Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 964–974

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speech enhancement based on soft audible noise masking and noise power estimation ☆

Rongshan Yu *

Department of Signal Processing, Institute for Infocomm Research (I2R), 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore

Received 28 December 2012; received in revised form 20 May 2013; accepted 22 May 2013
Available online 25 June 2013

Abstract

This paper presents a perceptual model based speech enhancement algorithm. The proposed algorithm measures the amount of the audible noise in the input noisy speech based on estimation of short-time spectral power of noise signal, and masking threshold calculated from the estimated spectrum of clean speech. An appropriate amount of noise reduction is chosen based on the result to achieve good noise suppression without introducing significant distortion to the clean speech. To mitigate the problem of "musical noise", the amount of noise reduction is linked directly to the estimation of short-term noise spectral amplitude instead of noise variance so that the spectral peaks of noise can be better suppressed. Good performance of the proposed speech enhancement system is confirmed through objective and subjective tests.

Keywords: Speech enhancement; Speech processing; Auditory model; Perceptual model; Noise estimation; Noise suppression; Noise tracking

1. Introduction

Speech enhancement technology has been widely used in telecommunication systems to improve the quality of voice communication in noisy environments. Usually it is performed directly on the output of the microphone in the transmission side to achieve the best speech enhancement quality. However it is possible to apply the speech enhancement system inside the telecommunication network or use it in the playback device. For economic reasons most systems are single microphone based solutions where the speech enhancement is done on the output of a single microphone, although better speech enhancement results can be achieved by using a microphone array system with more than one microphone.

In principle, a single microphone speech enhancement system uses some sort of adaptive filtering operation to attenuate the time/frequency (T/F) regions of the noisy speech signal that have low Signal-to-Noise-Ratios (SNR), and preserve those with high SNR. By doing so, the essential parts of the speech signal is thus preserved while the noise level is greatly reduced, leading to an enhanced signal with reduced noise level. Various speech enhancement systems along this line have been proposed in the literature, e.g., spectral subtraction (Boll, 1979), Wiener filter (Widrow and Stearns, 1985), MMSE-STSA (Ephraim and Malah, 1984), and MMSE-LSA (Ephraim and Malah, 1985). In these algorithms, some attenuation rules are used to decide which T/F regions of the noisy speech should be attenuated and by how much. Usually, these attenuation rules are optimized in such a way that the enhanced speech is as close as possible to the clean speech signal in the input; and the difference among these attenuation rules mainly result from different statistical models of the signals assumed as well as different distortion measurements used in the optimization.

☆ This research was done while the author was a staff engineer with Dolby Laboratories, Inc., San Francisco, CA 94103, USA.

* Tel.: +65 64082629.
E-mail addresses: ryu@i2r.a-star.edu.sg, rongshanyu@ieee.org.

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.05.006

Clearly, the quality of a single-microphone speech enhancement system described above is to a large extent determined by the suppression rule it uses. In general, a suppression rule with stronger attenuation will lead to less noisy output; however, the speech signal will become more distorted. Conversely, a suppression rule with more moderate attenuation will produce less distorted speech signal while it can only achieve limited amount of noise reduction. For this reason, a careful trade-off has to be made to balance the amount of the noise suppression with the speech distortion for optimal quality. To this end, auditory masking model, which has been successfully applied in wideband audio coding (Johnston, 1988), has recently been introduced in speech enhancement systems (Gustafsson et al., 1998; Lin et al., 2003; Virag, 1999; Tsoukalas et al., 1997; Hu and Loizou, 2004; Jabloun and Champagne, 2003; Hansen et al., 2006). In these systems, the masking properties of speech signal are employed to identify the perceptual significant noise components of the noisy speech signal, which are then subtracted from the speech signal for noise reduction. Various heuristics have been proposed in this system to incorporate the masking threshold of speech signal into the subtraction equations in order to obtain a trade-off between audible noise suppression and speech distortion.

In this paper we propose a perceptual model based speech enhancement algorithm. The proposed algorithm achieves good noise suppression qualities by using an attenuation rule that carefully balances the amount of reduction of the audible noise and the amount of distortion introduced to the clean speech. To this end, short-term spectral amplitudes of both the clean speech and noise signals are estimated continuously in the algorithm. Masking threshold of the estimated clean speech amplitude is then calculated by using a perceptual model. After that, the amount of audible noise in the noisy signal is calculated by contrasting the estimated noise amplitude to the masking threshold and an appropriate amount of attenuation is chosen based on a soft audible noise suppression principle that minimizes a cost function that explicitly includes the amount of audible noise and speech distortion in the enhanced speech signal. Since the auditory masking effect is only a short time phenomenon with a limited duration (Johnston, 1988), in the proposed algorithm the amount of attenuation of the proposed algorithm is linked directly to the estimation of the short-term noise spectral amplitudes instead of long-term noise variance. As a result, it provides superior suppression of the noise peaks in the frequency domain, leading to fewer “musical noise” artifacts (Cappe, 1994) in the enhanced speech signal.

Noise variance estimation plays an important role in determining the quality of a speech enhancement system, particularly in an environment with non-stationary noise. One popular choice of the noise variance estimator in both research literature and commercial speech enhancement implementations is the Voice Activity Detection (VAD) based approach, where the noise estimation is updated only when speech is not present in the input. The performance of the VAD approach strongly depends on the accuracy of the voice detection, which is a difficult task in particular for signals with low SNR. In addition, this method precludes the possibility of updating the noise estimation when the speech signal is present, which is inefficient since there may still be spectral bands where the speech level is weak even during speech segments. Another widely cited method is the Minimum Statistics (MS) noise estimator (Martin, 2001). In principle, the MS method keeps a record of historical samples for each spectral location, and the noise level is estimated based on the minimum signal level from the record. It is reported that the MS method achieves good tracking performance for non-stationary noises; however, it has a high memory demand and is not applicable to devices with limited memory resources.

To address these issues, in this paper we adopted a low-complexity noise variance estimation algorithm previously described in Yu (2009). In this algorithm, the instantaneous noise power is estimated each frame based on information from the incoming signal and the current estimated distribution parameters. After that, the distribution parameters, including the noise variance that we are interested in, are refined from the expectation results. Instead of using a trained gain function for noise power estimation as proposed in Erkelens and Heusdens (2008), naive minimum mean-square-error (MMSE) noise power expectation is used and the potential estimation bias problem is addressed by using a bias estimation correction method. The proposed algorithm has very low computational power and memory requirements while still delivering satisfactory performance for various noise types in our tests.

The rest of this paper is organized as follows. In Section 2, the principles of proposed soft audible noise suppression algorithms are described. Implementation issues of the proposed algorithm, including noise amplitude and variance estimation and the masking threshold calculation, are presented in detail in Section 3. The performance of the each of the proposed algorithms is evaluated in Section 4. Finally, Section 5 presents the conclusions drawn from this work.

