩年你們帶來的砥礪與笑聲。而與學弟徐偉智、王俊凱、吳俊慶、衛帝安、許書豪、馬瑞彬的朝夕相處，亦是值得回憶。

最後僅將此論文獻給我親愛的家人。感謝我的父親桂建中先生總是無條件支持我的決定，感謝母親馮志屏女士從小對我無微不至的呵護與諄諄教誨，感謝乾媽馮志琴女士這麼多年來不求回報的付出，感謝姐姐桂萱蓉和弟弟桂睿廷平時的打氣加油；感謝女友林吟盈總是陪在我身邊，聽我得意時的自吹自擂和在我失落的時候給予安慰與鼓勵。要感謝的人實在太多不及備載，如有疏漏，在此也一併致上最深的謝意。

TABLE OF CONTENTS

Table List

FIGURE LIST

I. INTRODUCTION

Robot industry has developed and changed with each passing day. There are security robot [1], military robot [2], household robot [3] and entertainment robot [4], etc. However, Robot with function of karaoke, like a nickelodeon, is rarely seen. Now we present a robot which can turn to user and sing the song what user ask for. So, this nickelodeon robot has to localize the sound made by user and then recognize commands by speech recognition system. Because the environment can be noisy (i.e. the signal to noise ratio (SNR) is low), we combine with speech enhancement to purify command voice. Therefore, it comes three sub-topics: source localization, speech enhancement and speech recognition. Among these three sub-topics, source localization and speech enhancement are based on microphone array technology [5] [6]. We introduce these three sub-topics as follows:

Microphone array have received increasing attention in past few years, especially in spatial filtering (beamforming) [7], and source localization [8] [9]. Microphone array techniques depend on many factors, including placement, geometrical configuration, number of microphones, as well as the conditions and the number of active acoustic sources in the environment under investigation. Acoustic localization is an important task in many practical applications such as videoconferencing [10], hands-free communication system [11], hearing aids [12] and human-machine interaction [13]. Different kinds of source localization methods were proposed in the literature [5] [24]-[27]. In this thesis, we employ object-related impulse response (ORIR) based method, cross correlation (CC) and generalized cross correlation (GCC) [14] method. ORIR-based method is motivated by the algorithm based on head-related transfer function (HRTF) proposed by McDonald [15].

Speech enhancement also exploits microphone array technology. The advantages of using an array to enhance the desired signal reception while simultaneously suppressing the undesired noise can be easily illustrated by a delay-and-sum (DAS) beamformer [5]. To achieve better performance, we optimize the beampattern of microphone array [5]. On the other hand, our work in speech enhancement is to increase the words recognition rate (RR). It is well known that the human binaural system is remarkable in its ability to separate sound sources even in a very difficult environment. Motivated by these observations, many models and algorithms have been developed using interaural time difference (ITD), interaural intensity difference (IID), interaural phase difference (IPD), and other cues [16] [17] [18]. IPD and ITD have been extensively used in binaural processing because this information can be easily obtained by spectral analysis [16]. In this thesis, we only focus on the ITD cue and we construct a binary model to mask the undesired sound and extract the purpose speech without distortion.

The last part of this thesis is speech recognition. Currently, there are largely two types for recognizers – DTW-based and HMM-based method. DTW is a technique of dynamic programming (DP) –matching [19] [20]. DTW-based method suffers by speaker independent (SI) recognition cases whereas it shows good performance for speaker-dependent (SD) cases [21] [22]. Besides, DTW is suitable for less than 50 vocabularies work. On the Contrary, HMM-based method is utilized for large vocabulary and continuous speech recognition and can let everyone use after the training progress [23]. Nevertheless, DTW still has various applications including menu-driven commanding and phone dialing due to it is uncomplicated and easy to implement. In this thesis, we adopt DTW for our speech recognition because we only have only a few commands to be recognized.

The organization of this paper is as follows: In section II, source localization based on ORIR, CC, and GCC method are involved. Section III introduces speech enhancement method such as array beamforming and phase difference (PD) algorithm. Section IV includes speech feature extraction and DTW algorithm. Some respective tests of the prior three sections and implementation of the singing robot are demonstrated in section V. The conclusion is provided in section VI.

II. SOURCE LOCALIZATION

2.1 ORIR-based method

Consider an array of *m* microphones mounted at arbitrary locations whose center is at point *P*. Imagine a sound that originates from azimuth θ and elevation ϕ relative to *P*. The task of any localization algorithm is to process each of the *m* microphone inputs $\{I_1,\ldots,I_m\}$ to generate azimuth and elevation estimates $\hat{\theta}$ and $\hat{\phi}$, respectively. Ideally, the algorithm should utilize all available location cues to maximize accuracy. Differences in times of arrival between the microphones will vary with the location of the sound source and then can be utilized to generate location estimates. Additional location cues are available if the frequency content of the microphone inputs varies with the location of the sound source. This can be achieved by inserting an object centered at *P* into the listening environment so that the filtering properties of the object will vary with the orientation of the sound source.

For illustrative purposes, consider the situation in which *m*=2 microphones are mounted on the two sides of a robot. Let the center of the hearing system of the robot be located at *P*. Consider a sound that originates at azimuth θ and elevation ϕ relative to *P*. The sound is altered by the head and torso of the robot before it arrives at the microphones. If I_j is a digital recording of the input to the *j*th microphone, then

$$I_j = O * F_j^{(\theta,\phi)} \tag{1}$$

where O is the sound that would arrive at point P if the robot were absent, * is the convolution operator, and $F_j^{(\theta,\phi)}$ is the object-related impulse response (ORIR) for microphone j when a sound originates from (θ,ϕ). The ORIR is a representation of the object-related transfer function (ORTF) in the time domain rather than the frequency domain and can therefore include both the time- and frequency-based filtering effects of the head and torso.

With the ORIR, the localization algorithm is motivated by the relationship of following two equations:

$$I_1 * F_2^{(\theta,\phi)} = (O * F_1^{(\theta,\phi)}) * F_2^{(\theta,\phi)} = O * F_1^{(\theta,\phi)} * F_2^{(\theta,\phi)} \tag{2}$$

and

$$I_2 * F_1^{(\theta,\phi)} = (O * F_2^{(\theta,\phi)}) * F_1^{(\theta,\phi)} = O * F_2^{(\theta,\phi)} * F_1^{(\theta,\phi)} = O * F_1^{(\theta,\phi)} * F_2^{(\theta,\phi)} \tag{3}$$

This follows from the commutativity and associativity of the convolution operator. If the correct location is chosen, then the operation will lead to the same result for both microphone inputs. If the ORIR associated with some other location (θ',ϕ') is chosen, however, then the results will differ:

$$I_1 * F_2^{(\theta',\phi')} = (O * F_1^{(\theta,\phi)}) * F_2^{(\theta',\phi')} = O * F_1^{(\theta,\phi)} * F_2^{(\theta',\phi')} \tag{4}$$

and

$$I_2 * F_1^{(\theta',\phi')} = (O * F_2^{(\theta,\phi)}) * F_1^{(\theta',\phi')} = O * F_2^{(\theta,\phi)} * F_1^{(\theta',\phi')} = O * F_1^{(\theta',\phi')} * F_2^{(\theta,\phi)} \tag{5}$$

Of course, a wide variety of similarity metrics are available; a moderate amount of testing suggested that the Pearson correlation maximized the accuracy and reliability of the “cross-channel” localization algorithm. Choosing $(\hat{\theta},\hat{\phi})$ as follows:

$$\max_{(\hat{\theta},\hat{\phi})} r(I_1 * F_2^{(\hat{\theta},\hat{\phi})}, I_2 * F_1^{(\hat{\theta},\hat{\phi})}) \tag{6}$$

In advance, we measured the ORIR database $\{F_1^{(\theta,\phi)}, F_2^{(\theta,\phi)}\}$ for doing the convolution operation in (6). Because we do not exactly know where the sound emits, we have to globally search ORIR database and then verify the correlation between $I_1 * F_2^{(\hat{\theta},\hat{\phi})}$ and $I_2 * F_1^{(\hat{\theta},\hat{\phi})}$. Ideally, if we make a correct choice, the Pearson correlation coefficient equals to 1. However, the coefficient can be affected by measurement accuracy, environment on the instant, quality of sound and so forth, we choose proximal one as estimated azimuth and elevation.

Even the “cross channel” approximation (i.e. finding the Pearson correlation coefficient which is most close to the value of 1) is directly perceived through the senses and accurate. But it is also computational consuming caused by global search. According to this drawback, we present a hybrid method. We utilize cross correlation (CC) and generalized cross correlation (GCC) method for roughly DOA estimation and then do the cross channel match for precision. Later on, we will introduce CC and GCC method.

2.2 Cross correlation

The DOA estimation is based on where the source is arranged to be in the array’s far-field, as illustrated in Fig. 1. In this situation, sound source radiates a plane wave in the condition of propagating through the non-dispersive medium air. The normal to the wavefront makes an angle θ with the line jointing the sensors in the linear array, so there exists time delay/advance between each microphone. In Fig. 1, we choose the right sensor as the reference point and the spacing between the microphones is denoted as d. Therefore, we can show that the plane wave needs more distance to get the second sensor. The distance can be easily calculated which equals to $d\cos\theta$. So we know the time difference is given by $\tau_{12} = d\cos\theta / c$, where *c* is the sound velocity in air. If the angle ranges between 0° and 180° and if τ_{12} is known then θ is uniquely determined, and *vice versa*. Therefore, estimating the incident angle θ is essentially identical to estimating the time difference τ_{12}. In other words, the DOA estimation problem is also called time-difference-of-arrival (TDOA) estimation problem in far-field case.

The speech source signal $s(k)$ propagates radiatively and the sound level drops as a function of distance from the source. If we choose the first microphone as the reference point, the signal received by *n*th microphone at time *k* can be expressed as follows:

$$\begin{aligned} y_n(k) &= \alpha_n s(k - t - \tau_{n1}) + \upsilon_n(k) \\ &= \alpha_n s[k - t - \lambda_n(\tau)] + \upsilon_n(k) \\ &= x_n(k) + \upsilon_n(k)\ , \quad n = 1, 2, \ldots, N \end{aligned} \tag{7}$$

where α_n are the attenuation factors, $s(k)$ is unknown source signal, t is the propagation time to sensor 1, $\upsilon_n(k)$ is an additive noise at the nth sensor, which is assumed to be uncorrelated with the signal and the noise captured by the other sensors, τ is the TDOA, and $\tau_{n1} = \lambda_n(\tau)$ is the TDOA between sensor 1 and n with $\lambda_1(\tau)=0$ and $\lambda_2(\tau)=\tau$. For n = 3, … , *N*, the function λ_n depends only on τ because of the microphone array geometric. We have

$$\lambda_n(\tau) = (n-1)\tau\ , \qquad n = 2, \ldots, N \tag{8}$$

Consider only two microphones case, the cross-correlation function (CCF) between the two observation signal $y_1(k)$ and $y_2(k)$ is defined as

$$r^{CC}_{y_1y_2}(p) = E[y_1(k)y_2(k+p)] \tag{9}$$

Substituting (7) into (9) , we can get

$$r^{CC}_{y_1y_2}(p) = \alpha_1\alpha_2 r^{CC}_{ss}(p-\tau) + \alpha_1 r^{CC}_{s\upsilon_2}(p+t) + \alpha_2 r^{CC}_{s\upsilon_1}(p-t-\tau) + r^{CC}_{\upsilon_1\upsilon_2}(p) \tag{10}$$

By the assumption of the signal and the noise are uncorrelated, (10) can easily checked that $r^{CC}_{y_1y_2}(p)$ in Fig. 2 (a) reaches maximum at $p = \tau$. Hence, we can obtain the TDOA between $y_1(k)$ and $y_2(k)$ as

$$\hat{\tau}^{CC} = \arg\max_{p} r^{CC}_{y_1y_2}(p) \tag{11}$$

where $p \in [-\tau_{\max}, \tau_{\max}]$, and $\tau_{\max}$ is the maximum possible delay.

In digital implementation of (11), some approximations are required because of the CCF is not known and must be estimated. A normal practice is to replace the CCF defined in (9) by time averaged estimate. Suppose that at time instant k we have a set of observation samples of x_n , $\{x_n(k), x_n(k+1), \ldots, x_n(k+K-1)\}$, $n = 1, 2$, the corresponding CCF can be estimated as either

$$r^{CC}_{y_1y_2}(p) = \begin{cases} \dfrac{1}{K}\displaystyle\sum_{i=0}^{K-p-1} y_1(k+i)y_2(k+i+p) \ , & p \geq 0 \\ r^{CC}_{y_1y_2}(-p) \ , & p < 0 \end{cases} \tag{12}$$

or

$$r^{CC}_{y_1y_2}(p) = \begin{cases} \dfrac{1}{K-p}\displaystyle\sum_{i=0}^{K-p-1} y_1(k+i)y_2(k+i+p) \ , & p \geq 0 \\ r^{CC}_{y_1y_2}(-p) \ , & p < 0 \end{cases} \tag{13}$$

where K is the block size. The difference between (12) and (13) is that the former leads to a biased estimator, while the latter is an unbiased one. However, since it has a lower estimation variance and is asymptotically unbiased, the former had been widely adopted in many applications.

2.3 Generalized cross correlation

Same as CC method, GCC employs free-field model (7) and considers only two microphones, i.e., *N*=2. Then TDOA estimate between the two microphones is obtained as the lag time that maximizes the CCF between the filtered signals of the microphone outputs which is often called the generalized CCF (GCCF):

$$\hat{\tau}^{\mathrm{CC}} = \arg\max_{p} r_{y_1 y_2}^{CC}(p) \tag{14}$$

where

$$\hat{\tau}^{GCC} = \arg\max_{p} r_{y_1 y_2}^{GCC}(p) \tag{15}$$

$$\begin{aligned} r_{y_1 y_2}^{GCC}(p) &= F^{-1}[\Psi_{y_1 y_2}(f)] \\ &= \int_{-\infty}^{\infty} \Psi_{y_1 y_2}(f) \mathrm{e}^{j2\pi fp} df \\ &= \int_{-\infty}^{\infty} \vartheta(f)\phi_{y_1 y_2}(f) \mathrm{e}^{j2\pi fp} df \end{aligned} \tag{16}$$

is the GCC function, $F^{-1}[\cdot]$ stands for the inverse discrete-time Fourier transform (IDTFT),

$$\phi_{y_1 y_2}(f) = E[Y_1\ Y_2^{*}(f)] \tag{17}$$

is the cross-spectrum with

$$Y_n(f) = \sum_{k} y_n(k) e^{-j2\pi fk}, \qquad n = 1, 2, \tag{18}$$

$\vartheta(f)$ is a frequency-domain weighting function, and

$$\Psi_{y_1 y_2}(f) = \vartheta(f)\phi_{y_1 y_2} \tag{19}$$

is the generalized cross-spectrum.

There are many different choices of the frequency-domain weighting function $\vartheta(f)$, leading to a variety of different GCC methods.

2.3.1 Classical cross correlation

If we set $\vartheta(f)=1$, it can be checked that the GCC degenerates to the cross-correlation method, as seen in Fig. 2 (b). The only difference is that now the CCF is estimated using the discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT), which can be implemented efficiently thanks to the fast Fourier transform (FFT).

We know from the free-field model (7) that

$$Y_n(f) = \alpha_n S(f) e^{-j2\pi f[t-\lambda_n(\tau)]} + V_n(f)\ , \quad n = 1, 2, \tag{20}$$

Substituting (20) into (19) and noting that the noise signal at one microphone is uncorrelated with the source signal and the noise signal at the other microphone by assumption, we have

$$\Psi^{GCC}_{y_1y_2}(f) = \alpha_1\alpha_2 e^{-j2\pi f\tau} E[|S(f)|^2] \tag{21}$$

The fact that $\Psi^{GCC}_{y_1y_2}(f)$ depends on the source signal can be detrimental for TDOA estimation since speech us inherently non-stationary.

