國立交通大學

電機與控制工程學系

博 士 論 文

以參考訊號架構為基礎之穩健語者定位與語音純化法

Robust Reference-signal-based Speaker's Location Detection and Speech Purification

研 究 生：鄭价呈

指導教授：胡竹生 教授

中 華 民 國 九 十 五 年 六 月

以參考訊號架構為基礎之穩健語者定位與語音純化法

Robust Reference-signal-based Speaker's Location Detection and Speech Purification

研 究 生：鄭价呈　　Student：Chieh-Cheng Cheng

指導教授：胡竹生　　Advisor：Jwu-Sheng Hu

國 立 交 通 大 學

電 機 與 控 制 工 程 學 系

博 士 論 文

A Dissertation

Submitted to Department of Electrical and Control Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Electrical and Control Engineering

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

國　立　交　通　大　學

博碩士論文全文電子檔著作權授權書

（提供授權人裝訂於紙本論文書名頁之次頁用）

本授權書所授權之學位論文，為本人於國立交通大學電機與控制工程所 94 學年度第 二 學期取得博士學位之論文。

論文題目：以參考訊號架構為基礎之穩健語者定位與語音純化法

指導教授：胡竹生

■ 同意　□不同意

本人茲將本著作，以非專屬、無償授權國立交通大學與台灣聯合大學系統圖書館：基於推動讀者間「資源共享、互惠合作」之理念，與回饋社會與學術研究之目的，國立交通大學及台灣聯合大學系統圖書館得不限地域、時間與次數，以紙本、光碟或數位化等各種方法收錄、重製與利用；於著作權法合理使用範圍內，讀者得進行線上檢索、閱覽、下載或列印。

論文全文上載網路公開之範圍及時間：	
本校及台灣聯合大學系統區域網路	■ 中華民國 95 年 9 月 30 日公開
校外網際網路	■ 中華民國 95 年 9 月 30 日公開

授 權 人：鄭价呈

親筆簽名： 鄭价呈

中華民國　95　年　8　月　22　日

國　立　交　通　大　學

博碩士紙本論文著作權授權書

（提供授權人裝訂於全文電子檔授權書之次頁用）

本授權書所授權之學位論文，為本人於國立交通大學電機與控制工程所 94 學年度第 二 學期取得博士學位之論文

論文題目：以參考訊號架構為基礎之穩健語者定位與語音純化法

指導教授：胡竹生

■ 同意

本人茲將本著作，以非專屬、無償授權國立交通大學，基於推動讀者間「資源共享、互惠合作」之理念，與回饋社會與學術研究之目的，國立交通大學圖書館得以紙本收錄、重製與利用；於著作權法合理使用範圍內，讀者得進行閱覽或列印。

本論文為本人向經濟部智慧局申請專利(未申請者本條款請不予理會)的附件之一，申請文號為：_____，請將論文延至____年____月____日再公開。

授 權 人：鄭价呈

親筆簽名： 鄭价呈

中華民國　95　年　8　月　22　日

以參考訊號架構為基礎之穩健語者定位與語音純化法

研究生：鄭价呈　　指導教授：胡竹生 博士

國立交通大學電機與控制工程學系(研究所)博士班

摘要

使用麥克風陣列來改善語音擷取的品質以及偵測語者方位在語音介面相關研究上非常重要。本研究的目的在於利用一組線性麥克風陣列以及參考訊號來定位某些特定所需之語者，並且提升語音辨識的正確性。本篇論文中所提出之方法皆是利用參考訊號為基礎的系統架構來間接地解決麥克風間匹配性問題。本篇論文所提出的語者定位法則利用高斯混合模型來針對每個位置所獨具的特徵(相位差分布)作出模型化的動作。此語者定位方法可以抵抗背景雜音與反射效應，並於近場與遮蔽的環境中提供準確的語者定位結果。

為了減低運算複雜度，本篇論文提出了兩種頻域語音純化法(SPFDBB 與 FDABB)。有一法則為：若在時域中語音訊號與通道之間的關係為捲積，則對應於頻域中這兩者的關係則變為一般的乘積。但是此法則並不適用於時域的濾波器階數大於轉換到頻域所取用的窗長度之情況中。因此，本論文所提出的語音純化法便將多個窗的資料結合在一起共同處理，以期能盡可能的逼近以上之法則。此外，還提出了一個參數以提供使用者可針對通道補償以及雜訊抑制來訂定不同的權重。提供上述功能的語音純化法稱之為 SPFDBB。但若同時將太多個窗之資料統一處理，則此語音純化法便不適用於一會經常性變動的環境中。因而本論文又更進一步地提出新參數稱為 *CBVI* 來自動調整窗之個數。結合此 *CBVI* 參數與 SPFDBB 之語音純化法則稱之為 FDABB。除了上述幾個議題外，模型化誤差亦為一重要課題。對此，本論文針對一著名理論稱之為 H_∞ 理論做出相關研究，進而將其套用於所提出之兩種語音純化法中。最終，本論文利用模擬以及實際環境下的實驗結果來說明所提出方法的可行性。

Robust Reference-signal-based Speaker's Location Detection and Speech Purification

Graduate Student: Chieh-Cheng Cheng Advisor: Dr. Jwu-Sheng Hu

Department of Electrical and Control Engineering
National Chiao-Tung University

Abstract

The use of microphone array to enhance speech reception and speaker localization is very important. The objective of this work is to locate speakers of interest and then provide satisfactory speech recognition rates using a linear microphone array. The proposed approaches utilize a reference-signal-based architecture to indirectly solve a practical issue, microphone mismatch problem. Additionally, the proposed speaker's location detection method utilizes Gaussian mixture model (GMM) to model a corresponding phase difference distribution for each specific location of the speaker. The proposed localization approach is useful in the presence of background noise and reverberations. Even under near-filed and non-line-of-sight environments, the approach can still provide high detection accuracy.

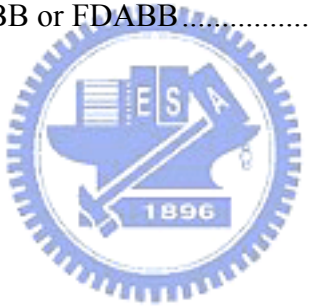
In terms of effectiveness, the proposed beamformers, soft penalty frequency-domain block beamformer (SPFDBB) and frequency-domain adjustable block beamformer (FDABB) are designed in the frequency domain. However, due to the fact that the convolution relation between channel and speech source in time-domain cannot be modeled accurately as a multiplication in the frequency domain with a finite window size, the proposed beamformers put several frames into a block to approximate the transformation. Furthermore, to put different emphases on

channel recovery and noise suppression, a parameter named soft penalty is designed. Note that for a highly variant environment, it is not suitable to allocate too many frames into one block. Therefore, the SFPDBB is extended to the FDABB with a measurement index, named *CBVI,* which enables the FDABB to automatically adjust the number of frames. An H_∞ adaptation criterion is also investigated and applied to enhance the robustness to the modeling error. Finally, the results from simulations and practical experiments are provided as proof of the effectiveness and usefulness of these proposed approaches.

Contents

Index

List of Figures



List of Tables

List of Notations

Common Notations:

M : Number of microphones

P : Number of filter taps

$\{s_1(n) \quad \cdots \quad s_M(n)\}$: Pre-recorded signal

$\{n_1(n) \quad \cdots \quad n_M(n)\}$: Environmental noise

$\{x_1(n) \quad \cdots \quad x_M(n)\}$: Online recorded noisy speech signal

$\{\hat{x}_1(n) \quad \cdots \quad \hat{x}_M(n)\}$: Training signal

$\boldsymbol{s}(n) = [\boldsymbol{s}_1(n) \quad \cdots \quad \boldsymbol{s}_M(n)]^T$: Pre-recorded signal vector

$\boldsymbol{n}(n) = [\boldsymbol{n}_1(n) \quad \cdots \quad \boldsymbol{n}_M(n)]^T$: Online recorded environmental noise vector

$\boldsymbol{x}(n) = [\boldsymbol{x}_1(n) \quad \cdots \quad \boldsymbol{x}_M(n)]^T$: Online recorded noisy speech signal vector

$\hat{\boldsymbol{x}}(n) = [\hat{\boldsymbol{x}}_1(n) \quad \cdots \quad \hat{\boldsymbol{x}}_M(n)]^T$: Training signal vector

$\boldsymbol{q} = [\boldsymbol{q}_1 \quad \cdots \quad \boldsymbol{q}_M]^T$: Filter coefficient vector

$r(n)$: Reference signal

$e(n)$: Error signal

$\hat{y}(n)$: Filtered output signal (Purified signal)

ω: Frequency index

k : Frame index

$\{S_1(\omega,k),\cdots,S_M(\omega,k)\}$: Pre-recorded speech signal

$\{N_1(\omega,k),\cdots,N_M(\omega,k)\}$: Online recorded environmental noise

$\{\hat{X}_1(\omega,k) \quad \cdots \quad \hat{X}_M(\omega,k)\}$: Training signal

$R(\omega,k)$: Reference signal

$\varepsilon_x(\omega,k)$: Error Signal

$\varepsilon_s(\omega,k)$: Distortion Signal

$\boldsymbol{Q}(\omega,k)$: Filter coefficient vector

$\hat{Y}(\omega,k)$: Filtered output signal (Purified signal)

$\mathbf{I}_L$: Identity matrix with dimension L

$\boldsymbol{R}(\omega,k)=\begin{bmatrix}R(\omega,k) & \cdots & R(\omega,k+L-1)\end{bmatrix}$: Reference signal vector

$\boldsymbol{Y}_b(\omega,k)=\begin{bmatrix}Y_b(\omega,k) & \cdots & Y_b(\omega,k+L-1)\end{bmatrix}$: Purified signal vector

f_s : Sampling rate

B_l : Length of STFT

B_i : Length of input data in a frame

B_s : Shift size of STFT

$\|.\|_\infty$: H_∞-norm

$\|.\|_2$: 2-norm

$\tilde{\boldsymbol{q}}(n)$: Coefficient vector estimation error

$\tilde{\boldsymbol{Q}}(\omega,k)$: Coefficient vector estimation error

$\hat{\boldsymbol{q}}(0)$: Initial guess

$\hat{\boldsymbol{Q}}(\omega,0)$: Initial guess

$|\boldsymbol{E}(\omega,k)|^2$: Energy of the disturbance

$e_f(n)$: Filtered output error

$e_r(n)$: Reference signal estimation error

v : Sound velocity

b : Specific band of b

J_b : Dimension in the band of b

$\boldsymbol{\lambda}(.)$: GM location model

$\boldsymbol{\lambda_0}(.)$: GM location initial model

$\boldsymbol{P}_{\hat{X}}(\omega,b,l)$: J_b-dimensional training phase difference vector

$g_i(.)$: Gaussian component densities

$\boldsymbol{\rho}(\omega,b,l)=[\rho_1(\omega,b,l) \quad \cdots \quad \rho_N(\omega,b,l)]$: Mixture weights

$\boldsymbol{\mu}(\omega,b,l)=[\boldsymbol{\mu}_1(\omega,b,l) \quad \cdots \quad \boldsymbol{\mu}_N(\omega,b,l)]$: Mean matrix in the band b at location l

$\boldsymbol{\Sigma}(\omega,b,l)=[\boldsymbol{\Sigma}_1(\omega,b,l) \quad \cdots \quad \boldsymbol{\Sigma}_N(\omega,b,l)]$: Covariance matrix in the band b at location l

$G_b(.)$: *Posteriori* probability

$\hat{l}$: Estimated location number

$p(.)$: Probability

$\mathbf{P}_{\hat{X},Q}(\omega,b,l,t)=\{\boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l),\cdots,\boldsymbol{P}_{\hat{X}}{}^{(t+Q-1)}(\omega,b,l)\}$: A sequence of Q trainingphase difference vectors

$Thd(l)$: Threshold at location l

$Up(.)$: Probability upper bound

$Lo(.)$: Probability lower bound

Chapter 1

K: K incident signals

$\{\theta_1,\cdots\theta_K\}$: K arrival angles

r_i: Radial distance from the ith sound source to the reference point of an arbitrary microphone array

L : Size of array

R_{xx} : Data correlation matrix

$\{x_1(n) \quad x_2(n) \quad \cdots \quad x_M(n)\}$: Signals received at the M microphones

$a(\theta)$: Steering vector

Chapter 2

γ : Small constant

Chapter 3

L : Number of frames in a block

$CBVI$: Changing block values index

$E\{.\}_L$: Taking an frame average over L frames

$\{\alpha_0, \alpha_1, \alpha_2\}$: Parameters of $CBVI$

$\boldsymbol{W}(\omega) = [W_1(\omega) \quad \ldots \quad W_M(\omega)]^T$: True channel response

λ : Step size

μ : Soft penalty parameter

Chapter 4

$\{u_i\}$: Input causal sequence

$\{y_i\}$: Output causal sequence

$\hat{\boldsymbol{d}}(n) = \Psi(r(1), \quad \cdots \quad r(n))$: Estimation of $\boldsymbol{d}(n)$

δ : Positive constant and lower than one

$eig(z)$: Minimum eigenvalue of z

Chapter 5

B : Geometrical volume

T : Length of the training phase difference sequence

N : Mixture number

L : Modeled location number

Q : Length of the testing sequence

$\hat{Q}(l)$: Most suitable length of testing sequences at location l

$[Q_{Lo}, Q_{UP}]$: Possible searching range of the length of the testing sequence

$\{\alpha, \beta, \gamma\}$: Parameters

Chapter 1

Introduction

Intelligent electronic devices for office, home, car, and personal applications are becoming increasingly popular. It is generally believed that the interface between human and electronic devices should not be restricted to keyboard, push-button, mouse or remote controller, touch panel, but nature language instead. One of the demands for intelligence is to enhance the convenience of operation, e.g., human-computer interaction (HCI) interfaces using speech communication. The speech-based HCI interface can be applied to robots, car computers, and teleconferencing applications. Moreover, the interface is particularly important when the use of hands and eyes puts the user in danger. For example, given concerns over driving safety and convenience, electronic systems in vehicles such as mobile phone, global positioning system (GPS), CD or VCD player, air conditioner, etc. should not be accessed by hands while driving.

However, speech communication, unlike push-button operation, suffers from unreliable problems because of environmental noises and channel distortion. The poor speech quality, acoustic echo of the far-end and near-end speech, and environmental

noises degrade the recognition performance, resulting in a low acceptance of hands free technology by consumers. Most speech acquisition system depends on the user to be physically close to the microphone to achieve satisfactory speech quality. In this case, this near-end speech significantly simplifies the acquisition problem by putting more emphasis on the desired speech signal than on environmental noises and other sound sources, and by reducing the channel effect on the desired speech signal. However, this limitation restricts the scope of applications, explaining why noise suppression approaches using single channel [1-2] and multiple microphones (i.e., microphone array) have been introduced to purify speech signals in noisy environments. Microphone-array-based approaches attempt to obtain a high speech quality without requiring the speaker to talk directly to a close-talking microphone or talk loudly to the microphone at a distance. In recent years, the microphone-array-based speech enhancement has received considerable attention as a means for improving the performance of traditional single-microphone systems.

The goal of this work is to locate speakers of interest and then provide satisfactory speech recognition rates and robustness to background noise and channel effects under both line-of-sight conditions and non-line-of-sight conditions using a uniform linear microphone array. The two main components of the general microphone-array-based speech enhancement systems are illustrated in Fig. 1-1 and the following two sections of this chapter present a brief introduction to each of them.

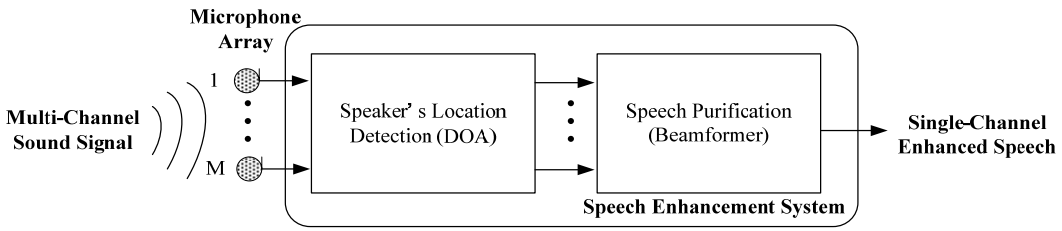


Figure 1-1 Block diagram of the general microphone-array-based speech enhancement system

1.1 Overview of Direction of Arrival Algorithms

Figure 1-2 shows the layout of a uniform linear microphone array consisting of M microphone with K incident signals from a set of arrival angles, $\{\theta_1,\cdots\theta_K\}$. The incident signals can be regarded as plane waves (far-field) or spherical waves (near-field). The definition of the far-field and the near-filed can refer to [3].

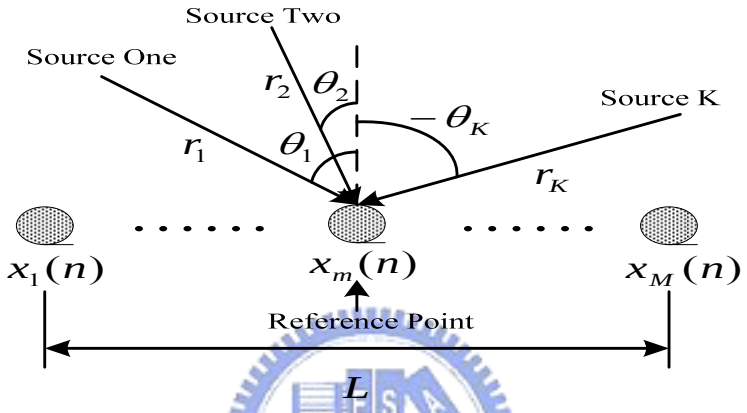


Figure 1-2 Microphone array configuration for plane or spherical wave hypothesis

where r_i is the radial distance from the ith sound source to the reference point of an arbitrary microphone array, L denotes the size of array.

Almost all microphone-array-based speech acquisition problems require a reliable active sound source location detection. Knowing the speaker's location (i.e., speaker localization) not only improves the purification results of a noisy speech signal, but also provides assistance to speaker identification. For speech enhancement applications, accurate location information of the speaker of interest or the interference sources is necessary to effectively steer the beampattern and enhance a desired speech signal, while suppressing interference and environmental noises simultaneously. Consequently, the speaker location detection is an integral part of the microphone-array-based speech enhancement system.

Location information can also be used as a guide for discriminating individual speakers in a multiple speakers' scenario. With this information available, it would be possible to automatically focus on and track a speaker of interest. Of particular interest recently is the video-conferencing system in which the speaker location is estimated for aiming a camera or series of cameras [4].

Existing microphone-array-based sound source localization algorithms may be loosely divided into three generally categories: steered-beamformer-based localization algorithms, high-resolution spectral-estimation-based localization algorithms, and time-difference of arrival (TDOA) based algorithms.

The steered-beamformer-based localization algorithms [5-7] utilize a certain beamformer to steer the array to various locations for obtaining the spatial responses and then search for a peak in the derived spatial responses. Hence, the location is derived directly from a filtered and summed data of the speech signals received at the multiple microphones. The task of computing the spatial responses for an appropriately dense set of possible locations is computationally expensive and highly dependent upon the spectral content of the sound source.

The high-resolution spectral-estimation-based localization algorithms [8-19] (eigenstructure-based DOA estimation algorithms) are based on the data correlation matrix R_{xx} derived from the signals received at the M microphones, $\{x_1(n) \quad x_2(n) \quad \cdots \quad x_M(n)\}$. These algorithms separate the eigenvectors of data correlation matrix into two parts - one is the signal subspace, and the other is the noise subspace. The steering vector $a(\theta)$ corresponds to sound source direction must be orthogonal to the noise subspace, so the inner product of steering vector and the noise subspace must be zero when it is consisted in the signal subspace. According to this phenomenon, the locations of the multiple sound sources can be simultaneous detected.

Figure 1-3 is a three-dimensional example which shows the relations between the signal subspace, the noise subspaces, and the steering vectors. The high-resolution algorithms suffer from a lack of robustness to the steering vector and environmental situations, especially reverberations, and have seen little practical use for general application.

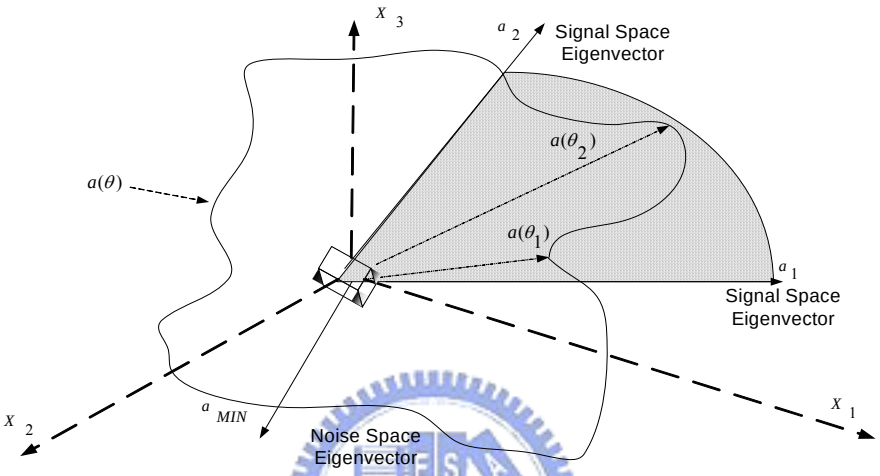


Figure 1-3 Relation of eigenspaces

The time-difference of arrival (TDOA) based procedures [20-27] locate sound source from a set of delay estimations measured across various combination of microphones. Generally, the TDOA–based procedures require the lowest computation power in these three categories. Accurate and robust time delay estimation between each microphone pair plays an important role in the effectiveness of this category. These TDOA-based procedures are sensitive to the weights of each microphone pair and various frequency bins. Notably, no mater what kinds of approaches are adopted, these three approaches cannot be applied to the fully non-line-of-sight case.

1.2 Overview of Beamformers

The second stage of the general speech enhancement system is a beamformer which deals with the suppression of interference signals and environmental noises. Normally, a beamformer (i.e., spatial filter) which uses the spatial information can generate directionally sensitive gain patterns that can be adjusted to increase sensitivity to the speaker of interest and decrease sensitivity in the direction of competing sound sources, interference signals and environmental noises. Consequently, beamformer can suppress undesired speech signals and enhance desired speech signals at the same time.

For a narrowband assumption, a beamformer is a linear combiner that produces an output by weighting and summing components of the snapshot of the received data, i.e.

$$\hat{y}(n) = \sum_{m=1}^{M} q_m^* x_m(n) = \mathbf{q}^H \boldsymbol{x}(n) \tag{1-1}$$

where $\boldsymbol{x}(n)$ is the snapshot of the received data, the superscript H denotes the Hermitian operation, and $\mathbf{q}$ is the filter coefficient vector given by $\mathbf{q} = [q_1 \quad \cdots \quad q_M]^T$. The diagram is shown in Fig. 1.4.

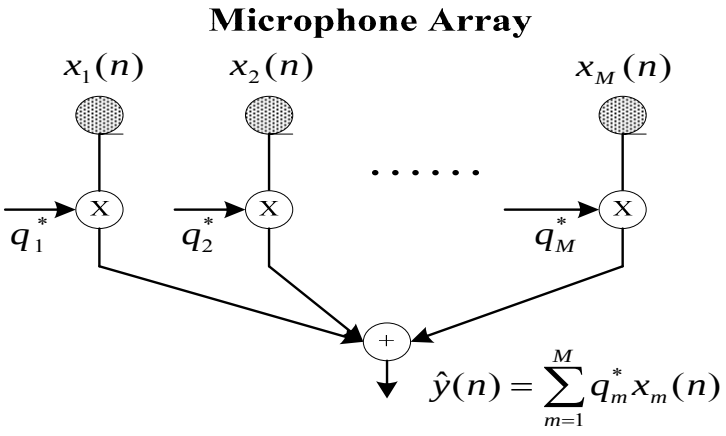


Figure 1-4 Diagram of the beamformer

However, the application in which microphone array is used is different from that of conventional array applications. This is because the speech signal has an extremely wide bandwidth relative to its center frequency. Therefore, the conventional narrowband beamformer is not suitable. Generally, for broadband signals or speech signals, a finite impulse response (FIR) filter is used on each microphone and the filter outputs are summed to generate a single-channel enhanced speech signal. For example, if a microphone array contains M microphones, each including a P taps filter, then there are MP free filter coefficients in time-domain wide-band beamformer architectures.

Existing beamformers may be approximately divided into four categories: fix-coefficients beamformers, post-filtering beamformers, subspace beamformers, and adaptive beamformers.

1.2.1 Fix-coefficients beamformers

The first category called fix-coefficients beamformers including constant directivity beamformer (CDB) [28-32] and superdirective beamformers [33-36], utilizes fixed coefficients to achieve a desired spatial response. CDB is designed to keep a constant beampattern over the all frequency bins of interest, e.g., the spatial response is approximately the same over a wide frequency band. The drawback of CDB is that the size of the microphone array is related to the lowest frequency bin and the number of microphone is relatively high. Consequently, for speech recognition applications, CDB and is impractical. Unlike CDB, superdirective beamformer attempts to minimize the power of the filtered output signal $\hat{y}(n)$ while keeping an undistorted signal response in the desired location with a finite array size. Moreover, the noise information, such as the noises' locations and noises' models can be included to improve the noise

suppression performance. Fix-coefficient beamformers generally assume the desired sound source, interference signals, and noises are slowly varying and at a known location. Therefore, these algorithms are sensitive to steering errors which limit their noises suppression performance and cause the desired signal distortion or cancellation. Furthermore, these algorithms have limited performance at enhancing a desired signal in highly reverberation environments.

1.2.2 Post-filtering beamformers

To enhance the noise suppression performance, post-filtering techniques [37-40] have been proposed. Post-filtering techniques perform post-processing to the filtered output signal $\hat{y}(n)$ by using a single channel filter, such as the Wiener filter and the normalized least mean square (NLMS) adaptation criterion. Notably, the filtered output signal is derived from a beamformer, including CDB, superdirective beamformer, and so on. Consequently, if the desired signal distortion or cancellation exists in the filtered output signal, the distortion or cancellation cannot be avoided even with a post-filtering beamformer.

1.2.3 Subspace beamformers

Subspace beamformers [41-44] [126-127] are developed based on the form of the optimal multidimensional Wiener filter. These algorithms utilize generalized singular value decomposition (GSVD) or singular value decomposition (SVD) to decompose the correlation matrix of the desired signal and the received signal when the desired signal cannot be observed. Because of the unobservable desired signal, these beamformers cannot deal with the channel distortion directly.

1.2.4 Adaptive beamformers

Instead of using fixed coefficients to suppress noises and interference signals, an adaptive beamformer [45-64] can adaptively forms its directivity beampattern to the desired signal and its null beampattern to the undesired signals. In the fix-coefficients beamformers, the null beampattern only exits when the noise's direction is known and remains unchanged. Adaptive beamformers can release this limitation and perform a better noise suppression performance than fix-coefficient beamformers. Although an adaptive beamformer can adapt itself to the change of environmental noises, the steering errors caused by microphone mismatch or DOA estimation errors deteriorate the performance. To develop a robust adaptive beamformer to cope with the steering errors is still an important issue in this field.

1.3 Outline of Proposed System

Figure 1-5 shows the block diagram of the proposed reference-signal-based speech enhancement system, which contains three main components: voice activity detection (VAD) algorithm, robust reference-signal-based speaker's location detection algorithm, and reference-signal-based frequency-domain adaptive beamformer.

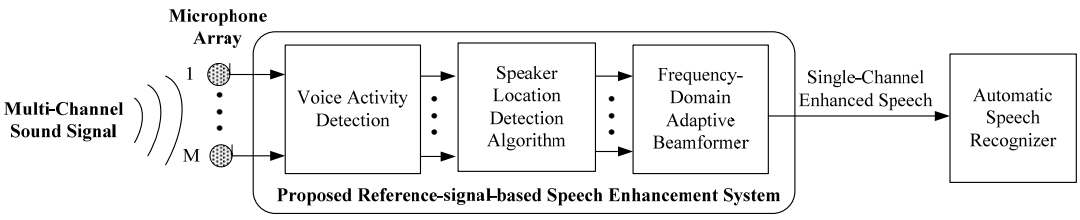


Figure 1-5 Block diagram of proposed reference-signal-based speech enhancement system

1.3.1 VAD Algorithm

An important issue in many speech processing applications is the determination of presence of speech segments in a given sound signal. To deal with this requirement, VAD was developed to detect silent and speech intervals. In the proposed reference-signal-based speech enhancement system, the VAD result drives the overall system to switch between two operational stages, the silent stage and the speech stage. Therefore, this first component in the proposed system is to provide an accurate silence detection mechanism. Figure 1-6 illustrates the flowchart of the fundamental VAD algorithm which is a two-step procedure: feature extraction and classification method.

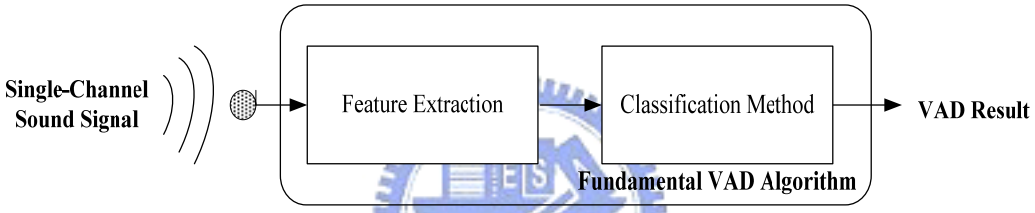


Figure 1-6 Flowchart of the fundamental VAD algorithm

Feature Extraction: Relevant features are extracted from the speech signal. To achieve a good detection of speech segments, the chosen features have to show a significant variation between speech and non-speech signals.

Classification Method: In general, a threshold is applied to the extracted features to distinguish between the speech and non-speech segments. The threshold can be a fixed value or an adjustable value. Moreover, decision rules using statistical properties [65-66] were also implemented to deal with the classification problem.

These features normally represent the variations in energy levels or spectral difference between noise and speech. There exist many discriminating features in speech detection, such as the signal energy [67-69], LPC [70-71], zero-crossing rates [72], the entropy [73-75], and pitch information [76]. Various features or feature vector,

the combinations of features, have been adopted in VAD algorithms [77-78]. To adapt to the changes of environmental noises or various noise characteristics, noise estimation method during non-speech periods should be added into the fundamental VAD algorithm [79-80]. The algorithm [81] is evaluated in Chapter 6 under vehicular and indoor environments. Based on the experimental results, the VAD algorithm in [81] is suitable for implementing the proposed reference-signal based speech purification system.

1.3.2 Reference-signal-based Speaker's Location Detection Algorithm

Because conventional sound source localization algorithms suffer from the uncertainties of environmental complexity and noise, as well as the microphone mismatch, most of them are not robust in real practice. Without a high reliability, the acceptance of speech-based HCI would never be realized. This dissertation presents a novel reference-signal-based speaker's location detection approach and demonstrates high accuracy within a vehicle cabinet and an office room using a single uniform linear microphone array.

Firstly, to perform single speaker's location detection, the proposed approach utilize Gaussian mixture models (GMM) to model the distributions of the phase differences among the microphones caused by the complex characteristic of room acoustic and microphone mismatch. The individual Gaussian component of a GMM represents some general location-dependent but content and speaker-independent phase difference distributions. Moreover, according to the experimental results in Chapter 6, the scheme performs well not only in non-line-of-sight cases, but also when the speakers are aligned toward the microphone array but at difference distances from it. This strong

performance can be achieved by exploiting the fact that the phase difference distributions at different locations are distinguishable in a non-symmetric environment.

However, because of the limitation of VAD algorithm, an unmodeled speech signal might trigger the algorithm and drive the system to a wrong stage. This unexpected signal, which is not emitted from one of the modeled locations, may come from radio broadcasting of the in-car audio system and the speaker's voices from unmodeled locations. Therefore, this dissertation proposes a threshold adaptation method to provide high accuracy in locating multiple speakers and robustness to unmodeled sound source locations.

1.3.3 Reference-signal-based Frequency-Domain Beamformer

This dissertation proposes two frequency-domain beamformers based on reference signals. They are *soft penalty frequency-domain block beamformer* (SPFDBB) and *frequency-domain adjustable block beamformer* (FDABB). Compared with the conventional reference-signal-based time-domain adaptive beamformers using NLMS adaptation criterion, these frequency-domain methods can significantly reduce the computational effort in speech recognition applications. Like other reference-signal-based techniques, SPFDBB and FDABB minimize microphone mismatch, desired signal cancellation caused by reflection effects and resolution due to the array's position. Additionally, these proposed methods are appropriate for both near-field and far-field environments. Generally, the convolution relation between channel and speech source in time-domain cannot be modeled accurately as a multiplication in the frequency- domain with a finite window size, especially in speech recognition applications. SPFDBB and FDABB can approximate this multiplication by treating several frames as a block to achieve a better beamforming result. Moreover,

FDABB adjusts the number of frames in a block on-line to cope with the variation of characteristics in both speech and interference signals. In Chapter 6, a better performance is found to be achievable by combining SPFDBB or FDABB with a speech recognition mechanism.

For a speech recognition application, another important issue in real-time beamforming of microphone arrays is the inability to capture the whole acoustic dynamics via a finite-length of data and a finite number of array elements. For example, the source signal coming from the side-lobe through reflection presents a coherent interference, and the non-minimal phase channel dynamics may require an infinite data to achieve perfect equalization (or inversion). All these factors appear as uncertainties or un-modeled dynamics in the receiving signals. Therefore, the proposed system attempts to adopt the H_∞ adaptation criterion, which does not require a *priori* knowledge of disturbances and is robust to the modeling error in a channel recovery process. The H_∞ adaptation criterion is to minimize the worst possible effects of the disturbances including modeling errors and additive noises on the signal estimation error. Consequently, using the H_∞ adaptation criterion can further improve the recognition performance.

