# 國立交通大學

機械工程學系

碩士論文

# Single-channel noise reduction algorithms for speech enhancement

研究生：陳俊宏

指導教授：白明憲

中華民國九十九年六月

# Single-channel noise reduction algorithms for speech enhancement

研究生：陳俊宏　　Student：Chun-hung Chen

指導教授：白明憲　　Advisor：Ming-Sian Bai

國立交通大學

機械工程學系

碩士論文

A thesis
Submitted to Department of Mechanical Engineering
Collage of Engineering
National Chiao Tung University
In Partial Fulfillment of Requirements
for the Degree of Master of Science
in
Mechanical Engineering
June 2010
HsinChu, Taiwan, Republic of China

中華民國九十九年六月

# Single-channel noise reduction algorithms for speech enhancement

**Student：Chun-hung Chen** **Advisor：Ming-Sian Bai**

**Department of Mechanical Engineering**

**National Chiao-Tung University**

## ABSTRACT

This paper will propose an optimized speech enhancement algorithm aimed at single-channel noise reduction (NR) ,and apply the NR algorithm in the speech recognition. The optimization process is based on an objective function obtained in a regression model and the simulated annealing (SA) algorithm that is well suited for problems with many local optima. The NR algorithm, minimum mean-square error noise reduction (MMSE-NR) algorithm, employs a time-recursive averaging (TRA) method for noise estimation. Objective tests were undertaken to compare the optimized MMSE-TRA-NR and MMSE-VAD-TRA-NR algorithm with several conventional NR algorithms. White noise and car noise at signal-to-noise ratio (SNR) 5 dB are used in these tests. As compared to conventional algorithms, the optimized MMSE-TRA-NR and MMSE-VAD-TRA-NR algorithm proved effective in enhancing noise-corrupted speech signals, without compromising the timbral quality. The optimized MMSE-TRA-NR algorithm also can be used in automatic speech recognition (ASR), the recognition rate will be enhance by the optimal parameters of the MMSE-TRA-NR algorithms.

# 誌謝

兩年的研究生生涯實在太短，時間轉眼即逝。在此先感謝白明憲教授的諄諄教誨與照顧，在白明憲教授的指導期間，深刻的感受到教授對於追求學問的熱忱，更是佩服教授淵博的學問與解決問題的方法。在教授豐富的專業知識以及嚴謹的治學態度下，使我能夠順利完成學業與論文，在此致上最誠摯的謝意。

在論文寫作方面，感謝本校電信工程系冀泰石教授和電子工程系桑梓賢教授在百忙中撥冗閱讀，並提出寶貴的意見與指導，使得本文的內容更趨完善與充實，在此學生致上無限的感激。

在這兩年的研究生生涯中，承蒙博士班林家鴻學長、陳勁誠學長、劉志傑學長、李雨容學姐，以及已畢業的謝秉儒學長、李志中學長、洪志仁學長、何克男學長、艾學安學長、劉冠良學長、王俊仁學長、郭育志學長在研究與學業上的適時指點，並有幸與劉嬰婷、張濬閣、廖國志、桂振益、廖士涵及曾智文同學互相切磋討論，讓我獲益甚多。此外學弟徐偉智、馬瑞彬、王俊凱、吳俊慶、衛帝安、許書豪在生活上的朝夕相處與砥礪磨練，亦值得細細回憶。因為有了你們，讓實驗室裡總是充滿歡笑。能順利取得碩士學位，要感謝的人很多，上述名單恐有疏漏，在此一併致上我最深的謝意。

最後僅以此篇論文，獻給我摯愛的家人，父親陳明輝先生、母親張秀英女士、哥哥陳俊成，這一路上，因為有你們的付出與支持，給了我最大的精神支柱，也讓我有勇氣面對更艱難的挑戰。

# TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# 1. INTRODUCTION

In recent years, applications of mobile communication, video conferencing, peer-to-peer internet telephony networks such as SKYPE®, hands-free car-kits, etc., are rapidly advancing in modern daily life. In these applications, effective communication in noisy environments is one of the pressing issues. Noise reduction (NR) technology has long been research interest in communication industry. How to achieve high reduction with impairing speech quality has been an imminent issue for NR algorithm design.

NR algorithms mostly can be divided into three primary classes:

(1) Spectral-subtraction algorithms[1-6]: The algorithms subtract an estimate of noise spectrum from the noisy speech spectrum. The noise spectrum can be estimated and updated during periods when the signal is absent. Therefore, an estimate of clean signal can be obtained.

(2) Statistical-model-based algorithms: The speech enhancement problem is posed in a statistical estimation framework. A linear (or nonlinear) estimator of the Fourier transform coefficients of the clean signal can be found if a set of the transform coefficients of the noisy signal are given. The Wiener algorithm[7-10] and minimum mean-square error (MMSE) [1], [11] algorithms fall in this category.

(3) Subspace algorithm: The subspace algorithms are based on linear algebra theory. The clean signal might be confined to a subspace of the noisy Euclidean space. Consequently, a method of decomposing the vector space of the noisy signal into "signal subspace" and "noise subspace" is given. The decomposition can be done using well-known orthogonal matrix factorization techniques from linear algebra and the singular value decomposition (SVD) or eigenvector-eigenvalue factorization. The Karhunen-Loéve transform (KLT) NR algorithm [11]-[12]

falls in this category.

These NR algorithms need to estimate the noise spectrum or noise covariance matrix. In fact, the residual noise will be audible and annoying if the noise estimate is too low. On the contrary, speech will be distortion if the noise estimate is too high. Various kinds of NR algorithms have been developed in terms of noise estimation such as voice activity detection (VAD), minimal tracking, time-recursive averaging (TRA), and histogram-based algorithms. A more detailed literature review can be found in a monograph on speech enhancement by Loizou [11].

In the paper, an improved MMSE NR algorithm based on VAD and TRA [11], [13] noise estimation, or abbreviated as MMSE-VAD-TRA-NR, is proposed. An optimization method employs a simulated annealing (SA) [14]-[16] to efficiently search the optimal parameters in the MMSE-VAD-TRA-NR algorithms. The SA method mainly finds the maximum of an objective function. The objective function is constructed by objective measures of the reduction performance and the incurred distortion of processed speech signals.

In order to assess those NR algorithms, objective tests and subjective listening tests were carried out. Those algorithms are simulated at the sampling rate of 8 kHz. The objective tests were employed according to the ITU-T standard P.862 [18]. The subjective listening tests were conducted according to the ITU-T standard P.835 [19]. The test data were processed by using analysis of variance (ANOVA) to justify the statistic significance of difference among algorithms. A post-hoc test, Tukey's HSD, was conducted to assess the significant differences between NR algorithms.

Besides the MMSE-VAD-TRA-NR algorithm, using the MMSE-VAD-TRA-NR algorithm to enhance the recognition rate is also proposed in the paper. A series of parameters in the NR algorithm will be optimal individually by different methods in order to improve the recognition rate.

## 2. NOISE REDUCTION ALGORITHMS

Figure 1 illustrates the general structure of NR algorithms. The procedures of NR algorithms commonly are that the noisy signals would be processed by some forward and inverse transform operations (e.g., the Fourier transform, the discrete cosine transform (DCT) or the KLT transform). Between the forward and inverse transforms, the major NR processes have been accomplished. In this section, a number of algorithms that generally have been proposed in literature for noise reduction (NR) are briefly reviewed.

### 2.1 Spectral subtraction method

Spectral subtraction is a widely used NR method whose original idea is based on the basic principle that as the noise $v(n)$ is additive, the spectral subtraction algorithm can subtract the noise spectrum $V(\omega)$ from the measurement signal $y(n)$. The noise spectrum can be estimated and updated during periods when the speech signal $s(n)$ is not present. The noisy signal $y(n)$ can be expressed as:

$$y(n) = s(n) + v(n)\,. \tag{1}$$

The estimate of clean speech power spectrum $\left|\hat{S}(\omega)\right|^2$ can be obtained as follows:

$$\left|\hat{S}(\omega)\right|^2 = H^2(\omega)\left|Y(\omega)\right|^2\,, \tag{2}$$

where $\left|Y(\omega)\right|^2$ is the noisy speech power spectrum, and $H(\omega)$ is known as the system's transfer function. The symbol "^" is used to indicate the estimated parameters of interest. From the subtraction rule, $H(\omega)$ can be given by

$$H(\omega) = \sqrt{1 - \frac{\left|\hat{V}(\omega)\right|^2}{\left|Y(\omega)\right|^2}}\,, \tag{3}$$

where $\left|\hat{V}(\omega)\right|^2$ is estimate of noise power spectrum, $H(\omega)$ can be considered to be a gain function in NR algorithm, and its value is always positive in the range of

$0 \le H(\omega) \le 1$. A general version of the spectral subtraction is expressed as follows:

$$\left|\hat{S}(\omega)\right|^p = \left|Y(\omega)\right|^p - \left|\hat{V}(\omega)\right|^p \tag{4}$$

where $p$ is the power exponent, with $p = 1$ yielding the original magnitude spectral subtraction, and $p = 2$ yielding the power spectral subtraction algorithm.

From Eq. (3) , it can be noticed that

$$\frac{\left|V(\omega)\right|^2}{\left|Y(\omega)\right|^2} = \frac{1}{1 + SNR(\omega)}. \tag{5}$$

Those spectral subtraction algorithms rely on accurate estimate of SNR in the frequency domain. However, accurate estimation of instantaneous SNR is generally difficult if not impossible. The estimation error causes the problem of *musical noise* that is a processing artifact plaguing most spectral subtraction methods. Musical noise is low-amplitude tonal components with rapidly varying frequencies.

### 2.2 Wiener filter-based NR algorithm

Wiener filter theory can be used to reduce noise by optimizing a mathematically error criterion and illustrated in Fig. 2 [8]. The noisy signal $y(n)$ consists of clean speech $s(n)$ and noise $v(n)$ as Eq. (1) showed. The error $e(n)$ between the desired signal $d(n) = s(n)$ and its estimate $\hat{s}(n)$ is minimized in the minimum mean-square error (MMSE) sense. The estimate $\hat{s}(n)$ can be obtained by the inner product of two vectors $\mathbf{w}$ and $\mathbf{y}$.

$$\hat{s}(n) = \mathbf{w}^T \mathbf{y}, \tag{6}$$

where $\mathbf{w}^T = [w_0, w_1, \ldots, w_{M-1}]$ is the Wiener filter coefficient vector, and $\mathbf{y}^T = [y(n), y(n-1), \ldots, y(n-M+1)]$ is the input vector containing the past $M$ samples of the input. The impulse response of Wiener filter $w(n)$ is usually a finite impulse response (FIR) filter, and the frequency response of Wiener filter is

$$W(\omega_k) = \frac{\xi_k}{\xi_k + 1} \quad . \tag{7}$$

The $\xi_k$ is defined by

$$\xi_k \;@ \frac{P_{ss}(\omega_k)}{P_{vv}(\omega_k)} = \frac{E\left\{\left|S(\omega_k)^2\right|\right\}}{E\left\{\left|V(\omega_k)^2\right|\right\}} \tag{8}$$

as the *a priori* SNR at frequency $\omega_k$, where $P_{ss}(\omega_k)$ and $P_{vv}(\omega_k)$ are power spectra of clean speech and additive noise, respectively, and $E\{\cdot\}$ is the expectation operator. From Eq. (7), it could be noticed that $0 \le W(\omega_k) \le 1$, and $W(\omega_k) \approx 0$ when $\xi_k$ approaches to zero and $W(\omega_k) \approx 1$ when $\xi_k$ approaches to infinity. We can get the estimate of clean speech signal by filtering the noisy signal through the Wiener filter.

### 2.3 Statistical-model-based noise reduction algorithm

Minimum mean-square-error noise reduction (MMSE-NR) algorithm yields a nonlinear estimator of the magnitude of the DFT coefficients of the signal not the complex spectrum of the signal done by the Wiener filter. The algorithm is based on statistical model. This model makes two assumptions: (1) The Fourier transform coefficients have a Gaussian probability distribution. The mean of the coefficients is zero, and the variances of the coefficients are time-varying owing to the nonstationarity of speech. (2) The Fourier transform coefficients are statistically independent and, hence, uncorrelated.