2. Principle

2.1. Signal model

We consider the following additive signal model for a noisy speech signal:

$$\mathbf{Y}_k(m) = \mathbf{X}_k(m) + \mathbf{D}_k(m), \quad k = 1, \ldots, K, \tag{1}$$

where $\mathbf{Y}_k(m), \mathbf{X}_k(m)$, and $\mathbf{D}_k(m)$ are complex-valued short-time Fourier Transform (STFT) coefficients of the noisy speech signal $y(n)$, clean speech signal $x(n)$, and noise signal $d(n)$ respectively. Here k is the subband index, K is the total number of subbands, and m is the frame index.

In most current speech enhancement algorithms the speech and the noise signals are usually modeled as independent, zero-mean, complex Gaussian variables due to the simplicity of this model and its relatively good performance (Ephraim and Malah, 1985; Ephraim and Malah, 1984; Wolfe and Godsill, 2003). Define the variances $\lambda_x(k) \triangleq E\{|\mathbf{X}_k|^2\}$ and $\lambda_d(k) \triangleq E\{|\mathbf{D}_k|^2\}$. From the Gaussian assumption we now have the following marginal and conditional distributions (Wolfe and Godsill, 2003):

$$p(A_k) = \begin{cases} \frac{2A_k}{\lambda_x(k)} \exp\left(-\frac{A_k^2}{\lambda_x(k)}\right) & A_k \in [0,\infty); \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

$$p(N_k) = \begin{cases} \frac{2N_k}{\lambda_d(k)} \exp\left(-\frac{N_k^2}{\lambda_d(k)}\right) & N_k \in [0,\infty); \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

$$p(\mathbf{Y}_k|\mathbf{D}_k) = \frac{1}{\pi\lambda_x(k)} \exp\left(\frac{-|\mathbf{Y}_k - \mathbf{D}_k|^2}{\lambda_x(k)}\right), \tag{4}$$

where $A_k \triangleq |\mathbf{X}_k|$and $N_k \triangleq |\mathbf{D}_k|$ are, respectively, the amplitudes of clean speech and noise. For conciseness, the frame index m is omitted in the above distribution functions as well as in the subsequent discussion where it is clear from context.

2.2. Soft audible noise masking

In principle, to reduce the amount of perceptual noise of the corrupt speech signal, a speech enhancement algorithm would apply a set of frequency dependent suppression gains $\{g_k\}$ where

$$0 < g_k \leqslant 1, \quad \forall k = 0, \ldots, K-1, \tag{5}$$

to each spectral component. The enhanced subband coefficients can now be written as:

$$\mathbf{Y}'_k \triangleq g_k \mathbf{Y}_k = \mathbf{X}'_k + \mathbf{D}'_k, \quad \forall k = 1, \ldots, K, \tag{6}$$

where $\mathbf{D}'_k = g_k \mathbf{D}_k$ is the remaining noise of reduced amplitudes in the enhanced speech; and $\mathbf{X}'_k = g_k \mathbf{X}_k$ is the distorted speech. Our goal here is to find a set of suppression gains that minimize the perceptible loudness of the remaining noise as well as the speech distortion. To this end, the following objective function is constructed:

$$\begin{aligned} C_k &\triangleq \underbrace{\lceil \log|\mathbf{D}'_k| - \log T_k, 0\rceil^2}_{\text{audible noise}} + \beta \underbrace{|\log \mathbf{X}'_k - \log \mathbf{X}_k|^2}_{\text{speech distortion}} \\ &= \lceil \log\ g_k + \log N_k - \log T_k, 0\rceil^2 + \beta|\log\ g_k|^2, \end{aligned} \tag{7}$$

where $\lceil a, b\rceil \triangleq \max(a, b)$ for real a and b, and T_k the masking threshold for subband k. The first term in this objective function measures the relative intensity of the remaining noise compared to T_k, and the second term measures the level of the speech distortion. Both elements are measured in the log domain which is more perceptually relevant than measurement in linear domain, and they are combined in the power domain so that the degradation of larger amplitude is given a heavier penalty in the cost function. A weighting factor $0 \leqslant \beta < \infty$ is also introduced, which determines the relative weighting of both elements in the objective function. Note that only a positive value of the audible noise term is considered here since otherwise the remaining noise will be totally masked and thus it shouldn't contribute to the objective function.

Our goal is to find a set of optimal suppression gains $\hat{g}_k$ that minimizes C_k for each subband. In order to render the problem more tractable, we note that the masking threshold T_k is mainly contributed by speech components of relatively large amplitudes, which are rarely strongly suppressed in the enhanced signal due to their relatively high Signal-to-Noise Ratio (SNR). For this reason, we may assume that T_k is calculated directly from the distortion-free speech and hence it is not related to the suppression gains g_k. With this simplification, the optimal $\hat{g}_k$ can be found by considering (7) for two cases. Firstly, for the case $\log\ g_k + \log N_k > \log T_k$, specialize (7) as

$$C_k = (\log\ g_k + \log N_k - \log T_k)^2 + \beta \log^2 g_k. \tag{8}$$

Define the noise-to-mask ratio $\Phi_k \triangleq N_k/T_k$, and set $\partial C_k/\partial g_k = 0$; the latter implies

$$\hat{g}_k = \Phi_k^{-1/(1+\beta)}. \tag{9}$$

Secondly, for the case $\log\ g_k + \log N_k \leqslant \log T_k$, we can specialize (7) as

$$C_k = \beta \log^2 g_k, \tag{10}$$

which implies

$$\hat{g}_k = 1. \tag{11}$$

Combining (9) and (11) the optimal suppression gains $\hat{g}_k$ are now given by:

$$\hat{g}_k = \begin{cases} \Phi_k^{-\frac{1}{1+\beta}}, & \Phi_k > 1; \\ 1, & \text{otherwise.} \end{cases} \tag{12}$$

The above suppression rule can also be interpreted as follows. When $T_k \geqslant N_k$, the noise is masked by the speech signal and hence no attenuation is required. In this case we have $\hat{g}_k = 1$. For subbands with audible noise, i.e., $T_k < N_k$, the optimal suppression gain $\hat{g}_k$ is actually given by Φ_k^{-1} which is the amount of attenuation that is necessary to suppress the noise amplitude to the level of the masking threshold, and further smoothing by a exponential factor $(1+\beta_k)^{-1}$ to avoid over-attenuating the speech signal. Therefore, we call the proposed algorithm soft audible noise suppression.

The value of β plays an important role in determining the quality of the final enhanced speech. Generally speaking, larger values of β will lead to less distorted speech, while there will be more noise remaining in the enhanced speech signal. Conversely, more noise suppression can be achieved by using a smaller value of β ; however the speech will be more distorted in the enhanced speech signal. In practical applications, the value of β can be adjusted based on the requirements of the applications, or based on user preferences.

Another well-known problem of speech enhancement systems is “musical noise” (Cappe, 1994). Musical noise is a direct consequence of the subband domain filtering operation used in the speech enhancement systems. In principle, noise signals in the subband domain can exhibit strong fluctuations in amplitudes and have a succession of randomly spaced peaks. If the speech enhancement system only suppresses noise signals with relatively small amplitudes in between the peaks, but fails to give enough attenuation to the peaks, those under-suppressed peaks will be transformed into short, bursty musical tones with random frequencies by the synthesis filterbank. In the proposed algorithm the musical noise problem is mitigated since the attenuation rule is tied explicitly to the estimated short-term noise amplitudes. As a result, it is able to track the spectral peaks of noise signal and suppresses them accordingly, in particular during low SNR regions where the noise amplitudes can be accurately located. A suppression gain defined in this way is more effective in suppressing the spectral peaks of the noise signal and hence is less susceptible to the musical noise problem.