It should be emphasized that DOA and beamformer are generally treated as two independent components and discussed respectively in general speech enhancement systems. However, the proposed reference-signal-based speaker's location detection algorithm and frequency-domain beamformer can be potentially integrated because they perform in the same operational architecture. Please refer to Chapter 7 for more detail.

1.4 Contribution of this Dissertation

The contribution of this dissertation is to propose and implement innovative algorithms for sound source localization and speech purification. Although a considerable number of studies have been made on these two fields over the past 40 years, only few attempts have so far been made at simultaneously targeting on practical issues such as microphone mismatch, near-field and far-filed, channel dynamics recovery, desired signal cancellation, and the resolution effect for speech purification. Further, for the sound source localization, important issues in practice are the non-line-of-sight and line-of-sight, reverberation, microphone mismatch, and noisy environment. This dissertation proposes two frequency domain beamformers, namely SPFDBB and FDABB, and speaker's location detection approach to simultaneously overcome the issues mentioned above.

1. SPFDBB and FDABB can flexibly adjust the emphasis on the channel dynamics recovery and noise suppression. Moreover, SPFDBB and FDABB reduce the computational effort significantly, and deal with the problem that the convolution relation between channel and speech source in time-domain cannot be modeled accurately as a multiplication in the frequency domain with a finite window size.

2. To cope with the variation of room acoustics, a frame number adaptation method is proposed using an index named *CBVI* in FDABB.

3. An H_∞ adaptation criterion is applied to the proposed SPFDBB and FDABB to reduce the effect of modeling error, which is common in the adaptive beamformers.

4. The proposed multiple speakers' locations detection approach is able to provide the suitable length of testing sequences and thresholds. Therefore, it can obtain the high

accuracy on detecting speaker's location and reduce the error caused by unmodeled locations and the overlapped speech segments.

1.5 Dissertation Organization

This chapter provides a brief introduction to the general microphone-array-based speech enhancement system, including the overviews of DOA algorithms, and beamformers. This chapter also briefly discusses three main components in the proposed reference-signal-based speech enhancement system. Chapter 2 introduces the reference-signal-based time-domain adaptive beamformer using NLMS adaptation criterion. Chapter 3 presents SPFDBB and FDABB using NLMS adaptation criterion and analyzes the computational efforts of the two proposed frequency-domain and the time-domain beamformers. Chapter 4 studies the robustness of the H_∞ adaptation criterion. Chapter 5 presents the proposed reference-signal-based speaker's location detection approach for single speaker's and multiple speakers' locations detection. Chapter 6 shows the simulation as well as the experimental results in real environment. Chapter 7 gives some concluding remarks and avenues for future research.

Chapter 2

Reference-signal-based Time-domain Adaptive Beamformer

2.1 Introduction

Speech enhancement systems are now becoming increasingly important, especially with the development of automatic speech recognition (ASR) applications. Although various solutions have been proposed to reduce the desired signal cancellation, particularly desired speech, in noisy environments, the recognition rate is still not satisfactory. Earlier approaches, such as delay-and-sum (DS) beamformer [82], Frost beamformer [45], and generalized sidelobe cancellation (GSC) [46], are only good in ideal cases, where the microphones are mutually matched and the environment is a free space. The causes of performance degradation include array steering vector mismodeling due to imperfect array calibration [47], and the channel effect (e.g., near-field or far-field problem [83], environment heterogeneity [84] and source local scattering [85]). To manage these limitations, the most common linearly constrained minimum variance (LCMV)-based techniques [33] and [57] have been developed to reduce uncertainty in the sound signal's look direction. However, these approaches are

limited by scenarios with look direction mismatch. Henry [58] employed the white noise gain constraint to overcome the problem of arbitrary steering vector mismatch. Unfortunately, no clear guidelines are available for choosing the parameters. Hoshuyama *et al.* [59] proposed two robust constraints on blocking matrix design. Cannot *et al.* [86] proposed a new channel estimation method for standard GSC architecture in the frequency domain, but loud noises, and in particular circuit noise, would heavily degrade its channel estimation accuracy. Sergiy A. *et al.* [60] proposed an approach based on optimizing the worst-case performance to overcome an unknown steering vector mismatch. However, a worst case is defined as a small random perturbation, which may not suitable for general cases.

Dahl *et al.* [61] proposed the reference-signal-based time-domain adaptive beamformer using NLMS adaptation criterion to perform indirect microphone calibration and to minimize the speech distortion due to the channel effect by using pre-recorded speech signals and a reference signal. Moreover, Yermeche *et al.* [62] and Low *et al.* [63] also utilized the reference signal to estimate the source correlation matrix and calibration correlation vector. These methods in [61-63] are fundamentally the same except that the works in [62-63] do not require a VAD. However, VAD is useful for finding speaker's location and enabling the speech purification system and speech recognizer. Therefore, the methods in [62-63] cannot offer any profit in ASR applications.

The following section describes the system architecture of the reference-signal-based time-domain adaptive beamformer and the corresponding dataflow. It also presents how to derive the pre-recorded and reference signals. Finally, conclusions are given in Section 2.3.

2.2 System Architecture

The architecture of the reference-signal-based time-domain adaptive beamformer is shown in Fig. 2-1. A speech signal passing through multi-acoustic channels is monitored at spatially separated sensors (a microphone array). Two kinds of signals, the pre-recorded signal, $\{s_1(n) \quad \cdots \quad s_M(n)\}$, and the reference signal, $r(n)$, are necessary before executing the reference-signal-based time-domain adaptive beamformer. Let M denotes the number of microphones. A set of pre-recorded speech signals are collected by placing a loudspeaker or a person in the desired position, and by letting the loudspeaker emit or the person speak a short sentence when the environment is quiet. Therefore, the pre-recorded speech signals provide *a priori* information between speakers and the microphone array. Additionally, the reference signal is acquired from the original speech source that emitted from the loudspeaker or using another microphone located near the person to record the speech. In practice, the loudspeaker or the person should move around the desired position slightly to obtain an effective recording. For example, Fig. 2-2 illustrates a vehicular environment with a headset microphone and a microphone array. The person is right on the desired location and speaks several sentences when the environment is quiet. The sentences are simultaneously recorded by the headset microphone and the microphone array. The speech signal collected by the headset microphone and the microphone array are called the reference signal and pre-recorded speech signals individually. Notably, the user does not need the headset microphone during the online applications.

After collecting the pre-recorded speech signals and the reference signal, the complete procedures of the reference-signal-based time-domain beamformer are divided in two stages, the silent stage and the speech stage, through the result of VAD algorithm.

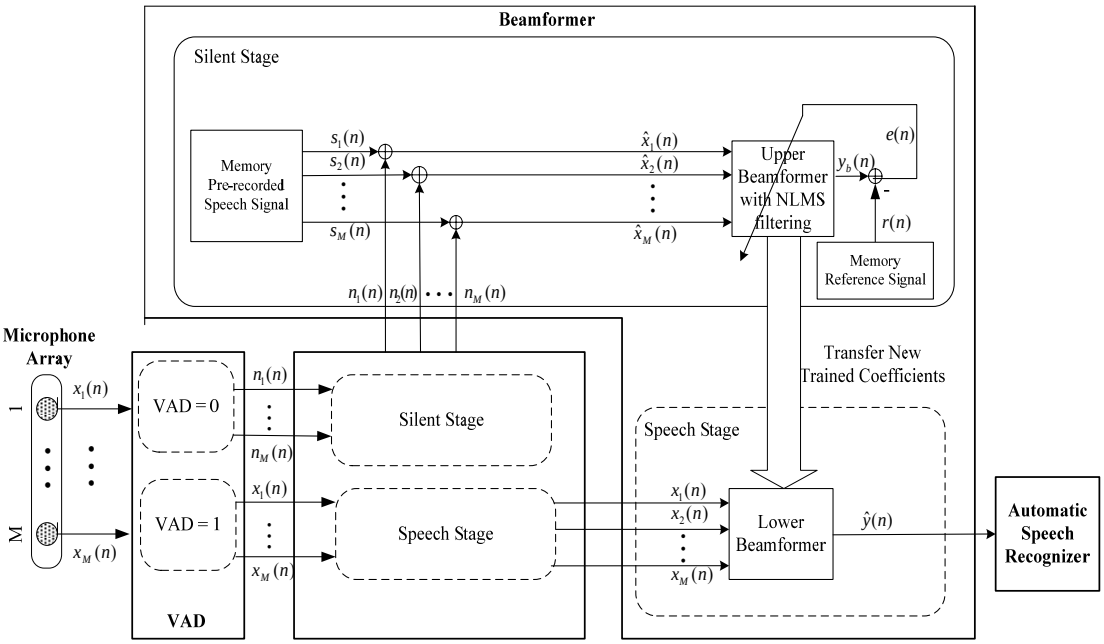


Figure 2-1 System architecture of the reference-signal-based time-domain adaptive beamformer

Figure 2-2 Installation of the array and headset microphone inside a vehicle

In other words, the VAD result decides whether to switch the system to the silent stage or speech stage. First, if VAD result equals to zero which means no speech signal contains in the received signals, $\{x_1(n) \cdots x_M(n)\}$ (i.e. the received signals are totally environmental noises denoted as $\{n_1(n) \cdots n_M(n)\}$), then the system is switched to the first stage: the silent stage. Given that the environmental noises are assumed to be additive, the acoustic behavior of the speech signal received when a

speaker is talking in a noisy environment can be expressed as a linear combination of the pre-recorded speech signal and the environmental noises. Therefore, in this stage, the system combines the online recorded environmental noise, $\{n_1(n) \quad \cdots \quad n_M(n)\}$, with the pre-recorded speech database, $\{s_1(n) \quad \cdots \quad s_M(n)\}$, to construct the training signals, $\{\hat{x}_1(n) \quad \cdots \quad \hat{x}_M(n)\}$, and to performs NLMS adaptation criterion to derive the filter coefficient vectors. Notably, the filter coefficient vectors are updated via the reference signal and the training signal thus implicitly solving the calibration.

Secondly, if the received sound signal is detected as containing speech signal, then the system is switched to the second stage, called speech stage. In this stage, the filter coefficient vectors obtained by the first stage are applied to the lower beamformer to suppress noises and enhance the speech signal in the speech stage. Finally, the single-channel purified speech signal $\hat{y}(n)$ is transformed to the frequency domain and then sent to the automatic speech recognizer. Because the variation between pre-recorded speech signals and the reference signal contains useful information about the dynamics of channel, electronic equipments uncertainties, and microphones' characteristics, the method potentially outperforms other un-calibrated algorithms in real applications. Figure 2-3 presents the flowchart of the reference-signal-based time-domain adaptive beamformer.

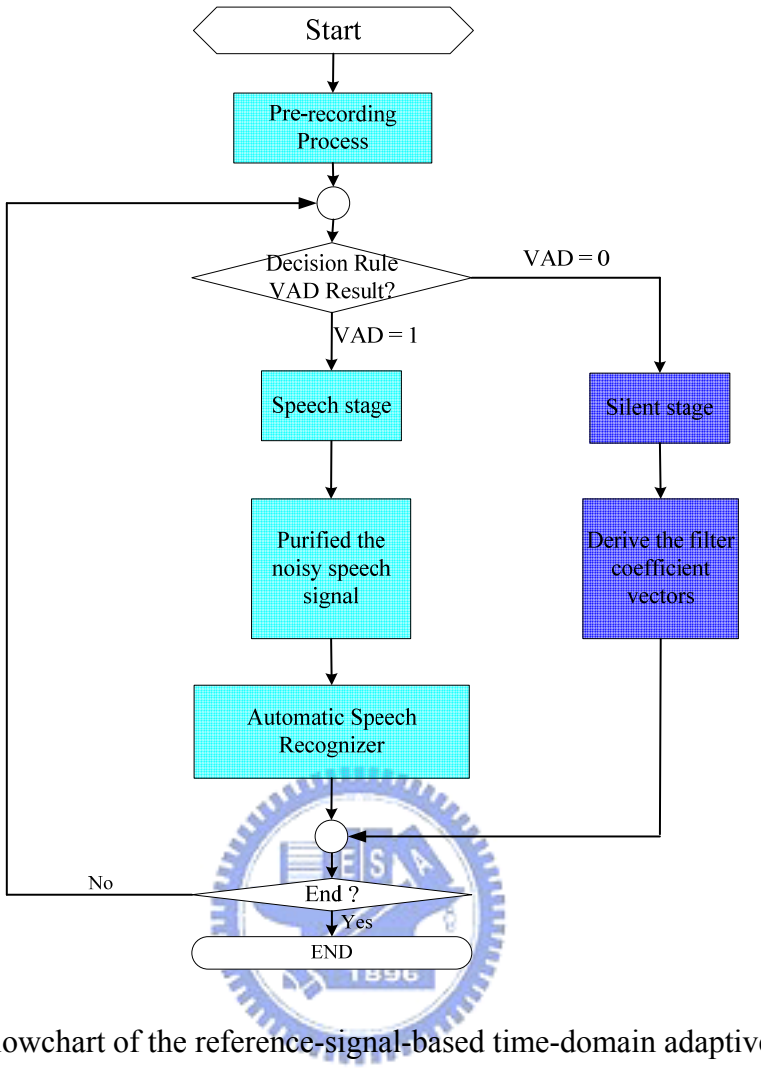


Figure 2-3 Flowchart of the reference-signal-based time-domain adaptive beamformer

While the speaker in the desired location is silent, the formulation of referenced-signal-based time-domain beamformer can be expressed as the following linear model:

$$r(n) = \boldsymbol{q}^T \hat{\boldsymbol{x}}(n) + e(n) = \boldsymbol{q}^T \left(\boldsymbol{s}(n) + \boldsymbol{n}(n) \right) + e(n) \qquad (2\text{-}1)$$

where the superscripts T denotes the transpose operation and $e(n)$ is the error signal in the time domain. Notice that italics fonts represent scalars, bold italics fonts represent vectors, and bold upright fonts represent matrices in this dissertation. Let the parameter P denote the FIR taps of the each estimated filter, and then

$\boldsymbol{s}(n) = [\boldsymbol{s}_1(n) \quad \cdots \quad \boldsymbol{s}_M(n)]^T$ denotes the pre-recorded signal vector;

$\boldsymbol{n}(n) = [\boldsymbol{n}_1(n) \quad \cdots \quad \boldsymbol{n}_M(n)]^T$ denotes the $MP \times 1$ online recorded environmental noise vector;

$\hat{\boldsymbol{x}}(n) = [\hat{\boldsymbol{x}}_1(n) \quad \cdots \quad \hat{\boldsymbol{x}}_M(n)]^T$ denotes the $MP \times 1$ training signal vector;

$\boldsymbol{q} = [\boldsymbol{q}_1 \quad \cdots \quad \boldsymbol{q}_M]^T$ denotes the $MP \times 1$ filter coefficient vector of the time-domain beamformer that we intent to estimate.

The corresponding vectors of the signals defined above are,

$\hat{\boldsymbol{x}}_i(n) = [\hat{x}_i(n) \quad \cdots \quad \hat{x}_i(n-P+1)]$; $\boldsymbol{s}_i(n) = [s_i(n) \quad \cdots \quad s_i(n-P+1)]$;

$\boldsymbol{n}_i(n) = [n_i(n) \quad \cdots \quad n_i(n-P+1)]$; $\boldsymbol{q}_i = [q_{i1} \quad \cdots \quad q_{iP}]$.

The well-known normalized LMS solution obtained by minimizing the power of error signal is represented in Eq. (2-2)

$$\boldsymbol{q}(n+1) = \boldsymbol{q}(n) + \frac{\hat{\boldsymbol{x}}(n)}{\gamma + \hat{\boldsymbol{x}}(n)^T \hat{\boldsymbol{x}}(n)} e(n) \tag{2-2}$$

where γ is a small constant included to ensure that the update term does not become excessively large when $\hat{\boldsymbol{x}}(n)^T \hat{\boldsymbol{x}}(n)$ temporarily become small. The purified signal can be calculated by

$$\hat{y}(n) = \boldsymbol{x}^T(n)\boldsymbol{q}(n) \tag{2-3}$$

where $\boldsymbol{x}(n) = [\boldsymbol{x}_1(n) \quad \cdots \quad \boldsymbol{x}_M(n)]^T$ is the $MP \times 1$ online recorded noisy speech signal vector acquired by the microphone array, and

$\boldsymbol{x}_i(n) = [x_i(n) \quad \cdots \quad x_i(n-P+1)]$.

2.3 Summary

This chapter presents the reference-signal-based system architecture which implicitly contains the information of the channel effect and microphones' characteristics. This architecture implicitly obtains the acoustic behavior from the desired location to microphone array and reduces the efforts of directly performing microphone calibration and channel inversion. Furthermore, it can be applied on both near-field and far-field situations which offers a significant advantage in speaker localization and beamformer algorithms. Extension of this idea to further improve the ASR rates will be described in the following chapters. Moreover, a novel speaker's location detection algorithm based on the reference-signal-based architecture is also proposed. In addition, Chapter 6 compares the performance of the reference-signal-based time-domain adaptive beamformer and other well-known non-reference-signal-based beamformers to show the effectiveness of the proposed method.

Chapter 3

Reference-signal-based Frequency-domain Adaptive Beamformer

3.1 Introduction

The required computational effort could be large when applying a large FIR filter coefficients, e.g., 256 to 512 taps in the time-domain adaptive beamformer introduced in the previous chapter. For subsequent ASR operation, another effort to compute Discrete Fourier transform (DFT) is required. One possible way to simplify the computational complexity is to compute the beamformer directly in the frequency domain because ideally the large FIR taps can be replaced by a simple multiplication at each frequency bin (e.g., the FIR filter with dimension of $MP \times 1$ is represented by filter coefficient vector $M \times 1$ in the frequency domain where M is the microphone number and P denotes the FIR taps). Moreover, the purified speech signal after a frequency-domain beamformer can be sent directly to the ASR. As explained later in this chapter and Chapter 6, the saving of computational effort is quite significant.

In a reference-signal-based beamformer, coefficients adjustment has two objectives: to minimize the interference signal and noises, and to equalize the channel effect (e.g. room acoustics). Channel equalization is important for ASR since the channel distortion may greatly reduce the recognition rate.

By formulating the same problem in the frequency domain, channel distortion can be emphasized using *a priori* information. In this chapter, a penalty function is incorporated into the performance index to calculate the filter coefficient vectors. This proposed algorithm is called SPFDBB.

A real-time frequency-domain beamformer is necessary to apply the short time Fourier transform (STFT). However, the corresponding window size of the STFT has to be fixed by the training data settings in ASR. For an environment with longer impulse response duration, the convolution relation between channel and speech source in time-domain cannot be modeled accurately as a multiplication in the frequency-domain with a finite window size. Therefore, the finite window size may not provide enough information for the coefficient adjustment and could not fit the assumptions that filter coefficient vector and the error signal should be independent to the input data in the NLMS adaptation criterion. In this case, SPFDBB takes the frame average over several frames as a block to improve the approximation of the linear model shown in Eq. (2-1). In other words, a block of windowed data is simultaneously adopted to calculate the filter coefficient vectors in the SPFDBB algorithm. The number of frames in a block is denoted as the frame number L. Intuitively, a large frame number could enhance the accuracy of the filter coefficient estimation. However, if the room acoustic dynamic changes suddenly, the channel response is difficult to be adjusted quickly when taking a large frame number for the updating process. Furthermore, the requirement of the length of the processed data would be too large when a large value

of L is chosen. Therefore, SPFDBB is further enhanced by allowing the frame number to be adapted on-line. A novel index called changing block values index ($CBVI$) is defined as the basis for adjusting the frame number. The overall algorithm is called FDABB.

The remainder of this chapter is organized as follows. Section 3.2 describes the system architecture and the corresponding dataflow. Section 3.3 represents SPFDBB, one of the reference-signal-based frequency-domain adaptive beamformers which utilizes NLMS adaptation criterion. Section 3.4 introduces the other proposed method, FDABB, and also analyzes the computing efforts of SPFDBB, FDABB and the reference-signal-based time-domain beamformer. Two frequency-domain performance indexes, the source distortion ratio (SDR) and the noise suppression ratio (NSR) are defined in Section 3.5. Finally, conclusions are given in Section 3.6.

3.2 System Architecture

Figure 3-1 shows the overall system architecture. The pre-recorded speech signals, $S_1(\omega,k),\cdots,S_M(\omega,k)$, and the reference signal, $R(\omega,k)$, can be recorded by the same way described in Chapter 2 when the environment is quiet. After acquiring the pre-recorded speech signal and the reference signal, the overall system automatically executes between the silent and speech stages based on the VAD result.

If the result of VAD equals to zero which means no speech signal contained in the received signal, $\{x_1(n) \quad \cdots \quad x_M(n)\}$, then the system is switched to the silent stage in which the adaptation of FDABB or SPFDBB is turned on. The filter coefficient vectors of FDABB or SPFDBB are adjusted through NLMS adaptation criterion in this stage. Notably, SPFDBB is a part of FDABB and can be executed separately.

On the other hand, if the received sound signal is detected as containing speech signal, then the system is switched to the second stage called speech stage. In this stage, the filter coefficient vectors obtained in the silent stage are applied to the lower beamformer to suppress the interference signals and noises, and enhance the speech signal. Finally, the purified speech signal $\hat{Y}(\omega,k)$ is directly sent to the ASR.

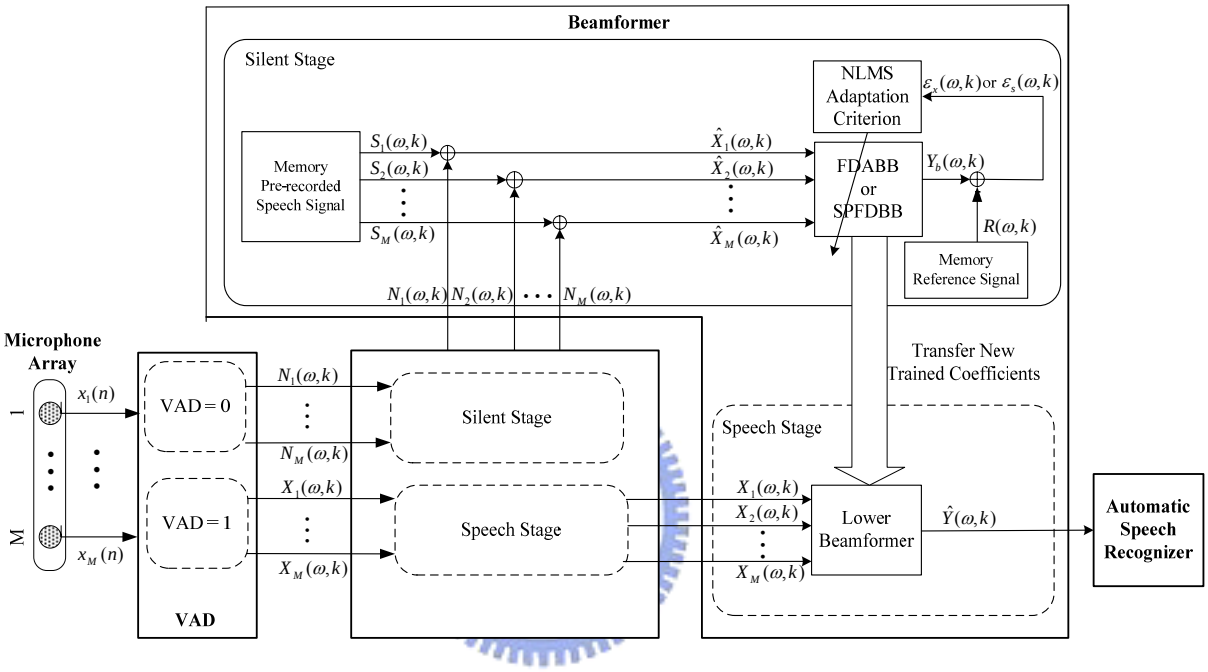


Figure 3-1 Overall system structure

3.3 SPFDBB Using NLMS Adaptation Criterion

The linear model in Eq. (2-1) is transformed to the frequency domain by padding the short-time Fourier transform of the error signal with zeros to make it twice as long as the window length. The error signal at frequency ω and frame k is written as:

$$\varepsilon_x(\omega,k) = R(\omega,k) - \mathbf{Q}^H(\omega,k)\hat{\boldsymbol{X}}(\omega,k) \qquad (3\text{-}1)$$

where $R(\omega,k)$ is the reference signal in the frequency domain, $\boldsymbol{Q}(\omega,k)$ denotes the filter coefficient vector we intend to find, and $\hat{\boldsymbol{X}}(\omega,k)=\left[\hat{X}_1(\omega,k) \quad \cdots \quad \hat{X}_M(\omega,k)\right]^T$ is the training signal vector at frequency ω and frame k. The optimal set of filter coefficient vectors can be found using the formula:

$$\begin{aligned}&\min_{Q}\varepsilon_x(\omega,k)\varepsilon_x^{*}(\omega,k)\\&=\min_{Q}\left[R(\omega,k)-\boldsymbol{Q}^H(\omega)\hat{\boldsymbol{X}}(\omega,k)\right]\left[R(\omega,k)-\boldsymbol{Q}^H(\omega)\hat{\boldsymbol{X}}(\omega,k)\right]^{*}\end{aligned} \tag{3-2}$$

where the superscripts $*$ denotes the complex conjugate.

The normalized LMS solution of Eq. (3-2) is given by:

$$\boldsymbol{Q}(\omega,k+1)=\boldsymbol{Q}(\omega,k)+\frac{\varepsilon_x(\omega,k)\hat{\boldsymbol{X}}^{*}(\omega,k)}{\gamma+\hat{\boldsymbol{X}}^{H}(\omega,k)\hat{\boldsymbol{X}}(\omega,k)} \tag{3-3}$$

Consequently, the purified output signal can be obtained by the following equation:

$$\hat{Y}(\omega,k)=\boldsymbol{Q}^H(\omega,k)\boldsymbol{X}(\omega,k) \tag{3-4}$$

where $\boldsymbol{X}(\omega,k)=\left[X_1(\omega,k) \quad \cdots \quad X_M(\omega,k)\right]$ is the received signal vector which contains speech source, interference and noise. From Eq. (3-1), the filter coefficient vector equalizes the acoustic channel dynamics and also creates the null space for the interference and noise.

As mentioned above, the filter coefficient vector $\boldsymbol{Q}(\omega,k)$ equalizes the channel response and rejects interference signals and noises. To emphasize these two objectives differently, a soft penalty function is added into the performance index as,

$$\min_{Q} \varepsilon_x(\omega,k)\varepsilon_x^{*}(\omega,k) + \mu\varepsilon_s(\omega,k)\varepsilon_s^{*}(\omega,k) \tag{3-5}$$

where μ is the soft penalty parameter and

$$\varepsilon_s(\omega,k) = R(\omega,k) - \boldsymbol{Q}^{H}(\omega,k)\boldsymbol{S}(\omega,k) \tag{3-6}$$

Then, the iterative equation utilizing the NLMS adaptation criterion can be shown as:

$$\boldsymbol{Q}(\omega,k+1) = \boldsymbol{Q}(\omega,k) + \frac{\lambda\{\mathbf{G}_x(\omega,k) + \mu\mathbf{G}_s(\omega,k)\}}{\gamma + \boldsymbol{X}^{H}(\omega,k)\boldsymbol{X}(\omega,k) + \mu\boldsymbol{S}^{H}(\omega,k)\boldsymbol{S}(\omega,k)} \tag{3-7}$$

where

$\boldsymbol{G}_x(\omega,k) = \varepsilon_x(\omega,k)\hat{\boldsymbol{X}}^{*}(\omega,k)$, $\boldsymbol{G}_s(\omega,k) = \varepsilon_s(\omega,k)\boldsymbol{S}^{*}(\omega,k)$ and λ is the step size. If the soft penalty is set to infinity, then the system only focuses on minimizing the channel distortion. On the other hand, the system returns to the formulation in Eq. (3-2) when the soft penalty is set to zero.

The problem of Eq. (3-7) is the window size has to equal that in ASR to ensure calculation accuracy. However, the window size may be too small for cases where the acoustic channel response duration is long (e.g., long reverberation path). Because the perturbation caused by channel model error is highly correlated to the reference signal instead of the uncorrelated noise, the updating process will not converge to a fixed channel response which is shown in Fig 6.3 and the relation between two sequential frames is highly relative. Taking the frame average over several frames (denoted as L) allows the channel response to be approximated; since information of channel response which is not contained in one frame could be regarded as an external noise in the next frame. Thus, the performance index can be written in a quadratic from as:

$$\min_{Q}\begin{bmatrix}\mathbf{V}(\omega,k)\\ \mathbf{U}(\omega,k)\end{bmatrix}^{H}\mathbf{\Lambda}\begin{bmatrix}\mathbf{V}(\omega,k)\\ \mathbf{U}(\omega,k)\end{bmatrix}$$

$$=\min_{Q}\left[\mathbf{V}^{H}(\omega,k)\mathbf{\Lambda}_{\mathbf{1}}\mathbf{V}(\omega,k)+\mathbf{V}^{H}(\omega,k)\mathbf{\Lambda}_{\mathbf{2}}\mathbf{U}(\omega,k)+\mathbf{U}^{H}(\omega,k)\mathbf{\Lambda}_{\mathbf{3}}\mathbf{V}(\omega,k)+\mathbf{U}^{H}(\omega,k)\mathbf{\Lambda}_{\mathbf{4}}\mathbf{U}(\omega,k)\right] \tag{3-8}$$

where

$$\boldsymbol{V}(\omega,k)=\left[R(\omega,k)-\boldsymbol{Q}^{H}(\omega)\boldsymbol{S}(\omega,k)\quad\cdots\quad R(\omega,k+L-1)-\boldsymbol{Q}^{H}(\omega)\boldsymbol{S}(\omega,k+L-1)\right],$$

$$\boldsymbol{U}(\omega,k)=\left[\boldsymbol{Q}^{H}(\omega)\boldsymbol{N}(\omega,k)\quad\cdots\quad\boldsymbol{Q}^{H}(\omega)\boldsymbol{N}(\omega,k+L-1)\right]^{T},$$

and

$\mathbf{\Lambda}=\begin{bmatrix}\mathbf{\Lambda}_{\mathbf{1}} & \mathbf{\Lambda}_{\mathbf{2}}\\ \mathbf{\Lambda}_{\mathbf{3}} & \mathbf{\Lambda}_{\mathbf{4}}\end{bmatrix}$ is a $2L\times 2L$ matrix.

The performance indexes denoted as Eqs. (3-2) and (3-5) are two special cases of Eq (3-8) with $\mathbf{\Lambda}_{\mathbf{1}}=1$, $\mathbf{\Lambda}_{\mathbf{2}}=\mathbf{\Lambda}_{\mathbf{3}}=-1$, $\mathbf{\Lambda}_{\mathbf{4}}=1$, $L=1$ and $\mathbf{\Lambda}_{\mathbf{1}}=1+\mu$, $\mathbf{\Lambda}_{\mathbf{2}}=\mathbf{\Lambda}_{\mathbf{3}}=-1$, $\mathbf{\Lambda}_{\mathbf{4}}=1$, $L=1$. Because of the long reverberation path and the fix window size, the error signal between different frames should be taken into consideration. Therefore, the soft penalty approach can be applied on several windows (frames) by choosing

$$\mathbf{\Lambda}_{\mathbf{1}}=\begin{bmatrix}1+\mu & \mu & \cdots & \mu\\ \mu & 1+\mu & \ddots & \vdots\\ \vdots & \ddots & \ddots & \mu\\ \mu & \cdots & \mu & 1+\mu\end{bmatrix},\quad \mathbf{\Lambda}_{\mathbf{4}}=\mathbf{I}_{L}\quad\text{and}\quad\mathbf{\Lambda}_{\mathbf{2}}=\mathbf{\Lambda}_{\mathbf{3}}=-\mathbf{I}_{L}$$

where $\mathbf{I}_{L}$ is an identity matrix with dimension L. Significantly, when the channel response is longer than the window size, the reference signal in previous windows is added in the following windows of the received signal. A good estimation should collect the information of several frames to eliminate this correlation effect. The

choice of $\mathbf{\Lambda}_1$ considers the cross-term as a factor to minimize this correlation effect. Using the performance index as Eq. (3-8), the SPFDBB can be summarized as:

$$\boldsymbol{Q}(\omega,k+1)=\boldsymbol{Q}(\omega,k)+\frac{\lambda\{\boldsymbol{H}_x(\omega,k)+\mu E_s(\omega,k)\boldsymbol{P}^*(\omega,k)\}}{\gamma+\hat{\mathbf{X}}^{\boldsymbol{H}}(\omega,k)\otimes\hat{\mathbf{X}}(\omega,k)+\mu\boldsymbol{P}^{\boldsymbol{H}}(\omega,k)\boldsymbol{P}(\omega,k)} \tag{3-9}$$

where $\boldsymbol{H}_x(\omega,k)=\hat{\mathbf{X}}^*(\omega,k)\boldsymbol{E}_x{}^T(\omega,k)$ and $\hat{\mathbf{X}}(\omega,k)=\left[\hat{\boldsymbol{X}}(\omega,k)\quad\cdots\quad\hat{\boldsymbol{X}}(\omega,k+L-1)\right]$ is the training signal matrix with dimension $M\times L$. The k th error signal vector can be denoted as:

$$\boldsymbol{E}_x(\omega,k)=\left[\varepsilon_x(\omega,k)\quad\cdots\quad\varepsilon_x(\omega,k+L-1)\right]=\boldsymbol{R}(\omega,k)-\boldsymbol{Y}_b(\omega,k) \tag{3-10}$$

$$E_s(\omega,k)=\sum_{j=k}^{k+L-1}\left[R(\omega,j)-\boldsymbol{Q}^{\boldsymbol{H}}(\omega)\boldsymbol{S}(\omega,j)\right] \tag{3-11}$$

where $\boldsymbol{R}(\omega,k)=\left[R(\omega,k)\quad\cdots\quad R(\omega,k+L-1)\right]$ is the reference signal vector, $\boldsymbol{Y}_b(\omega,k)=\left[Y_b(\omega,k)\quad\cdots\quad Y_b(\omega,k+L-1)\right]$ is the purified signal vector of the SPFDBB, and

$$\begin{aligned}&\hat{\mathbf{X}}^{\boldsymbol{H}}(\omega,k)\otimes\hat{\mathbf{X}}(\omega,k)\\&=\hat{\boldsymbol{X}}^H(\omega,k)\hat{\boldsymbol{X}}(\omega,k)+\cdots+\hat{\boldsymbol{X}}^H(\omega,k+L-1)\hat{\boldsymbol{X}}(\omega,k+L-1)\end{aligned} \tag{3-12}$$

where Eq. (3-12) means the sum of autocorrelations of the training signal at k th block and

$$\mathbf{P}(\omega,k)=\sum_{j=k}^{k+L-1}\boldsymbol{S}(\omega,j) \tag{3-13}$$

The purified speech signal at k th block can be represented as:

$$\hat{\boldsymbol{Y}}(\omega,k)=\boldsymbol{Q}^{H}(\omega,k)\mathbf{X}(\omega,k) \tag{3-14}$$

where $\hat{\boldsymbol{Y}}(\omega,k)=\left[\hat{Y}(\omega,k) \quad \cdots \quad \hat{Y}(\omega,k+L-1)\right]$ is the purified signal vector.

