Tracking of Nonstationary Noise Based on Data-Driven Recursive Noise Power Estimation

Jan S. Erkelens and Richard Heusdens

***Abstract*—This paper considers estimation of the noise spectral variance from speech signals contaminated by highly nonstationary noise sources. The method can accurately track fast changes in noise power level (up to about 10 dB/s). In each time frame, for each frequency bin, the noise variance estimate is updated recursively with the minimum mean-square error (mmse) estimate of the current noise power. A time- and frequency-dependent smoothing parameter is used, which is varied according to an estimate of speech presence probability. In this way, the amount of speech power leaking into the noise estimates is kept low. For the estimation of the noise power, a spectral gain function is used, which is found by an iterative data-driven training method. The proposed noise tracking method is tested on various stationary and nonstationary noise sources, for a wide range of signal-to-noise ratios, and compared with two state-of-the-art methods. When used in a speech enhancement system, improvements in segmental signal-to-noise ratio of more than 1 dB can be obtained for the most nonstationary noise sources at high noise levels.**

***Index Terms*—Discrete Fourier transform (DFT)-based speech enhancement, minimum mean-square error (mmse) estimation, noise spectrum estimation, noise tracking.**

I. Introduction

SINGLE-CHANNEL speech enhancement methods based on the discrete Fourier transform (DFT) have received significant interest due to their low complexity and relatively good performance, e.g., [1]–[8]. To work properly, speech enhancement methods need an estimate of the noise power spectral density. For nearly stationary noise, a voice activity detector (VAD) may be used to detect the speech pauses for estimation of the noise spectrum. However, VADs are not reliable at low input signal-to-noise ratio (SNR), and update their estimates not frequently enough for many nonstationary noise sources faced in practice. For better performance, it is necessary to make reliable noise spectrum estimates also during speech activity, but it is a challenging problem to avoid speech power leaking into the noise spectrum estimates. In recent years, several methods for tracking of nonstationary noise sources have appeared in the literature. Rangachari and Loizou [9] give an overview of several methods and discuss some of their limitations. These methods have different ways of coping with the speech leakage problem. An idea that has been proven quite successful is to track the minima of the smoothed noisy spectrum [10], [11].

Manuscript received August 28, 2007; revised May 16, 2008. Published July 16, 2008 (projected). This work was supported by MultimediaN. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Susanto Rahardja.

The authors are with the Department of Mediamatics, Delft University of Technology, 2628 CD, Delft, The Netherlands (e-mail: j.s.erkelens@tudelft.nl).

Digital Object Identifier 10.1109/TASL.2008.2001108

Several methods make use of some kind of minimum tracking procedure [9], [12]–[15]. Approaches based on Kalman filtering [16], [17] and subspace decompositions [15], [18] have also been explored.

The method of Sugiyama *et al.* [19, Ch. 6], is not based on minimum statistics, but uses a weighted noise estimation procedure. The squared noisy amplitudes are down-weighted, depending on the estimated SNR. This allows for continuous noise estimation without severe speech leakage, even in periods of speech presence.

We will briefly recapitulate the minimum statistics (MS) method of Martin [11], the improved minima controlled recursive averaging (IMCRA) method of Cohen [14], and the method of Sugiyama *et al.* in Section III. To reduce the amount of speech leakage, MS and IMCRA need a rather long time window for the minimum tracking. A long time window limits the algorithms' ability to follow a rapid increase in noise level, because in that case the minimum tracking will lag behind by the window length. In Section IV, we will propose to use minimum mean-square error (mmse) estimation of the noise power to update the noise spectrum estimates with a reduced risk of speech leakage. The mmse estimates are obtained with the standard method of multiplying the noisy powers by a spectral gain function. This removes most of the speech contribution from the noisy spectrum, allowing for fast and accurate tracking of changing noise levels. The spectral gain function for noise power estimation is found by an iterative data-driven method, as explained in Section IV-D. We will evaluate the proposed method in Section V and compare with the minimum statistics method and an improved version of the method of Sugiyama *et al.*, in terms of tracking performance and overall performance in a speech enhancement scheme. A summary and concluding remarks follow in Section VI.

II. Modeling Assumptions and Definitions

Before we discuss the MS method of Martin [11], the IMCRA method of Cohen [14], and the method of Sugiyama *et al.* [19, Ch. 6], we will introduce some modeling assumptions and notations and define some quantities of interest.

A. Spectral Modeling

We consider an additive-noise signal model of the form

$$X(k,m) = S(k,m) + N(k,m) \tag{1}$$

where $X(k,m)$,$S(k,m)$, and $N(k,m)$ are complex-valued random variables representing the short-time DFT coefficients obtained at frequency index k in signal frame m from the noisy speech, clean speech, and noise process, respectively.

We apply the standard assumption that $S(k,m)$ and $N(k,m)$ are statistically independent across time and frequency as well as from each other. For ease of notation, we therefore drop the time and/or frequency index when this does not cause confusion. The noisy amplitude is $R = |X|$, the speech spectral amplitude is $A = |S|$, and the noise amplitude $D = |N|$. The noise DFT coefficients N are assumed to follow a complex Gaussian distribution with variance λ_D. We will call D^2 the *(instantaneous) noise power*. Its expectation is λ_D. Similarly, the speech spectral variance λ_S is the expectation of the speech power A^2. The *prior SNR* ξ and the *posterior SNR* ζ are defined as

$$\xi(k,m) = \frac{\lambda_S(k,m)}{\lambda_D(k,m)}, \qquad \zeta(k,m) = \frac{R^2(k,m)}{\lambda_D(k,m)} \tag{2}$$

respectively.

B. Amplitude Estimation

A common method to estimate the speech amplitudes A is to multiply the noisy amplitudes R by a spectral gain function. In general, any power A^β of the speech amplitude can be estimated by applying a suitably chosen β-order gain function [20]

$$\widehat{A^\beta} = G_{A^\beta}(\xi,\zeta)R^\beta \tag{3}$$

where G_{A^β} depends on the assumed statistical models for the speech and the noise and on the criterion that is optimized for. Later on, we will estimate the noise power D^2 by means of a gain function G_{D^2}. Note that if the *same* type of prior distribution is assumed for the speech and noise DFT coefficients (for example, both Gaussian), G_{D^β} follows from G_{A^β} by interchanging the roles of λ_S and λ_D. We will not make this assumption in this paper, however.

III. State-of-the-Art Noise Tracking Algorithms

A. Minimum Statistics Method

The MS method [10], [11] uses the minima of the smoothed periodogram of the noisy speech to estimate the noise level for each frequency bin. The speech energy is frequently zero during speech pause and in between words and syllables. Also, in certain frequency bins, the speech power may be much smaller than the noise power. Therefore, the minima of the smoothed noisy periodogram in a finite window that is large enough to bridge high-power speech segments can be used to estimate the noise floor. A typical size of the window is in the order of 1 s. The method uses a time-varying smoothing parameter to reflect the degree of stationarity of the noisy signal. Furthermore, since the minimum values have an expected value that is smaller than the mean power level, a bias correction procedure is implemented. The method does not have to rely on a VAD.

The method has two main shortcomings. First, since the minimum value in a window is used, the estimates of the noise variance lag behind by about the window length in case of increasing noise power. This tracking delay can limit the performance of the method for very nonstationary noise sources. Second, it is difficult to find the correct bias compensation factor. The bias compensation factor can be derived under the assumptions that the window contains only noise. However, in practice, often large fractions of the window contain noisy speech. The periodogram values which contain speech power are less likely to be the minimum value in the window than those that contain only noise power. Therefore, a minimum is effectively the minimum of a fraction of the periodogram values. To find the correct bias compensation factor, an estimate of this effective number of noise periodogram values should be made.