3. Implementation

3.1. Overview

The block diagram of the proposed speech enhancement system is given in Fig. 1. As shown in the figure, the input noisy signal is decomposed into subbands of different frequencies using an analysis filterbank. An oddly stacked complex modulated filterbank is used so that both amplitude and phase are available for subband signals (Princen et al., 1987). The filterbank has a total of $K = 32$ subbands, and uses a 50% overlapping analysis window. For 8 kHz narrow-band speech signal this corresponds to a frequency resolution of 125 Hz and a total delay (including delay from both the analysis and the synthesis filterbanks) of 8 ms, which serves as a good compromise between the frequency resolution required for good noise suppression performance and the low-delay constraint from real-time telecommunication applications.

The subband signal output from the analysis filterbank is supplied to the speech and noise amplitude estimators to estimate the amplitudes of clean speech and noise respectively. Since the clean speech signal and the noise signal are already mixed in the input, such estimation is done on a “best-effort” basis, and relies on prior knowledge of the statistical distribution of the clean speech and noise signals. In the proposed algorithm, the Minimum Mean Square Error (MMSE) power estimator (Wolfe and Godsill, 2003) is used in both estimators. The variances of the clean speech signal and the noise signal, which are necessary for the MMSE power estimation, are also estimated from the input signal using the speech and noise variance estimators, respectively.

As a next step, a psychoacoustic model (Johnston, 1988) is used to calculate the masking threshold for different subbands from the estimated clean speech amplitudes. The desirable amount of suppression for each subband is then chosen from the resulting masking threshold and estimated noise amplitude as described in the previous section. The suppression gains are updated on a sample-by-sample basis in the subband domain for accurately tracking changes in the masking threshold and the amplitude of the noise signal. Finally, the suppression gains are applied to the subband signals to obtain the enhanced signals, which are sent to the synthesis filterbank to produce the time-domain enhanced speech signal.

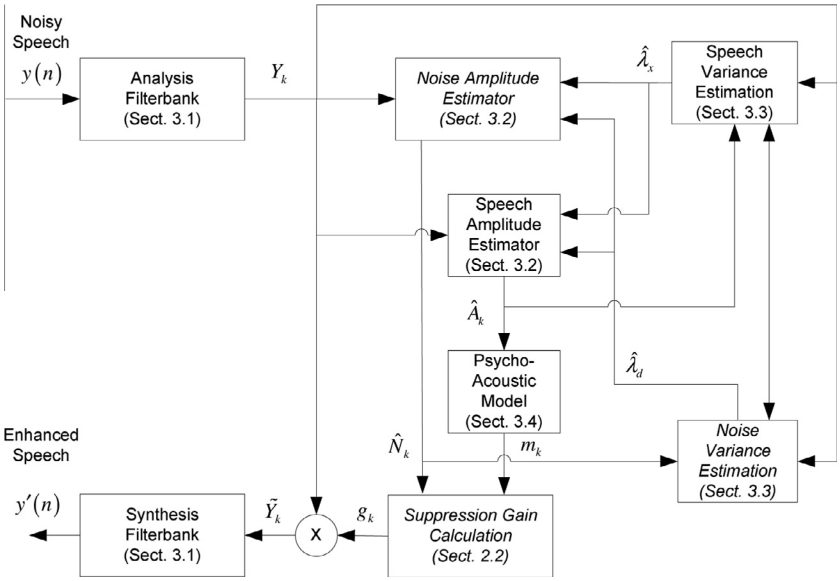

Noisy Speech
$y(n)$
Analysis Filterbank (Sect. 3.1)
Y_k
Noise Amplitude Estimator (Sect. 3.2)
$\hat{\lambda}_x$
Speech Variance Estimation (Sect. 3.3)
Speech Amplitude Estimator (Sect. 3.2)
$\hat{A}_k$
$\hat{\lambda}_d$
Psycho-Acoustic Model (Sect. 3.4)
Enhanced Speech
$\hat{N}_k$
m_k
Noise Variance Estimation (Sect. 3.3)
$y'(n)$
Synthesis Filterbank (Sect. 3.1)
$\tilde{Y}_k$
X
g_k
Suppression Gain Calculation (Sect. 2.2)

Fig. 1. Block diagram of the proposed speech enhancement system. Section number within each block indicates the location of the detailed description of this block. Blocks updated in this paper are highlighted with italic font.

3.2. Estimation of speech and noise amplitudes

Estimation of the amplitude of clean speech signal in a transformed domain from a noisy observed signal has been extensively studied in the speech enhancement literature, and various estimators have previously been proposed (Ephraim and Malah, 1984; Ephraim and Malah, 1985; Wolfe and Godsill, 2003). Most methods are based on the technique of statistical estimation which is conceptually very simple and meanwhile has proven very effective for this task. In those methods, *a posteriori* distribution of the amplitude of speech A_k after observing the noisy signal $\mathbf{Y}_k$ is determined based on the Gaussian assumptions, from which the optimal estimate of A_k that minimizes a certain distortion measurement is obtained.

More formally, following the definitions in Wolfe and Godsill (2003):

$$\frac{1}{\lambda(k)} \triangleq \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)} \tag{13}$$

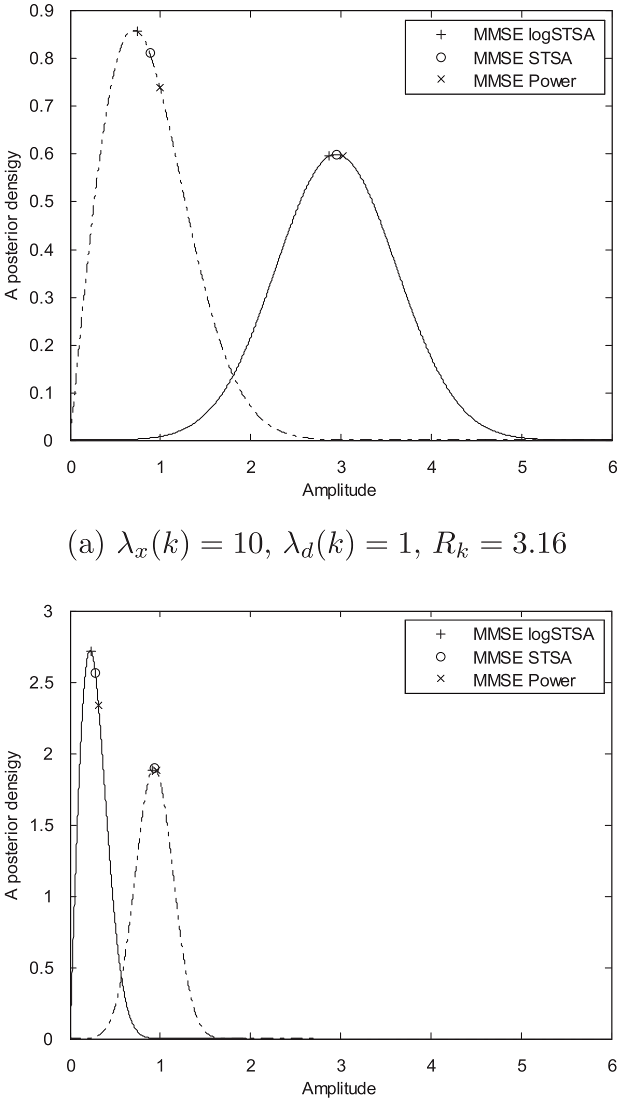
and

$$\upsilon_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k; \quad \xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)}; \quad \gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)}, \tag{14}$$

