Journal of
Mechanical
Science and
Technology

Journal of Mechanical Science and Technology 21 (2007) 708~722

Speech Enhancement via Mel-Scale Wiener Filtering with a Frequency-wise Voice Activity Detector

Hwa Soo Kim[b], Young Man Cho[b], Han-Jun Kim[a,*]

[a]*Finetec Century, 1002, Daechi-dong Gangnam-gu, Seoul, 135-280, Korea*
[b]*School of Mechanical and Aerospace Engineering, Seoul National University San 56-1, Shillim-dong Kwanak-gu, Seoul, 151-744, Korea*

(Manuscript Received March 27, 2006; Revised March 22, 2007; Accepted March 23, 2007)

Abstract

This paper presents a speech enhancement system that enables a comfortable communication inside an automobile. A couple of novel concepts are proposed in an effort to improve two major building blocks in the existing speech enhancement systems: a voice activity detector (VAD) and a noise filtering algorithm. The proposed VAD classifies a given data frame as speech or noise at each frequency, enabling the frequency-wise updates of noise statistics and thereby improving the effectiveness of the noise filtering algorithms by providing more up-to-date noise statistics. The celebrated Wiener filter is adopted in this paper as the accompanying noise filtering algorithm, which results in significant noise suppression. Yet, the musical noise present in most Wiener filter-based systems prompts the idea of applying the Wiener filter in the Mel-scale in which the human auditory system responds to the external stimulation. It turns out that the Mel-scale Wiener filter creates some masking effects and thereby reduces musical noise significantly, leading to smooth transition between data frames.

Keywords: Speech enhancement; Wiener filtering; Mel-scale; Voice activity detector

1. Introduction

An automobile has become one of the most com-on and accessible transportation media. Accordingly, the time spent in an automobile has been steadily increasing. As a result, the comfort level during the passenger-to-passenger conversation or dialogue via a cellular phone in an automobile cabin has emerged as one of the benchmark criteria in evaluating the performance of an automobile. Therefore, the auto-obile manufacturers have been putting significant efforts to reduce the noise in the automobile interior (Kim and Kim, 2005; Oh and Cha, 2000).

To reflect on its importance, various approaches have been proposed for speech enhancement up until now. The following techniques summarize the speech enhancement systems in their simplest forms (em-ploying a single microphone only), although the list is not meant to be exhaustive: comb filtering, noise masking, filter-model-based approach, enhancement-by-synthesis, statistical model-based approach, spec-tral subtraction, etc. (Rogan, 1998; Malah and Cox, 1982; Van-Compernolle, 1989; Deller et al., 1993). The comb filtering picks out the periodic components of the voiced speech. As a result, its success is limited when dealing with the unvoiced speech, which often degrades the intelligibility. The noise masking masks certain sounds by providing background noise floor. Yet, noise masking performs poorly at low signal-to-noise ratio (SNR) since it does not eliminate the existing sounds (Rogan, 1998). In the filter-based app-

*Corresponding author. Tel.: +82 2 2185 7800; Fax.: +82 2 2185 7088
E-mail address: hanjkim@finetec-century.com

roach, the speech is modeled as the output of a linear filter, typically as an autoregressive (AR) process whose parameters are estimated from a noisy speech. However, the filter-based approach provides enhancement in proportion to the quality of the model, which may often be quite poor (Rogan, 1998; Deller et al., 1993). The enhancement-by-synthesis first extracts clean speech parameters from the noisy observation, and synthesizes the enhanced speech from the estimated speech parameters. The main drawback of the enhancement-by-synthesis is that its performance does not improve gracefully as SNR changes. The statistical model-based technique first builds the statistical models of speech and noise and then minimizes the expected value of the distortion measure between the clean and estimated speeches. Hidden Markov Model (HMM) is generally used for speech model, which is difficult to implement on-line due to its complexity. In comparison to the aforementioned methodologies, the spectral subtraction technique stands out by virtue of its simplicity and relatively good performance. The primary assumption is that the noise and the speech signal are uncorrelated and, accordingly, the power spectrum of noisy speech is simply a sum of speech and noise power spectra. As a result, the speech power spectrum can be obtained by subtracting the noise power spectrum from the whole power spectrum (of noisy speech). The Wiener filter is similar to the spectral subtraction technique but statistically optimal in that it minimizes the mean squared error of the speech estimate. This paper builds on the framework of the Wiener filtering and improves its performance by interjecting some novel ideas, about which more will be said later. The effectiveness of the spectral subtraction and Wiener filtering is determined by two factors: 1) how well the 'so-called' musical noise is suppressed 2) how often the noise power spectrum is updated to provide most up-to-date noise statistics.

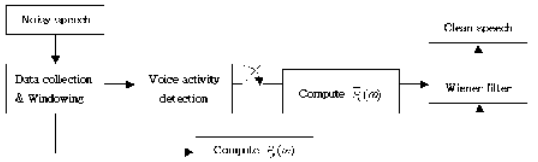
The musical noise sprouts from the inaccurate estimates of noise power spectrum, which tends to exhibit increasing variances. The conventional Wiener filtering algorithms adopt such concepts as the overestimation factor and the spectral floor as remedies for perceptually annoying musical noise (Berouti, et al., 1979; Boll, 1979). Although they mitigate the musical noise to certain extent, the conventional remedies ceases to be effective at low signal-to-noise ratio (SNR) due to the increasing speech distortion. This paper proposes that the Wiener filtering is performed in the Mel-scale. The power spectral estimates (noise and speech) are smoothed over adjacent frequency bins (Mel-filter bandwidth at the frequency of interest, to be more specific). It turns out that the human auditory system responds to the equal frequency difference in a gradually less sensitive manner as the frequency is increased. As a result, without incurring any noticeable speech distortion, the extra smoothing over adjacent frequency bins significantly reduces the variances of noise power spectral estimates (the root cause of the musical noise), which leads to smoother transition from frame to frame and subsequently much lower musical noise.

The existing speech enhancement systems provide most up-to-date noise statistics whenever a non-speech (i.e., noise) frame is detected (Martin, 1994). From the perspective of time-domain approaches, it seems impossible to provide more up-to-date noise statistics. The frequency domain perspective yet sheds light on a still better update scheme: frequency-bin-wise update of noise statistics. The proposed idea stems from the observation that even a speech frame often contains frequency bins with pre-dominant noise components. The resulting voice activity detector (VAD) is readily implemented and computationally efficient since it operates as the level of each frequency bin. Equipped with the Mel-scale Wiener filtering and the novel VAD, the proposed system has potential to enhance the quality of speech to great extent. The experimental results show that the speech quality is much improved without distorting speech in any noticeable manner.

This paper is organized as follows. Section 2 briefly introduces an existing VAD and an existing Wiener filter. In Section 3, the Mel-scale Wiener filter and the proposed VAD are explained. Section 4 presents the performance analysis of the proposed algorithm through experimental data, which shows its viability in real world applications.

2. Background

Speech enhancement systems may assume various structures depending on how enabling technologies are combined. Figure 1 shows the functional block diagram of a particular speech enhancement system, which mainly consists of a Wiener filter and a voice activity detector (VAD). This paper builds upon the structure in Fig. 1 and improves its performance by introducing a couple of novel concepts: Wiener filter-



$P_y(\omega)$: Noisy speech spectrum
$\overline{P}_n(\omega)$: Estimated noise spectrum
Fig. 1. Functional block diagram for a speech enhancement system with a voice activity detector.

ing in the Mel-scale and frequency-wise voice activity detection. The particular structure is chosen by virtue of its computational efficiency and superior performance. In the proposed structure, the stream of noisy speech data is first windowed, which is commonly denoted as a frame and is processed to generate the power spectrum of the windowed speech data. Then, a VAD determines whether the frame contains speech or noise, using such criteria as short time energy level, zero crossing rate, etc. Whenever a data frame is classified as a "noise" frame, the "noise-only" power spectrum is updated based upon the current and previous noise power spectra. Otherwise, a Wiener filter manipulates the noisy speech power spectrum and most up-to-date noise-only power spectrum estimate which in turn produces an enhanced speech. It should be noted in Fig. 1 that the frames are typically overlapped in order to ensure smooth transition.

