國立交通大學
電機與控制工程研究所

碩士論文

使用卡曼濾波器追蹤參考訊號之
適應性語音純化波束形成器

Adaptive Beamformer for Speech Enhancement
Using Kalman Filter with Reference Signal Tracking

研究生：　朱育成
指導教授：　胡竹生　博士

中華民國一百年九月

使用卡曼濾波器追蹤參考訊號之
適應性語音純化波束形成器

Adaptive Beamformer for Speech Enhancement Using Kalman Filter with Reference Signal Tracking

研究生：朱育成　Student: Yu-Cheng Chu

指導教授：胡竹生 博士　Advisor: Dr. Jwu-Sheng Hu

國立交通大學
電機與控制工程學系
碩士論文

A Thesis
Submitted to Institute of Electrical and Control Engineering
College of Electrical Engineering
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of Master
in

Electrical and Control Engineering

September 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年九月

使用卡曼濾波器追蹤參考訊號之
適應性語音純化波束形成器

研究生：朱　育　成　　　　　　　　指導教授：胡　竹　生　博士

國立交通大學電機與控制工程研究所碩士班

摘　要

本論文提出一套利用麥克風陣列來降低噪音及迴響效應的演算法。在實際環境中，目標音訊不只常受到穩態雜訊及非穩態雜訊的干擾，更常因為迴響效應而使語音品質遭到破壞。因此，本論文期望設計一個能濾除雜訊並減少迴響影響的適應性波束形成器。此演算法在波束形成器演算法中，引入參考訊號的觀念並輔以 Kalman 濾波器來進行演算。此外，經過些微的修改，本演算法也可以利用於偵測語音活動。藉由適當的語音活動偵測，可以幫助分辨目標與噪音在本質上的不同，並且加速 Kalman 濾波器的收斂。利用實際在車上錄得的音檔進行的實驗結果也在此篇論文中呈現。本論文並利用客觀的參數評估所提出的波束形成器與語音活動偵測的效能，並與其他已知的方法進行比較分析。

Adaptive Beamformer for Speech Enhancement Using Kalman Filter with Reference Signal Tracking

Student：Yu-Cheng Chu　　　　Advisor：Dr. Jwu-Sheng Hu

Institute of Electrical and Control Engineering

ABSTRACT

In this thesis, an algorithm that considers noise reduction and de-reverberation simultaneously using microphone array is proposed. In many practical environments, the desired speech signal is usually contaminated by stationary or non-stationary noises and distorted by reverberation. When considering noise reduction only, the desired speech signal could be distorted further due to the effect of desire signal cancellation etc. The objective of this thesis is to design an adaptive beamformer to incorporate de-reverberation into the noise reduction framework. The proposed method tracks a pre-recorded reference signal to compensate the reverberation effect. Consequently, the algorithm results in a trade-off between the two objectives. Further, a voice activity detection (VAD) algorithm is proposed by slightly modifying the proposed algorithm. An adequate VAD can help to identify the nature of signal and noise and accelerate the convergence rate of Kalman filter. The experiments on real car sound samples are processed. The performance of beamformer and voice activity detection are both evaluated and compared with existing algorithms.

致謝

這篇論文的完成，首先感謝我的指導教授胡竹生教授。他給予我許多栽培及建議、討論，讓我在思索、研究的過程中更能瞭解科學研究的精神，讓我收益良多。接著感謝我的父親朱榮洲以及母親陳秀珍，在這兩年的求學過程中，讓我可以無後顧之憂的完成學位，而父母親在心靈上以及感情上的支持，更是無時無刻支持我走下去的動力。有太多時候我都很挫折、很沮喪，但是只要聽到爸媽的聲音、想起你們的殷切盼望以及無私的栽培，我就又能鼓起勇氣繼續向前。

在兩年的過程中，X-Lab 實驗室就像我的第二個家，在這裡有許多共同求學的同學們以及親切又知識淵博的學長們。最先感謝的是李明唐學長，這兩年來，每次與他討論的過程總讓我進步許多，他給的建議也都能發揮莫大的助益，對學長的感激溢於言表。而已經畢業的楊佳興學長，在我研究的過程中，給我許多的文件，讓我等於直接吸收了他多年的功力，讓我的見識又更上一層。另外，溫柔而親切的永融學長、熱心也很有想法的阿吉學長、有趣但是運氣一直不太好的崇維學長、威武聰明但不失幽默的冠群學長、溫和愛運動的智謙學長、開朗陽光的庭昭學長，諸位學長們在我求學的過程中都讓育成的人生增添了許多的色彩與長進。另外一位不得不提的則是鎮宇學長，能進入 X-Lab 就先要感謝他兩三年前的建議。在實驗室的最後一年，能和鎮宇學長一起在實驗室做研究、到處吃喝、打球，甚至是傾吐心事，都讓我的碩二過得快樂許多。實驗室的同學們，感謝很 high 點子很多的湘筑、有活力有想法的偉庭、見解獨到很能幫助討論的建安、很健談幽默的新文、不健談但很幽默的昀軒、很健談但不幽默的學文，這兩年來有各位的陪伴，真是育成的榮幸。學弟們，很嘴砲很開朗的昭男、顯圖很閃的耕維、做事認真的哲鳴以及一樣有趣的建廷宗翰，這些日子以來感謝你們。

電資 98 的各位，各位的陪伴與支持也不可或缺。嘴砲的雄獅一哥宗憲、在國外努力生活的紹甫、一起打 BG 打球的紹丞、很嘴砲也很風趣的育瑞、就在對面很常遇到可以分享垃圾事聊天的冠甫、奕奇，這六年有你們真好！

謝謝交通大學這六年來的栽培，這絕對是我人生中精彩的旅程！

CONTENT

List Of Tables

List Of Figures

Chapter 1. INTRODUCTION

1.1 Motivation and Objective

Our hearing is perhaps the the most useful sense except vision. However, the information retrieved from hearing is usually contaminated by undesired sources. Although human beings are able to recognize desired speeches under interferences, it is still considered as a difficult task for computers or machines.

A common sensor for receiving sound wave is the microphone. Single microphone can collect spectral information but not spatial information. To retrieve more information among the sound wave, a collection of microphones, or microphone array, is applied to catch not only spectral information but also spatial information. Among several existing microphone-array-based enhancement algorithms, beamformer is one of the most popular methods and was extensively studied for hands-free speech communication or recognition.

Background noise and reverberation are the most common origin to signal quality degradation. The background noise is from undesired noise source or interferences. The spectral and spatial likelihood between noise and desired source will determine the difficulty of removing noise. The reverberation level will determine the distortion to the desired source. The reverberation level is commonly affected by the reflection ratio and openness of the environment.

The purpose of the thesis is to design a beamformer that handles both noise reduction and dereverberation. A scenario like car environment is possible to occur in real life, where the quality of sound is seriously deteriorated by engine noise and wind noise and reverberation from the narrow space of cabin.

1.2 Literature Review

The beamformers can be categorized in two types, fix beamformers and adaptive beamformers. Most of the fix beamformers are simpler than the adaptive beamformers. The implementation costs of fix beamformers are often lower than the adaptive counterparts.

Fix beamformers includes delay-and-sum beamformer (DSB) [11], constant directivity beamformer (CDB) [12] and fixed superdirective beamformers [13]. They utilize fixed coefficients to achieve a desired spatial response. The DSB is the simplest structure in fix beamformers. It first compensates to the relative time delay between distinct microphone signals and then sums the steered signal to form a single output. CDB is designed to maintain the spatial response equal over a wide frequency band while the fixed superdirective beamformer attempts to suppress noise coming from all directions without affecting the desired speech signal from a principal direction. Fix beamformers generally assume the desired sound source, interference signals, and noises are slowly varying and at known locations. Therefore, these algorithms are sensitive to steering errors, which limit their noise suppression capability and give rise to the desired signal distortion and cancellation. Furthermore, these algorithms also have limited performance under highly reverberation environments.

Instead of using fixed coefficients to suppress noises and interference signals, an adaptive beamformer can form its directivity beam-pattern to the desired signal and its null beam-pattern to the undesired signals. In the fixed beamformers, the beam-pattern of null only exists when the direction of noise is known and remains unchanged. To cope with environmental changes, various adaptive beamformers were proposed to improve the performance. One key issue in adaptive beamformers is the sensitivity due to the mismatch between the actual desired signal steering vector and the presumed one [10]. The mismatch can be induced by signal pointing errors [14], imperfect array calibration [15], or channel effect. In the presence of these effects, an adaptive beamformer can easily mix up the desired signal and interference components; that is, it suppress the desired signal instead of maintaining distortionless response. This phenomenon is commonly referred to signal self-nulling [16]. As a result, much effort has been devoted to the noise reduction and dereverberation.

Many adaptive beamformer techniques were extensively studied. The linearly constrained minimum variance (LCMV) beamformer was proposed in [17] to minimize the array output power under a look direction constraint. A form similar to LCMV is minimum variance distortionless response (MVDR) proposed by Capon in [1]. Another popular technique is the generalized sidelobe canceller (GSC) algorithm which essentially transforms the LCMV constrained minimization problem into an unconstrained one [18].

The formulation of MVDR is then revisited in [5] with Kalman filter by introducing the concept of state space. To improve the robustness against steering vector error, various methods are investigated [10]. The Kalman filter can be also substituted by H-infinity filter or Second Order Kalman filter or Second Order H-infinity filter [19] to enhance its robustness and reducing non-linearity.

Among adaptive beamformers, the usage of pre-recorded data is a solution to solve channel effect. The algorithm by Dahl et. al can be found in [20], which give rise to the reference signal concept in proposed algorithm.

1.3 Thesis Subject and Contribution

The contribution of this thesis is to propose and implement an innovative algorithm for speech enhancement. The subject of this thesis can be divided into two parts. The first part is to formulate a new beamformer considering given the information of pre-recorded data. The solution to the formulation is presented. The second part is to handle the resulting voice activity detection problem by the same formulation but only changes the parameters to render different results.

In the first part, the formulation using MVDR with pre-recorded signal is given. To solve the formulation, the linear first order Kalman filter is used. In the Kalman filter, the selection of parameters will pose different result among noise reduction and dereverberation. The tradeoff effect is discussed and explained.

In the second part, the same formulation is used to implement a voice activity detector. The design and parameter choosing technique are explained and discussed. Besides, the information given by the voice activity detector can be reused to finding the appropriate parameter in beamforming.

The experiment results are shown to verify the performance of the proposed algorithm, both in beamforming and voice activity detection.

1.4 Outlines of Thesis

The remainder of this thesis is organized as follows.

Chapter 2: The basic beamforming technique Minimum Variance Distortionless Response (MVDR) is introduced. The optimal solution of MVDR is presented. The method of incorporating state space formulation into solving MVDR and solve it with Kalman filter is investigated. These constructed the foundation of proposed algorithm.

Chapter 3: The detailed concept of reference signal based Kalman filter for beamformer is stated. It includes the beamforming formulation and voice activity detection. In beamforming, the formulation and its solution are presented. The technique of choosing the parameter and its effect are also discussed. In voice activity detection (VAD), the design and implementation are investigated. The method of utilizing the information from VAD to decide the parameters in beamforming is also described. Finally, the overall architecture is illustrated and explained.

Chapter 4: The experiment results are presented. It contains experiments regarding beamforming capability and voice activity detection. Some objective indices are calculated to compare the performance of proposed algorithm and former algorithms.

Chapter 5: The conclusion of this thesis and some issue that is still not clear is discussed is this chapter.

Chapter 2. BEAMFORMER USING KALMAN FILTER

2.1 Introduction

Kalman filter is a well-known optimal estimation filter in control theory. In this thesis, the use of Kalman filter in signal processing is more concerned. To begin with, a conventional beamformer MVDR proposed by Capon [1] is introduced. The main idea of MVDR is to minimize undesired noise while maintaining desired signal with known DOA, or Direction of Arrival, distortionless. Such idea can be formulated as a minimization problem with certain constraints. Conventional way to solve it is using Lagrange Multiplier and achieves optimal solution. Its optimal solution is presented in Section 2.2. In Section 2.3, the technique of incorporating state space concept and Kalman filter to solve MVDR problem is presented. The solution can be found using conventional Kalman filter solution. In later sections, another formulation to maintain the distortionless constraint will be presented and investigated.

2.2 Beamformer under MVDR Structure

The minimum variance distortionless response (MVDR) beamformer, also known as Capon beamformer [1], minimizes the output power of the beamformer under a single linear constraint on the response of the array towards the desired signal.

Consider the conventional signal model in which an M-element microphone array captures a convolved desired signal (speech source) in some noise field. The received signals are expressed as [2], [3], [4]

$$x_m(k) = a_m * s(k) + v_m(k) \qquad \mathrm{m} = 1,2,\ldots\mathrm{M}\,, \tag{2.1}$$

where a_m is the impulse response from the unknown (desired) source $s(k)$ to the m_{th} microphone, * stands for convolution, and $v_m(k)$ is the noise at the microphone m. The signals $s(k)$ and $v_m(k)$ are assumed as uncorrelated and zero mean.

In the frequency domain, (2.1) can be written as

$$X_m(jw) = A_m(jw) * S(jw) + V_m(jw) \qquad \mathrm{m} = 1,2,...\mathrm{M}, \tag{2.2}$$

where $A_m(jw)$, $S(jw)$, $X_m(jw)$, $V_m(jw)$ are the discrete-time Fourier transforms (DTFTs) of $a_m(k)$, $s(k)$, $x_m(k)$, $v_m(k)$, respectively, at angular frequency w $(-\pi < w \leq \pi)$ and j is the imaginary unit ($j^2 = -1$).

These M microphone signals in the frequency domain are summarized in a vector notation as

$$\mathbf{X}(jw) = \mathbf{A}(jw)S(jw) + \mathbf{V}(jw) \quad , \tag{2.3}$$

where

$$\begin{aligned}
\mathbf{X}(jw) &= [X_1(jw)\ X_2(jw)\ \cdots X_M(jw)]^T \\
\mathbf{A}(jw) &= [A_1(jw)\ \ A_2(jw)\ \ \cdots\ A_M(jw)]^{\mathrm{T}} \\
\mathbf{V}(jw) &= [V_1(jw)\ \ V_2(jw)\ \ \cdots\ V_M(jw)]^T
\end{aligned}$$

and superscript ^{T} denotes transpose of a vector or a matrix.

Consider finding a weight vector w_{MV} which satisfies the look direction constraint

$$\mathbf{w}_{MV}^H(jw)\mathbf{a}(\theta_s, jw) = 1 \tag{2.4}$$

while attempting to minimize beamformer output power

$$E\{|Y(jw)|^2\} = E\{|\mathbf{w}_{MV}^H(jw)\mathbf{X}(jw)|^2\} \equiv \mathbf{w}_{MV}^H(jw)\mathbf{R}_{XX}(jw)\mathbf{w}_{MV}(jw) \tag{2.5}$$

in order to suppress undesired interference from $\theta \neq \theta_s$ and noise. $Y(jw)$ is the beamformer output given by

$$Y(jw) = \mathbf{w}_{MV}^{H}(jw)\ \mathbf{X}(jw)\,. \tag{2.6}$$

$\mathbf{a}(\theta_s, jw)$ is the array manifold vector that points to the source direction.

With the consideration above, the following constrained optimization problem can be formulated:

$$\min \mathbf{w}_{MV}^{H}(jw)\mathbf{R}_{XX}(jw)\mathbf{w}_{MV}(jw) \text{ subject to } \mathbf{w}_{MV}^{H}(jw)\mathbf{a}(\theta_s, jw) = 1 \tag{2.7}$$

To solve this problem, the Lagrange Multiplier is incorporated.

$$\begin{cases} \nabla_{W_{MV}(jw)}\mathbf{w}_{MV}^{H}(jw)\mathbf{R}_{XX}(jw)\mathbf{w}_{MV}(jw) - \lambda\nabla_{W_{MV}(jw)}[\mathbf{w}_{MV}^{H}(jw)\mathbf{a}(\theta_s, jw) - 1] = 0 \\ \mathbf{w}_{MV}^{H}(jw)\mathbf{a}(\theta_s, jw) = 1 \end{cases} \tag{2.8}$$

(2.8) can be reduced to

$$\begin{cases} \mathbf{R}_{XX}(jw)\mathbf{w}_{MV}(jw) = \lambda\mathbf{a}(\theta_s, jw) \\ \mathbf{w}_{MV}^{H}(jw)\mathbf{a}(\theta_s, jw) = 1 \end{cases} \tag{2.9}$$

Assuming $\mathbf{R}_{XX}$ is nonsingular. Then

$$\mathbf{w}_{MV}(jw) = \frac{\mathbf{R}_{XX}^{-1}(jw)\mathbf{a}(\theta_s, jw)}{\mathbf{a}^{H}(\theta_s, jw)\mathbf{R}_{XX}^{-1}(jw)\mathbf{a}(\theta_s, jw)}, \tag{2.10}$$

which is the optimal solution to MVDR problem proposed by Capon[3] and is thoroughly evaluated in [4].

2.3 Beamformer Using Kalman Filter under MVDR Structure

The traditional formulation and solution to MVDR is presented in Section 2.2. In this section, The Kalman filter is introduced to solve the MVDR problem in a new formulation by Y.H. Chen and C.T. Chiang [5].