B. Improved Minima Controlled Recursive Averaging Method

In the IMCRA method [14], a noise variance estimate $\hat{\lambda}_D$ is obtained by recursive smoothing of the noisy power

$$\hat{\lambda}_D(k,m) = \alpha_s(k,m)\hat{\lambda}_D(k,m-1) + (1-\alpha_s(k,m))\,R^2(k,m). \tag{4}$$

The smoothing parameter $\alpha_s(k,m)$ depends on an estimate $\hat{p}(k,m)$ of the speech presence probability

$$\alpha_s(k,m) = \alpha_d + (1-\alpha_d)\hat{p}(k,m) \tag{5}$$

where α_d lies between 0 and 1. Equation (5) means that α_s always lies between α_d and 1. When the speech presence probability estimate $\hat{p}$ is near 1, so is α_s, and then the noise estimate is kept close to its previous value, preventing speech power to leak into the noise variance estimate. On the other hand, the lower the speech presence probability estimate, the faster the noise variance is updated.

An accurate estimate of $p(k,m)$ is needed to avoid speech leakage. The estimate of $p(k,m)$ in IMCRA is controlled by the minima values of a smoothed power spectrum of the noisy signal. Apart from exponential smoothing in the time direction, some averaging over neighboring frequency bins is performed, taking into account the strong correlation of speech presence in neighboring frequency bins of consecutive frames [21]. A fixed bias compensation factor is used for the minima of the smoothed spectrum. IMCRA uses two iterations of smoothing and minimum tracking, in order to make the minimum tracking during speech activity more robust.

As for the MS method, the IMCRA method reacts slowly to an increase in the noise level. Its performance is also influenced by the accuracy of the bias compensation applied, albeit in a less direct fashion than MS because (4) is used for noise variance estimation.

C. Method of Sugiyama, Kato, and Serizawa

The tracking methods described above are based on the use of minimum statistics. Sugiyama, Kato, and Serizawa [19, Ch. 6] proposed a method based on a different principle. In this method, to which we will refer as *the SKS method*, the noisy power $R^2(k,m)$ is weighted by a factor $W(k,m)$ that depends on the estimated posterior SNR, as follows:

$$W(k,m) = \begin{cases} 1 & : \ \hat{\zeta}(k,m) \le \zeta_1 \\ \left(\zeta_2 - \hat{\zeta}(k,m)\right)/(\zeta_2-\zeta_1) & : \ \zeta_1 < \hat{\zeta}(k,m) < \theta_Z \\ 0 & : \ \hat{\zeta}(k,m) \ge \theta_Z \end{cases} \tag{6}$$

where $\zeta_1 < \theta_Z < \zeta_2$. The estimated posterior SNR $\hat{\zeta}(k,m)$ follows from (2) by substituting an estimate $\hat{\lambda}_D(k,m)$ of the noise spectral variance. The estimated noise variance $\hat{\lambda}_D(k,m)$ is taken in the SKS method as the average (over time) of the last L_Z nonzero values of $W(k,m)R^2(k,m)$. Larger values of R^2 are down-weighted, reducing the amount of speech leakage. The weighting function $W(k,m)$ incorporates a hard decision about speech presence: when $\hat{\zeta}(k,m) \geq \theta_Z$, speech presence is implicitly assumed by setting $W(k,m) = 0$ and the corresponding $R^2(k,m)$ are not used for updating the noise variance estimates. This tracking method has been shown to perform well for nonstationary noise sources, because continuous updating is possible even in speech periods.

It has some drawbacks, however. First, the noise variance estimates produced by this method are biased low in noise-only regions because W is always less than or equal to 1. Fortunately, this problem is easily solved by applying a bias correction factor $C > 1$, which can be found from simulations. By generating realizations of R^2 with expected value 1, C can be calculated as 1 over the mean value of the nonzero weighted R^2 values. We will use this bias compensation in our evaluations, because it improves the performance of the method. A second drawback of the SKS method lies in the fact that the weighting function W is heuristic. Therefore, the noise variance estimator is not optimal in, for example, an mmse sense. Third, in case of a sudden large increase in noise level, the estimated noise variance will lag behind, causing $\hat{\zeta}$ to be severely overestimated. Most R^2 values are consequently down-weighted to zero and not used for updating the noise variance estimates. The result is a slow response of the algorithm at such sudden large increases in noise level. This problem can be satisfactory solved by means of a simple and effective safety net that we propose in Section IV-C.

IV. Using MMSE Estimation of the Noise Power to Reduce Speech Leakage

The slow response of MS and IMCRA to increasing noise levels is a result of using minimum values in a window of considerable length in order to prevent speech power from leaking into the noise variance estimates. The key idea to the method we will develop below is to avoid using the noisy power R^2 directly by removing as much as possible of the speech contribution from it, *before* smoothing with an equation like (4). In other words, we propose to replace $R^2(k,m)$ in (4) by an estimate of the noise power $\widehat{D^2}(k,m)$

$$\hat{\lambda}_D(k,m) = \alpha_s(k,m)\hat{\lambda}_D(k,m-1) + (1-\alpha_s(k,m))\,\widehat{D^2}(k,m). \tag{7}$$

This way, less speech power will leak into the noise variance estimate, and the noise variance estimate can be more reliably updated during speech activity. Consequently, the speech presence probability estimator does not need to be extremely accurate, the smoothing parameter needs to be close to 1 less frequently and faster tracking can be achieved. The principle is similar to that underlying the SKS method. The differences are that we use a time-varying smoothing parameter that is controlled by estimates of speech presence probability, and that we will use for $\widehat{D^2}$ the mmse estimator of the noise power D^2 instead of applying the heuristic weighting of (6). An iterative data-driven method is used to find the optimal gain function G_{D^2} (Section IV-D). The smoothing parameter α_s depends on an estimate of speech presence probability, as in (5), but a simplified estimation procedure for $\hat{p}$ is used that allows for faster tracking (Section IV-A).

In some sense, the problem of estimating D^2 is the reverse of that of estimating A^2. Estimation of speech characteristics is easiest at high SNR, while the opposite is true for the noise. The problem is not symmetrical, though. We are not interested in reconstructing the instantaneous noise power, but only in estimating its expectation, i.e., the noise spectral variance. A Gaussian noise model is often used, while super-Gaussian models for the speech give better results [5], [6], [8]. Furthermore, speech characteristics change constantly and speech contains many short pauses. We assume that the noise properties change more slowly than those of the speech. This allows for the use of exponential smoothers in (4) and (7), which respond slower to changes in noise level than the estimated prior SNR does to changes in speech variance. The advantage of using exponential smoothers is a reduction in the variance. There is a limit to how accurate we can estimate the noise variance; however, it depends on how reliable we can detect speech presence. There is also a limit to how fast noise can be tracked: if we react to the most abrupt changes in the noisy spectrum, we start to track the speech, resulting in overestimation of the noise variance (the speech leakage problem) and too much suppression in an enhancement setting.

A. Speech Presence Probability Estimation

Because we update the noise variance estimate with an estimate of the noise power instead of with the noisy power, the risk of speech leaking into the noise estimates is reduced. Therefore, errors in the speech presence probability estimates have less severe consequences and we can use a simpler speech presence probability estimator. First, the posterior SNR is smoothed over a few neighboring frequency bins to take into account the strong correlation of speech presence in neighboring frequency bins [14]

$$\tilde{\zeta}(k,m) = \sum_{i=-w}^{w} b(i)\zeta(k-i,m), \quad \text{with} \sum_{i=-w}^{w} b(i) = 1. \tag{8}$$

A rectangular window with $w = 1$ is used for $b(i)$ in this work (see Section V-C for the settings of all other parameters). Next, a hard decision about speech presence is made

$$\begin{aligned} &\text{if } \tilde{\zeta}(k,m) > T(k,m) \\ &\quad I(k,m) = 1 \quad \text{speech present} \\ &\text{else} \\ &\quad I(k,m) = 0 \quad \text{speech absent} \\ &\text{end.} \end{aligned} \tag{9}$$

The speech presence probability estimate is updated with a first order recursion:

$$\hat{p}(k,m) = \alpha_p \hat{p}(k,m-1) + (1-\alpha_p)I(k,m) \tag{10}$$

where α_p lies between 0 and 1. This estimate is used in (5) to find the smoothing parameter α_s in (7). This procedure for calculating α_s is similar to that in [9]. There, the ratio of the smoothed noisy spectrum and its local minimum is compared in (9) against a threshold. The local minimum in [9] is tracked by the method of Doblinger [12], with an adaptation time of the minimum tracking of about 0.5 s for nonstationary noise. Since we only use the posterior SNR of the current time frame in (9), we can react almost instantaneously to changing noise levels. The parameter T controls the tradeoff between the tracking speed and the amount of speech leakage. The higher the value, the faster the tracking speed, but the higher the risk of speech leakage.