2.3.2 Smoothed coherence transform

In order to overcome the impact of fluctuating levels of the speech source signal on TDOA estimation, an effective way is to pre-whiten the microphone outputs before their cross-spectrum is computed. This is equivalent to choosing

$$\vartheta(f) = \frac{1}{\sqrt{E[|Y_1(f)|^2]E[|Y_2(f)|^2]}} \tag{22}$$

which leads to so-called smoothed coherence transform (SCOT) method [28]. Substituting (20) and (22) into (19) produces the SCOT GCC function and can be seen in Fig. 2 (c):

$$
\begin{aligned}
\Psi_{y_1y_2}^{SCOT} &= \frac{\alpha_1\alpha_2 e^{-j2\pi f\tau} E[|S(f)|^2]}{\sqrt{E[|Y_1(f)|^2]E[|Y_2(f)|^2]}} \\
&= \frac{\alpha_1\alpha_2 e^{-j2\pi f\tau} E[|S(f)|^2]}{\sqrt{\alpha_1^2 E[|S_1(f)|^2]+\sigma_{\upsilon_1}^2(f)}\sqrt{\alpha_2^2 E[|S_2(f)|^2]+\sigma_{\upsilon_2}^2(f)}} \\
&= \frac{e^{-j2\pi f\tau}}{\sqrt{1+\frac{1}{SNR_1(f)}}\sqrt{1+\frac{1}{SNR_2(f)}}}
\end{aligned}
\tag{23}
$$

where

$$
\begin{aligned}
&\sigma_{\upsilon_n}^2(f) = E[|V_n(f)|^2] \\
&SNR_n(f) = \frac{\alpha_n^2 E[|S(f)|^2]}{E[|V_n(f)|^2]} \quad , \qquad n = 1, 2,
\end{aligned}
\tag{24}
$$

If the SNRs are the same at two microphones, then we get

$$
\Psi_{x_1x_2}^{SCOT}(f) = [\frac{SNR(f)}{1+SNR(f)}]\cdot e^{-j2\pi f\tau} \tag{25}
$$

Therefore, the performance of the SCOT algorithm for DOA estimation would vary with the SNR. But when the SNR is large enough,

$$
\Psi_{x_1x_2}^{SCOT}(f) = e^{-j2\pi f\tau} \tag{26}
$$

which implies that the estimation performance is independent of the power of the source signal. So, the SCOT method is theoretically superior to the CC method. But this superiority only holds when the noise level is low.

2.3.3 Phase transform

It becomes clear by examining (16) that the TDOA information is conveyed in the phase rather than the amplitude and only keep the phase. By setting

$$
\vartheta(f) = \frac{1}{|\phi_{y_1y_2}(f)|} \tag{27}
$$

we get the phase transform (PHAT) method [5]. In this case, the generalized cross-spectrum is given by

$$\Psi_{y_1 y_2}^{PHAT}(f) = e^{-j2\pi f\tau} \qquad (28)$$

which depends only on the TDOA τ. Substituting (28) into (19), we obtain an ideal GCC function:

$$\Psi_{y_1 y_2}^{PHAT}(f) = \int_{-\infty}^{\infty} e^{-j2\pi f(p-\tau)} df = \begin{cases} \infty, & p = \tau \\ 0, & otherwise \end{cases} \qquad (29)$$

The GCC function of PHAT is in Fig. 2 (d). As a result, the PHAT method performs in general better than CC and SCOT methods with respect to TDOA estimation.

III. Speech Enhancement

3.1 Array processing

In sensor arrays, a widely used signal model assumes that each propagation channel introduces some delay and attenuation only. With this assumption and in the scenario where we have an array consisting of *N* sensors, the array outputs, at time *k*, are expressed as

$$y_n(k) = \alpha_n s[k - t - F_n(\tau)] + v_n(k) = x_n(k) + v_n(k), \quad n = 1, 2, \ldots, N \qquad (30)$$

where α_n (*n* = 1, 2, . . .,*N*), which range between 0 and 1, are the attenuation factors due to propagation effects, *s*(*k*) is the unknown source signal (which can be narrowband or broadband), *t* is the propagation time from the unknown source to sensor 1, $v_n(k)$ is an additive noise signal at the *n*th sensor, *τ* is the relative delay or more often it is called the time difference of arrival (TDOA)] between sensors 1 and 2, and $F_n(\tau)$ is the relative delay between sensors 1 and *n* with $F_1(\tau)$ = 0 and $F_2(\tau)$ $= \tau$.

The most frequently and basically method that we use is delay-and-sum (DAS) beamformer. Such a beamformer consists of two basic processing steps. The first step is to time-shift each sensor signal by a value corresponding to the TDOA between that sensor and the reference one. With the signal model given above and after time shifting, we obtain

$$\begin{aligned} y_{a,n}(k) &= y_n[k + F_n(\tau)] \\ &= \alpha_n s(k-t) + v_{a,n}(k) \\ &= x_{a,n}(k) + v_{a,n}(k),\ n = 1, 2, \ldots, N \end{aligned} \qquad (31)$$

where

$$v_{a,n}(k) = v_n[k + F_n(\tau)] \qquad (32)$$

and the subscript ‘a’ implies an aligned copy of the sensor signal. The second step consists of adding up the time-shifted signals, giving the output of a DAS beamformer:

$$z_{DS}(k) = \frac{1}{N}\sum_{n=1}^{N} y_{a,n}(k) = \alpha_s s(k-t) + \frac{1}{N} v_s(k) \qquad (33)$$

where

$$\begin{aligned} \alpha_s &= \frac{1}{N}\sum_{n=1}^{N} \alpha_n, \\ v_s(k) &= \sum_{n=1}^{N} v_{a,n}(k) = \sum_{n=1}^{N} v_n[k + F_n(\tau)] \end{aligned} \qquad (34)$$

Next, we introduce an optimized microphone array. In order to reject the noise in an acoustic field, we need to optimize the way we combine multiple microphones. Specifically we need to consider the direction gain, i.e. the gain of the microphone array in a noise field over that of a simple omni-directional microphone. A common quantity used is the directivity factor Q, or equivalently, the directivity index (DI) $[10\log_{10}(Q)]$.

The directivity factor is defined as

$$Q(\omega, \theta_0, \phi_0)=\frac{|E(\omega, \theta_0, \phi_0)|^2}{\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}|E(\omega, \theta, \phi)|^2 u(\omega, \theta, \phi)\sin\theta d\theta d\phi} \tag{35}$$

where the angle θ and ϕ are the standard spherical coordinate angles, θ_0 and ϕ_0 are the angles at which the directivity factor is being measured, $E(\omega, \theta, \phi)$ is the pressure response of the array, and $u(\omega, \theta, \phi)$ is the distribution of the noise power. The function u is normalized such that

$$\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}u(\omega, \theta, \phi)\sin\theta d\theta d\phi = 1 \tag{36}$$

The directivity factor Q can be written as the ration of two Hermitian quadratic forms [35] as

$$Q = \frac{\mathbf{w}^H\mathbf{A}\mathbf{w}}{\mathbf{w}^H\mathbf{B}\mathbf{w}} \tag{37}$$

where

$$\mathbf{A} = \mathbf{S}_0\mathbf{S}_0^H \tag{38}$$

$\mathbf{w}$ is the complex weighting applied to the microphones and H is the complex conjugate transpose. The elements of the matrix $\mathbf{B}$ are defined as

$$b_{mn} = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}u(\omega, \theta, \phi)\exp[j\mathbf{k}(\mathbf{r}_m - \mathbf{r}_n)]\sin\theta d\theta d\phi \tag{39}$$

and the elements of the vector $\mathbf{S}_0$ are defined as

$$s_{0n} = \exp(j\mathbf{k}_0\cdot\mathbf{r}_n) \tag{40}$$

Note that for clarity we have left off the explicit functional dependencies of the above equations on the angular frequency ω . The solution for the maximum of Q, which is Rayleigh quotient, is obtained by finding the maximum generalized eigenvector of the homogeneous equation

$$\mathbf{A}\mathbf{w} = \lambda_M \mathbf{B}\mathbf{w} \tag{41}$$

The maximum eigenvalue of above equation is given by

$$\lambda_M = \mathbf{S}_0^H \mathbf{B}^{-1} \mathbf{S}_0 \tag{42}$$

The corresponding eigenvector contains the weights for combining the elements to obtain the maximum directional gain

$$\mathbf{w}_{opt} = \mathbf{B}^{-1} \mathbf{S}_0 \tag{43}$$

where $\mathbf{w}_{opt}$ is our array filter with maximum directivity index (maxDI).

Besides, there is another type of optimization of array beampatterns: constant beamwidth (constBW). Its cost function is defined as

$$J = \frac{\int_0^{2\pi} \int_0^{\theta_1} \left| H(\omega, \theta, \phi) \right|^2 \sin\theta \, d\theta \, d\phi}{\frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} \left| H(\omega, \theta, \phi) \right|^2 \sin\theta \, d\theta \, d\phi} \tag{44}$$

We can also take above equation as a Rayleigh quotient problem which was mentioned before. Both maxDI and constBW are so called “filter-and-sum” technique which is shown in Fig. 3

3.2 Phase difference approach

For promoting speech recognition accuracy, we apply a two-microphone approach to extract the speech which is masked by other interference. The algorithm separates signals based on differences in arrival time of signal components from two microphones. As we know that the human binaural system is primarily based on the use of interaural time difference (ITD) at low frequencies and interaural level difference (ILD) information at high frequencies. However, we only focus on the use of ITD cues. When multiple sound sources are presented, it is generally assumed that humans only want to hear the sound from one direction which is equivalent to the corresponding ITD.

First, the system performs a short-time Fourier transform (STFT) which decomposes the two input signals in time and in frequency. We get a subset of ITDs by comparing phase terms of the two input signals at each frequency. Through a time-frequency mask, we can extract the speech whose ITD is close to the target speaker and suppress unwanted sources.

The left-channel $x_L[n]$ and right-channel $x_R[n]$ are inputs of the system. We assume that the location of the desired target signal is known and its ITD is zero. For mathematical convenience, we refer to the number of interfering sources as *L*, with $\delta(l)$ being their respective ITDs. Note that both *L* and $\delta(l)$ are unknown. With the above formulations, the signals are the microphones are

$$x_L[n] = \sum_{t=0}^{L} x_l[n] \ , \qquad x_R[n] = \sum_{t=0}^{L} x_l[n-\delta(l)] \tag{45}$$

with $x_0[n]$ representing the target signal, $x_l(l \neq 0)$ representing interfering signals x_L and x_R, respectively, representing the signals at the left and right microphones. The corresponding short-time Fourier transforms can be represented as

$$X(k,m) = \sum_{n=-\infty}^{\infty} x[n]w[m-n]e^{-j2\pi kn/N} \tag{46}$$

$$\begin{aligned} X_L(k,m) &= \sum_{i=0}^{L} X_i(k,m) \\ X_R(k,m) &= \sum_{i=0}^{L} e^{-j2\omega_k d_i(k,m)} X_i(k,m) \end{aligned} \tag{47}$$

where $w[n]$ is a finite-duration Hamming window, *k* indicates one of *N* frequency bins, with positive frequency samples corresponding to $\omega_k = \dfrac{2\pi k}{N}$ for $0 \leq k \leq \dfrac{N}{2} - 1$. In (45), the difference between two microphones is a pure time delay, but it is more appropriate to consider the time delays are function of frequency. Correspondingly, we use the frequency-dependent ITD parameter $d(k,m)$ to replace the frequency-independent term δ in (45). Next, we assume that a specific time-frequency bin (k_0,m_0), is dominated by a single sound source *l*. This leads to

$$\begin{aligned} X_L(k_0;m_0) &\approx X_{l^*}(k_{0,}m_0) \\ X_R(k_0;m_0) &\approx e^{-j\omega_{k_0}d(k_0,m_0)}X_{l^*}(k_{0,}m_0) \end{aligned} \tag{48}$$

where the source l^* dominates the time-frequency bin (k_0,m_0). It becomes a simple binary decision to determine whether the time-frequency bin (k_0,m_0) belongs to the target speaker or not. The frequency-dependent ITD $d(k,m)$ for a particular time-frequency bin (k_0,m_0) is

$$|d(k_{0,}m_0)| \approx \frac{1}{|\omega_{k_0}|}\min_r |\angle X_R(k_0,m_0)| - \angle X_L(k_0,m_0) - 2\pi r| \tag{49}$$

then we derive the binary masking criterion

$$\mu(k_0,m_0) = \begin{cases} 1, & if \ \ |d(k_0,m_0)| \ \le \ \tau \\ \eta, & otherwise \end{cases} \tag{50}$$

In other words, we take only time-frequency bins with the condition of $|d(k_0,m_0)| \le \tau$ as the target speaker where τ can be considered as width of receiving beam. And we use a small value 0.01 for η to block the unwanted time-frequency bins. The mask $\mu(k,m)$ in (50) is applied to $\bar{X}(k,m)$, the averaged signal spectrogram from the two channels, and speech is reconstructed from the $\tilde{X}(k,m)$ where

$$\begin{aligned} \bar{X}(k,m) &= \frac{1}{2}\{X_L(k,m) + X_R(k,m)\} \\ \tilde{X}(k,m) &= \mu(k,m)\bar{X}(k,m) \end{aligned} \tag{51}$$

The PD method illustrated above is based on the sound source is located in the direction of the microphone array axis (i.e. 90 degrees). In actual application, we can’t guarantee that sound always originates at 90 degrees, so we employ beam-steering techniques [5] to compensate time delay between two sensors then do PD for speech enhancement. The beam-steering angle can be solved by DOA estimation which has been mentioned in section II.

IV. SPEECH RECOGNITION

4.1 Feature extraction

In speech recognition system, we have to extract the characteristic of the signal as templates. In this thesis, we choose the commonly used Mel-frequency ceptral coefficient (MFCC) to be our speech feature. The Mel-frequency ceptrum is a representation of the short-term power spectrum of a sound. The difference from the real cepstrum is that a nonlinear frequency scale is used. Davis and Mermelstein [36] showed the MFCC representation to be beneficial for speech recognition. Next, we briefly introduce how this kind of parameter is extracted out.

First of all, given the DFT of the input signal

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} \quad , \quad 0 \le k < N \tag{52}$$

Then, we design a filterbank with M filters (m=1, 2, … , M) , where filter m is triangular filter given by

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{(k - f[m-1])}{(f[m] - f[m-1])} & f[m-1] \le k \le f[m] \\ \dfrac{(f[m+1] - k)}{(f[m+1] - f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \tag{53}$$

which satisfies $\sum_{m=1}^{M} H_m[k] = 1$. As shown above, the frequency bands are equally spaced on the Mel-scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. These increasing bandwidths filters are displayed in Fig. 4. In (54), the boundary frequencies $f[m]$ are calculated by the following equation:

$$f[m] = (\frac{N}{F_s})B^{-1}[B(f_1) + m\frac{B(f_h) - B(f_1)}{M+1}] \tag{54}$$

where the Mel-scale B and B^{-1} are defined as

$$\begin{aligned} B(f) &= 1125 \cdot \ln(1 + \frac{f}{700}) \\ B^{-1}(b) &= 700 \cdot [\exp(\frac{b}{1125}) - 1] \end{aligned} \tag{55}$$

The input $X_a[k]$ passed through each filter then we get the log-energy output

$$S[m] = \ln(\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]) \quad , \quad 0 < m \le M \tag{56}$$

Finally, we take the discrete cosine transform (DCT) of the M filter outputs:

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos(\frac{\pi n(m - \frac{1}{2})}{M}) \quad , \quad 0 \le n \le M \tag{57}$$

where M varies from 24 to 40 in different kind of implementation. However, in speech recognition, we typically only use first 13 ceptrum coefficients.

4.2 DTW algorithm

DTW is based on overall distortion measure computed from the accumulated distance between the test and reference patterns along the aligned path. DTW is also called dynamic programming (DP) matching shown in Fig. 5. However, the alignment path is not apparent in the actual speech recognizers unless additional backtracking is performed.

The procedure for computing the accumulated distortion distance can be illustrated by the following procedure:

I) INITIALIZATION

```
$D(1,1) = d(1,1), \quad B(1,1) = 1,$
$for\ j = 2,\ldots,M$  {
        $compute\ D(1,j) = \infty$
    }
```

II) ITERATION

```
$for\ i = 2,\ldots,N$ {
    $for\ j = 1,\ldots,M\ \ compute$ {
        $D(i,j) = \min_{1\le p\le M}[D(i-1,p) + d(p,j)]$
        $B(i,j) = \arg\min_{1\le p\le M}[D(i-1,p) + d(p,j)]$
    }
}
```

III) BACKTRACKING AND TERMINATION

The optimal (minimum) distance is $D(N,M)$ and the optimal path is

$$(s_1, s_2, \ldots, s_N)$$

where $s_N = M$ and $s_i = B(i+1, s_{i+1})$, $i = N-1, N-2, \ldots, 1$

For reducing the computation load and we consider that normal speaking speed cannot be more than two times of the training speed; therefore, we introduce two kinds of constraints: local constraint and global constraint. They can be expressed in Fig. 6 and Fig. 7, respectively. Fig. 8 shows different local constraint may cause different path to do the DP.

V. SIMULATION AND EXPERIMENT

We divide this section into four parts: source localization, speech enhancement, speech recognition and finally, robot implementation. We give a briefly introduction which can be seen in Table 1.

In Table1, all the simulation and test are arranged by using two Knowles MEMS microphone (SPM0204HE5-PB) whose spacing is 0.05 m. However in experiment, our microphone spacing is changed by 0.2 m for increasing the resolution of the time delay estimation.