With the same formulation as (2.7), the two equations are written in model measurement equation as

$$\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}\mathbf{X}^H(k, jw)\\ \mathbf{a}^H(\theta_s, jw)\end{bmatrix}\mathbf{w}(k, jw) + \begin{bmatrix}v_1(k, jw)\\ v_2(k, jw)\end{bmatrix} \text{ or } \mathbf{Y} = \mathbf{B}^H(k, jw)\mathbf{w}(k, jw) + \mathbf{V}(k, jw), \qquad (2.8)$$

where $\mathbf{Y} = [0 \quad 1]^T$, and the input vector is given by

$$\mathbf{B}^H(k, jw) = \begin{bmatrix}\mathbf{X}^H(k, jw)\\ \mathbf{a}^H(\theta_s, jw)\end{bmatrix} \qquad (2.9)$$

and the measurement noise vector is

$$\mathbf{V}(k, jw) = \begin{bmatrix}v_1(k, jw)\\ v_2(k, jw)\end{bmatrix}. \qquad (2.10)$$

Here, $v_1(k, jw)$ is the residual error and $v_2(k, jw)$ is the constraint error. By the assumption that $v_1(k, jw)$ and $v_2(k, jw)$ are uncorrelated, the correlation matrix of $\mathbf{V}(k, jw)$ can be written as

$$Q = \begin{bmatrix}\sigma_{v_1}^2 & 0\\ 0 & \sigma_{v_2}^2\end{bmatrix} \qquad (2.11)$$

Since the optimum-constrained weight vector $\mathbf{w}(k, jw)$ is a constant all the time for the stationary environment [6], the truth-model process equation of the constrained Kalman algorithm may be written as

$$\mathbf{w}(k, jw) = \mathbf{w}(k-1, jw). \qquad (2.12)$$

With the process equation (2.12) and measurement equation (2.8), the constrained Kalman algorithm can minimize the residual error in the mean-square sense while maintaining a distortionless response along the look direction.

After applying the discrete Kalman filter theory with (2.12) and (2.8), the filtered estimate of the weight vector is recursively given by [6]

$$\hat{\mathbf{w}}(k, jw) = \hat{\mathbf{w}}(k-1, jw) + \mathbf{K}(k, jw)[\mathbf{Y}(k, jw) - \mathbf{B}^H(k, jw)\hat{\mathbf{w}}(k-1, jw)], \qquad (2.13)$$

where the Kalman Gain $\mathbf{K}(k, jw)$ can be calculated recursively by

$$\mathbf{K}(k, jw) = \mathbf{R}_{ee}(k-1, jw)\mathbf{B}(k, jw)[\mathbf{B}^H(k, jw)\mathbf{R}_{ee}(k-1, jw)\mathbf{B}(k, jw) + Q]^{-1}. \tag{2.14}$$

Here the filtered weight-error correlation matrix $\mathbf{R}_{ee}(k, jw)$ is

$$\mathbf{R}_{ee}(k, jw) = [\mathbf{I} - \mathbf{K}(k, jw)\mathbf{B}^H(k, jw)]\mathbf{R}_{ee}(k-1, jw), \tag{2.15}$$

where $\mathbf{I}$ is an m-by-m identity matrix. Using (2.8), (2.12), (2.13), the signal-flow graph of the constrained Kalman algorithm can be plotted as Fig. 1 [5].

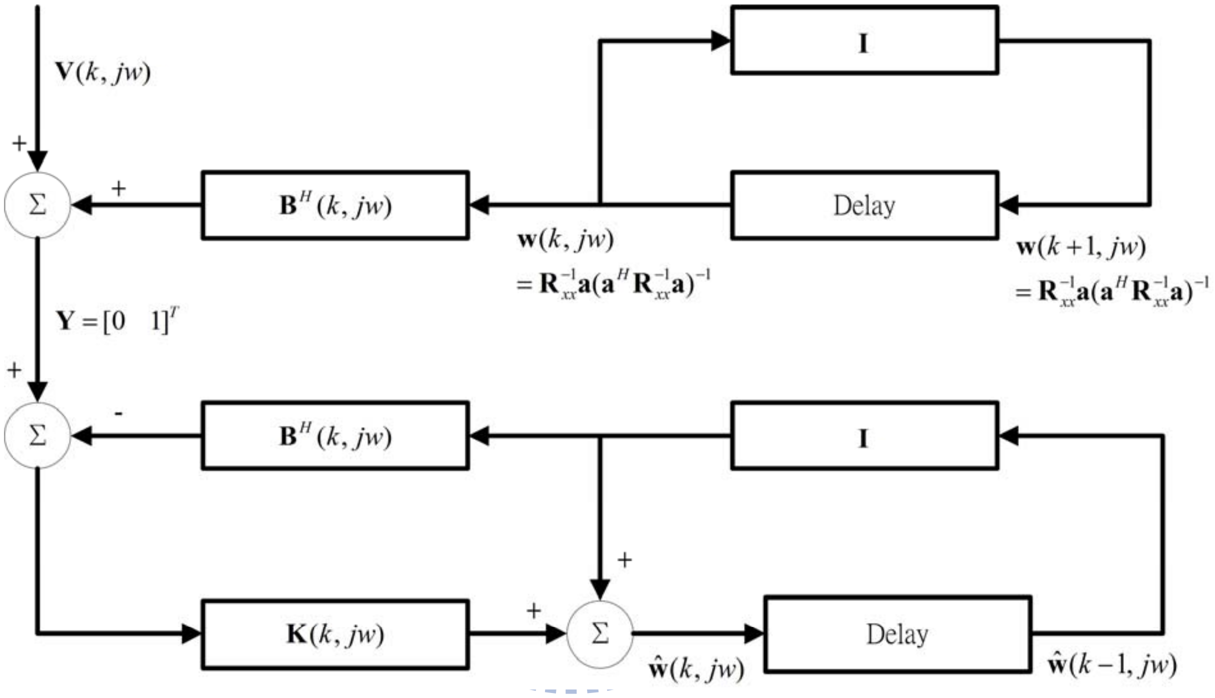

$\mathbf{V}(k, jw)$
+
Σ
+
$\mathbf{B}^H(k, jw)$
$\mathbf{I}$
Delay
$\mathbf{w}(k, jw) = \mathbf{R}_{xx}^{-1}\mathbf{a}(\mathbf{a}^H\mathbf{R}_{xx}^{-1}\mathbf{a})^{-1}$
$\mathbf{w}(k+1, jw) = \mathbf{R}_{xx}^{-1}\mathbf{a}(\mathbf{a}^H\mathbf{R}_{xx}^{-1}\mathbf{a})^{-1}$
$\mathbf{Y} = [0 \quad 1]^T$
+
Σ
-
$\mathbf{B}^H(k, jw)$
$\mathbf{I}$
+
$\mathbf{K}(k, jw)$
+
Σ
Delay
$\hat{\mathbf{w}}(k, jw)$
$\hat{\mathbf{w}}(k-1, jw)$

Fig. 1 Block diagram of the constrained Kalman algorithm without prior estimation of the desired signal

Chapter 3. REFERENCE-SIGNAL-BASED BEAMFORMER

3.1 Introduction

The MVDR structure can be modified by exploiting the distortionless response. To maintain desired signal distortionless, the algorithm incorporating pre-recorded signal as reference signal is proposed. The reference signal enhances the distortionless constraint by using more information and reduces the loading of carefully estimating the system parameters of the environment. More detailed explanation will be presented in this chapter.

In this chapter, the main algorithm of using reference signal to merge with MVDR and Kalman filter is presented. In Section 3.2, the formulation under that concept is proposed and described. In Section 3.3, the solution to solving proposed formulation is thoroughly investigated. In Section 3.4, the know-how of how to choose the parameters of the Kalman filter is discussed. The tradeoff phenomenon between the parameters is introduced and explained also in Section 3.4. The design and implementation of voice activity detection elaborating the same Kalman filter is presented in Section 3.5. The threshold decision method and parameter selection method is introduced in Section 3.6. The overall system architecture is illustrated and explained in Section 3.7.

3.2 Formulation of Referenced-Signal-Based Beamformer Using Kalman Filter

In this section, the proposed reference-signal-based beamformer using Kalman filter is presented.

In MVDR beamformer, the distortionless requirement is achieved by add a constraint that maintains the signal from a known direction unchanged. This constraint also avoids choosing the naïve solution of zero during the minimization process. In addition, this constraint also achieves the requirement of dereverberation since it not only preserves signal from desired direction, but also drop reverberation signals from other directions during the minimization process.

Another approach to maintain distortionless requirement is to estimate the acoustic transfer function (ATF) from the desired signal source to the microphone array. The ATFs can specifically describe the relationship from desired source to the microphone array including the effect of reverberation. With the ATFs, source signal can be regenerated with low distortion as long as the ATFs are estimated correctly and the surrounding environment is linearly time-invariant (LTI) and does not change during the filtering process.

However, estimating the ATFs is a cumbersome and tedious work. To avoid such process but still get useful knowledge of the environment, the concept of reference-signal is incorporated. The reference-signal is acquired by recording the signal while playing a known clip at the position of source. The received signal can be considered as the output of the known input processed by the surrounding environment functioned as the system. With the input and output information of the system, it can be considered as a reference to the environment and thus achieving the requirement of distortionless better and easier.

The conventional formulation of MVDR in Section 2.2 is

$$\min \mathbf{w}_{MV}^{H}(jw)\mathbf{R}_{XX}(jw)\mathbf{w}_{MV}(jw) \text{ subject to } \mathbf{w}_{MV}^{H}(jw)\mathbf{a}(\theta_s, w) = 1. \tag{2.7}$$

To incorporate the reference-signal, the formulation can be used to substitute the distortionless constraint and becomes

$$\begin{aligned} &\min \mathbf{w}_{MV}^{H}(jw)\mathbf{R}_{XX}(jw)\mathbf{w}_{MV}(jw) \\ &\text{subject to } \mathbf{w}_{MV}^{H}(jw)\mathbf{X}_{r}(jw) = s_r(jw), \end{aligned} \tag{3.1}$$

where $s_r(jw)$ is the discrete-time Fourier transforms (DTFTs) of the played known clip and $\mathbf{X}_r(jw)$ is discrete-time Fourier transforms (DTFTs) of the received signal while playing the known clip. $\mathbf{X}_r(jw) = [X_{r,1} \quad X_{r,2} \quad \cdots \quad X_{r,M}]^T$, where $X_{r,m}$ is the discrete-time Fourier transforms (DTFTs) of received signal at the m_{th} microphone. The subscript "$_r$" implies "pre-recorded" since the reference signal is recorded before the filtering process.

3.3 Solution to the Proposed Formulation

With the formulation above,

$$\begin{aligned} &\min \mathbf{w}_{MV}^{H}(jw)\mathbf{R}_{XX}(jw)\mathbf{w}_{MV}(jw) \\ &\text{subject to } \mathbf{w}_{MV}^{H}(jw)\mathbf{X}_{r}(jw) = s_r(jw), \end{aligned} \tag{3.1}$$

the solution to the formulation will be presented in this section [2].

From (3.1), state equations describing such formulation can be written as

Measurement Equation:

$$\begin{bmatrix} 0 \\ s_r(k,w) \end{bmatrix} = \begin{bmatrix} \mathbf{X}^H(k,w) \\ \mathbf{X}_r^H(k,w) \end{bmatrix} \mathbf{w}(k,w) + \mathbf{V}(k,w) \tag{3.2}$$

Process Equation:

$$\mathbf{w}(k+1,w) = \mathbf{w}(k,w) + \mathbf{Q}(k,w), \tag{3.3}$$

where k is the frame index and the superscript "^{H}" means conjugate-transpose. The noise $\mathbf{V}(k,w)$ and $\mathbf{Q}(k,w)$ are assumed with Gaussian distribution and thus the covariance matrix can be written as

$$\begin{aligned}&\mathbf{Q}(k,w) \sim N(0,\sigma_Q I)\\&\mathbf{V}(k,w) \sim N(0,\sigma_v \begin{bmatrix}1 & 0\\0 & \rho_v\end{bmatrix}),\end{aligned} \tag{3.4}$$

where "N" means Normal Distribution and σ_Q, σ_v, ρ_v are parameter to be chosen. $\mathbf{X}(k,w)$ is the received signal when desired signal is inactive, since the desired signal cannot always be guaranteed uncorrelated with the reference-signal. Once desired signal is correlated with reference signal, the phenomenon "desired signal cancelation" will occur and yield huge degradation to the desired signal.

Let the state estimation error is

$$\mathbf{e}(k|k-1,w) = \mathbf{w}(k,w) - \hat{\mathbf{w}}(k|k-1,w), \tag{3.5}$$

and the error covariance matrix is

$$\mathbf{R}_{ee}(k|k-1,w) = E[\mathbf{e}(k|k-1,w)\mathbf{e}^T(k|k-1,w)] \tag{3.6}$$

In the first step, no new observation is used. To predict $\mathbf{w}(k)$ using the state equation, the best possible predictor given no new information is available would be

$$\hat{\mathbf{w}}(k|k-1,w) = \hat{\mathbf{w}}(k-1|k-1,w). \tag{3.7}$$

The estimation error is

$$\begin{aligned}\mathbf{e}(k|k-1,w) &= \mathbf{w}(k,w) - \hat{\mathbf{w}}(k|k-1,w)\\&= \mathbf{w}(k-1,w) + \mathbf{Q}(k,w) - \hat{\mathbf{w}}(k-1|k-1,w)\\&= \mathbf{e}(k-1|k-1,w) + \mathbf{Q}(k,w)\end{aligned} \tag{3.8}$$

If requiring that $E[\mathbf{e}(k-1|k-1,w)] = 0$ (this zero-mean condition states that there is no constant bias in the optimal linear estimation [7]), $E[\mathbf{e}(k|k-1,w)]=0$. Since $\mathbf{e}(k-1|k-1,w)$ is uncorrelated with $\mathbf{Q}(k,w)$,

$$\mathbf{R}_{ee}(k|k-1,w)=\mathbf{R}_{ee}(k-1|k-1,w)+\sigma_Q^2\mathbf{I}. \tag{3.9}$$

This is the Riccati Equation.

In the second step, the new observation, $\begin{bmatrix} 0 \\ s_r(k,w) \end{bmatrix}=\mathbf{Y}(k,w)$ is incorporated to estimate $\mathbf{w}(k,w)$. A linear estimate that is based on $\hat{\mathbf{w}}(k|k-1,w)$ and $\mathbf{Y}(k,w)$ has the form

$$\hat{\mathbf{w}}(k|k,w)=\mathbf{K}'(k,w)\hat{\mathbf{w}}(k|k-1,w)+\mathbf{k}(k,w)\mathbf{Y}(k,w), \tag{3.10}$$

where $\mathbf{K}'(k,w)$ and $\mathbf{k}(k,w)$ are some matrix and vector to be determined. The vector $\mathbf{k}(k,w)$ is called the Kalman gain. Now, the estimation error is

$$\begin{aligned}\mathbf{e}(k|k,w)&=\mathbf{w}(k,w)-\hat{\mathbf{w}}(k|k,w)\\&=\mathbf{w}(k,w)-\mathbf{K}'(k,w)\hat{\mathbf{w}}(k|k-1,w)-\mathbf{k}(k,w)\mathbf{Y}(k,w)\\&=\mathbf{w}(k,w)-\mathbf{K}'(k,w)[\mathbf{w}(k,w)-\mathbf{e}(k|k-1,w)]-\mathbf{k}(k,w)[\boldsymbol{X}^H(k,w)\mathbf{w}(k,w)+\mathbf{V}(k,w)]\\&=[\mathbf{I}-\mathbf{K}'(k,w)-\mathbf{k}(k,w)\boldsymbol{X}^H]\mathbf{w}(k,w)+\mathbf{K}'(k,w)\mathbf{e}(k|k-1,w)-\mathbf{k}(k,w)\mathbf{V}(k,w),\end{aligned} \tag{3.11}$$

where $\boldsymbol{X}^H(k,w)=\begin{bmatrix}\mathbf{X}^H(k,w)\\ \mathbf{X}_r^H(k,w)\end{bmatrix}$.

Since $E[\mathbf{e}(k|k-1,w)]=0$, then $E[\mathbf{e}(k|k,w)]=0$ only if

$$\mathbf{K}'(k,w)=\mathbf{I}-\mathbf{k}(k,w)\boldsymbol{X}^H \tag{3.12}$$

With this constraint, it follows that

$$\begin{aligned}\hat{\mathbf{w}}(k|k,w)&=[\mathbf{I}-\mathbf{k}(k,w)\boldsymbol{X}^H]\hat{\mathbf{w}}(k|k-1,w)+\mathbf{k}(k,w)\mathbf{Y}(k,w)\\&=\hat{\mathbf{w}}(k|k-1,w)+\mathbf{k}(k,w)[\mathbf{Y}(k,w)-\boldsymbol{X}^H\hat{\mathbf{w}}(k|k-1,w)],\end{aligned} \tag{3.13}$$

and

$$\begin{aligned}\mathbf{e}(k|k,w) &= \mathbf{K}'(k,w)\mathbf{e}(k|k-1,w) - \mathbf{k}(k,w)\mathbf{V}(k,w) \\ &= [\mathbf{I} - \mathbf{k}(k,w)\boldsymbol{X}^H]\mathbf{e}(k|k-1,w) - \mathbf{k}(k,w)\mathbf{V}(k,w).\end{aligned} \tag{3.14}$$

Since $\mathbf{V}(k,w)$ is uncorrelated with $\mathbf{Q}(k,w)$ and with $\mathbf{Y}(k-1,w)$, then $\mathbf{V}(k,w)$ will be uncorrelated with $\mathbf{w}(k,w)$ and with $\hat{\mathbf{w}}(k|k-1,w)$; as a result $E[\mathbf{e}(k|k,w)\mathbf{V}(k,w)]$=0. Therefore, the error covariance matrix for $\mathbf{e}(k|k,w)$ is

$$\begin{aligned}\mathbf{R}_{ee}(k|k,w) &= E[\mathbf{e}(k|k,w)\mathbf{e}^T(k|k,w)] \\ &= [\mathbf{I} - \mathbf{k}(k,w)\boldsymbol{X}^H]\mathbf{R}_{ee}(k|k-1,w)[\mathbf{I} - \mathbf{k}(k,w)\boldsymbol{X}^H]^T + \mathbf{k}(k,w)\mathbf{R}_v(k,w)\mathbf{k}^T(k,w),\end{aligned} \tag{3.15}$$

Where $\mathbf{R}_v(k,w) = \sigma_v\begin{bmatrix} 1 & 0 \\ 0 & \rho_v \end{bmatrix}$.