B. Prior SNR Estimation

The gain functions take the prior and posterior SNRs as arguments. These parameters are unknown in practice and have to be estimated. We have found that the noise tracking performance depends on the particular prior SNR estimator used. While the "decision-directed" estimator [2] is very suitable for speech spectral amplitude estimation, we found that a modified estimator improves noise tracking performance. We will therefore use different estimators for the speech estimation and noise tracking tasks.

Prior SNR Parameter $\hat{\xi}_{NT}$ for Noise Tracking: The standard "decision-directed" estimator is the most commonly used estimator of prior SNR

$$\hat{\xi}(k,m) = \max\left[\alpha \frac{\hat{A}^2(k,m-1)}{\hat{\lambda}_D(k,m)} + (1-\alpha)\left[\frac{R^2(k,m)}{\hat{\lambda}_D(k,m)} - 1\right], \xi_{\min}\right] \quad (11)$$

where $\xi_{\min}$ is a small value larger than 0. This estimator leads to less musical noise, than, for example, the maximum-likelihood (ML) estimator [2], [22]. Because of the recursive nature of this estimator, $\hat{A}^2(k,m-1)$ depends on all previous noisy amplitudes $R(k,m-j), j>0$. We can say that $\hat{A}^2(k,m-1)$ *summarizes* the knowledge we have about the current speech spectral variance $\lambda_S(k,m)$ from previous noisy amplitudes. The estimator in (11) implicitly makes use of the significant correlation that exists between speech spectral amplitudes of consecutive frames [23]. However, speech spectral amplitudes are assumed independent of noise spectral amplitudes. This suggests that (11) may be less useful for the noise tracking task where we want to estimate $\lambda_D(k,m)$. However, there are more reasons to look for a modified prior SNR parameter for noise tracking.

First, (11) is delayed at speech onsets and offsets [22]. At an onset, $\hat{\xi}$ is too small and the SNR is underestimated. Therefore, the noise amplitude will be overestimated and speech power will leak into the noise variance estimates. Similarly, the noise variance will be underestimated at speech offsets. Second, any errors in the estimated noise variances will in turn affect the following prior and posterior SNR estimates. The decision-directed prior SNR estimates will be affected more than the posterior SNR estimates for the following reason. If λ_D is overestimated (underestimated), ξ and ζ will be underestimated (overestimated). This means that the gain function $G_A(\hat{\xi},\hat{\zeta})$ will suppress too much (too little), causing the errors in $\hat{\lambda}_D$ and $\hat{A}^2$ to be negatively correlated. Therefore, the error in $\hat{\lambda}_D$ will tend to amplify itself in the first term of (11). The error will also be amplified in the second term, because $R^2/\hat{\lambda}_D - 1$ equals $(R^2 - \hat{\lambda}_D)/\hat{\lambda}_D$, and therefore the errors in numerator and denominator are also negatively correlated in this term. As input to G_{D^2}, we will therefore use the following parameter $\hat{\xi}_{NT}$, which is less sensitive to errors in $\hat{\lambda}_D$:

$$\hat{\xi}_{NT}(k,m) = \max\left[\alpha_{NT}\frac{R^2(k,m-1)}{\hat{\lambda}_D(k,m)} + (1-\alpha_{NT})\frac{R^2(k,m)}{\hat{\lambda}_D(k,m)}, \xi_{\min}\right] \quad (12)$$

where we will use the latest available estimate of the noise variance $\hat{\lambda}_D(k,m)$. This estimator can be viewed as a smoothed version of posterior SNRs (using the latest available noise variance estimates). As such, it resembles the ML estimator without the -1 correction term. The ML estimator commonly uses less smoothing, though, when used for speech spectral amplitude estimation, because it has to react quickly to changing speech characteristics. In our noise tracking algorithm, we aim at tracking signals with spectra that vary more slowly than those of the speech, and we must therefore use more smoothing in (12).

$\hat{\xi}_{NT}$ would follow from $\hat{\xi}$ in (11) by using a gain function G_A for the first term that is completely flat, i.e., identically equal to 1, and by removing the -1 correction in the second term. It is clear that $\hat{\xi}_{NT}$ is not an unbiased estimator of λ_S. However, this is not a problem, because the gain function G_{D^2} will be adapted to this parameter by means of a data-driven method (Section IV-D) and the bias will be compensated for.

C. Safety Net

In Section V, we will show that our noise tracker can easily follow very fast changes in noise power level up to about 10 dB/s. However, if the noise level increases even much faster than that, for example, when it suddenly jumps to a high level and stays at that level, $\tilde{\zeta}$ in (8) will be calculated on the basis of a noise variance estimate which is too low. It therefore becomes more likely that a speech presence decision is made in (9), and the algorithm will react slowly. The SKS method also suffers from this problem. We therefore propose a simple and effective safety net, which ensures that these algorithms continue to work properly also under such extreme conditions. As will be shown in Section V, the safety net does not affect the performance of the algorithms under normal conditions in any negative way, while greatly improving the performance for sudden large increases in noise level.

The idea is to push the noise variance estimate into the right direction when we detect that its value is much too low. As a reference value, we use the minima $P_{\min}(k,m)$ of the smoothed values $\overline{P}(k,m)$ of the noisy power $R^2(k,m)$ in a short window of length $w_{\min}$, where $\overline{P}(k,m)$ is given by

$$\overline{P}(k,m) = \eta\overline{P}(k,m-1) + (1-\eta)R^2(k,m) \quad (13)$$

where η is a small smoothing parameter. After updating $\hat{\lambda}_D$ with (7), we check whether it fulfills the following condition:

$$B \cdot P_{\min}(k,m) < \hat{\lambda}_D(k,m) \quad (14)$$

where $B > 1$ is a correction factor. In case of a large increase in noise level that the algorithm cannot follow, $B \cdot P_{\min}(k,m)$ will become larger than $\hat{\lambda}_D(k,m)$ after a time of the order of the window length. If that happens, we reset the $\hat{\lambda}_D(k,m)$ values that violated (14) to $\max[B \cdot P_{\min}(k,m), \widehat{D^2}(k,m)]$, and the corresponding $\hat{p}(k,m)$ to 0. The factor B is taken larger than 1, but much smaller than the bias correction that would apply if the window $w_{\min}$ would contain only noise. This ensures that the safety net will not unintentionally come into action when some speech energy leaks into $P_{\min}$. We use very little smoothing of R^2 values (small η) to compute the minimum $P_{\min}$, because that allows us to keep the window $w_{\min}$ short. We have observed that the value of B and the window length are not very critical for good performance, but a window length of at least 0.5 s is required.

This safety net is based on a simplified minimum statistics scheme, but uses a much smaller smoothing parameter (η close to 0). We can do that because we are not trying to derive an accurate noise variance estimate from $P_{\min}$, for which much larger values of η (and B) are required (Martin [10] recommends $\eta \geq 0.9$ for window lengths $w_{\min} \geq 0.8$ seconds). Since we only use $P_{\min}$ to speed up the convergence of the algorithm in case of sudden large increases in noise level, this safety net does not lead to a significant increase in speech leakage, even at high SNRs, as can be seen from Figs. 2 and 3 in Section V-E.

D. Finding the Gain Function for Noise Power Estimation

For $\widehat{D^2}$, we would ideally like to use the mmse estimator, because that is an unbiased estimator [24] of λ_D with low variance. However, the optimal gain function is very hard to derive analytically because of several reasons. First, there are strong indications that the probability density function of speech DFT coefficients differs from a Gaussian model [5], [6], [8]. This may lead to mathematical complications when we want to derive the mmse estimator of D^2. Second, analytical gain functions are always derived under the assumption that speech and noise variances are known, but in practice these are estimated, affecting the optimality of the gain function. We therefore resort to a data-driven method to find the gain function. We will make use of the method in [25], but in an iterative fashion. This method makes no explicit assumptions about the speech statistics and can also take into account the influence of estimation inaccuracies in the estimated speech and noise variances. The following subsection starts with recapitulating very briefly the basic method.