The optimal MMSE nonlinear estimator[1] was searched that minimizes the mean-square error between the estimated and true magnitudes:

$$e = E\left\{\left(\hat{S}_k - S_k\right)^2\right\}, \tag{9}$$

where $\hat{S}_k$ and $S_k$ are the estimated and true spectral magnitudes of the clean speech signal at the frequency $\omega_k$, respectively. In particular, the expectation is finished by Bayesian mean-square error (MSE) approach, and the Bayesian MSE is given by:

$$Bmse\left(\hat{S}_k\right) = \int\int\left(S_k - \hat{S}_k\right)^2 p\left(\mathbf{Y}, S_k\right) d\mathbf{Y} dS_k\,, \tag{10}$$

where $\mathbf{Y} = \left[Y(\omega_0)Y(\omega_1)\cdots\ Y(\omega_{N-1})\right]$ is the noisy speech spectrum, and $p\left(\mathbf{Y}, S_k\right)$ is the joint probability density function (PDF). Minimization of Bayesian MSE with respect to $S_k$ leads to the optimal MMSE estimator given by:

$$\begin{aligned} \hat{S}_k &= \int S_k\, p\left(S_k \middle| \mathbf{Y}\right) dS_k \\ &= E\left[S_k \middle| \mathbf{Y}\right] \qquad\qquad . \\ &= E\left[S_k \middle| Y(\omega_0)Y(\omega_1)\cdots\ Y(\omega_{N-1})\right] \end{aligned} \tag{11}$$

In order to determine the MMSE estimator we first need to calculate the posterior PDF of $S_k$, i.e., $p\left(S_k \middle| Y(\omega_k)\right)$. Using Bayes' rule to determine it as:

$$\begin{aligned} p\left(S_k \middle| Y(\omega_k)\right) &= \frac{p\left(Y(\omega_k) \middle| S_k\right) p\left(S_k\right)}{p\left(Y(\omega_k)\right)} \\ &= \frac{p\left(Y(\omega_k) \middle| S_k\right) p\left(S_k\right)}{\int_0^\infty p\left(Y(\omega_k) \middle| s_k\right) p\left(s_k\right) ds_k}\,, \end{aligned} \tag{12}$$

where $s_k$ is a realization of the random variable $S_k$. Note that $p\left(Y(\omega_k)\right)$ is a normalization factor required to ensure that $p\left(S_k \middle| Y(\omega_k)\right)$ integrates to 1. Assuming statistical independence between the Fourier transform coefficients, i.e.,

$$E\left[S_k \middle| Y(\omega_0)Y(\omega_1)\cdots\ Y(\omega_{N-1})\right] = E\left[S_k \middle| Y(\omega_k)\right], \tag{13}$$

and using the preceding expression for $p\left(s_k \middle| Y(\omega_k)\right)$, the estimator in Eq. (11) simplifies to:

$$\begin{aligned}\hat{S}_k &= E\left[S_k \middle| Y(\omega_k)\right] \\ &= \int_0^\infty s_k\, p\left(s_k \middle| Y(\omega_k)\right) ds_k \\ &= \frac{\int_0^\infty s_k\, p\left(Y(\omega_k)\middle| s_k\right) p\left(s_k\right) ds_k}{\int_0^\infty p\left(Y(\omega_k)\middle| s_k\right) p\left(s_k\right) ds_k}\end{aligned} \tag{14}$$

Since

$$p\left(s_k \middle| Y(\omega_k)\right) ds_k = \int_0^{2\pi} p\left(Y(\omega_k)\middle| s_k,\theta_s\right) p\left(s_k,\theta_s\right) d\theta_s \;, \tag{15}$$

where $\theta_s$ is the realization of the phase random variable of $S\left(\omega_k\right)$, we get

$$\hat{S}_k = \frac{\int_0^\infty\int_0^{2\pi} s_k\, p\left(Y(\omega_k)\middle| s_k,\theta_s\right) p\left(s_k,\theta_s\right) d\theta_s ds_k}{\int_0^\infty\int_0^{2\pi} p\left(Y(\omega_k)\middle| s_k,\theta_s\right) p\left(s_k,\theta_s\right) d\theta_s ds_k} . \tag{16}$$

From the assumed statistical model, we know that $Y(\omega_k)$ is the sum of two zero-mean complex Gaussian random variables. Then the conditional PDF $p\left(Y(\omega_k)\middle| s_k,\theta_s\right)$ will also be Gaussian:

$$p\left(Y(\omega_k)\middle| s_k,\theta_s\right) = p_D\left(Y(\omega_k) - S(\omega_k)\right), \tag{17}$$

where $p_D\left(\cdot\right)$ is the PDF of the noise Fourier transform coefficients, $V(\omega_k)$. Then the Eq. (17) becomes:

$$p\left(Y(\omega_k)\middle| s_k,\theta_s\right) = \frac{1}{\pi P_{vv}(\omega_k)} \exp\left\{-\frac{1}{P_{vv}(\omega_k)}\left|Y\left(\omega_k\right) - S\left(\omega_k\right)\right|^2\right\}, \tag{18}$$

For complex Gaussian random variables, the magnitude $S_k$ and phase $\theta_s\left(k\right)$ random variables of $S\left(\omega_k\right)$ are independent, and the joint PDF as the product of the individual PDF's, i.e., $p\left(s_k,\theta_s\right) = p\left(s_k\right) p\left(\theta_s\right)$. The PDF of is uniform in $\left(-\pi,\pi\right)$, and therefore the joint probability is given by:

$$p(s_k,\theta_s)=\frac{s_k}{\pi P_{ss}(\omega_k)}\exp\left\{-\frac{{s_k}^2}{P_{ss}(\omega_k)}\right\}, \tag{19}$$

Substitute Eqs. (18) and (19) into Eq. (14), therefore the optimal MMSE magnitude estimator can be obtained as:

$$\hat{S}_k=\sqrt{\frac{P_{ss}(\omega_k)}{1+\xi_k}}\Gamma(1.5)\Phi(-0.5,1;-v_k), \tag{20}$$

where $\Gamma(\cdot)$ denotes the Gamma function, $\Phi(\cdot)$ denotes the confluent hyper-geometric function. Eq. (20) can be rewritten as

$$\hat{S}_k=\frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma_k}\exp\left(-\frac{v_k}{2}\right)\left[(1+v_k)I_0\left(\frac{v_k}{2}\right)+v_kI_1\left(\frac{v_k}{2}\right)\right]Y_k, \tag{21}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zero and the first order, respectively, $Y_k$ is the spectral magnitude of the noisy signal at the frequency $\omega_k$, and $v_k$ is defined by

$$v_k=\frac{\xi_k}{1+\xi_k}\gamma_k \tag{22}$$

where $\gamma_k$ denotes as the *a posteriori* SNR, and $\gamma_k$ is defined as

$$\gamma_k @ \frac{{Y_k}^2}{P_{vv}(\omega_k)}=\frac{{Y_k}^2}{E\left\{\left|V(\omega_k)\right|^2\right\}} \tag{23}$$

Generally, we do not have known the noise variance and the *a priori* SNR $\xi_k$ but measured noisy signal $y(n)$. However, the noise variance can be estimated and computed via a VAD in MMSE-NR algorithm if we assume the noise is stationary. A statistical-model-based VAD was used:

$$\frac{1}{N}\sum_{k=1}^{N-1}\log\left(\frac{1}{1+\xi_k}\exp\left(\frac{\gamma_k\xi_k}{1+\xi_k}\right)\right)\underset{H_0}{\overset{H_1}{\gtrless}}\Delta, \tag{24}$$

where *N* is the size of the fast Fourier transform, $H_1$ denotes the hypothesis of speech presence, *H*$_0$ denotes the hypothesis of speech absence, and $\Delta$ is a fixed threshold, which was usually set to 0.15.

As for *a priori* SNR $\xi_k$, a method for estimating the *a priori* SNR $\xi_k$ is called the "decision-directed". This method is assuming that *a priori* SNR $\xi_k$ is related to *a posteriori* SNR $\gamma_k$ by

$$\xi_k(m) = E\{\gamma_k(m)\} - 1 \quad , \tag{25}$$

where $m$ is the number of frame in the frame-based MMSE-NR algorithm. An estimator of the *a priori* SNR $\xi_k$ is given by

$$\hat{\xi}_k(m) = a\frac{\hat{S}_k^{\,2}(m-1)}{P_{vv}(\omega_k, m-1)} + (1-a)\max\left(\gamma_k(m)-1,\ 0\right) \quad . \tag{26}$$

where $a$ is a weighting factor commonly chosen to be $a = 0.98$ and $0 < a < 1$, and $\hat{S}_k^{\,2}(m-1)$ is the amplitude magnitude estimator of speech signal obtained in the past frame. Therefore, the estimate of clean speech signal magnitude can be calculated by Eq. (21). Finally, construct the clean speech signal spectrum $\hat{S}(\omega_k)$ by combing the estimate of clean speech signal magnitude spectrum $\hat{S}_k$ with the noisy signal phase spectrum $j\theta_y(k)$ and calculate the inverse DFT of $\hat{S}(\omega_k)$ to obtain the time-domain processed speech signal $\hat{s}(n)$. The processed signal spectrum can be showed as

$$\hat{S}(\omega_k) = \hat{S}\exp\left(j\theta_y(k)\right). \tag{27}$$

It can be shown that the optimal phase estimate is actually the noisy phase by Ephraim and Malah [11].

### 2.4 Karhunen-Loéve transform (KLT)-based noise reduction

A subspace algorithm, Karhunen-Loéve transform noise reduction (KLT-NR) algorithm [11]-[12], is rooted on linear algebra theory and can be also applied to enhance speech signal. First, a noisy speech vector $\mathbf{y}$ comprises clean speech and noise signal $\mathbf{v}$ vectors as

$$\mathbf{y} = \mathbf{s} + \mathbf{v} = \left[y(n), y(n-1), \ldots, y(n-M+1)\right]^T \tag{28}$$

containing $M$ samples of speech, where $\mathbf{s} = \left[s(n), s(n-1), \ldots, s(n-M+1)\right]^T$ is the noisy-free vector and $\mathbf{v} = \left[v(n), v(n-1), \ldots, v(n-M+1)\right]^T$ is the noise signal vector. Let $\hat{\mathbf{s}} = \mathbf{H} \cdot \mathbf{y}$ be a linear estimator of the clean speech vector $\mathbf{s}$, where $\mathbf{H}$ is a $M \times M$ matrix. The residual error $\varepsilon$ obtained by the estimation is given by:

$$\varepsilon = \hat{\mathbf{s}} - \mathbf{s} = \mathbf{H} \cdot \mathbf{y} - \mathbf{s} \tag{29}$$

The energy of the residual error $\overline{\varepsilon^2}$ is defined as

$$\overline{\varepsilon^2} = E\left[\varepsilon^T \varepsilon\right] = \mathrm{tr}\left(E\left[\varepsilon^T \varepsilon\right]\right) \tag{30}$$

The optimum linear estimator can be obtained by solving the unconstrained optimization problem:

$$\min_{\mathbf{H}_{KLT}} \overline{\varepsilon^2} \tag{31}$$

Substitute Eq. (29) in Eq. (30), we obtain:

$$\begin{aligned} \overline{\varepsilon^2} &= \mathrm{tr}\left(E\left[(\mathbf{H} \cdot \mathbf{y} - \mathbf{s})(\mathbf{H} \cdot \mathbf{y} - \mathbf{s})^T\right]\right) \\ &= \mathrm{tr}\left(E\left[(\mathbf{H} \cdot \mathbf{y}\mathbf{y}^T \cdot \mathbf{H}^T - \mathbf{s}\mathbf{y}^T\mathbf{H}^T - \mathbf{H} \cdot \mathbf{y}\mathbf{s}^T + \mathbf{s}\mathbf{s}^T\right]\right) \\ &= \mathrm{tr}\left(E\left[(\mathbf{H}\mathbf{R}_y\mathbf{H}^T - \mathbf{R}_{sy}\mathbf{H}^T - \mathbf{H}\mathbf{R}_{ys} + \mathbf{R}_s\right]\right) \end{aligned} \tag{32}$$

where $\mathbf{R}_s$ and $\mathbf{R}_y$ are the clean and noisy signal covariance matrices, respectively. Besides, $\mathbf{R}_{sy} \triangleq E\{\mathbf{s}\mathbf{y}^T\}$ and $\mathbf{R}_{ys} \triangleq E\{\mathbf{y}\mathbf{s}^T\}$. For white noise, the noise covariance matrix is given by

$$\mathbf{R}_v = \sigma_v^2 \mathbf{I}, \tag{33}$$

where $\sigma_v{}^2$ is the noise variance and $\mathbf{I}$ is an $M \times M$ identity matrix.

Furthermore, assume the clean speech and noise vectors are uncorrelated and zero mean, then the matrix $\mathbf{R}_y$ can be shown to be

$$\mathbf{R}_y = \mathbf{R}_s + \mathbf{R}_v \quad , \tag{34}$$

Take the derivative of previous $\overline{\varepsilon^2}$ with respect to $\mathbf{H}$ and set it equal to zero, in addition, make use of the fact that $\mathbf{R}_{ys}^T = \mathbf{R}_{sy}$ and substitute in Eq. (34), we obtain the optimal estimator:

$$\mathbf{H}_{opt} = \mathbf{R}_s \left( \mathbf{R}_s + \sigma_v^2 \mathbf{I} \right)^{-1} . \tag{35}$$

The estimator is simplified by using the eigenvalue decomposition (EVD) of $\mathbf{R}_s$:

$$\mathbf{R}_s = \mathbf{U} \mathbf{\Lambda}_s \mathbf{U}^T , \tag{36}$$

where $\mathbf{U}$ is the unitary eigenvector matrix, and the matrix is often called KLT transform. Substitute Eq. (36) into Eq. (35) and assume $\mathbf{R}_s$ has a rank $K\left(K < M\right)$, we obtain:

$$\mathbf{H}_{opt} = \mathbf{U}\mathbf{G}\mathbf{U}^T = \left[\mathbf{U}_1 \quad \mathbf{U}_2\right] \begin{bmatrix} \mathbf{G} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1{}^T \\ \mathbf{U}_2{}^T \end{bmatrix} = \mathbf{U}_1 \mathbf{G} \mathbf{U}_1{}^T , \tag{37}$$

where $\mathbf{G}$ is a diagonal matrix ( $K \times K$ )

$$\mathbf{G} = \Lambda_s \left( \Lambda_s + \sigma_v{}^2 \mathbf{I}_k \right)^{-1} \tag{38}$$

with diagonal matrix $\Lambda_s$ containing the eigenvalues $\sigma_s$ sorted in descending order $K$. The eigenvector matrix $\mathbf{U}$ can be partition as $\mathbf{U} = \left[\mathbf{U}_1 \quad \mathbf{U}_2\right]$ in Eq. (37), where $\mathbf{U}_1$ is a $M \times K$ matrix. Hence, the enhanced speech signal vector can be obtained by $\hat{\mathbf{s}} = \mathbf{H}_{opt}\mathbf{y}$.

## 3. ENHANCED MMSE-NR ALGORITHMS

In the section, three approaches of technical refinement that can be done to enhance the MMSE-NR algorithm are presented.

### 3.1 Noise reduction algorithms based on MMSE method

Two noise estimators are posed in this section as the time-recursive averaging (TRA) algorithm[13] and algorithm combined the TRA and the VAD. As mentioned earlier in MMSE-NR algorithm, the noise variance can be estimated and computed via a VAD if the noise is stationary. However, the majority of the VAD algorithms encounters problems in low-SNR conditions if the noise is nonstationary. An algorithm called time-recursive averaging (TRA) algorithm is suitable for highly nonstationary noisy environment to estimate noise variance. A NR algorithm that noise variance can be estimated via the TRA algorithm we call it "MMSE-TRA-NR" in this paper.