2.1 Existing voice activity detectors

A VAD plays an important role in a speech enhancement system based on the Wiener filtering since the quality of the estimated noise power spectrum essentially determines the performance of the Wiener filtering (Deller et al., 1993). Although it may be carried out in a straightforward manner when no noise is present, the task of detecting speech activity under background noise does not render any readily deployable solution. Even when the background noise is negligible, the complex nature of speech signals makes it difficult to discern speech from noise, as explained in what follows. Speech can be classified into two distinct categories: 'voiced' and 'unvoiced'. The voiced speech is generated through the vibrations of the vocal cord. Normally, the voiced speech is modeled as the output of a slowly time-varying linear system excited by a quasi-periodic pulse signal. On the contrary, the unvoiced speech does not require the vibrations of the vocal cord. The unvoiced speech is generated by forming a constriction at certain point in the vocal tract and passing air through the constriction at high velocity (Deller et al., 1993; Rabiner and Schafer, 1978). Spectral analysis shows that the former generally contains larger energy than the latter, while the latter exhibits more high frequency components than the former. Among various criteria for speech activity detection such as the short time energy level, the zero-crossing rate and cepstral coefficients etc., an existing methodology described in this section relies on the short time energy level in combination with the zero-crossing rate, which are popularly employed in practice by virtue of their computational efficiency and straightforward implementation (Deller, et al., 1993). The short time energy level is suitable for voiced speech signal, while the zero-crossing rate is effective for unvoiced speech signal. The short time energy E_k and the zero-crossing rate Z_k in the k^{th} frame are given as

$$E_k = \frac{1}{N}\sum_{i=1}^{N} x_k^2(t_i) \tag{1}$$

$$Z_k = \frac{1}{2}\sum_{i=1}^{N-1} \{\left|\text{sgn}[x_k(t_{i+1})] - \text{sgn}[x_k(t_i)]\right|\}$$
$$\text{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{2}$$

where t_i is the time index from t_1 to t_N, $x_k(t_i)$ is the sampled noisy speech signal at t_i in the k^{th} frame and N is the number of samples in a frame. Once E_k and Z_k are computed, the following two criteria determine the presence of speech activity in the k^{th} frame when either one of the following two conditions is satisfied:

$$E_k > T_{k,e} \tag{3}$$

$$(Z_k - T_{k,l}) \cdot (Z_k - T_{k,h}) > 0, \quad T_{k,l} < T_{k,h} \tag{4}$$

where $T_{k,e}$, $T_{k,l}$ and $T_{k,h}$ are the threshold on the energy, and the low and high thresholds on the zero-crossing rate in the k^{th} frame, respectively. The inequality (3) is based on the observation that speech has much more energy than noise. The inequality (4) comes from the fact that speech contains more high frequency components than noise, which results in higher zero-crossing rate (Deller et al., 1993). It must be noted that the success of the current speech

activity detection scheme relies heavily on $T_{k,e}$, $T_{k,l}$ and $T_{k,h}$. As a result, special attention must be paid to obtain their appropriate values. These thresholds are typically updated only when no speech activity is detected in the frame k according to the following equations (Rabiner and Schafer, 1978):

$$\begin{aligned} T_{k+1,e} &= p_e \cdot T_{k,e} + (1-p_e)\cdot E_k \\ T_{k+1,l} &= p_l \cdot T_{k,l} + (1-p_l)\cdot Z_k \qquad \text{if} \quad Z_k < T_{k,l} \\ T_{k+1,h} &= p_h \cdot T_{k,h} + (1-p_h)\cdot Z_k \qquad \text{if} \quad Z_k > T_{k,h} \end{aligned} \tag{5}$$

where p_e, p_l and p_h are forgetting factors (0< p_e, p_l, p_h <1). When a frame contains speech, the thresholds $T_{k,e}$, $T_{k,l}$ and $T_{k,h}$ maintain values from the previous frame (Rabiner and Schafer, 1978). This existing approach detects the presence of speech fully in the time domain. Consequently, the detection results are obtained at the level of frames and accordingly the noise statistics are updated when an entire frame is detected as "noise".

It seems inevitable that a speech enhancement system with this type of VAD suffers from out-of-date noise statistics. Indeed, when the detection rate drops at low SNR, the effectiveness of the speech enhancement system with the VAD in this section turns out to be significantly degraded. Since the musical noise sprouts from the discontinuities during transition from one speech frame to the adjacent one, the noise power spectrum must be updated as often as possible (barring from computational burden) in order to reflect on the time-varying nature of the noise statistics and to provide the most accurate noise spectral information.

2.2 Existing wiener filtering algorithms

Speech is assumed to be corrupted by an uncorrelated additive noise (Martin, 1994). During the initialization period when no speech is present, a VAD estimates the noise power spectrum at each frequency in the frame. With the noise parameter at hand, the Wiener filtering is applied to the subsequent noisy speech. Summarizing the previous assumptions and formulations yields (Martin, 1994)

$$x_k(t_i) = s_k(t_i) + n_k(t_i) \tag{6}$$

$$X_k(\omega_i) = S_k(\omega_i) + N_k(\omega_i) \tag{7}$$

$$P_{x,k}(\omega_i) = P_{s,k}(\omega_i) + P_{n,k}(\omega_i) \tag{8}$$

where $x_k(t_i)$, $s_k(t_i)$ and $n_k(t_i)$ denote the sampled noisy speech signal, the sampled speech signal and the sampled noise signal in the k^{th} frame. X_k, S_k and N_k are the short time Fourier transforms of x_k, s_k and n_k in the k^{th} frame respectively. ω_i is the i^{th} frequency where $i = 1,\cdots,(N/2-1)$ and T_s is the sampling rate. $P_{x,k}$, $P_{s,k}$ and $P_{n,k}$ are the power spectral densities (PSD) of x_k, s_k and n_k in the k^{th} frame, respectively. Then, the Wiener filter H_k in the k^{th} frame is given as

$$H_k = 1 - \frac{P_{n,k}}{P_{x,k}} \tag{9}$$

Since the noise power spectrum $P_{n,k}$ is directly unavailable, the estimate of the noise PSD $\overline{P}_{n,k}$ obtained by a VAD replaces $P_{n,k}$. Moreover, since the sudden change of the estimated noise PSD causes undesirable filtering results such as musical noise and speech distortion, it is computed recursively at each frequency between noise frames according to

$$\overline{P}_{n,k}(\omega_i) = \gamma \cdot \overline{P}_{n,k-1}(\omega_i) + (1-\gamma)\cdot P_{x,k}(\omega_i) \tag{10}$$

where γ is a forgetting factor ranging generally from 0.9 to 0.98 (Martin, 1994). When the k^{th} frame contains speech, $\overline{P}_{n,k}$ remains identical to $\overline{P}_{n,k-1}$. Since the power spectrum of the filtered speech signal is nonnegative, it seems reasonable to turn (9) into

$$H_k = \max\{\ 1 - \frac{\overline{P}_{n,k}}{P_{x,k}},\ 0\ \} \tag{11}$$

However, it is well-known that the direct implementation of (11) inevitably introduces the musical noise (Berouti et al., 1979; Boll, 1979). This perceptually annoying noise is composed of tones at random frequencies and has an increasing variance, which is caused by the incorrect estimate of the noise power spectrum. In order to reduce the musical noise, the overestimation factor α and the spectral floor β are typically introduced (Berouti et al., 1979). The overestimation factor α overestimates the noise power spectrum and increases the amount of the noise power spectrum subtracted from the power spectrum of noisy speech at low SNR. As a result, the peaks of tones at random frequencies disappear to certain extent.

The spectral floor determines the noise level remaining in the filtered speech signal. Setting the spectral floor to a certain non-zero value leads to masking the tones at random frequencies. With the help of these two parameters, the peaks and the valleys existing randomly in the frequency domain disappear considerably, which results in the modified Wiener filter

$$H_k = \max\{\ 1-\frac{\alpha\cdot\overline{P}_{n,k}}{P_{x,k}},\ \beta\cdot\overline{P}_{n,k}\ \} \tag{12}$$

In general, α is determined according to SNR in the k^{th} frame: $\alpha=\alpha_k$. For high SNR, α_k is close to 1 so that the noise power spectrum is hardly overestimated. Otherwise, α_k is much greater than 1 and the noise power spectrum is considerably overestimated. However, such modification of the Wiener filter gives rise to another problem, speech distortion at low SNR, which is addressed in the following section.

