EURASIP Journal on Applied Signal Processing 2005:18, 2954–2964

Speech Enhancement with Natural Sounding Residual Noise Based on Connected Time-Frequency Speech Presence Regions

Karsten Vandborg Sørensen
Department of Communication Technology, Aalborg University, DK-9220 Aalborg East, Denmark
Email: kvs@kom.aau.dk

Søren Vang Andersen
Department of Communication Technology, Aalborg University, DK-9220 Aalborg East, Denmark
Email: sva@kom.aau.dk

Received 13 May 2004; Revised 3 March 2005

We propose time-frequency domain methods for noise estimation and speech enhancement. A speech presence detection method is used to find connected time-frequency regions of speech presence. These regions are used by a noise estimation method and both the speech presence decisions and the noise estimate are used in the speech enhancement method. Different attenuation rules are applied to regions with and without speech presence to achieve enhanced speech with natural sounding attenuated background noise. The proposed speech enhancement method has a computational complexity, which makes it feasible for application in hearing aids. An informal listening test shows that the proposed speech enhancement method has significantly higher mean opinion scores than minimum mean-square error log-spectral amplitude (MMSE-LSA) and decision-directed MMSE-LSA.

Keywords and phrases: speech enhancement, noise estimation, minimum statistics, speech presence detection.

1. INTRODUCTION

The performance of many speech enhancement methods relies mainly on the quality of a noise power spectral density (PSD) estimate. When the noise estimate differs from the true noise, it will lead to artifacts in the enhanced speech. The approach taken in this paper is based on connected region speech presence detection. Our aim is to exploit spectral and temporal masking mechanisms in the human auditory system [1] to reduce the perception of these artifacts in speech presence regions and eliminate the artifacts in speech absence regions. We achieve this by leaving downscaled natural sounding background noise in the enhanced speech in connected time-frequency regions with speech absence. The downscaled natural sounding background noise will spectrally and temporally mask artifacts in the speech estimate while preserving the naturalness of the background noise.

In the definition of speech presence regions, we are inspired by the work of Yang [2]. Yang demonstrates high perceptual quality of a speech enhancement method where constant gain is applied in frames with no detected speech presence. Yang lets a single decision cover a full frame. Thus, musical noise is present in the full spectrum of the enhanced speech in frames with speech activity. We therefore extend the notion of speech presence to individual time-frequency locations. This, in our experience, significantly improves the naturalness of the residual noise. The speech enhancement method, proposed in this paper, thereby eliminates audible musical noise in the enhanced speech. However, fluctuating speech presence decisions will reduce the naturalness of the enhanced speech and the background noise. Thus, reasonably connected regions of the same speech presence decision must be established.

To achieve this, we use spectral-temporal periodogram smoothing. To this end, we make use of the spectral-temporal smoothing method by Martin and Lotter [3], which extends the original groundbreaking work of Martin [4, 5]. Martin and Lotter derive optimum smoothing coefficients for (generalized) χ^2-distributed spectrally smoothed spectrograms, which is particularly well suited for noise types with a smooth power spectrum. The underlying assumption in this approach is that the real and imaginary parts of the associated STFT coefficients for the averaged periodograms have the same means and variances. For the application of

spectral-temporal smoothing to obtain connected regions of speech presence decisions, we augment Martin and Lotters smoothing method with the spectral smoothing method used by Cohen and Berdugo [6].

For minimum statistics noise estimation, Martin [5] has suggested a theoretically founded bias compensation factor, which is a function of the minimum search window length, the smoothed noisy speech, and the noise PSD estimate variances. This enables a low-biased noise estimate that does not rely on a speech presence detector. However, as our proposed speech enhancement method has connected speech presence regions as an integrated component, this enables us to make use of a new, simple, yet efficient, bias compensation. To verify the performance of the new bias compensation, we objectively evaluate the noise estimation method that uses this bias compensation of minimum tracks from our spectrally temporally smoothed periodograms, prior to integrating this noise estimate in the final speech enhancement method.