5.1 Source localization

First of all, we measure the ORIR from the test input (swept sine) to two sensors mounted on the LEGO NXT robot inside the $4m \times 4m \times 3m$ anechoic chamber. Besides, for increasing the directivity of the microphone array, we combine with an acoustic device – horn in Fig. 9. The robot system is displayed in Fig. 10.We utilize B&K Pulse audio analyzer (3560C) to generate signal and receive sound data, Tannoy loudspeaker (V8) to play test input, B&K power amplifier (2176) to magnify test input and B&K turntable system (9640) to rotate the robot for measuring sound from different azimuth. Fig. 11 gives the whole experimental configuration. Fig. 12 (a) shows the 0 azimuth ORTF, (b) is the ORIR, and (c) is the magnified picture of (b). Fig. 13 (a) shows the 90 azimuth ORTF, (b) is the ORIR, and (c) is the magnified picture of (b). Comparing with Fig. 12 (c) and Fig. 13 (c), we can clearly find that there exists time delay due to azimuth of the source. So we employ the CC method as mentioned before to compute the ITD. We gather ITD database and the azimuth is 5 degrees per move, from 0 degree to 355 degrees. Finally, we utilize 2-norm to calculate the ILD database. The ITD and ILD database can be seen in Fig. 14.

Secondly, we present some DOA estimation simulation result calculated by CC, GCC (PHAT) and hybrid method (GCC_PHAT combined with ORIR based method) in Fig. 15. Note that sound source emits from 90 degrees and noise emits from 30 degrees (SNR is 12 dB). By comparing the estimation results, we choose GCC_PHAT to be to construct hybrid method because it performs better than the others. Fig. 16 – Fig. 18 show that we use hybrid method (ORIR based method combines with GCC_PHAT) to do DOA estimation in the circumstance of sound source always emits from 90 degrees and 3 types of noise (white noise, babble and exhibition) emit from 0 degree to 75 degrees every 15 degrees (SNR is 0 dB and 12 dB).

5.2 Speech enhancement

First, we show the beampatterns of DAS, maxDI and constBW in Fig. 19. When we input a clean speech mixed with white noise and SNR is 10 dB, Fig. 20 (a) – (c) show the outputs passed through these three types of array and we compare them with the original input.

Secondly, we present the output passed through the PD and identically compare with the original output in Fig. 20 (d). It is easily seen that PD provides much better noise reduction performance than DAS, maxDI and constBW methods, so we just apply PD method on our experimental robot which will be later shown in section 5.4.

5.3 Speech recognition

In order to ensure that our speech enhancement works, we utilize the corpus provided by ITRI which concludes 50 commands in Mandarin made by 6 men and 5 women. After extracting features of each command, we apply DTW to do the DP-matching and recognize. In Fig. 21, we show that the clean speech polluted by white noise, babble, car noise and movie in different conditions of SNR. We set speech at 90 degrees and noise at 0 degree. The RRs of outputs enhanced by DAS, maxDI, constBW and PD are shown in Fig. 22. By comparing Fig. 22, we can discover that PD performs significantly better than array processing.

Because PD provides better enhancement result, later we just employ PD method as our enhancement application. For detailed observation of the performance of PD, we set noise (white noise, babble and exhibition) every 15 degrees from 0 degree to 90 degrees and speech at 90 degrees always, the results are shown in Fig. 23 - Fig. 25. To achieve the best performance, we optimize τ in (50) for each noise angle mentioned above.

5.4 Robot implementation

Fig. 26 shows block diagram of whole system. When two microphones mounted on the NXT robot receive the 2-channel signals, they are sent to source localization system and speech enhancement system. After the process of speech enhancement, signal is sent to speech recognition system. Now, we get the information of user's direction and speaking words. Finally, the robot turns to user and sings what user asks. Fig. 27 provides the schematic diagram of robot implementation.

VI. CONCLUSION

The GCC methods are computationally efficient. They induce very short decision delays and hence have a good tracking capability: an estimate is produced almost instantaneously. We can also realize that GCC_PHAT has better performance than CC, classical GCC and GCC_SCOT in Fig. 2 because GCC_PHAT has the most obvious peak to make us find TDOA easily. Therefore, we only combine GCC_PHAT with ORIR-based technique to be our hybrid localization method. In Fig. 15, we can discover that hybrid method can catch the direction of target and interference more accurately than the others. Next we use this hybrid method to test numerous recorded wave files. When the difference between angle of the target and angle of the interference is small, hybrid method still has the ability to distinguish them. But sometimes we may make wrong judgments when the noise is too loud (i.e. SNR is low).

Secondly, in Fig. 19 and Fig. 20 (a) – (c), we can discover that DAS, maxDI and constBW can only provide a little improvement because of the microphone number. If we use 4 microphones or more, the performance would be better certainly. PD exhibits it powerful ability to separate the desired voice from the interference and also provides excellent RR which will be demonstrated later.

Finally, we make a conclusion about speech recognition by analyzing the RR results. By comparing Fig. 21 and Fig. 22 (a) – (c), it is can be expectable that the RR gets insignificant promotion by array processing. However, we obtain remarkable RR result shown in Fig. 22 (d) when we utilize PD algorithm. We can control the value of τ to decide the angle of the receiving beamwidth. When we choose an appropriate τ which makes the target source pass through yet the interference be suppressed, we can get a purified signal without distortion. Therefore, the RRs can always be above 79% even at the situation of 0 SNR. In order to cope with all circumstances that the spanning angle between the purposed sound and the interference is not fixed. We put the interference at different azimuth. By our localization method and do a simple subtraction, we get the difference between angle of the target and angle of the main interference. By this, we can adjust τ to optimize noise reduction performance. However, if the target sound and the interference originate from almost the same azimuth, we can’t find any value of τ to suppress unwanted noise which means PD method will fail in this case. It can be proved by Fig. 23 – Fig. 26. In Fig. 23 – Fig. 26, we set noise angle vary from 0 degree to 90 degrees and source angle always at 90 degrees. We can find that the RRs are above 58% except 90 degrees at 0 SNR. Unlike Fig. 22 (d), the RRs decrease because there exists DOA estimation error and leads to non-optimal τ selection.

REFERENCE

[1] Luo, R.C. and Su, K.L. "Autonomous Fire-Detection System Using Adaptive Sensory Fusion for Intelligent Security Robot," *Mechatronics, IEEE*, vol. 12, pp.274-281, June 2007.

[2] Voth, D., "A new generation of military robots," *Intelligent Systems, IEEE*, vol. 19, pp. 2-3, Jul-Aug 2004.

[3] Krose, B., Bunschoten, R. Hagen ST, Terwijn, B., Vlassis, N., "Household robots look and learn: environment modeling and localization from an omnidirectional svision system," *Robotics & Automation Magazine, IEEE*, vol. 11, pp. 45-52, Dec. 2004.

[4] Geppert, L, "Yoshihiro Kuroki: dancing with robots ," *Spectrum, IEEE*, vol. 41, pp. 34-35, Feb. 2004.

[5] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications* (Springer, New York, 2001).

[6] E. Hansler and G. Schmidt, *Speech and Audio Processing in Adverse Environments* (Springer, New York, 2008)

[7] J. G. Ryan and R. A. Goubran, "Optimum near-field performance of microphone arrays subject to a far-field beampattern constraint," *J. Acoust. Soc. Am.*, vol. 108, pp. 2248-2255, 2000.

[8] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Trans. Signal Process.*, vol. 50, pp. 1843-1854, 2002.

[9] X. Chen, Y. Shi, and W. Jiang, "Speaker tracking and identifying based on indoor localization system and microphone array," *21st International Conference on Advanced Information Networking and Applications Workshops*, vol. 2, pp.

347-352, 2007.

[10] H. Wang and P. Chu, “Voice source localization for automatic camera pointing system in video conferencing,” *Proceedings of the ICASSP*, vol. 1, pp. 187-190, 1997.

[11] S. Fischer and K. U. Simmer, “An adaptive microphone array for handsfree communication,” *Proceedings of the 4th International Workshop on Acoustic Echo and Noise Control, IWAENC-95*, pp. 44–47, 1995.

[12] M. R. Bai and C. Lin, “Microphone array signal processing with application in three-dimensional spatial hearing,” *J. Acoust. Soc. Am.*, vol. 117, pp. 2112-2121, 2005.

[13] K. Nakadai, H. Nakajima, M. Murase, S. Kaijiri, K. Yamada, T. Nakamura, Y. Hasegawa, H. G. Okuno, and H. Tsujino, “Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays,” *Proceedings of the ICASSSP* , vol. 4, pp. 929-932, 2006.

[14] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 320–327, Aug. 1976.

[15] Justin A. MacDonald, “A localization algorithm based on head-related transfer functions,” *J. Acoust. Soc. Am.*, vol. 123, pp. 4290-4296, June 2008.

[16] P. Arabi and G. Shi, “Phase-based dual-microphone robust speech enhancement,” *IEEE Tran. Systems, Man, and Cybernetics-Part B:*, vol. 34, no. 4, pp. 1763-1773, Aug. 2004.

[17] D. Halupka, S. A. Rabi, P. Aarabi, and A. Sheikholeslami, “Real-time dual-microphone speech enhancement using field programmable gate arrays,” *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. v/149-v/152,

March 2005.

[18] C. Kim, K. Kumar, B. Raj, and R. M. Stern, “Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain,” *INTERSPEECH-2009*, Sept. 2009.

[19] Hiroaki Sakoe and Seibi Chiba, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition,” *IEEE Transcations on Acoustics, Speech, Signal Processing* vol. 26 (1), pp. 43-49, February, 1978.

[20] Hiroaki Sakoe and Seibi Chiba, “Comparative study of DP-pattern matchingtechniques for speech recognition” (in Japanese), in *1973 Tech. Group Meeting Speech, Acoust. SOC. Japan, Preprints (S73-22),*Dec. 1973.

[21] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[22] C. Lévy, G. Linarès, and P. Nocera, “Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems,” *Workshop on DSP in Mobile and Vehicular Systems*, Nagoya, Japan, Apr. 2003.

[23] Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, *Spoken Language Processing*, Prentice Hall PTR, NJ, 2001

[24] L. Wang, N. Kitaoka, and S. Nakagawa, “Robust distance speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM,” *Speech Commun.*, vol. 49, 501-513, 2007.

[25] J. Benesty, “Adaptive eigenvalue decomposition algorithm for passive acoustic source localization,” *J. Acoust. Soc. Am.*, vol. 107, 384-391, 2000.

[26] R. Bucher and D. Misra, “A synthesizable vhdl model of the exact solution for three-dimensional hyperbolic positioning system,” *VLSI Des.*, vol. 15, 507-520, 2002.

[27] A. Brutti, M. Omologo, and P. Svaizer, “Speaker localization based on oriented global coherence field,” *Proceedings of the Interspeech*, pp. 2606-2609, 2006.

[28] G. C. Carter, A. H. Nuttall, and P.G. Cable, “The smoothed coherence transform,” *Proc. IEEE*, vol. 61, pp. 1497-1498, Oct, 1973.

[29] B. Champagne, S. Bédard, and A. Stéphenne, “Performance of time-delay estimation in presence of room reverberation,” *IEEE Trans. Speech Audio Process.*, vol. 4,pp. 148-152, Mar. 1996.

[30] J. P. Ianniello, “Time delay estimation via cross-correlation in the presence of large estimation errors, ” *IEEE Trans. Acoust., Speech, Signal Process.* ,vol. ASSP-30, pp.998-1002, Dec. 1982.

[31] M. S. Brandstein, “A pitch-based approach to time-delay estimation of reverberant speech,” in *Proc. IEEE WASPAA*, Oct. 1997.

[32] M. Omologo, and P. Svaizer, “Acoustic event localization using a crosspower-spectrum phase based technique,” in *Proc. IEEE ICASSP*, 1997, vol. 2, pp. 273-276.

[33] M. Omologo, and P. Svaizer, “Acoustic source location in noisy and reverberant environment using CSP analysis,” in *IEEE ICASSP*, 1996, vol. 2, pp. 921-924.

[34] C. Wang and M. S. Brandstein, “A hybrid real-time face tracking system,” *Proc. IEEE ICASSP*, 1998, vol. 6, pp. 3737-3741.

[35] D. K. Cheng, “Optimization techniques for antenna arrays,” *Proc. IEEE*, vol. 59, pp. 1664-1674, Dec. 1971.

[36] Davis, S. and P. Mermelstein, "Comparison of Parametric Representations for Monosyllable Word Recognition in Continuously Spoken Sentences," *IEEE Trans.on Acoustics, Speech and Signal Processing*, 1980, **28**(4), pp. 357-366.

	Simulation and Test	*Experiment*
Source localization	Localization methods simulation and test	ORIR measurement
Speech enhancement	Techniques simulation	
Speech recognition	Recorded file test	
Robot implementation		Robot implementation

Table 1 Introduction of section V.

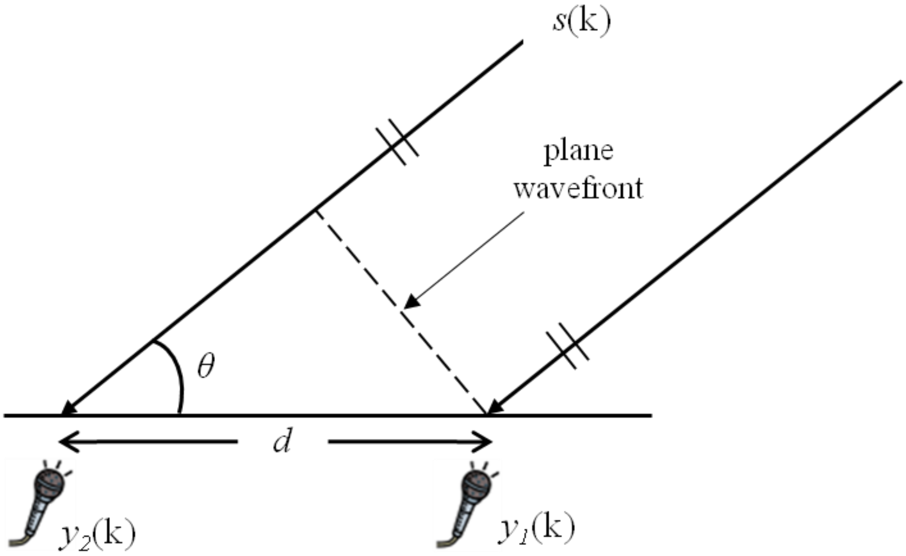
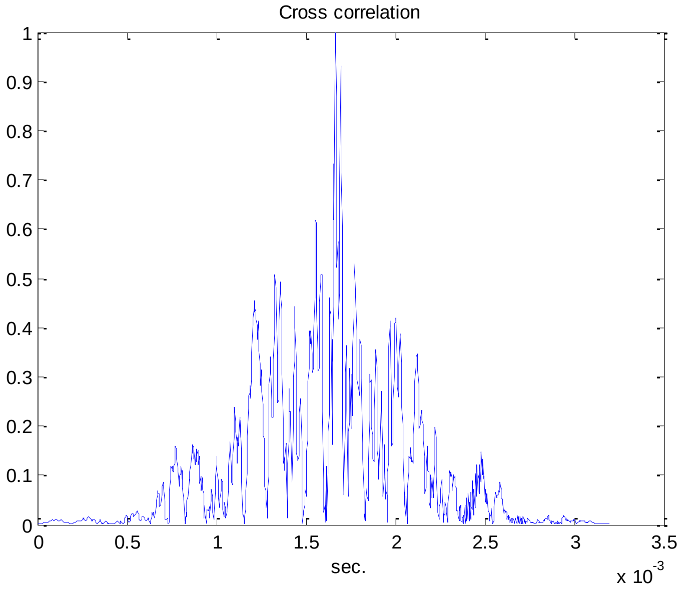
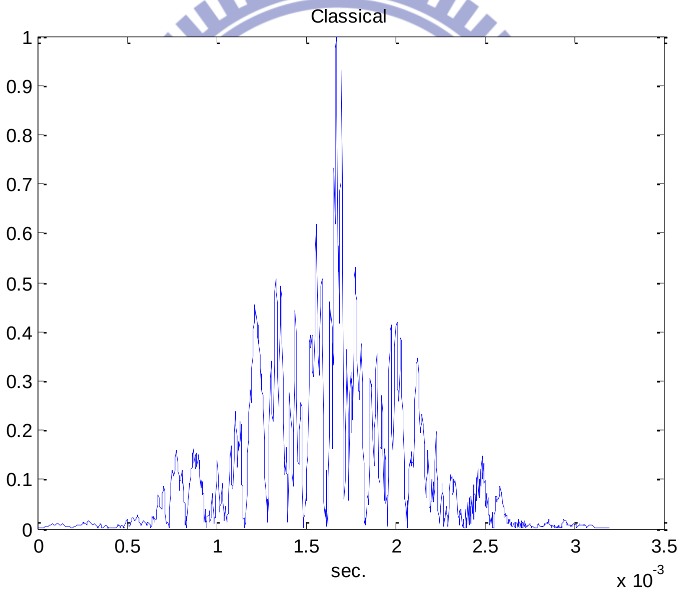