where $R_k \triangleq |\mathbf{Y}_k|$ is the amplitude of the noisy speech, and ξ_k and γ_k are usually referred to as *a priori* and *a posteriori* SNR in the literature. It can be shown that under the assumed Gaussian model, the *a posteriori* distribution of speech amplitude given the observed signal $\mathbf{Y}_k$ is Rice (1948) with parameters (σ^2, s^2):

$$p(A_k|\mathbf{Y}_k) = \frac{A_k}{\sigma^2}\exp\left(-\frac{A_k^2+s^2}{2\sigma^2}\right)I_0\left(\frac{A_k s}{\sigma^2}\right), \tag{15}$$

where $\sigma^2 \triangleq \frac{\lambda}{2}, s^2 \triangleq \upsilon\lambda$, and $I_i(\cdot)$ denotes the modified Bessel function of order i. It is now straightforward to derive statistically optimal estimators for speech amplitudes that incorporate different distortion measures such as Minimum-Mean–Square-Error Short-Time-Spectral–Amplitude (MMSE STSA) estimator (Ephraim and Malah, 1984), MMSE spectral power estimator (Wolfe and Godsill, 2003), and MMSE log-STSA estimator (Ephraim and Malah, 1985). Clearly, although those estimators were originally developed for speech amplitude estimation, they can be repurposed to estimate the amplitude of noise as well by exchanging the positions of $\lambda_x(k)$ and $\lambda_d(k)$ in the these estimators. Examples of *a posterior* distributions of the speech and noise amplitudes and their estimated values from different estimators are given in Fig. 2. It can be seen that all those estimators are effective in estimating speech and noise amplitudes. In addition, results from different estimators show only relatively small differences. Therefore, it is possible to use any estimator in the proposed system without materially affecting its performance, given that the small differences among those estimator are easily compensated for by tuning other parts of the overall system. However, the MMSE spectral power estimator (Wolfe and Godsill, 2003) is slightly preferred in our implementation due to its relatively low computational cost.



(a) $\lambda_x(k) = 10,\ \lambda_d(k) = 1,\ R_k = 3.16$

(b) $\lambda_x(k) = 0.1,\ \lambda_d(k) = 1,\ R_k = 1$

Fig. 2. A posteriori distribution of speech amplitude (solid line) and noise amplitude (dotted line) and their estimated values.

3.3. Estimation of speech and noise variances

Ideally, the variance of the noise signal can be tracked by using a recursive averaging process on the instantaneous noise power N_k^2, which is unfortunately not directly accessible since in most cases the noise signal is “corrupted” by the speech signal in the input signal. Instead, we may consider using the MMSE spectral power estimator for noise:

$$\hat{N}_k^2 = E\{N_k^2|\mathbf{Y}_k\} = \frac{\xi_k}{1+\xi_k}\left(\frac{1+\xi^{-2}\upsilon_k}{\gamma_k}\right)R_k^2, \tag{16}$$

which represents the best-effort estimate of N_k^2 given knowledge of the noisy signal. This idea leads to the following noise variance estimation algorithm:

$$\hat{\lambda}_d(k,m) = (1-\kappa)\hat{\lambda}_d(k,m-1) + \kappa\hat{N}_k^2(m-1), \tag{17}$$

where $0 < \kappa < 1$ is a constant, $\hat{\lambda}_d(k,m-1)$ is the noise variance estimation of the previous frame $m-1$ and $\hat{\lambda}_d(k,m)$ is the updated estimation of current frame m after incorporating noise power estimation $\hat{N}_k^2(m-1)$. The initial value $\hat{\lambda}_d(k,0)$ can be simply set to zero, or to the noise variances measured at the initialization stage of the noise variance estimator.

Once the noise variance $\lambda_d(k)$ is established, the speech variance can be obtained from the decision-directed method proposed in Ephraim and Malah (1984), which is normally represented in forms of estimation of *a priori* SNR ξ as follows:

$$\hat{\xi}_k(m) = \alpha\frac{\hat{A}_k^2(m-1)}{\hat{\lambda}_d(k,m)} + (1-\alpha)\lceil\gamma_k^2(m) - 1, 0\rceil. \tag{18}$$

Here $0 \ll \alpha < 1$ is a pre-selected constant, and$\hat{A}_k(m-1)$ is the estimated amplitude of clean speech of the previous analysis frame using MMSE spectral power estimator.

The noise power estimate in (16) is an unbiased estimator for $\lambda_d(k)$ only when we have perfect knowledge about the *a priori* SNR, i.e., when the $\hat{\xi}_k = \xi_k^*$ where ξ_k^* is the true SNR of the input signal. When $\hat{\xi}_k \neq \xi_k^*$ it becomes a bias estimator for $\lambda_d(k)$ where the estimation bias is given by:

$$b_k \triangleq \frac{E\{N_k^2 - \hat{N}_k^2\}}{E\{N_k^2\}} = \frac{\hat{\xi}_k - \xi_k^*}{(1+\hat{\xi}_k)^2}, \tag{19}$$

which may affect the accuracy of the noise variance estimation. As can be seen from Fig. 3(a) the estimation bias is very asymmetrical with respect to the error in the SNR estimation $\hat{\xi}_k$. Large negative bias (or over-estimation of noise power) occurs when $\hat{\xi}_k \ll 1$ and $\xi_k^* \gg \hat{\xi}_k$. This usually happens during speech onset where the estimated SNR from the decision-directed method (18) lags behind the true SNR, resulting in leakage of speech signal of large amplitudes into the noise variance estimation. To address this problem, we can simply skip samples with amplitudes that deviate too much from the assumed signal model in the noise estimation algorithm. More precisely, we exclude samples the amplitudes of which satisfy:

$$R_k^2 > \psi(1+\hat{\xi}_k)\hat{\lambda}_d(k), \tag{20}$$

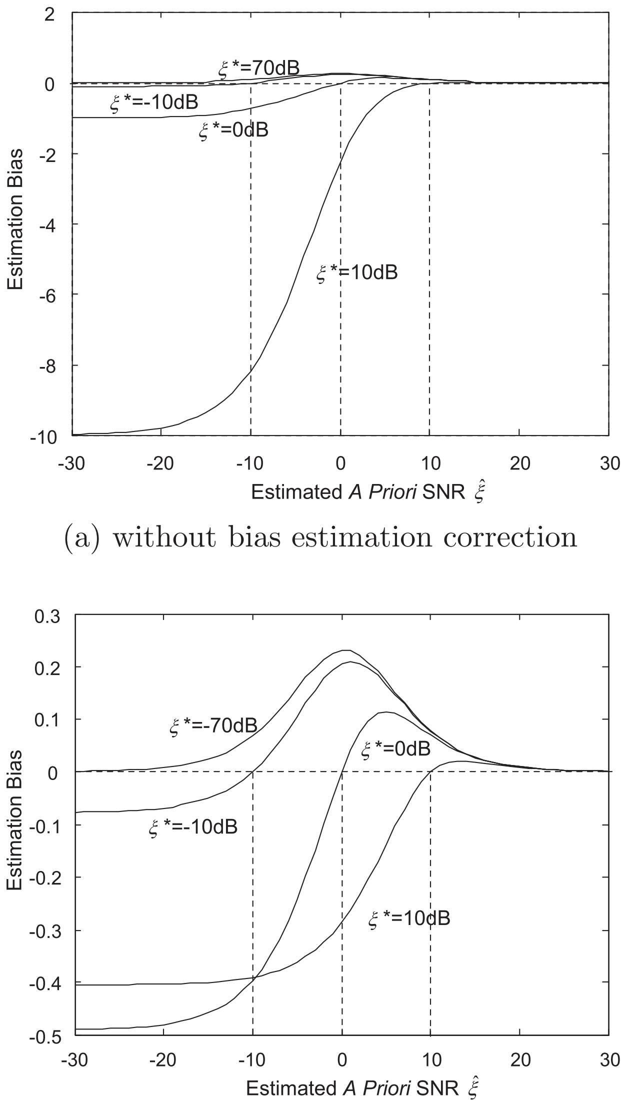
where $\psi > 0$ is a pre-defined constant. This is equivalent to replacing (16) with the following noise power estimator:

$$\hat{N}_k^2 = \begin{cases} E\{N_k^2|\mathbf{Y}_k\}, & R_k^2 \leqslant \psi(1+\hat{\xi}_k)\hat{\lambda}_d(k); \\ \hat{\lambda}_d(k), & \text{otherwise}. \end{cases} \tag{21}$$

This estimator effectively avoids the over-estimation problem. Unfortunately, it will introduce an (under) estimation bias even when $\hat{\xi}_k = \xi_k^*$:

$$b_k = \frac{1}{1+\hat{\xi}_k}\frac{\psi e^{-\psi}}{1-e^{-\psi}} > 0, \tag{22}$$

which, however, can be easily compensated for or simply neglected when ψ is sufficiently large. As evident in Fig. 3(b), the estimation bias of the updated noise estimator (21) is now well-bounded even for very inaccurate *a priori* SNR estimations.



(b) with bias estimation correction ($\psi = 4.5$)

Fig. 3. Estimation bias of MMSE noise estimator. Note that the MMSE is an unbiased estimator only when $\hat{\xi} = \xi^*$.

3.4. Calculation of masking threshold

Now we turn our attention to the masking threshold of the clean speech signal. Since the masking effect of the clean speech is largely determined by its spectral peaks, for which a very accurate estimation can be expected due to their relatively larger amplitudes, it is possible to estimate the masking threshold from the estimated clean speech amplitudes. To simplify the problem, we assume that the masker signals are pure tonal signals located at center frequencies of subbands, and have amplitudes of $\hat{A}_k, k = 1, \ldots, K$. Based on this simplification, the following procedure adopted Psychoacoustic Model I used in MPEG audio (ISO/IEC, 1992) is used to calculate the masking threshold from clean speech:

Step 1. Convert the speech power to the Sound Pressure Level (SPL) domain by

$$P_M(k) = \mathrm{PN} + 10\log_{10}(\hat{A}_k^2), \tag{23}$$

where the power normalization term PN is selected by assuming a reasonable playback volume.

Step 2. Calculate the masking threshold from individual maskers:

$$T_M(i,j) = P_M(j) - 0.275z(f_j) + \mathrm{SF}(i,j) - \mathrm{SMR}, \tag{24}$$

where $i, j = 1, \ldots, K$ is the subband indexes, f_j denotes the center frequency of subband j in Hz; $z(f)$ denotes the linear frequency f to Bark frequency mapping, which can be approximated as:

$$z(f) = 13\arctan(0.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right], \tag{25}$$

and $\mathrm{SF}(i,j)$ is the spreading of masking energy from subband j to subband i. As an example, the spreading function in Psychoacoustic Model I is given as follows:

$$\mathrm{SF}(i,j) = \begin{cases} 17\Delta_z - 0.4P_M(j) + 11, & -3 \leqslant \Delta_z < -1; \\ (0.4P_M(j) + 6)\Delta_z, & -1 \leqslant \Delta_z < 0; \\ -17\Delta_z, & 0 \leqslant \Delta_z < 1; \\ (0.15P_M(j) - 17)\Delta_z & \\ -0.15P_M(j), & 1 \leqslant \Delta_z < 8. \end{cases} \tag{26}$$

Here the maskee-masker separation in Bark Δ_z is given by:

$$\Delta_z \triangleq z(f_i) - z(f_j). \tag{27}$$

Step 3. Calculation of the global masking threshold:The contributions from all the maskers are summed up to produce the overall level of masking threshold for subband $k = 1, \ldots, K$:

$$T_M(k) = \sum_{l=1}^{K} 10^{0.1T_M(k,l)}. \tag{28}$$

Due to the non-normalized nature of the spreading function the calculated masking level is further normalized by:

$$T_N(k) = \frac{T_M(k)}{\sum_{l=1}^{M} 10^{0.1\mathrm{SF}\ (k,j)}}. \tag{29}$$

The normalized threshold is combined with the absolute hearing threshold (ISO/IEC, 1992) to produce the global masking threshold as follows:

$$T_g(k) = \max\{T_q(k), 10\log_{10}(T_N(k))\}, \tag{30}$$

where $T_q(k)$ is the absolute hearing threshold at center frequency of subband k in SPL.

Step 4. Transfer the global masking threshold back to the amplitude domain:

$$T_k = 10^{0.05[T_g(k) - \mathrm{PN}]}. \tag{31}$$

4. Experimental results

4.1. Performance of noise variance estimator

We evaluated the performance of the proposed noise variance estimator by measuring its estimation error for speech signal contaminated by various noise sources. The speech files used in the experiments contain concatenated sentences of eight short sentences from both male and female speakers. An approximate 0.5 second period of silence is inserted between each sentence pair in the speech files. The sentences are simple meaningful sentences in English extracted from the TIMIT database (Fisher et al., 1986). The speech files are filtered by the modified Intermediate Reference System (IRS) filter defined in ITU-T P.862 (ITU-T, 2001) and the speech levels as measured with the P.56 algorithm (ITU-T, 1993) are adjusted to -26 dBov. This processing ensures that the speech used in the tests is at the correct level and simulates the frequency characteristics of a telephone handset. Four different noise files were used in the tests and the noise levels were adjusted using the Root Mean Square (RMS) measure to the level commensurate with the SNR settings of the test conditions. The noise files were digitally mixed with the speech files to produce the testing files used in the tests. The sampling rate of the testing files is 8 kHz.