In result, our proposed speech enhancement algorithm has a low computational complexity, which makes it particularly relevant for application in digital signal processors with limited computational power, such as those found in digital hearing aids. In particular, the obtained algorithm provides a significantly higher perceived quality than our implementation of the decision-directed minimum mean-square error log-spectral amplitude (MMSE-LSA-DD) estimator [7] when evaluated in listening tests. Furthermore, the noise PSD estimate that we use to obtain a noise magnitude spectrum estimate for the attenuation rule in connected regions of speech presence is shown to be superior to estimates from minimum statistics (MS) noise estimation [5] and our implementation of χ^2-based noise estimation [3] for spectrally smooth noise types.

The rest of this paper is organized as follows. In Section 2, we describe the signal model and give an overview of the proposed algorithm. In Section 3, we list the necessary equations to perform the spectral-temporal periodogram smoothing. Section 4 contains a description of our detector for connected speech presence regions, and in Section 5, we describe how the spectrally temporally smoothed periodograms and the speech presence regions can be used to obtain both a noise PSD estimate and a noise periodogram estimate, which both rely on the new bias compensation. In the latter noise estimation method, we estimate the squared magnitudes of the noise short-time Fourier transform (STFT) coefficients. In Section 6, the connected region speech presence detector is introduced in a speech enhancement method with the purposes of reducing noise and augmenting listening comfort. Section 7 contains the experimental setup and all necessary initializations. Finally, Section 8 describes the experimental results and Section 9 concludes the paper with a discussion of the proposed methods and obtained results.

2. STRUCTURE OF THE ALGORITHM

After an introduction to the signal model, we give a structural description of the algorithm to provide an algorithmic overview before the individual methods, which constitute the algorithm, are described in detail.

2.1. Signal model

We assume that noisy speech $y(i)$ at sampling time index i consists of speech $s(i)$ and additive noise $n(i)$. For joint time-frequency analysis of $y(i)$, we apply the K-point STFT, that is,

$$Y(\lambda,k)=\sum_{\mu=0}^{L-1} y(\lambda R+\mu)h(\mu)\exp\left(-\frac{j2\pi k\mu}{K}\right), \tag{1}$$

where $\lambda \in \mathcal{Z}$ is the (subsampled) time index, $k \in \{0,1,\ldots,K-1\}$ is the frequency index, and L is the window length. In this paper, we have that L equals K. The quantity R is the number of samples that successive frames are shifted and $h(\mu)$ is a unit-energy window function, that is, $\sum_{\mu=0}^{L-1} h^2(\mu)=1$. From the linearity of (1), we have that

$$Y(\lambda,k)=S(\lambda,k)+N(\lambda,k), \tag{2}$$

where $S(\lambda,k)$ and $N(\lambda,k)$ are the STFT coefficients of speech $s(i)$ and additive noise $n(i)$, respectively. We further assume that $s(i)$ and $n(i)$ are zero mean and statistically independent, which leads to a power relation, where the noise is additive [8], that is,

$$E\left\{|Y(\lambda,k)|^2\right\}=E\left\{|S(\lambda,k)|^2\right\}+E\left\{|N(\lambda,k)|^2\right\}. \tag{3}$$

2.2. Structural algorithm description

The structure of the proposed algorithm and names of variables with a central role are shown in Figure 1. After applying an analysis window to the noisy speech, we take the STFT, from which we calculate periodograms $P_Y(\lambda,k) \triangleq |Y(\lambda,k)|^2$. These periodograms are spectrally smoothed, yielding $\mathcal{P}_Y(\lambda,k)$, and then temporally smoothed to produce $\mathcal{P}(\lambda,k)$. These smoothed periodograms are temporally minimum tracked, and by comparing ratios and differences of the minimum tracked values to $\mathcal{P}(\lambda,k)$, they are used for speech presence detection. As a distinct feature of the proposed method, we use speech presence detection to achieve low-biased noise PSD estimates $\widetilde{P}_{\hat{N}}(\lambda,k)$, but also for noise periodogram estimates $P_{\hat{N}}(\lambda,k)$, which equal $P_Y(\lambda,k)$ when $D(\lambda,k)=0$, that is, no detected speech presence. When $D(\lambda,k)=1$, that is, detected speech presence, the noise periodogram estimate equals the noise PSD estimate, that is, a recursively smoothed bias compensation factor applied on the minimum tracked values. The bias compensation factor is recursively smoothed power ratios between the noise periodogram estimates and the minimum tracks. This factor is only updated while no speech is present in the frames and kept fixed while speech is present. A noise magnitude spectrum estimate $|\hat{N}(\lambda,k)|$ obtained from the noise PSD