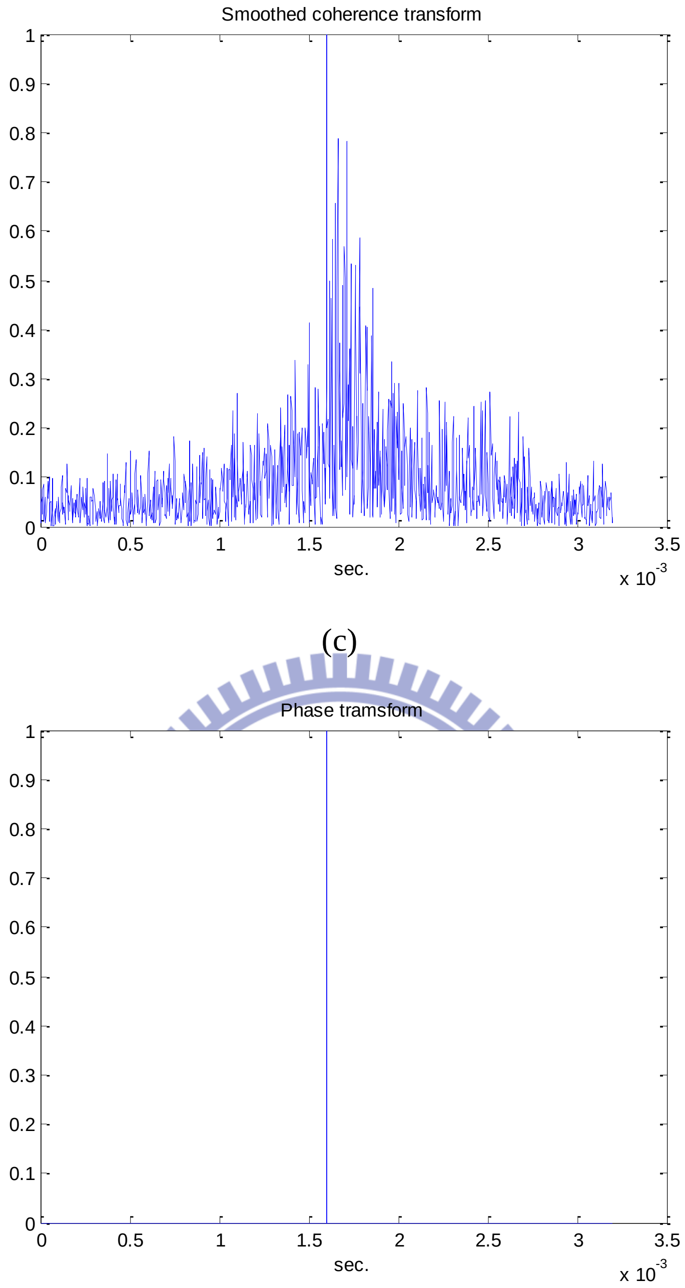
Fig. 1 Illustration of the DOA estimation problem in 2-dimensional space with two identical microphones: the source $s(k)$ is located in far-field, the incident angle is θ, and the spacing between two sensors is d.


Cross correlation
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
2
2.5
3
3.5
sec.
x 10^{-3}

(a)


Classical
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
2
2.5
3
3.5
sec.
x 10^{-3}

(b)



(c)

(d)

Fig. 2 (a) CCF, (b) GCCF of classical method, (c) GCCF of classical method of SCOT, (d) GCCF of PHAT method.

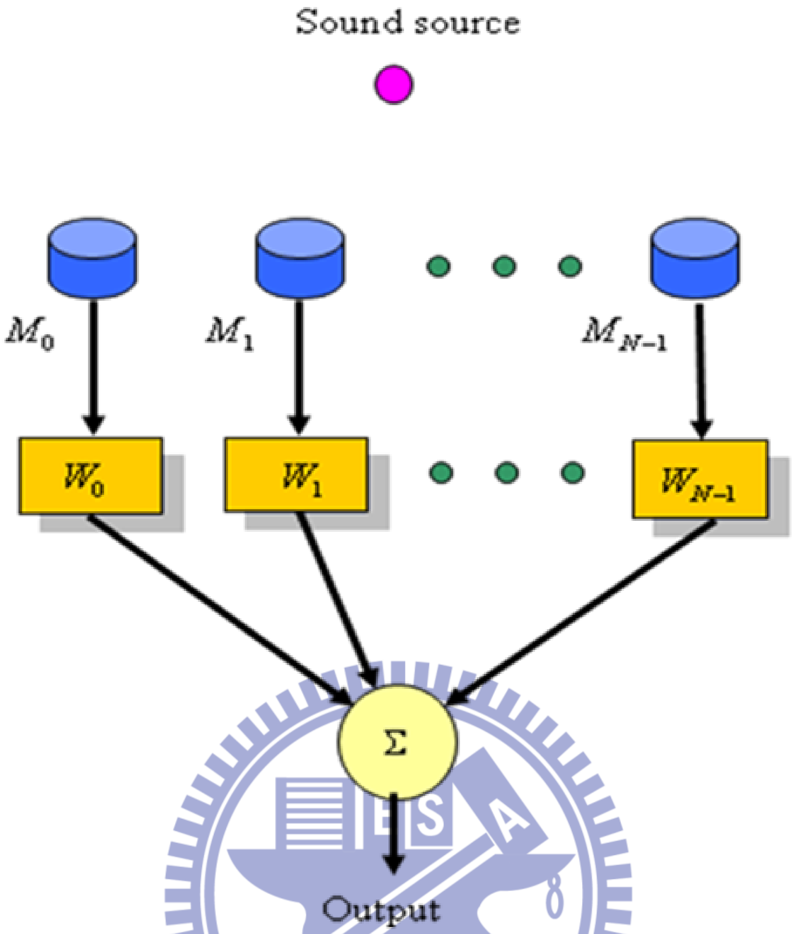


Fig. 3 Schematic diagram of filter-and-sum method.

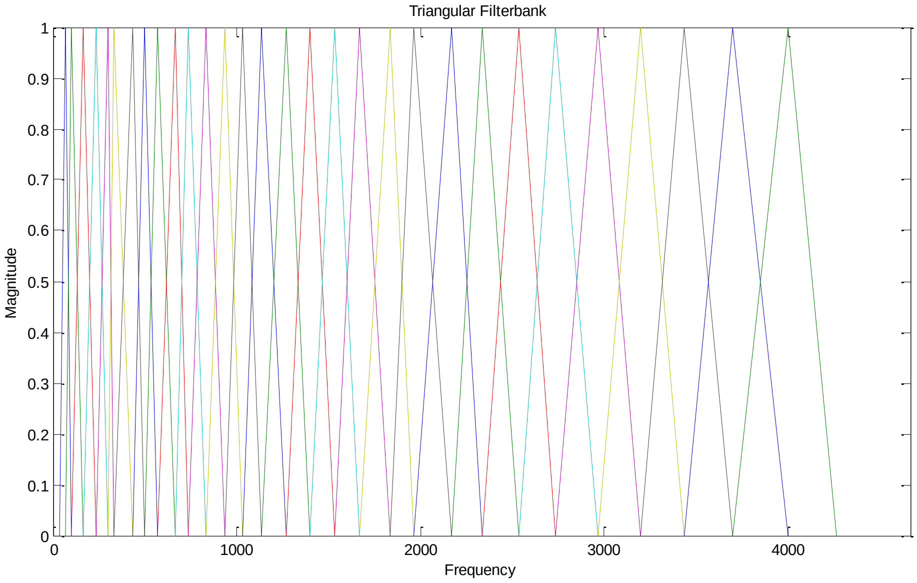


Fig. 4 30-channel triangular filterbank.

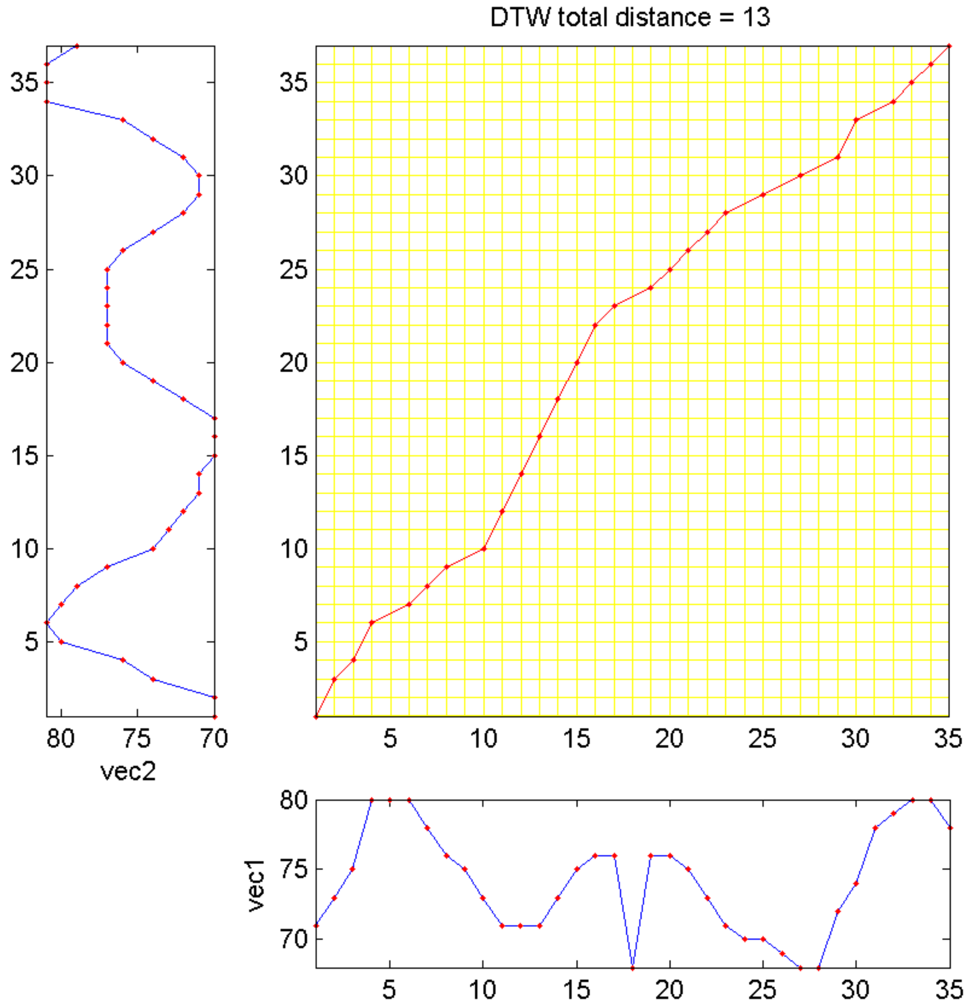


Fig. 5 The DP matching of two templates: the vertical axis stands for training speech template and the horizontal axis stands for test speech template.

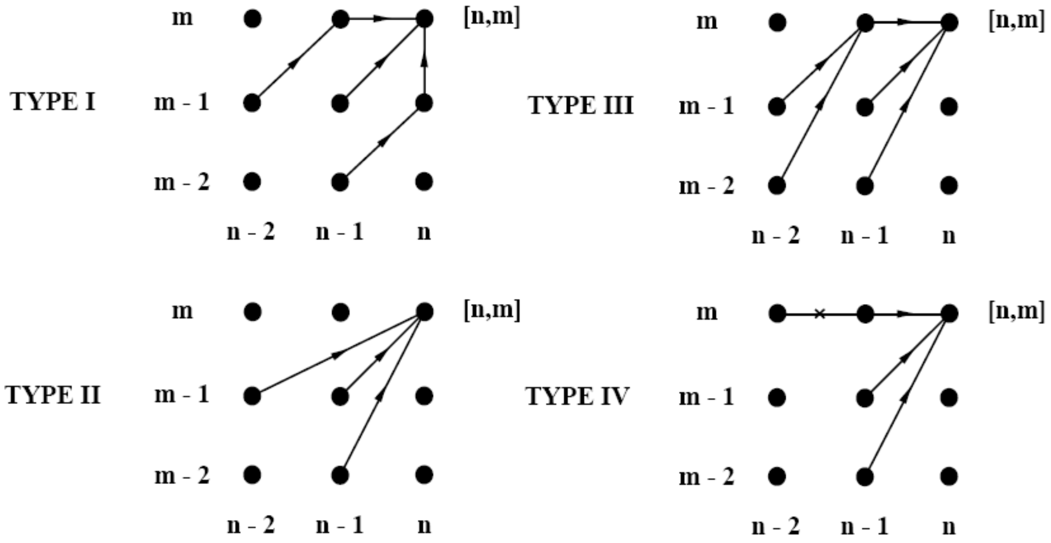


Fig. 6 Four kinds of local constraint for DTW.

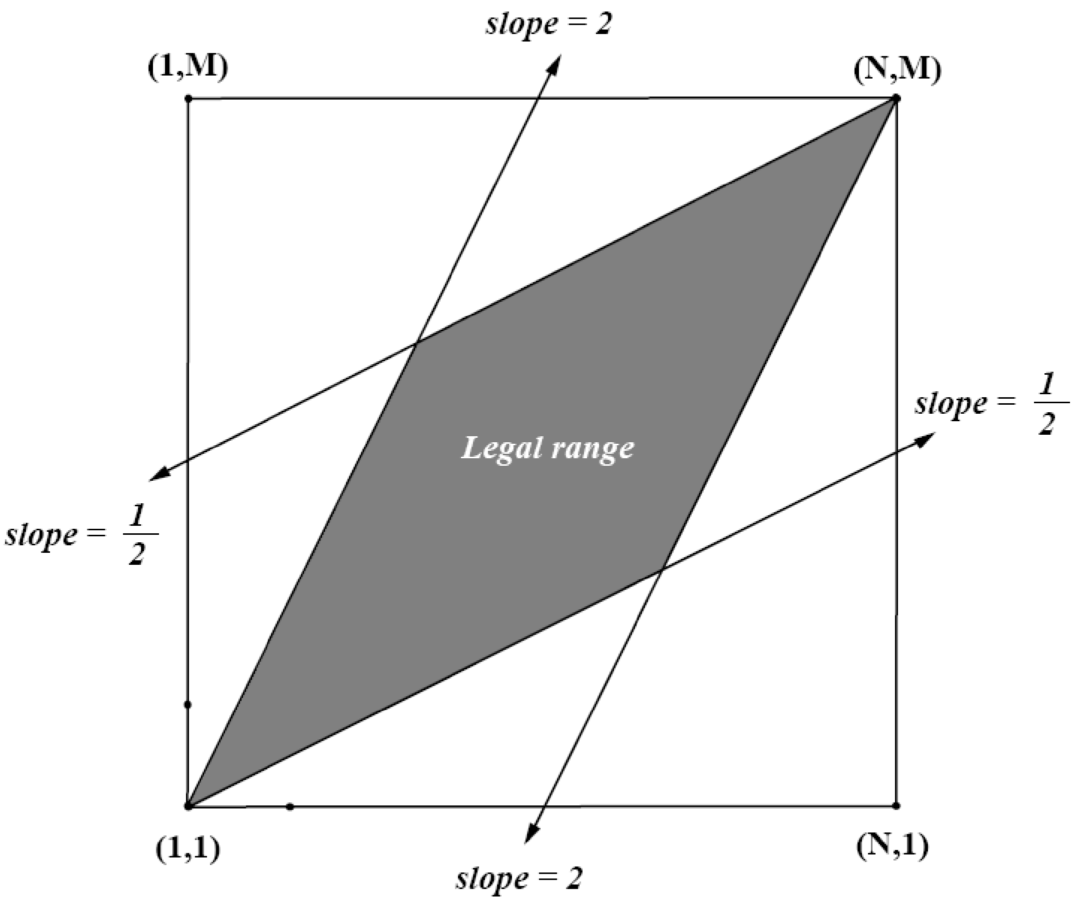


Fig. 7 Global constraint for DTW.