Iterative Data-Driven Gain Optimization: The method in [25] makes use of a large training database of speech material, contaminated with various levels of stationary white Gaussian noise of known SNR. For all training data, the prior and posterior SNRs are calculated for every time frame and every frequency index. Their values are discretized on a grid, typically in 1-dB steps. Each $(\hat{\xi}, \zeta)$-pair has a corresponding (D^2, R^2)-pair associated with it. Statistics are collected for all training data, and afterwards one scalar gain value G_{D^2} is computed for each grid cell such that the mean-square error between the D^2 and $\widehat{D^2}$ is minimized. The result is a two-dimensional table of optimal gain values, indexed by estimated prior SNR and posterior SNR.

The grid used in this paper covers the range [-19 dB, 40 dB] for ξ and [-30 dB, 40 dB] for ζ, both in steps of 1 dB. The training speech data consisted of about 25% of the TIMIT-TRAIN [26] database. To each file, white noise has been added at several SNRs, from -12.5 to 27.5 dB in steps of 5 dB. Noise only frames are not taken into account: to exclude the silence intervals, frames with a clean energy more than 40 dB below the maximum clean frame energy of a speech sentence are not taken into account for optimization of the gain function.

This data-driven method is not directly applicable for our noise estimation problem, however. During the training, the noise variance λ_D is known. When the resulting gain table G_{D^2} is used with the noise tracking method outlined earlier in this section, the noise variance is unknown but is estimated using G_{D^2}. The input parameters to the gain function depend on the quantity $\hat{\lambda}_D(k,m)$, which has been computed using the same gain function. In other words, a nonlinear recursion is introduced which was ignored in the training. Fortunately, the gain function can still be optimized while taking into account the recursion by means of an iterative scheme proposed next.

Let the value of any quantity in the ith iteration be denoted by a subscript i. For example, $G_{D^2,i}$ is the gain function for D^2-estimation in the ith iteration. To break the recursion, $G_{D^2,i}$ is only used to compute data to be used in the *next* iteration. The input parameters of $G_{D^2,i}$, $\hat{\xi}_{NT,i}(k,m)$ and $\hat{\zeta}_i(k,m)$, depend only on data computed in the *previous* iteration. The optimization procedure is as follows:

0) Initialization ($i = 0$):
$\widehat{D^2}_i(k,m) = D^2(k,m)$, $\hat{\lambda}_{D,i}(k,m) = \lambda_D(k,m)$
1) Compute $\hat{\zeta}_i$, $\hat{\xi}_{NT,i}$:
$\hat{\zeta}_i(k,m) = R^2(k,m)/\hat{\lambda}_{D,i}(k,m)$, $\hat{\xi}_{NT,i}(k,m) = \max[\alpha_{NT}(R^2(k,m-1)/\hat{\lambda}_{D,i}(k,m)) + (1 - \alpha_{NT})(R^2(k,m)/\hat{\lambda}_{D,i}(k,m)), \xi_{\min}]$
Collect (D^2, R^2) statistics per grid cell;
Update $\alpha_{s,i}(k,m)$ according to (5), (8)–(10), and $\hat{\lambda}_{D,i}$:
$\hat{\lambda}_{D,i}(k,m+1) = \alpha_{s,i}(k,m)\hat{\lambda}_{D,i}(k,m) + (1 - \alpha_{s,i}(k,m))\widehat{D^2}_i(k,m)$
$m := m + 1$;
Complete step 1) for all training data;
2) Minimize the mse in $\widehat{D^2}$ for each grid cell
$\Rightarrow G_{D^2,i+1}(\xi_{NT}, \zeta)$
3) Compute data for the next iteration:
$\widehat{D^2}_{i+1}(k,m) = G_{D^2,i+1}(\hat{\xi}_{NT,i}(k,m), \hat{\zeta}_i(k,m))R^2(k,m)$
4) $i := i + 1$;
Go to step 1) if $i < i_{\max}$.

This scheme typically converges in less than $i_{\max} = 7$ iterations. We do not apply the safety net of Section IV-C in step 1), which is unnecessary as no sudden large jumps in noise level occur during the training.

The $\widehat{D^2}_i$ are initialized with the true noise powers D^2 ("noise initialization"). Alternatively, they can be initialized with the noisy power R^2 ("noisy initialization") or even the

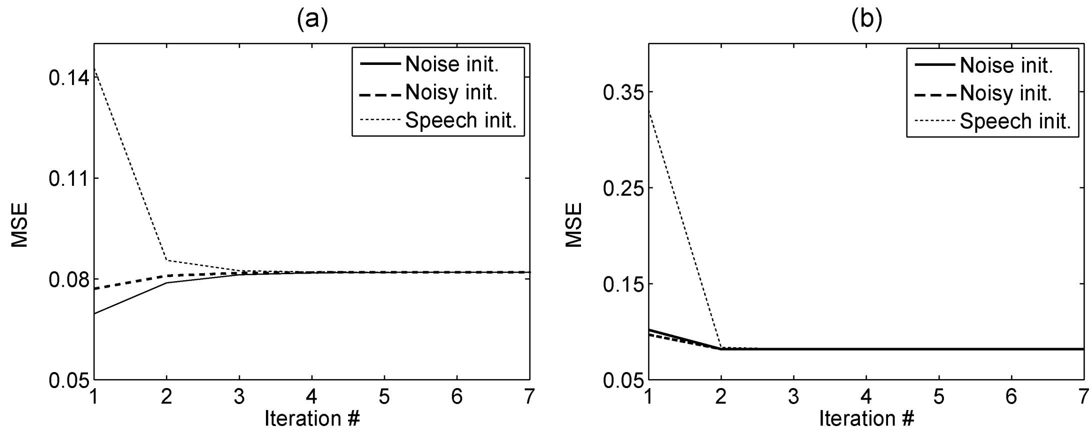


Fig. 1. Normalized mse for noise, noisy and speech initialization (a) during the iterative training procedure after step 2) and (b) when the optimized gain function is applied recursively.

speech power A^2 ("speech initialization"). Fig. 1(a) shows the normalized mse in $\widehat{D^2}$ after step 2), for these three different initializations. The normalized mse's have been computed as

$$\text{mse} = \frac{\sum_u \sum_v \sum_w \left\{ D^2(u,v,w) - \widehat{D^2}(u,v,w) \right\}^2}{\sum_u \sum_v \sum_w D^4(u,v,w)} \tag{15}$$

where the indices u and v run over all (ξ_{NT}, ζ)-cells, and the index w over all data collected in each cell. In the noise and noisy initialization cases, the mse increases with each iteration step, while for speech initialization, the mse decreases after each iteration. In all cases does the mse converge to the same end value 0.082. For noise and noisy initialization, the estimated noise power becomes a bit worse after each iteration, while it becomes better when speech initialization is used, but the gain function and mse converge to the same result.

The question arises how this iterative scheme succeeds in optimizing for the practical case when there are recursions (i.e., the output of the gain function in the current time frame is used in the calculation of the inputs for the next time frame). Convergence means that $G_{D^2,i}$ changes less and less from one iteration to the next when i increases. It also means that the differences between $\widehat{D^2}_{i+1}$ and $\widehat{D^2}_i$ become smaller and smaller. However, when $\widehat{D^2}_{i+1}$ and $\widehat{D^2}_i$ become almost equal, we have nearly the *same* situation as *with* the recursion. Although we cannot formally prove convergence, we can make a compelling argument that we find the lowest possible mse. When we use noise initialization, we know the mse after step 2) has to increase, because in practice our noise power estimates will always be contaminated to some extent by speech power. On the other hand, when we use the other extreme, i.e., speech initialization, the mse after step 2) will decrease because our estimates in practice will not be totally contaminated by the speech. We see in Fig. 1(a) that these two extremes converge to the same mse; there is no gap. This indicates that we cannot get more contamination in the noise initialization case and also not less contamination for speech initialization. In fact, when we use the optimized gain functions $G_{D^2,i}$ on the training data *recursively*, we see in Fig. 1(b) that the mse's become lower for all different initializations and converge to almost exactly the same value 0.082 as converged to with the iterative scheme. It seems therefore likely that the iterative optimization finds a gain function and corresponding data $\widehat{D^2}$ with the lowest possible amount of contamination.