The parameters used in the experiments are as follows: $\alpha = 0.02, \kappa = 0.04, \psi = 4.5$. For comparison, the MS noise estimator was included in our experiments, for which the settings from Martin (2001) were used. In our experiments the estimation error is defined as the absolute value of the difference between the estimated noise variance and the true variance in the logarithmic domain as follows:

$$\mathrm{LogErr} = \left| 10\log_{10}\left(\frac{\hat{\lambda}_d(k,m)}{\tilde{\lambda}_d(k,m)}\right) \right| \ (\mathrm{dB}), \tag{32}$$

where $\hat{\lambda}(k,m)$ is the estimated noise variance for frequency bin k and frame index m. The true noise variance $\tilde{\lambda}_d(k,m)$ is calculated separately from the noise files as follows:

$$\tilde{\lambda}_d(k,m) = \kappa_0\tilde{\lambda}(k,m-1) + (1-\kappa_0)N_k^2(m), \tag{33}$$

where $\kappa_0 = 0.02$.

Table 1 gives the mean and variance of logErr that are measured over all the frames and frequency bins of the test files for each experiment condition. It can be seen that the proposed algorithm achieves a smaller mean estimation error for all the noise conditions. In addition, its performance is more consistent as suggested by the smaller variances of estimation error, which is also important for the quality of a speech enhancement system given that it is relatively easy to compensate for a consistent estimation error by fine-tuning other elements of the system.

Table 1
Mean and variance (in parenthesis) of the estimation error of the proposed algorithm and MS.

Noise type	6 dB		15 dB	
	Proposed	MS	Proposed	MS
Street	1.14(1.46)	1.52(1.64)	1.47(1.69)	1.71(1.91)
Car	1.07(0.96)	1.36(1.22)	1.25(1.08)	1.64(1.68)
Babble	1.69(1.53)	3.33(2.52)	1.91(1.65)	3.24(2.49)
Train	0.94(1.09)	1.37(1.41)	1.21(1.25)	1.55(1.77)

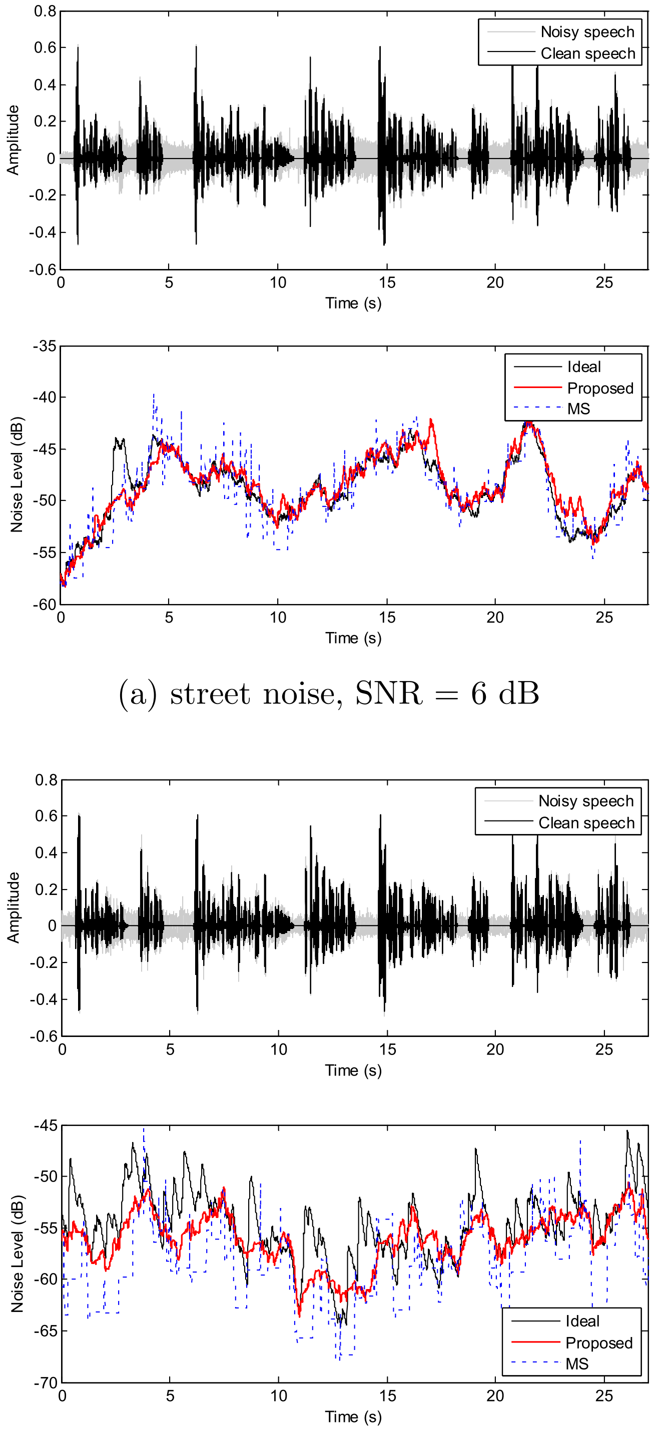
Instantaneous behavior of the proposed algorithm and MS algorithm are illustrated in Fig. 4 with two examples. It can be seen that in both cases the two algorithms are able to track the changing noise variance over time; neither of them, however, did a perfect job. The proposed algorithm tends to produce smoother estimates of the noise variance compared to the MS algorithm while the latter does a better job when the noise level is fluctuating significantly.

4.2. Performance of overall system

The performance of the proposed speech enhancement system is evaluated through ITU-T Recommendation P.835 (ITU-T, 2003) tests, which was developed for subjective evaluation of speech enhancement algorithms. The P.835 test was originally designed to reduce the listener's uncertainty as to which components of a noisy speech signal, i.e., the speech signal, the background noise, or both, should form the basis of their ratings of overall quality. This method instructs the listener to successively attend to and rate the enhanced speech signal based on: (a) the speech signal alone (SIG), (b) the background noise alone (BAK), and finally (c) the overall rating considering both speech and background quality (OVRL) using five-point rating scales shown in Table 2. More details about the testing methodology can be find in ITU-T (2003).

Two speech files (female and male) were used in our tests. Each speech file contains one pair of sentences and lasts $6 \sim 8$ s. An approximate 0.5 s period of silence precedes the first and second sentence. The sentences are simple meaningful sentences in English extracted from the TIMIT database. The speech files were digitally mixed to three different noises (car, street, and babble) to produce the noisy speech signals. Both the speech and noise files are processed using the same methods used in the noise estimation evaluation tests with two different SNR conditions: high SNR (10 dB) and low SNR (5 dB).