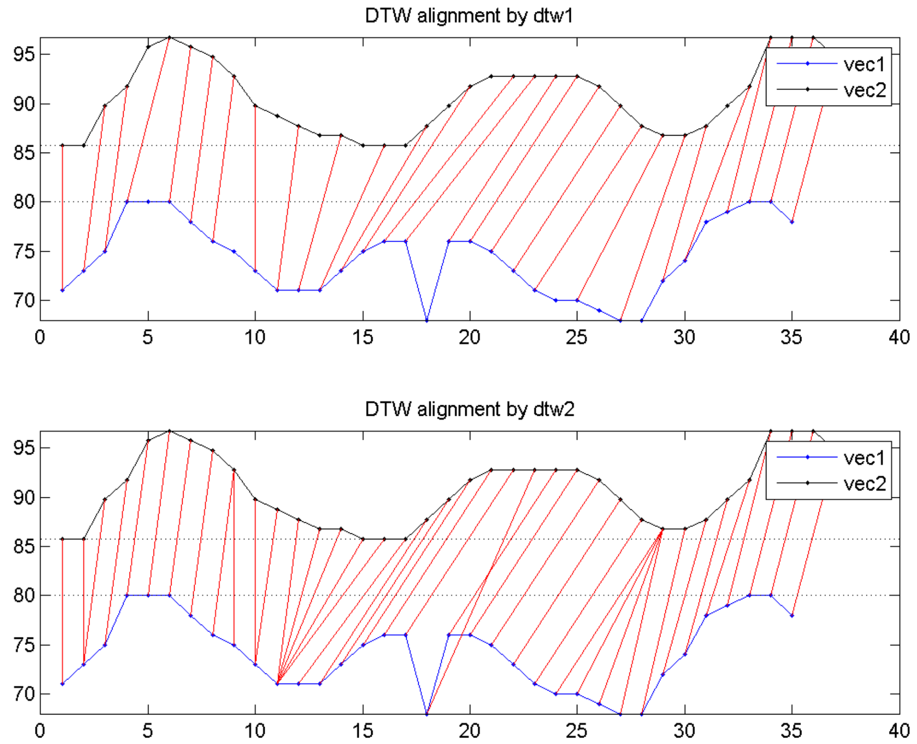


Fig. 8 Different local constraint may cause different DP result.

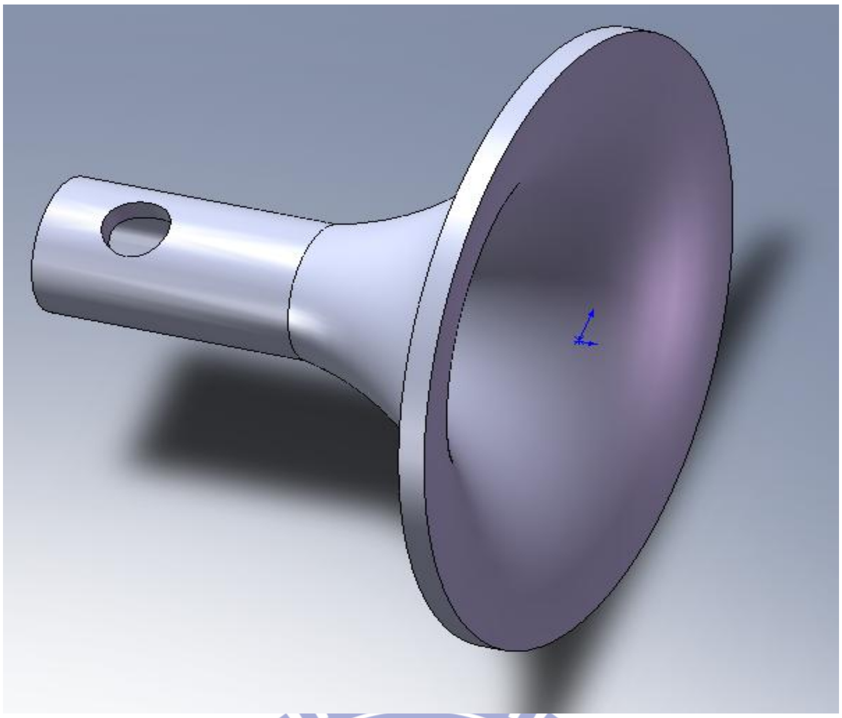


Fig. 9 Acoustic device – Horn.

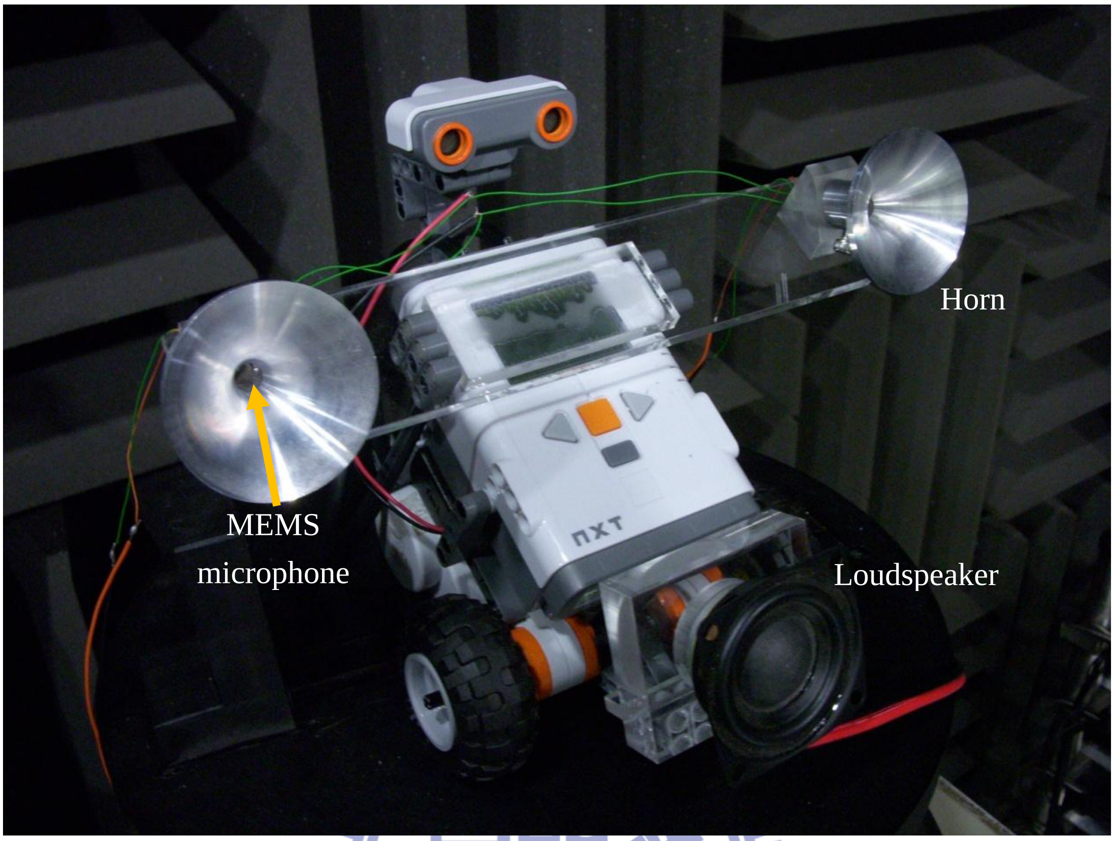


Fig. 10 Robot system.

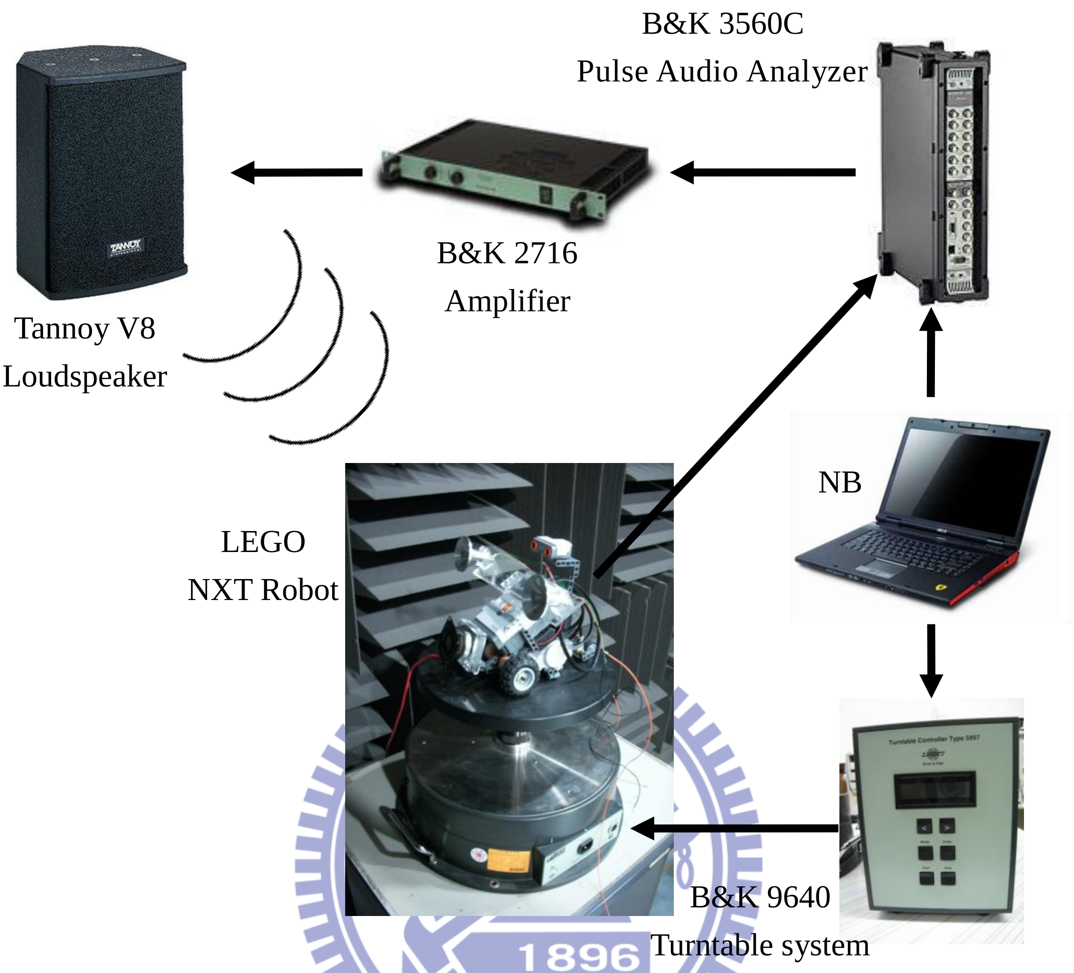
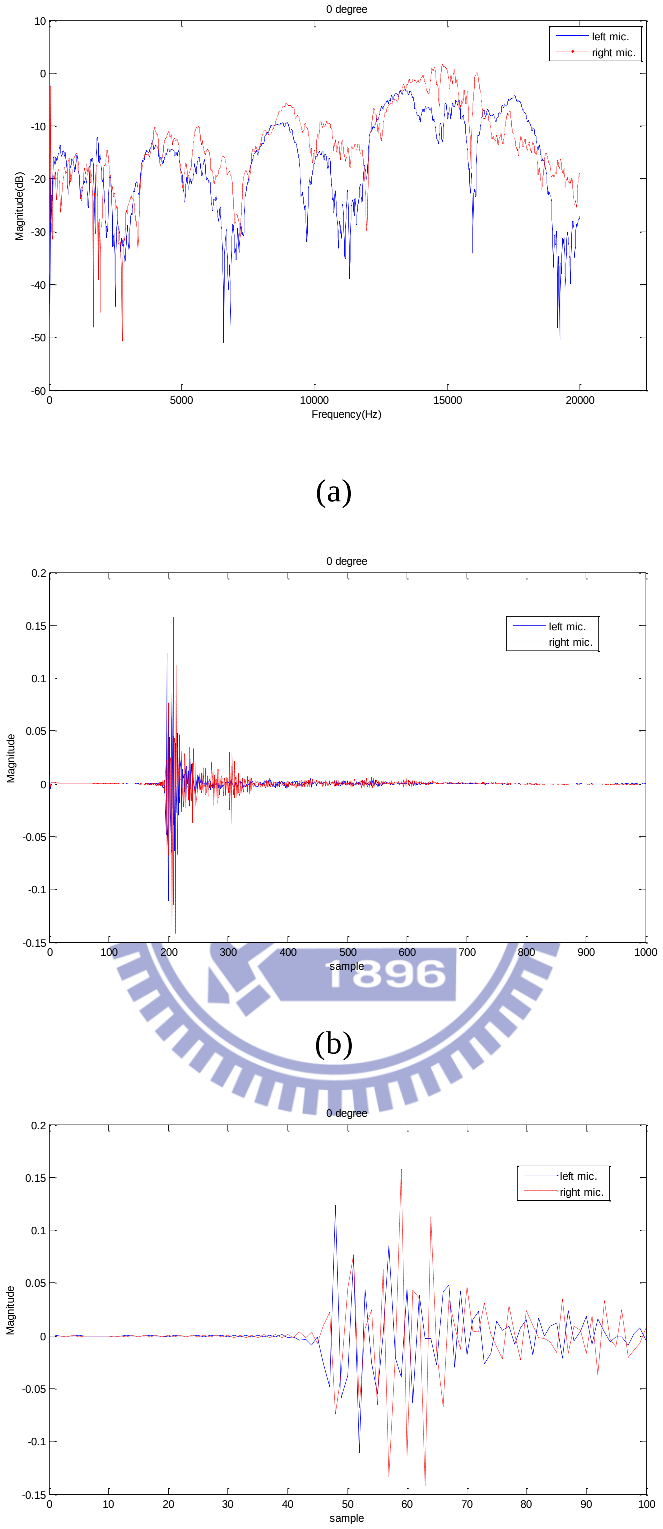


Fig. 11 Experimental configuration of ORIR measurement.

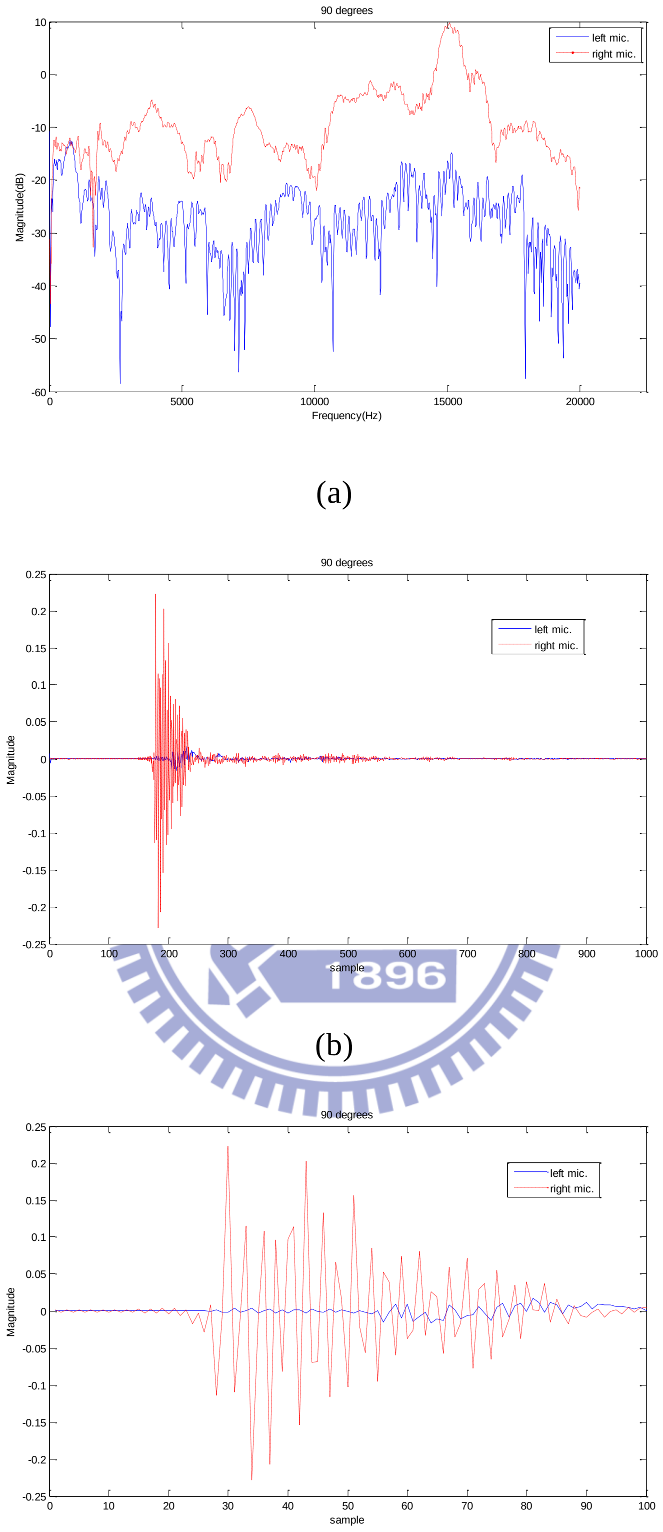


(a)

(b)

(c)

Fig. 12 (a) 0 degree ORTF, (b) 0 degree ORIR, (c) Magnified picture of (b).