The gain function found from the original noniterative data-driven scheme described in the beginning of this section, called $G_{D^2,0}$, is optimized using the true noise level. It achieves an mse of 0.23 when applied to the data recursively. This shows that our iterative optimization procedure clearly gives a better gain function, because it decreases the mse by as much as 65% over that obtained with $G_{D^2,0}$. In Section V-E, we will further quantify the improvement in noise tracking performance from iteratively updating the gain for various noise sources.

In the next section, the new noise tracking method using $G_{D^2,i_{\max}}$ will be compared with the MS and SKS methods.

V. Experimental Results

A. Experimental Setup

To evaluate the noise tracking performance of the proposed method, we concatenated sentences from the TIMIT-TEST database. Beginning and trailing silences were removed prior to concatenation, so there are no intervening pauses. The total length of the speech material is about 320 s. An equal number of male and female speakers have been used. All signals used in this work have been limited to 8-kHz sampling frequency and telephone bandwidth (300–3400 Hz). The noise recordings have been taken from the NOIZEUS corpus [27]. In addition, computer-generated white noise is also used. Noise tracking performance is measured directly and also in an enhancement system.

B. Enhancement System

The enhancement algorithm is based on mmse estimation in the DFT domain of speech spectral amplitudes. We used 50%-overlapping frames of 32 ms (256 samples) and a cosine-squared analysis window.

1) Prior SNR Estimator $\hat{\xi}_{SE}$ for Speech Enhancement: For speech estimation, we use the "decision-directed" estimator

[2], not in its original form (11), but with a bias correction [25]

$$\hat{\xi}_{SE}(k,m) = \max\left[\alpha_{SE}\frac{\widehat{A^2}(k,m-1)}{\hat{\lambda}_D(k,m)} + (1-\alpha_{SE})\left[\frac{R^2(k,m)}{\hat{\lambda}_D(k,m)} - 1\right], \xi_{\min}\right] \quad (16)$$

where we will use the latest available estimate of the noise variance $\hat{\lambda}_D(k,m)$. Note that the speech power estimate $\widehat{A^2}$ is used in the first term instead of the square of the amplitude estimate $\hat{A}^2$ (the original definition). An advantage of the alternative definition in (16) is that the estimate of prior SNR does not depend on the final amplitude estimate used for speech reconstruction. This prevents the prior SNR estimator from changing its behavior when another estimator for the speech amplitude is preferred, for example the log-spectral amplitude estimator [3], or any other perceptually relevant amplitude estimator [7].

Another advantage of using $\widehat{A^2}$ is that it reduces a bias that leads to the underestimation of prior SNR when α_{SE} is near 1 and the SNR is low [25]. An experimental comparison with the original definition showed that, for parameter settings for which both definitions have the same tradeoff between noise reduction and speech distortion, the new definition leads to less musical noise [28].

2) Amplitude Gain Functions: The gain functions for $\hat{A}$ and $\widehat{A^2}$ used in this paper are based on a generalized-Gamma speech amplitude prior. The generalized-Gamma prior is given by

$$f_A(a) = \frac{\gamma\beta^{\nu}}{\Gamma(\nu)}a^{\gamma\nu-1}\exp(-\beta a^{\gamma}),\ \beta>0, \gamma>0, \nu>0, a\geq 0 \quad (17)$$

where $\Gamma(.)$ is the gamma function, and β depends on γ, ν, and λ_S. The random variable A represents the DFT magnitude. This prior better models the heavy-tailed nature of observed speech amplitude distributions [5]. The corresponding MAP [5] and mmse [8] amplitude estimators improve on the traditional Rayleigh prior (corresponding to $\gamma = 2$, $\nu = 1$). We will use the mmse gain functions for $\gamma = 1$ and $\nu = 1$ for which the expressions can be found in [8]. For these parameter values, we have $\beta = \sqrt{2/\lambda_S}$.

C. Noise Tracking

We briefly recapitulate here the steps taken in our noise tracking algorithm: First, the prior SNR parameter $\hat{\xi}_{NT}(k,m)$ in (12) and the posterior SNR $\hat{\zeta}(k,m)$ are estimated, using the latest available noise variance estimate $\hat{\lambda}_D(k,m)$. Next, the speech presence probability estimate $\hat{p}$ is updated using (8)–(10). The speech presence probability determines the smoothing parameter α_s in (5). The noise variance estimate is now updated using (7), where $\widehat{D^2}$ is computed with the gain function found in Section IV-D

$$\widehat{D^2}(k,m) = G_{D^2,i_{\max}}(\hat{\xi}_{NT}(k,m), \hat{\zeta}(k,m))R^2(k,m). \quad (18)$$

Finally, the safety check of Section IV-C is performed. For the speech spectral amplitude estimation, we compute prior SNR $\hat{\xi}_{SE}$ from (16) and recompute posterior SNR using the new noise variance estimate.

Parameter Settings: The following parameter settings are used in the experiments: α_d in (5) is set to 0.85, $\alpha_p = 0.1$ in (10), and $T(k,m) = 4$ in (9) independent of time and frequency. We have used $w = 1$ and $b(i) = 1/(2w+1)$ in (8). The same value 0.98 is used for the smoothing parameters α_{NT} in (12) and α_{SE} in (16), and $\xi_{\min}$ is set to -19 dB. We use $\eta = 0.1$ in (13), $B = 1.5$ in (14), and the length of $w_{\min}$ spans 0.8 s.

For the minimum statistic method, we will use a minimum-search window of 1.5 s. For the SKS method, the parameter settings recommended by its authors are used: $\zeta_1 = 0$ dB, $\zeta_2 = 10$ dB, $\theta_Z = 7$ dB, $L_Z = 20$. In addition, we apply a bias compensation factor $C = 1.69$.

D. Performance Measures

Three quality measures will be used to evaluate the noise tracking algorithms. The noise tracking performance will be evaluated directly by means of the segmental logarithmic estimation error LogErr [18]. The performance in the speech enhancement system described above is expressed in terms of segmental SNR improvement (SSNR+) and perceptual evaluation of speech quality (PESQ) [29].

The segmental logarithmic estimation error LogErr is defined as [18]

$$\text{LogErr} = \frac{1}{|\mathcal{M}|K}\sum_{m\in\mathcal{M}}\sum_{k}\left|10\log_{10}\left[\frac{\overline{\lambda}_D(k,m)}{\hat{\lambda}_D(k,m)}\right]\right| \quad (19)$$

where K is the number of frequency bins. We left out frames which don't contain noise in the computation of LogErr, that is, frames with a noise energy more than 40 dB below the noise energy of the frame with maximum noise energy were not included in the index set $\mathcal{M}$. $|\mathcal{M}|$ is the cardinality of $\mathcal{M}$. Prior to evaluating the distortion, the true noise power is smoothed in time, lowering its variance [11]

$$\overline{\lambda}_D(k,m) = 0.9\overline{\lambda}_D(k,m-1) + 0.1D^2(k,m). \quad (20)$$

$\overline{\lambda}_D$ is used as the ideal reference in (19).