3. Mel-scale wiener filter with the frequency-wise VAD

Although the existing noise filtering approaches based on the Wiener filtering are easily implemented and effectively reduce the noise present in the corrupted speech signal, there still exist unresolved shortcomings: musical noise and speech distortion. While the former is the innate problem of the Wiener filter-based approach, the latter is a secondary problem stemming from an effort to reduce the musical noise. Excessive subtraction of the noise power spectrum would incur speech distortion, while insufficient subtraction would leave the noise unfiltered. The better the noise power spectrum is estimated, the less the musical noise remains in the filtered speech spectrum and the less the speech distortion results. Yet, since the noise power spectrum may not be promptly updated in an existing VAD, there may exist significant gap between the estimated noise power spectrum and the instantaneous noise power spectrum in a given data frame, which causes isolated tones (musical noise) with large variance to appear at random frequencies. The problem of the musical tones is tackled in two steps: a new algorithm is first proposed for VAD that leads to more accurate and up-to-date estimate of the noise power spectrum. Then, a novel Wiener filtering algorithm reduces the musical noise by performing filtering in the so-called Mel-scale (Deller et al., 1993; Rabiner and Juang, 1993).

3.1 Frequency-wise voice activity detector

Existing VADs detect voice activity using time-domain-based criteria such as the short time energy and zero-crossing rate. As a result, the detection is made on an individual time frame and correspondingly existing Wiener filtering algorithms perform noise filtering according to the frame-based information. Yet, it must be noted that even within the same frame, it is possible to contain speech and noise in certain frequency bands while pure noise occupies other frequency bands. In this respect, the changes in the noise statistics may not be updated on time based on the existing VADs. To overcome such a shortcoming, a novel VAD is proposed which detects the voice activity at each frequency within a given frame and, as a result, produces more up-to-date noise statistics or power spectrum.

It is a reasonable assumption that there is only noise during initialization period and that the noise power spectrum at each frequency bin is an independent and identical distributed (IID) random variable (RV) (Martin, 1994). The central limit theorem (CLT) states that the average of the noise power spectrum at each frequency bin becomes the Gaussian process as time goes on (Papoulis, 1981): for $j=1,2,\cdots,k$ and IID RVs $P_{n,j}(\omega_i)$ (noise power spectrum at ω_i), their average $z\equiv\overline{P}_{n,k}(\omega_i)=\frac{1}{k}\sum_{j=1}^{k}P_{n,j}(\omega_i)$ may be approximated as an Gaussian RV with the mean m and variance σ, given by

$$m=\frac{1}{k}\sum_{j=1}^{k}\mathrm{E}[P_{n,j}(\omega_i)] \tag{13}$$

$$\sigma^2=\frac{1}{k}\sum_{j=1}^{k}\mathrm{var}[P_{n,j}(\omega_i)]. \tag{14}$$

The probability density function (PDF) of the noise power spectrum at each frequency bin can be determined as

$$f(z)\equiv\frac{1}{\sigma\sqrt{2\pi}}e^{-(z-m)^2/2\sigma^2}. \tag{15}$$

Equipped with the statistical characteristics of noise power spectrum described by the mean m and variance σ, a novel scheme is proposed to detect

voice activity. During the initialization period from the 1st frame to the k^{th} frame, the mean and variance of the noise power spectrum at each frequency bin are calculated by (13) and (14). Though the average of the noise power spectrum can be approximated by a Gaussian RV, the power spectrum of the following frame even corresponding to noise may not be a Gaussian RV. Therefore, the threshold must cope with this irregularity and the threshold for detecting the presence of speech at the $(k+1)^{th}$ frame is set by

$$Th_k(\omega_i) = m_k(\omega_i) + \lambda \cdot \sqrt{\sigma_k(\omega_i)} \tag{16}$$

where $m_k(\omega_i)$ and $\sigma_k(\omega_i)$ are the mean and variance of the noise power spectrum at ω_i and λ is a constant larger than 1 to be determined. With this threshold, the new VAD detects the voice activity in the following manner.

If the calculated power spectrum $P_{x,\,k+1}(\omega_i)$ at ω_i in the $(k+1)^{th}$ frame is larger than the threshold $Th_k(\omega_i)$ at the corresponding frequency bin, this frequency bin is classified as speech. Otherwise, it is classified as noise. When noise is detected at some other frequencies, the corresponding noise power spectrum is updated at each frequency bin by

$$\overline{P}_{n,\,k+1}(\omega_i) = \gamma \cdot \overline{P}_{n,\,k}(\omega_i) + (1-\gamma) \cdot P_{x,\,k+1}(\omega_i) \tag{17}$$

where γ is a forgetting factor (<1) which guarantees the smoothed noise power spectrum and compensates for the uncertainty of the noise probability model (Berouti et al., 1979). When the speech is present, γ is set to be 1 in (17). Also, noise parameters such as the mean and variance of the noise power spectrum are is calculated by the following recursive equations (Papoulis, 1981)

$$\begin{aligned} m_{k+1}(\omega_i) &= m_k(\omega_i) + \frac{(\overline{P}_{n,\,k+1}(\omega_i) - m_k(\omega_i))}{k} \\ \sigma_{k+1}(\omega_i) &= (1-\frac{1}{k})\sigma_k(\omega_i) + \frac{1}{k}(1-\frac{1}{k})(\overline{P}_{n,\,k+1}(\omega_i) \\ &\quad - m_k(\omega_i))^2. \end{aligned} \tag{18}$$

Compared with existing VADs, the proposed VAD can detect the voice activity in the frequency-wise manner even within a frame and provide more up-to-date noise statistics crucial to the noise filtering algorithm. Furthermore, the frequency-wise VAD can offer various options to the noise filtering algorithm. Unlike the existing Wiener filters, the overestimation factor α and the spectral floor β in (12) may be chosen to be different at each frequency even within a frame so that $\alpha = \alpha_k(\omega_i)$ and $\beta = \beta_k(\omega_i)$. It should be noted that all these advantages do not entail any additional computational burden. The major computational burden in the existing VAD is to compute the Eqs. (1) and (2), which amounts to O(6N) flops for the number of samples N in a data frame while (16), (17) and (18) requires O(7N) flops in the proposed VAD. The order of the flops counts remains unchanged and the increase in computation burden is minimal, if at all, which does not affect the real-time realization of the proposed VAD.

It is worthwhile to compare the proposed algorithm to the one based on the minimum statistics by Martin (Martin, 1994; Martin, 2001). It is based on the observation that the power spectrum of a noisy speech signal exhibits distinct peaks that correspond to speech and valleys that correspond to noise. It uses so-called the minimum floor to estimate the noise power spectrum and introduces a factor to reduce the bias between the true noise power spectrum and estimated noise power spectrum. However, for rapidly varying noise, minimum statistics always shows undesirable bias and the effectiveness of estimation considerably diminishes.

3.2 Wiener filtering in the Mel-scale

Existing Wiener filters adopt the overestimation factor α_k and the spectral floor β in order to reduce the musical noise. Since the peaks and the valleys in the estimated speech spectrum cause the musical noise, the overestimation factor α_k tend to eliminate the broadband peaks, while the spectral floor β fills the valleys; they together render the residual noise "perceptually white" and create some masking effects (Berouti et al., 1979; Boll, 1979). Granted that they attenuate the musical noise to certain extent, these two parameters entail some undesirable side effects such as speech distortion. To reduce these side effects, the transfer functions of the Wiener filter are generally smoothed over the adjacent noise frames in the existing approaches (Boll, 1979; Martin, 1994): if at the frequency ω_i, the $(k+1)^{th}$, … $(k+q)^{th}$ frames are consecutively identified as noise frames, the smoothed Wiener filter in these frames becomes

$$\bar{H}_{k+j+1} = \mu \cdot \bar{H}_{k+j} + (1-\mu) \cdot H_{k+j+1}, \quad \bar{H}_{k+1} = H_{k+1}, \; j = 1, \ldots, q-1 \tag{19}$$

where $\bar{H}_{k+j+1}$ and $\bar{H}_{k+j}$ denote the smoothed Wiener filters in the $(k+j+1)^{\text{th}}$ frame and the $(k+j)^{\text{th}}$, respectively (Martin, 1994). H_{k+j+1} is the existing Wiener filter in the $(k+j+1)^{\text{th}}$ frame. μ is a forgetting factor (<1). When the transfer function of the Wiener filter over the adjacent noise frames is smoothed, the variance of the Wiener filter transfer function is greatly reduced and the speech distortion will be somewhat lessened. However, this approach has its own limitation since the smoothing takes place only at the frequencies where the consecutive frames contain noise.