The final task is to find the Kalman gain vector $\mathbf{k}(k,w)$, that minimizes the MSE

$$J(k) = tr[\mathbf{R}_{ee}(k|k,w)] \tag{3.16}$$

Differentiating $J(k)$ with respect to $\mathbf{k}(k,w)$, we get

$$\frac{\partial J(k)}{\partial \mathbf{k}(k,w)} = -2[\mathbf{I} - \mathbf{k}(k,w)\boldsymbol{X}^H]\mathbf{R}_{ee}(k|k-1,w)\boldsymbol{X} + 2\mathbf{k}(k,w)\mathbf{R}_v(k,w) \tag{3.17}$$

and equating it to zero, we deduce the Kalman gain

$$\mathbf{k}(k,w) = \mathbf{R}_{ee}(k|k-1,w)\boldsymbol{X}\left[\boldsymbol{X}^H\mathbf{R}_{ee}(k|k-1,w)\boldsymbol{X} + \mathbf{R}_v(k,w)\right]^{-1} \tag{3.18}$$

The expression for the error covariance matrix can be simplified as

$$\begin{aligned}\mathbf{R}_{ee}(k|k,w) &= [\mathbf{I} - \mathbf{k}(k,w)\boldsymbol{X}^H]\mathbf{R}_{ee}(k|k-1,w) - \\ &\{[\mathbf{I} - \mathbf{k}(k,w)\boldsymbol{X}^H]\mathbf{R}_{ee}(k|k-1,w)[\mathbf{I} - \mathbf{k}(k,w)\boldsymbol{X}^H] + \mathbf{R}_v(k,w)\mathbf{k}(k,w)\}\mathbf{k}^T(k,w),\end{aligned} \tag{3.19}$$

Where, by using (3.17), the second term in (3.19) is equal to zero. Hence

$$\mathbf{R}_{ee}(k|k,w) = [\mathbf{I} - \mathbf{k}(k,w)\boldsymbol{X}^H]\mathbf{R}_{ee}(k|k-1,w) \tag{3.20}$$

In conclusion, the Kalman filter can be summarized as

State Equation:

$$\hat{\mathbf{w}}(k+1|k,w)=\hat{\mathbf{w}}(k|k,w)+\mathbf{Q}(k,w)$$

Observation Equation (or Measurement Equation):

$$\mathbf{Y}(k,w)=\begin{bmatrix}0\\ s_r(k,w)\end{bmatrix}=\begin{bmatrix}\mathbf{X}^H(k,w)\\ \mathbf{X}_r^H(k,w)\end{bmatrix}\mathbf{w}(k,w)+\mathbf{V}(k,w)=\boldsymbol{X}^H(k,w)\mathbf{w}(k,w)+\mathbf{V}(k,w)$$

Initialization:

$$\hat{\mathbf{w}}(0|0,w)=E[\mathbf{w}(0,w)] \qquad \mathbf{R}_{ee}(0|0,w)=E[\mathbf{w}(0)\mathbf{w}^T(0)]$$

Computation for $k=1,2,\cdots$

$$\hat{\mathbf{w}}(k|k-1,w)=\hat{\mathbf{w}}(k-1|k-1,w)$$

$$\mathbf{R}_{ee}(k|k-1,w)=\mathbf{R}_{ee}(k-1|k-1,w)+\sigma_Q^2\mathbf{I}$$

The Kalman gain:

$$\mathbf{k}(k,w)=\mathbf{R}_{ee}(k|k-1,w)\boldsymbol{X}\left[\boldsymbol{X}^H\mathbf{R}_{ee}(k|k-1,w)\boldsymbol{X}+\mathbf{R}_v(k,w)\right]^{-1}$$

$$\hat{\mathbf{w}}(k|k,w)=\hat{\mathbf{w}}(k|k-1,w)+\mathbf{k}(k,w)[\mathbf{Y}(k,w)-\boldsymbol{X}^H(k,w)\hat{\mathbf{w}}(k|k-1,w)]$$

$$\mathbf{R}_{ee}(k|k,w)=[\mathbf{I}-\mathbf{k}(k,w)\boldsymbol{X}^H(k,w)]\mathbf{R}_{ee}(k|k-1,w)$$

One more point needs to be mentioned is that the weighting retrieved in proposed formulation is not normalized yet. It makes the weighting differs in length and gain among each frame. The result is the output waveform looks blurred in frequency spectrum. To solve this problem, the weighting has to be normalized before multiplying the input.

3.4 Parameter Selection and Tradeoff

In the formulation above, three parameters are to be determined: σ_Q, σ_v and ρ_v. σ_Q control the error covariance of the Process Equation. σ_v control the error covariance of the Measurement Equation. ρ_v control the error proportion between the upper line and lower line of the Measurement Equation.

Process Equation:

$$\mathbf{w}(k+1,w)=\mathbf{w}(k,w)+\mathbf{Q}(k,w), \tag{3.2}$$

Measurement Equation:

$$\begin{bmatrix} 0 \\ s_r(k,w) \end{bmatrix}=\begin{bmatrix} \mathbf{X}^H(k,w) \\ \mathbf{X}_r^H(k,w) \end{bmatrix}\mathbf{w}(k,w)+\mathbf{V}(k,w) \tag{3.3}$$

$$\begin{aligned} &\mathbf{Q}(k,w)\sim N(0,\sigma_Q I) \\ &\mathbf{V}(k,w)\sim N(0,\sigma_v\begin{bmatrix} 1 & 0 \\ 0 & \rho_v \end{bmatrix}) \end{aligned} \tag{3.4}$$

The value $\frac{\sigma_Q}{\sigma_v}$, which is the ratio between σ_Q and σ_v, controls the adaption speed. If $\frac{\sigma_Q}{\sigma_v}$ is large, the filter adapts to the variation in environment faster. By (3.2), it can be observed that if σ_Q is large, the change between $\mathbf{w}(k,w)$ and $\mathbf{w}(k+1,w)$ will be larger and leads to faster adaption in $\mathbf{w}(k,w)$. By (3.3), it can be observed that if σ_v is small, the $\mathbf{V}(k,w)$, or the Measurement Error, has small variations between each step, which means $\mathbf{w}(k,w)$ has to adapt fast if $\begin{bmatrix} \mathbf{X}^H(k,w) \\ \mathbf{X}_r^H(k,w) \end{bmatrix}$ varies fast .

In the case of the environment is a Linearly Time-Invariant (LTI) system, there is no need to do adaption to those variations in the system. Therefore, the best choose of $\frac{\sigma_Q}{\sigma_v}$ will be zero by setting σ_Q to zero.

The parameter ρ_v controls the tradeoff between noise reduction and dereverberation. Large ρ_v leads to strong noise reduction and little dereverberation while small ρ_v leads to strong dereverberation and little noise reduction. If ρ_v is small, that means the error variation in the lower line of (3.4) is relatively small compared with the upper line, which leads to closer tracing in the lower line and looser tracing in the upper line, achieving strong dereverberation and weak noise reduction. If ρ_v is large, that means the error variation in the upper line of (3.4) is relatively small compared with the lower line, which leads to closer tracing in the upper line and looser tracing in the lower line, achieving strong noise reduction and weak dereverberation.

Extreme choose of ρ_v in either cases will decrease the signal quality since too much distortion or too much noise are both degrading reasons to the quality of the signal. The optimal choose of ρ_v should be related to the signal-to-noise ratio (SNR) since ρ_v can be treated as a leverage that distributes the total effort of filtering between signal dereverberation and noise reduction. If the noise level is relatively small to the signal, or the SNR is high, more effort should be emphasized on signal dereverberation while if the noise level is relatively large to the signal, or the SNR is low, more effort should be emphasized on noise reduction. Experiments on this tradeoff will be presented in Section 4.

3.5 Voice Activity Detection under Proposed Formulation

As mentioned before in Section 3.3, the vector $\mathbf{X}(k,w)$ is the data recorded when the desired signal is inactive, or desired signal cancelation phenomenon will occur. Thus, a voice activity detector is required. A feasible option is to incorporate other algorithm that detects voice activity or signal activity. However, some parameters during the filtering procedure can be utilized to implement as voice activity detector. The procedure regarding such implementation will be presented in this section.

Starting again from the formulation:

Measurement Equation:

$$\begin{bmatrix} 0 \\ s_r(k,w) \end{bmatrix} = \begin{bmatrix} \mathbf{X}^H(k,w) \\ \mathbf{X}_r^H(k,w) \end{bmatrix} \mathbf{w}(k,w) + \mathbf{V}(k,w) \qquad (3.2)$$

Process Equation:

$$\mathbf{w}(k+1,w) = \mathbf{w}(k,w) + \mathbf{Q}(k,w). \qquad (3.3)$$

The vector $\mathbf{V}(k,w)$ is the Measurement Error. By observing the value of the Measurement Error, the voice activity detector can be implemented. In (3.2), the upper line can be regarded as suppressing noise while the lower line can be regarded as preserving the desired signal. If $\mathbf{X}(k,w)$ is purely noise, it will be minimized by both the upper line and lower line of (3.2). The Measurement Error with such $\mathbf{X}(k,w)$ is small and has low variance. If $\mathbf{X}(k,w)$ contains desired signal, it is prone to be preserved by the lower line but also prone to be minimized by the upper line, which constitutes a dilemma. The filtering result is that the first element of the Measurement Error, corresponding to the error in the upper line, is large, which means such $\mathbf{X}(k,w)$ cannot be minimized by the upper line and leads to large residual error.

In summary, the Measurement Error of noise reduction is employed as a feature to detect voice activity under this algorithm. It can be considered as a data rejection procedure before filtering [8]. If the Measurement Error is larger than the threshold, the current frame is regarded as voice activity and thus the parameters update is abandoned with respect to current frame. If the Measurement Error is smaller than the threshold, the current frame is regarded as voice inactivity and thus the parameters update is preserved with respect to current frame. The flow chart of the voice activity detection procedure is as Fig 2.

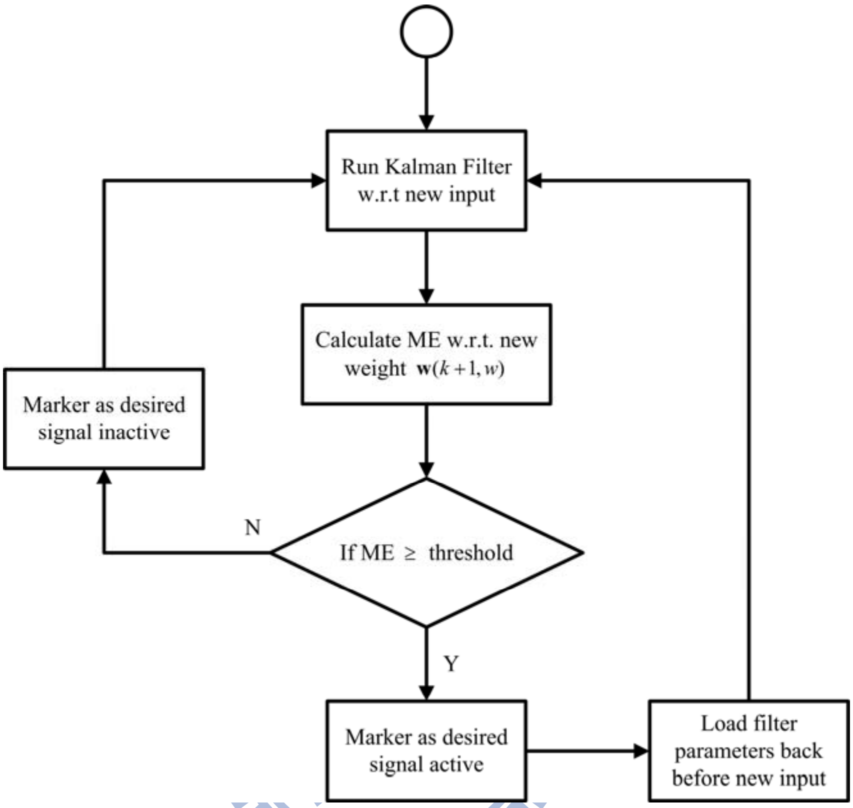

Run Kalman Filter w.r.t new input
Calculate ME w.r.t. new weight $\mathbf{w}(k+1,w)$
Marker as desired signal inactive
N
If ME $\geq$ threshold
Y
Marker as desired signal active
Load filter parameters back before new input

Fig. 2 Flow Chart of Voice Activity Detection Procedure

It has to be noted that the Measurement Error is not discriminative enough if ρ_v is ill-chosen. Since the the critical error term is the Measurement Error on noise reduction, ρ_v should be chosen large enough to spare efforts on noise reduction. However, the best ρ_v should consider both noise reduction and dereverberation, so the appropriate ρ_v should not be chosen extremely large. To overcome such dilemma, two Kalman filters should be executed, one with large ρ_v that executing noise reduction and detecting signal activity while another one with medium ρ_v that computes optimal weight $\mathbf{w}(k,w)$ to achieve best tradeoff between noise reduction and dereverberation.

3.6 Threshold Decision and SINR Estimation

In Section 3.5, the threshold that discriminates the voice activity is not determined. In Section 3.6.1, the procedure that determines the threshold will be presented. In Section 3.6.2, the result of the detection procedure can be further reused to estimated current SINR and help choosing the best ρ_v, which is undetermined in Section 3.4.

3.6.1 Gaussian Mixture Model and EM Algorithm

The Gaussian Mixture Model (GMM) is incorporated to guide the data classification [9]. The distribution of Measurement Error when signal is inactive is modeled as a Gaussian distribution and the distribution of Measurement Error when signal is active is modeled as another Gaussian distribution as Fig. 3. This model is described by the following equations. Let x_k denote the first element of the Measurement Error at time k. z is the speech/nonspeech label, $z \in \{0,1\}$, where 0 denotes nonspeech and 1 for speech. According to Bayes' Rule, it can be written that

$$p(x_k|\lambda) = \sum_z p(x_k, z|\lambda) = \sum_z p(x_k|z,\lambda)p(z), \tag{3.21}$$

where $p(z)$ is the prior probability of speech/nonspeech, and is actually equal to the weight coefficient w_z ($w_0 + w_1 = 1$). $p(x_k|z,\lambda)$ represents the likelihood of x_k given speech/nonspeech model.

$$p(x_k|z,\lambda)=\frac{1}{\sqrt{2\pi\kappa_z}}\exp\{-(x_k-\mu_z)^2/2\kappa_z\} \tag{3.22}$$

where μ_z and κ_z denotes the mean and variance respectively. $\lambda\equiv\{\mu_z,\kappa_z,w_z|z=0,1\}$ is the parameter set of the GMM.

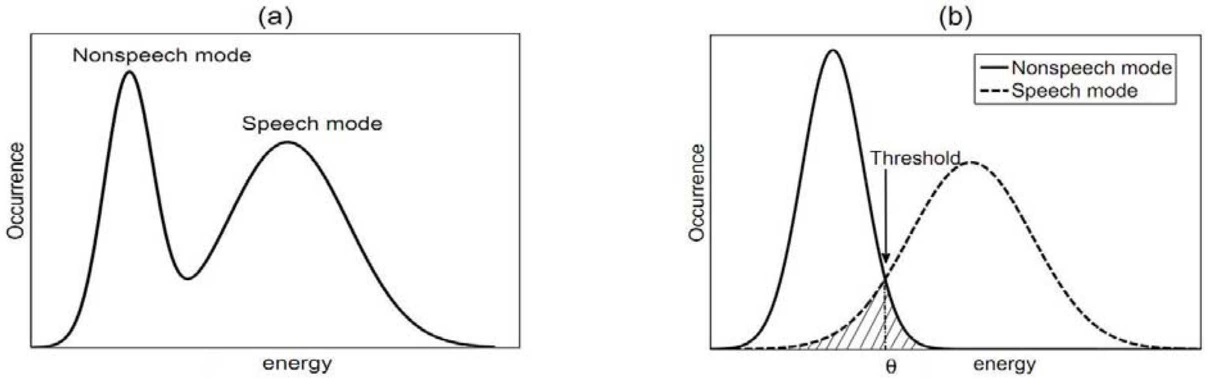


Fig. 3 Schematic illustration of error distribution: (a) Distribution of noisy speech; (b) Distributions of speech and nonspeech (This Figure is modified from [9])

Let $\mathbf{x}\equiv\{x_0,x_1,x_2\cdots x_M\}$ be a sequence of the first element of the Measurement Error. The probability density function (PDF) is given by

$$p(\mathbf{x}|\lambda)=\prod_{k=0}^{M}p(x_k|\lambda) \quad . \tag{3.23}$$

The parameter set λ is estimated by maximizing the above PDF function.