We prefer LogErr over the (segmental) relative estimation error [14] which is sometimes used, for the following reasons. The relative estimation error is very sensitive to outliers [9] and may be dominated by just a few frames with low λ_D and high λ_S. Furthermore, errors in the estimated noise level due to tracking delays are penalized less for increasing noise levels than for decreasing noise levels. However, many noise estimation methods react slowly to increasing noise levels. Improvements in tracking speed are therefore of much interest but are not very well accounted for with the relative estimation error. The LogErr penalizes errors at increasing and decreasing noise levels more symmetrically. This could be one reason why the relative estimation error has been found to correlate poorly with

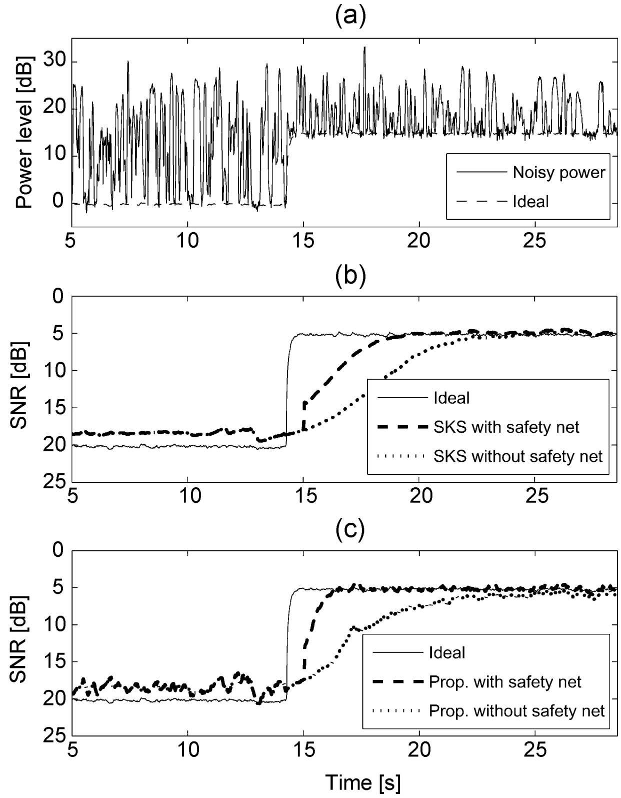


Fig. 2. Reaction of the SKS method (b), and the proposed algorithm (c) to an instantaneous 15-dB increase in noise level, with and without application of the safety net. Panel (a) shows the power level of the noisy speech and the smoothed noise level (ideal reference in (20)).

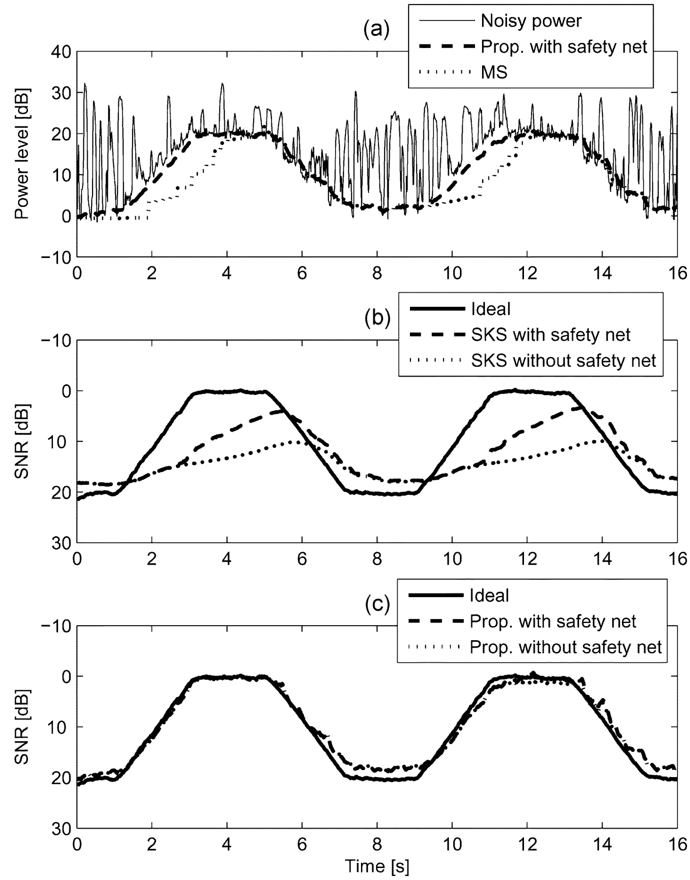


Fig. 3. Noise tracking performance of MS, SKS, and the proposed method on speech contaminated with highly nonstationary white Gaussian noise. The noise level varies between SNRs of 20 and 0 dB, changing at a rate of 10 dB/s. Panel (a) shows the power level of the noisy speech and the noise level estimated by MS and the proposed method. Panels (b) and (c) show the influence of the safety net on the tracking performance of SKS and the proposed method, respectively.

subjective preference tests [9]. One should also consider the correlation of noise estimation errors with the speech power level. The tracking errors are likely to be larger for high speech levels, but smaller for weaker speech components. Therefore, the perceptual influence of speech leakage is probably overestimated by the relative estimation error.

Objective enhancement quality was measured by means of the improvement in Segmental SNR over the noisy signal ($SSNR+$) and $PESQ$. For the computation of $SSNR+$, only frames that contain speech are taken into account, i.e., frames with a clean energy more than 40 dB below the maximum clean frame energy of a speech sentence are not considered.

E. Evaluations

1) Effectiveness of the Safety Net: As was explained in Section IV-C, the SKS method and the proposed method will react quite slowly to sudden, large jumps in the noise level. For these cases, the safety net ensures that the algorithms continue to work properly. The effectiveness of the safety net is illustrated in Fig. 2, where the noise level suddenly increases with 15 dB from one frame to the next, and stays at that higher level. Panel (a) shows the power level of the noisy speech and the ideal noise reference. Panels (b) and (c) show the ideal reference and the response of the SKS and proposed method, respectively, with and without the application of the safety net. All levels are averages over all frequency bins. The response of the algorithms is much accelerated, while it is seen that the noise estimation in stationary-noise parts is not negatively affected, even at an SNR[1] of 20 dB. It can be seen that in the first part of the signal where the SNR is high, both the SKS method and the proposed method show signs of some speech energy leaking into the noise variance estimates. It is clearly seen that the SKS method uses much more smoothing in time than the proposed method, limiting its tracking speed.

Highly Nonstationary Noise: The noise tracking method presented above allows for very fast noise tracking. As a first example, a highly nonstationary white Gaussian noise has been added to a speech fragment consisting of the concatenation of sentences from the TIMIT-TEST database. The noise starts at 20-dB SNR (compared with the average speech power level) and stays at that level for 1 s. It then increases in 2 s to 0-dB SNR (at a rate of 0.16 dB/frame, i.e., 10 dB/s) and stays at that level for 2 s. The noise level then decreases in 2 s to 20-dB SNR, stays at that level for 1 s, and the entire pattern is repeated.

Fig. 3(a) shows the power level of the noisy speech and the estimated noise levels from our method (dashed line) and the MS method (dotted line). As expected, MS cannot track the rapid increases in noise level. The performance measures are

[1]The SNR is measured with respect to the average clean speech power level.

TABLE I
PERFORMANCE MEASURES FOR MS, THE SKS METHOD, AND THE PROPOSED METHOD FOR SPEECH CONTAMINATED BY NONSTATIONARY WHITE GAUSSIAN NOISE (FIG. 3).* MEANS THAT THE SAFETY NET HAS BEEN APPLIED

$LogErr$ [dB]					$SSNR+$ [dB]					$PESQ$				
MS	SKS	SKS*	Prop.	Prop.*	MS	SKS	SKS*	Prop.	Prop.*	MS	SKS	SKS*	Prop.	Prop.*
3.84	7.07	5.09	1.74	1.46	3.01	1.55	2.06	4.05	4.36	1.78	1.66	1.72	1.99	2.01

TABLE II
PERFORMANCE MEASURES FOR MS, THE SKS METHOD, AND THE PROPOSED METHOD FOR SPEECH CONTAMINATED BY TRAFFIC NOISE (FIG. 4). THE SAFETY NET HAS BEEN APPLIED TO THE SKS METHOD AND THE PROPOSED METHOD

$LogErr$ [dB]			$SSNR+$ [dB]			$PESQ$		
MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.
2.84	2.02	0.98	3.79	4.31	6.04	1.40	1.51	1.64

shown in Table I. Our method can handle both fast increases and decreases in noise level, resulting in much better performance figures. Panel (b) shows the ideal noise reference (solid line) of (20), and the estimated noise level from the SKS method with safety net (dashed line) and without it (dotted line). The safety net clearly improves the tracking of increasing noise levels, which is also reflected in the performance measures. Panel (c) and Table I show that our method also benefits slightly from the safety net for this example.