In TRA algorithm, the individual frequency bands of noise spectrum can be updated by noisy spectrum when SNR is extremely low, or the estimate of noise variance at the last frame will be kept on in estimating noise variance. The TRA algorithm has the form:

$$\hat{\sigma}_v^{\,2}(\lambda,k)=\alpha(\lambda,k)\hat{\sigma}_v^{\,2}(\lambda-1,k)+\left(1-\alpha(\lambda,k)\right)\left|Y(\lambda,k)\right|^2 \tag{39}$$

where $\left|Y(\lambda,k)\right|$ is the noisy speech magnitude spectrum, $\hat{\sigma}_v^{\,2}(\lambda,k)$ is the estimate of noise variance at frame $\lambda$ and frequency $k$, and $\alpha(\lambda,k)$ is the smoothing factor, which is the time and frequency dependent. Different algorithms were developed depending on the selection of the smoothing factor $\alpha(\lambda,k)$. Some chose to calculate $\alpha(\lambda,k)$ based on the estimated SNR of each frequency bin, whereas others chose to calculate $\alpha(\lambda,k)$ based on the probability of speech being present/absent at frequency $k$. Others chose to use a fixed value for $\alpha(\lambda,k)$, but updated $\hat{\sigma}_v^2(\lambda,k)$ only after a certain condition was met. In the paper, the smoothing factor $\alpha(\lambda,k)$ is chosen to be a sigmoid function of the posteriori SNR $\gamma_k(\lambda)$ as:

$$\alpha(\lambda,k)=\frac{1}{1+e^{-\beta(\gamma_k(\lambda)-\delta)}} \tag{40}$$

where $\beta$ and $\delta$ are parameters, and *a posteriori* SNR $\gamma_k(\lambda)$ can be given by

$$\gamma_k(\lambda)=\frac{|Y(\lambda,k)|^2}{\frac{1}{10}\sum_{m=1}^{10}\hat{\sigma}_v^2(\lambda-m,k)} \tag{41}$$

In Eq. (41), the denominator is the average of the estimated noise variance in the past ten frames. Figure 3 plots the smoothing factor $\alpha$ calculated according to Eq. (40) for different values of the parameter $\beta$ when $\delta=1$. A noisy speech signal (dotted line) corrupted by a non-stationary noise that consists of three different level of loudness is shown in Fig. 4. Top panel shows the noise (solid line) estimated using the aforementioned VAD algorithm, and bottom panel shows the noise (solid line) estimated using the TRA algorithm. From Fig. 4 we can notice that the TRA algorithm works better than the VAD algorithm for non-stationary noise.

To enhance the TRA algorithm, we combine the TRA algorithm and the VAD algorithm and call it “MMSE-VAD-TRA-NR” in this paper. Using the VAD algorithm to verify the current frame of the input signal is the speech or the noise. According to the verification of the VAD algorithm, the speech signal will through the TRA algorithm process, and the noise signal will be deleted.

### 3.2 Intelligent tuning of the parameters in enhanced MMSE-NR algorithm

As mentioned previously, the parameters $\beta$ and $\delta$ are used in the sigmoid function of the TRA algorithm for noise estimation. Conventionally, choices such as $\delta = 1.5, 15 \le \beta \le 30$ are recommended in the literature [11]. To our surprise, we found that these two parameters $\beta$ and $\delta$ have profound effects on the NR performance of the MMSE-TRA-NR algorithm and the MMSE-VAD-TRA-NR algorithm. It is then worth exploring how to adjust these two parameters such that noise reduction performance can be maximized without too much quality degradation. The SA method provides an optimal way to intelligently tune those parameters.

### 3.2.1 SA method

SA [14]-[16] is a generic probabilistic meta-algorithm for the global optimization problem, namely locating a good approximation to the global optimum of a given function in a large search space. SA has demonstrated to be a good technique for solving global optimization problems with many local optima. The flow diagram of the SA is illustrated in Fig. 5. In SA method, each point of the search space is analogous to a state of some physical system, and the objective function $Q$ to be maximized is analogous to the internal energy of the system in that state. The goal is to bring the system from an initial state to a new state with the minimum possible energy.

Two conditions can transfer state and update the objective function in accepting rule of SA method. One is the objective function increasing. The other is the objective function decreasing but the acceptance probability function is more than a random number $\varphi$ which is randomly generated subject to the uniform distribution on the interval $(0,1)$. The acceptance probability function is given by

$$p_{SA} = \exp(\Delta Q / T) \tag{42}$$

where $\Delta Q$ denotes the variation of the objective function, and $T$ is a control parameter called the temperature. It follows that the system actually may move to the new state even when it is worse than the current one. This feature prevents the method from staying in a local maximum – a state that is worse than the global maximum but better than any of its neighbors. Initially the high temperature $T$ causes the high probability of accepting a move that decreases the objective function. Finally, the probability of accepting a move becomes extremely small when the objective function is decreasing continuously, and the temperature is getting lower in accordance with an annealing schedule.

The most generally employed annealing schedule is exponential cooling which begins at some initial temperature $T_0$ and decreases temperature in steps according to

$$T_{k+1} = \alpha_c T_k \tag{43}$$

where $0 < \alpha_c < 1$ is a cooling factor. Typically, a fixed number of moves must be accepted at each temperature before proceeding to the new state. A way of SA action is stopped either when the temperature reaches some final value $T_f$ or the system is not transformed to a new state after some times. An empirical choice for $\alpha_c$ is 0.95 and that $T_0$ should be chosen so that the initial acceptance probability is higher than 0.8. The initial solution is generated typically at random.

### 3.2.2 Objective function Q

An appropriate objective function is very important for optimizing the performance in MMSE-TRA-NR algorithm and MMSE-VAD-TRA-NR algorithm. Two objective indices, the segmental SNR (SNRseg) and the perceptual evaluation of sound quality (PESQ) [18], were applied to construct the objective function. The SNRseg is a basic objective measure to evaluate noise reduction algorithms, and has the form:

$$SNRseg = \frac{10}{M_s} \sum_{m=0}^{M_s - 1} \log_{10} \frac{\sum_{n=N_s m}^{N_s m + N_s - 1} s^2(n)}{\left(s(n) - \hat{s}(n)\right)^2} \quad , \tag{44}$$

where $N_s$ is the frame length, and $M_s$ is the number of frames in the signal. The SNRseg can reflect the SNR level of the enhanced speech by NR algorithms. As for the PESQ, it is widely used for automated assessment of the speech quality in telephony industry. The structure of PESQ is complicated that the original and degraded signals are first level-equalized to a standard listening level and filtered by a filter with response similar to a standard telephone handset. The signals are aligned in time to correct for time delays, and then processed through an auditory transform to obtain the loudness spectra. More details of the PESQ can be found in ITU-T P. 862 [18]. In a word, the SNRseg and the PESQ reflect the SNR level and the sound quality, respectively, of the processed signals via NR algorithms. The higher values of SNRseg and PESQ both indicate the better noise reduction performance.

Consider the objective function as a linear combination of the SNRseg and the PESQ, the weights between the SNRseg and the PESQ can be found from a subjective listening test. Three subjective indices including *noise reduction*, *sound quality* and *total preference* were employed in this listening test. The grading scale is set to be -3~3. Five NR algorithms are applied to two kinds of noise at SNR level 5 dB: (1) white noise (2) car noise. Figure 6 shows the waveforms of the test sentence corrupted by white noise (top panel) and car noise (bottom panel) respectively. Figure 7(a)-(b) show the spectrograms of the test sentence corrupted by white noise and car noise. NR algorithms including spectral subtraction, Wiener filtering, MMSE-NR, MMSE-TRA-NR, MMSE-VAD-TRA-NR and KLT-NR algorithms were examined. The sampling rate is 8 kHz, the frame length is about 20~32 ms, and the amount of overlap is 50%. The loudness of all reproduced signals was adjusted to be the same level. A headset was used as the means of audio rendering. Multiple regression analysis enables us to establish the relationship between several independent variables (*noise reduction* and *sound quality*) and a dependent variable (*total preference*). There are thirty-two experienced listeners participating in the subjective test. The results of multiple regression analysis provide the weights between the SNRseg and the PESQ for the objective function. Hence, the objective function is experimentally constructed as

$$Q = 1.867 * SNRseg + PESQ \quad . \tag{45}$$

We can achieve the optimal performances in MMSE-TRA-NR algorithm and MMSE-VAD-TRA-NR algorithm according to Eq. (45) by using SA method in terms of objective measures.

### 3.2.3 Compare with and without optimization MMSE-NR algorithms

The parameters $\beta$ and $\delta$ in the MMSE-TRA-NR algorithm can be randomly chosen to $\beta = 1.6$ and $\delta = 1$ for processing the previous noisy speech signals in two noise conditions. However, the optimal parameters can be obtained by using SA method. For white noise condition, the optimal $\beta = 0.6117$ and the optimal $\delta = 0.5214$ in the MMSE-TRA-NR algorithm, the optimal $\beta = 0.5671$ and the optimal $\delta = 0.2606$ in the MMSE-VAD-TRA-NR algorithm. For car noise condition, the optimal $\beta = 0.7128$ and the optimal $\delta = 0.5265$ in the MMSE-TRA-NR algorithm, the optimal $\beta = 0.6896$ and the optimal $\delta = 0.1724$ in the MMSE-VAD-TRA-NR algorithm. Table I shows the NR performance of the MMSE-TRA-NR algorithm and the MMSE-VAD-TRA-NR algorithm in terms of the SNRseg and PESQ for different values of parameters $\beta$ and $\delta$. Form table I, we can notice that there are higher values of the SNRseg and the PESQ in MMSE-TRA-NR and MMSE-VAD-TRA-NR with optimal parameters $\beta$ and $\delta$.

Another subjective listening test was conducted to assess the NR performance between the random and optimal parameters in MMSE-TRA-NR and MMSE-VAD-TRA-NR. The test conditions are similar to the subjective listening test for constructing objective function of SA method. The grading scale is set to be 1~5, as recommended in ITU-T P.835 [19]. Three subjective indices including *Scale of Signal Distortion (SIG)*, *Scale of background Intrusiveness (BAK)* and *Scale of Overall Quality (OVL)* were employed in the listening test. Every subject participating in the test is instructed with the definitions of the preceding subjective indices and the procedure prior to the listening test. Figures 8(a)-(d) show the results of the subjective listening test in white noise and car noise. The scores from all subjects were also processed by using the MANOVA [20] to justify the statistical significance of the test results. The average, 5%-95% bracket and the significance level of the grades were shown in the analysis. Cases with significance levels below 0.05 indicate that statistically significant difference exists among methods. From Figs. 8(a)-(d) and Table II, there is no significant difference in *OVL* but *SIG* and *BAK* between the random and optimal parameters in MMSE-TRA-NR algorithm. The optimal parameters lead to the worse values in *SIG*, however, the random parameters lead to the worst values of *BAK* that the values almost are the lowest about 1 in the two noise conditions. According to the values of BAK, there is almost no NR performance in the MMSE-TRA-NR and MMSE-VAD-TRA-NR by using random parameters. Form the results of the objective and subjective tests, we always chose the optimal parameters in MMSE-TRA-NR and MMSE-VAD-TRA-NR algorithms that optimized by SA method. Furthermore, the optimal MMSE-VAD-TRA-NR will be compared to some NR algorithms in objective and subjective tests later.

## 4. OBJECTIVE AND SUBJECTIVE EVALUATIONS

### 4.1 Performance evaluation of NR algorithms in objective measure

Two objective measures, the SNRseg and the PESQ, are employed to assess the performance of six reduction algorithms in two kinds of background noise (white noise and car noise) at SNR levels 5dB. Six NR algorithms are spectral subtraction, Wiener filtering, MMSE-NR, MMSE-TRA-NR, MMSE-VAD-TRA-NR, and KLT-NR algorithms. All test conditions are similar to compare with and without optimization MMSE-TRA-NR algorithms by using SA method in terms of the SNRseg and PESQ. These measures assess speech quality by estimating the “distortion” between the clean and processed signals and then mapping the estimated distortion value to a quality metric.

Figure 9 (a) shows the noise-free speech signal used for a computer simulation, where the sampling rate is 8 kHz. The noisy and the processed speech signals by those NR algorithms are shown in Figs. 9(b)-(c). Computational requirement (processing time) and objective NR performance are compared in Table III. The test signals and conditions are similar to the performance evaluation of NR algorithms in objective test. The SNRseg and the PESQ are employed in the objective NR performances. In terms of the SNRseg, the less noise estimation causes more residual noise in order to avoid serious speech distortion in the Wiener filtering algorithm. Therefore, the Wiener filtering algorithm leads to the lowest values of SNRseg in all noise conditions. Opposition to the Wiener filtering, there are the highest values of SNR in the KLT-NR algorithm. As for PESQ, the result indicated that there is no significant difference between those NR algorithms in speech quality of the processed signals.

### 4.2 Performance evaluation of NR algorithms by subjective listening tests

In order to compare the preceding NR algorithms, subjective listening tests were conducted in terms of sound quality. The listening tests were conducted according to the standards ITU-T P.835 [19]. Thirty-two experienced listeners participated in the subjective tests. The grading scale is set to be 1~5, as recommended in ITU-T P.835 [19]. Six noise reduction algorithms, spectral subtraction, Wiener filtering, MMSE-NR, MMSE-TRA-NR, MMSE-VAD-TRA-NR and KLT-NR are compared in the test. The sampling rate is 8 kHz, the frame length is about 20~32 ms, and the amount of overlap is 50%. Three subjective indices including *Scale of Signal Distortion (SIG), Scale of background Intrusiveness (BAK)* and *Scale of Overall Quality (OVL)* were employed in the listening test. Every subject participating in the test is instructed with the definitions of the preceding subjective indices and the procedure prior to the listening test. In the listening tests, those NR algorithms are applied to two kinds of noise: (1) white noise (2) car noise at SNR level 5 dB. The design of the subjective tests is completely the same with the preceding subjective test for comparing with and without optimization MMSE-TRA-NR algorithms using SA method. Listening tests were conducted for the noise corrupted speech and the results are shown in Figs. 10(a)-(b). Not only the mean grades but also the significance levels were shown in the analysis for different NR algorithms. The vertical bars indicate 95% confidence intervals. The test results were processed using MANOVA [20]. The significance level in the MANOVA output is summarized in Table V. Cases with significance levels below 0.05 indicate that statistically significant difference exists among methods. According to Table V, the difference in all cases of the test results was found to be statistically significant. Furthermore, multiple paired comparisons according to a post-hoc Tukey's HSD test [20] were conducted to assess significant differences between NR algorithms. Table VI shows the results from the Tukey's HSD test for the signal distortion *SIG*, noise distortion *BAK* and overall quality *OVL* comparisons. Asterisks in the table indicate absence of statistically significant difference ($p$ > 0.05) between the algorithm with the highest score and denoted algorithm. That is to say, the NR algorithms denoted by asterisks in Table VI performed equally well. On the contrary, the algorithms with no asterisks performed poorly.