From the GMM, both of the PDFs of speech and nonspeech can be obtained, namely $p(\theta|z=1,\lambda)p(z=1)$ and $p(\theta|z=0,\lambda)p(z=0)$. These two PDFs are shown in Fig. 3(b). From the two PDFs, the optimal threshold θ can be obtained to minimize the classification error. The threshold θ satisfies

$$p(\theta|z=1,\lambda)p(z=1)=p(\theta|z=0,\lambda)p(z=0) \tag{3.24}$$

Eq. (3.24) is a quadratic equation with one unknown θ. The threshold is one of its roots location between the two means, namely $\mu_1>\theta>\mu_0$. The samples with error less than θ are determined as nonspeech, and otherwise as speech. The shadow in Fig.

3(b) denotes the classification error.

The crucial issue of the above model is to estimate the parameter set λ. The estimation consists of an initialization and a sequential updating process. The initial GMM is first established by the EM algorithm, and then incrementally updated with coming data. The parameter set at time k is denoted as $\lambda_k \equiv \{\mu_{k,z}, \kappa_{k,z}, w_{k,z} | z = 0,1\}$. λ_0 is the initial parameter set estimated from the first M samples by EM algorithm. According to [9], the following are the typical EM re-estimation formulas,

$$p(z|x_j,\lambda) = \frac{w_z\, p(x_j|z,\lambda)}{\sum_z w_z\, p(x_j|z,\lambda)} \tag{3.25}$$

$$w_z' = \frac{1}{M}\sum_{j=0}^{M-1} p(z|x_j,\lambda) \tag{3.26}$$

$$\mu_z' = \frac{\sum_{j=0}^{M-1} x_j\, p(z|x_j,\lambda)}{Mw_z'} \tag{3.27}$$

$$\kappa_z' = \frac{\sum_{j=0}^{M-1} (x_j - \mu_z')^2\, p(z|x_j,\lambda)}{Mw_z'} \tag{3.28}$$

, where $\lambda' \sim \{w_z', \mu_z', \kappa_z'\}$ is the new parameter set re-estimated from λ. In the next iteration, λ is replaced by λ'. This iteration continues until EM algorithm converges. The final λ' is the initial parameter set λ_0 required to GMM initialization and the threshold θ can be obtained by solving (3.24).

According to [9], it assumes the GMM varies with time slowly, $\lambda_k \approx \lambda_{k-1}$ at time k. Accordingly, the relationship $\sum_{j=k-K+1}^{k} p(z|x_j,\lambda_k) \approx \sum_{j=k-K+1}^{k} p(z|x_j,\lambda_{k-1})$. The summation is approximated by the zero-order moment, $\sum_{j=k-K+1}^{k} p(z|x_j,\lambda_k) \approx Kw_{k,z}$, where K is a parameter defined by user which determines the adaption speed. Therefore, the adaption formulas can be written as follows,

$$w_{k+1,z} = \alpha w_{k,z} + (1-\alpha)p(z|x_{k+1},\lambda_k) \tag{3.29}$$

$$\mu_{k+1,z} = \frac{\alpha w_{k,z}\mu_{k,z} + (1-\alpha)p(z|x_{k+1},\lambda_k)x_{k+1}}{w_{k+1,z}} \tag{3.30}$$

$$\kappa_{k+1,z} = \frac{\alpha w_{k,z}\kappa_{k,z} + (1-\alpha)p(z|x_{k+1},\lambda_k)(x_{k+1}-\mu_{k+1,z})^2}{w_{k+1,z}}, \tag{3.31}$$

where α stands for forgetting factor. Besides, some constraints are required during the adaption process as follows.

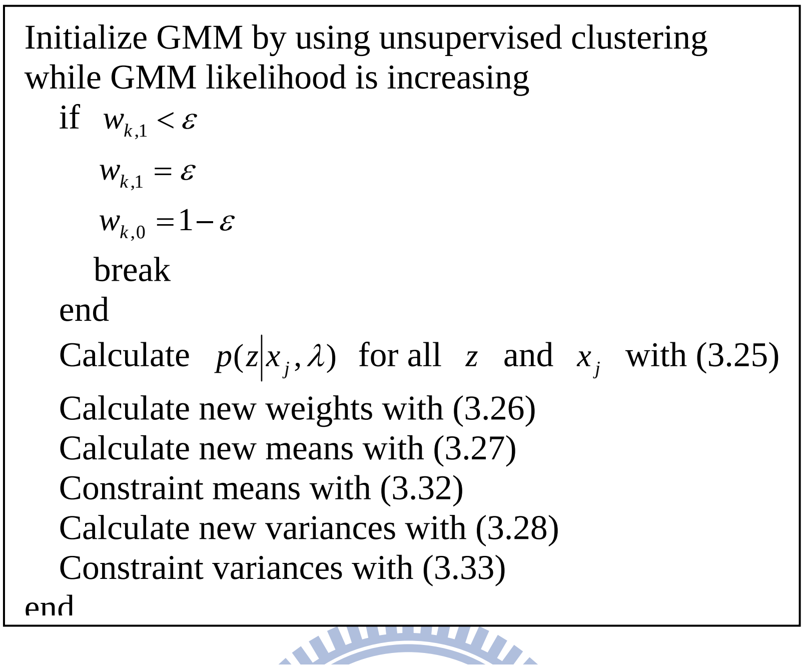
$$\mu_{k,1} = \max\{\mu_{k,1}, \mu_{k,0} + \delta\} \tag{3.32}$$

$$\kappa_{k,1} = \max\{\kappa_{k,0}, \kappa_{k,1}\} \tag{3.33}$$

$$\begin{aligned} w_{k,1} &= \max\{w_{k,1}, \varepsilon\} \\ w_{k,0} &= 1 - w_{k,1} \end{aligned} \tag{3.34}$$

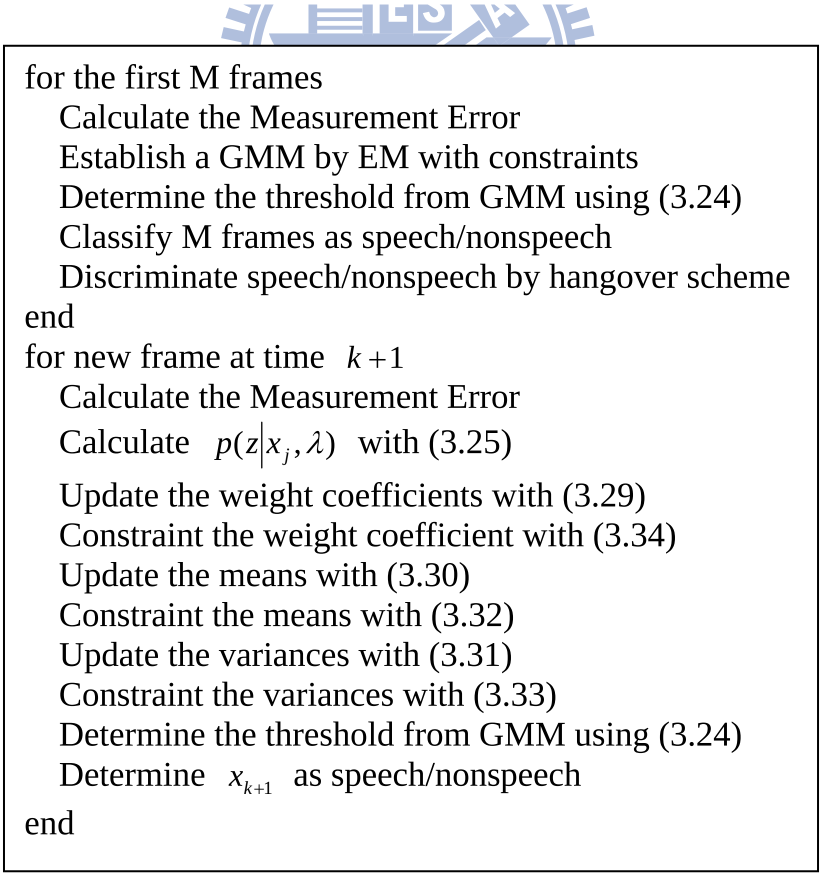
The reason for constraint (3.32) is based on the inspection that the mean of the Measurement Error when speech is always larger than nonspeech, thus a lower bound for $\mu_{k,1}$ is implemented by adding a gap δ to $\mu_{k,0}$ and choose the larger one. The reason for constraint (3.33) is based on the inspection that the variance of the Measurement Error when speech is always not smaller than the the variance of the Measurement Error when nonspeech. The reason for constraint (3.34) is to stem the minimum prior probability of speech from becoming 0 and inducing no adaption afterwards, where ε is also a parameter to be chosen.

After building the GMM model, the threshold θ can be determined after EM initialization and adaption. The process of EM algorithm is written in Fig. 4 and the total procedure of VAD decision is written as Fig. 5.



```
Initialize GMM by using unsupervised clustering
while GMM likelihood is increasing
   if  $w_{k,1} < \varepsilon$
      $w_{k,1} = \varepsilon$
      $w_{k,0} = 1 - \varepsilon$
      break
   end
   Calculate  $p(z|x_j, \lambda)$  for all  $z$  and  $x_j$  with (3.25)
   Calculate new weights with (3.26)
   Calculate new means with (3.27)
   Constraint means with (3.32)
   Calculate new variances with (3.28)
   Constraint variances with (3.33)
end
```

Fig. 4 EM algorithm with constraints (revised from [9])



```
for the first M frames
   Calculate the Measurement Error
   Establish a GMM by EM with constraints
   Determine the threshold from GMM using (3.24)
   Classify M frames as speech/nonspeech
   Discriminate speech/nonspeech by hangover scheme
end
for new frame at time  $k+1$
   Calculate the Measurement Error
   Calculate  $p(z|x_j, \lambda)$  with (3.25)
   Update the weight coefficients with (3.29)
   Constraint the weight coefficient with (3.34)
   Update the means with (3.30)
   Constraint the means with (3.32)
   Update the variances with (3.31)
   Constraint the variances with (3.33)
   Determine the threshold from GMM using (3.24)
   Determine  $x_{k+1}$  as speech/nonspeech
end
```

Fig. 5 The process of VAD decision (revised from [9])

3.6.2 SINR Estimation

In Section 3.4, the best ρ_v that determines the tradeoff between noise reduction and dereverberation is undetermined. It is mentioned that it should be related to the current SINR since ρ_v leverages the effort to reduce noise and enhance signal while SINR stands for the ratio of signal power and noise power. From the result of Section 3.6.1, the two Gaussian Models stand for the Measurement Error of signal part and noise part, which is also can be related to SINR. The mean of the Gaussian Model for signal and noise can be regarded as two indices describing the signal power and noise power after adaptive filtering. Therefore, the mean difference of the two Gaussian Models can be interpreted as an index describing current SINR. Fig. 6 shows the relationship from Mean Difference in VAD to the best estimation of ρ_v.

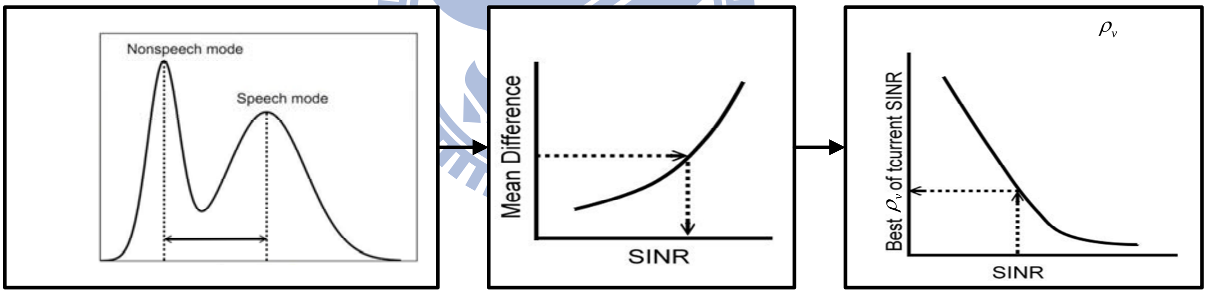


Fig. 6 The relationship from Mean Difference in VAD to the best estimation of ρ_v

In Fig. 6, there are three blocks used to determine the best estimation of ρ_v. The first block is calculating the mean difference from current GMM, which is trivial after building the Gaussian Mixture Models.

The second block is estimating the current SINR by current Mean Difference. Although the conceptual relationship can be imagined, there is still no concrete equation to describe the relationship between them. To solve that problem, the relationship can be pre-trained. The curve, or the relationship, can be found by mixing signal clip and noise clip recorded on testing scenario with various amplitudes to acquire clips with different SINRs. With those clips, the computation of computing Measurement Error with Kalman filter and perfect VAD are preceded. After the computation and building GMM modles, the Mean Difference can be found corresponding to the testing clips. Finally, rearranging the correspondence from SINR to Mean Difference, the relationship can be trained. An example showing the result of a series of training is in Fig. 7. With the relationship from SINR to Mean Difference, it can be used to inversely look up when requiring current SINR given Mean Difference.

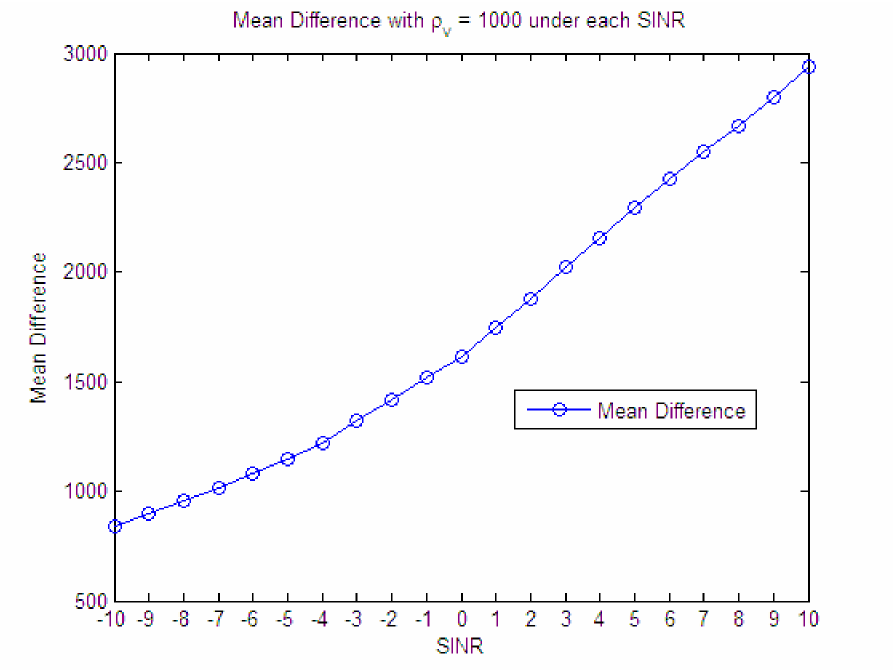


Fig. 7 An example of trained relationship from SINR to Mean Difference

The third block is estimating the best ρ_v corresponding to current SINR. It can also be trained to build the relationship. The pre-training procedure is varying ρ_v from 0.01 to 100 with multiplication of $10^{0.2}$ for each sample clip of different SINR and finding the best output. The "best output" can be measured by some combination of objective indices like output SINR or log spectrum distortion (LSD). An example of giving the best output by minimizing the LSD through various ρ_v and various SINR is presented in Fig. 8. Note that small LSD stands for less distortion and high signal quality.

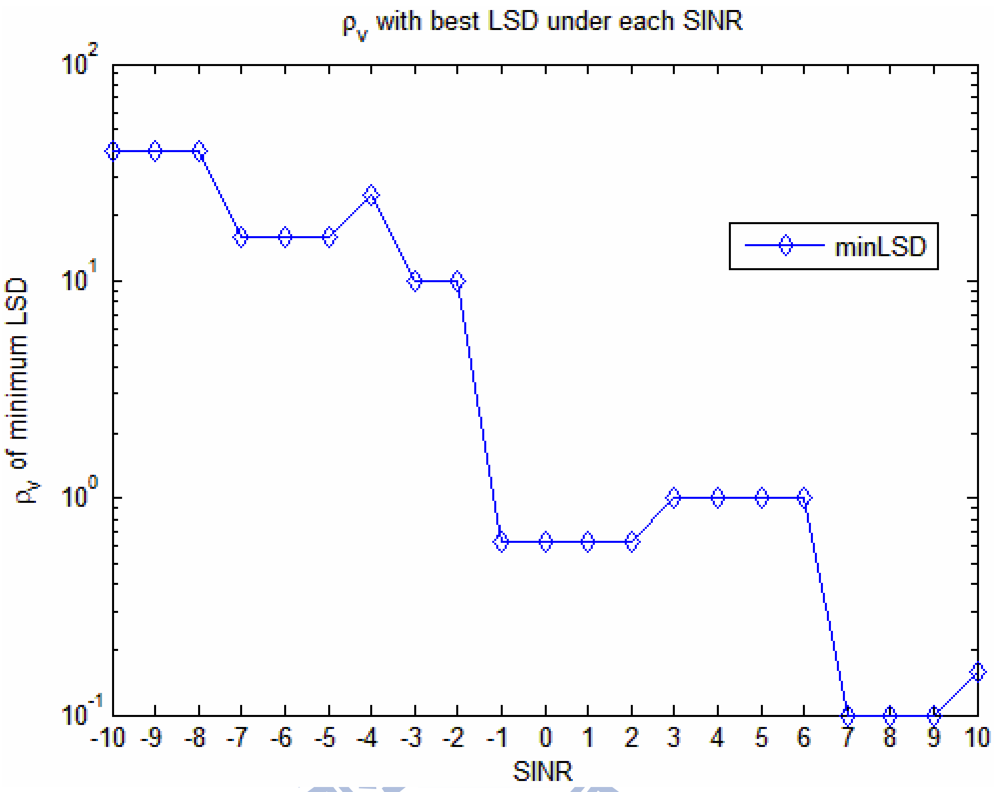


Fig. 8 SINR vs. the ρ_v giving Best LSD

With the Gaussian Mixture Models and the two pre-trained blocks, the best ρ_v under that trained scenario can be founded.