In order to take advantage of the smoothing effect further, a novel idea of smoothing is proposed, where the smoothing is applied over the adjacent frequencies in addition to adjacent frames. The proposed idea utilizes the psychophysical characteristics of the human auditory system, which is known to respond in the Mel-scale. The Mel-scale is the frequency scale that approximates the sensitivity of the human auditory system to differences in two frequencies. It is well-known that the human auditory system exhibits varying sensitivity to the frequency difference as the frequency changes (Deller et al., 1993; Rabiner and Schafer, 1978; Rabiner and Juang, 1993). The relation between the Mel-scale and the physical frequency is

$$f_{Mel} = 1127 \cdot \ln\left(1 + \frac{f}{700}\right) \tag{20}$$

where f_{Mel} denotes the Mel-scale and f denotes the physical frequency (Deller et al., 1993). This relation shows that the mapping from the physical frequency to the Mel-scale is approximately linear below 1000 Hz and logarithmic above that frequency. In order to better understand the Mel-scale, two exemplary cases are shown in Fig. 2: one when the frequency changes from 500 Hz to 1000 Hz and the other when the frequency changes from 3000 Hz to 3500 Hz. Even though the difference between two changes is identical in terms of a linear scale, a human being perceives two cases differently. Actually a human being discriminates two frequencies in the first case better because the difference in the first case is larger than in the second case in the Mel-scale. Such an observation leads to an idea that the perceptual quality of the filtered speech may be enhanced by smoothing the transfer function over the adjacent frequencies where Mel-scales are similar.

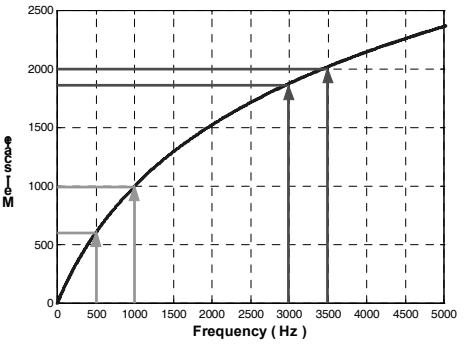


Fig. 2. Relation between Mel scale and the physical frequency.

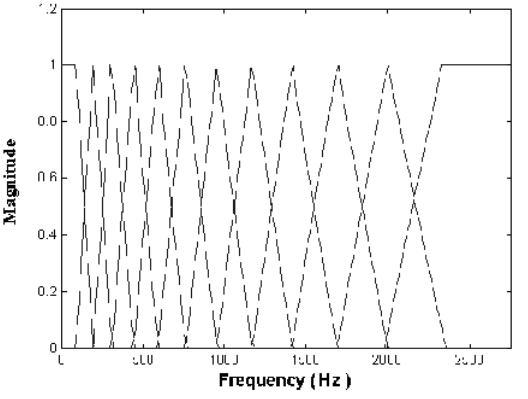


Fig. 3. Frequency response of the Mel-filter bank.

Suppose that m and T_s are the number of Mel-filters in the filter bank and the sampling rate. The first step is to calculate $f_{Mel}(T_s/2)$ Mel that is equivalent to the frequency band of interest $f = T_s/2$ by using the Eq. (20) and then, divide the Mel-scale frequency band up to $f_{Mel}(T_s/2)$ into m sub-bands of the same length, where the center of each sub-band is equal to the Mel-scale center frequency f_{Mel}^c of the Mel-filter in the filter bank. Finally, these Mel-scale center frequencies f_{Mel}^c are transformed into the general center frequencies f^c by using the conversion formula derived from the Eq. (20), that is, $f^c = 700 \cdot \left[\exp\left(f_{Mel}^c / 1127\right) - 1 \right]$ and as a result, the multiple sub-frequency bands B_j, $j = 1, \cdots, m$ are obtained: for each sub-band B_j, the Mel-filter f_j is designed to have a triangle band-pass frequency response as shown in Fig. 3, which corresponds to the case of m =12 and T_s =5.5 kHz. Note that there is an overlap between adjacent sub-bands B_j and B_{j+1} to minimize the energy loss and to guarantee the smooth transition. The bandwidth of

the Mel-filter becomes large as the physical frequency increases because a human auditory system is more sensitive at lower frequency in the Mel-scale. The first and last windows assure special shapes to minimize the energy loss. Then, the Mel-scale-based averaging operator Mel_f for any frequency response h with $\omega_i \in B_j$, for some $j \in \{1,\cdots,m\}$ may be defined by

$$Mel\_f(h)(\omega_i) = \sum_{\substack{i=1 \\ \omega_i \in B_j}} f_j(\omega_i) \cdot h(\omega_i) \tag{21}$$

From (12), (19) and (21), the Mel-scale Wiener filter in the k^{th} frame is given as follows

$$\hat{H}_k = Mel\_f\left[\bar{H}_k\right]. \tag{22}$$

Recall that the frequency-wise VAD enables the overestimation factor α_k to be and the spectral floor β_k to be functions of the frequency even within the same frame, *i.e.*, $\alpha_k(\omega_i)$ and $\beta_k(\omega_i)$.

The following steps summarize the Mel-scale Wiener filtering algorithm in conjunction with the proposed VAD.

[Initialization]

[I1]: Determine the sampling rate T_s, the number of samples N in a frame, λ (>1) in (16), the forgetting factors γ (<1) in (17) and μ (<1) in (19) and the number of sub-frequency band m. Design the Mel-filter f_j, $j=1$, $\cdots$, m with a triangular bans-pass response as shown in Fig. 3.

[I2] Based on the assumption that only noise is present during the initialization period from the 1st frame to the k^{th} frame, calculate the power spectra $P_{x,1}(\omega_i)$, $\cdots$, $P_{x,k}(\omega_i)$ at each frequency ω_i, $i=1,\cdots,(N/2-1)$. Recall that there is 50 % overlap between adjacent frames.

[I3] The estimate of noise power spectrum $\bar{P}_{n,1}(\omega_i)$ is equal to $P_{x,1}(\omega_i)$ and $\bar{P}_{n,k}(\omega_i)$ is computed recursively according to (17) with $\bar{P}_{n,1}(\omega_i)$, $P_{x,2}(\omega_i)$, $\cdots$, $P_{x,k}(\omega_i)$. Then, calculate the mean $m_k(\omega_i)$ and variance $\sigma_k(\omega_i)$ of noise power spectrum by (13) and (14) and construct the threshold $Th_k(\omega_i)$.

[Recursion]

[Step 1]: With 50 % overlap, calculate the power spectrum $P_{x,k+1}(\omega_i)$ of the $(k+1)^{\text{th}}$ frame.

[Step 2]: By comparing $P_{x,k+1}(\omega_i)$ with the threshold $Th_k(\omega_i)$, determine whether the spectrum $P_{x,k+1}(\omega_i)$ at each frequency ω_i contains speech or not. $Th_k(\omega_i) < P_{x,k+1}(\omega_i)$ implies that voice is present at the frequency ω_i for some $i \in \{1,\cdots,(N/2-1)\}$.

[Step 3]: For each frequency ω_i where no speech is present, the noise power spectrum $\bar{P}_{n,k+1}(\omega_i)$ is updated with $\bar{P}_{n,k}(\omega_i)$ in the **I3** step and $P_{n,k+1}(\omega_i) = P_{x,k+1}(\omega_i)$ according to (17). Then, the mean $m_{k+1}(\omega_i)$ and variance $\sigma_{k+1}(\omega_i)$ is calculated recursively by (18) and the new threshold $Th_{k+1}(\omega_i)$ for the next frame is obtained.

[Step 4]: For frequencies where speech is present, $\bar{P}_{n,k+1}(\omega_i)$ and $m_{k+1}(\omega_i)$ are set equal to $\bar{P}_{n,k}(\omega_i)$ and $m_k(\omega_i)$.

[Step 5]: Determine the overestimation factor $\alpha_{k+1}(\omega_i)$ and the spectral floor $\beta_{k+1}(\omega_i)$ differently at each frequency ω_i considering the presence of speech and the SNR.