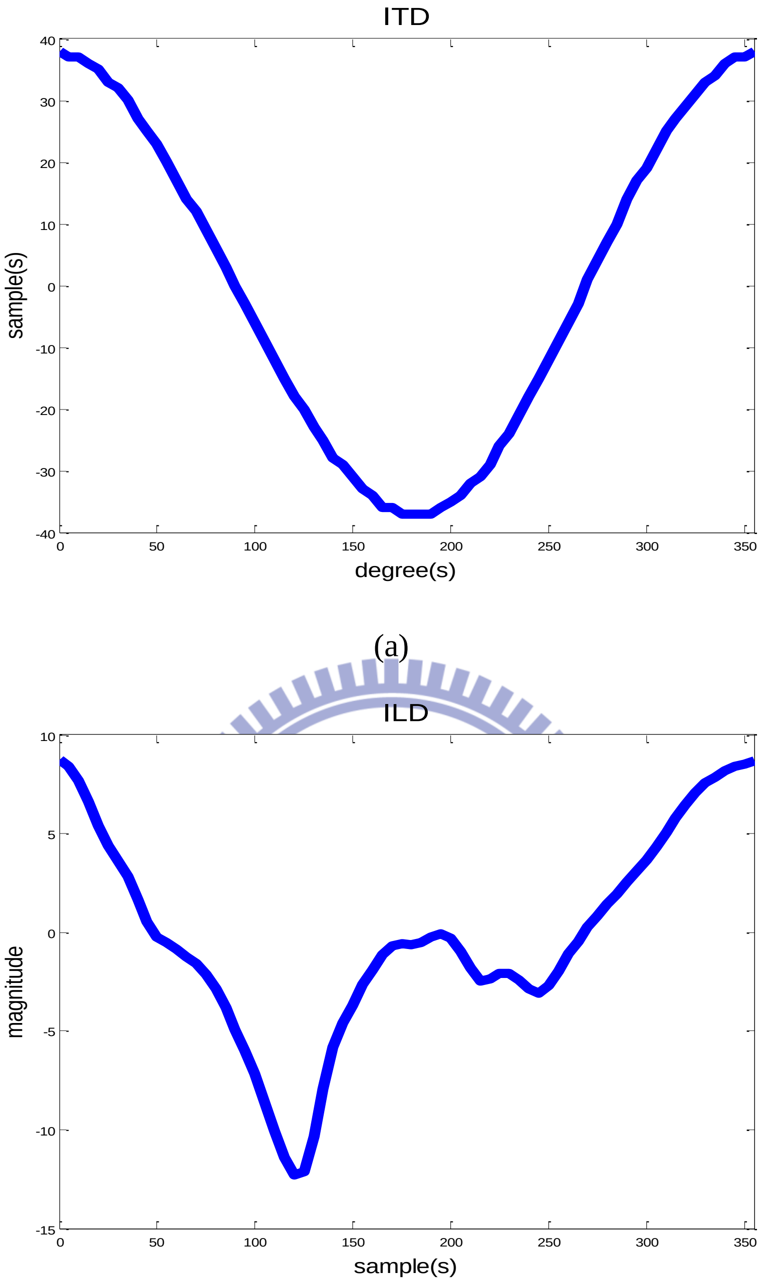


(a)

(b)

(c)

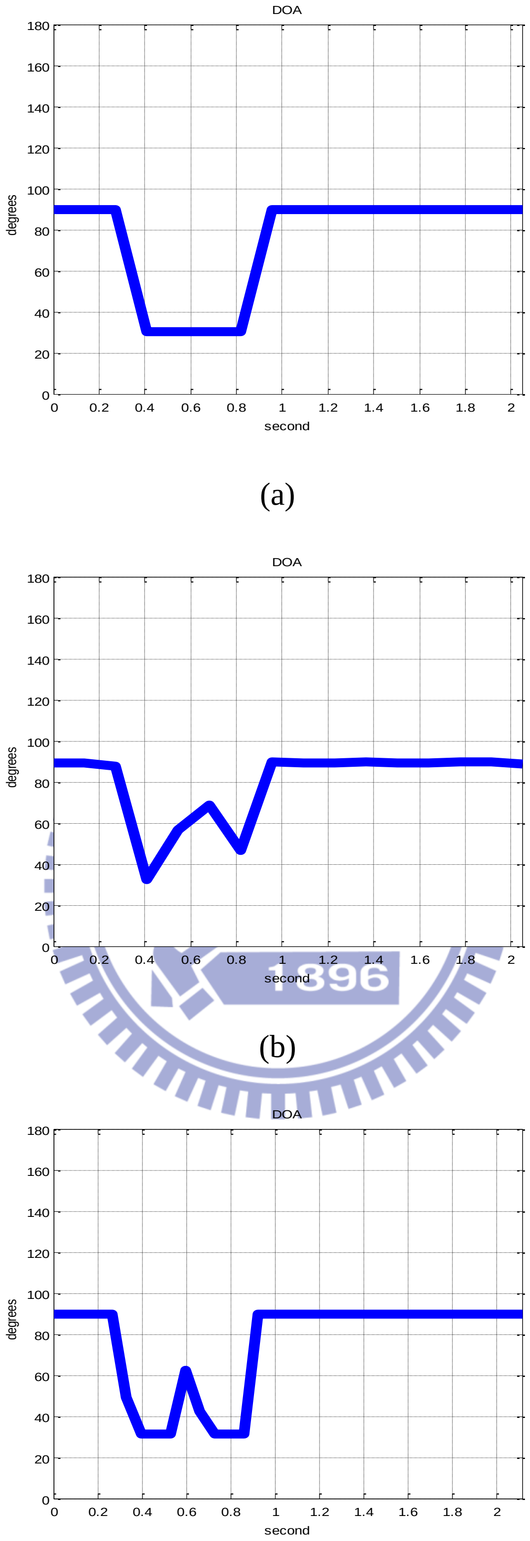
Fig. 13 (a) 90 degree ORTF, (b) 90 degree ORIR, (c) Magnified picture of (b).



(a)

(b)

Fig. 14 (a) ITD database, (b) ILD database.

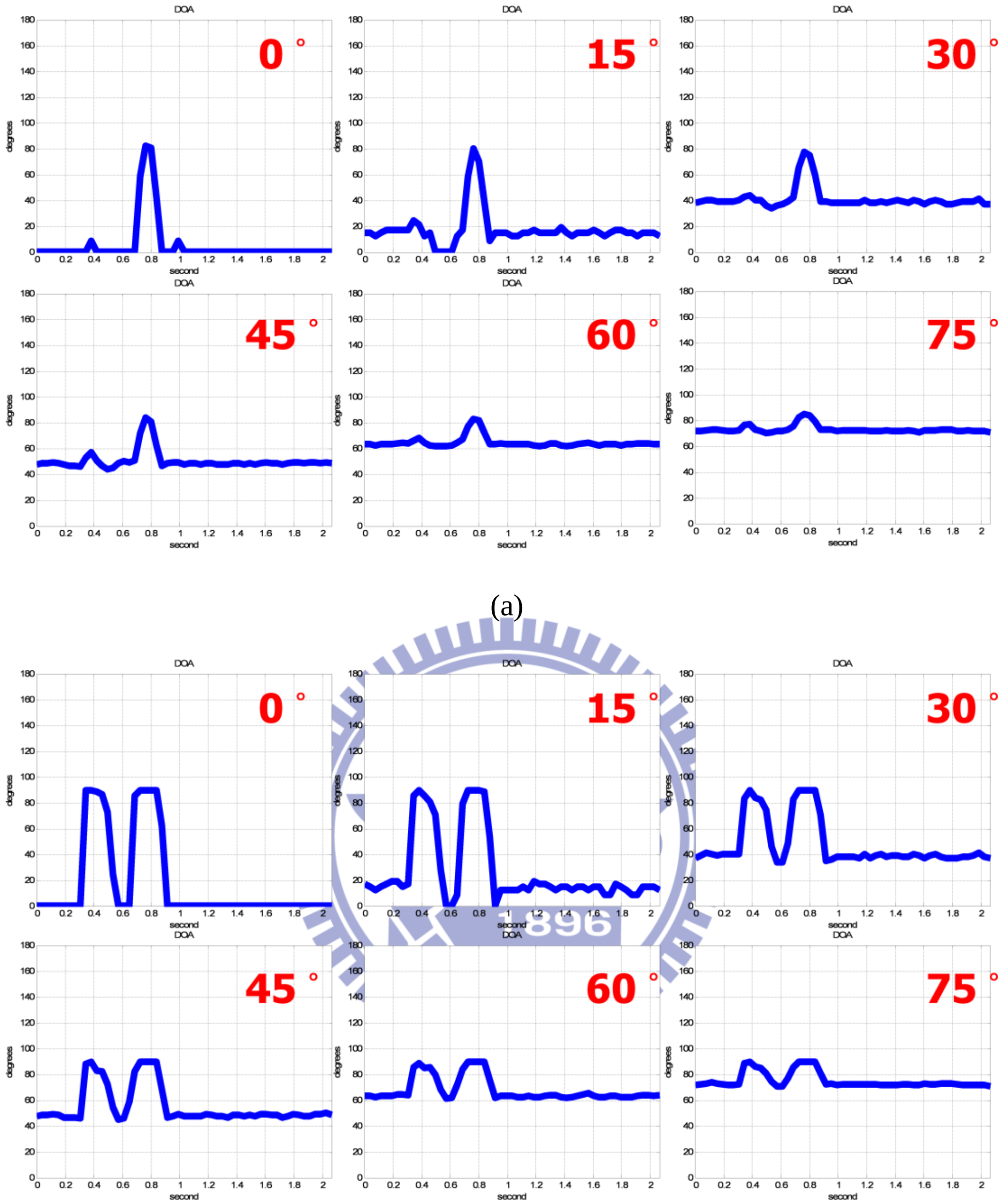


(a)

(b)

(c)

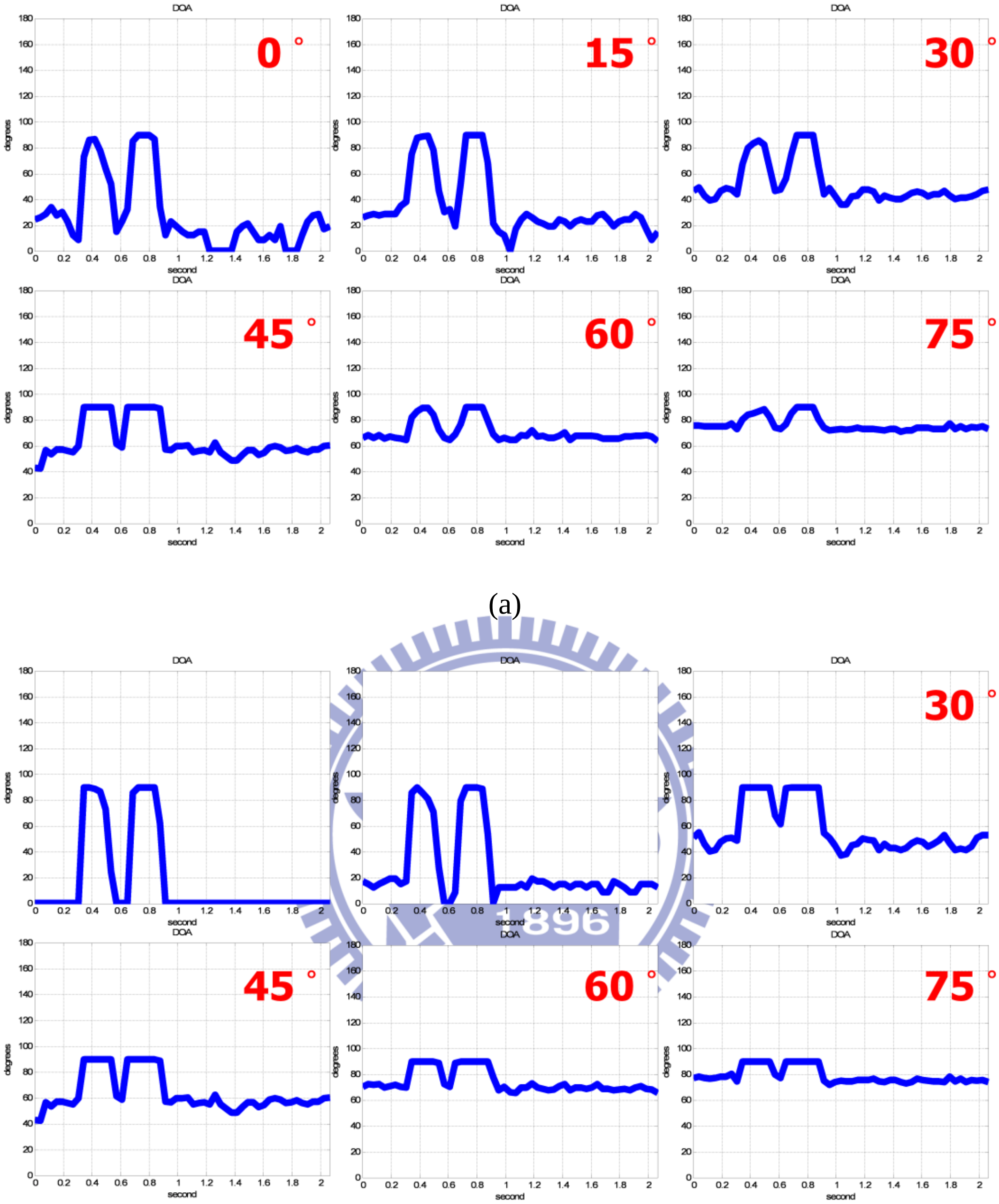
Fig. 15 (a) DOA estimation by CC method, (b) DOA estimation by GCC_PHAT method, (c) DOA estimation by hybrid method.



(a)

(b)

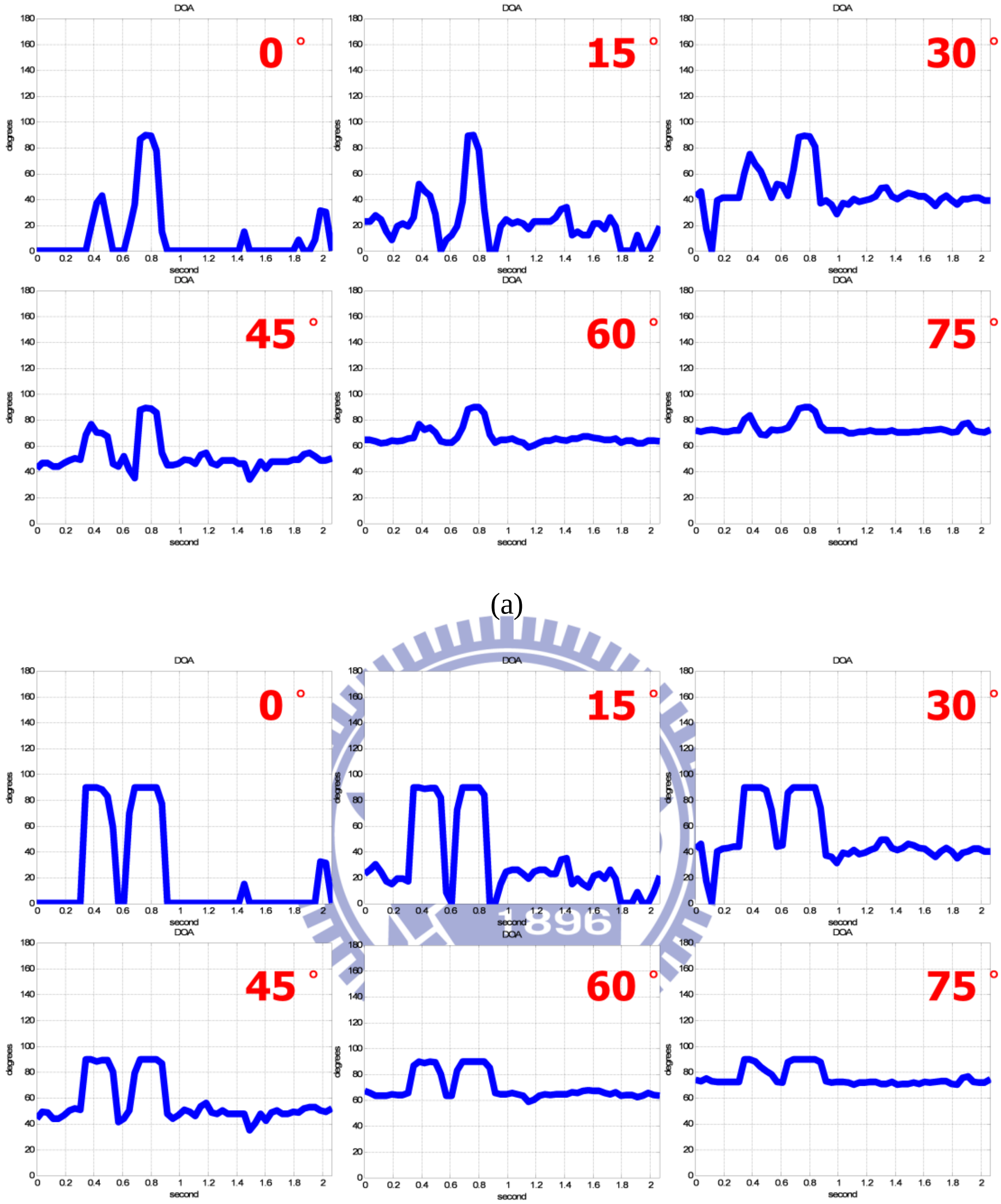
Fig. 16 (a) DOA estimation of 0 dB SNR noisy speech (white noise case) at different emitted angle, (b) DOA estimation of 12 dB SNR noisy speech (white noise case) at different emitted angle.