3.7 Overall System Architecture

Combining the beamforming technique proposed in Section 3.3, the voice activity detection in Section 3.5 and the parameter determinism in Section 3.6, the overall system architecture is presented in this section.

The flow chart Fig. 9 is plotted to elaborate the overall system architecture. The main processing can be separated to two Kalman filters, written as Kalman filter 1 and Kalman filter 2 in Fig. 9 The Kalman filter 1 is operated as the voice activity detector, thus its ρ_v should be chosen large enough to place appropriate efforts on noise reduction. By a large ρ_v, the Measurement Error will be discriminative enough to separate the signal part and noise part. The Kalman filter 2 serves as the beamformer, so its ρ_v should be chosen appropriately to balance the tradeoff between noise reduction and dereverberation.

To start with, new speech samples in time domain are collected in frames with fixed overlap to the previous frame and transformed to frequency domain after zero padding and Hanning windowing. Before feeding the new frame to Kalman filter 1, the old parameters of Kalman filter 1 is preserved in case later the Measurement Error shows the Kalman filter 1 should not adapt to the new frame since it contains desired signal. After saving current parameters of Kalman filter 1, the Kalman filter 1 tries to adapt itself to the new frames and calculate the Measurement Error with respect to the new frame. The Measurement Error is compared with the threshold and used to determine the new frame is desired signal active or inactive.

If the new frame is determined as desired signal active, it should be weighted and summed by the weightings given by Kalman filter 2. As mentioned before, the Kalman filter 2 serves as beamformer and filters out undesired noise and maintains desired signal undistorted. After giving filtered result, the parameters of Kalman filter 1 should be loaded by the parameters before adapting to new frame, since the new frame contains desired signal and should not be adapted by Kalman filter 1.

If the new frame is determined as desired signal inactive, it should be fed to Kalman filter 2 to adapt to the noise contained in the new frame. During the adaption phase, the parameters will be meanwhile updated.

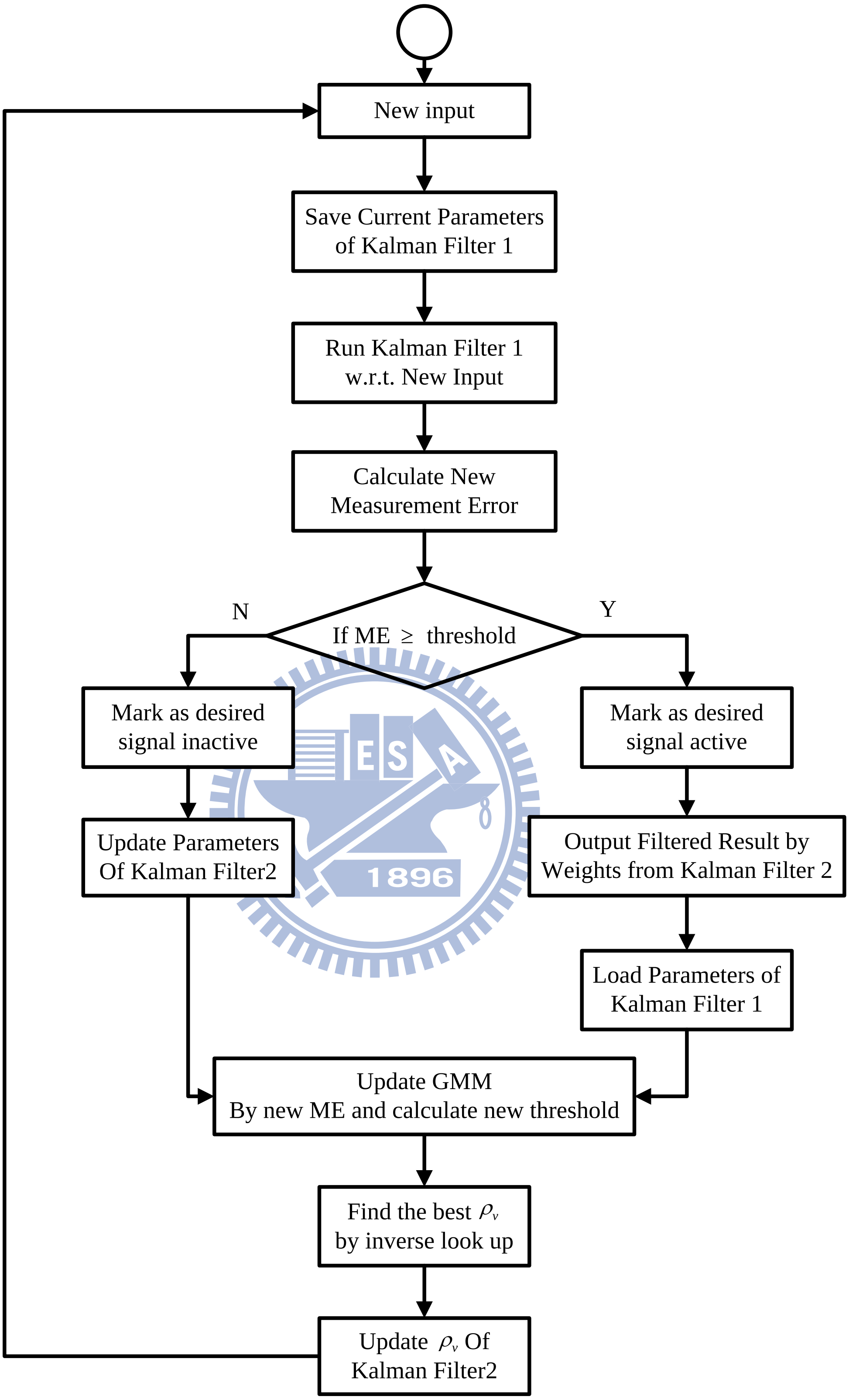

New input
Save Current Parameters of Kalman Filter 1
Run Kalman Filter 1 w.r.t. New Input
Calculate New Measurement Error
If ME ≥ threshold
N
Y
Mark as desired signal inactive
Update Parameters Of Kalman Filter2
Mark as desired signal active
Output Filtered Result by Weights from Kalman Filter 2
Load Parameters of Kalman Filter 1
Update GMM By new ME and calculate new threshold
Find the best ρ_v by inverse look up
Update ρ_v Of Kalman Filter2

Fig. 9 The Flowchart of Overall System

After determining the voice activity, the new Measurement Error is used to update the GMM and calculate for new threshold. The Mean Difference of the two Gaussian Models can be used to look up for current SINR and the best ρ_v for Kalman filter 2.

To sum up with, the overall algorithm contains two Kalman filters to handle the two issues of voice activity detection and beamforming respectively. The two Kalman filters differ in its crucial parameter ρ_v and thus render different functions and scenarios. The GMM is incorporated to help detecting voice activity and separate the signal and noise as two groups, which gives more information to retrieve the best ρ_v corresponding to current SINR.

Chapter 4. EXPERIMENT RESULTS

4.1 Introduction of the Experiment Condition

In the experimental results presented afterward, the original sound samples are recorded in a Ford Fiesta car by a microphone array placed at the sun shield of driver's seat. The desired male speech is played by the Head and Torso Simulator (HATS) by Brüel & Kjær on the driver's seat. The speech data is extracted from a listening comprehension test by an English learning center, thus giving high SNR. The interfering female speech is played by the same HATS on the copilot's seat. It is also extracted from an English listening comprehension test. The noise is recorded when the car is driving on road with speed at around 50 km/hr. More specifications about the experiment are presented in Table 1. The photos illustrating the recording environment are as Fig. 10 and Fig. 11. Fig. 12 and Fig. 13 are the time-frequency plots from the known clips played and the signal clips recorded, both of which are used as reference signal in this experiment.

Microphone Number	4	Microphone Displacement	7 cm
Sampling rate	8000 Hz	FFT size	512 samples
Shift number	160 samples	Zero padding	32 samples

Table 1 Parameters in experiment

The sound data is recorded by a digital microphone array, which uses digital microphones to receive signal and collects 16-bits array data in an Altera FPGA development board. The received data is visible for an embedded network hardware NetBurner through shared memory. Finally, the array data is transferred to PC or Laptop through Local Area Network (LAN).

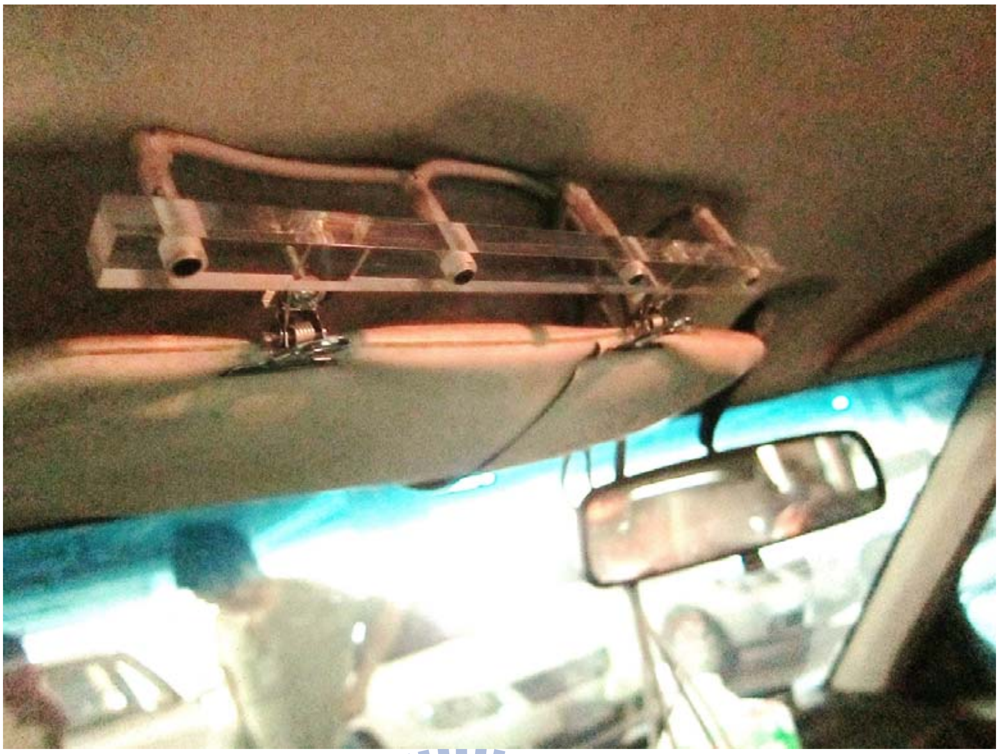


Fig. 10 The photo for the microphone array at the sun shield of the driver's seat.

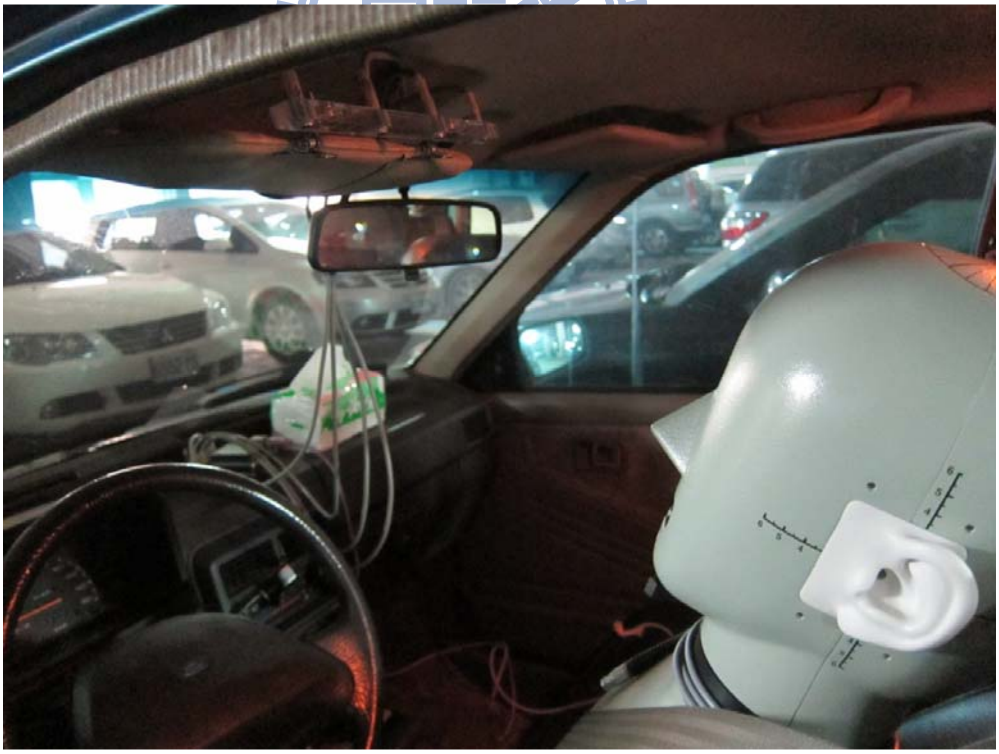


Fig. 11 The photo for the HATS at the driver's seat

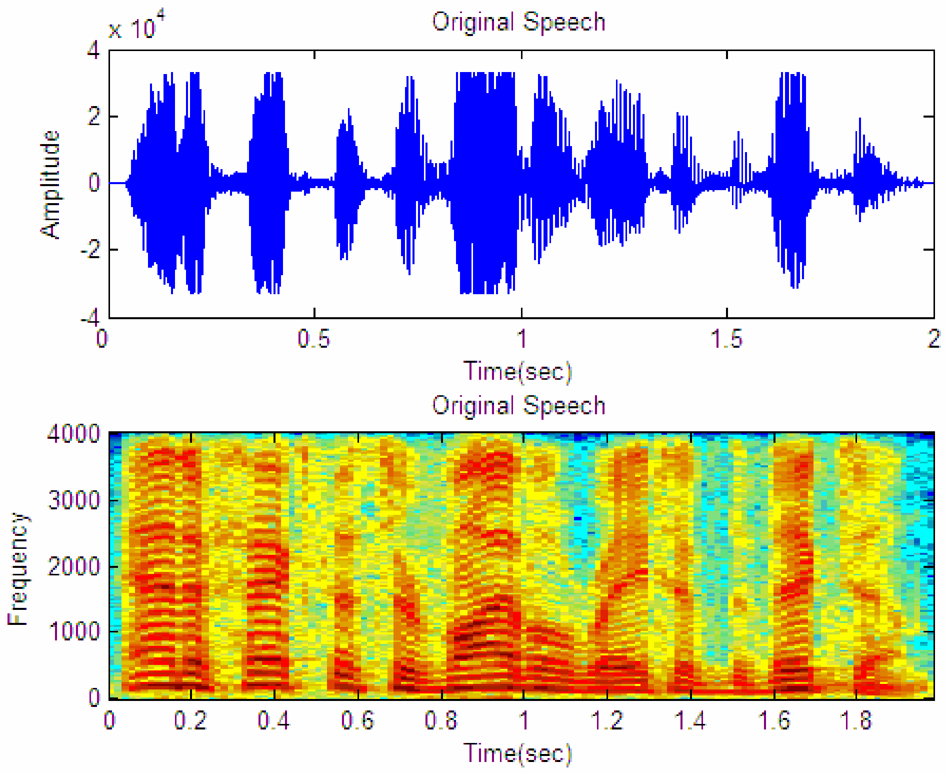


Fig. 12 The time-frequency plot for original speech

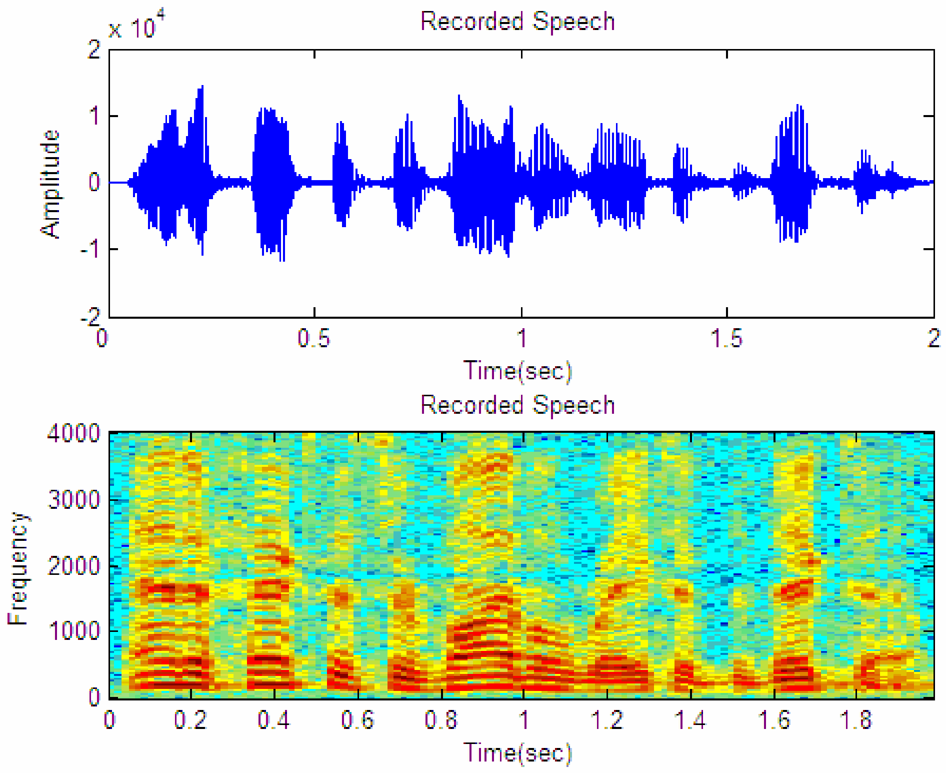


Fig. 13 The time-frequency plot for recorded speech

4.2 Experiments on Performance of Noise Reduction and Its Tradeoff with Dereverberation

In this section, the tradeoff phenomenon between noise reduction and dereverberation is exhibited. The experiment environment is as mentioned in Section 4.1. Three speech enhancement algorithms, MVDR, MVDR with Kalman filter solution, DSB (Delay and Sum Beamformer) are implemented to compare with proposed algorithm. In this section, perfect voice activity detection is assumed for MVDR, MVDR with Kalman filter and proposed algorithm to avoid sample matrix inverse (SMI) problem [10]. For the MVDR filter, the forgetting factor of sample covariance matrix is 0.99. In proposed beamformer, the parameter ρ_v ranges from 0.001 to 1000 with ration of increase 10.