Notably, the step k is chosen as $0,L,2L,3L,\cdots$ to perform the adaptation process every L frames. In this way, the relation between the two sequential block data may be lower than the one by overlapping $(L-1)$ frames to perform the next adaptation process (e.g. $k=0,1,2,3,\cdots$). Moreover, the approach which the step k is chosen as $0,1,2,3,\cdots$ significantly increases the computational effort and the memory consumption as the value L increases. The goal of carrying out the beamformer in the frequency domain couldn't achieve. As a result, the proposed SPFDBB (e.g. $k=0,L,2L,3L,\cdots$) is a kind of efficient approach to obtain low computational effort in the frequency domain.

3.4 FDABB and Computational Effort Analysis

3.4.1 FDABB Using NLMS Adaptation Criterion

The number of frames L in the SPFDBB greatly influences the performance shown in Chapter 6. However, the bigger value of L needs more training data sequences. To cope with the window size problem, an index-based algorithm is proposed to adjust the value L automatically. Using Eq. (3-1), the error signal could be separated as:

$$\begin{aligned}\varepsilon_{x}(\omega,k)&=R(\omega,k)-\boldsymbol{Q}^{H}(\omega,k)\hat{\boldsymbol{X}}(\omega,k)\\&=R(\omega,k)-\boldsymbol{Q}^{H}(\omega,k)\left(\boldsymbol{S}(\omega,k)+\boldsymbol{N}(\omega,k)\right)\end{aligned} \tag{3-15}$$

Taking an frame average over L frames

$$E\{\varepsilon_x(\omega,k)\boldsymbol{N}^{\boldsymbol{H}}(\omega,k)\}_L$$
$$= E\{R(\omega,k)\boldsymbol{N}^{\boldsymbol{H}}(\omega,k)\}_L - \boldsymbol{Q}^H(\omega,k)E\{\boldsymbol{S}(\omega,k)\boldsymbol{N}^{\boldsymbol{H}}(\omega,k)+\boldsymbol{N}(\omega,k)\boldsymbol{N}^{\boldsymbol{H}}(\omega,k)\}_L \quad (3\text{-}16)$$

If no correlation exists between speech source and interference signals or noises, the first term and second term of the Eq. (3-16) could be zero. Consequently, this equation can be rewritten as:

$$E\{\varepsilon_x(\omega,k)\boldsymbol{N}^{\boldsymbol{H}}(\omega,k)\}_L = -\boldsymbol{Q}^H(\omega,k)E\{\boldsymbol{N}(\omega,k)\boldsymbol{N}^{\boldsymbol{H}}(\omega,k)\}_L \quad (3\text{-}17)$$

The optimal coefficient vectors should be the null space of $E\{\boldsymbol{N}(\omega,k)\boldsymbol{N}^{\boldsymbol{H}}(\omega,k)\}_L$ and then the norm of Eq. (3-17) should be zero. The large norm of Eq. (3-17) indicates a strong negative or positive correlation between the error signal and interference signals or noises. In other words, the large norm of Eq. (3-17) means that the interference signals or noises affect the error significantly and that the convergence period is not achieved with the frame number L. If the norm of Eq. (3-17) is small, then it means the present value L has lesser help for finding better coefficients. Consequently, the value of L is increased to improve the performance of the algorithm. Conversely, if the room acoustic varies temporarily, then the present norm of Eq. (3-17) would become much larger than the last one. Then, the value of L should be reset to the initial value to handle this sudden change. The $CBVI$ at frequency ω and block i are defined in Eq. (3-18). The parameters in Eq. (3-18) are chosen as $\alpha_0 = 3$, $\alpha_1 = 2$, and $\alpha_2 = 1$ to be an high pass filter to increase the sensitivity of the temporal variation of room acoustic and the convergence. Figure 3-2 summarizes the proposed FDABB algorithm.

$$
\begin{aligned}
&CBVI(\omega,i)\\
&=\sum_{j=0}^{2}\alpha_j\left(\left\|E\left\{\varepsilon_x(\omega,(i-j)L)\boldsymbol{N}^{\boldsymbol{H}}(\omega,iL)\right\}_L\right\|_2^2-\left\|E\left\{\varepsilon_x(\omega,(i-j-1)L)\boldsymbol{N}^{\boldsymbol{H}}(\omega,(i-j-1)L)\right\}_L\right\|_2^2\right)\\
&=\alpha_0\left(\left\|E\left\{\varepsilon_x(\omega,iL)\boldsymbol{N}^{\boldsymbol{H}}(\omega,iL)\right\}_L\right\|_2^2\right)+(\alpha_1-\alpha_0)\left(\left\|E\left\{\varepsilon_x(\omega,(i-1)L)\boldsymbol{N}^{\boldsymbol{H}}(\omega,(i-1)L)\right\}_L\right\|_2^2\right)\\
&+(\alpha_2-\alpha_1)\left(\left\|E\left\{\varepsilon_x(\omega,(i-2)L)\boldsymbol{N}^{\boldsymbol{H}}(\omega,(i-2)L)\right\}_L\right\|_2^2\right)-\alpha_2\left(\left\|E\left\{\varepsilon_x(\omega,(i-3)L)\boldsymbol{N}^{\boldsymbol{H}}(\omega,(i-3)L)\right\}_L\right\|_2^2\right)
\end{aligned}
\tag{3-18}
$$

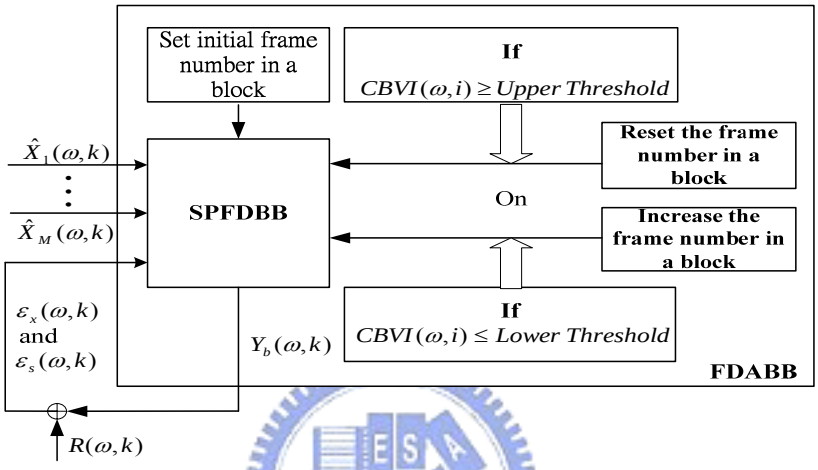


Figure 3-2 FDABB using NLMS adaptation criterion

3.4.2 Computational Effort Analysis

This section analyzes the computational effort of the reference-signal-based time-domain adaptive beamformer, SPFDBB, and FDABB from two different viewpoints: the coefficients adaptation phase and the lower beamformer phase. In the coefficients adaptation phase, the speaker is silent and the coefficients are updated with the iteration equation (2-2) for the reference-signal-based time-domain adaptive beamformer, with Eq. (3-9) for SPFDBB, and with Eqs. (3-9) and (3-18) for FDABB. The computational effort is based on a one-second length input datum for each phase, and is shown in Table 3-1. The sampling rate is denoted as f_s, meaning that the considered input data contains f_s samples. The length of STFT is represented as B_l;

the length of input data in a frame is represented as B_i and the shift size of STFT is represented as B_s. The filter tap of the reference-signal-based time-domain adaptive beamformer is assumed to be P, and the dimension of filter coefficient vector in the time domain is given by $MP \times 1$. The function $I(z)$ takes the integer part of z only. With the increasing value of L, the computational efforts of SPFDBB and FDABB decrease. The computational effort of STFT is estimated by using radix-2 decimation-in-frequency FFT [87]. Notably, the STFT of M microphones is necessary for VAD and speaker's location detection algorithm. Therefore, the computational loading could be reasonably omitted to decrease the amount of real multiplication requirement when the overall system architecture is considered. Furthermore, SPFDBB and FDABB could be implemented with parallelism to decrease the beamformer loading. Furthermore, Chapter 6 lists the computational effort of the proposed algorithms in the two simulations.

Table 3-1 Real Multiplication Requirement in One Second Input Data

	Multiplication Requirement	
	Adaptation Phase	**Lower Beamformer Phase**
Time-domain adaptive beamformer	$f_s(2MP+1)$	$Mf_sP + I\left(\frac{f_s - B_i}{B_s}+1\right)\frac{(B_i+2)\log_2(B_i+2)}{2}$
FDABB	$\frac{(18ML+10M+7)}{L} I\left(\frac{f_s - B_i}{B_s}+1\right)\left(\frac{B_i}{2}+1\right)$	$4MI\left(\frac{f_s - B_i}{B_s}+1\right)\left(\frac{B_i}{2}+1\right)$
SPFDBB	$\frac{(14ML+8M+3)}{L} I\left(\frac{f_s - B_i}{B_s}+1\right)\left(\frac{B_i}{2}+1\right)$	$4MI\left(\frac{f_s - B_i}{B_s}+1\right)\left(\frac{B_i}{2}+1\right)$

3.5 Frequency-domain Performance Indexes

From Eq. (3-1), the filter coefficient vector should equalize the acoustic channel dynamics and also creates the null space for the interference signals or noises. Two

frequency-domain performance indices, SDR and NSR, are defined for these effects. SDR means the decreased level of source distortion produced by inexact filter coefficient vector estimation and NSR means the improved level of noise reduction. Assuming that the true channel response is $\boldsymbol{W}(\omega) = [W_1(\omega) \quad ... \quad W_M(\omega)]^T$, then the source distortion can be represented as:

$$\boldsymbol{Q}^H(\omega,k)\boldsymbol{S}(\omega,k) - R(\omega,k) = \boldsymbol{Q}^H(\omega,k)\begin{bmatrix} W_1(\omega)R(\omega,k) \\ \vdots \\ W_M(\omega)R(\omega,k) \end{bmatrix} - R(\omega,k) \tag{3-19}$$

Thus, the optimal filter coefficient vector should satisfy two equality equations:

$$\boldsymbol{Q}^H(\omega)\boldsymbol{W}(\omega) = 1 \quad \text{and} \quad \boldsymbol{Q}^H(\omega)\boldsymbol{N}(\omega,k) = 0 \tag{3-20}$$

The SDR is defined as:

$$SDR(\omega,k) = 20\log\left(\frac{\left|\boldsymbol{Q}^H(\omega,k)\boldsymbol{W}(\omega) - 1\right|}{\left|\boldsymbol{Q}^H(\omega,1)\boldsymbol{W}(\omega) - 1\right|}\right) \qquad \forall\, k = 1,... \tag{3-21}$$

and the NSR is defined as:

$$NSR(\omega,k) = 20\log\left(\frac{\left|\boldsymbol{Q}^H(\omega,k)\boldsymbol{N}(\omega,k)\right|}{\left|\boldsymbol{Q}^H(\omega,1)\boldsymbol{N}(\omega,1)\right|}\right) \qquad \forall\, k = 1,... \tag{3-22}$$

Since the denominators of Eqs. (3-21) and (3-22) are initial reference values, SDR or NSR represents the performance enhancement derived by comparing certain iteration with the initial filter coefficient vector. Smaller values of SDR and NSR indicate smaller source distortion and higher noise suppression performance.

3.6 Summary

This chapter presents two novel reference-signal-based frequency-domain beamformers, FDABB and SPFDBB, to overcome problems such as calibration, near-field or far-field cases, resolution and desired signal cancellation etc. These approaches not only reduce the computational effort significantly in the ASR-based application as compared with the reference-signal-based time domain adaptive beamformer, but also improve performance in a noisy environment. Moreover, FDABB which can automatically adjust the frame number to different environments is particularly suitable for practical applications.

Chapter 4

H_∞ Adaptation Criterion

4.1. Introduction

The NLMS adaptation criterion used in the preceding chapter does not make any assumption about the pre-recoded signals and the disturbance, unlike the exact least square algorithm such as recursive least square (RLS). The solution of the NLMS adaptation criterion recursively updates the filter coefficient vector along the direction of the instantaneous gradient of the squared error. Therefore, the NLMS adaptation criterion is more robust to disturbance variation than the RLS algorithm. For example, it has been observed that the NLMS has better tracking capabilities than the RLS algorithm in the presence of non-stationary inputs [88]. However, the performance of the NLMS depends upon the properties of the modeling errors which may lead to large coefficient vector estimation error. Consequently, it is necessary to design a robust adaptive algorithm to guarantee that if the disturbance energy is small, the coefficient vector estimation error will be small as well (in energy).

There has been an increasing interest in the mini-max estimation method [89-97] called H_∞ algorithm which is more robust and less sensitive to model uncertainties and

parameter variations than the H_2 adaptation criterion (such as the Kalman filter). This is because no a *priori* knowledge of the disturbance statistics is required in the H_∞ algorithm. It means that the H_∞ algorithm can accommodate for all conceivable disturbances which have a finite energy. Moreover, the estimation criterion of the H_∞ algorithm is to minimize the worst possible effects of the disturbances (modeling errors and additive noises) on the signal estimation error. Actually, the NLMS adaptation criterion is the central a *posteriori* H_∞ optimal filter [94]. However, it is the H_∞ optimal filter that minimizes the worst possible effects of the disturbances on the filtered output error. But the goal of the reference-signal-based beamformer is to estimate the filter coefficient vector itself instead of in minimizing the filtered output error. In this case, the criterion of the H_∞ optimal filter has to be modified to address the problem of filter coefficient vector estimation (eg. minimizing the coefficient vector estimation error).

The remainder of this chapter is organized as follows. Section 4.2 describes the definition of the H_∞-norm and the recursive solution of the time-domain adaptive beamformers which utilizes H_∞ adaptation criterion. Section 4.3 applies H_∞ adaptation criterion to the two proposed frequency-domain adaptive beamformers, SPFDBB and FDABB as well as analyzes the computing effort of the two frequency-domain beamformers and the time-domain beamformer using H_∞ adaptation criterion. Section 4.4 defines time-domain performance indices for the experiments in Chapter 6 to measure the robustness of H_∞ adaptation criterion. Finally, a conclusion is given in Section 4.5.

4.2. Time-Domain Adaptive Beamformer Using H_∞ Adaptation Criterion

4.2.1 Definition of H_∞-norm

The H_∞-norm defines the worst case response of a system. If Z denotes a transfer operator that maps an input causal sequence $\{u_i\}$ to an output causal sequence $\{y_i\}$ as shown in Fig. 4-1, the H_∞-norm of Z is defined as,

$$\|Z\|_\infty = \sup_{u \neq 0} \frac{\|y\|_2}{\|u\|_2} \tag{4-1}$$

where the notation $\|.\|_2$ denotes the 2-norm. Obviously, the H_∞-norm can be regards as the maximum energy gain from the input u to the output y.

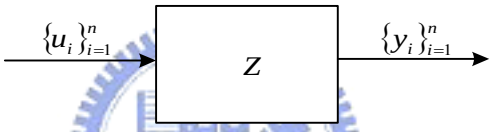


Figure 4-1 Transfer operator Z from input $\{u_i\}$ to output $\{y_i\}$

4.2.2 Formulation of Time-Domain Adaptive Beamformer

Figure 4-2 shows the overall architecture of the time-domain adaptive beamformer using H_∞ adaptation criterion. The data flow and architecture are almost the same with those introduced in the Chapter 2 except the H_∞ adaptation criterion is used. In the silent stage (VAD = 0), the filter coefficient vectors are adapted through H_∞ adaptation criterion. In the speech stage (VAD = 1), the computed filter coefficient vectors are applied to the lower beamformer to suppress the interference signals and noises, and derive the purified speech signal.

Based on the system architecture shown in Fig. 4-2, the formulation of speech enhancement system can be expressed as the following linear model:

$$r(n) = \hat{\boldsymbol{x}}^T(n)\boldsymbol{q} + e(n) \tag{4-2}$$

where $r(n)$ is the reference signal and $\hat{\boldsymbol{x}}(n) = [\hat{\boldsymbol{x}}_1(n) \quad \cdots \quad \hat{\boldsymbol{x}}_M(n)]^T$ is a $MP \times 1$ training signal vector. $\hat{\boldsymbol{x}}_i(n) = [\hat{x}_i(n) \quad \cdots \quad \hat{x}_i(n-P+1)]$ is a $1 \times P$ training signal vector and each component in the silent stage is constructed from the linear combination of the pre-recorded speech signals and the online recorded interference signals or noises as $\hat{x}_i(n) = s_i(n) + n_i(n)$. M denotes the number of microphones and P denotes the number of filter tap. Additionally, $\boldsymbol{q} = [q_{11} \quad \cdots \quad q_{1P} \quad \cdots \quad q_{M1} \quad \cdots \quad q_{MP}]^T$ is the $MP \times 1$ unknown filter coefficient vector in the time domain that we intent to estimate. $e(n)$ is the unknown disturbance, which may also include modeling error.

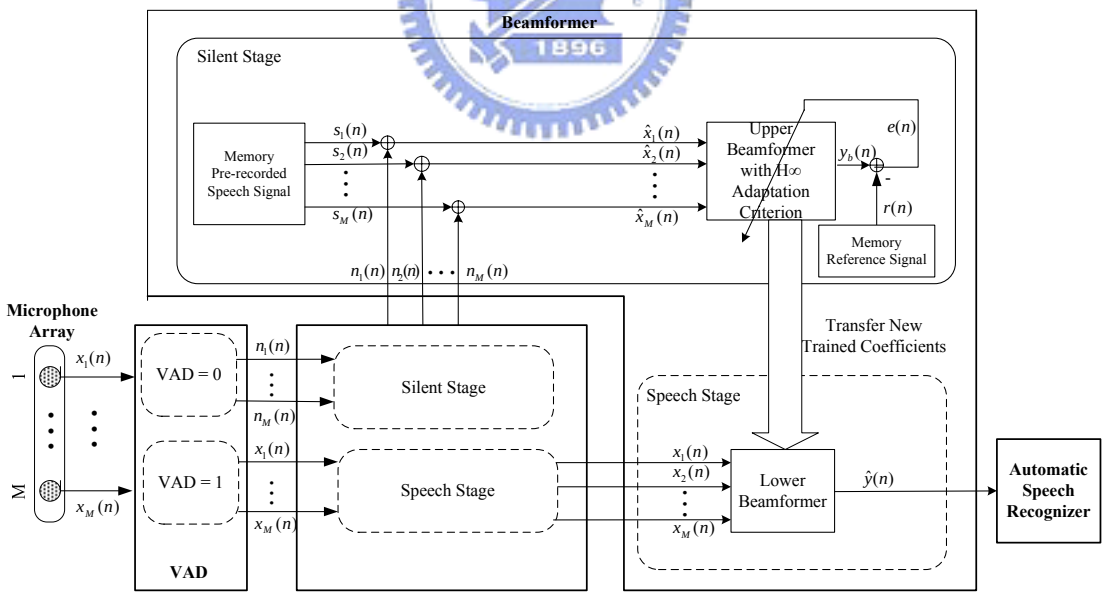


Figure 4-2 System Architecture of the time-domain adaptive beamformer using H_∞ adaptation criterion

The problem of the proposed speech enhancement system is how to use a strategy called $\Psi(r(1), \cdots r(n); \hat{\boldsymbol{x}}(1), \cdots \hat{\boldsymbol{x}}(n))$ to estimate the filter coefficient vector $\boldsymbol{q}$

using all the information available from time 1 to time n such that the H_∞-norm from the disturbances $\left\{\mu_0^{-\frac{1}{2}}\left\|\boldsymbol{q}-\hat{\boldsymbol{q}}(1)\right\|_2,\ \{e(i)\}_{i=1}^n\right\}$ to the coefficient vector estimation error $\{\tilde{\boldsymbol{q}}(i)\}_{i=1}^n$ minimized, where $\hat{\boldsymbol{q}}(1)$ denotes the initial guess for $\boldsymbol{q}$ and μ_0 is a positive constant that reflects to how close $\boldsymbol{q}$ is to the initial guess $\hat{\boldsymbol{q}}(1)$. This transfer operator shown in Fig. 4-3 is designate by $T(\Psi)$. $\tilde{\boldsymbol{q}}(n)$ is the $MP\times 1$ coefficient vector estimation error defined as:

$$\tilde{\boldsymbol{q}}(n)=\boldsymbol{q}-\hat{\boldsymbol{q}}(n) \tag{4-3}$$

where $\hat{\boldsymbol{q}}(n)$ is the $MP\times 1$ estimated filter coefficient vector.

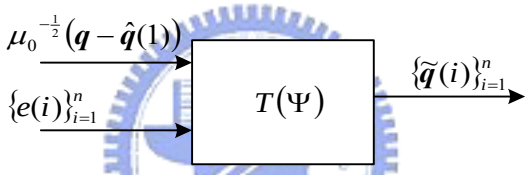


Figure 4-3 Transfer operator from disturbances to coefficient vector estimation error

To apply the adaptive H_∞ adaptation criterion, the linear model, as in Eq. (4-2), is transformed into an equivalent state-space form:

$$\begin{cases}\boldsymbol{q}(n+1)=\boldsymbol{q}(n)\\ r(n)=\hat{\boldsymbol{x}}^T(n)\boldsymbol{q}(n)+e(n)\end{cases}\quad\text{with}\quad \boldsymbol{q}(1)=\boldsymbol{q} \tag{4-4}$$

To find the optimal H_∞ estimation, the criterion in the sense of H_∞-based filtering is:

$$\gamma_{qo}^2=\inf_{\Psi}\left\|T(\Psi)\right\|_\infty^2=\inf_{\Psi}\sup_{e,\boldsymbol{q}}\frac{\sum_{i=1}^n\left|\tilde{\boldsymbol{q}}(i)\right|^2}{\mu_0^{-1}\left|\boldsymbol{q}-\hat{\boldsymbol{q}}(0)\right|^2+\sum_{i=1}^n\left|e(i)\right|^2} \tag{4-5}$$

where $\|\cdot\|^2$ denotes the square of the 2-norm. However, a closed form solution of Eq. (4-5) is unavailable for general cases. Therefore, it is common in the literature to relax the minimization condition and settle for a suboptimal solution. Given a scalar $\gamma_q > 0$, find a H_∞ suboptimal estimation strategy called $\hat{\boldsymbol{q}}(n) = \Psi\big(r(1), \cdots r(n); \hat{\boldsymbol{x}}(1), \cdots \hat{\boldsymbol{x}}(n)\big)$ that achieves $\|T(\Psi)\|_\infty < \gamma_q$. In other words, the suboptimal solution is to find a strategy that achieves

$$\sup_{e,\boldsymbol{q}} \frac{\sum_{i=1}^{n} |\tilde{\boldsymbol{q}}(i)|^2}{\mu_0^{-1} |\boldsymbol{q} - \hat{\boldsymbol{q}}(1)|^2 + \sum_{i=1}^{n} |e(i)|^2} < \gamma_q^2, \qquad \text{for } 1 \le i \le n \tag{4-6}$$

4.2.3 Solution of suboptimal H_∞ Adaptation Criterion

Consider a time-variant state-space model of the form

$$\begin{cases} \boldsymbol{q}(n+1) = \mathbf{F}(n)\boldsymbol{q}(n) + \mathbf{G}(n)\boldsymbol{u}(n) \\ \boldsymbol{r}(n) = \mathbf{H}(n)\boldsymbol{q}(n) + e(n) \end{cases} \quad \text{with } n \ge 0 \tag{4-7}$$

where $\mathbf{F}(n) \in C^{MP \times MP}$, $\mathbf{G}(n) \in C^{MP \times J}$, $\mathbf{H}(n) \in C^{B \times MP}$ are known matrices, $\boldsymbol{u}(n) \in C^{J \times 1}$ are unknown process noises, and $\boldsymbol{r}(n) \in C^{B \times 1}$ are observations. In general, the H_∞-formulation estimates some arbitrary linear combination of the states, say

$$\boldsymbol{d}(n) = \mathbf{B}(n)\boldsymbol{q}(n) \tag{4-8}$$

where $\mathbf{B}(n) \in C^{N \times MP}$. Let $\hat{\boldsymbol{d}}(n) = \Psi\big(r(1), \cdots r(n)\big)$ denote the estimation of $\boldsymbol{d}(n)$ given observations $\{r(i)\}|_{i=1}^{n}$. Define the estimation error as

$$\tilde{\boldsymbol{d}}(n) = \hat{\boldsymbol{d}}(n) - \mathbf{B}(n)\boldsymbol{q}(n) \tag{4-9}$$

Given a scalar $\gamma_d > 0$, find a suboptimal H_∞ estimation strategy $\hat{\boldsymbol{d}}(n) = \Psi(r(1), \cdots r(n))$ that achieves $\|T(\Psi)\|_\infty < \gamma_d$. In other words, find a strategy that achieves

$$\sup_{e,\boldsymbol{q},\boldsymbol{u}} \frac{\sum_{i=1}^{n} \left|\tilde{\boldsymbol{d}}(i)\right|^2}{\mu_0^{-1}\left|\boldsymbol{q} - \hat{\boldsymbol{q}}(1)\right|^2 + \sum_{i=1}^{n} \boldsymbol{u}(i)^H \boldsymbol{u}(i) + \sum_{i=1}^{n} \left|e(i)\right|^2} < \gamma_d^2 \quad , \qquad \text{for } 1 \le i \le n \tag{4-10}$$

In this case, the suboptimal solution [96] is recursively computed as

$$\hat{\boldsymbol{q}}(n+1) = \mathbf{F}(n)\hat{\boldsymbol{q}}(n) + \mathbf{K}(n)\left(\boldsymbol{r}(n) - \mathbf{H}(n)\hat{\boldsymbol{q}}(n)\right) \tag{4-11}$$

$$\mathbf{K}(n) = \mathbf{F}(n)\mathbf{P}(n)\mathbf{H}^H(n)\left(\mathbf{I} + \mathbf{H}(n)\mathbf{P}(n)\mathbf{H}^H(n)\right)^{-1} \tag{4-12}$$

$$\mathbf{P}(n+1) = \mathbf{F}(n)\left[\mathbf{P}^{-1}(n) + \mathbf{H}^H(n)\mathbf{H}(n) - \gamma_d^{-2}\mathbf{B}^H(n)\mathbf{B}(n)\right]^{-1}\mathbf{F}^*(n) + \mathbf{G}(n)\mathbf{G}^H(n) \tag{4-13}$$

$$\hat{\boldsymbol{q}}(1) = \boldsymbol{0} \quad \text{and} \quad \mathbf{P}(1) = \mu_0 \mathbf{I} \tag{4-14}$$

where $\mathrm{P}(n)$ is an $MP \times MP$ matrix, $(\cdot)^{-1}$ denotes the matrix inverse operation, and $(\cdot)^*$ denotes the transpose operation. If the $\mathbf{F}(n)$ is invertible and $\mu_0 > 0$, then the following alternative condition can be used to guarantee the existence of Eq. (4-10)

$$\mathbf{P}^{-1}(n) + \mathbf{H}^H(n)\mathbf{H}(n) - \gamma_d^{-2}\mathbf{B}^H(n)\mathbf{B}(n) > 0 \text{, for all } n \tag{4-15}$$

4.2.4 Solution of Time-domain Adaptive Beamformer

Let's apply Eqs. (4-11), (4-12), (4-13), and (4-14) to the state-space model Eq. (4-4) where $\mathbf{F}(n)=\mathbf{I}$, $\mathbf{G}(n)=\mathbf{0}$, $\mathbf{H}(n)=\hat{\boldsymbol{x}}^T(n)$, and $\mathbf{B}(n)=\mathbf{I}$. Thus the solution of $\hat{\boldsymbol{q}}(n)$ can be found by the following iterative equations:

$$\hat{\boldsymbol{q}}(n+1)=\hat{\boldsymbol{q}}(n)+\mathbf{K}(n)\left(r(n)-\hat{\boldsymbol{x}}^T(n)\hat{\boldsymbol{q}}(n)\right) \tag{4-16}$$

$$\mathbf{K}(n)=\mathbf{P}(n)\hat{\boldsymbol{x}}(n)\left(1+\hat{\boldsymbol{x}}^T(n)\mathbf{P}(n)\hat{\boldsymbol{x}}(n)\right)^{-1} \tag{4-17}$$

$$\mathbf{P}^{-1}(n+1)=\mathbf{P}^{-1}(n)+\hat{\boldsymbol{x}}(n)\hat{\boldsymbol{x}}^T(n)-\gamma_q^{-2}\mathbf{I} \tag{4-18}$$

$$\hat{\boldsymbol{q}}(1)=\boldsymbol{0} \quad \text{and} \quad \mathbf{P}(1)=\mu_0\mathbf{I} \tag{4-19}$$

To ensure the existence of Eq. (4-6), γ_q should be chosen such that $\mathbf{P}^{-1}(n)+\hat{\boldsymbol{x}}(n)\hat{\boldsymbol{x}}^T(n)-\gamma_q^{-2}\mathbf{I}>0$. For this reason, γ_q^{-2} is selected as $\delta\times eig\left(\mathbf{P}^{-1}(n)+\hat{\boldsymbol{x}}(n)\hat{\boldsymbol{x}}^T(n)\right)$ during the iteration, where $eig(z)$ denotes the minimum eigenvalue of z. δ is a positive constant and lower than one to ensure that Eq. (4-18) is positive definite.

The adaptation of the filter coefficient vector is performed in the silent stage. When the system is switched to speech stage, the adaptation stops and the filter coefficient vector is passed to lower beamformer. The purified speech signal can be calculated by

$$\hat{y}(n)=\boldsymbol{x}^T(n)\hat{\boldsymbol{q}}(n) \tag{4-20}$$

where $\boldsymbol{x}(n)=\left[\boldsymbol{x}_1(n) \quad \cdots \quad \boldsymbol{x}_M(n)\right]^T$ is the $MP\times 1$ online recorded noisy speech signal vector acquired by the microphone array, where $\boldsymbol{x}_i(n)=\left[x_i(n) \quad \cdots \quad x_i(n-P+1)\right]$.

4.3. SPFDBB, FDABB and Computational Effort Analysis

4.3.1 SPFDBB Using H_∞ Adaptation Criterion

Figure 4-4 shows the overall architecture of the proposed speech enhancement using H_∞ adaptation criterion in the frequency domain. For the ASR application, the purified spectrum data should be computed directly to save computational effort, since most speech recognition algorithms are performed in the frequency domain. In this case, the filter coefficient vectors can be updated on a block of data. Hence, the problem is transformed into the frequency domain by using STFT. In conjunction with the spectrum-based ASR, the window size in the STFT has to equal to that in ASR in order to obtain a more accurate result. However, the window size may be too small to capture the acoustic channel response. For this reason, Chapter 4 proposed an approach called SPFDBB which takes the frame average over several frames as a block improving the approximation of the channel response. The number of frames in a block is denoted as the frame number L. In this chapter, the H_∞ adaptation criterion is adopted to improve the performance further.