The resulting noisy speech signals were processed by speech enhancement systems to produce the testing sequences. Two speech enhancement algorithms were selected as reference systems in our tests. The first one was the MMSE-logSTSA (LSA) algorithm (Ephraim and Malah, 1985) due to its popularity as well as its good performance according to subjective evaluation results reported in Hu and Loizou (2006). The second system is the built-in speech enhancement algorithm in Enhanced Variable Rate Codec (EVRC-NS) (3GPP2, 2004), which is the top performer in Comparison Category Rating (CCR) test within 3GPP evaluation for noise suppression systems (3GPP, 2001b). The LSA system in our test uses the same parameters as those published except for the noise variance estimator, which is the same as the one used in the



(a) street noise, SNR = 6 dB

(b) babble noise, SNR = 6 dB

Fig. 4. Outputs of noise variance estimation algorithms as a function of time (of frequency bin $k = 10$).

Table 2
SIG, BAK and OVRL scales defined in ITU-T P.835 test.

SIG: Signal rating (attending only to spech)	BAK: Noise rating (attending only to noise)	OVRL: Overall rating (attending to both speech and noise)
5: Not distorted	5: Not noticeable	5: Excellent
4: Slightly distorted	4: Slightly noticeable	4: Good
3: Somewhat distorted	3: Noticeable but not intrusive	3: Fair
2: Fairly distorted	2: Somewhat intrusive	2: Poor
1: Very distorted	1: Very intrusive	1: Bad

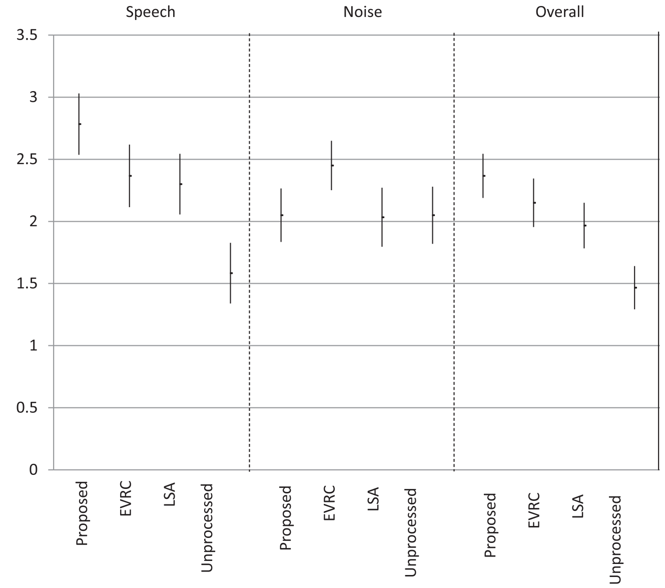


Fig. 5. Listening test results (SNR = 5 dB). Results are shown with 95% confidence interval.

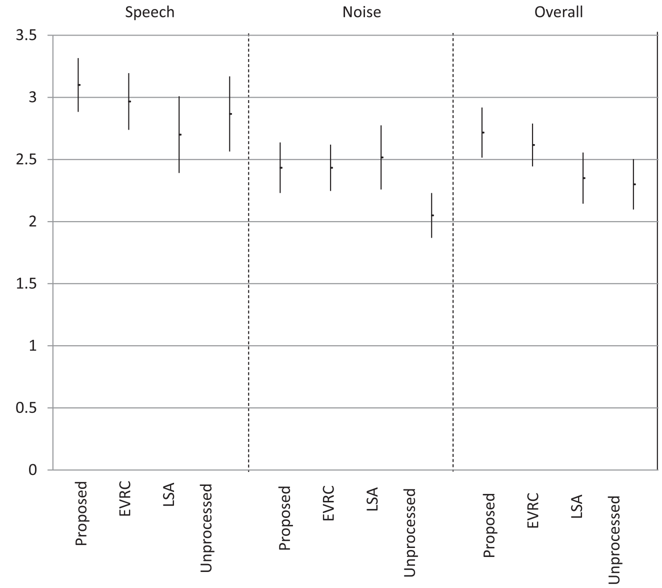


Fig. 6. Listening test results (SNR = 10 dB). Results are shown with 95% confidence interval.

proposed system for fair comparison. For EVRC-NS, C floating-point implementation available from 3GPP2 reference software (ver. 4.2, Jan 2004) was used. No further sound level adjustment was performed to the processed speech sequences. The unprocessed noisy speech (original) signals were also included in the tests for comparison.

In total 10 listeners participated in this listening test and results are summarized in Fig. 5 and Fig. 6 for different systems and testing conditions. Results are shown with 95% confidence interval. It is evident from those results that the proposed algorithm performed better compared to other systems. The only exception is that EVRC-NS performed best in BAK rating in 5 dB SNR tests. It is interesting to note that in most cases, the results do not show improved background quality from speech enhancement systems even though the background noise levels have been significantly reduced by those systems. In fact, in 10 dB tests the LSA algorithm even resulted in a lower mean BAK rating compared to that of unprocessed signals. One possible explanation is that the residual background noises from a speech enhancement system, albeit with lower levels, may perceived to be less natural compared to the original ones, which may be annoying to human subjects when they attend to the background only. However, most listeners reported improved speech quality from speech enhancement systems when they attend to speech only or to both speech and noise. In particular, informal interviews with listeners after the listening test revealed that compared to other systems, the proposed algorithm produced signals with less speech signal distortion, which is more favorable to most listeners. On the contrary, the LSA algorithm tends to over-suppress the speech signal, and produce noticeable amounts of musical noises in the background, which contributed to its relatively low ratings in the tests.

We conducted a further listening test to compare the performance of the proposed system with that of EVRC under a tandem configuration where the enhanced speech signals are further processed by a speech codec. Since in many practical applications such as mobile phones, speech enhancement algorithms are typically used in conjunction with a speech codec, it is important to evaluate the performances of speech enhancement algorithms under such a configuration. In this test, the noisy speech was firstly processed with different speech enhancement algorithms, and then encoded using a narrow band Adaptive Multi-Rate (AMR) speech codec (3GPP, 2001a) working at a fixed bit-rate of 12.2 kbit/s to produce the testing sequences. The noisy speech previously used in the objective evaluation of noise variance estimator as described in Section 4.1 were reused in this test. The Comparison Mean Opinion Score (CMOS) method as prescribed in Annex E of ITU-

Table 3
Comparison category rating used in the CMOS test.

Rating	Score (Quality of second stimulus compared to the first one)
3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

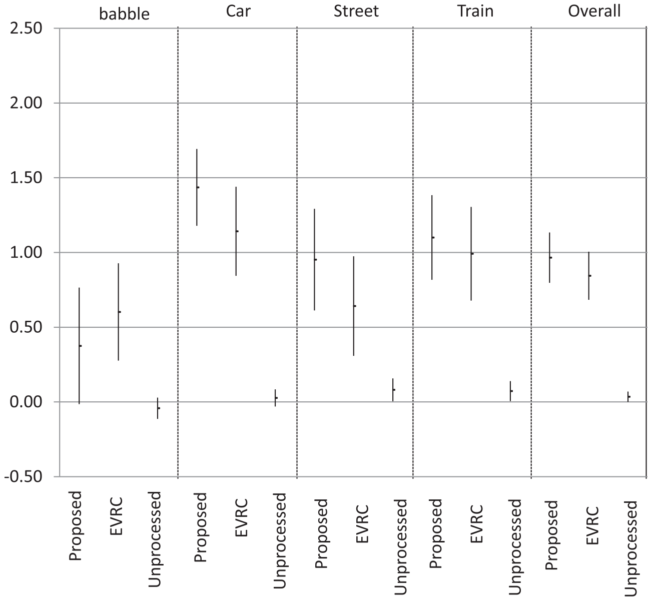


Fig. 7. Listening test results for tandem system (SNR = 6 dB). Results are shown with 95% confidence interval.