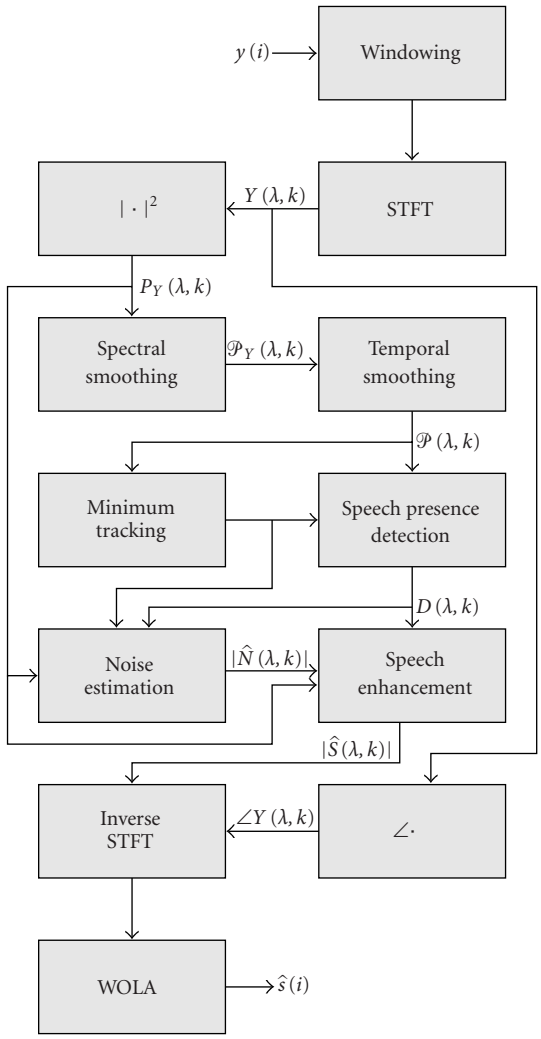

Figure 1: A block diagram of the proposed speech enhancement algorithm. Only the most essential variables are introduced in the figure.

estimate and the speech presence decisions are used in a speech enhancement method that applies different attenuation rules for speech presence and speech absence. For speech synthesis, we take the inverse STFT of the estimated speech magnitude spectrum with the phase from the STFT of the noisy speech. The synthesized frame is used in a weighted overlap-add (WOLA) method, where we apply a synthesis window before overlap and add.

3. SPECTRAL-TEMPORAL PERIODOGRAM SMOOTHING

In this section, we briefly describe the spectral-temporal periodogram smoothing method.

3.1. Spectral smoothing

First, the noisy speech periodograms $P_Y(\lambda,k)$ are spectrally smoothed by letting a spectrally smoothed periodogram bin $\mathcal{P}_Y(\lambda,k)$ consist of a weighted sum of $2D+1$ periodogram bins, spectrally centered at k [6], that is,

$$\mathcal{P}_Y(\lambda,k) = \sum_{\nu=-D}^{D} b(\nu) P_Y\left(\lambda, ((k-\nu))_K\right), \tag{4}$$

where $((m))_K$ denotes m modulus K, and K is the length of the full (mirrored) spectrum. The window function $b(\nu)$ used for spectral weighting is chosen such that it sums to 1, that is, $\sum_{\nu=-D}^{D} b(\nu) = 1$, and therefore preserves the total power of the spectrum.