(a)

(b)

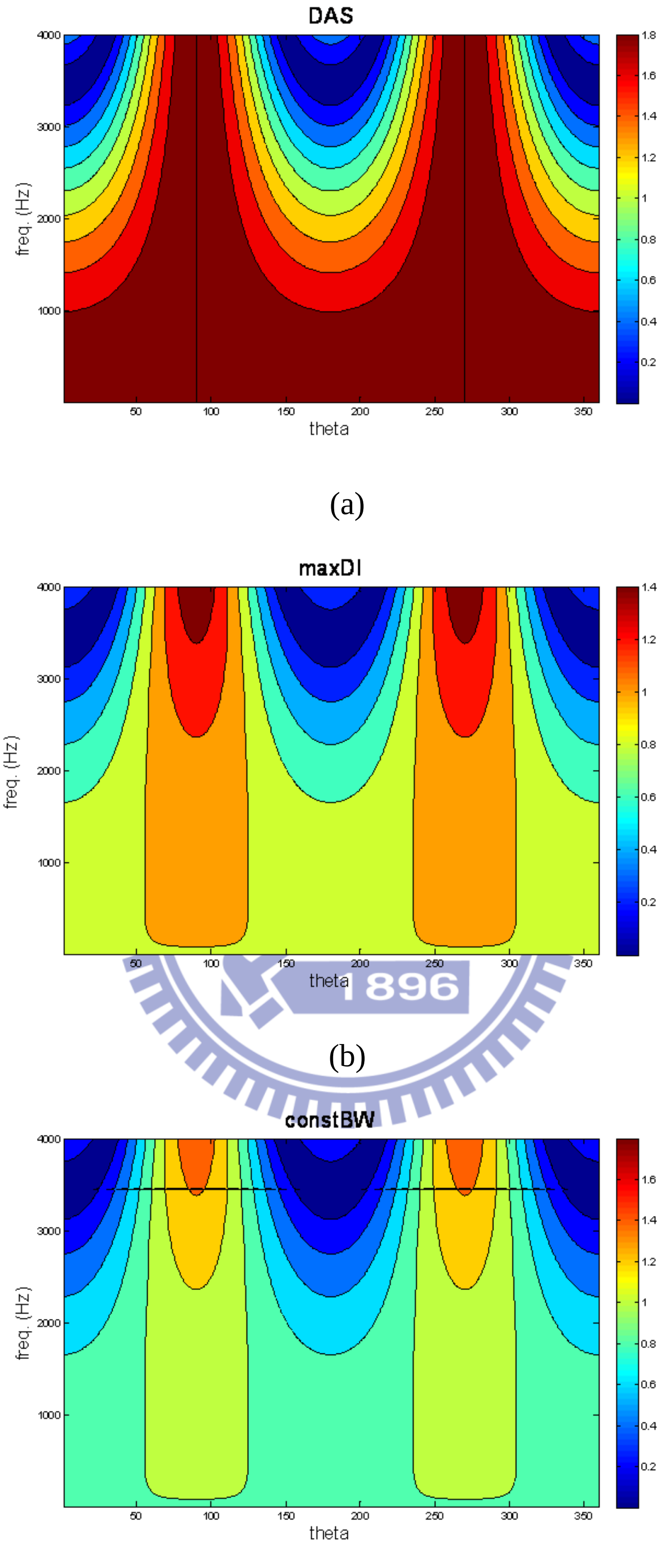
Fig. 17 (a) DOA estimation of 0 dB SNR noisy speech (babble case) at different emitted angle, (b) DOA estimation of 12 dB SNR noisy speech (babble case) at different emitted angle.



(a)

(b)

Fig. 18 (a) DOA estimation of 0 dB SNR noisy speech (exhibition case) at different emitted angle, (b) DOA estimation of 12 dB SNR noisy speech (exhibition case) at different emitted angle.

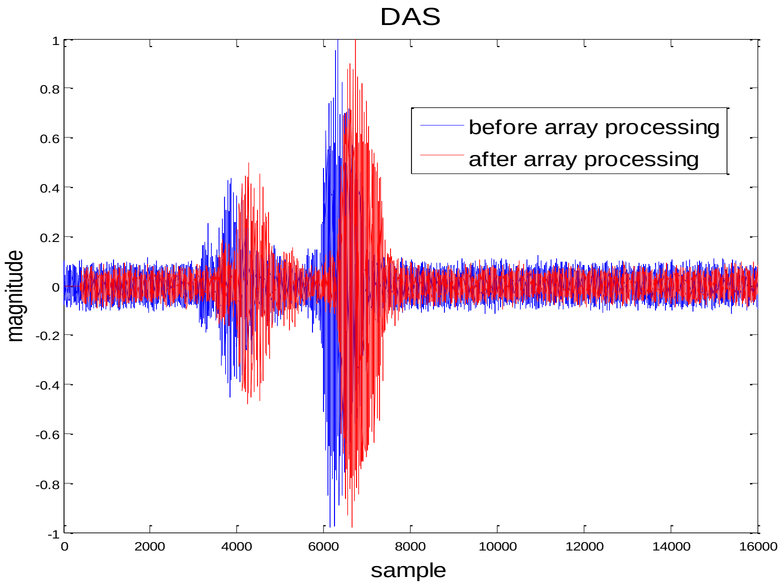


(a)

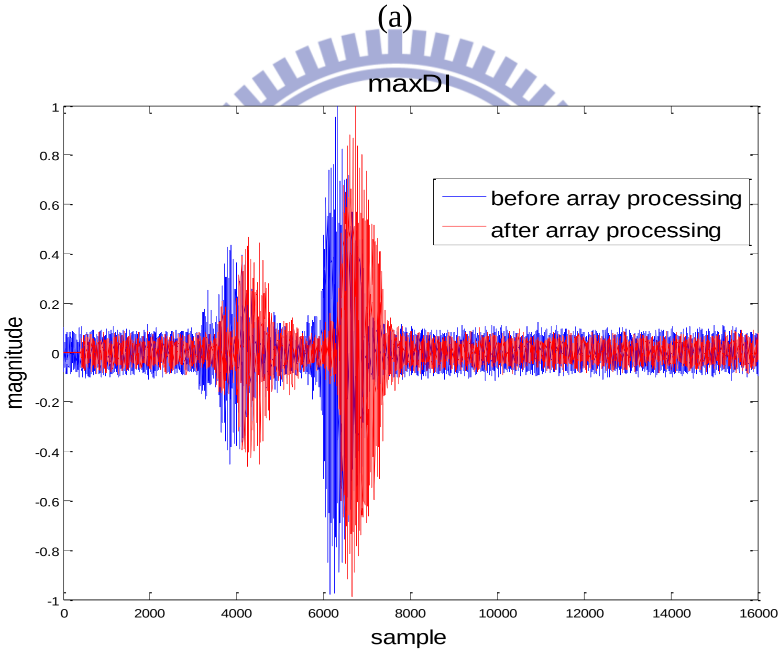
(b)

(c)

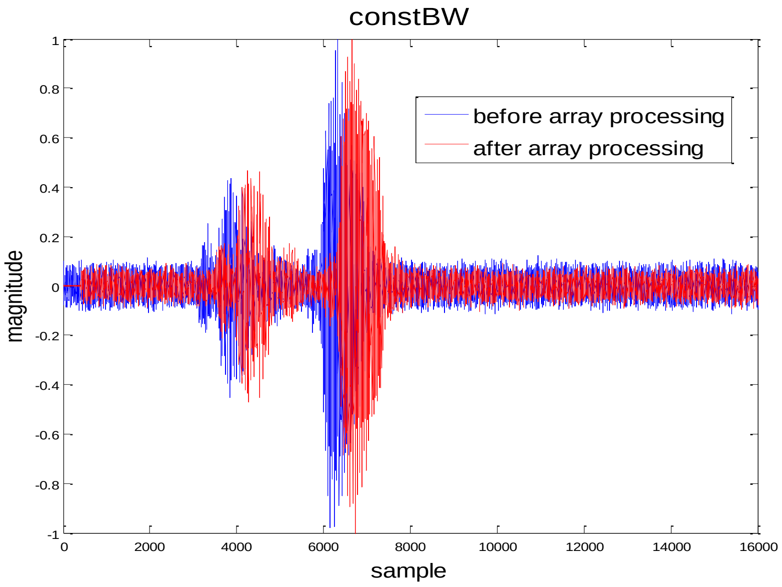
Fig. 19 (a) Beampattern of DAS, (b) Beampattern of maxDI, (c) Beampattern of constBW.


DAS
before array processing
after array processing
magnitude
sample
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
2000
4000
6000
8000
10000
12000
14000
16000

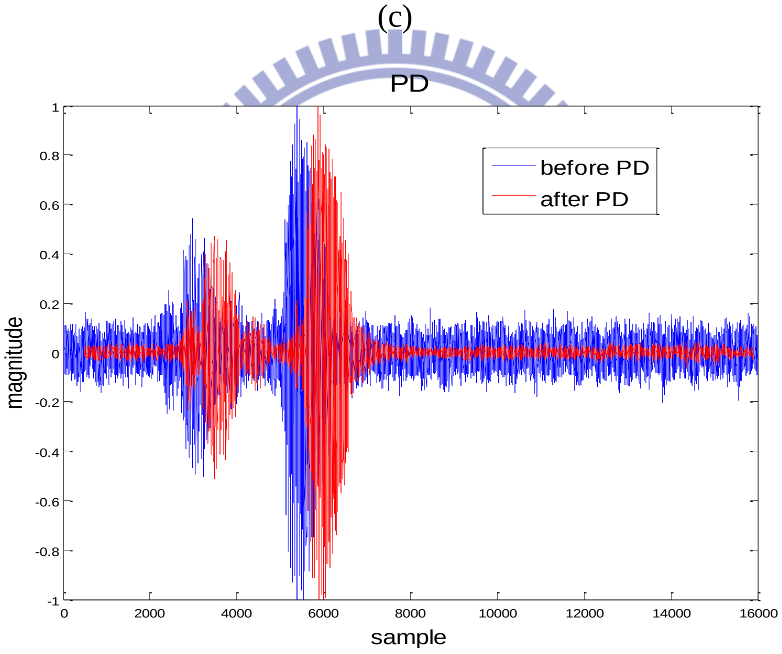
(a)


maxDI
before array processing
after array processing
magnitude
sample
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
2000
4000
6000
8000
10000
12000
14000
16000

(b)



(c)



(d)

Fig. 20 (a) Waveforms of before DAS processing and after DAS processing, (b) Waveforms of before maxDI processing and after maxDI processing, (c) Waveforms of before constBW processing and after constBW processing, (d) Waveforms of before PD processing and after PD processing.

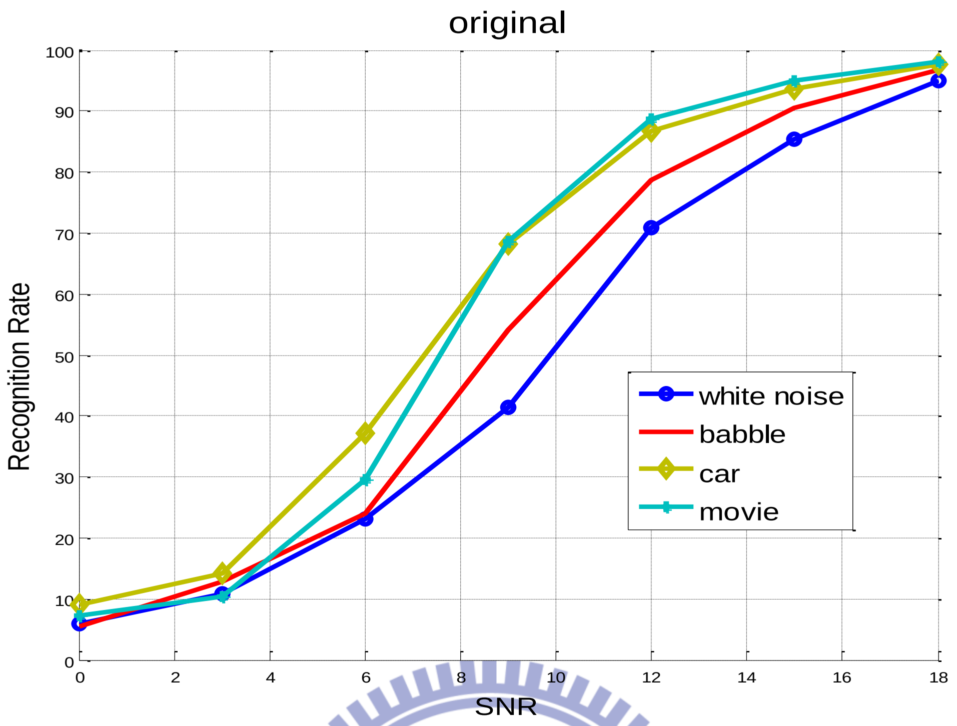
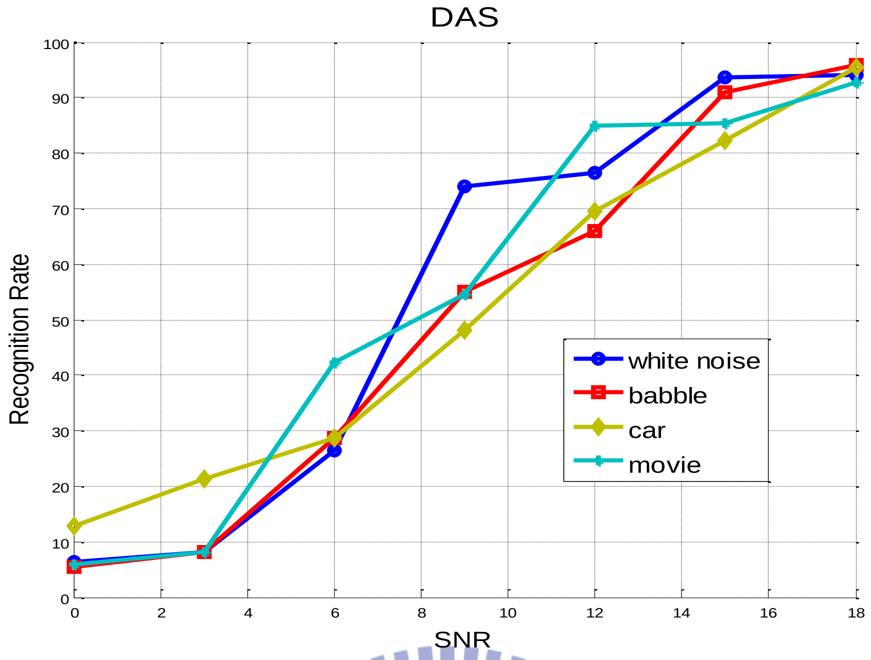
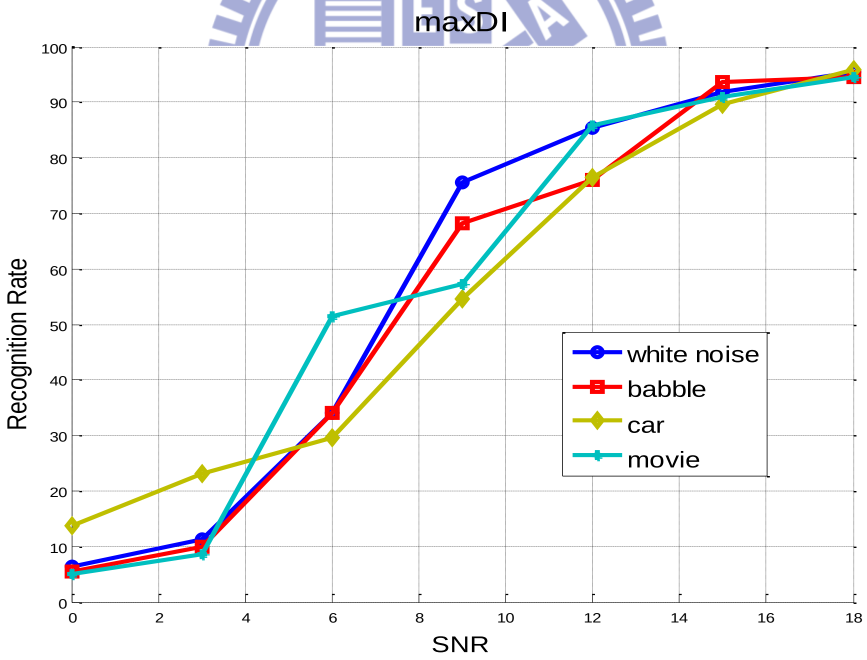