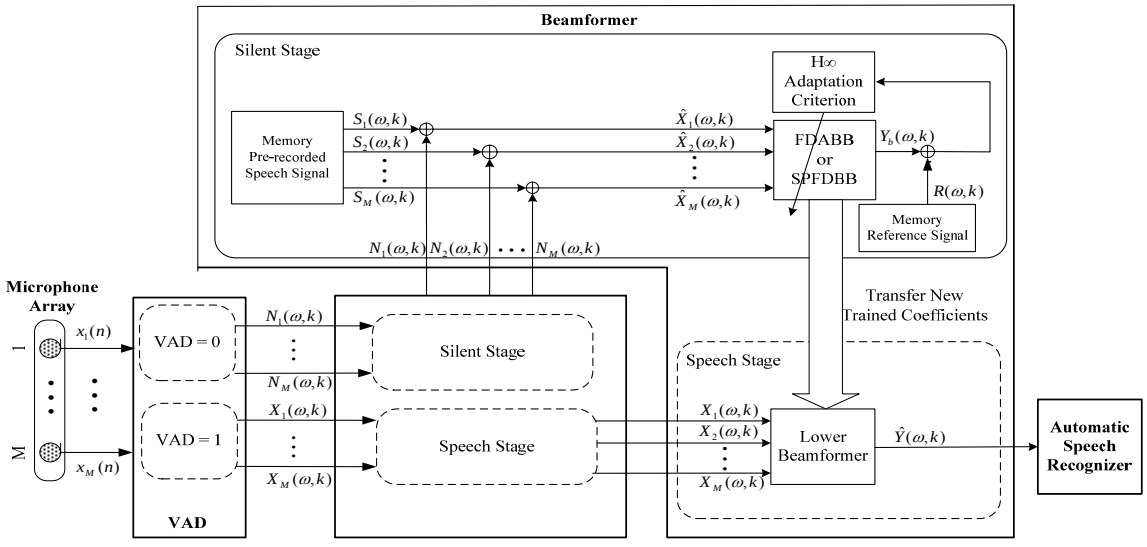


Figure 4-4 System Architecture of SPFDBB and FDABB using H∞ adaptation criterion

The strategy of the SPFDBB using H_∞ adaptation criterion can be formulated as:

$$\sup_{E,\boldsymbol{Q}} \frac{\sum_{i=1}^{k}\left|\tilde{\boldsymbol{Q}}(\omega,i)\right|^2}{\mu_0^{-1}\left|\boldsymbol{Q}(\omega)-\hat{\boldsymbol{Q}}(\omega,1)\right|^2+\sum_{i=1}^{k}\left|\boldsymbol{E}(\omega,i)\right|^2} < \gamma_Q^2, \qquad \text{for } 1\le i\le k \qquad (4\text{-}21)$$

where $\boldsymbol{Q}(\omega)$ denotes the $M\times 1$ unknown filter coefficient vector at frequency ω and $\hat{\boldsymbol{Q}}(\omega,1)$ is the initial guess. $\tilde{\boldsymbol{Q}}(\omega,k)$ is the $M\times 1$ coefficient vector estimation error at kth block defined as:

$$\tilde{\boldsymbol{Q}}(\omega,k)=\boldsymbol{Q}(\omega)-\hat{\boldsymbol{Q}}(\omega,k) \qquad (4\text{-}22)$$

where $\hat{\boldsymbol{Q}}(\omega,k)$ is the $M\times 1$ estimated filter coefficient vector in the frequency domain. The energy of the disturbance $\left|\boldsymbol{E}(\omega,k)\right|^2$ is defined as:

$$\left|\boldsymbol{E}(\omega,k)\right|^2=\begin{bmatrix}\boldsymbol{V}(\omega,k)\\ \boldsymbol{U}(\omega,k)\end{bmatrix}^H\boldsymbol{\Lambda}\begin{bmatrix}\boldsymbol{V}(\omega,k)\\ \boldsymbol{U}(\omega,k)\end{bmatrix} \qquad (4\text{-}23)$$

where

$$\boldsymbol{V}(\omega,k)=\left[R(\omega,k)-\boldsymbol{S}^H(\omega,k)\hat{\boldsymbol{Q}}(\omega,k) \quad \cdots \quad R(\omega,k+L-1)-\boldsymbol{S}^H(\omega,k+L-1)\hat{\boldsymbol{Q}}(\omega,k)\right]^T$$

and $\boldsymbol{U}(\omega,k)=\left[\boldsymbol{N}^H(\omega,k)\hat{\boldsymbol{Q}}(\omega,k) \quad \cdots \quad \boldsymbol{N}^H(\omega,k+L-1)\hat{\boldsymbol{Q}}(\omega,k)\right]^T$.

$\boldsymbol{\Lambda}=\begin{bmatrix}\boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_2\\ \boldsymbol{\Lambda}_3 & \boldsymbol{\Lambda}_4\end{bmatrix}$ is a $2L\times 2L$ matrix.

$$\mathbf{\Lambda_1}=\begin{bmatrix}1+\mu & \mu & \cdots & \mu\\ \mu & 1+\mu & \ddots & \vdots\\ \vdots & \ddots & \ddots & \mu\\ \mu & \cdots & \mu & 1+\mu\end{bmatrix},\ \mathbf{\Lambda_4}=\mathbf{I_L},\ \text{and}\ \mathbf{\Lambda_2}=\mathbf{\Lambda_3}=-\mathbf{I_L}$$

where $\mathbf{I_L}$ is an identity matrix with dimension $L\times L$ and μ is the soft penalty. $\boldsymbol{N}(\omega,k)=\begin{bmatrix}N_1(\omega,k) & \cdots & N_M(\omega,k)\end{bmatrix}^T$, $\boldsymbol{S}(\omega,k)=\begin{bmatrix}S_1(\omega,k) & \cdots & S_M(\omega,k)\end{bmatrix}^T$, and $R(\omega,k)$ represent the frequency-domain online recorded environmental noise vector, the pre-recorded speech signal vector, and the reference signal respectively.

Let's apply Eqs. (4-11), (4-12), (4-13), and (4-14) to the SPFDBB, where $\mathbf{F}(\omega,k)=\mathbf{I}$, $\mathbf{G}(\omega,k)=\mathbf{0}$, $\mathbf{H}(\omega,k)=\begin{bmatrix}\hat{\boldsymbol{X}}(\omega,k) & \cdots & \hat{\boldsymbol{X}}(\omega,k+L-1) & \mu^{\frac{1}{2}}\sum_{j=k}^{k+L-1}\boldsymbol{S}(\omega,j)\end{bmatrix}^H$, $\mathbf{B}(\omega,k)=\mathbf{I}$, and $\hat{\boldsymbol{X}}(\omega,k)$ is the training signal vector with dimension $M\times 1$. Thus the state space model can be represented by

$$\begin{cases}\boldsymbol{Q}(\omega,k+1)=\boldsymbol{Q}(\omega,k)\\ \boldsymbol{R}_L(\omega,k)=\mathbf{H}(\omega,k)\boldsymbol{Q}(\omega,k)+\boldsymbol{E}(\omega,k)\end{cases}\quad\text{with}\ k\geq 0\ \text{and}\ \boldsymbol{Q}(\omega,k)=\boldsymbol{Q}(\omega)$$

where $\boldsymbol{R}_L(\omega,k)=\begin{bmatrix}R(\omega,k) & \cdots & R(\omega,k+L-1) & \mu^{\frac{1}{2}}\sum_{j=k}^{k+L-1}R(\omega,j)\end{bmatrix}^T$. The solution of $\hat{\boldsymbol{Q}}(\omega,k)$ can be approximated by the iteration:

$$\hat{\boldsymbol{Q}}(\omega,k+1)=\hat{\boldsymbol{Q}}(\omega,k)+\mathbf{K}(\omega,k)\left[\boldsymbol{R}_L(\omega,k)-\mathbf{H}(\omega,k)\hat{\boldsymbol{Q}}(\omega,k)\right] \tag{4-24}$$

$$\mathbf{K}(\omega,k)=\mathbf{P}(\omega,k)\mathbf{H}^H(\omega,k)\left(\mathbf{I}+\mathbf{H}(\omega,k)\mathbf{P}(\omega,k)\mathbf{H}^H(\omega,k)\right)^{-1} \tag{4-25}$$

$$\mathbf{P}^{-1}(\omega,k+1)=\mathbf{P}^{-1}(\omega,k)+\mathbf{H}^H(\omega,k)\mathbf{H}(\omega,k)-\gamma_Q^{-2}\mathbf{I} \tag{4-26}$$

$$\hat{\boldsymbol{Q}}(\omega,1)=\boldsymbol{0} \quad \text{and} \quad \mathbf{P}^{-1}(\omega,1)=\mu_0\mathbf{I} \tag{4-27}$$

The value of γ_Q^{-2} during the iteration is chosen as $\delta eig\left(\mathbf{P}^{-1}(\omega,k)+\mathbf{H}^H(\omega,k)\mathbf{H}(\omega,k)\right)$ where $eig(\mathbf{z})$ denotes the minimum eigenvalue of z. δ is a positive constant and lower than one to ensure that Eq. (4-26) is positive definite. Consequently, the purified speech signal at k th block can be obtained by the following equation:

$$\hat{\boldsymbol{Y}}(\omega,k)=\hat{\boldsymbol{Q}}^H(\omega,k)\boldsymbol{X}(\omega,k) \tag{4-28}$$

where $\hat{\boldsymbol{Y}}(\omega,k)=\left[\hat{Y}(\omega,k) \quad \cdots \quad \hat{Y}(\omega,k+L-1)\right]$ is the purified result and $\boldsymbol{X}(\omega,k)$ is the $M\times L$ online recorded noisy speech signal matrix. The step k is chosen as $0,L,2L,3L,\cdots$ to perform the adaptation process every L frames.

4.3.2 FDABB Using H_∞ Adaptation Criterion

The requirement of the length of the training data would be too large and the SPFDBB could not respond to the change of room acoustics when a large value of L is chosen. Therefore, the method called FDABB is proposed in Chapter 4 to further enhance SPFDBB through the index *CBVI* to allowing the frame number to be adapted on-line. In other words, *CBVI* defined in Eq. (3-18) is the basis for adjusting the frame number. Figure 4-5 summarizes the proposed FDABB algorithm using H_∞ adaptation criterion.

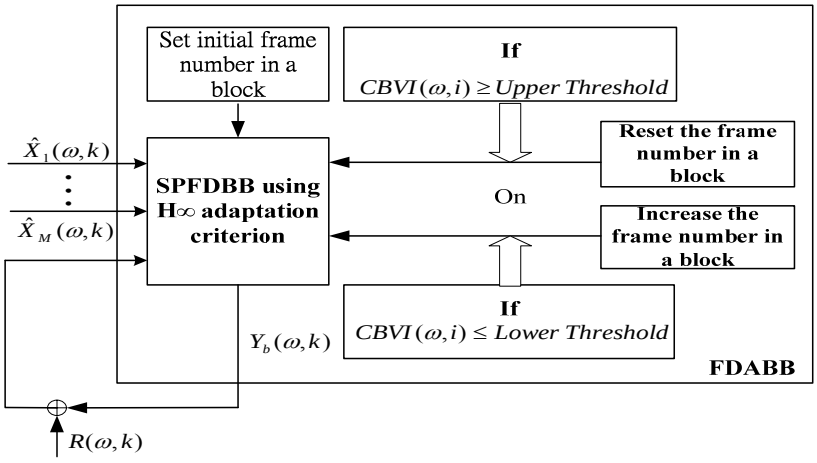


Figure 4-5 FDABB using H_∞ adaptation criterion

4.3.3 Computational Effort Analysis

This section analyzes the computational efforts of the time-domain and the two frequency-domain adaptive beamformers using H_∞ adaptation criterion from two different phases: the coefficients adaptation phase and the lower beamformer phase. In the coefficients adaptation phase, the desired speaker is silent and the coefficients are updated with the iteration equations from (4-16) to (4-19) for the time-domain adaptive beamformer, with Eqs. (4-24) to (4-27) for SPFDBB, and with Eqs. (4-24) to (4-27) and (3-18) for FDABB. The computational efforts are calculated according to a one-second length input datum for each phase, and are shown in Table 4-1. The meanings of the parameters are defined in Section 3.4.2. Notably, the computational effort of a matrix inversion was given in [98]. On the other hand, the eigenvalue must be found to determine the value of γ_q or γ_Q . However, the computation of eigenvalues is complex and the computational effort varies with the precision required. To ensure a steady performance, a general method based on Householder method and the shifted QR algorithm is considered; the computational effort associated with finding the eigenvalues is given in [99-100].

Table 4-1 Real Multiplication Requirement in One Second Input Data

	Multiplication Requirement	
	Adaptation Phase	**Lower Beamformer Phase**
Time-domain beamformer	$\left(2M^3P^3+\frac{9M^2P^2+13MP+10}{4}\right)f_s$	$Mf_sP+I\left(\frac{fs-B_i}{B_s}+1\right)\frac{(B_l+2)\log_2(B_l+2)}{2}$
FDABB	$\left((L+1)^3+5M^3+16M(L+1)^2+4M^2(L+\frac{5}{4})+4M(L+\frac{1}{4})+7\right)I\left(\frac{f_s-B_i}{B_s}+1\right)\left(\frac{B_l}{2}+1\right)/L$	$4MI\left(\frac{f_s-B_i}{B_s}+1\right)\left(\frac{B_l}{2}+1\right)$
SPFDBB	$\left((L+1)^3+5M^3+16M(L+1)^2+4M^2(L+\frac{5}{4})+4M(L+\frac{1}{4})+7\right)I\left(\frac{f_s-B_i}{B_s}+1\right)\left(\frac{B_l}{2}+1\right)/L$	$4MI\left(\frac{f_s-B_i}{B_s}+1\right)\left(\frac{B_l}{2}+1\right)$

4.4. Time-domain Performance Indexes

In this section, six time-domain performance indexes are defined. The first two parameters, SDR and NSR, instead of SNR are defined to evaluate the performances of the NLMS and H_∞ adaptation criterions. This is because a lower SNR may not correspond to a higher ASR rate and these two indexes facilitate to directly separate two main issues: the inverse issue and noise suppression issue. SDR is defined as:

$$SDR(n)=20\log\left(\frac{\left|r(n)-\boldsymbol{s}^T(n)\hat{\boldsymbol{q}}(n)\right|}{\frac{1}{V}\sum_{n=1}^{V}\left|r(n)\right|}\right) \tag{4-29}$$

and NSR is defined as

$$NSR(n)=20\log\left(\frac{\left|\hat{\boldsymbol{q}}^T(n)\boldsymbol{n}(n)\right|}{\frac{1}{MV}\sum_{n=1}^{V}\sum_{i=1}^{M}\left|n_i(n)\right|}\right) \tag{4-30}$$

where V denotes the length of signals in both equations. SDR represents the degree of source distortion that caused by channel effect and noises. Moreover, NSR is the degree of noise reduction.

To observe different characteristics between the NLMS and H_∞ adaptation criterion, four performance indexes named filtered output error $e_f(n)$, reference signal estimation error $e_r(n)$, filter coefficient estimation error ratio, and filtered output error ratio are defined in Eqs. (4-31), (4-32), (4-33), and (4-34) individually.

$$e_f(n) = \hat{\boldsymbol{x}}^T(n)\boldsymbol{q} - \hat{\boldsymbol{x}}^T(n)\hat{\boldsymbol{q}}(n) \tag{4-31}$$

$$e_r(n) = r(n) - \hat{\boldsymbol{x}}^T(n)\hat{\boldsymbol{q}}(n) \tag{4-32}$$

$$\frac{\sum_{i=1}^{n}\left|\tilde{\boldsymbol{q}}(i)\right|^2}{\mu_0^{-1}\left|\boldsymbol{q}-\hat{\boldsymbol{q}}(1)\right|^2 + \sum_{i=1}^{n}\left|e(i)\right|^2} \tag{4-33}$$

$$\frac{\sum_{i=1}^{n}\left|e_f(i)\right|^2}{\mu_0^{-1}\left|\boldsymbol{q}-\hat{\boldsymbol{q}}(1)\right|^2 + \sum_{i=1}^{n}\left|e(i)\right|^2} \tag{4-34}$$

where $\tilde{\boldsymbol{q}}(n)$ is the coefficient vector estimation error defined in Eq. (4-3).

4.5. Summary

In this chapter, the H_∞ adaptation criterion is investigated to enhance the robustness to the modeling error caused by an inadequate window size to capture the acoustic channel dynamics. This chapter utilizes the H_∞ adaptation criterion to replace the NLMS in the reference-signal-based time-domain adaptive beamformer. However, due to the intensive computational effort requirement in the time domain, SPFDBB

and FDABB, using H_∞ adaptation criterion are proposed to significantly reduce the computational effort which is analyzes in Section 4.3.3. As shown in Chapter 6, H_∞ adaptation criterion outperforms the NLMS with the same filter order in terms of SDR, NSR and ASR results.

Chapter 5

Reference-signal-based Speaker's Location Detection

5.1 Introduction

A speech enhancement system using microphone array usually requires a speaker's location estimation capability. With the information of the desired speaker's location, the speech enhancement system can suppress the interference signals and noises from the other locations. For example, in vehicle applications, a driver may wish to exert a particular authority in manipulating the in-car electronic systems through spoken language. Consequently, a better receiving beam using a microphone array can be formed to suppress the environmental noises and enhance the driver's speech signal if the driver's location is known.

In a highly reflective or scattering environment, conventional delay estimation methods such as GCC-based (TDOA-based) algorithms [21-23] or previous works [24-25] do not yield satisfactory results. Although Brandstein *et al.* [101] proposed Tukey's Biweight to redefine the weighting function to deal with the reflection effect; it

is not suitable for a noisy environment. To overcome this limitation, Nikias *et al.* [102] adopted the alpha-stable distribution, instead of a single Gaussian model, to model ambient noise and to obtain a robust speaker's location detection in advance. In recent years, several works have introduced probability-based methods to eliminate the measurement errors caused by uncertainties, such as those associated with reverberation or low energy segments. Histogram-based TDOA estimators such as time histograms [20] and weighted time-frequency histograms [26-27] have been proposed to reduce direction-of-arrival root-mean square errors. The algorithm in [27] performs well especially under low SNR conditions. Moreover, Potamitis *et al.* [103] proposed the probabilistic data association (PDA) technique with the interacting multiple model (IMM) estimator to conquer these measurement errors. Ward *et al*. [7] developed a particle filter beamforming (steered-beamformer-based location approach) in which the weights and particles can be updated using a likelihood function to solve the reverberation problem. Although these statistical based methods [7], [20], [26-27] and [103] can improve the estimation accuracy further, they cannot distinguish from the locations using a single linear microphone array under a totally non-line-of-sight condition which is common in vehicular environments.

Another approach (spectral-estimation-based location approach), proposed by Balan *et al*. [8], explores the eigenstructure of the correlation matrix of the microphone array by separating speech signals and noise signals into two orthogonal subspaces. The DOA is then estimated by projecting the steering vectors onto the noise subspace. MUSIC [9-10] combined with spatial smoothing [11] and [104] is one of the most popular methods for eliminating the coherence problem. However, as the experiment in Chapter 6 indicates, its robustness is still poor in a vehicular environment when the

SNR is low. Furthermore, the near-field effect [105-107] should also be considered in applications in real environments.

In some environments, especially in vehicular environments, the line-of-sight condition may not be available because, for example, barriers may exist between the speaker and the microphone array. Therefore, when a single linear array is employed, the aforementioned methods cannot distinguish speakers under non-line-of-sight conditions. Hence, multiple microphone arrays must be considered [108-109]. Further, the microphone mismatch problem often arises when such methods as steered-beamformer-based, GCC-based or spectral-estimation-based algorithms are used since these methods require the microphones to be calibrated in advance. Several sound source localization works [110-111] also mentioned the importance of calibration and the influence of microphone mismatch problem. However, accurate calibration is not easy to obtain since the characteristics of microphones vary from the sound source directions.

The relationship between a sound source and a receiver (microphone) in a complicated enclosure is almost impossible to characterize with a finite-length data in real-time applications (such as in frame-based calculations). According to the investigation of room acoustics [112], the number of eigen-frequencies with an upper limit of $f_s/2$ kHz can be obtained by the following equation:

$$\frac{4\pi}{3}B\left(\frac{f_s}{2\nu}\right)^3 \tag{5-1}$$

where f_s denotes the sampling frequency, ν represents the sound velocity ($\nu \approx 340 m/s$) and B is the geometrical volume. This equation indicates that the number of poles is too high when the frequency is high, and that the transient response

occurs in almost any processing duration when the input signal is a speech signal. For example, the number of poles is about 96435 when the sampling frequency is 8 *kHz* and the volume is $14.1385\ m^3$. Hence, the non-stationary characteristics of speech signals make the phase differences between the signals received by two elements of a single linear microphone array from a fixed sound source vary among data sets. Moreover, the stochastic nature of the phase difference is more prominent when the sound source is moving slightly and environmental noises are present. Consequently, the proposed method in this dissertation does not explicitly utilize the information of direct path from sound source to microphones to detect speaker's location, nor attempt to suppress the effect of reverberations, interference signals, and noises. Instead, this proposed method utilizes the sound field features obtained when the speaker is at different location in an indoor environment. In other words, this dissertation proposes the use of the distributions of phase differences, rather than their actual values, to locate the speaker, because the phase difference distributions vary among locations and can be distinguished by pattern matching methods. Previous researches [113-114] also showed that common acoustic measures vary significantly with small spatial displacements of the sound source or the microphone.

The experimental results in Chapter 6 indicate that the GMM [115] is very suitable for modeling these distributions. Furthermore, the model training uses the distributions of phase differences among microphones as a location-dependent but content and speaker-independent sound field feature. In this case, the geometry of the microphone array should be considered to cope with the aliasing problem and maximize the phase difference of each frequency band to detect the speaker's location accurately. Consequently, the microphone array can be decoupled into several pairs with various distances between the microphones to deal with different frequency bands. The location

detector integrates the overall probability information from different frequency bands to detect the speaker's location.

The reminder of this chapter is organized as follows. The next section introduces the overall system architecture and data flow. Section 5.3 presents the design of the location model and the model parameters estimation approach. Section 5.4 presents the proposed reference-signal-based single speaker's location detection criterion. Section 5.5 discusses an approach to find each location's testing sequence length and threshold. Section 5.6 describes the proposed reference-signal-based multiple speakers' locations detection criterion using the information of testing sequence length and threshold of each location. Conclusions are made in Section 5.7.

5.2 System Architecture

5.2.1 System Architecture

Figure 5-1 illustrates the overall system architecture. A voice activity detector divides the system into two stages, the silent stage and the speech stage. Before the proposed system is training online, a set of pre-recorded speech signals can be required via the description in Chapter 2 to obtain *a priori* information between speakers and the microphone array. The pre-recorded speech database can represent the acoustical characteristic of each location. After collecting the pre-recorded speech signals, the system switches automatically between the silent and speech stages according to the VAD result.

The first stage is called the silent stage in which speakers are silent. In this stage, environmental noises without speech are recorded online. The system combines the

online recorded environmental noise, $N_1(\omega),\cdots,N_M(\omega)$, with the pre-recorded speech database, $S_1(\omega),\cdots,S_M(\omega)$, to construct training signals, $\hat{X}_1(\omega),\cdots,\hat{X}_M(\omega)$. After that, the GM location models are derived via the location model training procedure described in Section 5.3 or 5.5. Since the environmental noise alters, the GM location models that contain the characteristics of environmental noise are updated to ensure the detection accuracy and robustness in this stage. The second stage is the speech stage, in which the parameters of GM location models derived from the first stage are duplicated into the location detector to detect the speaker's location.

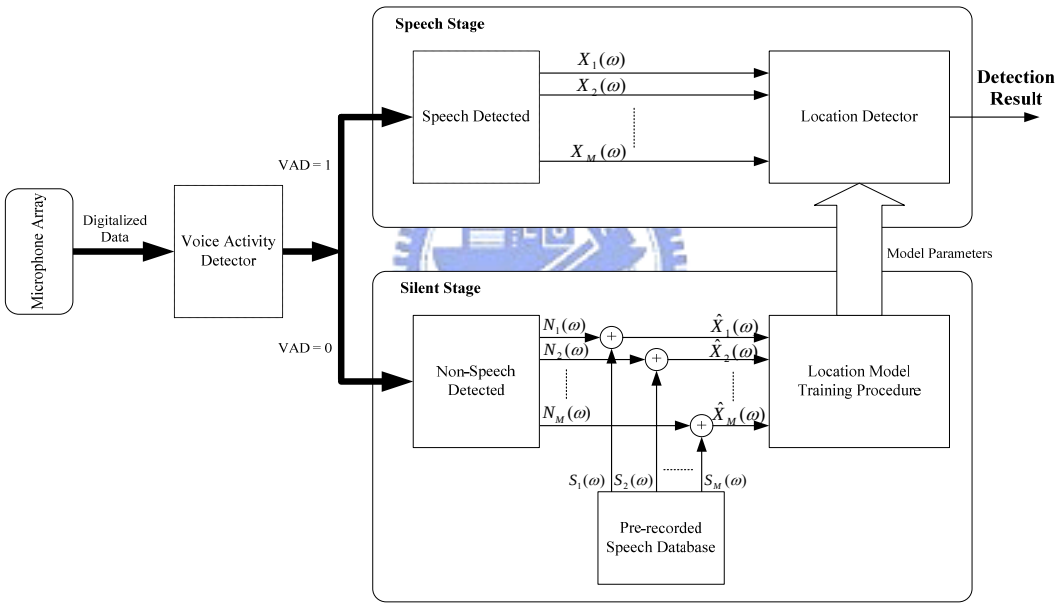


Figure 5-1 Proposed reference-signal-based speaker's location detection system architecture

5.2.2 Frequency Band Divisions based on a Uniform Linear Microphone Array

The phase difference of the received signal becomes more significant as the distance between microphones increases. However, the aliasing problem occurs when this distance exceeds half of the minimum wavelength of the received signal [116]. The

distance between pairs of microphones should be chosen based on the selected frequency band to obtain clear phase difference data to enhance the accuracy of location detection and prevent aliasing.

Figure 5-2 illustrates a uniform microphone array with M microphones and the distance of d. According to the geometry, the training frequency range is divided into $(M-1)$ bands listed in Table 5-1, where m denotes the mth microphone; b represents the band number, v denotes the sound velocity, and J_b is the number of microphone pairs in the band of b. The phase differences measured by the microphone pairs at each frequency component, ω (belonging to a specific band, b) are utilized to generate a GM location model with the dimension of J_b.

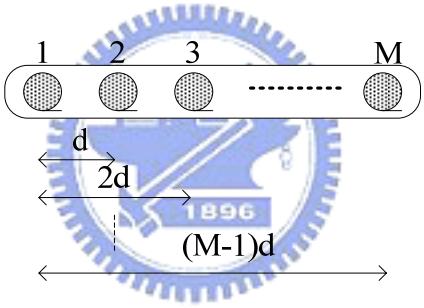


Figure 5-2 Microphone array geometry

Table 5-1 Relationship of Frequency Bands to the Microphone Pairs

Frequency Band	Microphone Pairs	The Number of Microphone Pair	The Range of Frequency Band
Band 1 ($b=1$)	$(m, m+M-1)$ with $m=1$	$J_b = J_1 = 1$	$0 \le \omega \le \frac{v}{2(M-1)d}$
Band 2 ($b=2$)	$(m, m+M-2)$ with $1 \le m \le 2$	$J_b = J_2 = 2$	$\frac{v}{2(M-1)d} < \omega \le \frac{v}{2(M-2)d}$
⋮	⋮	⋮	⋮
Band $M-1$ ($b=M-1$)	$(m, m+1)$ with $1 \le m \le M-1$	$J_b = J_{M-1} = M-1$	$\frac{v}{4d} < \omega \le \frac{v}{2d}$

5.3 Location Model Description and Parameters Estimation

5.3.1 GM Location Model Description

If the GM location model at location l is represented by the parameter $\boldsymbol{\lambda}(l)=\{\boldsymbol{\lambda}(\omega,1,l),\cdots,\boldsymbol{\lambda}(\omega,M-1,l)\}$, then a group of L GM location models can be represented by the parameters, $\{\boldsymbol{\lambda}(1),\cdots,\boldsymbol{\lambda}(L)\}$. A Gaussian mixture density in the band b at location l can be denoted as a weighted sum of N Gaussian component densities:

$$G_b\left(\boldsymbol{P}_{\hat{X}}(\omega,b,l)\,|\,\boldsymbol{\lambda}(\omega,b,l)\right)=\sum_{i=1}^{N}\rho_i(\omega,b,l)g_i\left(\boldsymbol{P}_{\hat{X}}(\omega,b,l)\right) \tag{5-2}$$

where $\boldsymbol{P}_{\hat{X}}(\omega,b,l)=\left[P_{\hat{X}}(\omega,1,l)\ \cdots\ P_{\hat{X}}(\omega,J_b,l)\right]^T$ is a J_b-dimensional training phase difference vector derived from the training signals, $\hat{X}_1(\omega),\cdots,\hat{X}_M(\omega)$. $\rho_i(\omega,b,l)$ is the ith mixture weight and each component of the training phase difference vector can be obtained as follows:

$$P_{\hat{X}}(\omega,m,l)=phase\left(\hat{X}_{m+M-b}(\omega,b,l)\right)-phase\left(\hat{X}_m(\omega,b,l)\right)\quad \text{with } 1\le m\le b \tag{5-3}$$

where $\hat{X}_m(\omega,b,l)$ denotes constructed training signal of mth microphone in the band b at location l. The GM location model parameter in the band b at location l, $\boldsymbol{\lambda}(\omega,b,l)$, is constructed by the mean matrix, covariance matrices and mixture weights vector from N Gaussian component densities:

$$\boldsymbol{\lambda}(\omega,b,l)=\{\boldsymbol{\rho}(\omega,b,l),\boldsymbol{\mu}(\omega,b,l),\boldsymbol{\Sigma}(\omega,b,l)\} \tag{5-4}$$

where

$\boldsymbol{\rho}(\omega,b,l)=\left[\rho_1(\omega,b,l)\ \cdots\ \rho_N(\omega,b,l)\right]$ denotes the mixture weights vector in the

band b at location l.

$\boldsymbol{\mu}(\omega,b,l)=[\boldsymbol{\mu}_1(\omega,b,l) \quad \cdots \quad \boldsymbol{\mu}_N(\omega,b,l)]$ denotes the mean matrix in the band b at location l.

$\boldsymbol{\Sigma}(\omega,b,l)=[\boldsymbol{\Sigma}_1(\omega,b,l) \quad \cdots \quad \boldsymbol{\Sigma}_N(\omega,b,l)]$ denotes the covariance matrix in the band b at location l.

The i th corresponding vector and matrix of the parameters defined above are

$\boldsymbol{\mu}_i(\omega,b,l)=[\mu_i(\omega,1,l) \quad \cdots \quad \mu_i(\omega,J_b,l)]^T$ and

$$\boldsymbol{\Sigma}_i(\omega,b,l)=\begin{bmatrix}\sigma_i^2(\omega,1,l) & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_i^2(\omega,J_b,l)\end{bmatrix}.$$

$g_i\left(\boldsymbol{P}_{\hat{X}}(\omega,b,l)\right)$ denotes the i th Gaussian component density in the band b at location l:

$$g_i\left(\boldsymbol{P}_{\hat{X}}(\omega,b,l)\right)=\frac{\exp\left(-\frac{1}{2}\left[\boldsymbol{P}_{\hat{X}}(\omega,b,l)-\boldsymbol{\mu}_{\mathbf{i}}(\omega,b,l)\right]^T\boldsymbol{\Sigma}_{\mathbf{i}}(\omega,b,l)^{-1}\left[\boldsymbol{P}_{\hat{X}}(\omega,b,l)-\boldsymbol{\mu}_{\mathbf{i}}(\omega,b,l)\right]\right)}{(2\pi)^{J_b/2}\left|\boldsymbol{\Sigma}_{\mathbf{i}}(\omega,b,l)\right|^2} \tag{5-5}$$

Notably, the mixture weight must satisfy the constraint that

$$\sum_{i=1}^{N}\rho_i(\omega,b,l)=1 \tag{5-6}$$

The covariance matrix, $\boldsymbol{\Sigma}_i(\omega,b,l)$, is selected as a diagonal matrix. Although the phase differences of the microphone pairs may not be statistically independent of each

other, GMMs with diagonal covariance matrices have been observed to be capable of modeling the correlations within the data by increasing mixture number [117].

5.3.2 Parameters Estimation via EM Algorithm

The purpose is to determine the L GM location models, $\{\boldsymbol{\lambda}(1),\cdots,\boldsymbol{\lambda}(L)\}$, from the measured phase differences between each microphone pair in band b. Several techniques are available for estimating $\boldsymbol{\lambda}(l)$, of which the most popular is the EM algorithm [115] that estimates the parameters by using an iterative scheme to maximum the log-likelihood function. The EM algorithm can guarantee a monotonic increase in the model's log-likelihood value, and its iteration equations corresponding to frequency band selection can be arranged as:

Expectation step:

$$G_b\left(i \mid \boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l),\boldsymbol{\lambda}(\omega,b,l)\right)=\frac{\rho_i(\omega,b,l)g_i\left(\boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l)\right)}{\sum_{i=1}^{N}\rho_i(\omega,b,l)g_i\left(\boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l)\right)} \tag{5-7}$$

where $G_b\left(i \mid \boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l),\boldsymbol{\lambda}(\omega,b,l)\right)$ is a *posteriori* probability and $\mathbf{P}_{\hat{X}}(\omega,b,l)=\left\{\boldsymbol{P}_{\hat{X}}{}^{(1)}(\omega,b,l),\cdots,\boldsymbol{P}_{\hat{X}}{}^{(T)}(\omega,b,l)\right\}$ is a sequence of T input phase difference vectors.

Maximization step:

(i). Estimate the mixture weights:

$$\rho_i(\omega,b,l)=\frac{1}{T}\sum_{t=1}^{T}G_b\left(i \mid \boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l),\boldsymbol{\lambda}(\omega,b,l)\right) \tag{5-8}$$

(ii). Estimate the mean vector:

$$\boldsymbol{\mu}_i(\omega,b,l)=\frac{\sum_{t=1}^{T} G_b\left(i \mid \boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l),\boldsymbol{\lambda}(\omega,b,l)\right)\boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l)}{\sum_{t=1}^{T} G_b\left(i \mid \boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l),\boldsymbol{\lambda}(\omega,b,l)\right)} \tag{5-9}$$

(iii). Estimate the variances:

$$\sigma_i^2(\omega,j,l)=\frac{\sum_{t=1}^{T} G_b\left(i \mid \boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l),\boldsymbol{\lambda}(\omega,b,l)\right)P_{\hat{X}}{}^{(t)2}(\omega,j,l)}{\sum_{t=1}^{T} G_b\left(i \mid \boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l),\boldsymbol{\lambda}(\omega,b,l)\right)}-\mu_i{}^2(\omega,j,l) \tag{5-10}$$

where $i=\{1,\cdots,N\}$ and $j=\{1,\cdots,J_b\}$.