From Table VI, we notice that the KLT-NR algorithm performed poorly in terms of *SIG* in all noise conditions. The result indicates that the overestimate of noise in KLT-NR algorithm brings to more signal distortion. In terms of noise distortion *BAK*, as mentioned earlier in the objective test, the Wiener filtering algorithm obtained the worst scores in all noise condition in order to avoid serious signal distortion. Besides, the residual noise that commonly is musical noise results in the worse value of *BAK* in the spectral subtraction algorithm for the two noise conditions. Another surprising thing is that the MMSE-TRA-NR performed poorly in real-world noise (car noise) condition. However, the MMSE-VAD-TRA-NR has better perform than the MMSE-TRA-NR in car noise condition. As for overall quality *OVL*, there is no significant difference between those NR algorithms in car noise condition. However, the spectral subtraction and KLT-NR algorithms obtained the worse scores of *OVL* than the other NR algorithms for white noise case. The result shows that listeners might be influenced more by speech distortion *SIG* in terms of *OVL* when making judgments. A summary from Table VI is that there is no much difference between the MMSE-NR, MMSE-TRA-NR and MMSE-VAD-TRA-NR algorithms in terms of all subjective indices in the two noise scenarios.

## 5. ENHANCED MMSE-NR ALGORITHMS FOR AUTOMATIC SPEECH RECOGNITION

We use the MMSE-TRA-NR algorithm to enhance the acoustic speech recognition (ASR). The database of the speech for the ASR is 50 short chinese commands, each command has 6 male and 5 female speakers. The ASR is using the Hidden Markov Model (HMM). Figure 11 shows that the recognition rates at different noise types and noise level. It is obivous that the recognition rates of processed signals are lower than the recognition rates of original signals in babble and movie noise conditions. Beside, the recognition rates at 6 to 12 dB in the movie condition noise are significant lower than the recognition rates of original signals. Therefore, we want to optimal recognition rates of the movie noise condition at 6 to 12 dB and get the optimal parameters to optimal the recognition rate of babble and movie noise conditions.

First, we optimal the window length. There is a trade-off between time-resolution and frequency-resolution when selecting the window length for frequency-domain analysis. A longer frame length results in more accurate spectral represenation when we want to obtain better frequency domain resolution. As a tradeoff between these two competing criteria, a frame length between 20 ms and 30 ms has been widely used in speech analysis. Even though a window of such short duration is optimal for analyzing speech signal, there is no guarantee that the optimal length would be the same for estimating the noise component. It is widely known that noise changes more slowly than speech signal, thus based on the above discussion, it is quite obvious that longer windows might be better for estimating the noise. Whereas, the recognition rate is the most important thing that we concern in our condition. Therefore, we only have to find a window length that can balance the time-resolution and frequency-resolution and has higher recognition rate. Figure 12 shows that the recognition rate of movie noise condition will change with different window length. In movie noise condition at 6 to 12 dB, we search the window length from 20 to 100 ms and find that the best window length is 50 ms.

Beside the window length, we also optimal the parameters $\beta$ and $\delta$. We use the SA method, and the cost function is the recognition rate. Figure 13 shows that the recognition rates are higher when the optimal parameters $\beta = 3.0964$ and $\delta = 11.9103$ is used. But, the recognition rates of original signal with SA are lower than the recognition rates of original signal.

After the passing through the MMSE-TRA algorithm, the features of the speech may be lost. Therefore, we add some of original signals to the processed signals in order to restruct the features of the speech. Fig 14 shows that the recognition rates with adding different ratio of processed signals in movie noise condition at 9 dB, and the best ratio is 70 %.

With these new optimal parameters $\beta = 3.0964$ , $\delta = 11.9103$ , and the window length is 50 ms, and ratio of the processed signals is 70 %, the Fig 15 shows the result. After a series of tuning the parameters, the MMSE-TRA algorithm has higher recognition rates than the original signals.

## 6. CONCLUSIONS

An optimization method to efficiently search the optimal parameters in the MMSE-VAD-TRA-NR algorithms has been proposed. In order to obtain optimal NR performances, the optimization method employs a SA method and constructs an appropriate objective function to achieve the goal. We observe that the parameters $\beta$ and $\delta$ need to be chosen carefully because they can affect the estimate of the noise spectrum obviously; that is, $\beta$ and $\delta$ are the most important parameter to affect the NR performance of the MMSE-VAD-TRA-NR significantly.

The comparisons and research of some NR algorithms have been represented in computation complexity, objective tests, and subjective listening tests. The results of the processing time reflect the calculation and processing data complexity in those NR algorithms. The results of objective and subjective tests do not only imply that the

Wiener filtering algorithm yield the more residual noise in order to avoid serious signal distortion, but also shows that the overestimate of noise results in the lowest scores of signal distortion *SIG* in the KTL-NR algorithm. The results of the subjective listening tests nearly indicate that for all subjective indices, the MMSE-NR, MMSE-TRA-NR and MMSE-VAD-TRA-NR algorithms perform equally well in the white and car noise scenarios. Therefore, it can be concluded that the MMSE-NR, MMSE-TRA-NR and MMSE-VAD-TRA-NR algorithms are better NR algorithms than others according to the aforementioned comparisons and research in the paper.

To enhance recognition rate is not the main propose of the general NR algorithm. After the general NR algorithm processing, the signal will enhance speech and reduce the noise. But, sometimes the speech will be distortion because the noise reduction of the NR algorithm is too aggressive. However, MMSE-TRA-NR can change the parameters to enhance the recognition rate and avoid the trade-off between the distortion and noise reduction.

Future research is planned on integrating the noise reduction algorithms with the microphone arrays to exploit its full potential of noise suppression in telecommunication applications such as peer-to-peer internet telephony networks, hands free car-kits, wireless earphones, and so forth.

## REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short time spectral amplitude estimator," IEEE Trans on Acoustic, Speech, Signal Process. **32**(6), 1109-1121 (1984).

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. Acoust., Speech, Signal Process. **28**(2), 137-145 (1980).

[3] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control a Practical Approach*, (John Wiley, New York, 2004)

[4] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," IEEE Trans on Acoustics, Speech, Signal Process. **281**(1), 99-102 (1980).

[5] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," IEEE Trans on Acoustics, Speech, Signal Process, **24**(3), 243-248 (1976).

[6] U. Zölzer, *DAFX – Digital Audio Effects*, (John Wiley, New York, 2002)

[7] S. L. Gay, J. Benesty, *Acoustic Signal Processing for Telecommunication* (Kluwer Academic Publishers, Norwell, MA, 2000)

[8] N. Wiener, Extrapolation, *Interpolation, and Smoothing of Stationary Time Series with Engineering Applications* (John Wiley, New York, 1949)

[9] B. Farhang-Boroujeny, *Adaptive Filters Theory and Application* (John Wiley, New York, 2000)

[10] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction* (John Wiley, New York, 1996)

[11] P. C. Loizou, *Speech Enhancement Theory and Practice* (CRC, New York, 2007)

[12] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech

corrupted by colored noise," IEEE Trans on Acoustics, Speech, Signal Process, **11**(4), 334-341 (2003).

[13] L. Lin ,W. Holmes and E. Ambikairajah, "Adaptive Noise Estimation Algorithm for Speech Enhancement," Electronics Lett., **39**(9), 754-755 (2003).

[14] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, "Equations of State Calculations by Fast Computing Machines," JCP, 1087-1092 (1953).

[15] A. Das and B. K. Chakrabarti (Eds.), *Quantum Annealing and Related Optimization Methods* (Springer, Heidelberg, 2005)

[16] J. De Vicente, J. Lanchares, R. Hermida, "Placement by Thermodynamic Simulated Annealing," Physics Letters A, **317**(5-6), 415-423 (2003).

[17] S. J. Orfanidis, *Optimum Signal Processing, An Introduction*, (McGraw Hill, New York, 1996).

[18] ITU-R Rec. P.862, "*Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*," (International Telecommunications Union, Geneva, Switzerland, 2000).

[19] ITU-R Rec. P.835, "*Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*," (International Telecommunications Union, Geneva, Switzerland, 2003).

[20] G. Keppel and S. Zedeck, *Data analysis for research designs*. (Freeman, New York, 1989).

TABLE I. The NR performance of the MMSE-TRA-NR algorithm and MMSE-VAD-TRA algorithm in terms of the SNRseg and PESQ for different values of parameters $\beta$ and $\delta$

| **Noise type** | **Algorithms** | $\beta$ | $\delta$ | **SNRseg** | **PESQ** |
|---|---|---|---|---|---|
| white noise | TRA | 1.6 | 1 | -1.0942 | 1.9639 |
| | optimal-TRA | 0.6117 | 0.5214 | 1.5155 | 2.1619 |
| | VAD-TRA | 1.6 | 1 | -1.1833 | 1.7369 |
| | optimal-VAD-TRA | 0.5671 | 0.2606 | 1.5899 | 2.1582 |
| car | TRA | 1.6 | 1 | -1.5609 | 2.2168 |
| | optimal-TRA | 0.7128 | 0.5265 | 0.7061 | 2.3145 |
| | VAD-TRA | 1.6 | 1 | -1.4524 | 2.1396 |
| | optimal-VAD-TRA | 0.6896 | 0.1724 | 0.7666 | 2.3219 |

TABLE II. The MANOVA results of the subjective listening test in white noise and car noise conditions for compare with and without optimization in MMSE-TRA-NR and MMSE-VAD-TRA-NR.

| **Algorithm** | **Noise type** | **Significance value** | | |
|---|---|---|---|---|
| | | **SIG** | **BAK** | **OVL** |
| MMSE-TRA | White noise | 0.040 | 0.000 | 0.117 |
| | Car noise | 0.017 | 0.000 | 0.784 |
| MMSE-VAD-TRA | White noise | 0.042 | 0.000 | 0.126 |
| | Car noise | 0.015 | 0.000 | 0.631 |

TABLE III. Comparison of computational requirement and objective noise reduction performance of the six noise reduction algorithms.

| | SNRseg | | PESQ | |
|---|---|---|---|---|
| Noise condition<br>NR algorithms | White | Car | White | Car |
| Spectral subtraction | 2.115 | 1.450 | 2.224 | 2.118 |
| Wiener filtering | 0.878 | 0.073 | 2.162 | 2.322 |
| MMSE-NR | 2.215 | 1.224 | 2.250 | 2.394 |
| MMSE-TRA-NR | 1.515 | 0.7061 | 2.161 | 2.314 |
| MMSE-VAD-TRA-NR | 1.5899 | 0.7666 | 2.1582 | 2.3219 |
| KLT-NR | 3.177 | 1.856 | 2.400 | 2.367 |

TABLE IV. The MANOVA output of the listening test of the NR algorithms. Cases with significance value *p* below 0.05 indicate that statistically significant difference exists among all methods.

| **Noise type** | **Significance value *p*** | | |
|---|---|---|---|
| | **SIG** | **BAK** | **OVL** |
| White noise | 0.007 | 0.000 | 0.006 |
| Car noise | 0.012 | 0.000 | 0.083 |

TABLE V. The result from Tukey’s HSD test for *SIG*, *BAK* and *OVL* between NR algorithms. (The denoted NR algorithms by asterisks have equally good performance. The algorithms with no asterisks have poor performance.)

| | ***SIG*** | | ***BAK*** | | ***OVL*** | |
|---|---|---|---|---|---|---|
| **Noise condition**<br>**NR algorithms** | **White** | **Car** | **White** | **Car** | **White** | **Car** |
| Spectral subtraction | * | * | | | | * |
| Wiener filtering | * | * | | | * | * |
| MMSE-NR | * | * | * | * | * | * |
| MMSE-TRA-NR | * | * | * | | * | * |
| MMSE-VAD-TRA-NR | * | * | * | * | * | * |
| KLT-NR | | | * | * | | * |

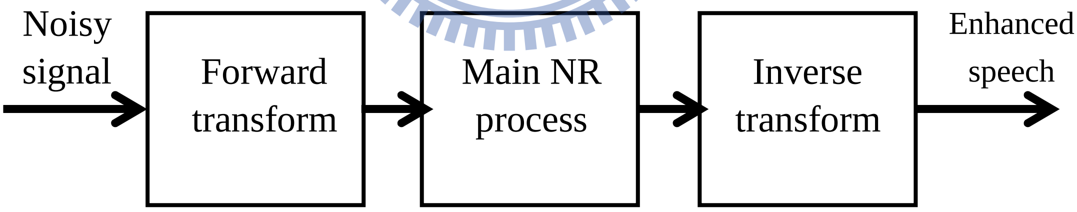


FIG. 1. General structure of NR algorithms.