[Step 6]: Construct the Wiener filter H_{k+1} at each frequency ω_i using (12) based on $\alpha_{k+1}(\omega_i)$, $\beta_{k+1}(\omega_i)$ and $\bar{P}_{n,k+1}(\omega_i)$ (obtained in **Step 2**) and set H_{k+1} to $\bar{H}_{k+1}$ in (19). Note that from **Step 1** to **Step 7**, the smoothed Wiener filter is calculated using (19) only for the frequency containing consecutively the noise component.

[Step 7]: Using the Mel-filter designed in the initialization step **I1**, establish the Mel-Scale Wiener filter $\hat{H}_{k+1}$ from H_{k+1} in **Step 6** by (21) and (22). Obtain the filtered signal by applying the Mel-Scale Wiener filter $\hat{H}_{k+1}$ to the $(k+1)^{\text{th}}$ frame. Recall that the filtered signal has the delay corresponding to 50 % overlap, which is tunable by changing the overlap size between the adjacent frames. Move to the $(k+2)^{\text{th}}$ frame with 50 % overlap and return to **Step 1**.

4. Performance analysis

The performances of the proposed VAD and the Mel-scale Wiener filter are evaluated in this section. Since the performance of the Mel-scale Wiener filter depends on that of the VAD, the proposed VAD is considered first. The proposed VAD determines the presence of speech at the frequency level within a given data frame, enabling the frequency-wise updates of noise statistics while existing VADs distinguish between speech activity and pause on the entire data frame so that noise statistics are updated only at the frame level. An idea based on the minimum

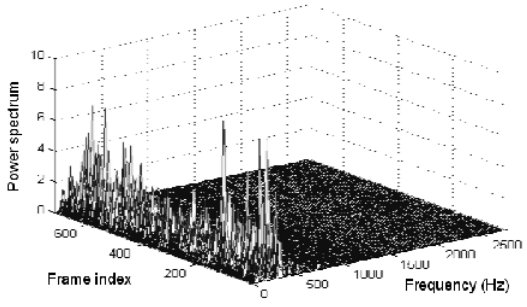
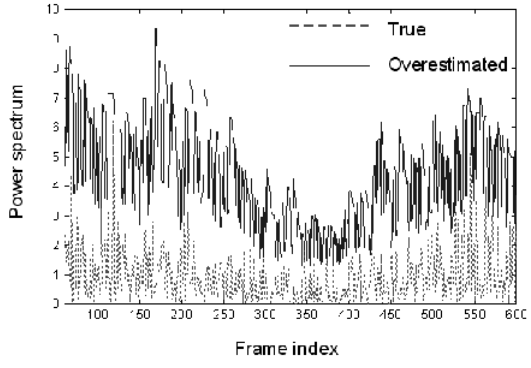
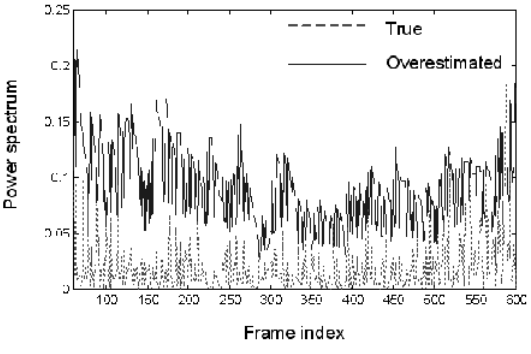


Fig. 4. The power spectrum of automobile interior noise.



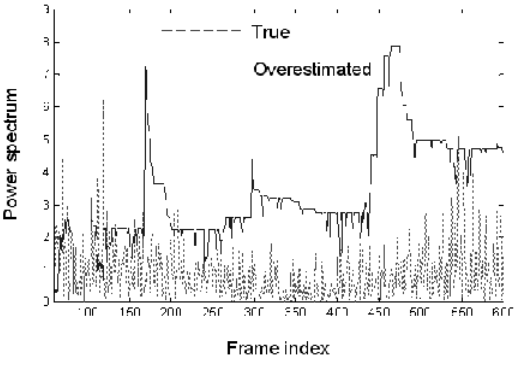
(a)



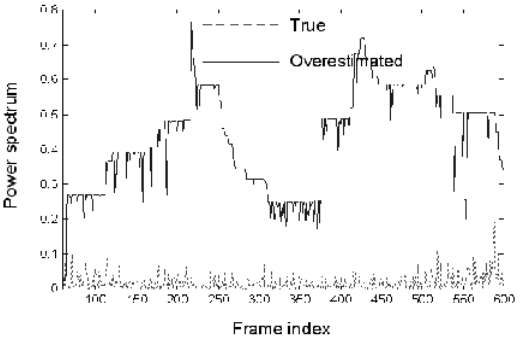
(b)

Fig. 5. Real and estimated noise power spectra using the proposed voice activity detector: (a) ω_i =320 Hz (b) ω_i =925 Hz.

statistics by Martin (Martin, 1994; Martin, 2001) is quite similar to the proposed VAD in that it also handles the noise power spectrum during speech activity. It is conceptually simple and suited for the real-time implementations. However, it needs to compensate for the bias of the estimated noise power spectrum since the minimum of the estimated noise power spectrum is typically smaller than the true one. As a result, it tends to suffer from the non-stationary noise. Since the accuracy of the estimated noise power



(a)



(b)

Fig. 6. Real and estimated noise power spectra using minimum statistics (a) ω_i = 320 Hz (b) ω_i = 925 Hz.

spectrum is essential to the noise filtering algorithm incorporated with the VAD, the per-formance of the proposed VAD is evaluated in the light of its capability of capturing the spectral characteristics of noise, specifically the noise power spectrum.

The performance of the proposed VAD is compared with that of the minimum statistics by Martin using the automobile interior noise. The automobile interior noise is collected while driving the car at 60 km/hour on a high way (approximately 14 seconds) at the sampling rate 5.5 kHz. The number of samples in a data frame is 256 (about 30 ms) and there is 50 % overlap between the adjacent frames. Figure 4 shows the power spectrum of the automobile interior noise with respect to the data frame index. As shown in Fig. 4, the automobile noise has significant power in the low frequency range (below 500 Hz). The proposed VAD estimates the noise power spectrum following **Steps 1, 2, 3** and **4** after initialization steps **I1, I2** and **I3** as described in Section 3. It is noted that the noise power spectra using the proposed VAD and the minimum statistics are overestimated by the over-

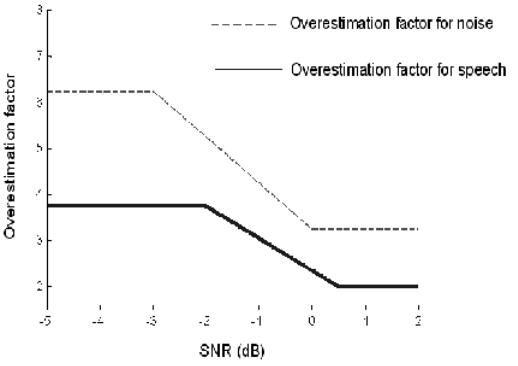


Fig. 7. The overestimation factor α as a function of SNR.

estimation factor $\alpha_k(\omega_i)$ in Fig. 7. Figures 5 and 6 show the traces of the noise power spectra at 320 Hz and 925 Hz estimated using the proposed VAD and the minimum statistics, respectively. The noise power spectrum at 320 Hz manifests the non-stationary nature of the automobile noise with the maximal variation of 15.5 dB, while the one at 925 Hz seems to remain quite stationary at much lower power level. The solid and dashed lines in Fig. 5 and 6 denote the overestimated and real noise power spectra. Since the noise power spectrum is overestimated, the solid line must always lie above the dashed one. Otherwise, the probability of the false detection increases, i.e., the noise component may be mistaken for speech component. It is obvious that at both 320 Hz and 925 Hz, the proposed VAD tracks the noise profile much better than the minimum statistics. It must be specially noted that despite some delays and errors in certain frames, the proposed VAD provides the estimated noise spectrum that tracks the true noise spectrum at 320 Hz much better, which shows how well the proposed VAD performs under the rapidly varying noise. In Fig. 6(a), it is observed that the estimated noise power spectrum sometimes lies below the true one, which results in the underestimated noise power spectrum that may cause false detection of speech activity. Although it outperforms the one based on the minimum statistics at both 320 Hz and 925 Hz, the tracking capability of the proposed VAD becomes more pronounced at the high noise level or low SNR (320 Hz), which implies that the proposed VAD works well even under the low SNR situation. Compared with Figs. 5(a) and (b), Figs. 6(a) and (b) indicate that the estimated noise power spectra based on the minimum statistics deviate from the true one to much larger extent. As a result, it may be deduced that the excessive overestimation of the noise power spectrum with minimum statistics inevitably degrades the quality of the filtered speech signal, granted that the overestimation is essential to certain extent.