However, the EM algorithm only guarantees to find a local maximum log-likelihood model. A different choice of initial model $\boldsymbol{\lambda}_0(\omega,b,l)$ leads to various local maximum models. This work considered two initialization methods to find out the initial model. K-means [118] is by far the most widely-used method. Elkan [119] proposed an accelerated K-means algorithm which utilizes the triangle inequality to decrease significantly the computational effort. Charles' method is also suitable for finding a good initial model to lower the iteration number of the EM algorithm. The first method utilizes the accelerated K-means clustering method. The second method separates phase difference range, $\{-\pi,\pi\}$, into N segments to obtain a fixed initial mean model since the phase difference range is small enough. Consequently, the initial mean model is $\{-\pi \quad \frac{2\pi}{N-1}-\pi \quad \frac{4\pi}{N-1}-\pi \quad \cdots \quad \pi\}$. The location detection performances of the two initial approaches have slightly different performance and no one is always the best.

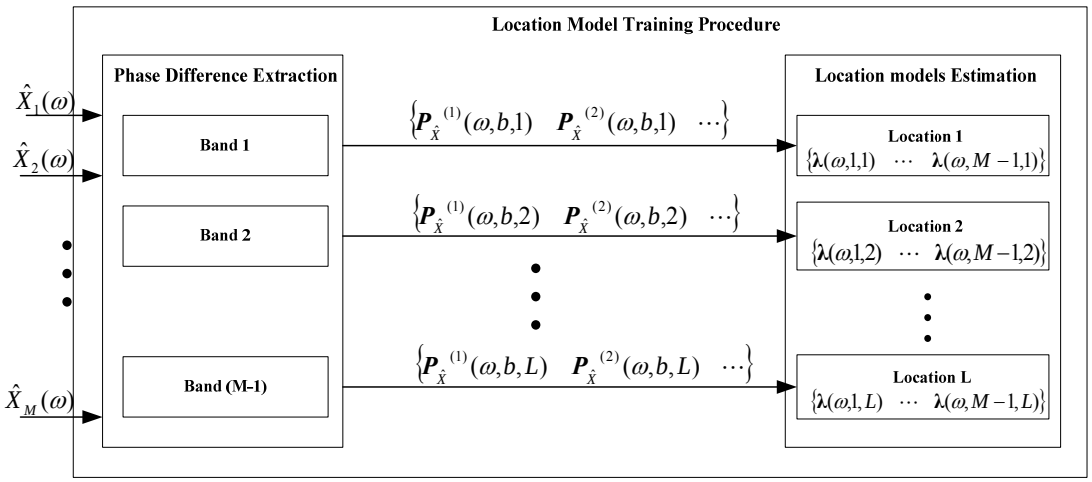


Figure 5-3 Location model training procedure with the total location number L

5.4 Single Speaker's Location Detection Criterion

The location is determined by finding the GM location model which has the maximum *posteriori* probability for a given observation sequences:

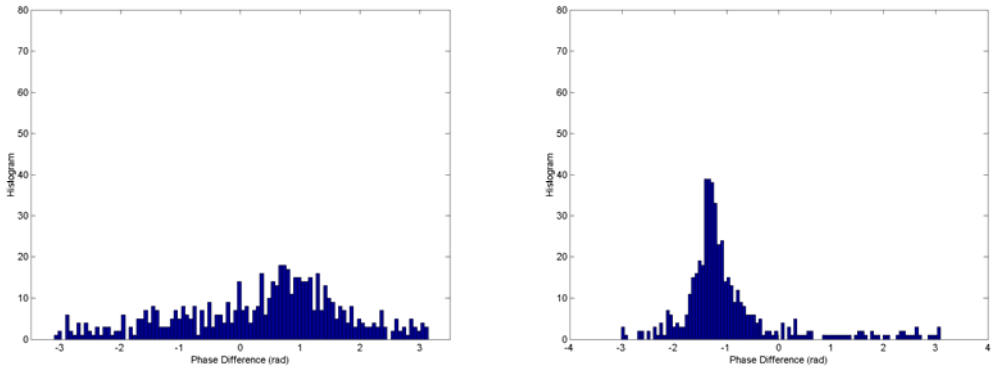
$$\begin{aligned}\hat{l} &= \arg\max_{1\le l\le L}\sum_{b=1}^{M-1}\log\left[G_b\left(\boldsymbol{\lambda}(\omega,b,l)\,|\,\mathbf{P}_X(\omega,b)\right)\right] \\ &= \arg\max_{1\le l\le L}\sum_{b=1}^{M-1}\log\frac{G_b\left(\mathbf{P}_X(\omega,b)\,|\,\boldsymbol{\lambda}(\omega,b,l)\right)p\left(\boldsymbol{\lambda}(\omega,b,l)\right)}{p\left(\mathbf{P}_X(\omega,b)\right)}\end{aligned} \tag{5-11}$$

where $\mathbf{P}_X(\omega,b)=\left\{\boldsymbol{P}_X{}^{(1)}(\omega,b),\cdots,\boldsymbol{P}_X{}^{(Q)}(\omega,b)\right\}$ is a phase difference testing sequence derived from $X_1(\omega),\cdots,X_M(\omega)$, and Q denotes the length of the testing sequence. If the probability densities at all locations are equally likely, then $p\left(\boldsymbol{\lambda}(\omega,b,l)\right)$ could be chosen as $1/L$. The probability $p\left(\mathbf{P}_X(\omega,b)\right)$ is the same for all location models and the detection rule can be rewritten as:

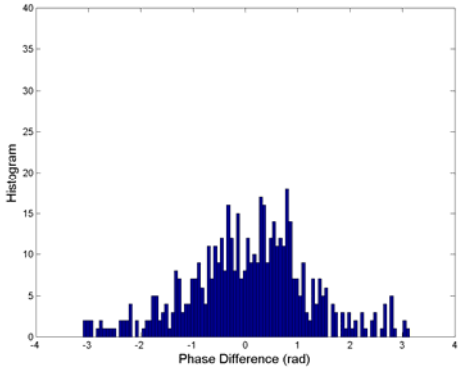
$$\hat{l} = \arg\max_{1\le l\le L}\sum_{b=1}^{M-1}\sum_{q=1}^{Q}\log G_b\left(\boldsymbol{P}_X{}^{(q)}(\omega,b)\,|\,\boldsymbol{\lambda}(\omega,b,l)\right) \tag{5-12}$$

5.5 Testing Sequence Lengths and Thresholds Estimation

First, the work in Section 5.4 assumed that the speech signals are emitted from one of the previously modeled locations. Consequently, an unmodeled speech signal which is not emitted from one of the modeled locations, such as the radio broadcasting from the in-car audio system and the speaker's voices from unmodeled locations, degrades the performance. The unmodeled speech signal could trigger the VAD, resulting in an incorrect detection of the speaker location. Therefore, a method that can prevent the detection errors without modifying the VAD approach is necessary. Second, the work in Section 5.4 cannot detect multiple speakers' locations. If the speech signals from various modeled locations are mixed together, then the derived phase difference distribution becomes an unmodeled distribution, leading to a detection error. Figure 5-4 shows the example of the phase difference distribution from two simultaneously speaking passengers at locations No. 1 and 2 which is not similar to the one from location No. 1 or 2, and thus may lead to a detection error.



(a). Location No. 1 (frequency = 0.9375 *kHz*) (b). Location No. 2 (frequency = 0.9375 *kHz*)



(c). Locations No. 1 and 2 (frequency = 0.9375 *kHz*)

Figure 5-4 The histograms of phase differences at locations No. 1, 2, and 1 and 2 between the third and the sixth microphones at a frequency of 0.9375 kHz.

Moreover, how to find a suitable length of testing sequence that could significantly affect the location detection performance is not discussed in Section 5.4. This section proposes a new threshold-based location detection approach that utilizes the training signals and the trained GM location model parameters to determine the length of testing sequence and then obtain a threshold of the *a posteriori* probability for each location to resolve the two issues mentioned above.

Since conversational speech contains many short pauses, Potamitis *et al.* [103] locates multiple speakers by detecting the direction of individual speaker when the frame is originated from a single speaker. For example, Fig. 5-5 shows a two people conversation condition. Based on this concept, this dissertation proposes a threshold-based location detection approach to determine whether a speech frame originates from a single speaker or from simultaneously active speakers. This approach identifies the frames in which probably only one speaker is talking, and returns a valid location detection result. Moreover, because each location has specific acoustical characteristics, the threshold at each location can be used to determine whether it

represents the radio broadcasting or speech signals coming from unmodeled or modeled locations.

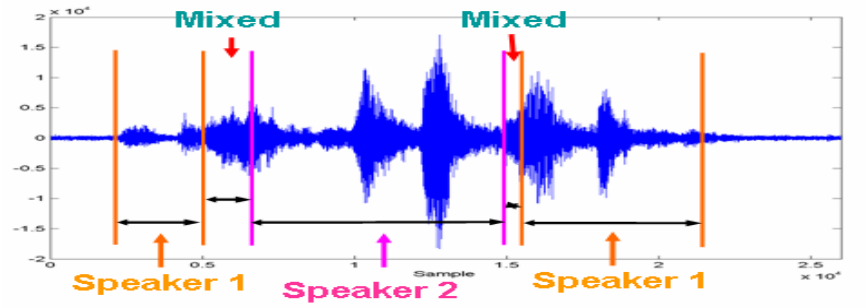


Figure 5-5 A two people conversation condition

The lengths of testing sequences and thresholds can be derived using the estimated parameters of the L GM location models. The most suitable length of testing sequences at location l is denoted as $\hat{Q}(l)$, the threshold at location l is denoted as $Thd(l)$, and the possible searching range of the length of the testing sequence is set to $\lfloor Q_{Lo}, Q_{Up} \rfloor$. T is the total length of the training phase difference sequence. $\mathbf{P}_{\hat{X},Q}(\omega,b,l,t) = \{\boldsymbol{P}_{\hat{X}}{}^{(t)}(\omega,b,l), \cdots, \boldsymbol{P}_{\hat{X}}{}^{(t+Q-1)}(\omega,b,l)\}$ is a sequence of Q training phase difference vectors, where $1 \le t \le T-Q+1$. $\hat{Q}(l)$ and $Thd(l)$ can be obtained using the following criterions:

$$\hat{Q}(l) = \arg \max_{Q_{Lo} \le Q \le Q_{Up}} \{C(Q)\} \tag{5-13}$$

where

$$C(Q) = \alpha\left[Lo\left(\boldsymbol{\lambda}(l), \mathbf{P}_{\hat{X}}(l), Q\right) - Up\left(\boldsymbol{\lambda}(l), \mathbf{P}_{\hat{X}}(l), Q\right)\right] + \beta \sum_{\substack{i=1 \\ i \ne l}}^{L} I\left[Lo\left(\boldsymbol{\lambda}(l), \mathbf{P}_{\hat{X}}(l), Q\right) - Up\left(\boldsymbol{\lambda}(i), \mathbf{P}_{\hat{X}}(l), Q\right)\right] + \gamma Lo\left(\boldsymbol{\lambda}(l), \mathbf{P}_{\hat{X}}(l), Q\right)$$

with $\alpha+\beta+\gamma=1$ (5-14)

and

$$Thd(l)=Lo\left(\boldsymbol{\lambda}(l),\mathbf{P}_{\hat{X}}(l),\hat{Q}(l)\right)/\hat{Q}(l) \tag{5-15}$$

where α,β,γ are weights, and $I(k)=\begin{cases}k & if\ k\geq 0\\ -\infty & if\ k<0\end{cases}$.

$Up\left(\boldsymbol{\lambda}(l),\mathbf{P}_{\hat{X}}(l),Q\right)$ and $Lo\left(\boldsymbol{\lambda}(l),\mathbf{P}_{\hat{X}}(l),Q\right)$ denote the probability upper bound and lower bound when the length of the training phase difference sequence is Q. They are derived from the following equations:

$$Up\left(\boldsymbol{\lambda}(l),\mathbf{P}_{\hat{X}}(l),Q\right)=\max_{\forall t}\sum_{b=1}^{M-1}\log\left[G_b\left(\boldsymbol{\lambda}(\omega,b,l)\right)|\,\mathbf{P}_{\hat{X},Q}(\omega,b,l,t)\right] \tag{5-16}$$

$$Lo\left(\boldsymbol{\lambda}(l),\mathbf{P}_{\hat{X}}(l),Q\right)=\min_{\forall t}\sum_{b=1}^{M-1}\log\left[G_b\left(\boldsymbol{\lambda}(\omega,b,l)\right)|\,\mathbf{P}_{\hat{X},Q}(\omega,b,l,t)\right] \tag{5-17}$$

where

$$\log\left[G_b\left(\boldsymbol{\lambda}(\omega,b,l)|\,\mathbf{P}_{\hat{X},Q}(\omega,b,l,t)\right)\right]=\log\left[\frac{G_b\left(\mathbf{P}_{\tilde{X},Q}(\omega,b,l,t)|\,\boldsymbol{\lambda}(\omega,b,l)\right)p(\boldsymbol{\lambda}(\omega,b,l))}{p\left(\mathbf{P}_{\hat{X},Q}(\omega,b,l,t)\right)}\right].$$

The term $p(\boldsymbol{\lambda}(\omega,b,l))$ could be eliminated because $p(\boldsymbol{\lambda}(\omega,b,l))$ is independent to t and the probability $p\left(\mathbf{P}_{\hat{X},Q}(\omega,b,l,t)\right)$ is the same for all t. Therefore, Eqs. (5-16) and (5-17) can be rewritten as:

$$Up\left(\boldsymbol{\lambda}(l),\mathbf{P}_{\hat{X}}(l),Q\right)=\max_{\forall t}\sum_{b=1}^{M-1}\sum_{q=0}^{Q-1}\log G_b\left(\boldsymbol{P}_{\hat{X}}{}^{(t+q)}(\omega,b,l)|\,\boldsymbol{\lambda}(\omega,b,l)\right) \tag{5-18}$$

$$Lo\left(\boldsymbol{\lambda}(l),\mathbf{P}_{\hat{X}}(l),Q\right)=\min_{\forall t}\sum_{b=1}^{M-1}\sum_{q=0}^{Q-1}\log G_b\left(\boldsymbol{P}_{\hat{X}}^{(t+q)}(\omega,b,l)\,|\,\boldsymbol{\lambda}(\omega,b,l)\right) \tag{5-19}$$

The first term of Eq. (5-14) represents the negative maximum probability variation of the trained model when the length of the training phase difference sequence is Q. As the value of this term increases, the corresponding selection of Q yields a more robust result under the trained GM location model. The second term of Eq. (5-14) is the sum of the probability differences of the location l versus other locations and a larger value means the corresponding selection of Q has a higher discrimination level between the location l and the other trained GM locations. Finally, a high discrimination level between the location l and other unmodeled locations can be achieved if the third term of Eq. (5-14) is large. Figure 5-6 shows the GM location model training procedure with the total location number L.

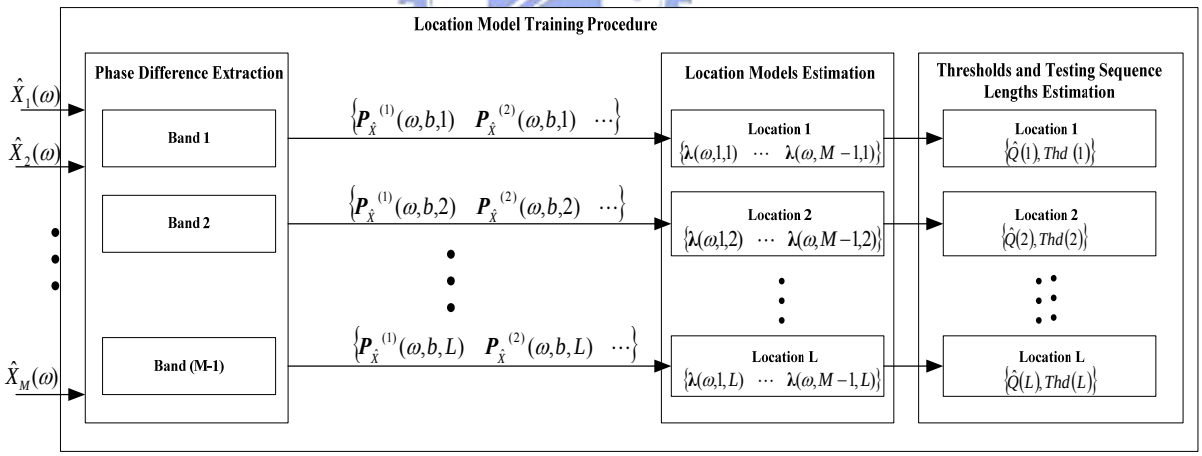


Figure 5-6 Location model training procedure with testing sequence length and thresholds estimation

5.6 Multiple Speakers' Locations Detection Criterion

The location is detected as,

$$\hat{l} = \arg\max_{1\le l\le L} \frac{1}{\hat{Q}(l)} \sum_{b=1}^{M-1} \log\left[G_b\left(\boldsymbol{\lambda}(\omega,b,l) \mid \mathbf{P}_X(\omega,b,l)\right)\right]$$
$$= \arg\max_{1\le l\le L} \sum_{b=1}^{M-1} \log \frac{G_b\left(\mathbf{P}_X(\omega,b,l) \mid \boldsymbol{\lambda}(\omega,b,l)\right) p\left(\boldsymbol{\lambda}(\omega,b,l)\right)}{\hat{Q}(l)\, p\left(\mathbf{P}_X(\omega,b,l)\right)}$$
$$\text{if } Thd\left(\arg\max_{1\le l\le L} \frac{1}{\hat{Q}(l)} \sum_{b=1}^{M-1} \log\left[G_b\left(\boldsymbol{\lambda}(\omega,b,l) \mid \mathbf{P}_X(\omega,b,l)\right)\right] \right) \le \max_{1\le l\le L} \frac{1}{\hat{Q}(l)} \sum_{b=1}^{M-1} \log\left[G_b\left(\boldsymbol{\lambda}(\omega,b,l) \mid \mathbf{P}_X(\omega,b,l)\right)\right]$$

(5-20)

where $\mathbf{P}_X(\omega,b,l) = \left\{ \boldsymbol{P}_X{}^{(1)}(\omega,b), \cdots, \boldsymbol{P}_X{}^{(\hat{Q}(l))}(\omega,b) \right\}$ is a testing sequence derived from $X_1(\omega), \cdots, X_M(\omega)$. If the probability densities at all locations are equally likely, then $p\left(\boldsymbol{\lambda}(\omega,b,l)\right)$ could be chosen as $1/L$. The probability $p\left(\mathbf{P}_X(\omega,b,l)\right)$ is the same for all location models and then the detection rule can be rewritten as

$$\hat{l} = \arg\max_{1\le l\le L} \frac{1}{\hat{Q}(l)} \sum_{b=1}^{M-1} \sum_{q=1}^{\hat{Q}(l)} \log\left[G_b\left(\boldsymbol{P}_X{}^{(q)}(\omega,b) \mid \boldsymbol{\lambda}(\omega,b,l)\right)\right]$$
$$\text{if } Thd\left(\arg\max_{1\le l\le L} \frac{1}{\hat{Q}(l)} \sum_{b=1}^{M-1} \sum_{q=1}^{\hat{Q}(l)} \log\left[G_b\left(\boldsymbol{P}_X{}^{(q)}(\omega,b) \mid \boldsymbol{\lambda}(\omega,b,l)\right)\right] \right) \le \max_{1\le l\le L} \frac{1}{\hat{Q}(l)} \sum_{b=1}^{M-1} \sum_{q=1}^{\hat{Q}(l)} \log\left[G_b\left(\boldsymbol{P}_X{}^{(q)}(\omega,b) \mid \boldsymbol{\lambda}(\omega,b,l)\right)\right]$$

(5-21)

If the value of $\max_{1\le l\le L} \sum_{b=1}^{M-1} \sum_{q=1}^{\hat{Q}(l)} \log\left[G_b\left(\boldsymbol{P}_X{}^{(q)}(\omega,b) \mid \boldsymbol{\lambda}(\omega,b,l)\right)\right] \Big/ \hat{Q}(l)$ is not larger than the corresponding threshold, then the frames may contain speech components that come simultaneously from multiple modeled locations or from unmodeled locations.

5.7 Summary

This chapter proposes reference-signal-based single speaker's location and multiple speakers' locations detection methods. The GM location models which are constructed by the location dependent features, phase differences. The proposed methods can overcome practical issues, such as the microphone mismatch, near-field effect, local

scattering, and coherence problems. The proposed methods are found out to work even under non-line-of-sight conditions and when speakers are in the same direction but different distances from the microphone array.

Additionally, the proposed threshold adaptation approach computes a suitable length of testing sequence and a threshold for each modeled location. Experimental results in Chapter 6 show that the speaker's location detection approach with these two adapted parameters performs well on detecting multiple speakers' locations and reducing the average error rates caused by the unmodeled locations at various SNRs.

Chapter 6

Experimental Results

This chapter provides simulation and practical environmental results to assess the capability of the reference-signal-based adaptive beamformers and speaker's location detection approaches proposed in this dissertation. In these experiments, the sampling frequency is set to 8 *kHz* and the amplified microphone signals are digitized by 16-bit AD converters. The processed frame window for STFT contained 256 zero padding samples and 32ms speech signals, totaling 512 samples. Figure 6-1 illustrates the processed frame window and the overlapping condition. The pre-recorded speech signals are acquired by locating a loudspeaker on the speech location and the reference signal is obtained from the original speech source that emitted from the loudspeaker.

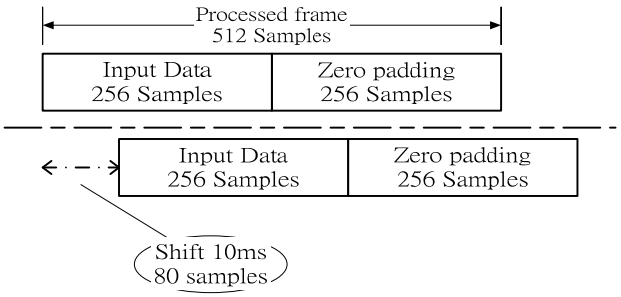


Figure 6-1 Processed frame window and overlapping condition

The remainder of this chapter is organized as follows. Section 6.1 utilizes the ASR rates and the two frequency-domain performance indexes introduced in Chapter 3 to show the advantages of the proposed reference-signal-based frequency-domain beamformer, SPFDBB and FDABB. Section 6.2 compares the robustness of the NLMS and H_∞ adaptation criterions in the time domain through the simulation results. Section 6.2 also utilizes the ASR rates in both vehicular and indoor environments to prove the advantages of H_∞ adaptation criterion. The experimental results containing the location detection performance in single and multiple speakers' cases, and the cases of radio broadcasting and speech from unmodeled locations are discussed in Section 6.3. Finally, conclusions are made in Section 6.4

6.1 Adaptive Beamformers Using NLMS Adaptation Criterion

6.1.1 Simulation Results

In this simulation, a speech source and two noises, a white noise and a music signal are considered and the linear array contains six microphones. The speech source comes from 0° relative to the linear array, and the white noise and the music signal come from -30° and 60° respectively. Figure 6-2 illustrates the arrangement of the microphone array and the sources. The value of γ is 10^{-6} and the step size λ of 0.4 is selected. Two simulations are shown: the first one is performed to compare the performance among different parameters, and the second one is performed to observe the adaptation performance of FDABB in a sudden change of the noise channel, which the noise moves from -30° to -60°. Moreover, the two simulations are executed in three environments specified by different channel response durations: 1024, 2048, and 3072 taps.

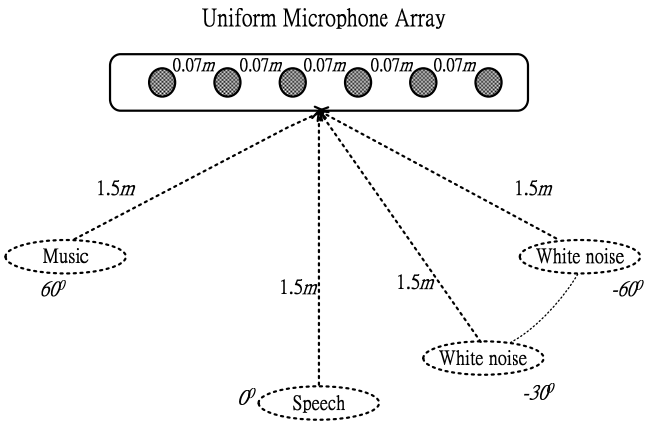


Figure 6-2 Arrangement of microphone array, noies and speech source in simulation experiments

In the first simulation, the locations of the speech source and the noises are fixed in the overall training data length. The soft penalty parameter μ has three options, which are 0, 2, and 4. The frame number L in a block varies from 10 to 20, and 20 to 30 corresponding to different soft penalty parameters and channel response durations. Two frequency-domain performance indexes, NSR and SDR, of the most significant frequency, 410Hz, are shown in Tables 6-1, 6-2, and 6-3. The values shown in Tables 6-1, 6-2, and 6-3 are computed by averaging the last 120 frames. The notation *ADL* indicates that the value of L is adjusted by the *CBVI* with a lower threshold of 0.02, an upper threshold of 1.2, and the initial frame number 10. In other words, if the *CBVI* is smaller than 0.02, the value of L will be increased. On the contrary, if the *CBVI* is larger than 1.2, the value of L will be reset to the initial frame number. Additionally, Table 6-1 summarizes the related parameters of FDABB. Figure 6-3 depicts the NSR and the SDR from C6 to C9 with channel response duration 1024 shown in Table 6-2. Figure 6-3 shows that the measurement index in the condition with $L=1$ varies heavily than the one in the conditions with $L=10$, $L=20$, and $L=30$; that is, the performance of the NSR and the SDR cannot be guaranteed even when the algorithm is run for a long time. From Tables 6-1 to 6-3, the SDR and the NSR become worse as the

channel response duration grows, but the proposed beamformer with a larger value of L would have smaller performance decay and have better convergence performance. The SDRs of SPFDBB with $L=10$, $L=20$, and $L=30$ in the condition of $\mu=2$ has decreased from about 1.84dB to 4.52dB as compared with those in the conditions of $\mu=0$. Although the NSR increases at the same time as the SDR fell, the SDR decreasing rate is more important for ASR applications when the NSR is very low, especially when a larger value of L is chosen.

Table 6-1 The First Simulation Experiment: Soft Penalty Parameter is 0

		Channel response duration 1024		Channel response duration 2048		Channel response duration 3072	
Condition	*L*	**NSR(dB)**	**SDR(dB)**	**NSR(dB)**	**SDR(dB)**	**NSR(dB)**	**SDR(dB)**
C1	*L* = 1	-73.88	-46.67	-60.92	-42.97	-57.55	-16.97
C2	*L* = 10	-102.82	-47.98	-91.53	-46.33	-90.60	-45.37
C3	*L* = 20	-111.00	-48.85	-98.45	-47.28	-97.12	-46.08
C4	*L* = 30	-122.92	-50.39	-112.57	-49.60	-105.32	-48.69
C5	*ADL*	-122.40	-50.10	-113.42	-49.87	-106.32	-48.50

Table 6-2 The First Simulation Experiment: Soft Penalty Parameter is 2

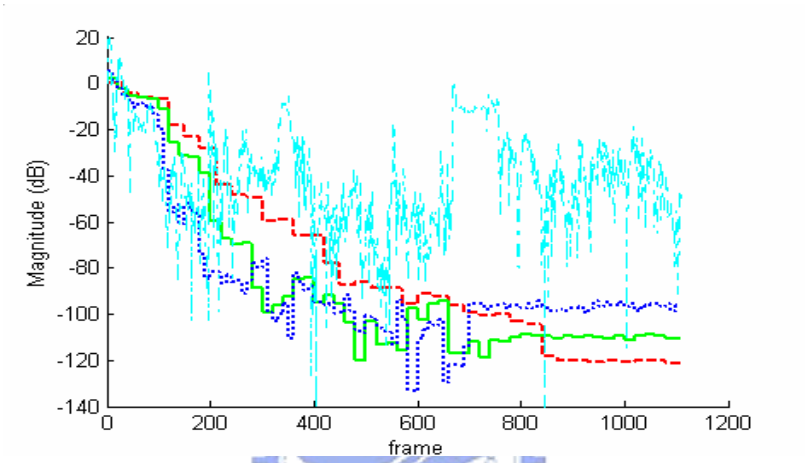
		Channel response duration 1024		Channel response duration 2048		Channel response duration 3072	
Condition	*L*	**NSR(dB)**	**SDR(dB)**	**NSR(dB)**	**SDR(dB)**	**NSR(dB)**	**SDR(dB)**
C6	*L* = 1	-65.03	-44.20	-45.77	-42.35	-46.96	-14.50
C7	*L* = 10	-97.97	-49.82	-89.67	-49.80	-92.87	-47.59
C8	*L* = 20	-110.32	-51.16	-100.32	-50.62	-96.25	-48.96
C9	*L* = 30	-120.92	-52.80	-109.27	-52.24	-105.24	-52.13
C10	*ADL*	-125.34	-52.31	-110.27	-52.21	-105.07	-52.11

Table 6-3 The First Simulation Experiment: Soft Penalty Parameter is 4

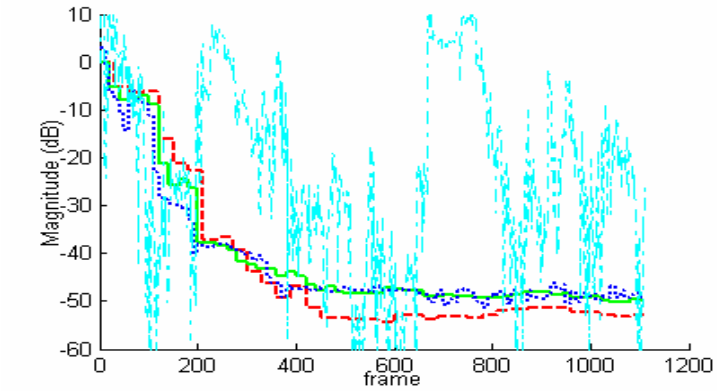
		Channel response duration 1024		Channel response duration 2048		Channel response duration 3072	
Condition	*L*	**NSR(dB)**	**SDR(dB)**	**NSR(dB)**	**SDR(dB)**	**NSR(dB)**	**SDR(dB)**
C11	*L* = 1	-46.08	-29.40	-42.91	-27.39	-37.27	3.06
C12	*L* = 10	-93.07	-50.06	-85.35	-49.85	-91.03	-48.08
C13	*L* = 20	-109.65	-52.54	-95.39	-52.01	-96.00	-50.37
C14	*L* = 30	-120.56	-54.02	-102.04	-53.87	-105.18	-53.21
C15	*ADL*	-121.71	-53.97	-102.62	-53.92	-105.59	-53.59

Table 6-4 Parameters of the FDABB

Length of STFT	512 Samples
Length of Input data in a frame	256 Samples
Shift of STFT	80 Samples
Window function	Hamming
Initial block value	10
Block value increment	10
Threshold of *CBVI*	0.02 and 1.2



(a) NSR



(b) SDR

Figure 6-3 NSR and SDR form C6 to C9 with channel response duration 1024. The dash-dot line represents C6 ($L = 1$), the dot line represents C7 ($L = 10$), the straight line represents C8 ($L = 20$), and the dash line represents C9 ($L = 30$)

In the first simulation, FDABB adjusts the frame number twice from 10 to 30; first at frame 261 and then at frame 621. Figure 6-4 illustrates the adaptation of *CBVI*.

Figure 6-5 shows the NSR and the SDR from C7 to C10 with channel response duration 1024 shown in Table 6-2. Since the initial frame number of FDABB is 10, the SDR and the NSR of FDABB are equivalent to the dash line in the first 261 samples. Obviously, the FDABB could not only perform well in a shorter adaptation process but could also obtain a good convergence result. Since SPFDBB adopts the soft penalty, it emphasize on the SDR improvement than the NSR. Consequently, the SDR of $L = 30$ is better than the SDR of $L = 10$ after frame 300 and the convergence period of the SDR is shorter than that of the NSR.