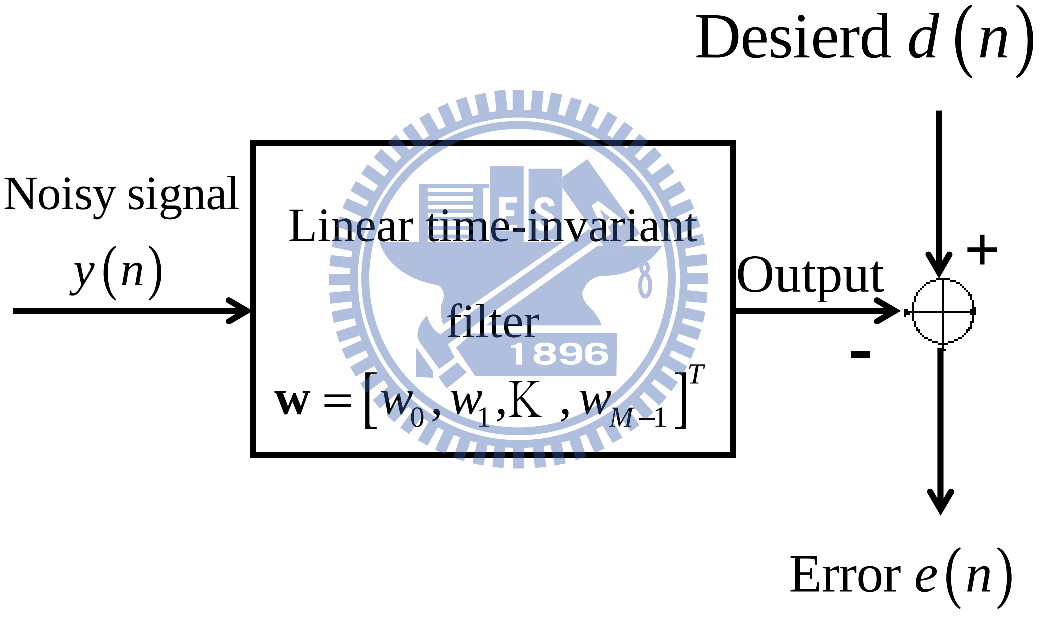


FIG. 2. Block diagram of the filtering problem.

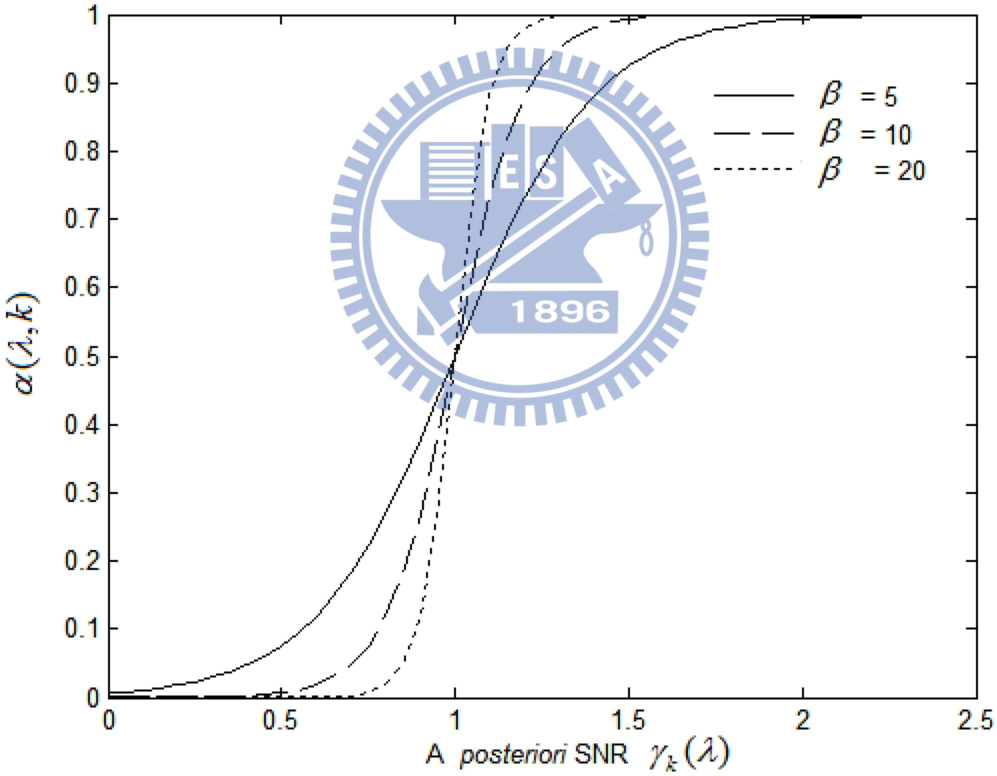


FIG. 3. The smoothing factor $\alpha(\lambda,k)$ calculated according to Eq. (40) for different values of the parameter $\beta$ when $\delta=1$. (Solid line: $\beta=5$; Dash line: $\beta=10$; Dotted line: $\beta=20$)

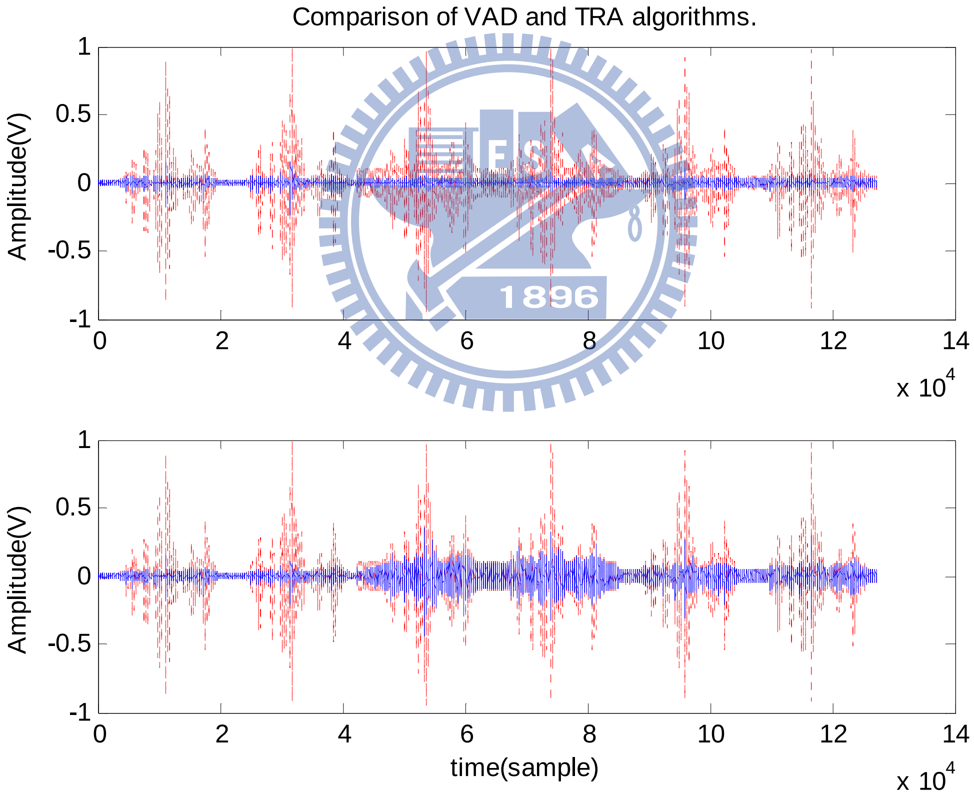


FIG. 4. Plots of the non-stationary noise (solid line) estimated using the VAD (top panel) and TRA (bottom panel) algorithms from noisy speech signal (dotted line).

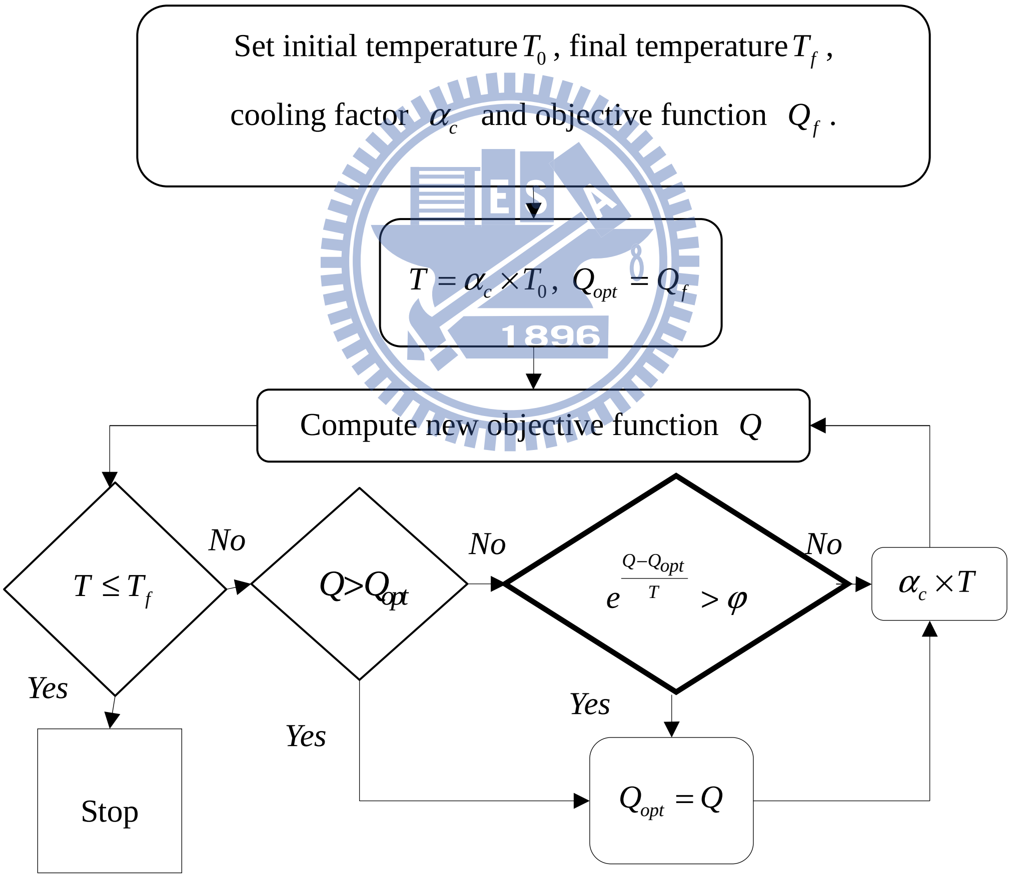

Set initial temperature $T_0$, final temperature $T_f$,
cooling factor $\alpha_c$ and objective function $Q_f$.
$T = \alpha_c \times T_0$, $Q_{opt} = Q_f$
Compute new objective function $Q$
$T \le T_f$
No
$Q>Q_{opt}$
No
$e^{\frac{Q-Q_{opt}}{T}} > \varphi$
No
$\alpha_c \times T$
Yes
Stop
Yes
Yes
$Q_{opt} = Q$

FIG. 5. The flow diagram of the SA method.

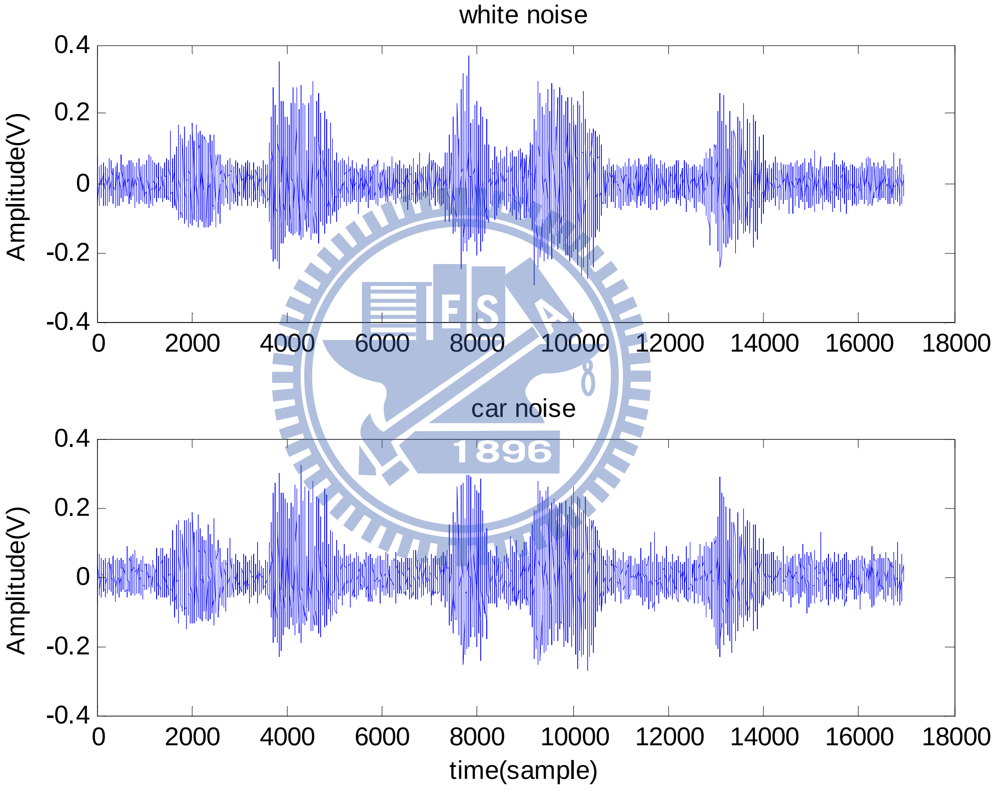
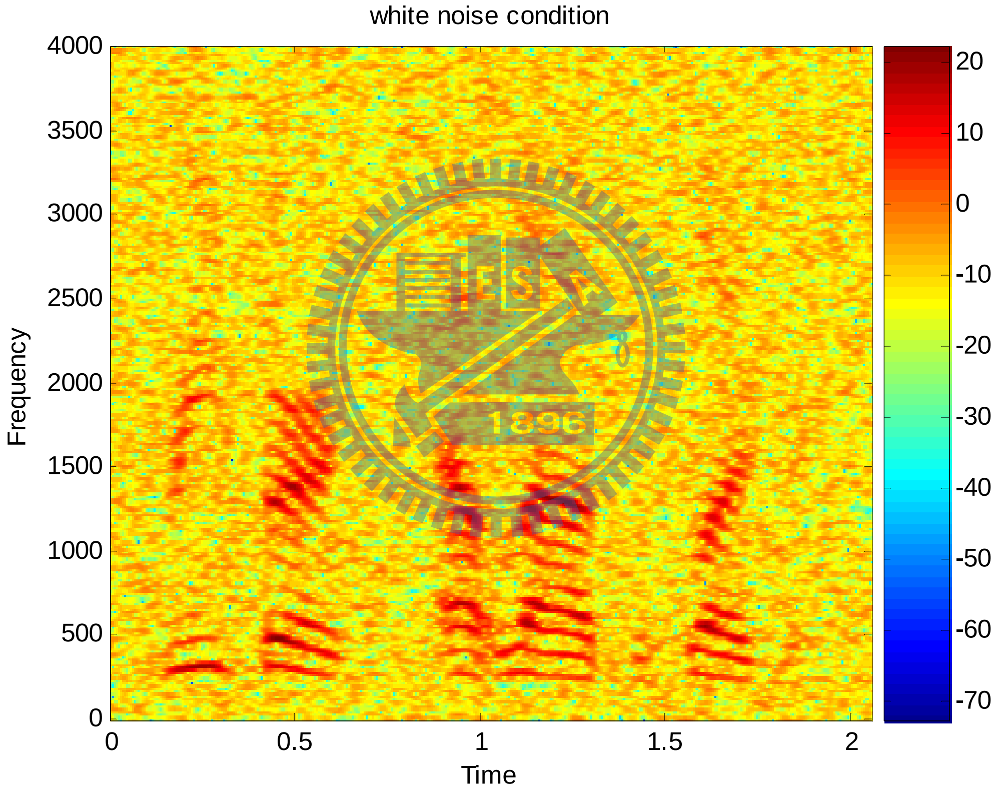
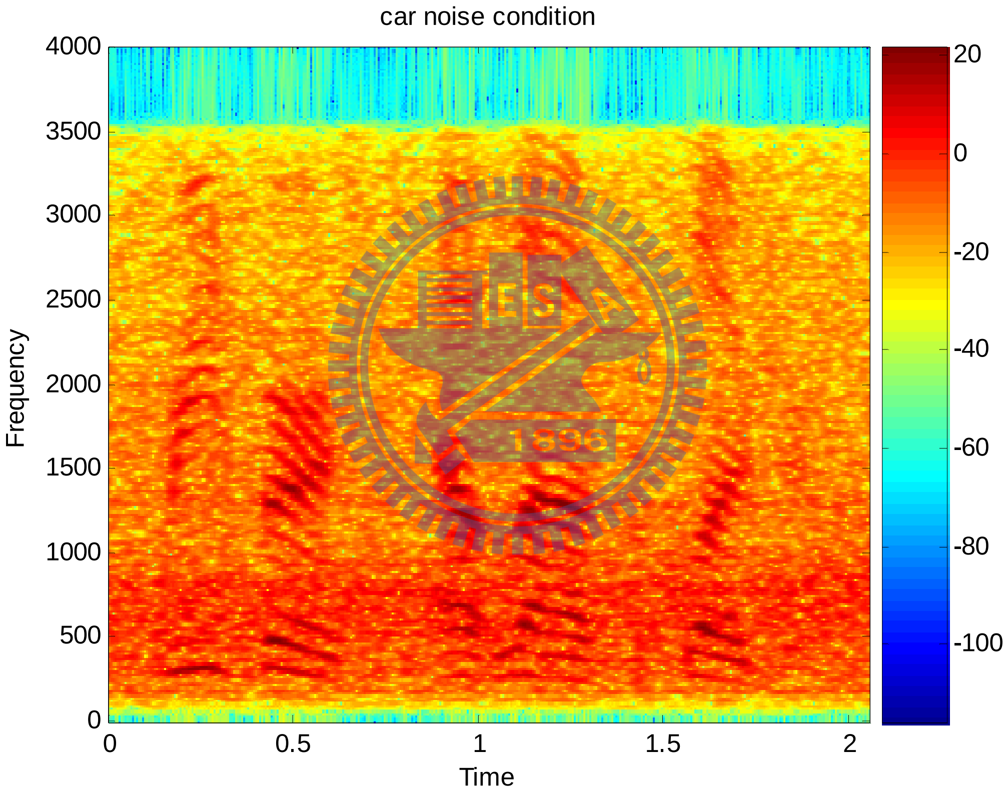