Now equipped with the proposed VAD, the performance of the Mel-scale Wiener filter is established with two types of experimental data. The first set consists of synthesized noisy speech signals that are artificially generated by adding the independently collected automobile interior noise to the clean speech. The second set consists of noisy speech signals collected in an automobile cabin while the vehicle speed is marked at 60 km/hour. For the first set, the speech signals and the automobile noise are obviously uncorrelated while the same may not be true for the second one. Yet, the performance analysis with these two sets of experimental data must shed light on the viability of the proposed algorithm in practice.

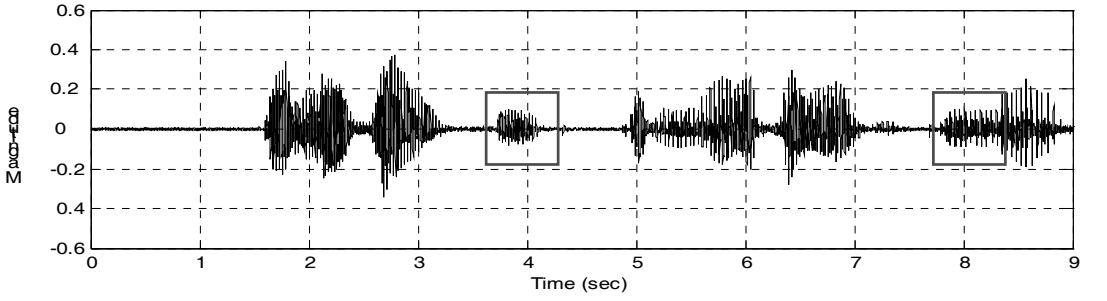
The experimental data are again sampled at the sampling rate of 5.5 kHz. The number of samples N in a data frame is 256 and the total duration of experimental data is 9 seconds. The overestimation factor $\alpha_k(\omega_i)$ varies within a frame, depending on whether a frame contains speech or noise at a given frequency. Figure 7 shows the overestimation factor as a function of SNR. The solid and dashed lines denote the overestimation factors for speech and noise, respectively. The overestimation factor $\alpha_k(\omega_i)$ for speech is smaller and shifted to the higher frequency in order to minimize the speech distortion by preventing the excessive subtraction of noise power spectrum. The spectral floor $\beta_k(\omega_i)$ is set to 0.001 over the entire frequency range and data frames.

In order to evaluate the performance of the Mel-scale Wiener filter in conjunction with the proposed VAD, several measures are introduced: segmental/overall SNRs and Itakura measure (Deller et al., 1993). The segmental/overall SNRs are defined as

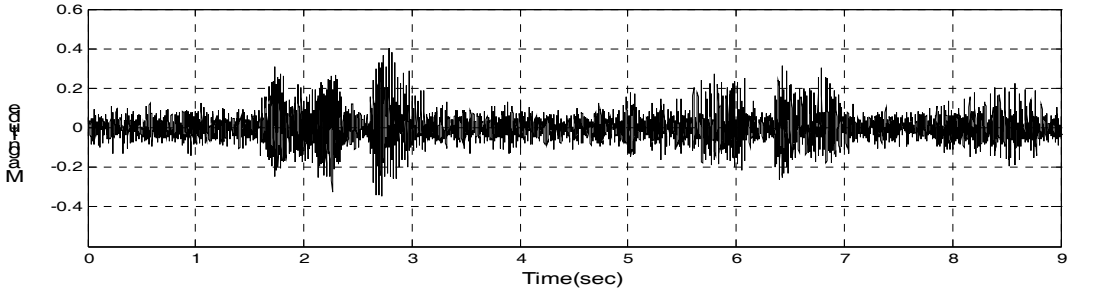
$$\mathrm{SNR}_{\mathrm{over}} = 10\log_{10}\frac{E_s^{\mathrm{over}}}{E_n^{\mathrm{over}}} = 10\log_{10}\frac{\sum_i s^2(t_i)}{\sum_i \left[s(t_i)-\hat{s}(t_i)\right]^2} \tag{23}$$

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\sum_{i=m_j-N+1}^{m_j}\frac{s^2(t_i)}{\left[s(t_i)-\hat{s}(t_i)\right]^2}\right] \tag{24}$$

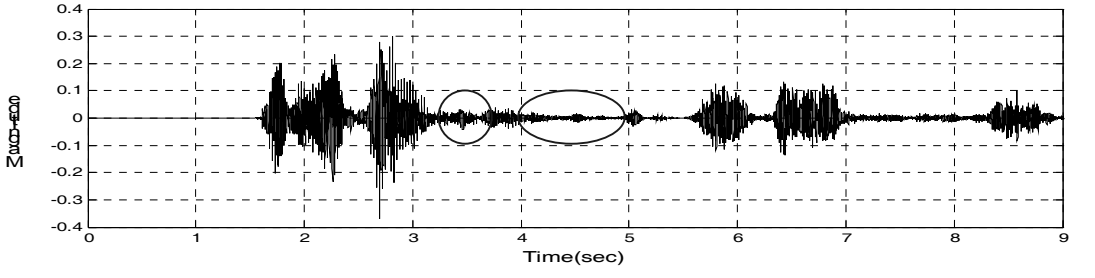
where $s(t_i)$ and $\hat{s}(t_i)$ denote the clean and filtered speech, respectively. E_s^{over} and E_n^{over} represent the energies of clean speech and noise over the entire frames. m_0 , m_1 ,…, m_{M-1} are the ending time index for the M frames, each with N samples. The segmental/overall SNRs provide indicators for average errors over time and frequency for a filtered signal. Note that the SNR of -10 dB represents rather



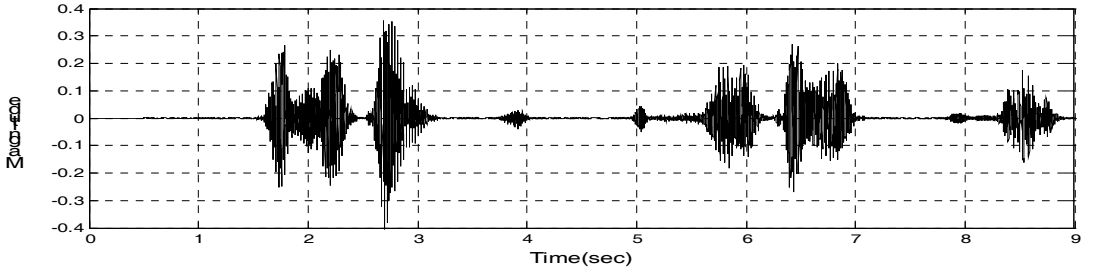
(a)



(b)



(c)



(d)

Fig. 8. Synthesized speech signal: (a) clean speech (b) clean speech + noise (c) filtered speech with the existing Wiener filtering (d) filtered speech with the Mel-scale Wiener filtering.

harsh conditions unlikely to occur often in the real-world applications. On the other hand, at the SNR of 10dB, the speech quality may be acceptable even without a speech enhancement. For most applications, the input SNR lies between -5~5 dB ranges. Yet, another measure, the segmental/overall Itakura measures, are introduced. Let $\sigma_c^2/|A_c(\omega)|^2$ and $\sigma_f^2/|A_f(\omega)|^2$ represent the power spectra of the clean and filtered speech signals, which are all assumed to be modeled as autoregressive models (AR) $\sigma/A(z)$. The Itakura measure [0] is given by

$$d_{IS}\left(\frac{\sigma_c^2}{|A_c(\omega)|^2},\frac{\sigma_f^2}{|A_f(\omega)|^2}\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma_c^2/|A_c(\omega)|^2}{\sigma_f^2/|A_f(\omega)|^2}d\omega - \ln\left|\frac{\sigma_c^2}{\sigma_f^2}\right| - 1 \quad (25)$$

where $A_c(\omega) = A_c(z)_{z=e^{j\omega}} = 1+\sum_{i=1}^{P} a_c(i)e^{-j\omega}$, $A_f(\omega) = A_f(z)_{z=e^{j\omega}} = 1+\sum_{i=1}^{P} a_f(i)e^{-j\omega}$. σ_c^2 and σ_f^2, $a_c(i)$ and $a_f(i)$ are the gains and i^{th} LPC prediction coefficients of two *P-order* LPC models, respectively. The segmental Itakura measure is calculated on a given frame while the overall Itakura measure is over the whole frames. The Itakura measure is sensitive to variations in speech spectrum and, as a result, is heavily influenced by the spectral dissimilarity due to mismatch in formant locations, whereas errors in matching spectral valleys do not contribute significantly to the measure. Such a property is highly desirable because the auditory system is more sensitive to errors in formant location and bandwidth than to the spectral valleys between peaks.