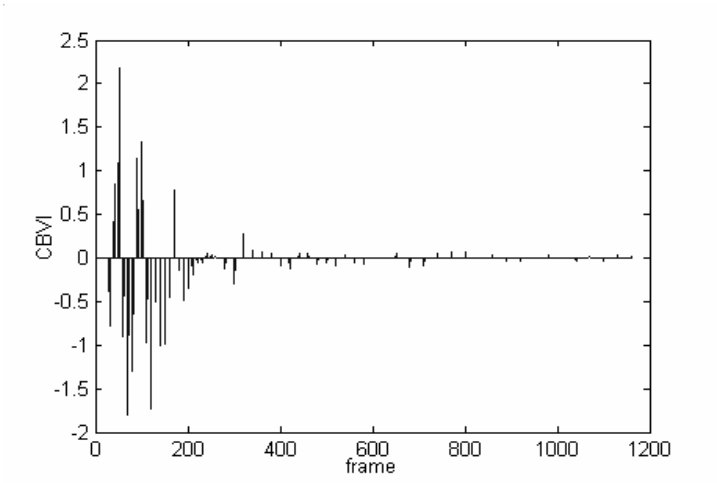
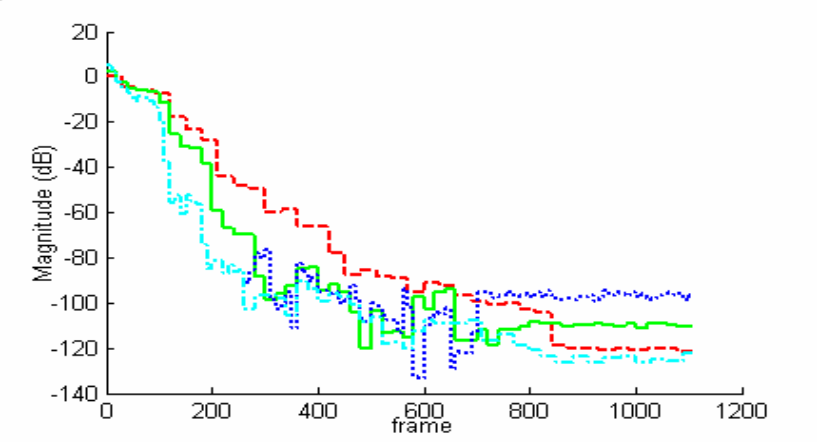
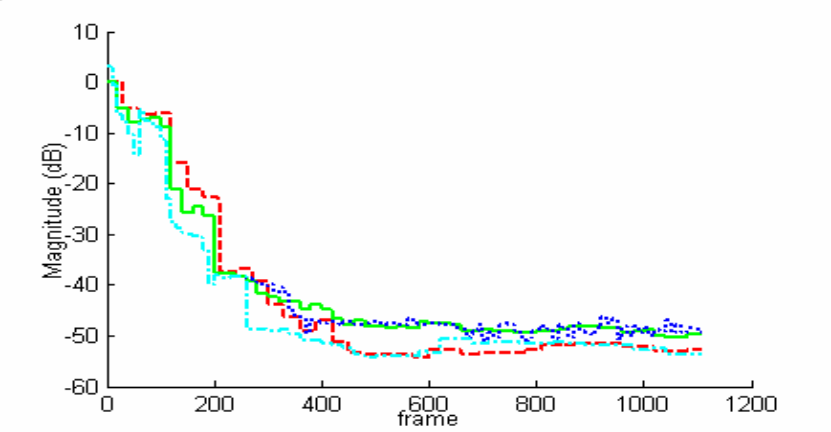


Figure 6-4 *CBVI* in the first simulation experiment



(a) NSR



(b) SDR

Figure 6-5 NSR and SDR form C7 to C10 with channel response duration 1024. The dash-dot line represents C10 (*ADL*), the dot line represents C7 ($L = 10$), the straight line represents C8 ($L = 20$), and the dash line represents C9 ($L = 30$)

In the second simulation, the location of white noise varies from -30° to -60° during the training data sequence. As shown in the Figs. 6-6 and 6-7, *CBVI* and the NSR both exhibit a big jump at frame 601 in response to the noise channel variation. Since the impulse response of the speech source is fixed, the SDR has a little variation. After this sudden change is detected, FDABB resets the value of L to the initial frame number to perform advanced adaptation of the noise channel and changes the frame number at frame 771 and frame 851 to maintain convergence.

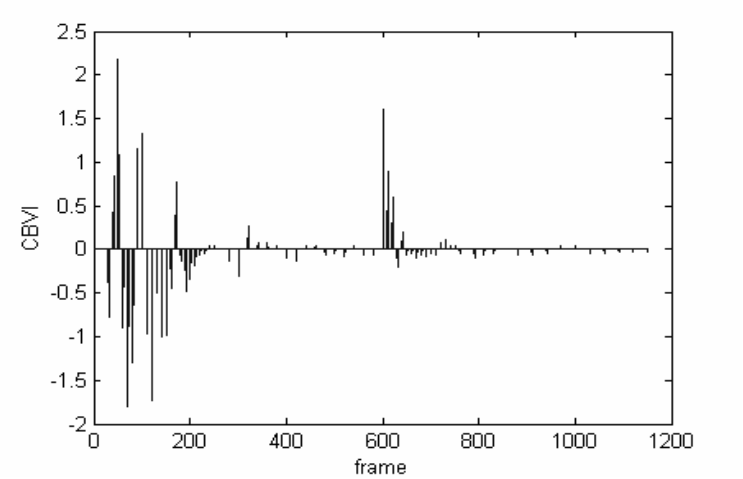
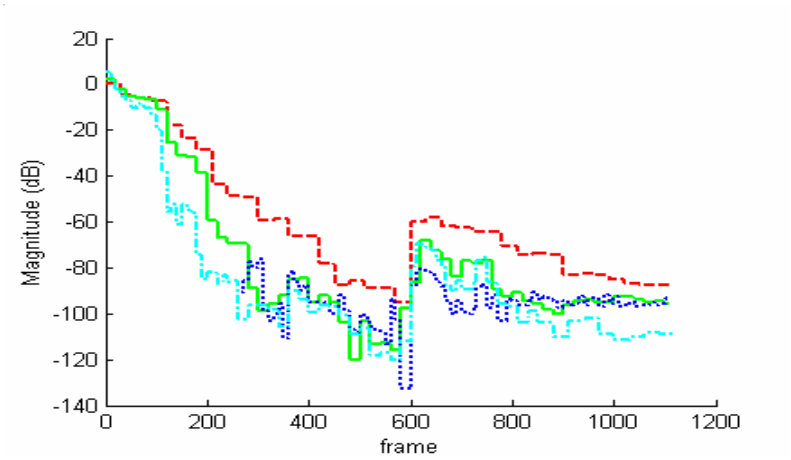
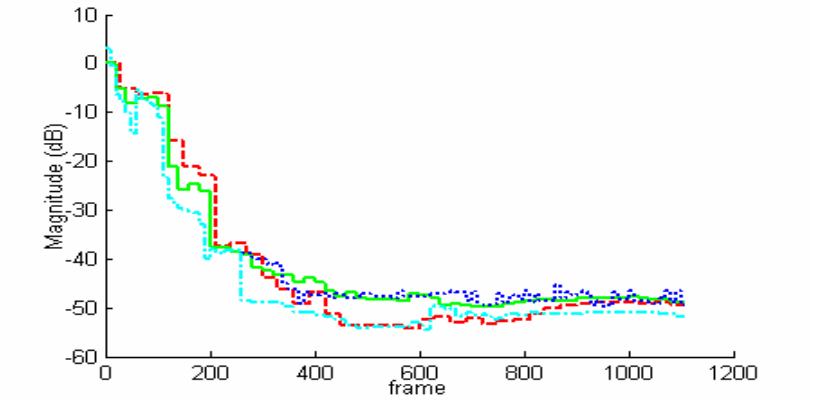


Figure 6-6 *CBVI* in the second simulation experiment



(a) NSR



(b) SDR

Figure 6-7 NSR and SDR in the second simulation experiment. The dash-dot line represents C10 (*ADL*), the dot line represents C7 ($L = 10$), the straight line represents C8 ($L = 20$), and the dash line represents C9 ($L = 30$)

Table 6-5 shows the number of multiplications ratios of FDABB and SPFDBB to the reference-signal-based time-domain adaptive beamformer. Significant saving of computing power can be achieved as these data indicated.

Table 6-5 Real Multiplication Requirement Ratio

	Multiplication Requirement Ratio	
	Adaptation Phase	Lower Beamformer Phase
FDABB with $\mu = 2$ in the first simulation case	1 : 8.57	1 : 20.72
FDABB with $\mu = 2$ in the second simulation case	1 : 8.48	1 : 20.72
SPFDBB with $L = 10$	1 : 10.69	1 : 20.72
SPFDBB with $L = 20$	1 : 11.04	1 : 20.72
SPFDBB with $L = 30$	1 : 11.17	1 : 20.72

6.1.2 Indoor Environment

A uniform, linear array using 6 microphones is constructed for this experiment with microphones spaced 0.07 *m* apart. The array is mounted on an easel which is one meter in height and two meters to the nearest wall. The environment is a 20 *m* x 15 *m* x 4 *m* room full of office furniture to simulate a practical environment and its reverberation time at 1000 *Hz* is around 0.52 second. The interference signals in this experiment are mutually uncorrelated white noise. The speech signal comes from 0° or 30° with a distance of 1.5 *m* and the configuration is shown as Fig. 6-8.

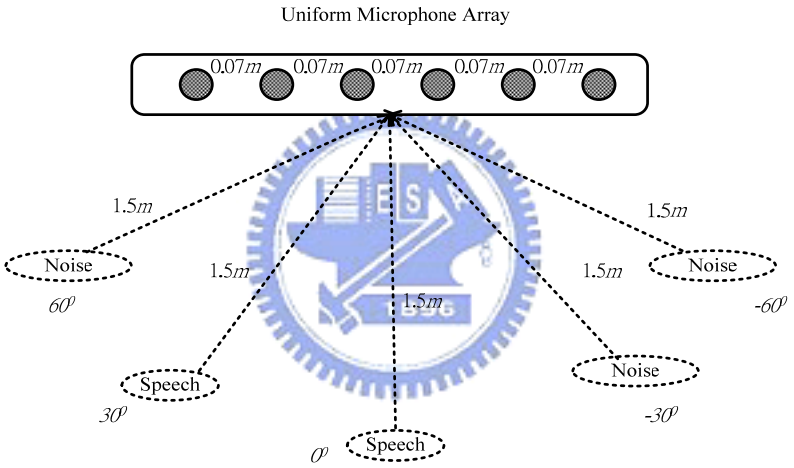


Figure 6-8 Arrangement of microphone array, noises and speech source in a noisy environment

This experiment utilizes the ASR rates to measure the performances of FDABB, SPFDBB, the reference-signal-based time-domain adaptive beamformer, DS beamformer, GSC, robust adaptive beamformer, and minimum variance beamformer (MV) under a fixed speech source and different number of interference signals. To measure the ASR rate, 500 pairs of the vehicle identification numbers pronounced in Chinese are used. An HTK software package [120] is adopted as a speech recognizer. Here, the FDABB parameters are the same as those in Table 6-4 except that the soft

penalty is set at a constant value 2. The value of γ is 10^{-6}. Table 6-6 shows the ASR system parameters. The filter tap of the reference-signal-based time-domain adaptive beamformer is chosen as 2560.

Table 6-6 Parameters of the ASR

Recognition kernel	HTK ver.3.0
Model	HMM
Feature Vector	12 th order MFCC + 12th order ΔMFCC
Training data Set	1001 clean pairs of the vehicle identification numbers
Recognition Task	500 pairs of the vehicle identification numbers

Figures 6-9 and 6-10 present the ASR rates of two different speech sources and the notations used in the two figures are shown in Table 6-7. Figure 6-9 shows that the ASR rate decreases as the number of interference source increases (see Table 6-7). Because DS beamformer, GSC, robust adaptive beamformer, and MV beamformer do not take the calibration problem into consideration, the improved ASR rates of speech source in 30° are lower than in 0°. For example, these traditional beamformers perform better in C1-C3 than in C4-C6. On the other hand, the proposed methods, SPFDBB with $L = 20$ and $L = 30$, and FDABB with $u = 2$, and the reference-signal-based time-domain adaptive beamformer shown in Fig. 6-10 can overcome this effect. For example, the improved ASR rate of SPFDBB with $L = 20$ in C4, 32.23%, is better than that in C1, 29.57%. These experimental results in Figs. 6-9 and 6-10 show that the value of L could affect the recognition rate. In this experiment, the FDABB with the $\mu = 2$ has the best performance in all conditions and the SPFDBB with $L = 30$ performs better than the reference-signal-based time-domain adaptive beamformer.

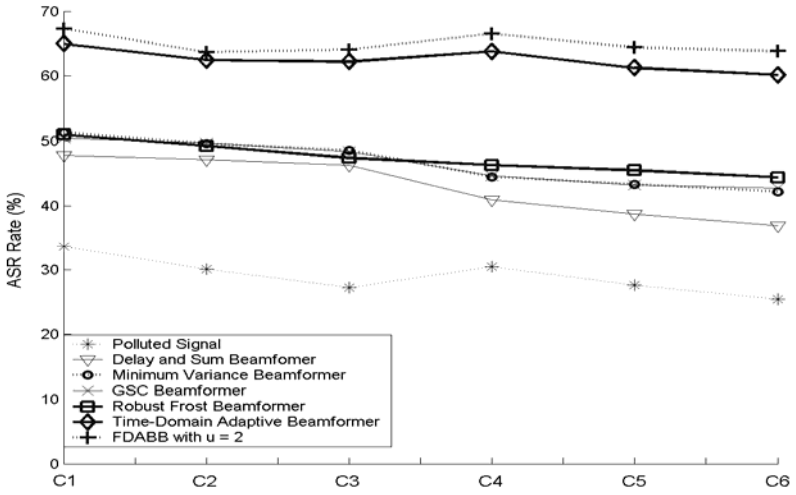


Figure 6-9 ASR rates of different kinds of beamformer outputs versus different experiment conditions

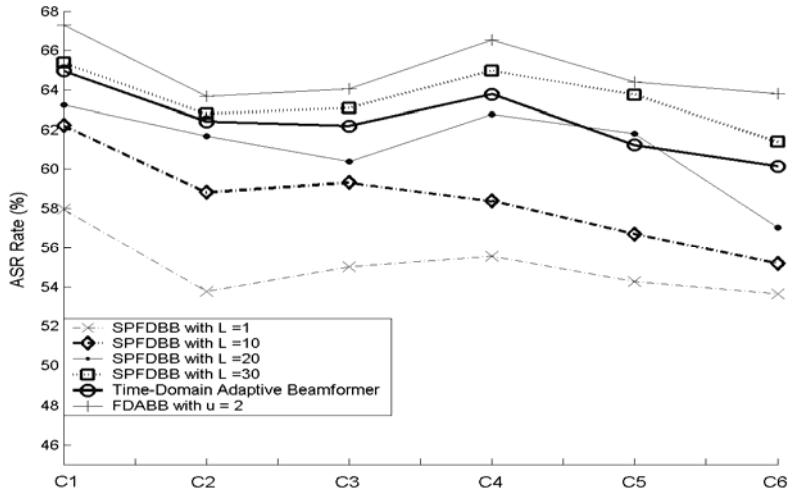


Figure 6-10 ASR rates of SPFDBB and FDABB versus different experiment conditions

Table 6-7 Meaning of Notations in Figs 6-9 and 6-10

C1	Speech source in 0˚and noise source in -60˚
C2	Speech source in 0˚and noise source in 60˚and -30˚
C3	Speech source in 0˚and noise source in 60˚, -30˚, and -60˚
C4	Speech source in 30˚and noise source in 60˚
C5	Speech source in 30˚and noise source in 60˚and -30˚
C6	Speech source in 30˚and noise source in 60˚, -30˚ , and -60˚

6.1.3 Vehicular Environment

The experiment is performed on passenger seat of a mini-van vehicle instead of the driver's seat due to the driving safety consideration. A uniform linear microphone array of six un-calibrated microphones with 0.05 m spacing is mounted in front of the passenger seat. Additionally, the distance between the microphone array and the speaker in the passenger seat is about 0.62 *m*. During the experiment, all windows are closed to prevent the microphones from saturating and the cabinet temperature is set to be 24ºC using the in-car air conditioner. Off-the-shelf, low-cost and non-calibrated microphones are used for the array. The performance of the proposed approaches is evaluated by ASR rates with the parameters shown in Table 6-6 under ten conditions (C1-C10 of Table 6-8). Table 6-8 shows the average SNRs in the ten conditions. A music piece containing vocal sound is played repeatedly from six build-in loudspeakers when the in-car audio system is turned on. This experiment utilizes the ASR rates shown in Fig. 6-11 to measure the performances of FDABB, SPFDBB, and the reference-signal-based time-domain adaptive beamformer. In the vehicular environment, the FDABB does not always outperform the SPFDBB with $L = 30$. However, the FDABB with the soft penalty parameter of 2 and the SPFDBB with $L = 30$ outperform the reference-signal-based time-domain adaptive beamformer.

Table 6- 8 Ten Experimental Conditions and Isolated Average SNRs

Condition Number	Speed	Power of In-car Audio System	Average SNR (dB)	Condition Number	Speed	Power of In-car Audio System	Average SNR (dB)
C1	20 km/h	Off	4.20	C6	20 km/h	On	-0.08
C2	40 km/h	Off	2.84	C7	40 km/h	On	-2.19
C3	60 km/h	Off	2.72	C8	60 km/h	On	-2.28
C4	80 km/h	Off	-1.90	C9	80 km/h	On	-4.75
C5	100 km/h	Off	-3.04	C10	100 km/h	On	-5.40

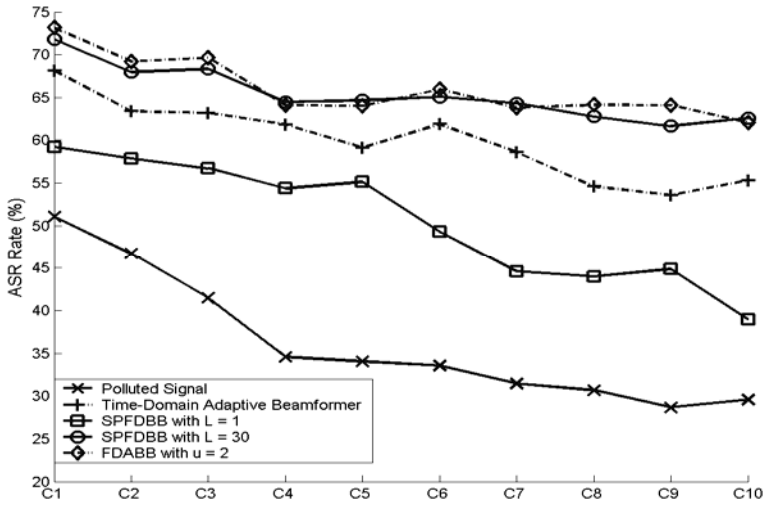


Figure 6-11 ASR rates of reference-signal-based time-domain beamformer, SPFDBB and FDABB

6.2 Comparison of NLMS and H_∞ Adaptation Criterions

6.2.1 Simulation Results

6.2.1.1 Single-channel Case

The four time-domain performance indexes: filtered output error $e_f(n)$ in Eq. (4-31), reference signal estimation error $e_r(n)$ in Eq. (4-32), filter coefficient estimation error ratio in Eq. (4-33), and filtered output error ratio in Eq. (4-34) are utilized in this case. The simulation signal is constructed through the linear model shown in Eq. (4-2). To reflect the modeling error, the unknown disturbance $e(n)$ is constructed as:

$$e(n) = \rho\left[\left(\hat{\boldsymbol{x}}^T(n)\boldsymbol{q}\right)^3 + v(n)\right] \tag{6-1}$$

where $\{v(n)\}$ is a white noise sequence and ρ is a scalar that can produce various SNR simulation cases. Therefore, the linear model can be rewritten as,

$$r(n) = \hat{\boldsymbol{x}}^T(n)\boldsymbol{q} + \rho\left[\left(\hat{\boldsymbol{x}}^T(n)\boldsymbol{q}\right)^3 + v(n)\right] \tag{6-2}$$

The tap number of the single channel, P, is chosen to be 10 and 20, μ_0 is set to 1, δ is set to 0.001, and there are a total of seven SNRs, denoted from C1 to C7, as shown in Table 6-9. Figure 6-12 illustrates the filtered output error ratio obtained by executing NLMS adaptation criterion at C7. The filtered output error ratio of the H_∞ adaptation criterion at C7 is presented in Fig 6-13. Obviously, the filtered output error ratio derived via the NLMS adaptation criterion never exceed one which fits the fact that the NLMS adaptation criterion guarantees the energy of the filtered output error will never exceed the energy of disturbance. Furthermore, the filtered output error, coefficient vector estimation error, and reference signal estimation error versus seven conditions shown in Figs. 6-14, 6-15, and 6-16 are derived from averaging the last one thousand runs of the total adaptation runs of 30000. Although the NLMS adaptation criterion is robust to the disturbance, the filtered output error and coefficient vector estimation error of H_∞ adaptation criterion still outperform those of the NLMS adaptation criterion in this simulation. Notably, because the basic concept of the NLMS and H_∞ adaptation criterions are to minimize the energy of the reference signal estimation error, $e_r(n)$, the reference signal estimation errors of the two adaptation criterion are similar.

Table 6-9 Seven Kinds of SNRs

Condition	C1	C2	C3	C4	C5	C6	C7
Average SNR	5.53 dB	1.99 dB	-0.57 dB	-2.47 dB	-4.09 dB	-5.42 dB	-6.56 dB

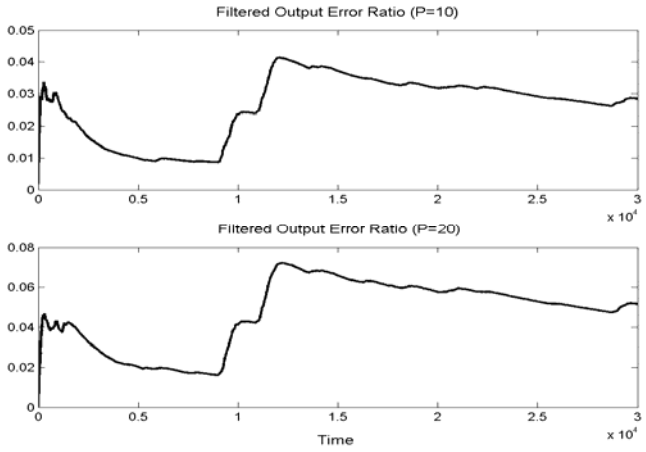


Figure 6-12 Filtered output error ratios of NLMS adaptation criterion with the tap number of 10 and 20

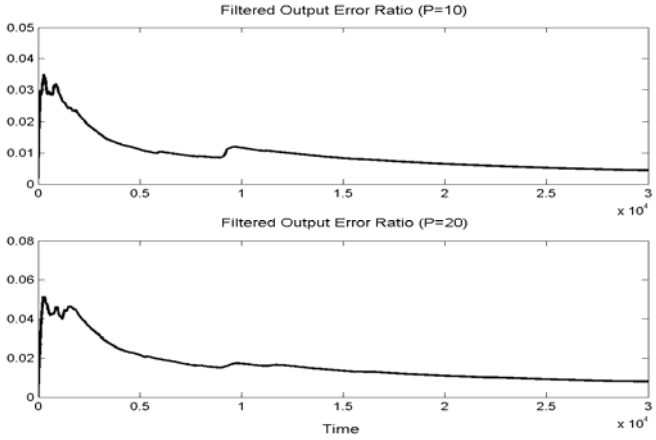


Figure 6-13 Filtered output error ratios of H_∞ adaptation criterion with the tap number of 10 and 20

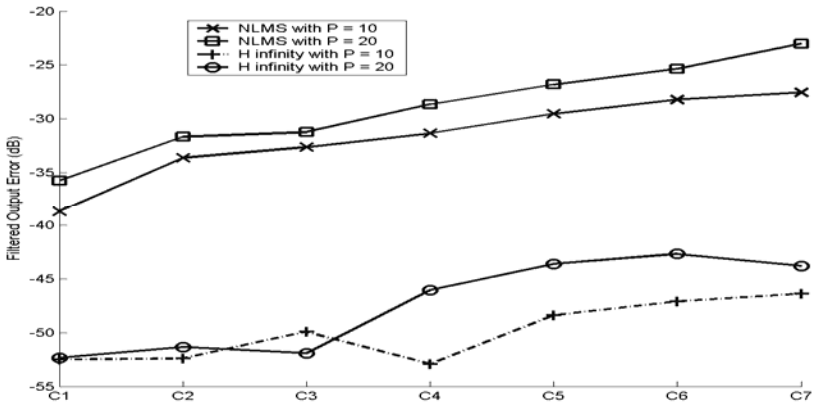


Figure 6-14 Filtered output error versus seven conditions

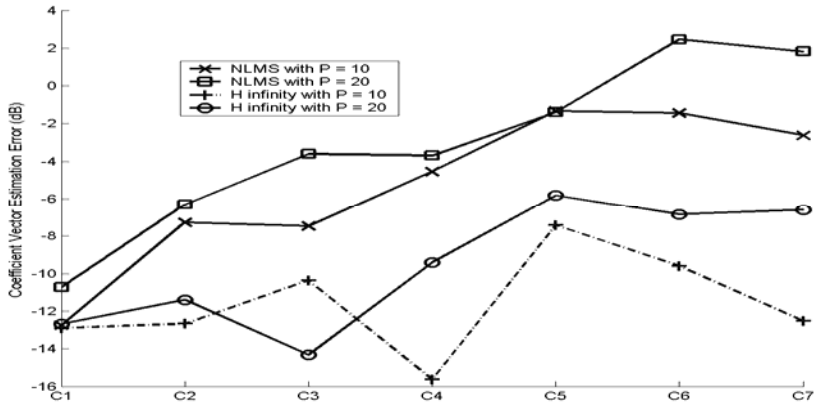


Figure 6-15 Coefficient vector estimation error versus seven conditions

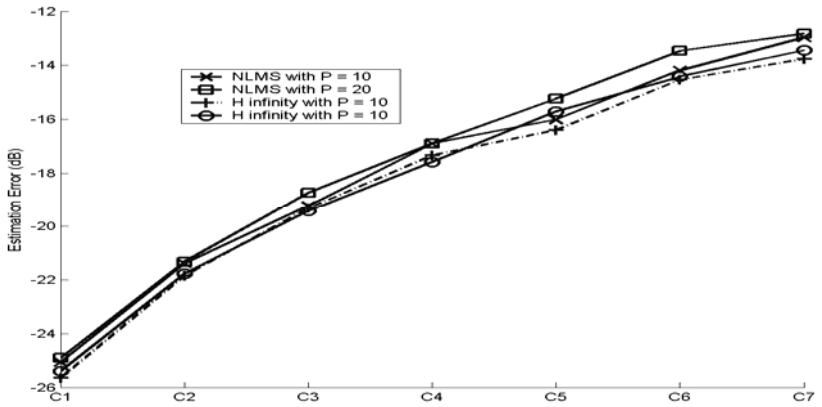


Figure 6-16 Reference signal estimation error versus seven conditions

Although the H_∞ adaptation criterion can guarantee Eq. (4-6) holds when the Riccati recursion $\mathbf{P}^{-1}(n)+\hat{\boldsymbol{x}}(n)\hat{\boldsymbol{x}}^T(n)-\gamma_q^{-2}\mathbf{I}$ is larger than 0, how to find a most suitable value of γ_q is still an issue. According to Eq. (4-6), a smaller value of γ_q^2 can guarantee a smaller upper bound of Eq. (4-6) under the same disturbances. It means that a smaller value of γ_q^2 limits the maximum filter coefficient estimation error ratio to a smaller value. However, a smaller value of γ_q^2 does not always achieve a better performance. In this simulation, three different values of γ_q are compared at C2 with the SNR of 1.99dB to prove that the selection of γ_q could affect the estimation

performance. P is set to 10. γ_q^{-2} in the three cases are selected as $\text{eig}(\mathbf{P}^{-1}(n)+\hat{\boldsymbol{x}}(n)\hat{\boldsymbol{x}}^T(n))\times 10^{-3}$, $\text{eig}(\mathbf{P}^{-1}(n)+\hat{\boldsymbol{x}}(n)\hat{\boldsymbol{x}}^T(n))\times 10^{-4}$, and $\text{eig}(\mathbf{P}^{-1}(n)+\hat{\boldsymbol{x}}(n)\hat{\boldsymbol{x}}^T(n))\times 10^{-5}$ individually. Notably, the three values of γ_q satisfy $\mathrm{P}^{-1}(n)+\hat{\boldsymbol{x}}(n)\hat{\boldsymbol{x}}^T(n)-\gamma_q^{-2}\mathbf{I}>0$, so the upper bound of Eq. (4-6) exists. Table 6-10 lists the corresponding minimum values of γ_q^2, filter output errors, and coefficient estimation errors in three cases from averaging the last one thousand runs. Figure 6-17 illustrates the filter coefficient estimation error ratios in three cases. Clearly, the filter coefficient estimation error ratios in three cases do not exceed the minimum value of γ_q^2. Although, case one has the smallest filter coefficient estimation error ratio and the fastest convergence rate, it does not converge to a smaller coefficient estimation error as compared with case two.

Table 6-10 Experimental Results in Three Different Selection Cases of γ_q^2

	Case One	Case Two	Case Three
Minimum value of γ_q^2	629.08	2837.90	14220
Filtered output error (dB)	-53.48 dB	-55.31 dB	-51.78 dB
Coefficient Estimation error (dB)	-12.27 dB	-16.98 dB	-11.37 dB

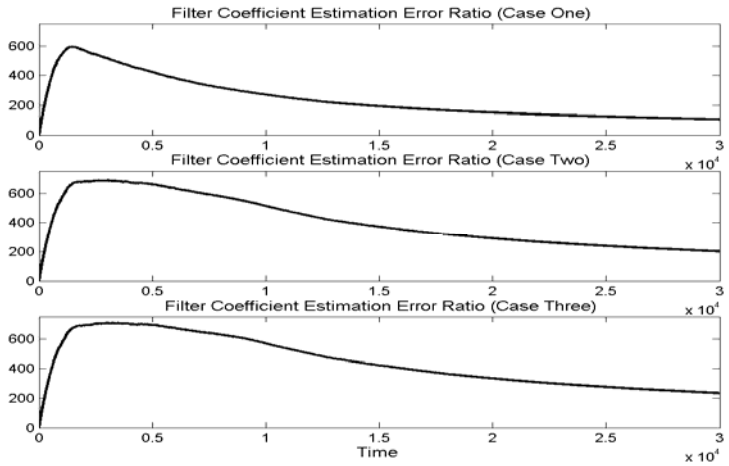


Figure 6-17 Filter coefficient estimation error ratios in three cases

6.2.1.2 *Multiple-channel Case*

To simulate a noisy environment, a speech source and a white noise are passed through each individual channel to the six microphones and the channel response duration is set to 30 taps. The tap number of the estimated coefficient vector has three selections of 10, 20, and 30 taps which can simulate the conditions under which the filter order is lower than or equal to the channel response duration. The time-domain SDR and NSR defined in Eqs. (4-29) and (4-30) are adopted as performance measurements in this section. Table 6-11 depicts the values of the SDR and the NSR versus the number of taps. Notably, the total adaptation runs is 30000 and the values of time-domain SDR and NSR are derived from averaging the last one thousand runs. Clearly, the H_∞ adaptation criterion outperforms the NLMS adaptation criterion. In Table 6-11, the performance of SDR and NSR degrades with the decrease of the filter order, especially for the H_∞ adaptation criterion. Although increasing the filter order enlarges the degree of freedom, the improvement provided by NLMS adaptation criterion is insignificant, since the approach continuous to suffer from the modeling error problem. On the contrary, the H_∞ adaptation criterion provides the robustness to modeling error, and thus can utilize the same increment of the degree of freedom to provide higher performance improvements.

Table 6-11 SDR and NSR at the SNR of -5.16 dB

	P = 10		P = 20		P = 30	
	SDR	NSR	SDR	NSR	SDR	NSR
NLMS adaptation criterion	-52.91 dB	-58.79 dB	-54.50 dB	-60.21 dB	-55.49 dB	-61.21 dB
H∞ adaptation criterion	-67.50 dB	-85.62 dB	-75.90 dB	-116.51 dB	-79.09 dB	-119.73 dB

6.2.2 *Indoor environment*

The indoor environment is arranges as Fig. 6-8 and the parameters of ASR is shown in Table 6-6. The FDABB parameters are the same as those in Table 6-4 and the soft penalty is 2. Figure 6-18 presents the ASR rates of SPFDBB and FDABB using NLMS and H_∞ adaptation criterions in an indoor environment. Clearly, H_∞ adaptation criterion outperforms the NLMS adaptation criterion in this case. Notably, although FDABB can adjust the frame number, *L*, this is not to say that the performance of FDABB is always better than the one of SPFDBB. For example, under the H_∞ adaptation criterion, the ASR rates of SPFDBB are better than those of FDABB in C3 and C6.

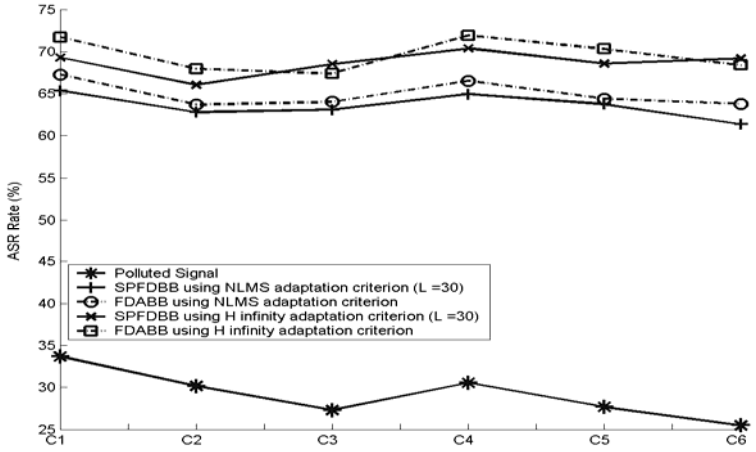


Figure 6-18 ASR rates of SPFDBB and FDABB using NLMS and H_∞ adaptation criterions in an indoor environment

6.2.3 Vehicular Environment

The experiment conditions are described in Section 6.1.3. Figure 6-19 shows the ASR rates of SPFDBB and FDABB using NLMS and H_∞ adaptation criterions in a vehicular environment. Obviously, the observation of that H_∞ adaptation criterion outperforms the NLMS remains true in a vehicular environment.

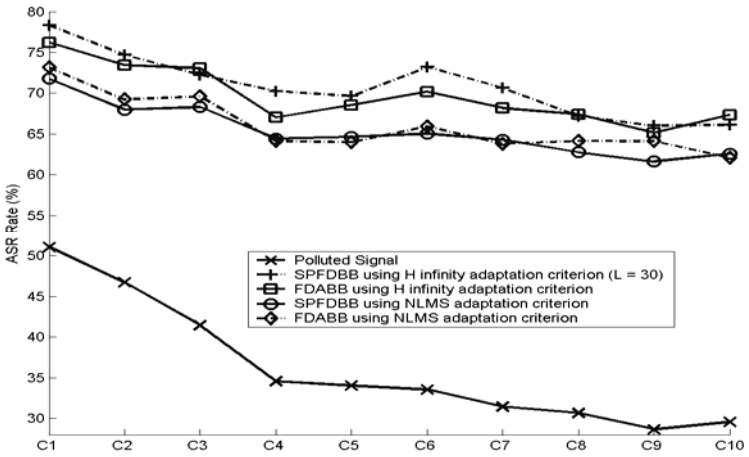


Figure 6-19 ASR rates of SPFDBB and FDABB using NLMS and H_∞ adaptation criterions in a vehicular environment

6.3 Reference-signal-based Speaker's Location Detection

6.3.1. Vehicular Environment

The experiment is performed in a mini-van vehicle with six separated seats [121]. Figure 6-20 presents the locations of the seats. During the experiment, the speakers at these locations slightly move around to mimic real usage scenarios. A uniform linear array of six microphones with 0.05 *m* spacing is mounted in front of location No. 2. The experiment is performed in various noisy environments. The environmental noise signals changed at various speeds. Table 6-12 lists the SNR ranges at various speeds, corresponding to the six locations. Table 6-13 presents the frequency bands that correspond to the pairs of microphones.