FIG. 6. The waveforms of a test sentence corrupted by white noise (top panel) and car noise (bottom panel).

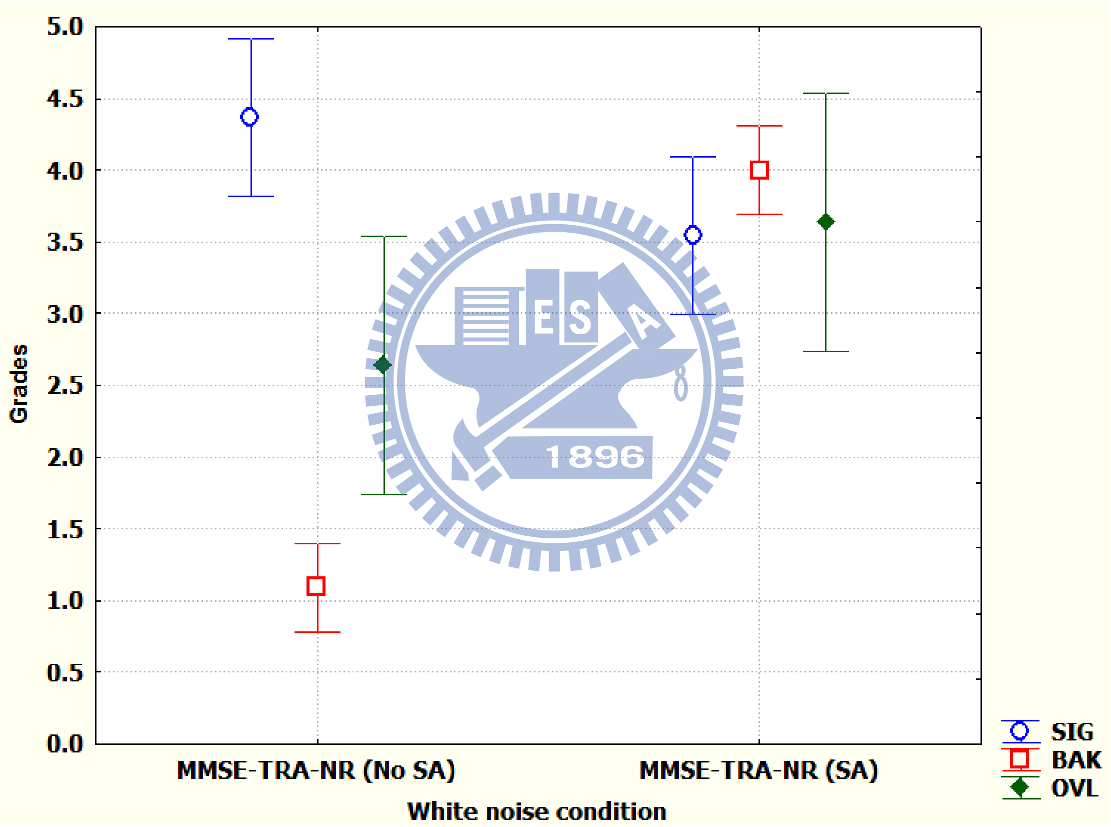

white noise condition
4000
3500
3000
2500
2000
1500
1000
500
0
Frequency
0
0.5
1
1.5
2
Time
20
10
0
-10
-20
-30
-40
-50
-60
-70

(a)

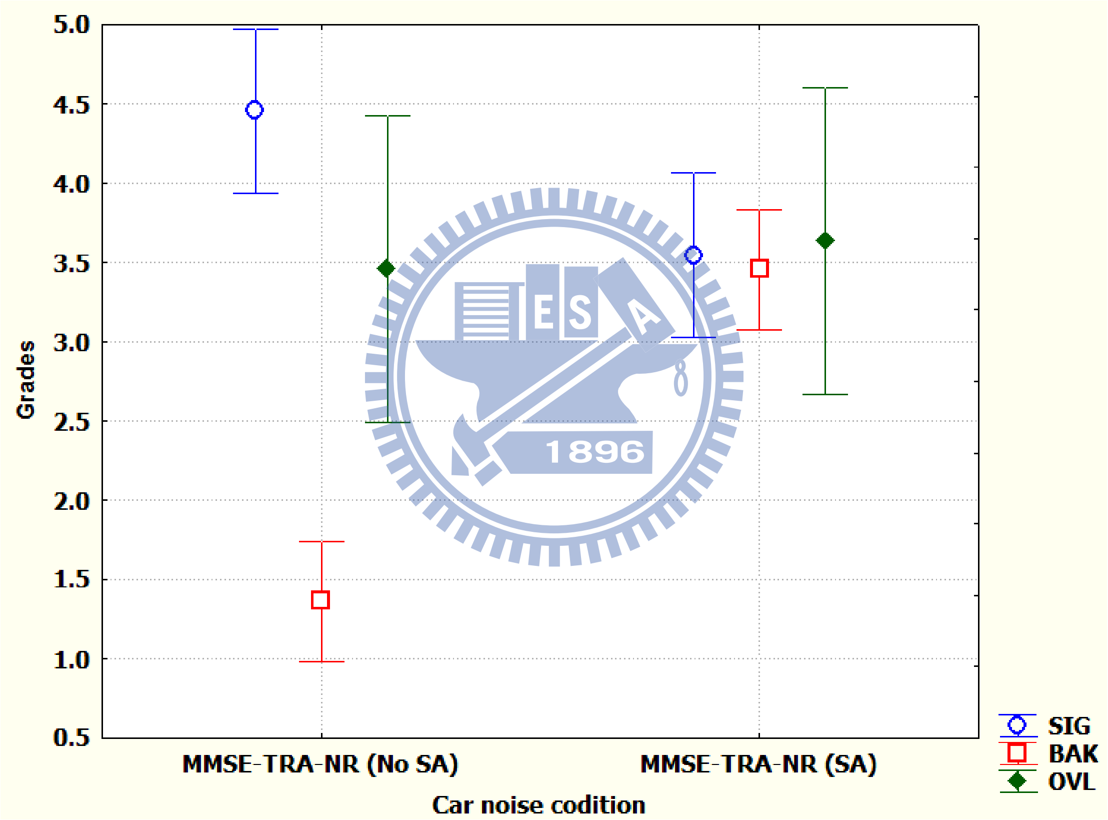


(b)

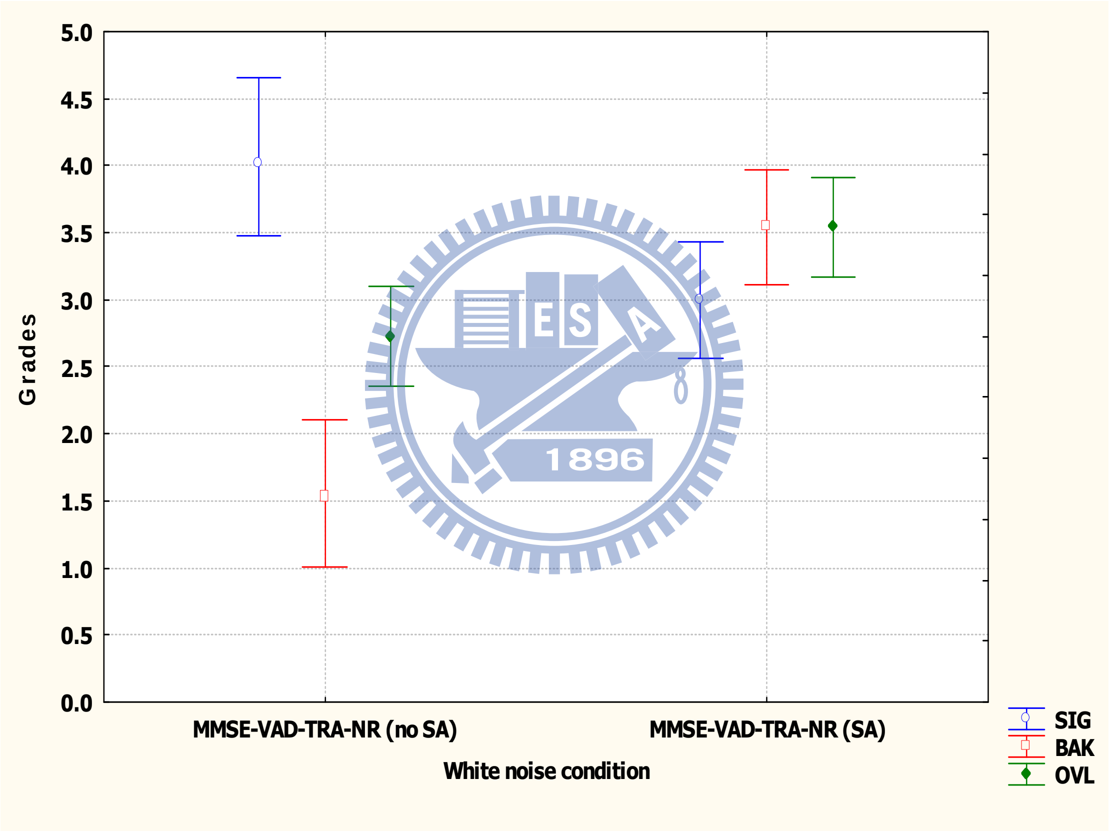
FIG. 7 The spectrograms of the test sentence corrupted by two noise conditions. (a) White noise. (b) Car noise.