The SNR of a synthesized speech signal is 0.55 dB. Figures 8(a), (b), (c) and (d) show in the time domain the clean and synthesized speeches, the filtered speech signals via an existing and Mel-scale Wiener filters, respectively. It is obvious that the filtered speech signal via the Mel-scale Wiener filter looks most similar to the clean speech. Over the time intervals where only the noise is present such as over the initial 1.5 sec interval, both Wiener filtering algorithms seem to perform reasonably well. However, the advantage of the Mel-scale Wiener filter becomes obvious at other time intervals when the speech signal is active. Note that the clean speech signal in the time intervals from 3.5 sec to 4 sec and from 7.6 sec to 8.4 sec (indicated by rectangles) is completely inundated by the automobile noise, which makes it difficult to detect the presence of the speech with the naked eye or through the hearing test. Although the speech volume has decreased, the Mel-scale Wiener filter separates the speech signal from the background noise much better than the exiting one. Table 1 summarizes the segmental/overall SNRs and the Itakura distance measures between the clean and filtered speech signals. Recall that the high overall/segmental SNR and the low overall/segmental Itakura measures imply the better performance of the filtering algorithm. Compared with an existing Wiener filter, the Mel-scale Wiener filter noticeably enhances the quality of the noisy speech signal; the overall SNR improvement (=8.58 dB) via the Mel-scale Wiener filter are larger than that (=4.47dB) via an existing Wiener filter. The worst (=0.77 dB) and the best (=12.7 dB) segmental SNR improvements via the proposed Wiener filter outnumber those (-1.27 dB and 11.3 dB, respectively) via an existing Wiener filter. In addition, the segmental/overall Itakura measures demonstrate that the Mel-scale Wiener filtering excels in reproducing the clean speech as well as eliminating noise from the noisy speech signal. Compared with the existing Wiener filtering that accompanies the annoying musical noise indicated by the circles in Fig. 8(c), the Mel-scale Wiener filtering provides considerable reduction in musical noise and much smooth transition in Fig. 8(d).

To confirm the capability of the proposed Wiener filter under a real-world situation, the performance analysis through a real noisy speech signal follows.

Table 1. Segmental/overall SNRs and Itakura measures of the Mel-Scale Wiener filtering and existing Wiener filtering.

	Mel-Scale Wiener filtering			Existing Wiener filtering		
Segmental SNR(dB)	Min	Mean	Max	Min	Mean	Max
	0.77	8.05	12.7	-1.23	5.58	11.3
Overall SNR(dB)	8.58			4.47		
Segmental Itakura measure	Min	Mean	Max	Min	Mean	Max
	0.10	2.39	17.2	0.31	4.06	26.2
Overall Itakura measure	2.37			4.03		

Figure 9(a) shows a real noisy speech signal in time domain. Figures 9(b) and (c) shows the filtered speech signals obtained from an existing and Mel-scale Wiener filters, respectively. In comparison with the synthesized noisy speech signal in Fig. 8(b), the real noisy speech signal seems to be collected at lower SNR and it is expected that the filtering results may be poor. At the time intervals corresponding to the relatively high SNR (indicated by the rectangles), it is obvious that both algorithms successfully separate speech from noise. However, at the low SNR conditions, Fig. 9(b) and (c) shows that the musical

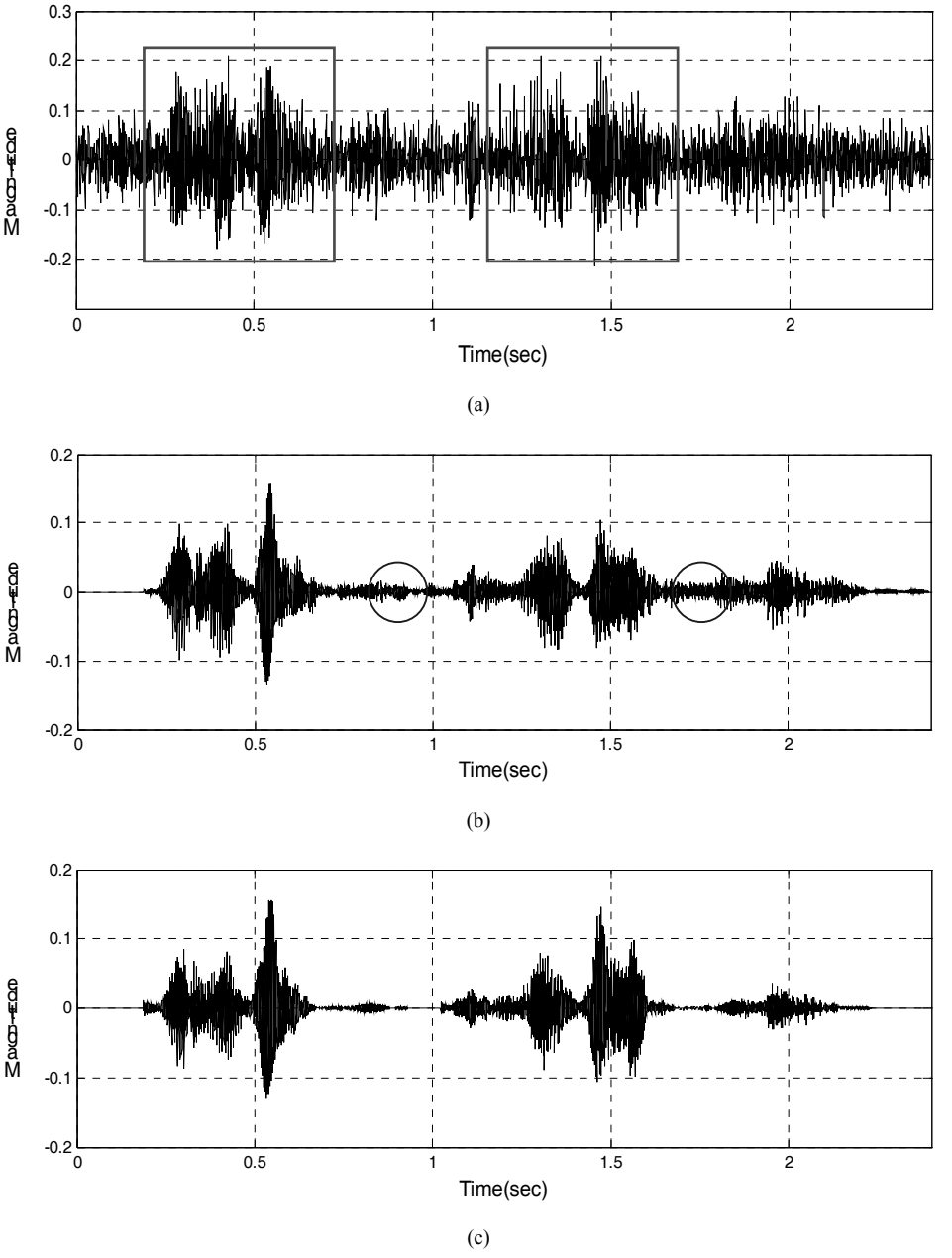
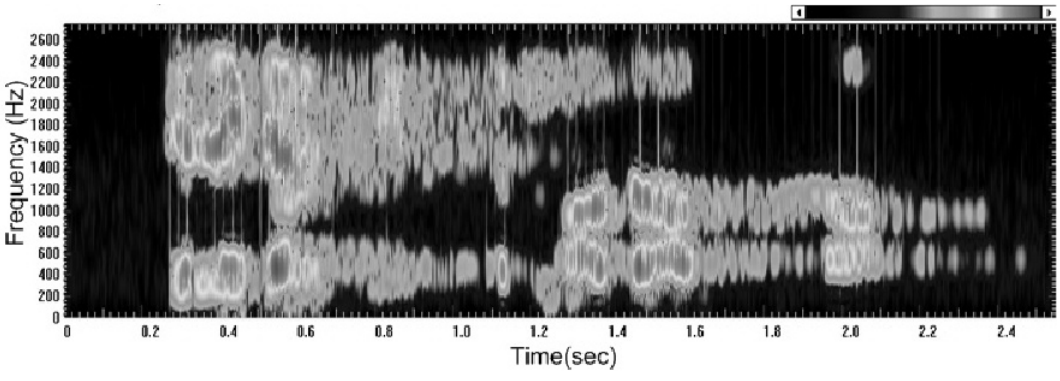


Fig. 9. Real speech signal: (a) noisy speech (b) filtered speech with the existing Wiener filtering (c) filtered speech with the Mel-scale Wiener filtering.