3.2. Temporal smoothing

The spectrally smoothed periodograms $\mathcal{P}_Y(\lambda,k)$, see Figure 1, are now temporally smoothed recursively with time and frequency varying smoothing parameters $\alpha(\lambda,k)$ to produce a spectrally temporally smoothed noisy speech periodogram $\mathcal{P}(\lambda,k)$, that is,

$$\mathcal{P}(\lambda,k) = \alpha(\lambda,k)\mathcal{P}(\lambda-1,k) + (1-\alpha(\lambda,k))\mathcal{P}_Y(\lambda,k). \tag{5}$$

We use the optimum smoothing parameters proposed by Martin and Lotter [3]. Their method consists of optimum smoothing parameters for χ^2-distributed data with some modifications that makes it suited for practical implementation. The optimum smoothing parameters are given by

$$\alpha^\star(\lambda,k) = \frac{2}{2+\widetilde{K}\left(\mathcal{P}(\lambda-1,k)/E\left\{|N(\lambda,k)|^2\right\}-1\right)^2}, \tag{6}$$

with

$$\widetilde{K} = (4D+2)\frac{\left[\sum_{\mu=0}^{L-1} b^2(\mu)\right]^2}{L\sum_{\mu=0}^{L-1} b^4(\mu)} \tag{7}$$

"equivalent" degrees of freedom of a χ^2-distribution [3]. For practical implementation, the noise PSD, which is used in the calculation of the optimum smoothing parameters, is estimated as the previous noise PSD estimate, that is,

$$E\left\{\widehat{|N(\lambda,k)|^2}\right\} = \widetilde{P}_{\hat{N}}(\lambda-1,k). \tag{8}$$

3.3. Complete periodogram smoothing algorithm

Pseudocode for the complete spectral-temporal periodogram smoothing method is provided in Algorithm 1. A smoothing parameter correction factor $\alpha_c(\lambda,k)$, proposed by Martin [5], is multiplied on $\tilde{\alpha}(\lambda,k)$. Additionally, in this paper, we lower-limit the resulting smoothing parameters to ensure a minimum degree of smoothing, that is,

$$\alpha(\lambda,k) = \max\left(\alpha_c(\lambda,k)\tilde{\alpha}(\lambda,k), 0.4\right). \tag{9}$$

(1) {Initialize as listed in Tables 3 and 1}
(2) *for* $\lambda = 0$ to $M-1$ *do*
(3) *for* $k = 0$ to $K-1$ *do*
(4) $P_Y(\lambda,k) \leftarrow \text{abs}(\sum_{\mu=0}^{L-1} y(\lambda R+\mu)h(\mu) \exp(-j2\pi k\mu/K))^2$
(5) $\mathcal{P}_Y(\lambda,k) \leftarrow \sum_{\nu=-D}^{D} b(\nu)P_Y(\lambda, \mod(k-\nu,K))$
(6) $\alpha^\star(\lambda,k) \leftarrow 2/(2+\widetilde{K}(\mathcal{P}(\lambda-1,k)/\widetilde{P}_{\hat{N}}(\lambda-1,k)-1)^2)$
(7) *end for*
(8) $R \leftarrow (\sum_{k=0}^{K-1} \mathcal{P}(\lambda-1,k))/(\sum_{k=0}^{K-1} P_Y(\lambda,k))$
(9) $\tilde{\alpha}_c \leftarrow 1/(1+(R-1)^2)$
(10) $\alpha_c(\lambda) \leftarrow 0.7\alpha_c(\lambda-1) + 0.3\max(\tilde{\alpha}_c, 0.7)$
(11) *for* $k = 0$ to $K-1$ *do*
(12) $\alpha(\lambda,k) \leftarrow \max(\alpha_c(\lambda)\tilde{\alpha}(\lambda,k), 0.4)$
(13) $\mathcal{P}(\lambda,k) \leftarrow \alpha(\lambda,k)\mathcal{P}(\lambda-1,k) + (1-\alpha(\lambda,k))\mathcal{P}_Y(k)$
(14) {Obtain a noise PSD estimate $\widetilde{P}_{\hat{N}}(\lambda,k)$, e.g., as proposed in Section 5.}
(15) *end for*
(16) *end for*

Algorithm 1: Periodogram smoothing.

In the next section, we use temporal minimum tracking on the spectrally temporally smoothed noisy speech periodograms in a method for detection of connected speech presence regions, which later will be used for noise estimation and speech enhancement.

4. CONNECTED SPEECH PRESENCE REGIONS

We now base a speech presence detection method on comparisons, at each frequency, between the smoothed noisy speech periodograms and temporal minimum tracks of the smoothed noisy speech periodograms.