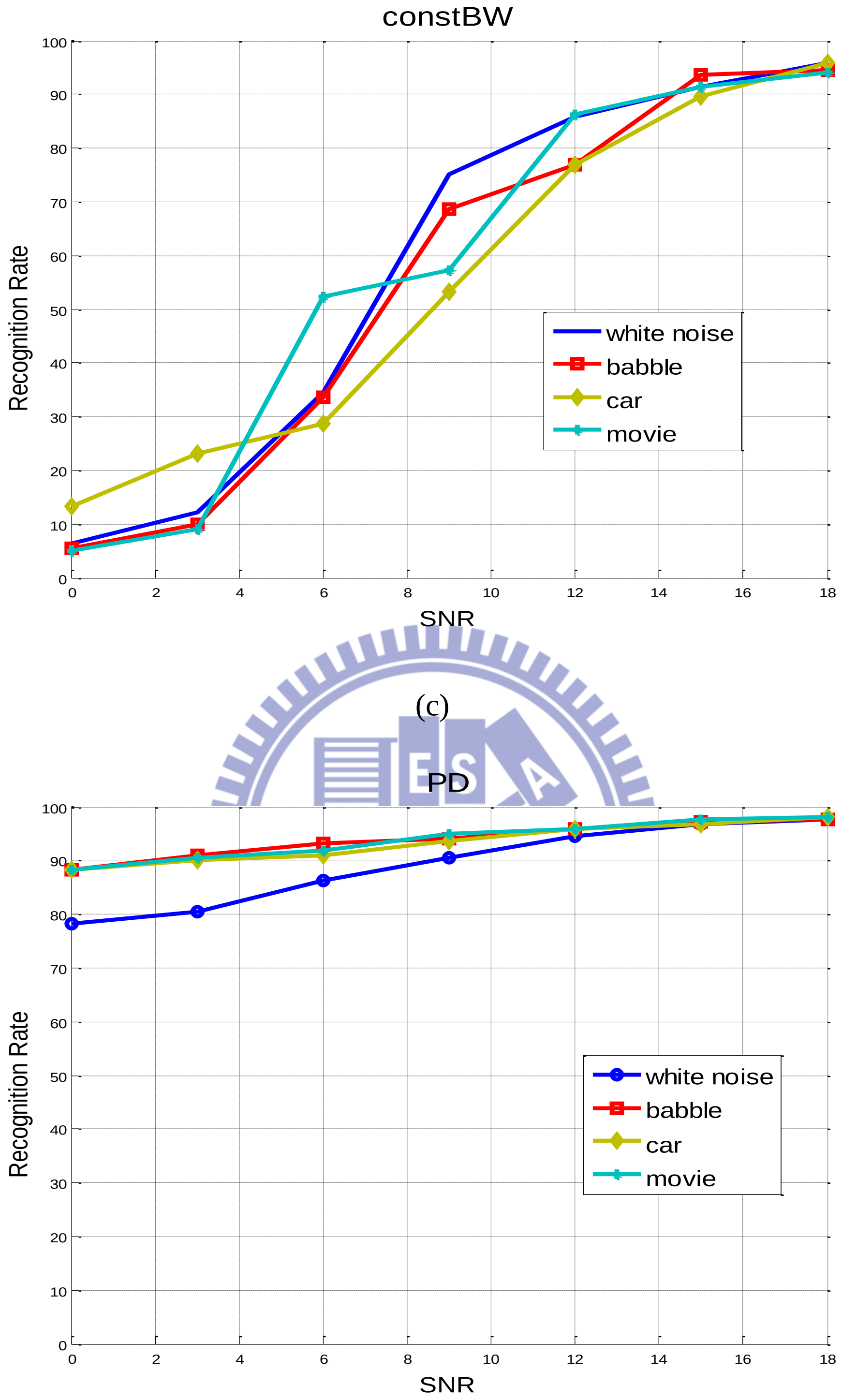
Fig. 21 RRs of clean speech polluted by white noise, babble, car noise and movie in different conditions of SNR.


DAS
Recognition Rate
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
18
SNR
white noise
babble
car
movie

(a)


maxDI
Recognition Rate
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
18
SNR
white noise
babble
car
movie

(b)



(c)

(d)

Fig. 22 (a) RRs of polluted speech enhanced by DAS, (b) RRs of polluted speech enhanced by maxDI, (c) RRs of polluted speech enhanced by constBW, (d) RRs of polluted speech enhanced by PD.

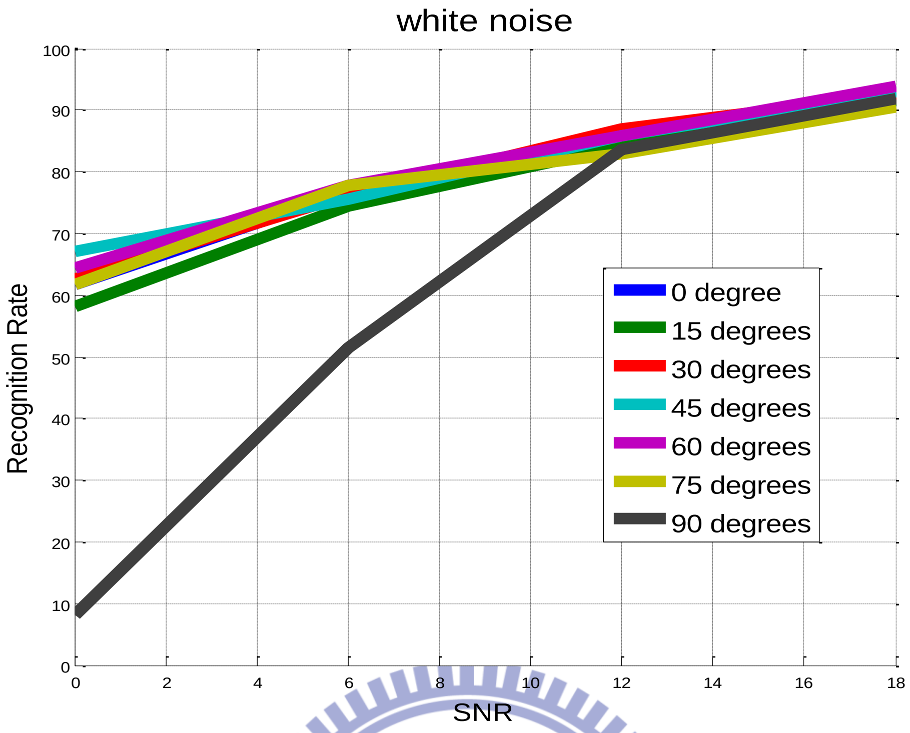


Fig. 23 RRs of noisy speech (white noise case) at different emitted angle.

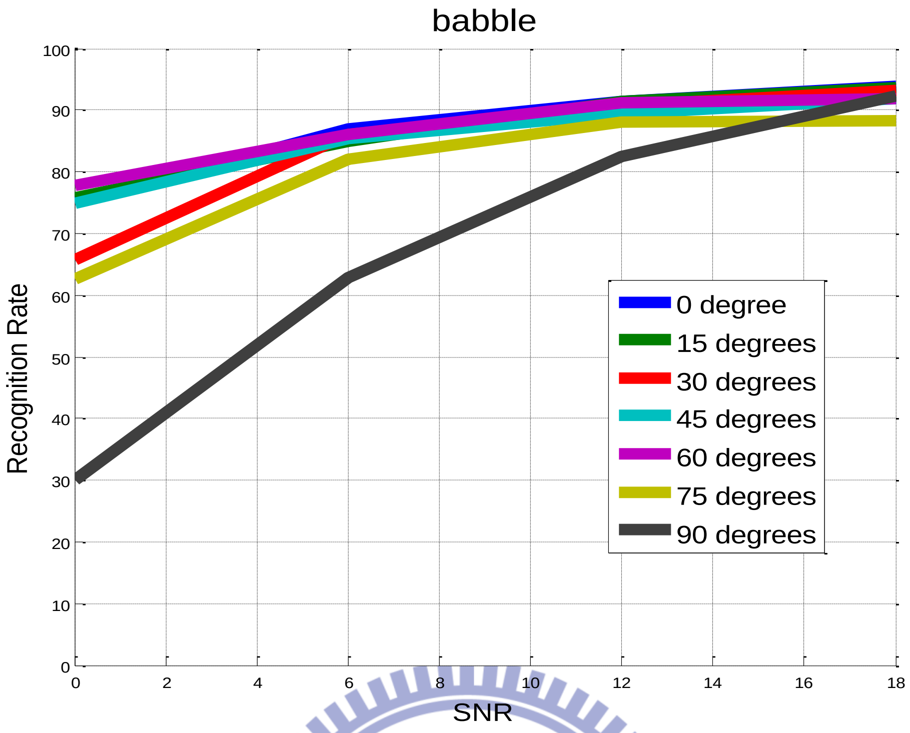


Fig. 24 RRs of noisy speech (babble case) at different emitted angle.

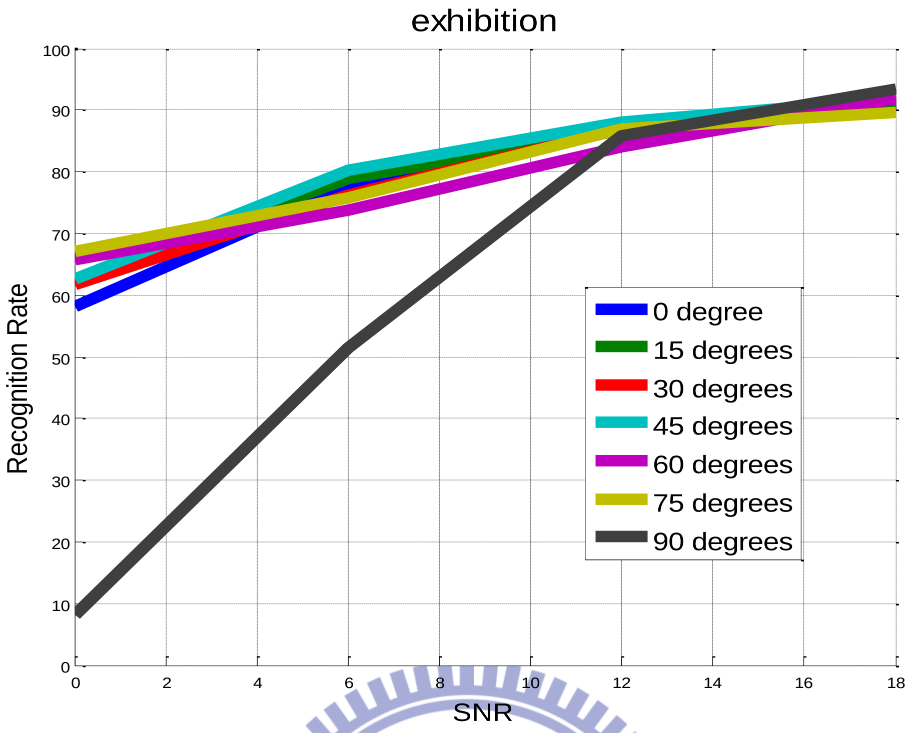


Fig. 25 RRs of noisy speech (exhibition case) at different emitted angle.

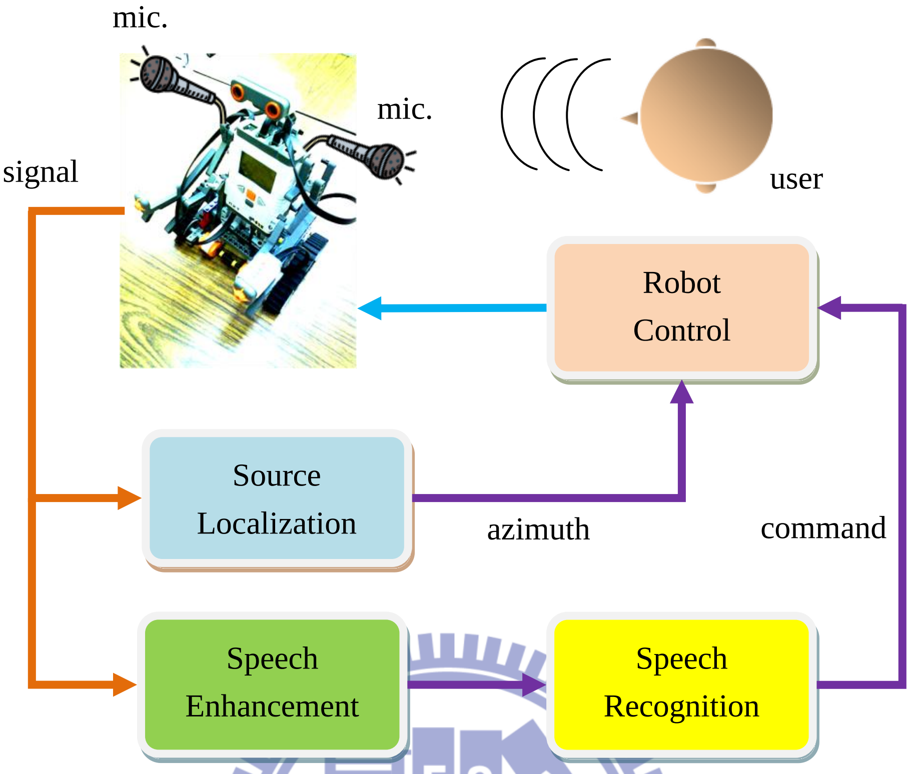


Fig. 26 Block diagram of whole system

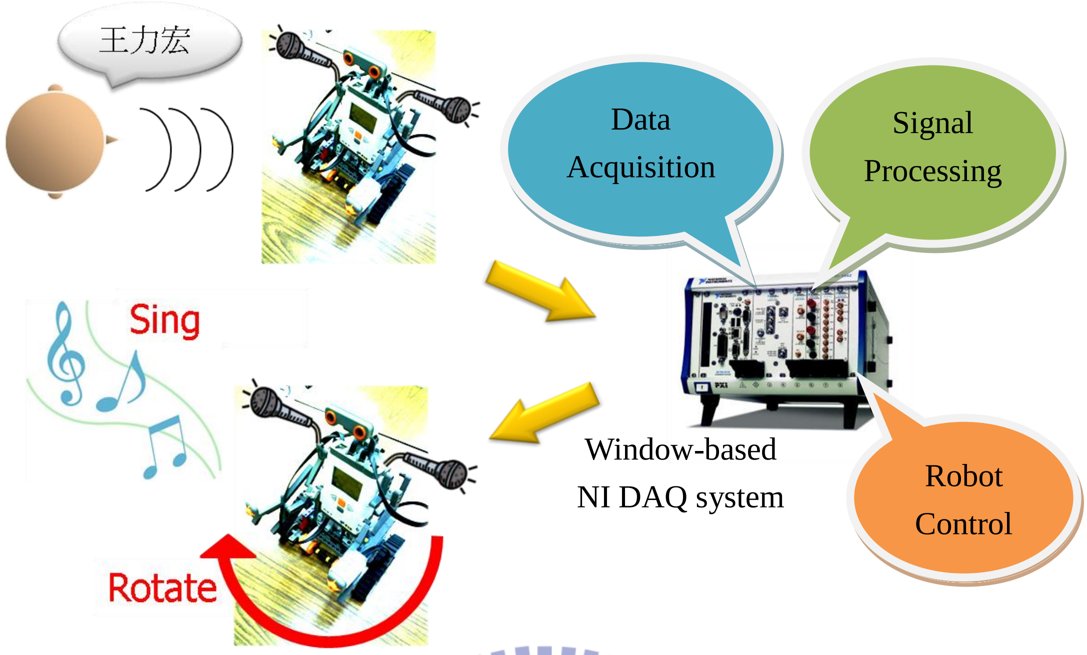


Fig. 27 Demonstration of robot implementation