5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Grades
MMSE-TRA-NR (No SA)
MMSE-TRA-NR (SA)
White noise condition
SIG
BAK
OVL

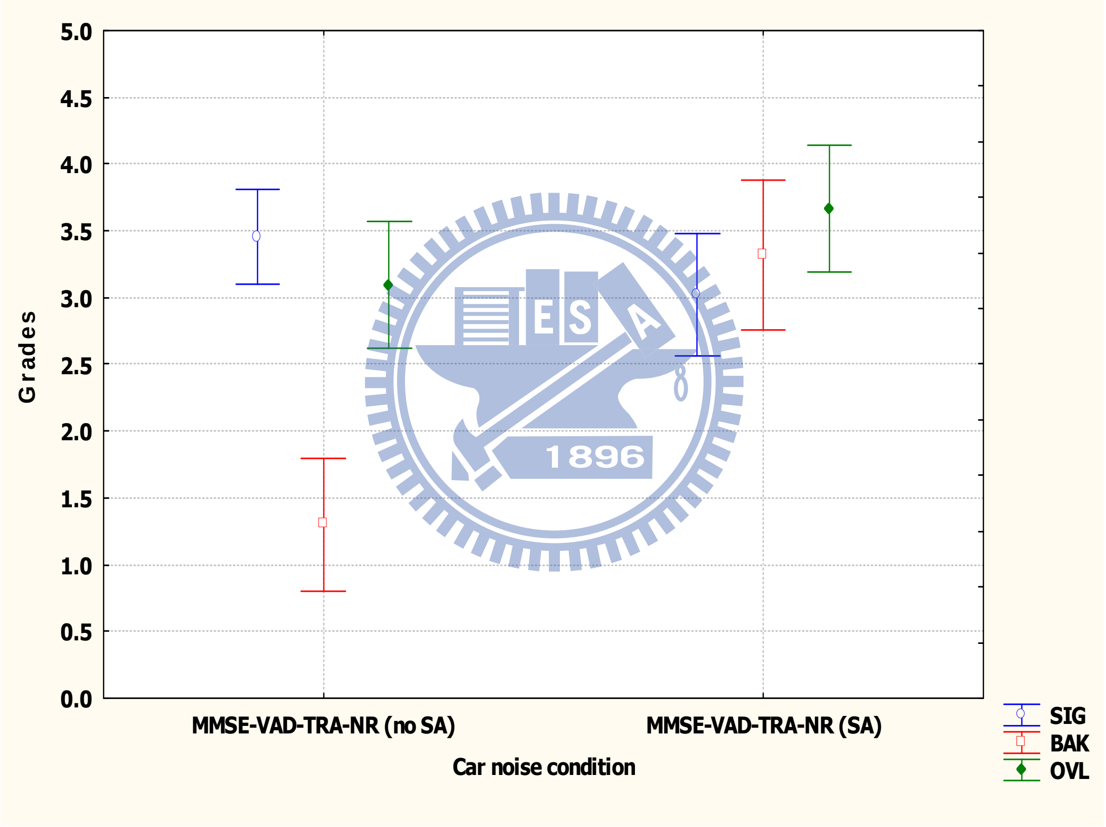
(a)


5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Grades
MMSE-TRA-NR (No SA)
MMSE-TRA-NR (SA)
Car noise codition
SIG
BAK
OVL

(b)

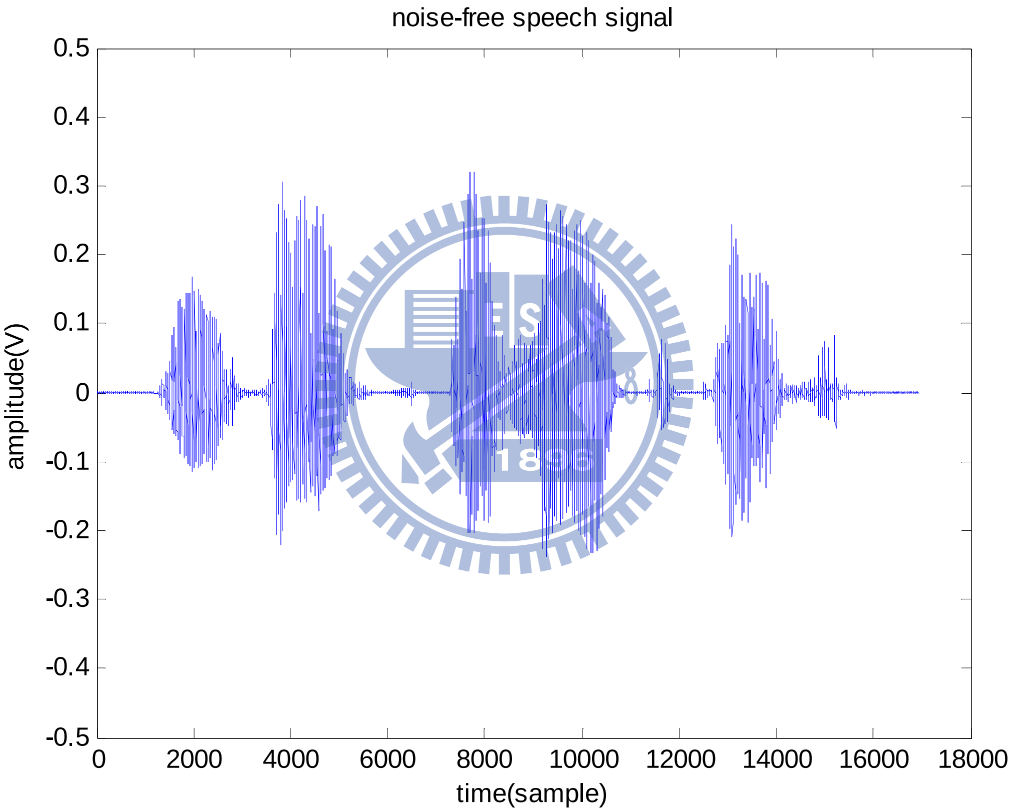

5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Grades
MMSE-VAD-TRA-NR (no SA)
MMSE-VAD-TRA-NR (SA)
White noise condition
SIG
BAK
OVL

(c)

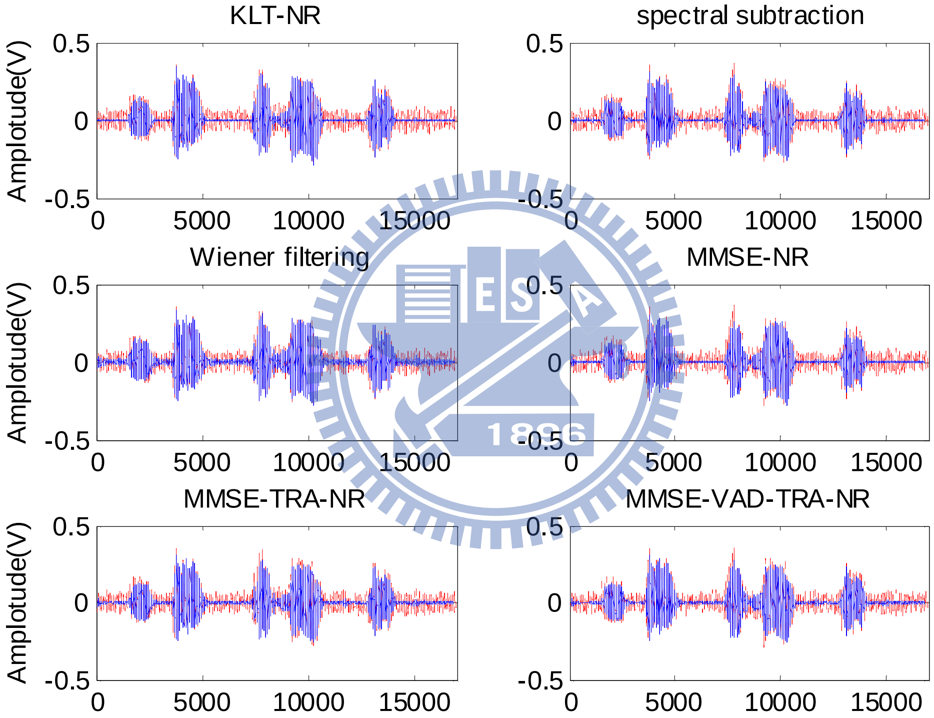


(d)

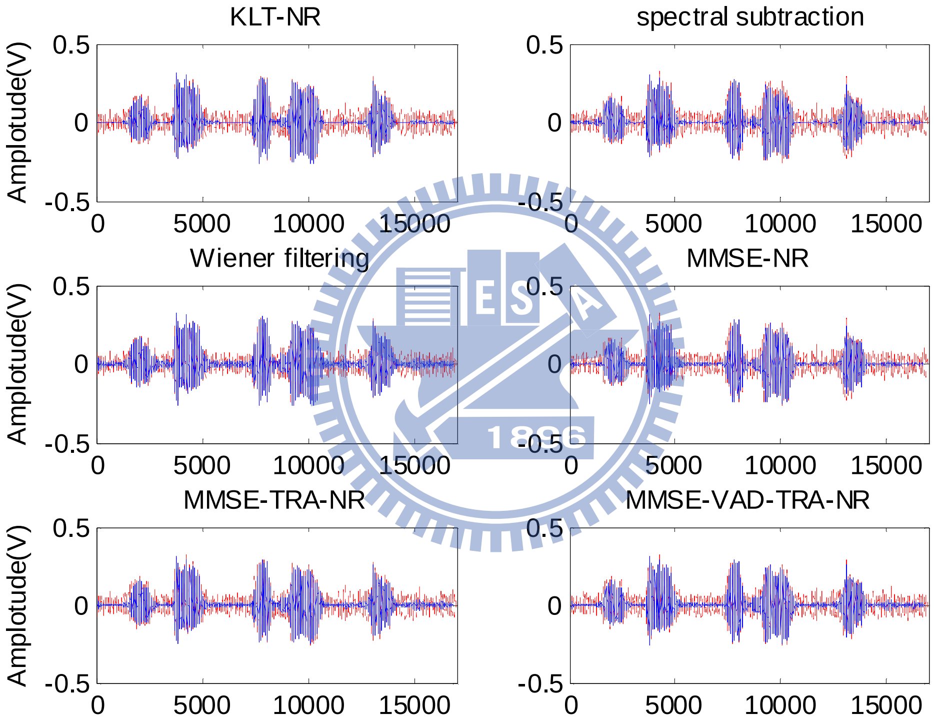
FIG. 8 The results of the listening test analyzed by using the MANOVA. (a) MMSE-TRA-NR in white noise condition. (b) MMSE-TRA-NR in car noise condition. (c) MMSE-VAD-TRA-NR in white noise condition. (d) MMSE-VAD-TRA-NR in car noise condition.


noise-free speech signal
amplitude(V)
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
time(sample)

(a)

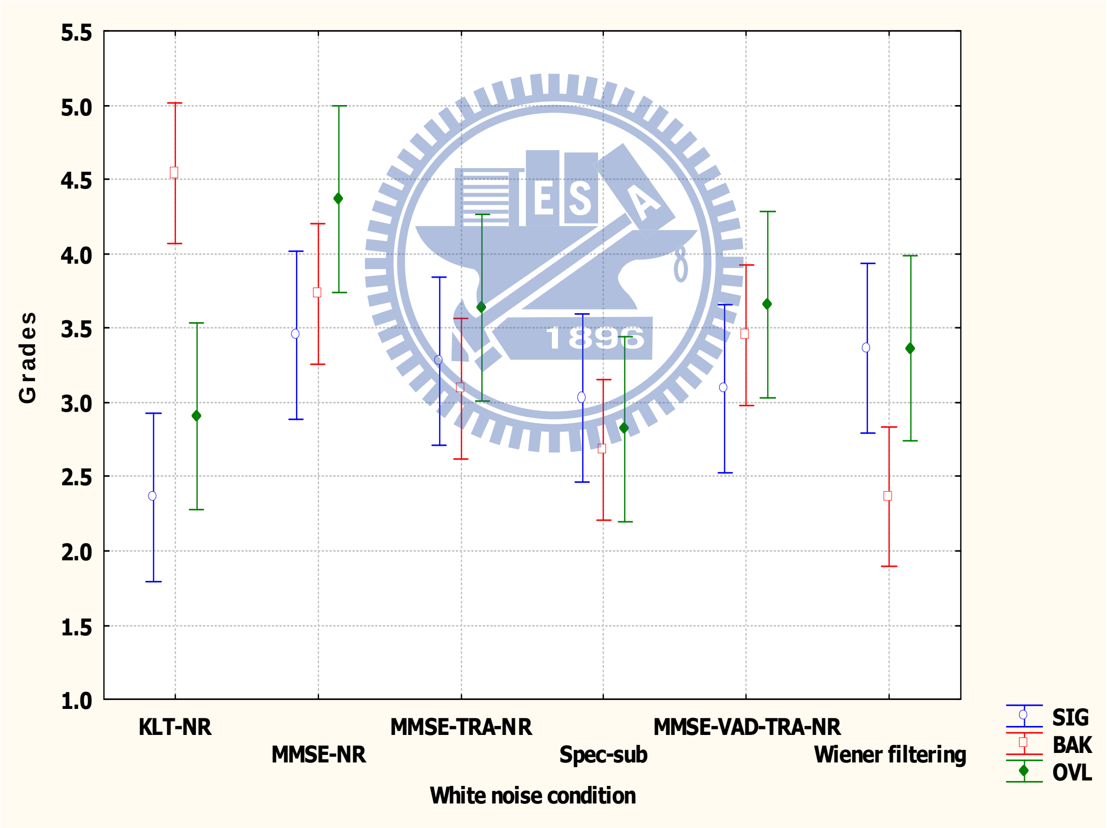

KLT-NR
spectral subtraction
Wiener filtering
MMSE-NR
MMSE-TRA-NR
MMSE-VAD-TRA-NR
Amplotude(V)
0.5
0
-0.5
0
5000
10000
15000

(b)