4.1. Temporal minimum tracking

From the spectrally temporally smoothed noisy speech periodograms $\mathcal{P}(\lambda,k)$, we track temporal minimum values $\mathcal{P}_{\min}(\lambda,k)$, within a minimum search window of length $D_{\min}$, that is,

$$\mathcal{P}_{\min}(\lambda,k) = \min\left(\mathcal{P}(\psi,k) \mid \lambda - D_{\min} < \psi \le \lambda\right), \quad (10)$$

with $\psi \in \mathcal{Z}$. $D_{\min}$ is chosen as a tradeoff between the ability to bridge over periods of speech presence [5], which is crucial for the minimum track to be robust to speech presence, and the ability to follow nonstationary noise. Typically, a window length corresponding to 0.5–1.5 seconds yields an acceptable tradeoff between these two properties [5, 6]. We now have that $\mathcal{P}_{\min}(\lambda,k)$ is approximately unaffected by periods of speech presence, but on the average, biased towards lower values, when no spectral smoothing is applied [5]. Memory requirements of the tracking method can be reduced at the cost of lost temporal resolution, see, for example, [5]. In the following, the temporal minimum tracks $\mathcal{P}_{\min}(\lambda,k)$ are used in a speech presence decision rule.

For the spectral-temporal periodogram smoothing, we use the settings and algorithm initializations given in Table 1. The decision rules, that are used for speech presence detection, have the threshold values listed in Table 2. For noise estimation, we use the two parameters from Table 4. The speech enhancement method uses the parameter settings that are listed in Table 5.

4.2. Binary speech presence decision rule

We have shown in previous work [9] that temporally smoothed periodograms and their temporal minimum tracks can be used for speech presence detection. Also shown in [9] is that including terms to compensate for bias on the minimum tracks improves the speech presence detection performance (measured as the decrease in a cost function) by less than one percent. In this paper, we therefore do not consider a bias compensation factor in the speech presence decision rule. Rather, as we show later in this paper, the speech presence decisions can be used in the estimation of a simple and very well-performing bias compensation factor for noise estimation. Similar to our previous approach for temporally smoothed periodograms [9], we now exploit the properties of spectrally temporally smoothed periodograms $\mathcal{P}(\lambda,k)$ in a binary decision rule for the detection of speech presence. The presence of speech will cause an increase of power in $\mathcal{P}(\lambda,k)$ at a particular time-frequency location, due to (3). Thus, the ratio between $\mathcal{P}(\lambda,k)$ and a noise PSD estimate, given by a minimum track $\mathcal{P}_{\min}(\lambda,k)$ with a bias reduction, yields a robust (due to the smoothing) estimate of the signal-plus-noise-to-noise ratio at the particular time-frequency location. Our connected region speech presence detection method is based on the smooth nature of $\mathcal{P}(\lambda,k)$ and $\mathcal{P}_{\min}(\lambda,k)$. The smoothness will ensure that spurious fluctuations in the noisy speech power will not cause spurious fluctuations in our speech presence decisions. Thus, we will be able to obtain connected regions of speech presence and of speech absence. This property is fundamental for the proposed noise estimation and speech enhancement methods. As a rule to decide between the two speech presence hypotheses, namely,

$$\begin{aligned} H_0(\lambda,k) &: \text{“speech absence,”} \\ H_1(\lambda,k) &: \text{“speech presence,”} \end{aligned} \quad (11)$$

which can be written in terms of the STFT coefficients, that is,

$$\begin{aligned} H_0(\lambda,k) &: Y(\lambda,k) = N(\lambda,k), \\ H_1(\lambda,k) &: Y(\lambda,k) = N(\lambda,k) + S(\lambda,k), \end{aligned} \quad (12)$$

we use a combination of two binary initial decision rules. First, let $D(\lambda,k) = i$ be the decision to believe in hypothesis $H_i(\lambda,k)$ for $i \in \{0,1\}$. We define two initial decision rules, which will give two initial decisions $D'(\lambda,k)$ and $D''(\lambda,k)$. The initial decision rules are given by a rule, where the smoothed noisy speech periodograms $\mathcal{P}(\lambda,k)$ are compared with the temporal minimum tracks $\mathcal{P}_{\min}(\lambda,k)$,

Table 1: Smoothing setup and initializations for Algorithm 1.