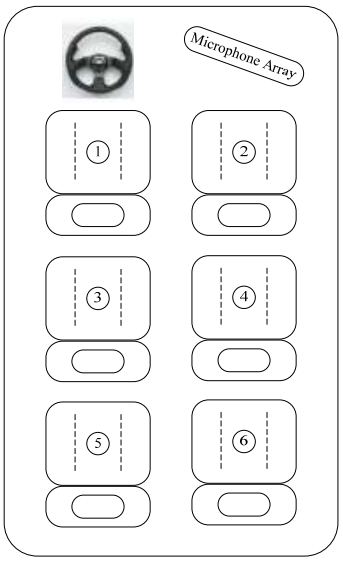


Figure 6-20 Seat number and microphone array position

Table 6-12 The SNR Ranges at Various Speeds

Speed	Speed = 0 *km/h*	Speed = 20 *km /h*
The SNR Range (dB)	10.8204 ~ 17.2664	4.1762 ~ 10.6222
Speed	Speed = 40 *km/h*	Speed = 60 *km/h*
The SNR Range (dB)	-4.5320 ~ 1.9140	-6.2526 ~ 0.1934
Speed	Speed = 80 *km/h*	Speed = 100 *km/h*
The SNR Range (dB)	-8.4709 ~ -2.0249	-13.0531 ~ -6.6071

Table 6-13 The Frequency Bands Correspond to the Microphone Pairs

Frequency Band	Microphone Pairs	The Number of Microphone Pair	The Range of Frequency Band
Band 1 ($b=1$)	(1,6)	$J_1=1$	$0 \le f \le 680Hz$
Band 2 ($b=2$)	(1,5); (2,6)	$J_2=2$	$680Hz < f \le 850Hz$
Band 3 ($b=3$)	(1,4); (2,5); (3,6)	$J_3=3$	$850Hz < f \le 1100Hz$
Band 4 ($b=4$)	(1,3); (2,4); (3,5); (4,6)	$J_4=4$	$1100Hz < f \le 1700Hz$
Band 5 ($b=5$)	(1,2); (2,3); (3,4); (4,5); (5,6)	$J_5=5$	$1700Hz < f \le 3400Hz$

6.3.1.1 MUSIC Algorithm - Single Speaker's Location Detection

A wideband incoherent MUSIC algorithm [10] with arithmetic mean is implemented and the results are compared with those of the proposed approach. Ten major frequencies, ranging from 0.1 *kHz* to 3.4 *kHz*, are adopted for the MUSIC algorithm.

Outliers are removed from the estimated angles by utilizing the method provided in [122]. Moreover, the angle errors needed for outlier rejection is derived from the estimated angles and real angles. Locations No. 2, 4 and 6 had the same DOA to the microphone array. Therefore, only locations No. 1, 2, 3 and 5 are considered for online testing. A frequently used classification method KNN (K-nearest-neighbor classification rule [123]) is used to construct a flexible boundary to improve the accuracy of detection to cope with the slight movement of the source, microphone mismatch, transient response and environmental noise. The estimated angles following outlier rejection are used as reference data in online location detection to illustrate further the performance of the MUSIC algorithm in the car cabinet. Suppose that the l th location contains β_l estimated angles and that $\sum_l \beta_l = \beta$ is the reference data set. Assume that the l th location contains K_l points in the $K-nearest$ results of a new estimate $\hat{r}$ derived from MUSIC with outlier rejection. The *a posteriori* probability is then given as

$$p(l \mid \hat{r}) = \frac{K_l}{K} \tag{6-3}$$

To minimize the probability of a false classification of $\hat{r}$, the estimated location, denoted as $\hat{l}_{MUSIC}$, is decided by using following equation:

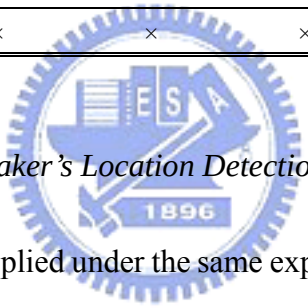
$$\hat{l}_{MUSIC} = \arg \max_{l=1,2,3,5} p(l \mid \hat{r}) \tag{6-4}$$

Notably, the new estimate will not be classified if it is an outlier. The parameters are set to $\beta_l = 200$, $l = \{1,2,3,5\}$, $\beta = 800$ and $K = 30$ and the number of trials is 100. Table 6-14 presents the correct rate after KNN classification with outlier rejection. The

correct rates at locations No. 3 and 5 are too low to be useful. In summary, these experimental results demonstrate that the MUSIC algorithm is not sufficiently reliable in a vehicle environment, even a classification method is applied and outliers are rejected to cope with the uncertainties.

Table 6-14 Correct Rate of MUSIC Method Utilizing KNN with Outlier Rejection

LOCATION	The correct rates at various speeds (*km/h*)					
	Speed 0 *km/h*	Speed 20 *km/h*	Speed 40 *km/h*	Speed 60 *km/h*	Speed 80 *km/h*	Speed 100 *km/h*
1	94 %	85 %	74 %	79 %	84 %	91 %
2	93 %	90 %	92 %	89 %	81 %	89 %
3	60 %	44 %	63 %	70 %	36 %	52 %
4	×	×	×	×	×	×
5	59 %	46 %	17 %	26 %	78 %	22 %
6	×	×	×	×	×	×



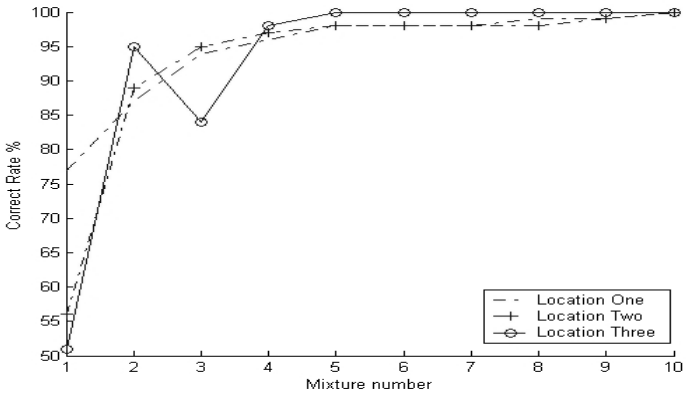
6.3.1.2 Proposed Single Speaker's Location Detection Approach

The proposed approach is applied under the same experimental conditions as Section 6.3.1.1 to detect the speaker's location. The second initial approach mentioned in section 5.3.2 is utilized to initialize the mean values. The covariance update may lead to numerical difficulties, as the covariance matrices become nearly singular. Consequently, the practical solution is to limit the minimum variance $\sigma^2_{\min}$. In this experiment, the value of $\sigma^2_{\min}$ is set to 0.02. The lengths of the training sequence T and the testing sequence Q are set to 200 and 50; in other words, a two-second length input datum is set for training, and a half-second length input datum is set for testing. The mixture number of GMM model has ten choices, from one to ten. Figure 6-21 plots the experimental result of the correct rate versus the mixture numbers at 100 *km/h*. As shown in Fig. 6-21 (a), a single Gaussian distribution (where the mixture number is one)

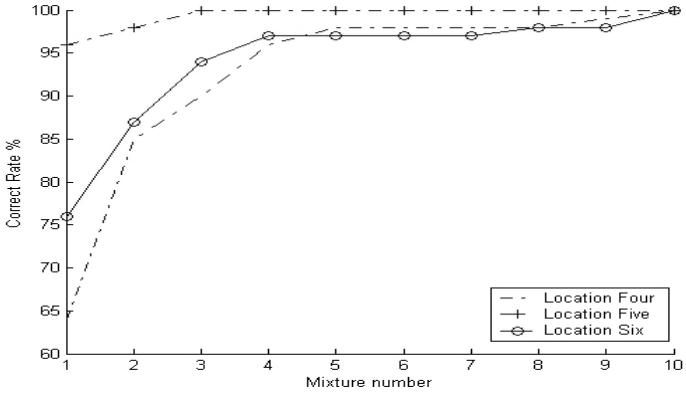
could not yield a satisfactory experimental performance. The correct rates are 100% at all locations with the mixture number is ten. This finding justifies the assumption that GMM is suitable for this application. Although the experimental performance improved as the mixture number increased, the improvement in performance is not significant when the mixture number exceeded five. Table 6-15 lists the experimental results with a mixture number of five. Clearly, the proposed method outperforms the MUSIC algorithm. Even at locations No. 4 and 6, the proposed method could distinguish them with significant accuracy. Figure 6-22 shows the histograms of phase differences at locations No. 2, 4, and 6 between the third and the sixth microphones at a frequency of 0.9375 *kHz* and between the fourth and the sixth microphones at a frequency of 1.5 *kHz*, e.g., in third and fourth frequency bands. The speed that corresponds to this figure is 100 *km/h*. Although the locations had the same angle to the microphone array, their phase difference distributions are quite different, as indicated by several research reports [113-114]. Additionally, the proposed method combined five frequency bands, each of which contained different phase difference distributions. As a result, the proposed method is able to distinguish all of the locations by exploiting their implicit diversities. Moreover, under low SNR conditions, the proposed approach still yielded a high correct rate and is robust against in-vehicle noise.

Table 6-15 Experimental Result of the Proposed Method with a Mixture Number of Five

Location	The correct rates at various speeds (*km/h*)					
	Speed 0 *km/h*	Speed 20 *km/h*	Speed 40 *km/h*	Speed 60 *km/h*	Speed 80 *km/h*	Speed 100 *km/h*
1	99 %	100 %	99 %	99 %	99 %	98 %
2	99 %	100 %	99 %	99 %	97 %	98 %
3	100 %	100 %	99 %	100 %	100 %	100 %
4	99 %	99 %	99 %	98 %	98 %	98 %
5	100 %	100 %	100 %	100 %	100 %	100 %
6	99 %	99 %	98 %	99 %	98 %	97 %

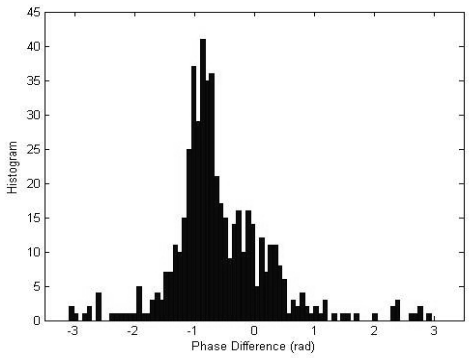


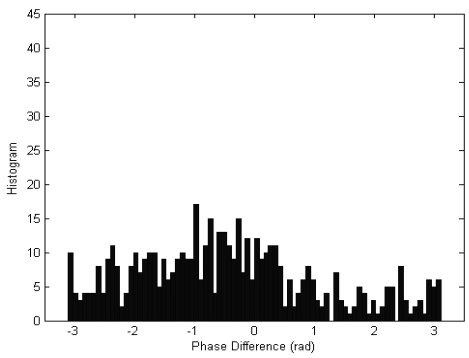
(a). The location number is chosen from 1 to 3



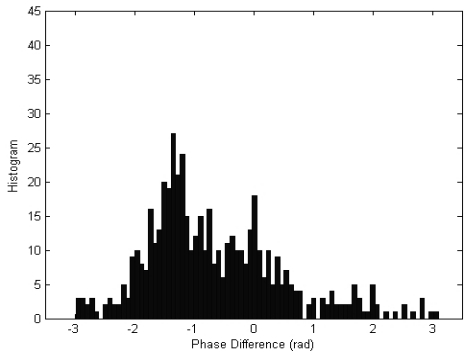
(b). The location number is chosen from 4 to 6

Figure 6-21 Correct rate versus the different mixture numbers in 100 *km/h*

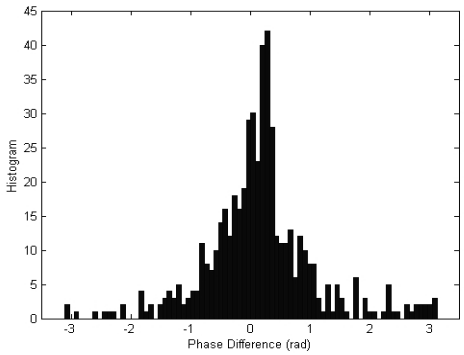




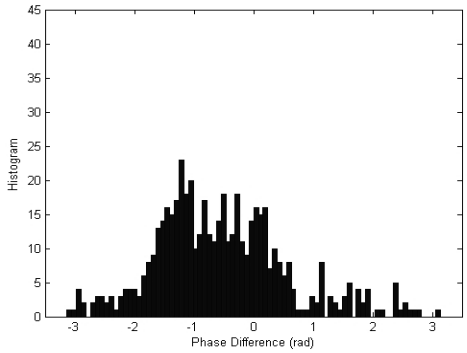
(a). Location No. 2 (frequency = 0.9375*kHz*) (b). Location No. 4 (frequency = 0.9375*kHz*)



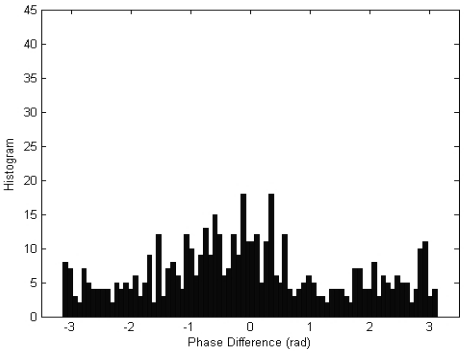
(c). Location No. 6 (frequency = 0.9375 *kHz*)



(d). Location No. 2 (frequency = 1.5 *kHz*)



(e). Location No. 4 (frequency = 1.5 *kHz*)



(f). Location No. 6 (frequency = 1.5 *kHz*)

Figure 6-22 The histograms of phase differences at locations No. 2, 4, and 6 between the third and the sixth microphones at a frequency of 0.9375 kHz and between the fourth and the sixth microphones at a frequency of 1.5 kHz, e.g., in third and fourth frequency bands (speed = 100 km/h)

6.3.1.3 Proposed Multiple Speakers' Locations Detection Approach

Figure 6-23 shows the locations of the six in-car loudspeakers, and the locations that are tested for the experiment. The first six locations correspond to modeled locations, and the radio broadcasting emits from the six in-car loudspeakers, locations no. 7, 8, and 9 correspond to unmodeled locations. The total length of the training phase difference sequence T is set to 300 (3-second duration). The values of Q_{Lo}, Q_{UP}, α, β, and γ are set to 10, 35, 0.3, 0.4, and 0.3 respectively.

Table 6-16 lists the SNR ranges at various speeds. The mixture number of GMM model has six choices, 1, 3, 5, 7, 9, and 11. The trial number for localization detection is 300 for each mixture number at each speed. For the condition of a single speaker, Fig. 6-24 plots the average correct rates versus mixture numbers, and indicates that a single Gaussian distribution, $M = 1$, could not yield a satisfactory performance, and that increasing the mixture number improves the performance.

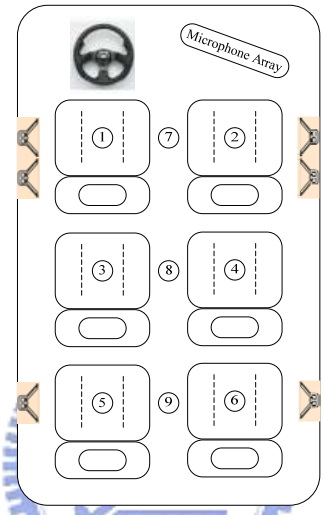
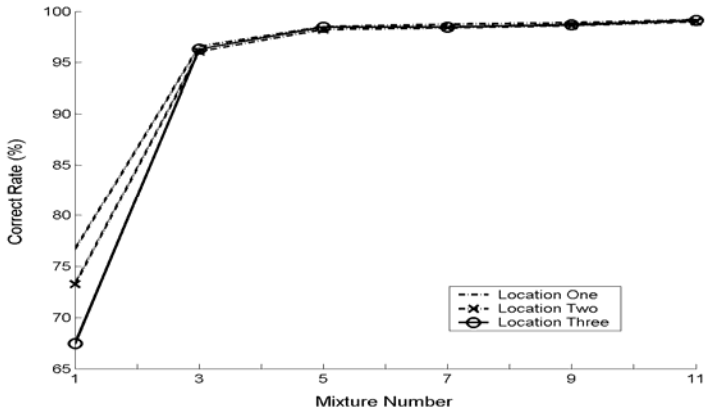


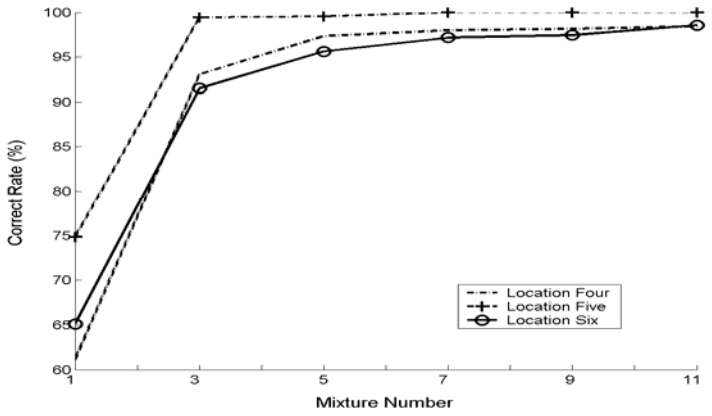
Figure 6-23 Locations number of the seats

Table 6-16 SNR Ranges at Various Speeds

Speed (*km/h*)	SNR Ranges (dB)				
	Multiple Speakers at locations no. 1 to 6	Radio broadcasting	Single Speaker at location no. 7	Single Speaker at location no. 8	Single Speaker at location no. 9
Speed = 0 *km/h*	10.81 – 18.15 dB	13.10 dB	14.96 dB	13.18 dB	17.31 dB
Speed = 20 *km/h*	5.62 – 12.96 dB	7.20 dB	10.15 dB	9.37 dB	11.50 dB
Speed = 40 *km/h*	0.19 – 7.54 dB	2.18 dB	4.53 dB	2.76 dB	6.89 dB
Speed = 60 *km/h*	-0.54 – 6.81 dB	1.75 dB	3.81 dB	2.03 dB	5.16 dB
Speed = 80 *km/h*	-5.32 – 2.02 dB	-3.04 dB	-0.98 dB	-2.76 dB	1.37 dB
Speed = 100 *km/h*	-7.28 – 0.07 dB	-5.99 dB	-2.93 dB	-4.71dB	-0.58 dB



(a). Locations number 1 to 3



(b). Locations number 4 to 6

Figure 6-24 Average correct rates versus the mixture numbers

Fifteen possible combinations, such as locations No. 1 and 2, and locations No. 1 and 3, exist with two speakers talking. Three, four, and five speakers talking yield 20, 15, and 6 possible combinations respectively. Table 6-17 lists the average error rates of these conditions with a mixture number of 11. Notably, an error is defined as a detection result that does not give the location of any of these speakers. For example, if the speech signals come from locations No. 2 and 3, then an error occurs when the detection result is neither 2 nor 3. Table 6-18 lists the average error rates of radio broadcasting and the speech signals coming from locations No. 7, 8, and 9 with a mixture number of 11. The error in the table is defined as the detection result pointing to

one of the modeled locations. The experimental results indicate that the proposed method can successfully deal with multiple speakers and unmodeled speech sources.

Table 6-17 Average Error Rates at Various Speeds under Multiple Speakers' Conditions

Speaker Number	Average Error Rates (%)					
	Speed 0 *km/h*	Speed 20 *km/h*	Speed 40 *km/h*	Speed 60 *km/h*	Speed 80 *km/h*	Speed 100 *km/h*
2	0.67 %	1.11 %	0.44 %	0.67 %	1.56 %	1.78 %
3	0.50 %	1.00 %	0.67 %	0.50 %	1.17 %	1.83 %
4	0.89 %	0.89 %	0.66 %	0.44 %	1.11 %	1.56 %
5	0.11 %	0.05 %	0 %	0 %	0.05 %	0.11 %

Table 6-18 Average Error Rates of Unmodeled Locations at Various Speeds

Speed (km/h)	Average Error Rates (%)			
	Radio broadcasting	Single Speaker at Location No. 7	Single Speaker at Location No. 8	Single Speaker at Location No. 9
Speed = 0 *km/h*	0.22 %	0 %	0.06 %	0.22 %
Speed = 20 *km/h*	0.28 %	0 %	0.17 %	0 %
Speed = 40 *km/h*	0 %	0 %	0 %	0 %
Speed = 60 *km/h*	0.06 %	0 %	0 %	0.33 %
Speed = 80 *km/h*	0.28 %	0.33 %	0.33 %	0.33 %
Speed = 100 *km/h*	0.33 %	0 %	0.39 %	0.67 %

6.3.2. Indoor Environment

The dimensions of the experimental room and the arrangement of microphone array are the same with those in the Section 6.1.2.

6.3.2.1 Proposed Single Speaker's Location Detection Approach

Figure 6-25 presents the real configuration. The four speech signals are located at different angles, 0°, 30°, and −60°, with various distances to the array. Noises are located at 60°, −30° and −60°. Notably, speeches No. 1 and 3, and speech No. 4 and

noise No. 1 have the same DOA to the microphone array. All of the speech signals and noises are played by loudspeakers during the experiment. The interference signals in this experiment are white Gaussian noises and mutually uncorrelated. The distances between the microphone array and the noises are all 1.5 m. There are a total of twelve experimental conditions, denoted from C1 to C12, as shown in Table 6-19.

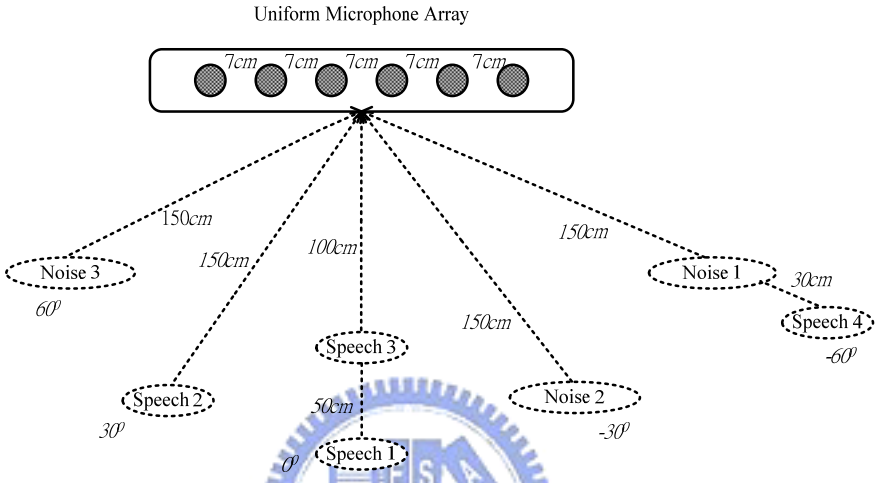
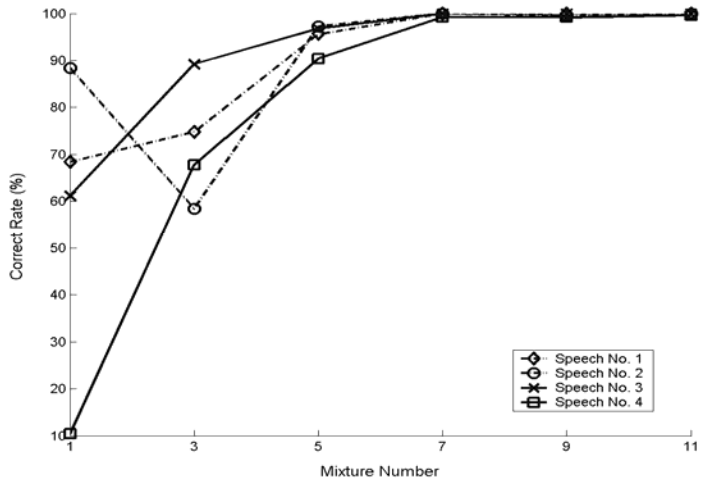


Figure 6-25 Configuration of microphone array, noises and speech sources in noisy environment

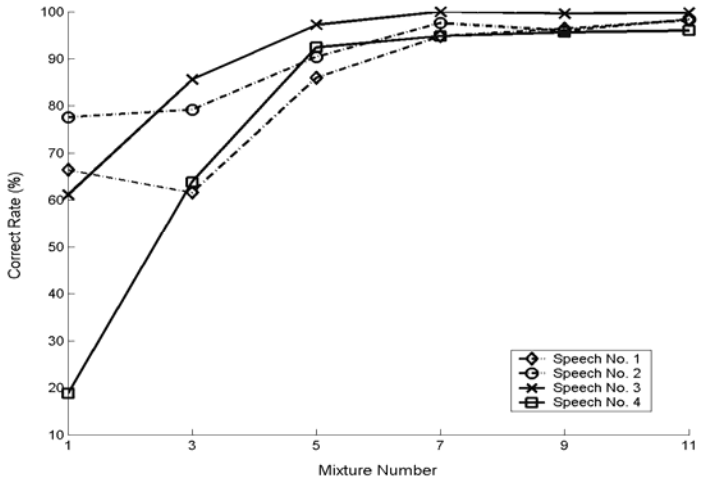
Table 6-19 Twelve Kinds of Experimental Conditions

	Condition	SNR (dB)
C1	Speech No. 1 and noise No. 1	15.46 dB
C2	Speech No. 2 and noise No. 1	16.45 dB
C3	Speech No. 3 and noise No. 1	13.93 dB
C4	Speech No. 4 and noise No. 1	15.67 dB
C5	Speech No. 1, noises No. 1 and 2	10.08 dB
C6	Speech No. 2, noises No. 1 and 2	11.07 dB
C7	Speech No. 3, noises No. 1 and 2	8.55 dB
C8	Speech No. 4, noises No. 1 and 2	10.28 dB
C9	Speech No. 1, noises No. 1, 2, and 3	5.95 dB
C10	Speech No. 2, noises No. 1, 2, and 3	6.93 dB
C11	Speech No. 3, noises No. 1, 2, and 3	4.42 dB
C12	Speech No. 4, noises No. 1, 2, and 3	6.15 dB

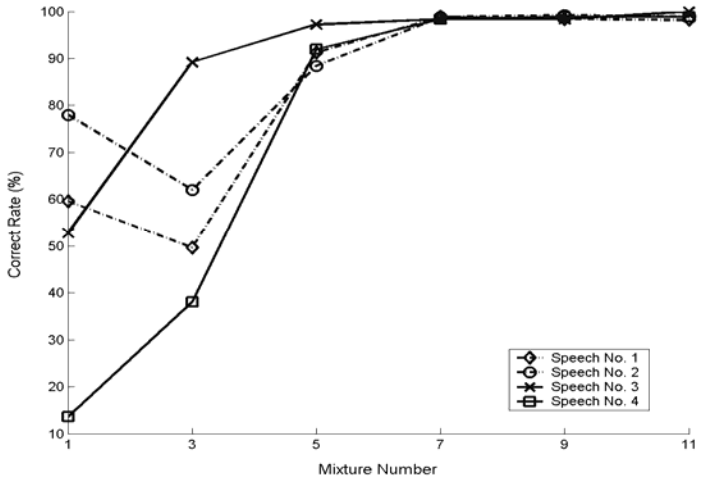
The lengths of the training sequence T and the testing sequence Q are set to 300 and 50; in other words, a three-second length input datum is set for training, and a half-second length input datum is set for testing. The mixture number of GMM model has six choices, 1, 3, 5, 7, 9, and 11. The trial number for localization detection is 250 for each mixture number at each condition. Figure 6-26 plots the experimental result of the correct rates versus the mixture numbers under various conditions. Although the speech No. 4 and noise No. 1 come from the same direction, speech No. 4 still can be distinguished in C4, C8, and C12 with a higher mixture number. Generally, the correct rates in the indoor environment are lower than those in the vehicular environment when the mixture number is low, such as 1, 3, and 5. This phenomenon means that the phase difference distributions of different locations in the vehicular environment are more distinguishable.



(a). Conditions one to four



(b) Conditions five to eight



(c) Conditions nine to twelve

Figure 6-26 Correct rates versus the different mixture numbers

6.3.2.2 Proposed Multiple Speakers' Locations Detection Approach

Figure 6-27 presents the real configuration. Clearly, three unmodeled speech signals, speeches No. 5, 6, and 7, are added in the experiment. It means that speeches No, 5, 6, and 7 can be regard as undesired speech signals or interference signals. The total length of the training phase difference sequence T is set to 300 (3-second duration). The values of Q_{Lo}, Q_{UP}, α, β, and γ are set to 10, 35, 0.3, 0.4, and 0.3 respectively. Table 6-20 lists the SNR ranges at three different noisy environments. The mixture

number of GMM model also has six choices, 1, 3, 5, 7, 9, and 11. The trial number for localization detection is 250 for each mixture number. For the condition of a single speaker, Fig. 6-28 plots the average correct rates versus mixture numbers.

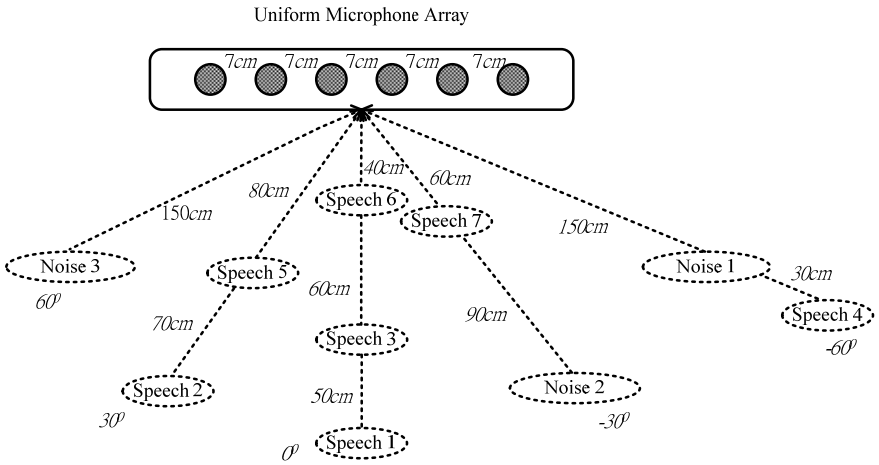


Figure 6-27 Configuration of microphone array, noises and speech sources in noisy environment

Table 6-20 SNR Ranges at Three Different Noisy Environments

Noisy Environments	SNR Ranges (dB)			
	Multiple Speakers at Speeches No. 1 to 4	Single Speech Signal at Speech No. 5	Single Speech Signal at Speech No. 6	Single Speech Signal at Speech No. 7
Noise No. 1	13.93 – 26.9 dB	14.97 dB	18.17 dB	16.00 dB
Noises No. 1 and 2	8.55 – 21.54 dB	9.57 dB	12.78 dB	10.62 dB
Noises No. 1, 2, and 3	4.42 – 15.66 dB	3.68 dB	6.88 dB	4.71 dB

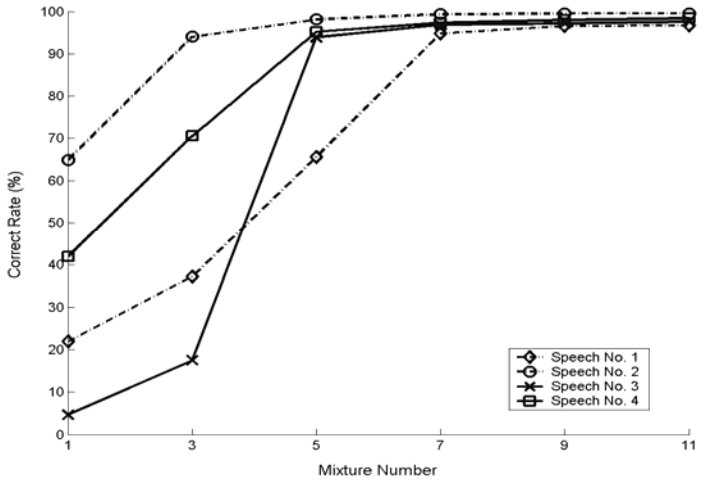


Figure 6-28 Average correct rates versus the mixture numbers

Six possible combinations, such as speeches No. 1 and 2, and locations No. 1 and 3, exist with two speakers talking. Three speakers talking yield four possible combinations respectively. Table 6-21 lists the average error rates of these conditions and Table 6-22 lists the average error rates of the speech signals coming from speeches No. 5, 6, and 7 with a mixture number of 11. Notably, speeches No. 1, 3 and 6 and speeches No. 2 and 5 have the same DOA to the microphone array. The experimental results indicate that the proposed method can also successfully deal with multiple speakers and unmodeled speech signals.