Two objective performance indices are used to measure the waveform property. The first is the average SINR (avgSINR) defined as

$$\text{avgSINR} = \frac{\frac{1}{T_s}\sum_{t\in Ts} x^2(t) - \frac{1}{T_n}\sum_{t\in T_n} x^2(t)}{\frac{1}{T_n}\sum_{t\in T_n} x^2(t)} \tag{4.1}$$

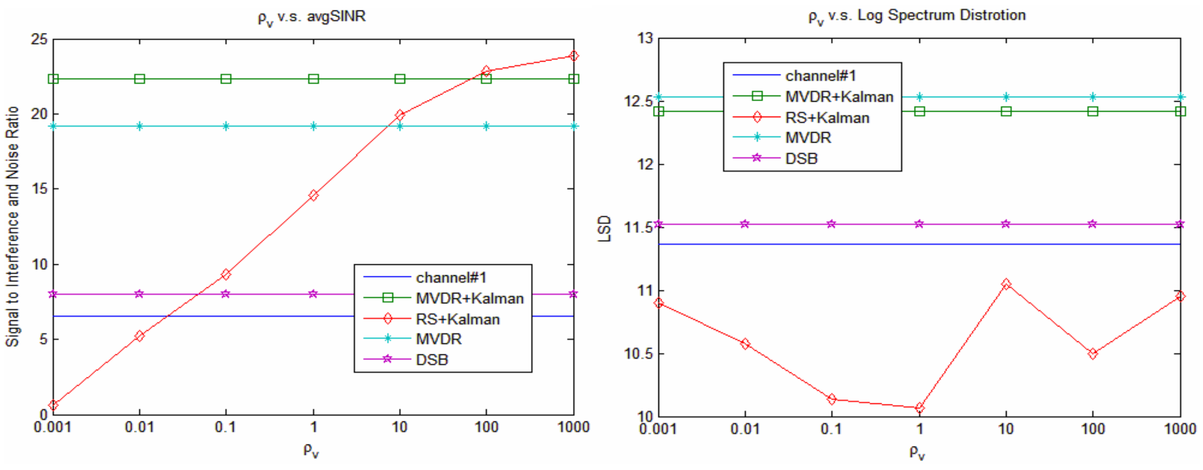
where T_s and T_n denote periods in time when only the desired speech is active and only the interference-plus-noise signals are active respectively. The second quality measure is log spectral distortion (LSD) defined as

$$\text{LSD} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{W}\sum_{w=1}^{W}\left(10\log_{10}\left|s_r(k,w)\right| - 10\log_{10}\left|Y(k,w)\right|\right)^2} \tag{4.2}$$

where $s_r(k,w)$ is the Short-time Fourier transform (STFT) of the original sound played by HATS and $Y(k,w)$ is the STFT of the beamformer output. LSD means the speech distortion in frequency domain. Note that a lower LSD level corresponds to better performance.

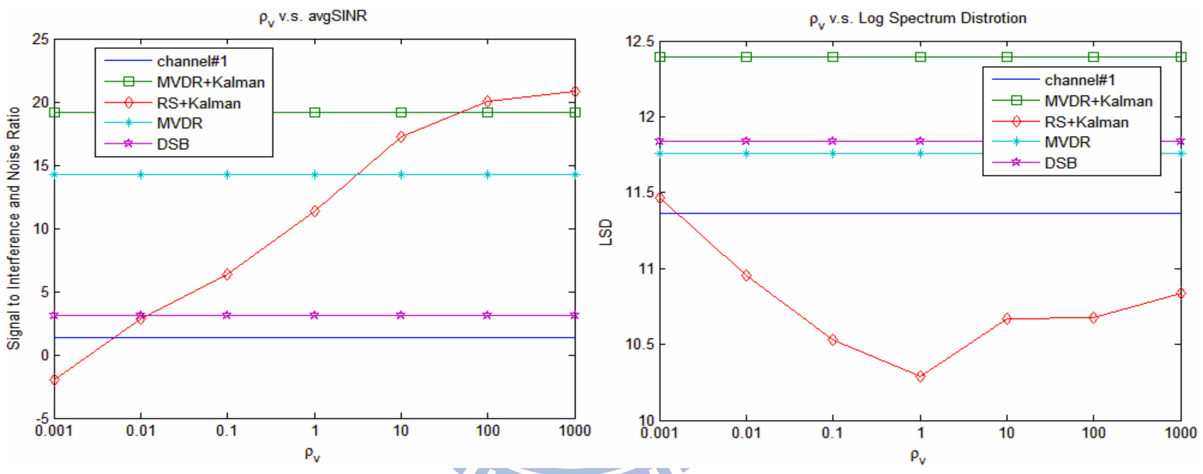
In Fig. 14(a), Fig. 15(a) and Fig. 16(a), the effect of ρ_v regarding SINR is as expected. Higher ρ_v gives higher noise reduction level and thus giving better performance. In contrast, small ρ_v gives low noise reduction level and thus giving bad result in LSD since noise and distortion both worsen the LSD. Since the perfect voice activity detection is assumed, other methods like MVDR and MVDR with Kalman filter both performs well. With perfect voice activity detection, the MVDR works on perfect situation that signal correlation matrix and noise correlation matrix are perfectly identified. In that case, its solution is close to optimal solution for maximizing output SINR. However, in Fig. 14(b), Fig. 15(b), Fig. 16(b) the LSD shows that MVDR and MVDR with Kalman filter are suffered from distortion while proposed algorithm works better if ρ_v is chosen appropriately.

For subjective evaluations, Fig. 17, Fig. 18 and Fig. 19 show the waveforms and spectrograms at different SINR -2 dB, 4 dB and 7 dB. It can be observed that both in MVDR and DSB, the voice pattern in frequency domain is still preserved while proposed method and MVDR with Kalman filter somehow blurred the voice pattern in frequency domain. Regarding the noise reduction, it can be observed that most of the noises are eliminated in proposed method, MVDR and MVDR with Kalman filter.