Variable	Value	Description
D	7	Spectral window length: $2D+1$
$b(\nu)$	$G_b\text{TRIANG}(2D+1)$[i]	Spectral smoothing window
M	154–220[ii]	Number of frames
$\widetilde{K}$	20.08[iii]	"Equivalent" degrees of freedom of χ^2-distribution
$P_{\hat{N}}(-1,k)$	$P_Y(0,k)$	Initial noise periodogram estimate
$\alpha_c(-1)$	1	Initial correction variable

[i] $G_b = (\text{sum}(\text{TRIANG}(2D+1)))^{-1}$ scales the window to unit sum.
[ii] Calculated at run time as $M = \text{round}\,(\text{length}\,(y(i))/R - 1/2) - 1$.
[iii] Calculated at run time as $\widetilde{K} = (4D+2)([\sum_{\mu=0}^{L-1} b^2(\mu)]^2)/(L\sum_{\mu=0}^{L-1} b^4(\mu))$ [3, 14].

Table 2: Speech presence detection setup.

Variable	Value	Description
γ'	6	Constant for ratio-based decision rule
γ''	0.5	Constant for difference-based decision rule

weighted with a constant γ', that is,

$$D'(\lambda,k) : \mathcal{P}(\lambda,k) \underset{D'(\lambda,k)=0}{\overset{D'(\lambda,k)=1}{\gtrless}} \gamma' \mathcal{P}_{\min}(\lambda,k), \tag{13}$$

and one where, at time λ, the difference is compared to the average of the minimum tracks scaled by γ'', that is,

$$D''(\lambda,k) : \mathcal{P}(\lambda,k) \underset{D''(\lambda,k)=0}{\overset{D''(\lambda,k)=1}{\gtrless}} \mathcal{P}_{\min}(\lambda,k) + \gamma'' \frac{1}{K}\sum_{k=0}^{K-1} \mathcal{P}_{\min}(\lambda,k). \tag{14}$$

For the initial decision rules, we have adopted the notation, used by Shanmugan and Breipohl [8]. Because the minimum tracks are representatives of the noise PSDs [5], the first initial decision rule classifies time-frequency bins based on the estimated signal-plus-noise-to-noise power ratio. Note that this can be seen as a special case of the indicator function proposed by Cohen [10] (with $\zeta_0 = \gamma'/B_{\min}$ and $\gamma_0 = \infty$). The second initial decision rule $D''(\lambda,k)$ classifies bins from the estimated power difference between the noisy speech and the noise using a threshold that adapts to the minimum track power level in each frame. Multiplication of the two binary initial decisions corresponds to the logical AND-operation, when we define true as deciding on $H_1(\lambda,k)$ and false as deciding on $H_0(\lambda,k)$. We therefore propose a decision that combines the two initial decisions from the initial decision rules above, that is,

$$D(\lambda,k) = D'(\lambda,k)\cdot D''(\lambda,k). \tag{15}$$

In effect, the combined decision allows detection of speech in low signal-to-noise ratios (SNRs) without letting low-power regions with high SNRs contaminate the decisions. Thereby, we obtain connected time-frequency regions of speech presence. The constant γ' is not sensitive to the type and intensity of environmental noise [11] and it can be adjusted empirically. This is also the case for γ''. For applications where a reasonable objective performance measure can be defined, the constants γ' and γ'' can be obtained by interpreting the decision rule as artificial neural network and then conduct a supervised training of this network [9].

Speech at frequencies below 100 Hz is considered perceptually unimportant, and bins below this frequency are therefore always classified with speech absence. Real-life noise sources often have a large part of their power at the low frequencies, so this rule ensures that this power does not cause the speech presence detection method to falsely classify these low-frequency bins as if speech is present. If less than 5% of the K periodogram bins are classified with speech presence, we expect that these decisions have been falsely caused by the noise characteristics, and all decisions in the current frame are reclassified to speech absence. When the speech presence decisions are used in a speech enhancement method, as we propose in Section 6, this reclassification will ensure the naturalness of the background noise in periods of speaker silence.