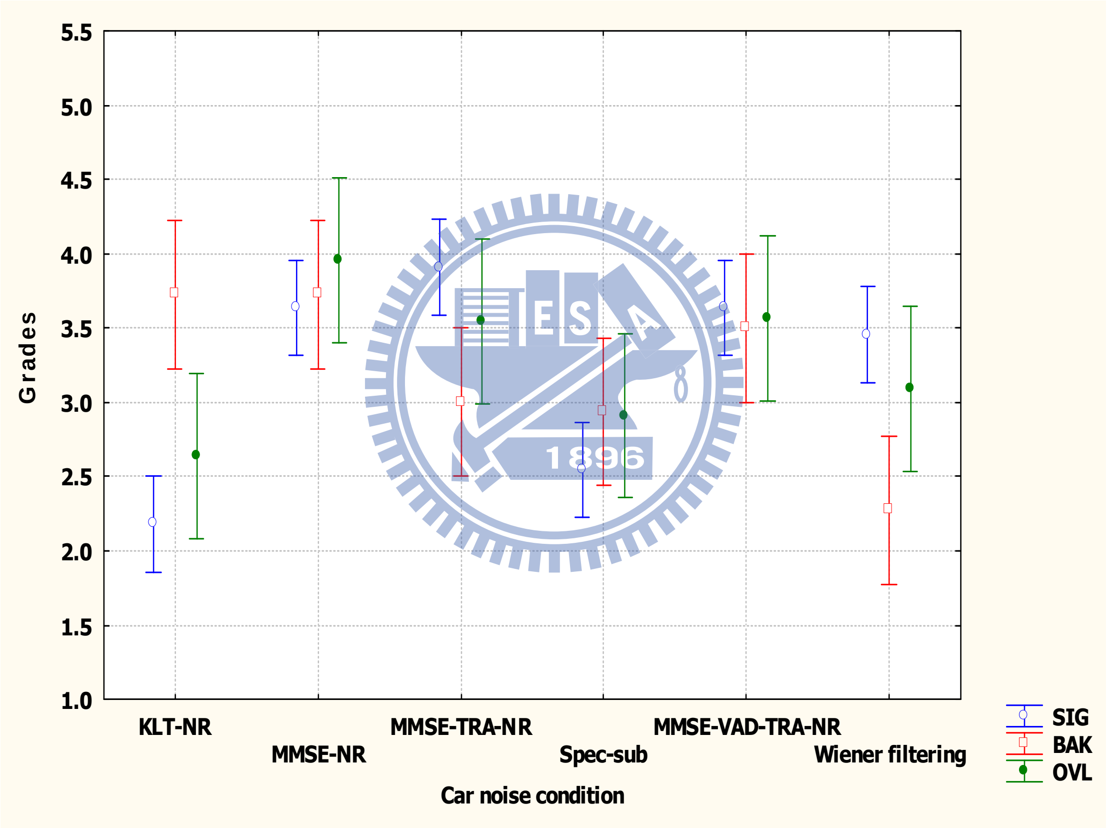
(c)

FIG. 9. Simulation results for six NR algorithms. (a) The noise-free speech signal used for a computer simulation. (b) Waveforms of the noisy and processed speech signals via six NR algorithms in white noise condition. (c) Waveforms of the noisy and processed speech signals via six NR algorithms in car noise condition. (Dotted line: noisy speech signals; Solid line: processed speech signals)



5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Grades
KLT-NR
MMSE-NR
MMSE-TRA-NR
Spec-sub
MMSE-VAD-TRA-NR
Wiener filtering
White noise condition
SIG
BAK
OVL

(a)



(b)

FIG. 10. The results of the listening test analyzed by using the MANOVA. (a) White noise case. (b) Car case.

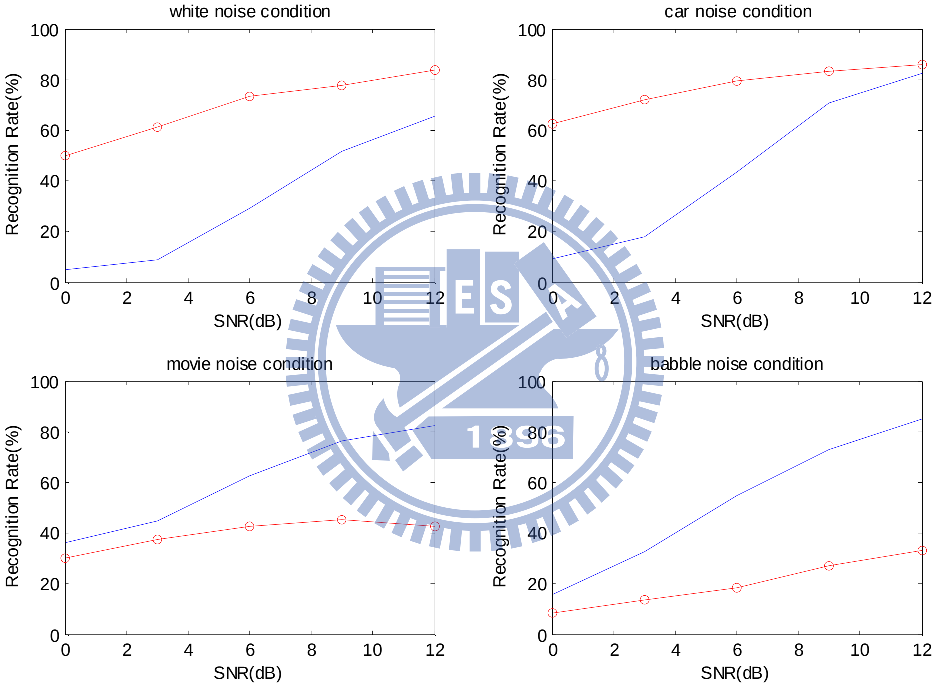


FIG. 11. The recognition rate in different noise condition and SNR level.

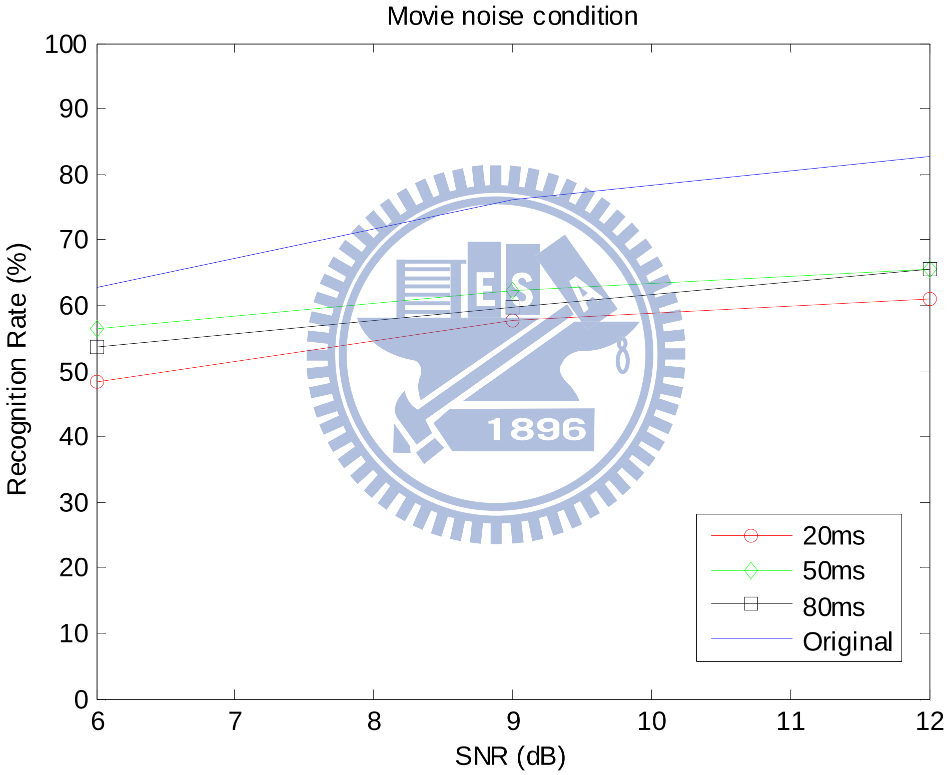


FIG. 12. The recognition rate in movie noise condition with different window length and SNR level.

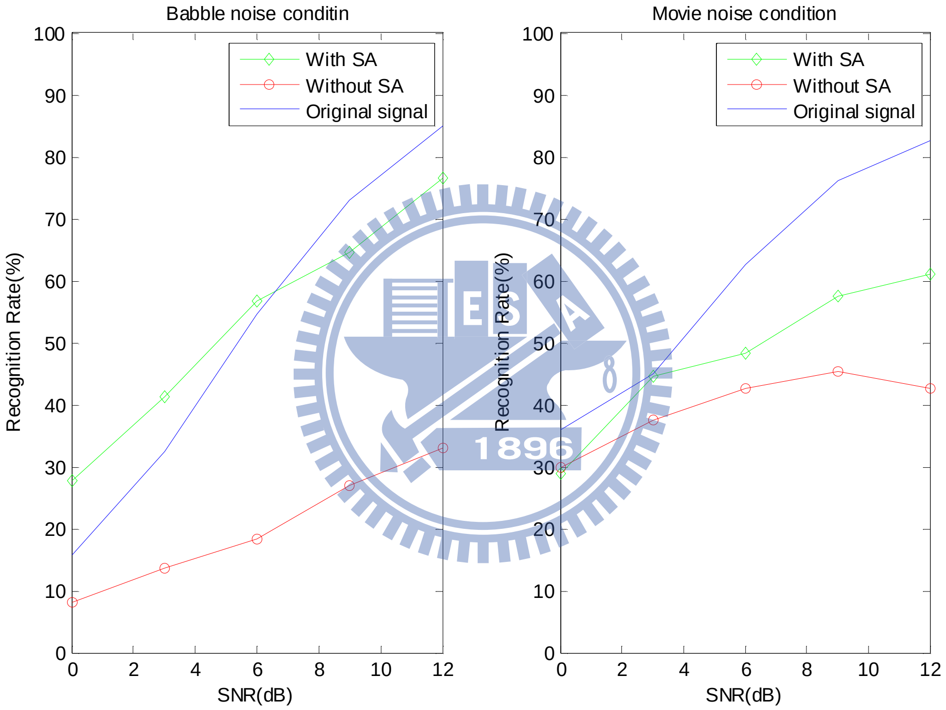


FIG. 13. The comparison of recognition rate between the with SA and without SA in the babble noise condition (the left figure) and in the movie noise condition (the right figure).

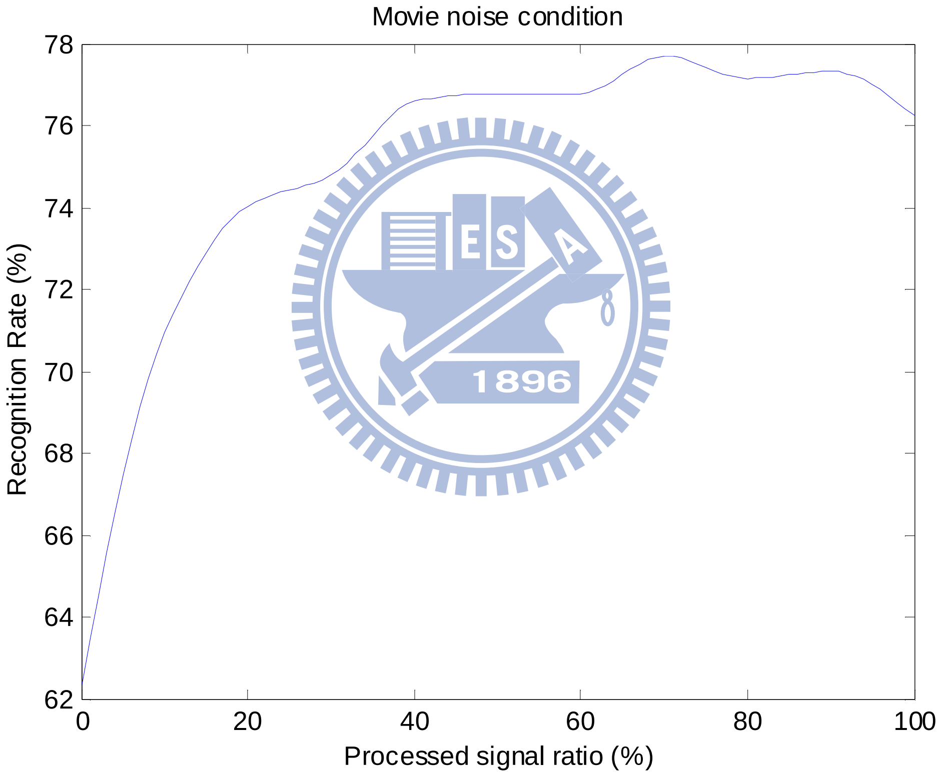


FIG. 14. The recognition rate in movie noise condition with different processed signal ratio.

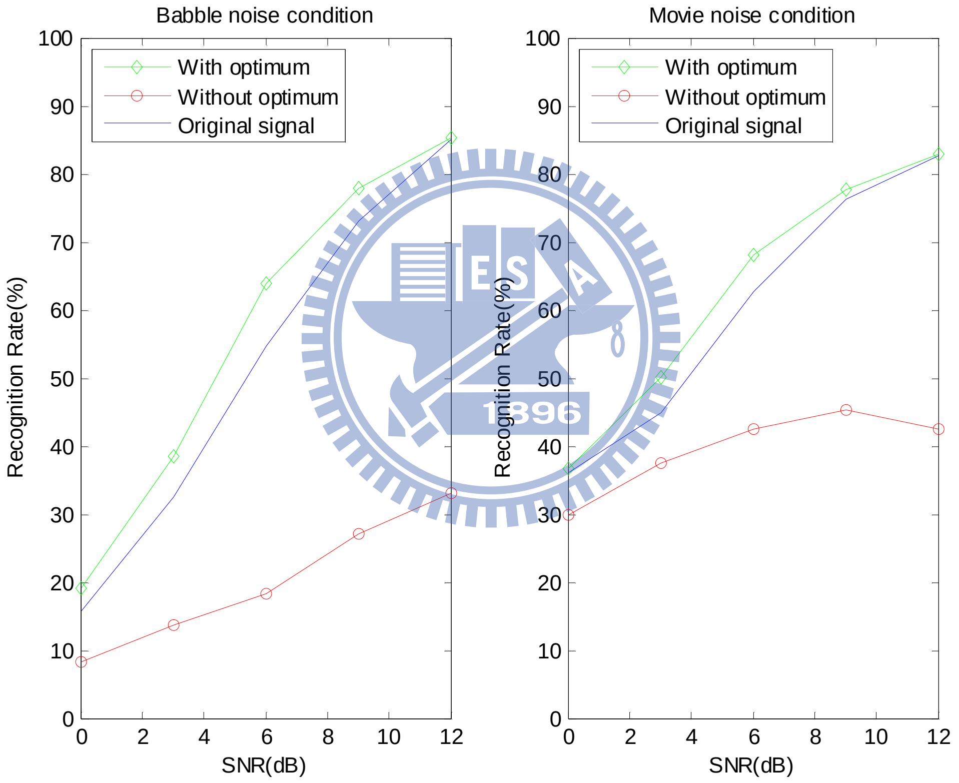


FIG. 15. The comparison of recognition rate between the with optimum and without optimum in the babble noise condition (the left figure) and in the movie noise condition (the right figure).