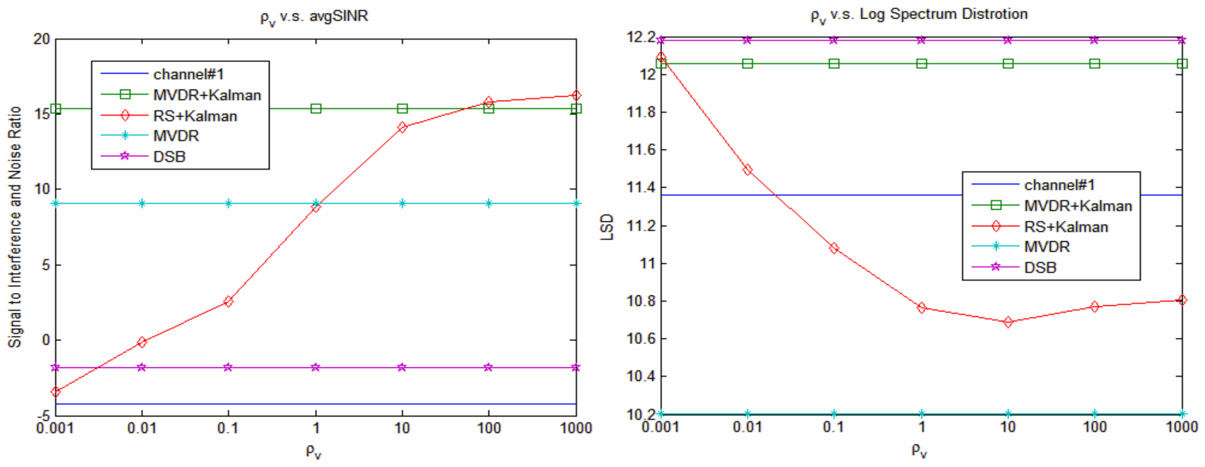
(a) avgSINR result (b) LSD result

Fig. 14 Experiment results in car environment with input SNR 7 dB



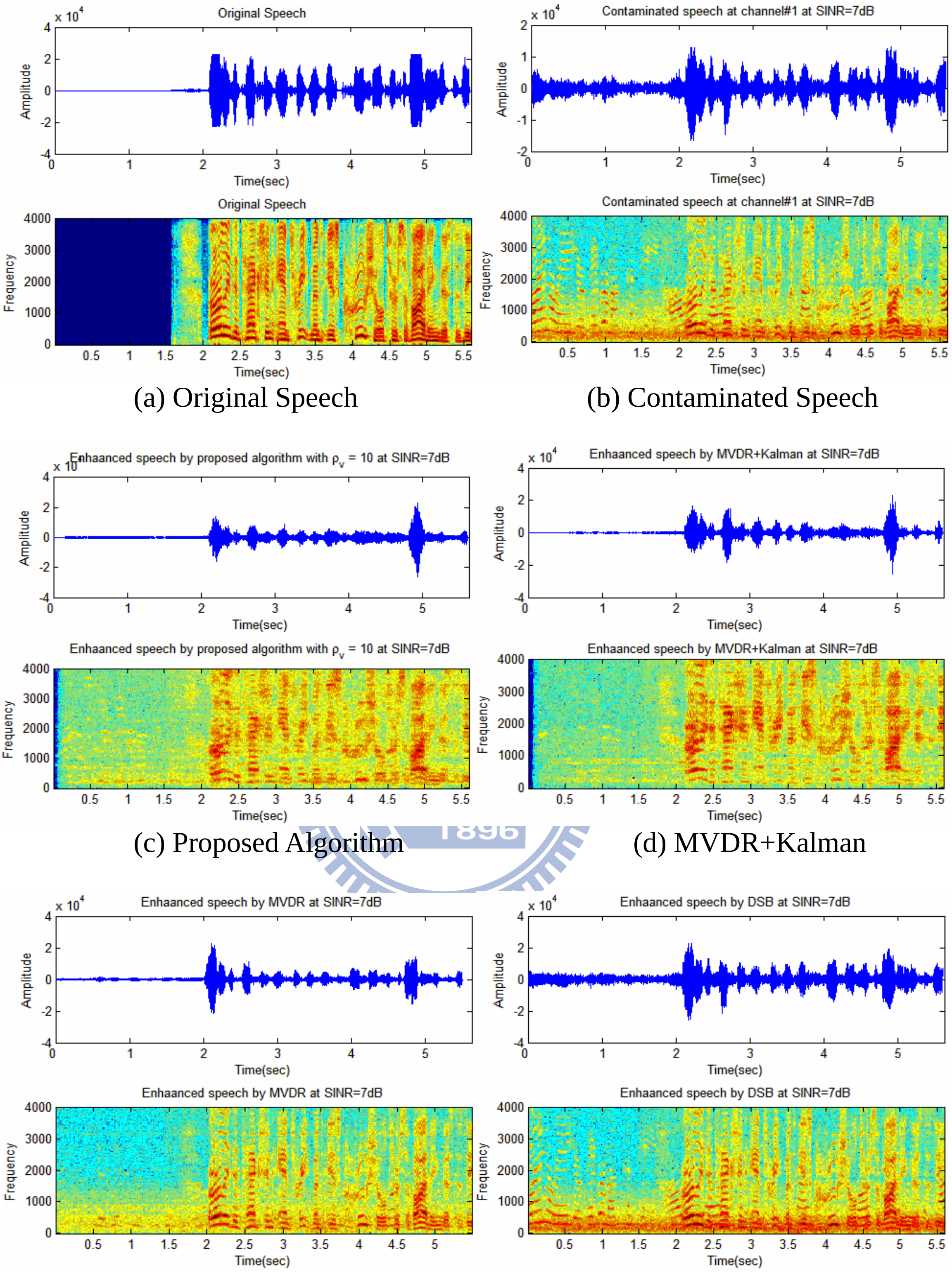
(a) avgSINR result (b) LSD result

Fig. 15 Experiment results in car environment with input SNR 2 dB



(a) avgSINR result (b) LSD result

Fig. 16 Experiment results in car environment with input SNR -4 dB



(a) Original Speech

(b) Contaminated Speech

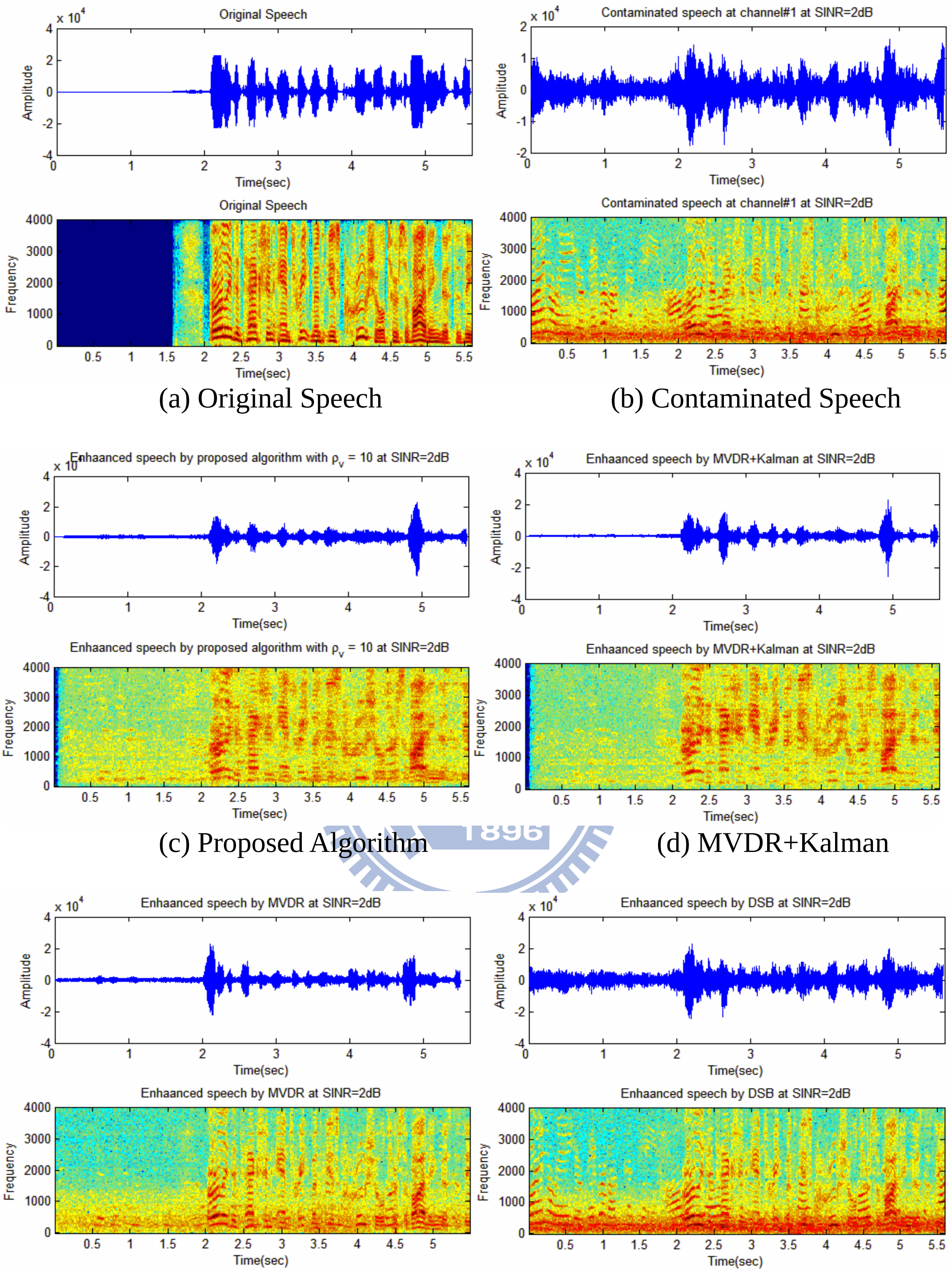
(c) Proposed Algorithm

(d) MVDR+Kalman

(e) MVDR

(f) DSB

Fig. 17 Experiment results in car environment with input SINR 7 dB



(a) Original Speech

(b) Contaminated Speech

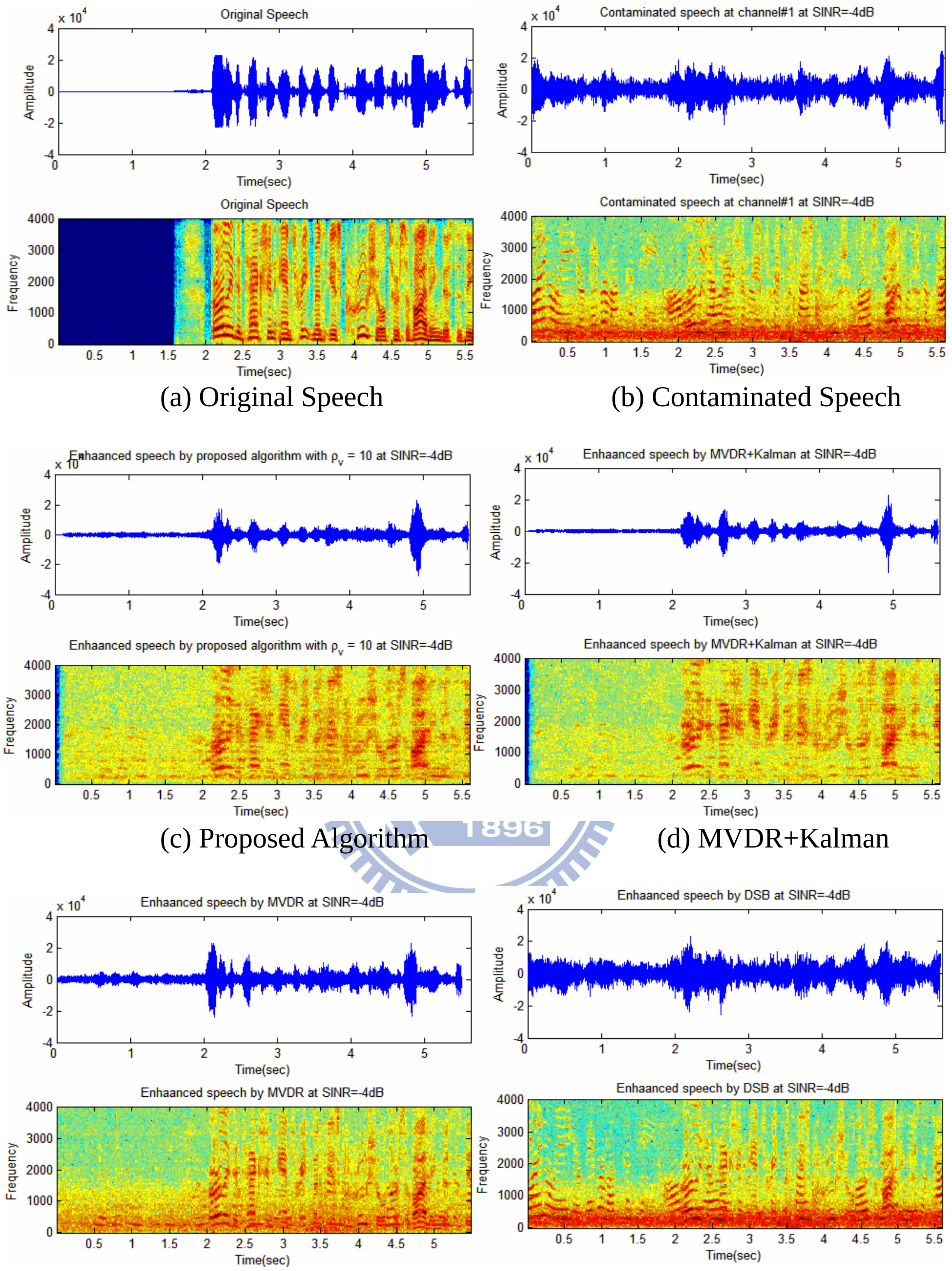
(c) Proposed Algorithm

(d) MVDR+Kalman

(e) MVDR

(f) DSB

Fig. 18 Experiment results in car environment with input SINR 2 dB



(a) Original Speech

(b) Contaminated Speech

(c) Proposed Algorithm

(d) MVDR+Kalman

(e) MVDR

(f) DSB

Fig. 19 Experiment results in car environment with input SINR -4 dB

4.3 Performance on Voice Activity Detection

In this section, experiments regarding voice activity detection are investigated. To compare the detection ability of proposed algorithm, the perfect voice activity detection is made artificially as base line. In addition, the Target-Jammer Ratio (TJR) and energy threshold algorithm are implemented for comparison. The TJR algorithm with target at the front is as follows

$$\mathrm{TJR}(k) = \frac{[\sum_{m=1}^{M} x_m(k)]^2}{[\max(x_1(k)-x_4(k), x_2(k)-x_3(k))]^2} \tag{4.3}$$

, where M is the microphone number and $x_m(k)$ is the data received at m_{th} microphone at snapshot k .The numerator is to form a beam toward target direction and thus representing the intensity of the target. The denominator is to form a null on target and thus retrieves the intensity of the noise. The energy algorithm calculates the energy in all frequency bands. The observation is calculated as

$$Energy(k) = \frac{1}{W}\sum_{w=1}^{W} \left|X(k,w)\right|^2 \tag{4.5}$$

The proposed VAD, TJR VAD and energy VAD are then classified by the GMM model with EM initialization. The forgetting factor α is chosen as 0.999 among all of the three methods. In the proposed VAD, the parameter ρ_v is chosen as 1000 to impose strong noise reduction.

To measure the correctness of the detectors, the coverage with perfect detector is calculated for objective index. The correct rate, false positive rate and false negative rate can be calculated as

$$
\begin{aligned}
P_c &= \frac{1}{K}\sum_{k=1}^{K}[VAD(1|1)+VAD(0|0)],\\
P_{FP} &= \frac{1}{K}\sum_{k=1}^{K}VAD(1|0),\\
P_{FN} &= \frac{1}{K}\sum_{k=1}^{K}VAD(0|1),
\end{aligned}
\tag{4.6}
$$

where $VAD(1|1)$ means given active by perfect VAD the detection is active.

The result of correct rate, false positive rate and false negative rate under SINR from -5~10 are list in Table 2.

In Table 2, it can be observed that proposed method performs better than TJR VAD and Energy VAD, especially under low SINR case. In high SINR case, all the algorithms can extract the voice activity and represent the activation by its feature. However, in low SINR case, it becomes tougher to separate signal and noise since the noise group and signal group are getting closer. In proposed method, high ρ_v places much emphasis on noise reduction, so the Measurement Error is still distinguishable between signal and noise under low SINR case.

The reason to choose α as 0.999, which is large than normal, is to avoid close tracking and over-damping for the threshold.

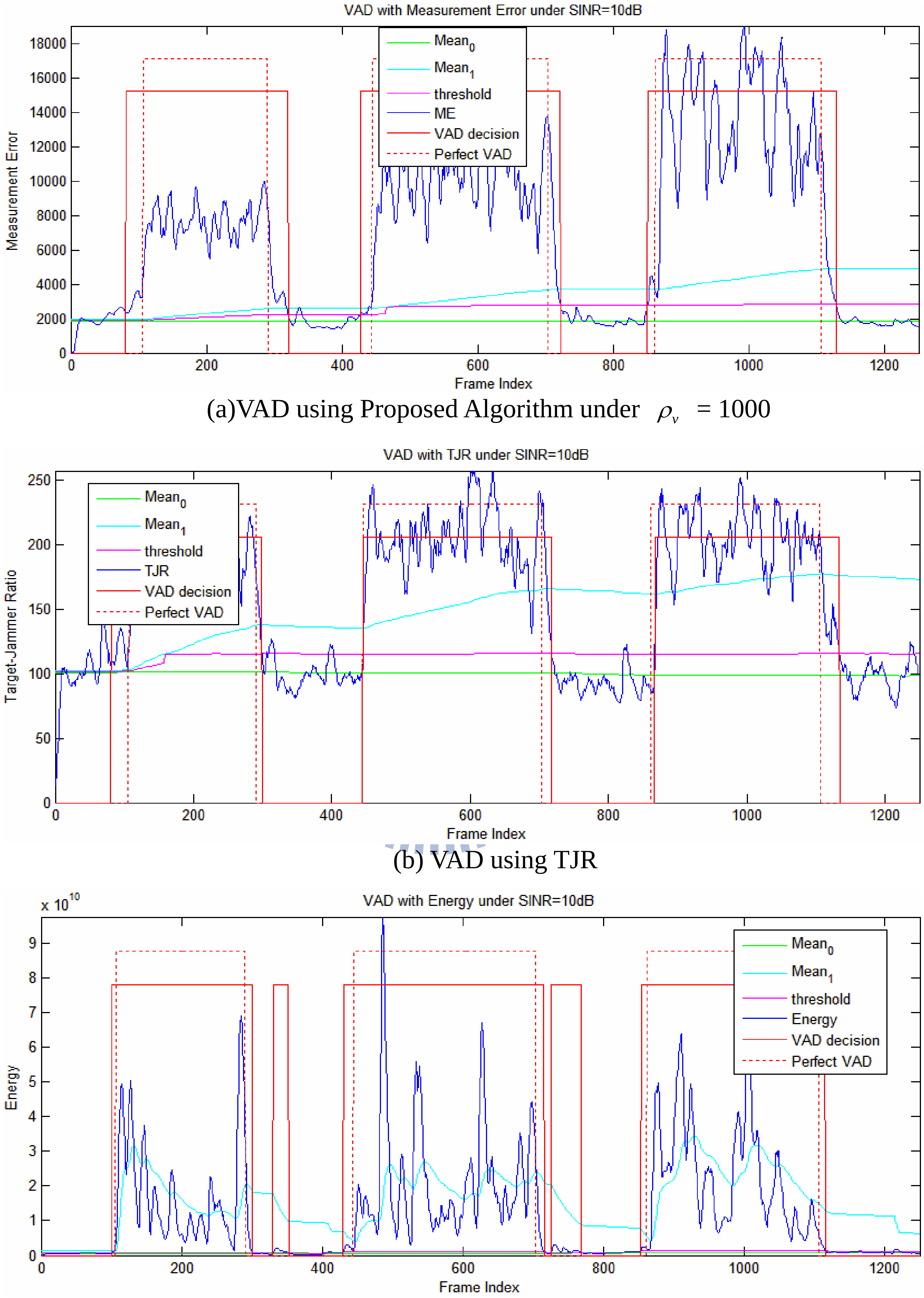
From the amplitude of Fig. 20(a)~Fig. 24(a), it can be observed that the relationship between the Mean Difference and SINR is proportion. Such phenomenon exists throughout the figures from other algorithms. Actually, in situations of VAD using energy, the Mean Difference from energy after taking logarithm direct maps to SINR under each frequency band.

In Fig. 20~Fig. 24, the Mean$_0$ and Mean$_1$ represents for the mean value of the Gaussian Model of noise and speech respectively.

It has to be mentioned that proposed algorithm is prone to crash down thoroughly under very low SINR like -15 dB. The reason is from the looping architecture of proposed algorithm. If a frame is detected as noise frame but actually signal, i.e. false negative case, the Kalman filter will adapt to minimize the signal term, which makes next signal frame more likely to be treated as noise frame and never come back. To get rid of such disaster, the forgetting factor should be chosen large to avoid close tracking.

SINR=10 dB			
	Correct(%)	False Positive(%)	False Negative(%)
Proposed	90.08	9.92	0
TJR	93.36	6.16	0.48
Energy	89.92	10.08	0
SNR= 5dB			
Proposed	90.24	9.36	0.40
TJR	93.52	5.92	0.56
Energy	87.68	12.32	0
SNR= 0dB			
Proposed	90.88	8.64	0.48
TJR	93.60	5.76	0.64
Energy	68.80	31.20	0
SNR= -5dB			
Proposed	95.28	4.24	0.48
TJR	92.88	1.92	5.20
Energy	67.28	30.80	1.92
SNR= -10dB			
Proposed	95.20	3.76	1.04
TJR	78.00	2.40	19.60
Energy	53.28	14.00	32.72