Table 6-21 Average Error Rates at Three Noisy Environments under Multiple Speakers' Conditions

Speaker Number	Average Error Rates (%)		
	Noise No. 1	Noises No. 1 and 2	Noises No. 1, 2, and 3
2	2.93 %	1.33 %	0.87 %
3	0.1 %	0.1 %	0 %

Table 6-22 Average Error Rates of Unmodeled Locations at Three Noisy Environments

Noisy Environments	Average Error Rates (%)		
	Single Speech Signal at Speech No. 5	Single Speech Signal at Speech No. 6	Single Speech Signal at Speech No. 7
Noise No. 1	0.8 %	0 %	0 %
Noises No. 1 and 2	0 %	0 %	0.2 %
Noises No. 1, 2, and 3	0.3 %	0.4 %	0.2 %

6.4 Summary

This chapter evaluates the proposed SPFDBB, FDABB, and speaker's location detection approaches through simulation and real experimental results. Section 6.1 proves the proposed SPFDBB and FDABB not only outperform the reference-signal-based time-domain adaptive beamformer and several famous

beamformers, but also reduce the computational effort. Section 6.2 simulates the single-channel and multiple-channel cases, while performing vehicular environment and indoor environment experiments, to show the robustness of the H_∞ adaptation criterion. Moreover, Section 6.3 executes the proposed speaker's location detection approach in noisy vehicular and indoor environments to prove the high detection accuracy and the robustness to the unmodeled or unexpected speech sources.

Chapter 7

Conclusions and Future researches

7.1. Conclusions

This dissertation presents reference-signal-based methods of sound source localization and speech purification using microphone array. Specifically, frequency-domain adaptive beamformers, namely SPFDBB and FDABB, are proposed to cope with the computation issues in real-time. Under the architecture, the proposed approaches can be applied to both near-field and far-field environments and overcome microphone mismatch problem.

Other than the advantages mentioned above, SPFDBB and FDABB can minimize the channel effects, the desired signal cancellation, and the resolution effect due to the array's position. FDABB and SPFDBB not only reduce the computational effort, but also deal with the problem of inaccurate channel representation. That is to say, the convolution relation between channel and speech source in time-domain cannot be modeled accurately as a multiplication in the frequency domain with a finite window size. According to the computational effort analysis in Chapters 4, 5, and 6, FDABB or SPFDBB requires a lower computational effort as compared with the

reference-signal-based time-domain adaptive beamformer. Additionally, FDABB utilizes an index named *CBVI* to adjust the frame number L automatically, so it is more suitable than SPFDBB for applications with a small training data length or variations of the channel dynamics.

FDABB and SPFDBB attempt to simultaneously suppress the noise signals and recover the channel dynamics. However, according to Eq. (5-1), the finite filter coefficient vectors are not sufficient to perform the perfect equalization in general environments, thus leading to the modeling error. To reduce the effect of modeling error, this dissertation further studies the robustness of H_∞ adaptation criterion and applies the criterion to the proposed FDABB and SPFDBB.

To overcome the non-line-of-sight problem in the sound source localization field, this dissertation proposes an approach utilizing GMM to model the distributions of the phase differences among the microphones caused by the complex characteristic of room acoustic and microphone mismatch. According to the experimental results in Chapter 6, the scheme performs well not only in non-line-of-sight cases, but also when the speakers are aligned toward the microphone array but at difference distances from it. However, an unmodeled speech signal which is not emitted from one of the modeled locations degrades the detection performance. Therefore, this dissertation further proposes multiple speakers' location detection approach to provide an accurate localization of multiple speakers and robustness to unmodeled sound source locations.

7.2. Future researches

To improve the current speech enhancement system, this dissertation proposes two possible further research directions; the first one is to combine SPFDBB or FDABB

with the speech recognizer, and the second one is to combine multiple speakers' location detection approach with the proposed SPFDBB or FDABB.

Currently, the proposed reference-signal-based frequency-domain beamformers are performed in two independent phases: speech purification and then recognition as shown in Figs. 1-5, 3-1 and 4-2. The proposed beamformers designed to reduce the speech distortion and suppress the noise effects assume that improving the quality of the speech waveform will result in better recognition performance and are independent of the recognition system. Although the proposed beamformers can conquer many practical issues, the beamformers still cannot compete with the microphone in a close distance in terms of the ASR rates. Generally, a speech recognizer is a statistical pattern classifier that operates on a sequence of features derived from the waveform. To increase the recognition accuracy in distant-talking environments, the architecture of connecting the proposed beamformers and the speech recognizer as shown in Fig. 7-1 is worth a further study in the further. This architecture enables the beamformer to use the data transmitted from the recognizer and ensures the beamformer enhances those signal components important for ASR. In other words, this architecture enables the designed filters not to undue emphasis on unimportant components.

For example, Seltzer *et al.* [124-125] proposed a likelihood-maximizing beamformer (LIMABEAM) that integrates the speech recognition system into the filter design process. They proved that incorporating the statistical models of the recognizer into the array processing stage can improve the ASR rates. The goal of the LIMABEAM is not to generate an enhanced output waveform but rather to generate a sequence of features which maximizes the likelihood of generating the correct hypothesis. In other words, the filter coefficient vectors are chosen to maximize the likelihood of the training signal as measured by the recognizer, rather than to improve its SNR or perceptual quality.

Nishiura *et al.* [128] also proposed a method which combines HMM model and microphone to enhance the speech recognition further.

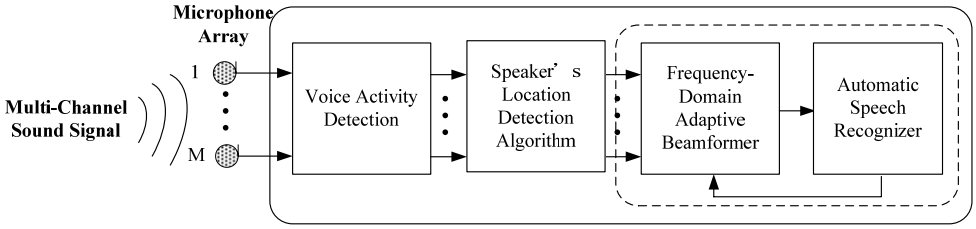


Figure 7-1 Speech enhancement system with a combination of beamformer and recognizer

Furthermore, under the same architecture, it is highly possible to combine the proposed speaker's location detection approach and SPFDBB or FDABB to significantly reduce the possibility of the wrong operation of SPFDBB or FDABB. The wrong operation means that the unmodeled or unexpected speech signal triggers the VAD to switch the beamformers to the speech stage, and thus obtain an error output. According to the system architecture of frequency-domain beamformers in Chapters 3 and 4, this wrong operation is unavoidable due to the limitation of VAD. However, the proposed multiple speakers' locations detection approach in Chapter 5, which can detect the unmodeled sound signals, is highly possible to avoid the wrong operation. Therefore, the unmodeled speech signal could hardly cause the error output under the new system architecture in Fig. 7-2. Based on the system architecture, both the VAD result and speaker's location detection system are combined to decide the stage of the SPFDBB or FDABB. It means that if the received sound signal is detected as containing speech signal and coming from one of the desired locations, then the system is switched to the speech stage. Consequently, it is also worthwhile to study how to construct the two components with a suitable integration procedure. Figure 7-3 shows the flowchart of the integrated architecture.

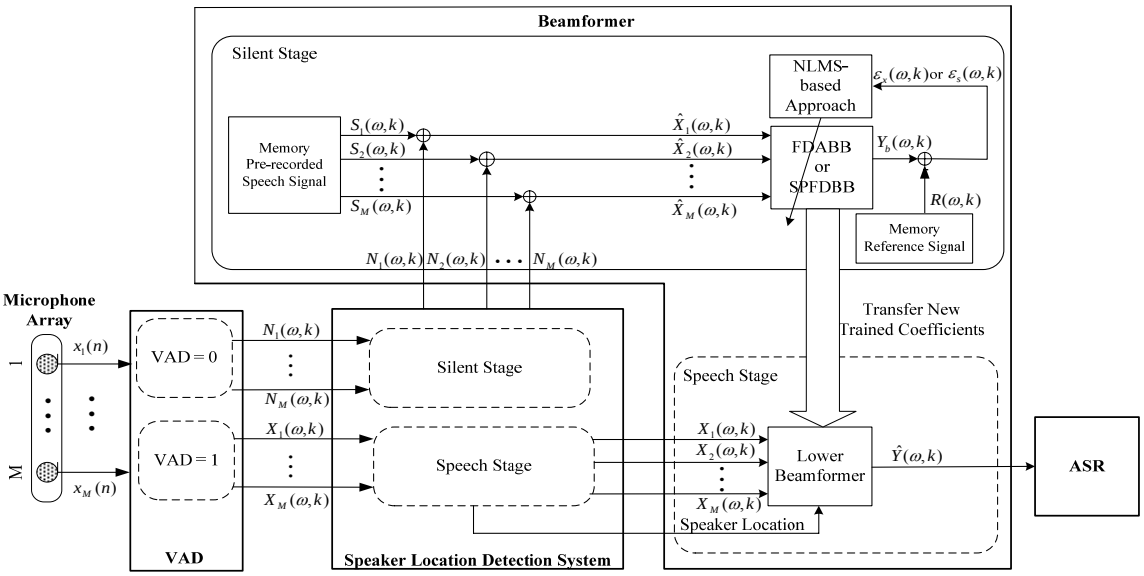


Figure 7-2 Overall system structure which integrates the speaker's location detection approach and SPFDBB or FDABB

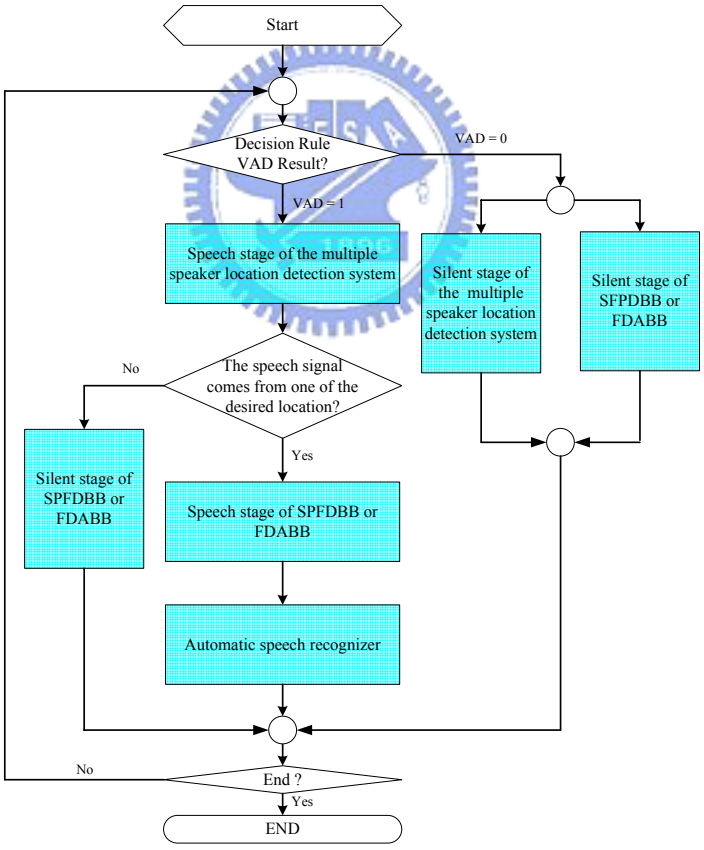


Figure 7-3 Flowchart of the architecture which integrates the speaker's location detection approach and SPFDBB or FDABB

Reference

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. ASSP-27, pp. 113-120, Apr. 1979.

[2] A. Kawamura, Y. Iiguni, and Y. Itoh, "A noise reduction method based on linear prediction with variable step-size," *IEICE Tran. Fundamentals*, vol. E88-A, no. 4, pp. 855-861, April 2005.

[3] J. G. Ryan, R. A. Goubran, "Near-field beamforming for microphone arrays," *IEEE Conference on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 21-24, Apr. 1997.

[4] P. L. Chu, "Superdirective microphone array for a set-top video conferencing system," *IEEE International Conference on Acoust., Speech, and Signal Processing*, vol. I, pp. 235-238, Apr. 1999.

[5] M. Wax and T. Kailath, "Optimal localization of multiple sources by passive arrays," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. ASSP-31, pp. 1210-1217, Oct. 1983.

[6] H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone-array data," *Computer, Speech, and Language*, vol. 6, pp. 129-152, Apr. 1992.

[7] D. B. Ward, and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," *International Conference on Acoustics, Speech, and Signal Processing,* vol. 2, pp. 1777-1780, May 2002.

[8] R. V. Balan and J. Rosca, "Apparatus and method for estimating the direction of Arrival of a source signal using a microphone array," *European Patent, No US2004013275,* 2004.

[9] M. Wax, T. J. Shan, and T. Kailath. " Spatio-temporal spectral analysis by eigenstructure methods," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol ASSP-32, pp. 817–827, Aug 1984.

[10] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation,* vol. 34, pp. 276-280. Mar 1986.

[11] H. Wang and M. Kaveh, "Coherent signal subspace processing for detection and estimation of angle of arrival of multiple wideband sources," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol ASSP-33, pp. 823–831, Aug. 1985.

[12] G. Bienvenu, “Eigensystem properties of the sampled space correlation matrix,” *IEEE International Conference on Acoust., Speech, and Signal Processing*, pp. 332–335, 1983.

[13] K. M. Buckley and L. J. Griffiths, “Eigenstructure based broadband source location estimation,” *IEEE International Conference on Acoust., Speech, and Signal Processing*, pp. 1869–1872, 1986.

[14] M. A. Doron, A. J. Weiss, and H. Messer, “Maximum likelihood direction finding of wideband sources,” *IEEE Trans. on Signal Processing,* vol. 41, pp. 411–414, Jan 1993.

[15] M. Agarwal and S. Prasad, “DOA estimation of wideband sources using a harmonic source model and uniform linear array,” *IEEE Trans. on Signal Processing*, vol. 47, pp. 619-629, Mar. 1999.

[16] H. Messer, “The potential performance gain in using spectral information in passive detection/localization of wideband sources,” *IEEE Trans. on Signal Processing*, vol. 43, pp. 2964-2974, Dec. 1995.

[17] M. Agrawal and S. Prasad, “Broadband DOA estimation using spatial-only modeling of array data,” *IEEE Trans. on Signal Processing*, vol. 48, pp. 663-670, Mar. 2000.

[18] J.-H Lee, Y.-M Chen, and C.-C Yeh, “A covariance approximation method for near-field direction finding using a uniform linear array,” *IEEE Trans. on Signal Processing*, vol. 43, pp. 1293-1298, May 1995.

[19] K. Buckley and L. Griffiths, “Broad-band signal-subspace spatial-spectrum (BASS-ALE) estimation,” *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. 36, pp. 953-964, July 1988.

[20] N. Strobel and R. Rabenstein, “Classification of time delay estimates for robust speaker localization,” *International Conference on Acoustics, Speech, and Signal Processing,* vol. 6, pp. 15-19, March 1999.

[21] G. C. Carter, A. H. Nuttall, and P. G. Cable, “The smoothed coherence transform(SCOT),” *Naval Underwater Systems Center, New London Lab., New London, CT, Tech. Memo* TC-159-72, Aug 8 1972.

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

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

[24] J. Hu, T.M. Su, C.C. Cheng, W.H. Liu, and T.I. Wu, “A self-calibrated speaker tracking system using both audio and video data,” *IEEE Conference on Control Applications*, vol.2, pp. 731-735, Sept 2002.

[25] J. Hu, C.C. Cheng, W.H. Liu, and T.M. Su, "A speaker tracking system with distance estimation using microphone array," *IEEE/ASME International Conference on Advanced Manufacturing Technologies and Education,* Aug 2002.

[26] S. Mavandadi, P. Aarabi, "Multichannel nonlinear phase analysis for time-frequency data fusion," *Proceedings of the SPIE, Architectures, Algorithms, and Applications VII (AeroSense 2003),* vol. 5099, pp. 222-231, April 2003.

[27] P. Aarabi and S. Mavandadi, "Robust sound localization using conditional time–frequency histograms," *Information Fusion*, vol. 4, pp. 111-122, June 2003.

[28] C. Winter, "Using continuous apertures discretely," *IEEE Trans. on Antennas and Propagation,* vol. AP-25, pp. 695-700, Sep. 1977.

[29] G. W. Elko, T. C. Chou, R. J. Lustberg, M. M. Goodwin, "A constant-directivity beamforming microphone array," *J. Acoust. Soc. Amer.,* vol. 96, no. 5, pp.3244, Nov. 1994.

[30] J. H. Doles III and F. D. Benedict, "Broad-band array design using the asymptotic theory of unequally spaced arrays," *IEEE Trans. on on Antennas and Propagation,* vol. 36, no. 1, pp. 27-33, Jan. 1988.

[31] A. Ishimaru and Y. S. Chen, "Thinning and broadbanding antenna arrays by unequal spacings," *IEEE Trans. on Antennas and Propagation,* vol. AP-13, pp. 34-42, Jan. 1965.

[32] T. Chou, "Frequency-independent beamformer with low response error," *IEEE Trans. on Speech and Audio Processing,* pp. 2995-2998, May 1995.

[33] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Acoustic, Speech, Signal Processing Magazine*, pp. 4-24, Apr. 1988.

[34] M. Doerbecker, "Speech enhancement using small microphone array with optimized directivity," *International Workshop on Acoustic, Echo, and Noise Control*, pp. 100-103, Sep. 1997.

[35] J. G. Ryan and R. A. Goubran, "Optimal near-field response for microphone arrays," *IEEE International Workshop on Application Signal Processing to Audio Acoustics*, pp. 19-22, Oct. 1997.

[36] J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "An alternative implementation of the superdierective beamformer,"in Proc. *IEEE International Workshop on Application Signal Processing to Audio Acoustics*, pp. 7-10, New Paltz, NY, USA, Oct. 1997.

[37] R. Zelinski, "Noise reduction based on microphone array with LMS adaptive post-filtering," *Electronic letter*, vol. 26, pp. 2036-2037, Nov. 1990.

[38] M. Dorbecker and S. Ernst, "Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation," *European Signal Processing Conference*, pp. 995-998, Sep. 1996.

[39] C. Marro, Y. Mahieux, and K. U. simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. on Acoustic, Speech, Signal Processing,* vol. 6, pp. 240-259, May 1998.

[40] I. a. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," *IEEE International Conference on Acoust., Speech, and Signal Processing*, pp. 1723-1726, Apr. 2000.

[41] S. Doclo and M. Moonen, "SVD-based optimal filtering with applications to noise reduction in speech signals," *IEEE International Workshop on Application Signal Processing to Audio Acoustics* pp. 143-146, Oct. 1999.

[42] S. Doclo and M. Moonen, "Robustness of SVD-based optimal filtering noise reduction in multi-microphone speech signals," *IEEE International Work*shop on Acoustic echo and noise control, pp. 80-83, Sept. 1999.

[43] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. 8, no. 5, pp. 497-507, Sept. 2000.

[44] S. Doclo, E. De Clippel, and M. Moonen, "Combined acoustic echo and noise reduction using GSVD-based optimal filtering," *IEEE International Conference on Acoust., Speech, and Signal Processing*, vol. 2, pp. 1061-1064, June 2000.

[45] O.L. Frost, "An Algorithm for Linear Constrained Adaptive Array Processing," *Proc. IEEE*, vol.60, no.8, pp.926-935, Aug. 1972.

[46] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas Propagation,* vol. AP-30, pp. 27-34, Jan. 1982.

[47] N.K. Jablon, "Adaptive beamforming with the generalized sidelobe canceller in the presence of array imperfections," *IEEE Trans. on Antennas Propagation.*, vol. AP-34, pp. 996–1012, Aug. 1986.

[48] I. Claessen, S. Nordholm, B. A. Begtsson, and P. Eriksson, "A multi-DSP implementation of a broad-band adaptive beamformer for use in a hands-free mobile radio telephone," *IEEE Trans. on Vehicular Tech.*, vol. 40, no. 1, pp. 194-202, Feb. 1991.

[49] Y. Grenier, "A microphone array for car environments," *Speech Communication*, vol. 12, no. 1, pp. 25-39, Mar. 1993.

[50] M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," *IEEE International*

Conference on Acoust., Speech, and Signal Processing, vol. I, pp. 239-242, Apr. 1997.

[51] P. M. Peterson, "Using linearly-constrained adaptive beamforming to reduce interference in hearing aids from competing talkers in reverberant rooms," *IEEE International Conference on Acoust., Speech, and Signal Processing*, pp. 2364-2367, Apr. 1987.

[52] P. M. Zurek, J. E. Greenberg, and P. M. Peterson, "Sensitivity to design parameters in an adaptive-beamforming hearing aid," *IEEE International Conference on Acoust., Speech, and Signal Processing*, A1.10, pp. 1129-1132, Apr. 1990.

[53] W. Soede, F. Bilson, and A. J. Berkhout, "Assignment of a directional microphone array for hearing-impaired listeners" *J. Acoust. Soc. Amer.,* vol. 94, no. 2, pp. 799-808, Aug. 1993.

[54] J. M. Kates, "Superdirective arrays for hearing aids," *J. Acoust. Soc. Amer.,* vol. 94, no. 2, pp. 1930-1933, Oct. 1993.

[55] J. M. Kates, "A comparison of hearing-aid array-processing techniques," *J. Acoust. Soc. Amer.,* vol. 99, no. 5, pp. 313803148, May 1996.

[56] A. Wang, K. Yao, R. E. Hudson, D. Korompis, S. F. Soli, and S. Gao, "A high performance microphone array system for hearing aid applications," *IEEE International Conference on Acoust., Speech, and Signal Processing*, pp. 3197-3200, May 1996.

[57] K. L. Bell, Y. Ephraim, and H. L. Van Trees. **"**A Bayesian approach to A Bayesian approach to robust adaptive beamforming," *IEEE Trans. on Signal Processing*, vol. 48, pp. 386–398, Feb. 2000.

[58] C. Henry, "Robust Adaptive Beamforming," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. ASSP-35, pp. 1365-1376, Oct. 1987.

[59] O. Hoshuyama, A. Sugiyama, and A. Hirano**, "**A Robust Adaptive Beamformer for Microphone Arrays with Blocking Matrix Using Constrained Adaptive Filters," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, Oct 1999.

[60] S. A. Vorobyov, A. B. Gershman, Z. Q. Luo, **"**Robust adaptive beamforming using worst-case performance optimization: a solution to the signal mismatch problem," *IEEE Trans. on Signal Processing*, vol. 51, pp. 313-324, Feb. 2003.

[61] M. Dahl, and I. Claesson, "Acoustic noise and echo canceling with microphone array", *IEEE Trans. on Vehicular Technology*, vol. 48, pp.1518 -1526, Sept. 1999.

[62] Z. Yermeche, P. M. Garcia, N. Grbic, and I. Claesson, "A calibrated subband beamforming algorithm for speech enhancement," *Sensor Array and Multichannel Signal Processing Workshop Proceedings,* pp. 4-6, Aug. 2002

[63] S.Y. Low, N. Grbic, and S. Nordholm, "Robust microphone array using subband adaptive beamformer and spectral subtraction," *IEEE International Conference on Communication Systems,* vol. 2, pp. 25-28, Nov. 2002.

[64] J. S. Hu and Chieh-Cheng Cheng, "Frequency domain microphone array calibration and beamforming for automatic speech recognition,", *IEICE Trans. Fundamentals*, vol. E88-A, no. 9, pp. 2401-2411, Sep. 2005.

[65] L. J. Siegel and A. C. Bessey, "Voiced/Unvoiced/Mixed Excitation Classification of Speech," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol. ASSP-30, no. 3, June 1982.

[66] Y. Zhang and M. S. Scordilis, "Robust voiced/unvoiced/mixed/silence classifier with maximum a posteriori channel/background adaptation," *Proc. IEEE*, pp. 229-232, 2005.

[67] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G.Wilson, "An improved endpoint detector for isolated word recognition" , *IEEE Trans. on Acoust., Voice, Signal Processing,* v29, pp. 777–785, Aug. 1981.

[68] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., speech, signal processing,* vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[69] Dong Enqing; Liu Guizhong; Zhou Yatong; Cai Yu, "activity detection based on short-time energy and noise spectrum adaptation," *IEEE International Conference on Signal processing,* vol. 1, pp. 26-30, Aug. 2002.

[70] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," *IEEE International Conference on Acoust., Speech, and Signal Processing*, pp. 323–326, May 1977.

[71] Nemer, E., Goubran, R, and Mahmoud, S., "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. on Speech and audio processing,* vol. 9, pp. 217-231, March 2001.

[72] J. Zhang, W. Ward, and B. Pellom, "Phone based voice activity detection using online Bayesian adaptation with conjugate normal distributions," *IEEE International Conference on Acoust., Speech, and Signal Processing*, vol. 1, pp. 321-324, 2002.

[73] J. L. Shen, J. W. Hung, and L. S. Lee, " Robust entropy based endpoint detection for voice recognition in noisy environments" , *in Proc. ICSLP'96*, 1996.

[74] Junqua, J.-C., Mak, B., and Reaves, B. A robust algorithm for word boundary detection in presence of noise. *IEEE Trans. on Speech and Audio Processing,* vol. 2(3):406–412, July 1994.

[75] R. Chengalvarayan, “Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition,” *in Proc. Eurospeech*, pp. 61–64, Sep. 1999.

[76] P. Renevey and A. Drygajlo, “Entropy Based Voice Activity Detection in Very Noisy Conditions,” Proceedings of 7th European Conference on Speech Communication and Technology, *EUROSPEECH'2001*, pp. 1887-1890, Sep. 2001

[77] C. T. Lin, J. Y. Lin, and G.. D. Wu, “A robust word boundary detection algorithm for variable noise-level environment in cars,” *IEEE Trans. on Intelligent Transportation System*, vol. 3, pp. 89-101, Mar. 2002.

[78] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” *IEEE International Conference on Acoust., Speech, and Signal Processing*, vol. 1, pp. 365-368, 1998.

[79] D. V. GOKHALE, Maximum Entropy Characterization of Some Distributions. In Statistical Distributions in Scientific Work, Patil, Kotz and Ord. Eds., Boston, M.A. Reidel, Vol. 3, 299-304., 1975.

[80] R. Martin, “An efficient algorithm to estimate the instantaneous SNR of speech signals,” *EUROSPEECH'1993*, vol. 1, pp. 1093-1096, 1993.

[81] J. Ramírez, J.C. Segura, C. Benítez, d.l. Torre, Ángel, and R. Antonio, “Efficient voice activity detection algorithms using long-term speech information,” *Speech Communication*, vol. 42, pp. 271-287, April 2004.

[82] D. H. Johnson and D. E. Dudgeon, Array Signal Processing, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[83] Y. J. Hong, C. C. Yeh, and D. R. Ucci, “The effect of finite-distance signal source on a far-field steering Applebaum array—Two dimensional array case,” *IEEE Trans. on Antennas Propagation*, vol. 36, pp.468–475, Apr. 1988.

[84] J. Ringelstein, A. B. Gershman, and J. F. Böhme, “Direction finding in random inhomogeneous media in the presence of multiplicative noise,” *IEEE Signal Processing Letter*, vol. 7, pp. 269–272, Oct. 2000.

[85] D. Astely and B. Ottersten, “The effects of local scattering on direction of arrival estimation with MUSIC,” *IEEE Trans. on Signal Processing,* vol. 47, pp. 3220–3234, Dec. 1999.

[86] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. on Signal Processing*, vol. 49, pp. 1614-1626, Aug. 2001.

[87] A. P. Oppenheim, R. W. Schafer, and J. R. Buck, Discrete-time signal processing, second edition, Prentice-Hall, 1999, chapter 9.

[88] S. S. Haykin, Adaptive Filter Theory. Prentice Hall, Englewood Cliffs, NJ, 3rd edition, 1996.

[89] M. J. Grimble and A. E. SAYED, "Solution of the H∞ optimal linear filtering problem for discrete-time systems", *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 38, no. 7, pp. 1092-1104, July 1990.

[90] K. M. Nagpal and P. P. Khargonekar, "Filtering and smoothing in an H∞ setting", *IEEE Trans. On Automatic control*, vol. 37, no. 2, pp. 152-166, Feb. 1991.

[91] I. Yaesh and U. Shaked, "A transfer function approach to the problems of discrete-time systems H∞-optimal linear control and filtering", *IEEE Trans. On Automatic control*, vol. 36, no. 11, pp. 1264-1271, Nov. 1991.

[92] U. Shaked and Y. Theodor, "H∞ optimal estimation a tutorial", *Proceeding of the 31st conference on decision and control*, pp. 2278-2286, Dec. 1992.

[93] B. Hassibi, and T. Kailath, "H∞ adaptive filtering," *IEEE International Conference on Acoust., Speech, and Signal Processing*, vol. 2, pp. 949-952, May 1995.

[94] B. Hassibi, A. H. Sayed and T. Kailath, "H-infinity optimality of the LMS algorithm**",** *IEEE Trans. on Signal Processing,* vol. 44, pp. 267-280, Feb. 1996.

[95] W. Zhuang, "Adaptive H infinity channel equation for wireless personal communications," *IEEE Trans. on Vehicular Technology*, vol. 48, no. 1, pp. 126-136, January 1999.

[96] X. Shen and L. Deng, "A dynamic system approach to speech enhancement using the H∞ filtering algorithm," *IEEE Trans. on Speech and audio processing*, vol. 7, no. 4, July 1999.

[97] A.T. Erdogan, B. Hassibi and T. Kailath, "FIR H-infinity equalization of communication channels," *IEEE Trans.* on Signal Processing, vol.81, no.5, pp. 907-17, May. 2001.

[98] R. L. Burden, J. D. Faires, A. C. Reynolds, Numerical analysis, PWS-Kent Pub. Co., 1993.

[99] T. E. Shoup, Applied numerical methods for the microcomputer. Prentice-Hall, 1984, chapter 4.

[100] M. J. Maron and R. J. Lopez, Numerical analysis a practical approach, third edition. Wadsworth publishing, 1991, chapter 9.

[101] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," *International Conference on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 375-378, April 1997.

[102] C. L. Nikas. and M. Shao, Signal Processing with Alpha-Stable Distributions and Applications. New York: Wiley, 1995.

[103] I. Potamitis, H. Chen, and G. Tremoulis., "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 520-529, Sep. 2004.

[104] S. U. Pillai and B. H. Know, "Forward/Backward spatial smoothing techniques coherent signal identification," *IEEE Trans. on Acoust. Speech, Signal Processing,* vol. 37, pp. 8-14, Jan. 1989.

[105] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. on Systems, Man and Cybernetics, Part B*, vol. 34, pp, 1526-1540, Jun. 2004.

[106] J. G. Ryan and R. A. Goubran, "Array optimization applied in the near field of a microphone array," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. 8, pp. 173-176, Mar. 2000.

[107] Y. R. Zheng, R. A. Goubran, and M. K El-Tanany, "Robust near-field adaptive beamforming with distance discrimination", *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. 12, pp. 478-488, Sep. 2004.

[108] M. S. Brandstein, J. E. Adcock and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays", *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. 5, pp. 45-50, Sep. 1997.

[109] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Acoust. Speech, Signal Processing,* vol. 5, pp. 288-292, May 1997.

[110] N. B. Chong and C. M. S. See, "Sensor-array calibration using a maximum-likelihood approach," *IEEE Trans. on Antennas and Propagation*, vol. 44, pp. 827-835, June, 1996.

[111] D. B. Ward, E. A. Lehmann, and R. C Williamson,, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 826-836, Nov. 2003.

[112] H. Kuttruf, Room acoustics. London: Elsevier, 1991, chapter 3, pp. 56.

[113] D. D. Vries, E. M. Hulsebos, and J. Baan, "Spatial fluctuations in measures for spaciousness," *J. Acoust. Soc. Amer.,* vol. 110, no. 2, pp. 947-954, Aug. 2001.

[114] X. Pelorson, J-P. Vian, and J-D. Polack, "On the variability of room acoustical parameters: reproducibility and statistical validity," *Applied Acoustics*, vol. 37, pp. 175-198, 1992.

[115] G. Xuan, W. Zhang, and P. Chai, "EM algorithms of Gaussian mixture model and hidden Markov model," *IEEE Conference on Image Processing*, vol. 1, pp. 145-148, Oct 2001.

[116] M. Brandstein and D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, New York: Springer-Verlag, 2001, chapter 2, p.26.

[117] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing,* vol. 3, pp. 72-83, Jan. 1995.

[118] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, 1967.

[119] C. Elkan, “Using the Triangle Inequality to Accelerate k-Means,” *Proceedings of the Twentieth International Conference on Machine Learning, ICML*, 2003.

[120] Hidden Markov Model Toolkit (http://htk.eng.cam.ac.uk/)

[121] http://www.sym-motor.com.tw/savrin-1.htm

[122] Pham T., Sadler B. M., “Adaptive wideband aeroacoustic array processing,” *IEEE Conference on Statistical Signal and Array Processing*, Cat. No. 96TB10004, pp. 295-298, June 1996.

[123] M. Friedman and A. Kandel, Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy Logic Approaches, World Scientific Publishing Company, 1999.

[124] M. L. Seltzer and B. Raj, “Speech-recognizer-based filter optimization for microphone array processing,” *IEEE Signal Processing Letters*, vol. 10, pp. 69-71, Mar. 2003.

[125] M. L. Seltzer, B. Raj, and R. M. Stern, “ Likelihood-maximizing beamforming for robust hands-free speech recognition,“ *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 489-498, Sep. 2004.

[126] F. Asano and S. Hayamizu, “Speech enhancement using array signal processing based on the coherent-subspace method,” “*IEICE Trans. Fundamentals*, vol. E80-A, pp. 2276-2285, Nov. 1997.

[127] T. Murakami, T. Hoya, and Y. Ishida, “Speech enhancement by spectral subtraction based on subspace decomposition,” *IEICE Trans. Fundamentals*, vol. E88-A, pp. 690-701, Mar. 2005.

[128] T. Nishiura, K. Miki, S. Nakamura, K. Shikano, "Complimentary combination of microphone array and HMM composition for noisy speech recognition", *IEEE Conference on Hands-Free Speech Communication*, pp. 167-170, Apr. 2001.