As a second, real-life, example, consider Fig. 4 where a clean speech fragment (a) has been contaminated (b) with the noise from passing traffic. The traffic noise has been taken from the ETSI EG 202 396-1 Background Noise Database [30]. Panel (c) shows the ideal noise level, and the estimated noise levels from MS, SKS, and the proposed method. The safety net has been applied to the SKS method and the proposed method. The corresponding enhanced signals are shown in panels (d)-(f). Clearly, the proposed method performs better than the other two methods. The objective performance measures are shown in Table II.

Such fast changes in noise characteristics may, however, not occur very often in other real-life situations. We will therefore now compare both methods also for various other noise sources. We will apply the safety net to the SKS method and the proposed method but not to the MS method, because the MS method principally does not need it.

2) Other Noise Sources: Tables III–V show the performance measures for the MS, the SKS, and the proposed method, for speech contaminated with stationary white Gaussian noise (WGN), and nonstationary airport, babble, street, and train noise. Our method clearly outperforms the MS and SKS methods in terms of LogErr and SSNR+. We achieve the largest improvements for the more nonstationary noise sources, as expected. We have tested on some other noise sources as well, obtaining similar improvements [31]. The SKS method and the proposed method score higher than MS in terms of the PESQ measure (Table V). The proposed method is slightly better than SKS for the nonstationary noise sources. The more nonstationary the noise source, the larger improvements we can expect, as we have already seen with the nonstationary white Gaussian noise in Fig. 3 and the traffic noise in Fig. 4.

In our last experiment, we will compare the performance of the gain function $G_{D^2,0}$, obtained with the original data-driven method [25], to the performance of $G_{D^2,i_{\max}}$, found iteratively (Section IV-D). Table VI shows LogErr and SSNR+ obtained with $G_{D^2,0}$. Comparing with Tables III and IV clearly shows that taking into account the recursive nature of the noise estimation by means of the iterative optimization is necessary. We also find large improvements for the nonstationary WGN of Fig. 3. For $G_{D^2,0}$ we have LogErr $= 4.06$ dB, SSNR+ $= 2.63$ dB, while we had LogErr $= 1.46$ dB and SSNR+ $= 4.36$ dB for the gain function found iteratively.

VI. SUMMARY AND CONCLUDING REMARKS

We have developed a fast noise tracking algorithm for speech contaminated with nonstationary additive noise. The method is based on recursive averaging of the mmse estimates of the noise power. The value of the smoothing parameter is controlled by an estimate of speech presence probability. The advantage of using an estimate of the noise power instead of the noisy power is that it strongly reduces the amount of speech power leaking into the noise variance estimates, in case of errors in the speech presence probability estimates. As a result, a simpler speech presence probability estimator can be used that allows the noise tracker to react almost instantaneously to changes in noise level.

The gain function for noise power estimation is found by an iterative data-driven method. The optimization procedure takes into account the recursive nature of the noise variance estimation, where the input parameters of the gain function in the current time frame have been computed in the previous frame on the basis of that gain function. Once the optimal gain function has been found, the noise tracker can be implemented with very low computational complexity; a result of storing the gain function in a lookup table.

In the development of the proposed method, we have aimed at a high tracking speed. In applications with noise that does not change very rapidly, we could sacrifice some of the tracking speed for a lower variance and increased accuracy. For example, we could lower the threshold $T(k,m)$ in (9) or increase the value of α_d in (5). A more advanced speech presence probability estimation method such as the one of IMCRA could further reduce speech leakage. Then the tradeoff between accuracy

TABLE III
LogERR [dB] FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING MS, THE SKS METHOD, AND THE PROPOSED METHOD

SNR	Stat. WGN			Airport			Babble			Street			Train		
[dB]	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.
0	1.02	1.00	0.69	3.22	2.73	2.19	3.55	2.88	2.34	3.72	3.18	2.17	3.14	2.43	1.93
5	1.15	1.08	0.79	3.13	2.70	2.25	3.35	2.74	2.30	3.68	3.15	2.22	3.15	2.38	1.94
10	1.28	1.20	0.92	3.07	2.73	2.40	3.13	2.65	2.35	3.57	3.16	2.34	3.04	2.35	1.98
15	1.47	1.41	1.15	3.15	2.84	2.66	3.04	2.66	2.52	3.49	3.23	2.58	2.92	2.39	2.12
20	1.78	1.75	1.52	3.58	3.07	3.03	3.29	2.80	2.83	3.61	3.40	2.96	2.98	2.56	2.41

TABLE IV
SSNR+ [dB] FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING MS, THE SKS METHOD, AND THE PROPOSED METHOD

SNR	Stat. WGN			Airport			Babble			Street			Train		
[dB]	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.
0	8.28	7.83	7.83	3.43	3.51	4.26	3.52	3.66	4.27	3.72	3.77	4.88	4.88	5.18	5.69
5	6.56	6.59	6.56	2.77	2.89	3.36	2.92	3.06	3.42	2.95	3.06	3.85	3.90	4.25	4.55
10	4.88	5.16	5.14	2.12	2.23	2.49	2.30	2.40	2.57	2.19	2.32	2.82	2.97	3.28	3.44
15	3.32	3.70	3.66	1.43	1.57	1.67	1.60	1.73	1.77	1.41	1.55	1.83	2.02	2.29	2.37
20	1.93	2.32	2.27	0.64	0.90	0.93	0.80	1.05	1.02	0.57	0.82	0.89	1.01	1.35	1.33

TABLE V
PESQ FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING MS, THE SKS METHOD, AND THE PROPOSED METHOD

SNR	Stat. WGN			Airport			Babble			Street			Train		
[dB]	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.	MS	SKS	Prop.
0	1.35	1.39	1.37	1.44	1.45	1.46	1.38	1.39	1.39	1.30	1.32	1.37	1.29	1.31	1.32
5	1.73	1.79	1.77	1.81	1.83	1.86	1.74	1.76	1.78	1.64	1.68	1.76	1.61	1.66	1.68
10	2.13	2.24	2.22	2.23	2.28	2.31	2.17	2.21	2.24	2.05	2.11	2.20	1.99	2.07	2.10
15	2.52	2.70	2.69	2.69	2.78	2.81	2.63	2.71	2.74	2.49	2.58	2.66	2.42	2.54	2.58
20	2.93	3.17	3.15	3.14	3.27	3.29	3.09	3.21	3.24	2.93	3.06	3.12	2.89	3.05	3.08

TABLE VI
LogErr [dB] AND SSNR+ [dB] FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING $G_{D^2,0}$ INSTEAD OF $G_{D^2,i_{\max}}$ FOR NOISE POWER ESTIMATION

SNR	Stat. WGN		Airport		Babble		Street		Train	
[dB]	$LogErr$	$SSNR+$	$LogErr$	$SSNR+$	$LogErr$	$SSNR+$	$LogErr$	$SSNR+$	$LogErr$	$SSNR+$
0	1.25	6.64	3.55	2.94	4.08	2.90	3.68	3.40	3.25	4.40
5	1.21	5.71	3.40	2.42	3.84	2.42	3.57	2.77	3.12	3.66
10	1.17	4.63	3.25	1.89	3.58	1.92	3.48	2.12	2.97	2.89
15	1.15	3.48	3.12	1.38	3.33	1.44	3.39	1.49	2.81	2.12
20	1.19	2.27	2.98	0.89	3.07	0.97	3.32	0.92	2.68	1.39

and tracking speed is also controlled by the length of the minimum-search window.