Table 2 Results for voice activity detection under various SINR

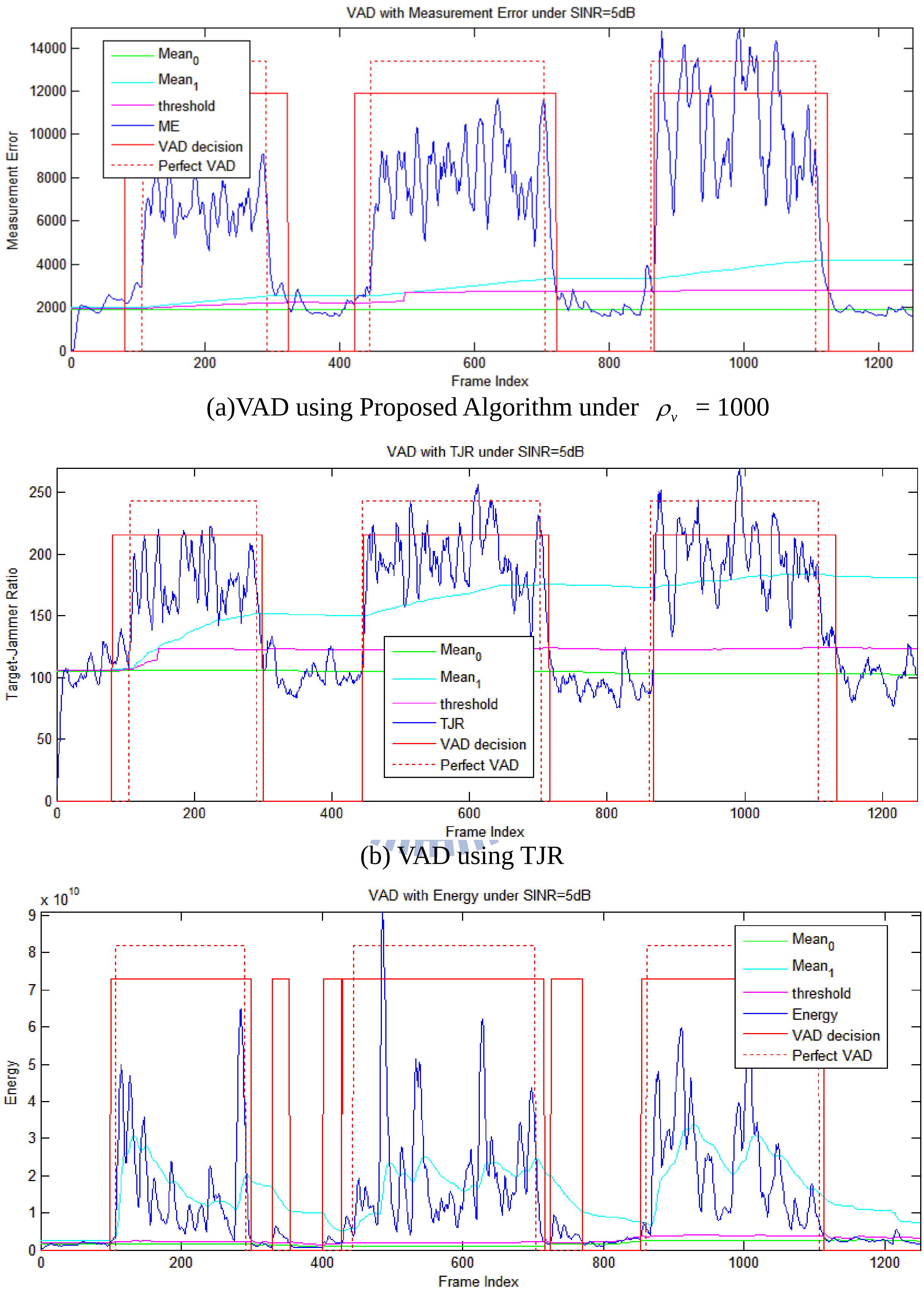


(a)VAD using Proposed Algorithm under ρ_v = 1000

(b) VAD using TJR

(c) VAD using Energy

Fig. 20 Voice Activity Detection under SINR = 10 dB

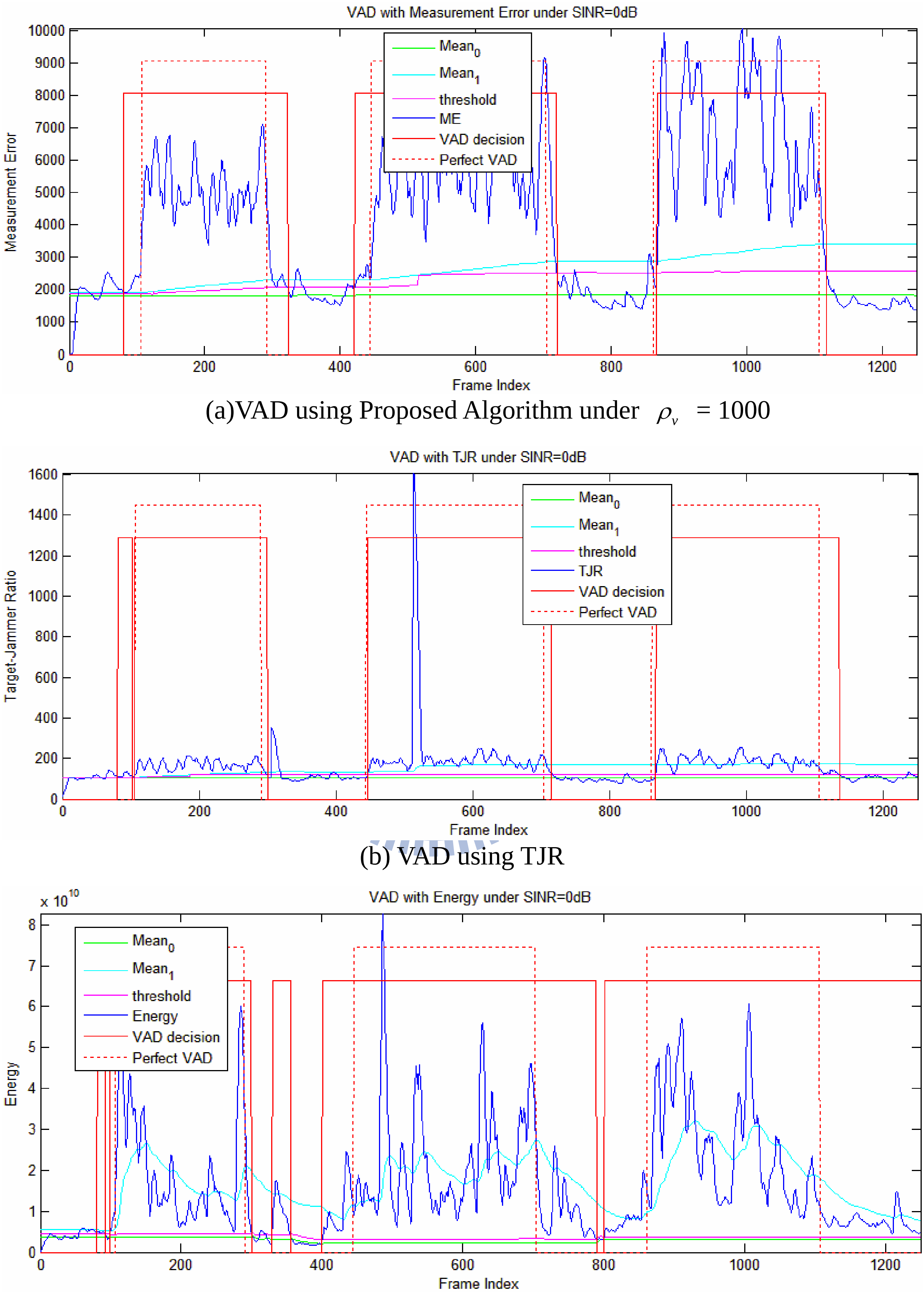


(a)VAD using Proposed Algorithm under ρ_v = 1000

(b) VAD using TJR

(c) VAD using Energy

Fig. 21 Voice Activity Detection under SINR = 5 dB

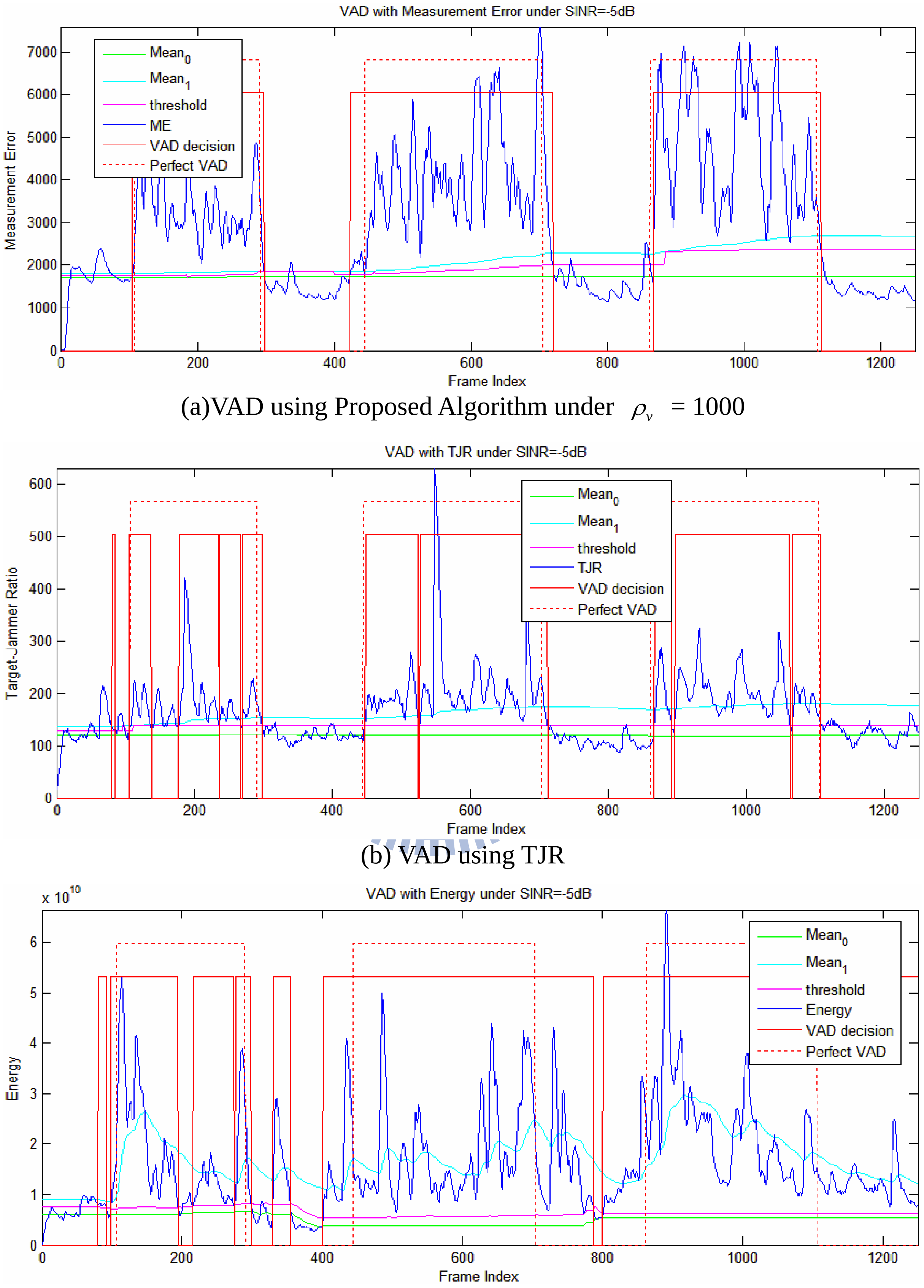


(a)VAD using Proposed Algorithm under ρ_v = 1000

(b) VAD using TJR

(c) VAD using Energy

Fig. 22 Voice Activity Detection under SINR = 0 dB

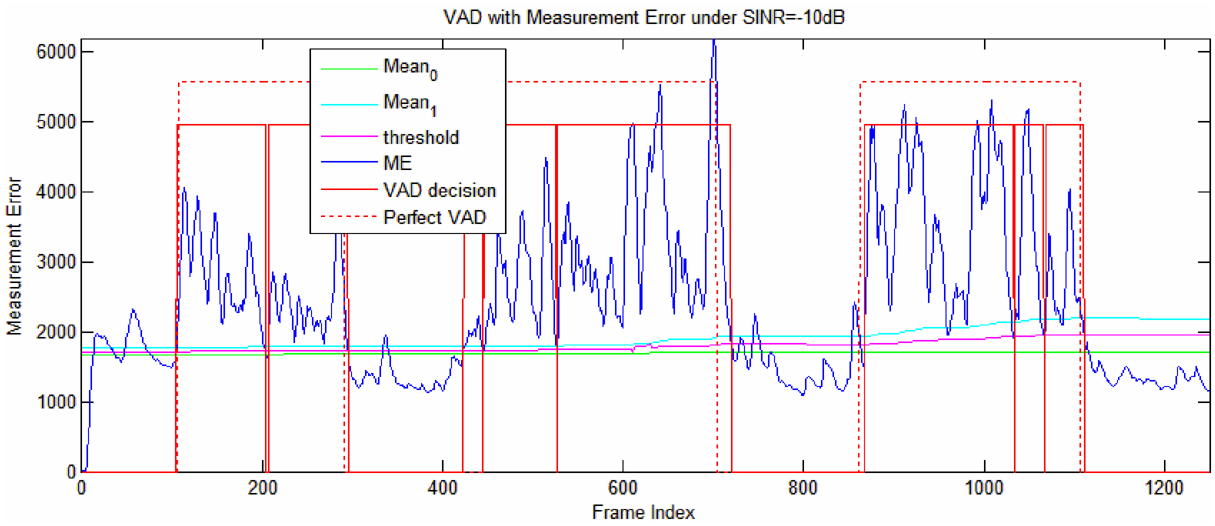


(a)VAD using Proposed Algorithm under $\rho_v = 1000$

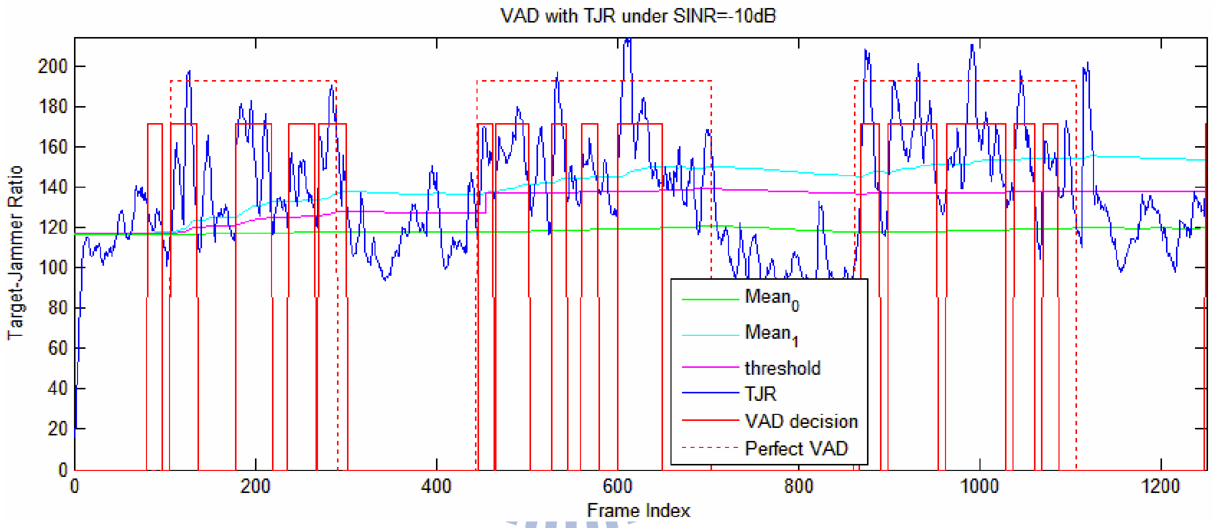
(b) VAD using TJR

(c) VAD using Energy

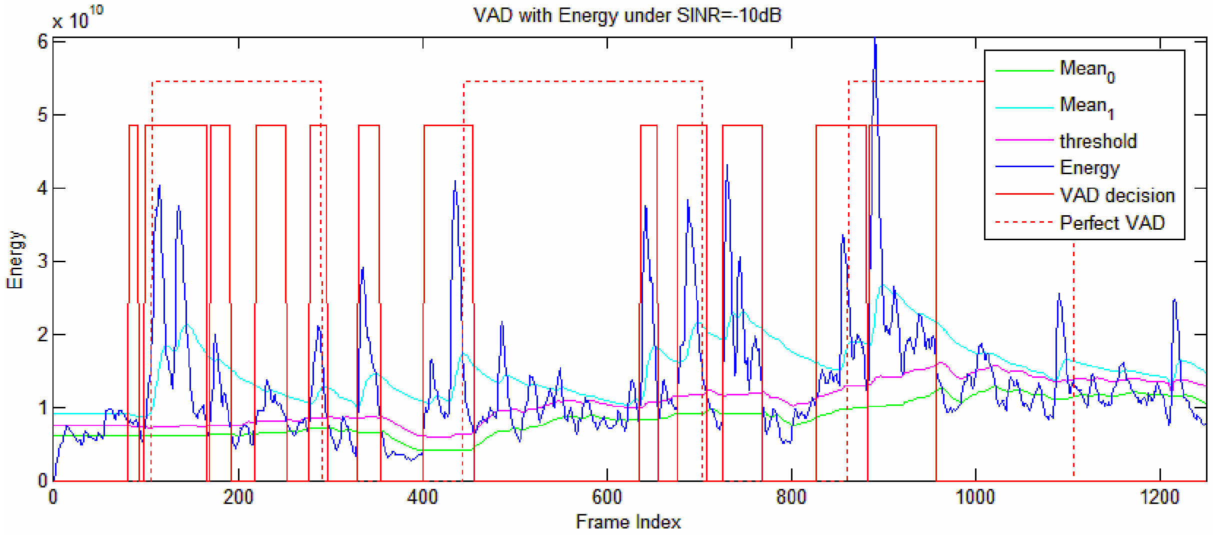
Fig. 23 Voice Activity Detection under SINR = -5 dB



(a)VAD using Proposed Algorithm under ρ_v = 1000



(b) VAD using TJR



(b) VAD using TJR

Fig. 24 Voice Activity Detection under SINR = -10 dB

Chapter 5. CONCLUSION AND FUTURE STUDY

The experiment results presented in Chapter 4 show that the algorithm is capable of processing noise reduction and dereverberation issues at the same time and performs better than general beamformers as in Fig. 14, Fig. 15 and Fig. 16. The capability of utilizing proposed algorithm to implement VAD is experimented in and Fig. 20~Fig. 24. The correction rate shows it functions well under high SINR with correction rate 93.27%. It works better than other VAD algorithms under low SINR case with correction rate 96.11%.

SNR= -10dB			
Proposed	95.20	3.76	1.04
TJR	78.00	2.40	19.60
Energy	53.28	14.00	32.72

For choosing the best ρ_v, the strategy of train and look-up is used. However, the relationship requires more complicated training to ensure theresult is robust enough. Also, the intensity of how the reference signal contributes to maintaining the response distortionless seems crucial to the relationship, since the intensity will effect the decision of ρ_v.

For the VAD, the robustness under very low SNR case is not strong enough. In low SNR case, once a few signal parts are treated as noise, the filter will take signal as noise and proceed on minimization. As a result, the desired signal cancellation phenomenon will occur and destroy the whole algorithm, which is totally undesired. To solve this problem, a more intelligent grouping method should be helpful to overcome such situation.

REFERENCE

[1] J. Capon, "High resolution frequency-wavenumber spectrum analysis," Proc. IEEE, vol. 57, no. 8, pp. 1408-1418, August. 1969.

[2] J. Benesty, J. Chen, and Y. Huang, Microphone Array Signal Processing, Springer-Verlag, Berlin Germany, 2008.

[3] M. Brandstein , Darren B. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Springer-Verlag, Berlin Germany, 2001.

[4] E. Habets, J. Benesty, I. Cohen, S. Gannot, J. Dmochowski, "New Insights into the MVDR Beamformer in Room Acoustics", IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 1, pp. 158–170, Jan. 2010

[5] Y. H. Chen, C. T. Chiang, "Adaptive beamforming using the constrained Kalman filter," IEEE Trans. Antennas Propag., vol. 41, no. 11, pp. 1576–1580, Nov. 1993.

[6] S. Haykin, Adaptive Filter Theory, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[7] S. Gannot and A. Yeredor, "The Kalman filter," in Springer Handbook of Speech Processing, J. Benesty, M. M. Sondhi, and Y. Huang, eds., Springer-Verlag, Berlin, 2007.

[8] Mohinder S. Grewal, Angus P. Andrews, Kalman filtering: Theory and Practice Using Matlab®, Third Edition, Section 8.3, John Wiley & Sons Inc., Hoboken NJ, 2008.

[9] D. Ying, Y. Yan, J. Dang, F. Soong,"Voice Activity Detection Based On An Unsupervised Learning Framework", Volume: PP, Issue: 99, IEEE Transactions on Audio, Speech, and Language Processing, 2011.

[10] S.A. Vorobyov, A.B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization", IEEE Trans. Signal Processing, vol. 51, pp.313 - 324, 2003.

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

[12] 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.

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

[14] J.W. Kim and C.K. Un, "An adaptive array robust to beam pointing error," IEEE Trans. Signal Processing, vol. 40, no. 6, pp. 1582-1584, Jun. 1992.

[15] N.K. Jablon, "Adaptive beamforming with the generalized sidelobe canceller in the presence of array imperfections," IEEE Trans. Antennas Propag., vol. AP-34, no. 8, pp. 996-1012, Aug. 1986.

[16] A.B. Gershman, Y. Hua and Q. Cheng, Eds., "Robustness issues in adaptive beamforming and high-resolution direction finding, in High-Resolution and Robust Signal Processing," pp. 63-110, Marcel Drekker, New York, 2003.

[17] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," Proc. IEEE, vol. 60, no. 8, pp. 926-935, Aug. 1972.

[18] 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.

[19] J.S. Hu and C.H. Yang, "Second-Order Extended H_{infinity} Filter for Nonlinear Discrete-Time Systems Using Quadratic Error Matrix Approximation," IEEE Transactions on Signal Processing, vol. 59, pp. 3110 – 3119, 2011.

[20] S. Nordholm, I. Claesson, M. Dahl, "Adaptive microphone array employing calibration signals: an analytical evaluation," IEEE Transactions on Speech and Audio Processing, vol. 7, pp. 241-245, May 1999.