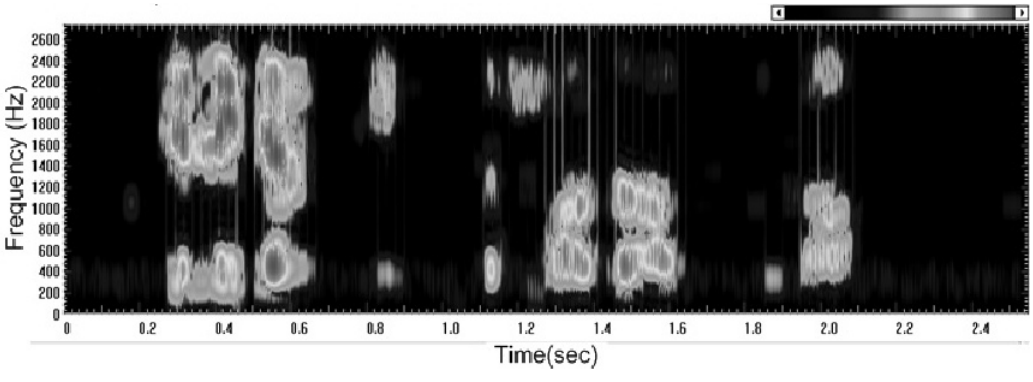
noise is present in the existing Wiener filtering while it is completely eliminated through the Mel-scale Wiener filtering. These observations are explicitly confirmed from Fig. 10(a) and (b) which show the corresponding spectrograms of filtered speech signals in Fig. 9(b) and (c), respectively. As shown in the time intervals from 0.7 sec to 0.8 sec, from 0.9 sec to 1.1 sec and from 1.6 sec to 1.8, the proposed Mel-scale Wiener filter successfully mitigates the musical noise widely spread in the existing Wiener filter without decreasing the energy level of the filtered speech signal. Since a clean speech signal cannot be completely separated from a real noisy speech signal as in the synthesized speech, the informal listening tests by several listeners are conducted, which show that the Mel-scale Wiener filter attenuates the noise considerably while minimizing the speech distortion. The previous two experiments exemplifies that the Mel-scale Wiener filtering improves the efficiency of the existing Wiener filtering by eliminating the residual noise and decreasing the speech distortion.

5. Conclusion

In this paper, new approaches for speech enhancement in an automobile cabin are presented: the Mel-scale Wiener filter and the frequency-wise voice activity detector. Through smoothing over the adjacent frequencies, the Mel-scale Wiener filter based on the characteristics of a human auditory system improves upon an existing Wiener filter that exhibits an inevitable nuisance, "musical noises". For the more up-to-date and accurate estimate of noise statistics, the frequency-wise voice activity detector detects the presence of speech not at the frame level but at the frequency level within a data frame so that the detection error may decrease and provide more accurate information on the noise power spectrum, which plays an important role in the Mel-scale Wiener filtering. The proposed VAD is shown to



(a)



(b)

Fig. 10. Spectrograms of (a) filtered speech with the existing Wiener filtering (b) filtered speech with the Mel-scale Wiener filtering.

detect the presence of the speech signal successfully even at low SNR and under the rapidly varying noise, while computational complexity remains acceptable. Combined with the proposed VAD, the Mel-scale Wiener filter is shown to reduce the musical noise and result in smooth transition over adjacent frames without distorting the speech quality. The proposed approach has demonstrated its viability in practice through extensive experiments, which will lead to its deployment in the commercial speech enhancement system in an automobile cabin. It goes without saying that the same idea has potential to impact the speech recognition systems and hands-free devices in an automobile cabin.

References

Berouti, M., Schwartz, R. and Makhoul, J., 1979, "Enhancement of Speech Corrupted by Acoustic Noise," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 208~211.

Boll, S. F., 1979, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. 27, No. 2, pp. 113~120.

Cappe, O., 1994, "Elimination of Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, pp. 345~349.

Chen, G., Koh S. N. and Soon, I. Y., 2003, "Enhanced Itakura Measure Incorporating Masking Properties of Human Auditory System," *Signal Processing*, Vol. 83, No. 7, pp. 1445~1456.

Deller, J., Proakis, J. and Hansen, J., 1993, "Discrete-Time Processing of Speech Signals," Macmillan Publishing Co, Englewood Cliffs, NJ, USA.

Doblinger, G., 1995, "Computationally Efficient Speech Enhancement By Spectral Minima Tracking In Subbands," *EUROSPEECH '95*, Vol. 2, pp. 1513~1516, Madrid, Spain.

Kailath, T., 1981, "Lectures on Wiener and Kalman Filtering," Springer-Verlag, NY, USA.

Kim K. C. and C. M. Kim, 2005, "A Study on the Body Attachment Stiffness for the Road Noise," *Journal of Mechanical Science and Technology*, Vol. 19, No. 6, pp. 1034~1312.

Kybic, Jan, 1998, "Kalman filtering and Speech enhancement," Master thesis, Czech Technical Univ, Czech Republic.

Malah D. and Cox, R. V., 1982, "A Generalized Comb Filtering Technique for Speech Enhancement," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Vol. 7, pp. 160~163, Paris, France.

Martin, R., 1994, "Spectral Subtraction Based on Minimum statistics," *Proc. EUSIPCO'94*, pp. 1182~1185, Edinburgh, UK.

Martin, R., 2001, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech, Audio Processing*, Vol. 9, No. 5, pp. 504~512.

Oh J. E. and Cha, K. J., 2000, "Noise Reduction of Muffler by Optimal Design," *Journal of Mechanical Science and Technology*, Vol. 14, No. 9, pp. 947~955.

Papoulis, A., 1981, "Probability, Random Variables, and Stochastic Process," McGraw Hill, NY, USA.

Pollak, P., Sovka, P. and Uhlir, J. "Cepstral Speech/Pause Detectors," 1995, *Proc. IEEE Workshop on Nonlinear Signal and Image Processing*, Neos Marmaras, Greece.

Puder, H., 1999, "Single Channel Noise Reduction Using Time-Frequency Dependent Voice Activity Detection," *Proc. 6th Int. Workshop on Acoustic Echo and Noise Control*, pp. 68~71, Pocono Manor, USA.

Rabiner L. and Juang, B., 1993, "Fundamentals of Speech Recognition," Prentice Hall, Englewood cliffs, NJ, USA.

Rabiner L. R. and Schafer, R. W., 1978, "Digital Processing of Speech Signals," Prentice Hall, Englewood cliffs, NJ, USA.

Rogan, B., 1998, "Adaptive Model-Based Speech Enhancement," Ph. D thesis, Cambridge University, Cambridge, UK.

Sohn, H. S., Kim N. S. and Sung, W., 1999, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Processing Letters*, Vol. 6, No.1, pp. 365~368.

Sovka, P., Pollak P. and Kybic, J., 1996, "Extended Spectral Subtraction," *Proc. of European Conference on Signal processing and Communication*, Trieste, Italy.

Van-Compernolle, D., 1989, "Noise Adaptation in a Hidden Markov Model Speech Recognition System," *Computer Speech Language*, Vol. 3, pp. 151~167.