Our noise tracker can be easily integrated into existing speech enhancement schemes. It shows excellent noise tracking capabilities and noise reduction performance for a variety of stationary and nonstationary noise sources for a wide range of SNRs. The techniques proposed in this paper could be of interest for many applications where estimation of the noise spectrum is required, such as automatic speech recognition and speaker verification, speech coding, hearing aids, and restoration of old

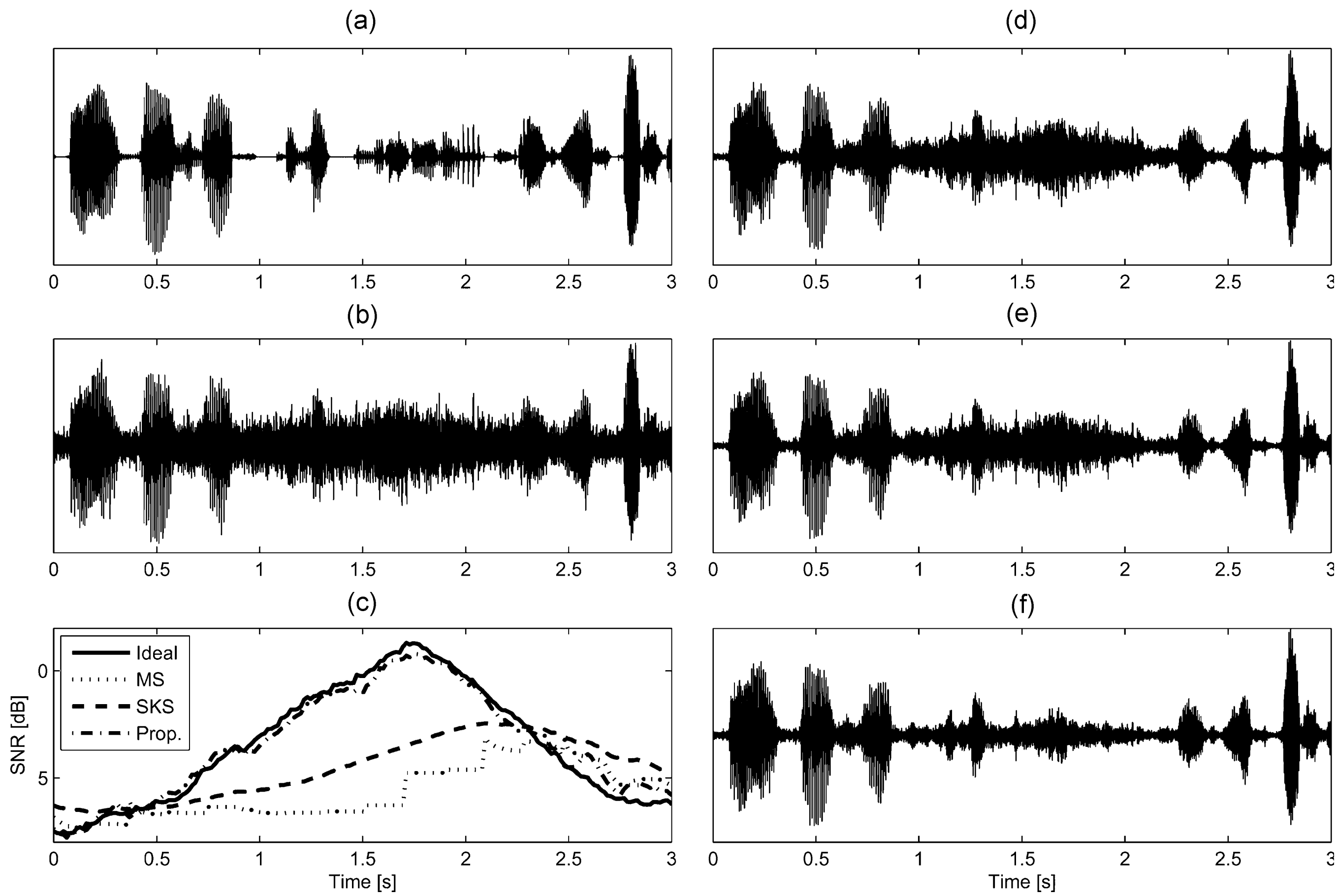


Fig. 4. Panel (a) shows a clean speech signal, and panel (b) the signal contaminated by traffic noise. In panel (c), the ideal reference noise level is shown (solid line), along with the estimated noise levels from the MS (dotted), the SKS (dashed), and the proposed method (dash-dotted). The enhanced signals are shown in the remaining panels: (d) MS method, (e) SKS method, (f) proposed method.

recordings, to name a few. It falls outside the scope of this paper to investigate to what extent the improvements shown carry over to these specialized areas of research, but such investigations are necessary, of scientific and practical importance, and could lead to further progress and understanding in these fields.

Acknowledgment

The authors would like to thank the anonymous reviewers whose comments greatly improved the clarity of the presentation.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[4] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[5] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 1110–1126, 2005.

[6] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[7] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[8] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[9] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, pp. 220–231, Feb. 2006.

[10] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. EUSIPCO*, 1994, pp. 1182–1185.

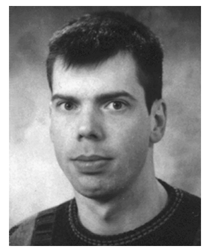
[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[12] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. Eurospeech*, 1995, pp. 1513–1516.

[13] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, pp. 1875–1878.

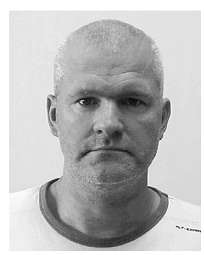
[14] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[15] A. Borowicz and A. Petrovsky, "Minima controlled noise estimation for KLT-based speech enhancement," in *Proc. Eur. Signal Process. Conf. EUSIPCO*, Florence, Italy, 2006, CD-ROM.
[16] S. Gannot, "Speech enhancement: Application of the Kalman filter in the estimate-maximize (EM framework)," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. New York: Springer, 2005, pp. 161–198.
[17] I. Batina, J. Jensen, and R. Heusdens, "Noise power spectrum estimation for speech enhancement using an autoregressive model for speech power spectrum dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, vol. III, pp. 1064–1067.
[18] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.
[19] *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. New York: Springer, 2005.
[20] C. H. You, S. N. Koh, and S. Rahardja, " β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.
[21] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Speech Commun.*, vol. 81, pp. 2403–2418, 2001.
[22] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
[23] I. Cohen, "Relaxed statistical model for speech enhancement and *a priori* SNR estimation," *Speech Commun.*, vol. 86, pp. 698–709, Apr. 2006.
[24] P. Ishwar and P. Moulin, "On the equivalence of set-theoretic and maxent MAP estimation," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 698–713, Mar. 2003.
[25] J. S. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Commun., Special Iss. Speech Enhancement*, vol. 49, pp. 530–541, Jul.–Aug. 2007.
[26] "Timit, Acoustic-Phonetic Continuous Speech Corpus," DARPA, NIST Speech Disc 1-1.1, Oct. 1990.
[27] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, vol. I, pp. 153–156.
[28] J. S. Erkelens, J. Jensen, and R. Heusdens, "Improved speech spectral variance estimation under the generalized Gamma distribution," in *IEEE BENELUX/DSP Valley Signal Process. Symp.*, Antwerp, Belgium, 2007, pp. 43–46.
[29] J. G. Beerends, "Extending p.862 PESQ for assessing speech intelligibility," White Contribution COM 12-C2 to ITU-T Study Group 12, Oct. 2004.
[30] "ETSI EG 202 396-1: Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database," ETSI [Online]. Available: http://portal.etsi.org/docbox/STQ/Open/EG 202 396-1 Background noise database/
[31] J. S. Erkelens and R. Heusdens, "Fast noise tracking based on recursive smoothing of MMSE noise power estimates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, 2008, pp. 4873–4876.



Jan S. Erkelens received the Ph.D. degree from the Applied Physics Department, Delft University of Technology, Delft, The Netherlands, in 1996. The subject of his thesis was low bit-rate speech coding. He also has experience in the atmospheric sciences and speech recognition.

He currently holds a Postdoctoral position in the Information and Communication Theory Group, Department of Mediamatics, Delft University of Technology, where he is working on speech enhancement.



Richard Heusdens received the M.Sc. and Ph.D. degrees from the Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively.

Since 2002, he has been an Associate Professor in the Department of Mediamatics, Delft University of Technology. In the spring of 1992, he joined the Digital Signal Processing Group, Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group, Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio and speech processing activities within the ICT group. He is involved in research projects that cover subjects such as audio and speech coding, speech enhancement, and digital watermarking of audio.