T Recommendation P.800 (ITU-T, 1996) was adopted, which can be summarized as follows. For each trial in the test, the subjects listened to the reference speech and the speech to be assessed, and made their assessment by comparing the perceived quality of the two. The ordering of the two speech was changed randomly in the test and the subjects did not know which speech is the one to be assessed. The subjects report their opinions of the quality of the second speech sample compared with the first speech sample using a Comparison category Rating (CCR) as illustrated in Table 3. In our test, the original noisy speech signal encoded with the same speech codec was used as the reference speech.

In total 16 subjects participated in this listening test; the results are summarized in Fig. 7 and Fig. 8 for different noise types and SNR settings. Results are shown with 95% confidence interval. It can be seen from these results that at low SNR setting (SNR = 6 dB), the proposed system achieves better mean CMOS ratings compared to EVRC at most noise types except for babble noise. In addition, it achieves better overall performance when results from all four noise types are combined. Similar observations can be made from results of high SNR setting (SNR = 15 dB), where the proposed algorithm achieves better performance for babble and train noises, and comparable performance for car and street noises when compared to EVRC. In addition, it also achieves better overall performance when results from all four noise types are combined. It can also been seen from those results that both speech enhancement algorithms significantly improved the perceived quality of the decoded speech when compared with system with unprocessed noisy speech, which clearly justified the usefulness of the speech enhancement algorithms when they are used in conjunction with speech codecs.

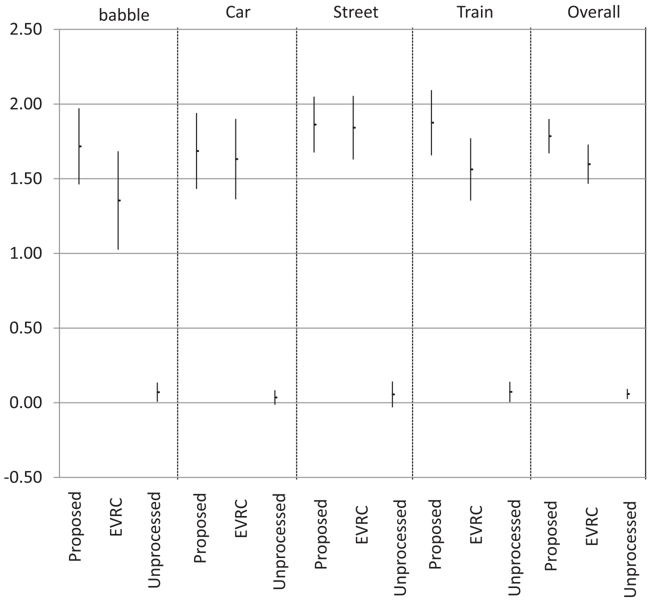


Fig. 8. Listening test results for tandem system (SNR = 15 dB). Results are shown with 95% confidence interval.

5. Conclusions

A speech enhancement algorithm based on a soft audible noise masking principle is proposed in this paper. The proposed algorithm tries to achieve an optimal balance between the audible noise level reduction and the speech distortion by incorporating a psychoacoustic model to determine the exact amount of audible noise to be suppressed, and using a suppression gain that takes both the noise reduction level and speech distortion into account. Good perceptual quality of the proposed algorithm is confirmed by the subjective listening test results using two state-of-art systems as references.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments.

References

3GPP, 2001. Mandatory speech codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; transcoding functions, 3GPP TS 26.090.

3GPP, 2001. Results of the Adaptive Multi-Rate (AMR) noise suppression selection phase, 3GPP TR 26.978.

3GPP2, 2004. Enhanced variable rate codec standard, 3GPP2 C.S0014-0.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustic, Speech and Signal Processing 27, 113–120.

Cappe, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Transactions on Speech and Audio Processing 2, 345–349.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean square error short time spectral amplitude estimator. IEEE Transactions on Acoustic, Speech and Signal Processing 32, 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean square error Log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 33, 443–445.

Erkelens, J.S., Heusdens, R., 2008. Fast noise tracking based on recursive smoothing of MMSE noise power estimates. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Fisher, W.M., Doddington, G.R., Goudie-Marshall, K.M., 1986. The DARPA speech recognition research database: specifications and status. In: Proceedings of DARPA Workshop on Speech Recognition, pp. 93–99.

Gustafsson, S., Jax, P., Vary, P., 1998. A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Hansen, J., Radhakrishnan, V., Arehart, K., 2006. Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system. IEEE Transactions on Audio, Speech, and Language Processing 14, 2049–2063.

Hu, Y., Loizou, P., 2004. Incorporating a psychoacoustical model in frequency domain speech enhancement. IEEE Signal Processing Letters 11, 270–273.

Hu, Y., Loizou, P., 2006. Subjective comparison of speech enhancement algorithms. In: 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

ISO/IEC, 1992. Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part3: Audio, ISO/IEC JTC1/SC29/WG11 IS 11172–3.

ITU-T, 1993. Objective measurement of active speech level – telephone transmission quality objective measuring apparatus, ITU-T Rec. P.56.

ITU-T, 1996. Methods for subjective determination of transmission quality, ITU-T Rec. P.800.

ITU-T, 2001. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, ITU-T Rec. P.862.

ITU-T, 2003. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, ITU-T Rec. P.835.

Jabloun, F., Champagne, B., 2003. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Transacations on Speech and Audio Processing 11, 700–708.

Johnston, J., 1988. Transform coding of audio signals using perceptual noise criteria. IEEE Journal on Selected Areas in Communications 6, 314–323.

Lin, L., Holmes, W., Ambikairajah, E., 2003. Subband noise estimation for speech enhancement using a perceptual wiener filter. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing 9, 504–512.

Princen, J., Johnson, A., Bradley, A., 1987. Subband/transform coding using filter bank designs based on time domain aliasing cancellation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, pp. 2161–2164.

Rice, S.O., 1948. Statistical properties of a sine wave plus random noise. Bell System Technical Journal 27, 109–157.

Tsoukalas, D., Mourjopoulos, J., Kokkinakis, G., 1997. Speech enhancement based on audible noise suppression. IEEE Transactions on Speech and Audio Processing 5, 497–514.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Transactions on Speech and Audio Processing 7, 126–137.

Widrow, B., Stearns, S.D., 1985. Adaptive Signal Processing. Prentice Hall, Englewood Cliffs, NJ.

Wolfe, P.J., Godsill, S.J., 2003. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP JASP on Digital Audio for Multimedia Communications 10, 1043–1051.

Yu, R., 2009. A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4421–4424.