5. NOISE ESTIMATION

The spectral-temporal smoothing method [3], which we use in this paper, reduces the bias between the noise PSD and the minimum track $\mathcal{P}_{\min}(\lambda,k)$ if the noise is assumed to be ergodic in its PSD. That is, it reduces the bias compared to minimum tracked values from periodograms, smoothed temporally using Martin's first method [5]. Martin gives a parametric description of a bias compensation factor, which depends on the minimum search window length, the smoothed noisy speech, and the noise PSD estimate variances. The spectral smoothing lowers the smoothed noisy speech periodogram variance, and as a consequence, a longer minimum search window can be applied when the noise spectrum is not changing rapidly. This give the ability to bridge over longer speech periods.

We propose to use the speech presence detection method from Section 4 to obtain two different noise estimates, that is, a noise PSD estimate and a noise periodogram estimate. The PSD estimate will be used in the speech enhancement methods and the noise periodogram estimate will illustrate

Table 3: General setup.

Variable	Value	Description
F_s	8 kHz	Sample frequency
K	256	FFT size
L	256	Frame size
R	128	Frame skip
$h(\mu)$	$G_h^{-1}\sqrt{\text{Hanning}(K)}$[i]	Analysis window
$h_s(\mu)$	$G_h\sqrt{\text{Hanning}(K)}$[ii]	Synthesis window

[i] G_h is the square root of the energy of $\sqrt{\text{Hanning}(K)}$, which scales the analysis window to unit energy. This is to avoid scaling factors throughout the paper.
[ii] G_h scales the synthesis window $h_s(\mu)$ such that the analysis window $h(\mu)$, multiplied with $h_s(\mu)$, yields a Hanning(K) window.

some of the properties of the residual noise from the speech enhancement method we propose in Section 6.

5.1. Noise periodogram estimation

The noise periodogram estimate is equal to a time-varying power scaling of the minimum tracks $\mathcal{P}_{\min}(\lambda,k)$, for $D(\lambda,k)=1$. For $D(\lambda,k)=0$, it is equal to the noisy speech periodogram $P_Y(\lambda,k)$, that is,

$$P_{\widehat{N}}(\lambda,k)=\begin{cases}R_{\min}(\lambda)\mathcal{P}_{\min}(\lambda,k) & \text{if } D(\lambda,k)=1,\\ P_Y(\lambda,k) & \text{if } D(\lambda,k)=0.\end{cases} \tag{16}$$

In the above equation, a bias compensation factor $R_{\min}(\lambda)$ scales the minimum. The scaling factor is updated in frames where no speech presence is detected and kept fixed while speech presence is detected in the frames. We let $\widetilde{R}_{\min}(\lambda)$ be given by the ratio between the sums of the previous noise periodogram estimate $P_{\widehat{N}}(\lambda-1,k)$ and the minimum tracks $\mathcal{P}_{\min}(\lambda,k)$, that is,

$$\widetilde{R}_{\min}(\lambda)=\frac{\sum_{k=0}^{K-1}P_{\widehat{N}}(\lambda-1,k)}{\sum_{k=0}^{K-1}\mathcal{P}_{\min}(\lambda,k)}, \tag{17}$$

which is recursively smoothed when speech is absent in the frame, and fixed when speech is present in the frame, that is,

$$R_{\min}(\lambda)=\begin{cases}R_{\min}(\lambda-1) & \text{if } \sum_{k=0}^{K-1}D(\lambda,k)>0,\\ \alpha_{\min}R_{\min}(\lambda-1)+(1-\alpha_{\min})\widetilde{R}_{\min}(\lambda) & \text{if } \sum_{k=0}^{K-1}D(\lambda,k)=0,\end{cases} \tag{18}$$

where $0\le\alpha_{\min}\le 1$ is a constant recursive smoothing parameter. The magnitude spectrum, at time index λ, is obtained by taking the square root of noise periodogram estimate, that is,

$$|\widehat{N}(\lambda,k)|=\sqrt{P_{\widehat{N}}(\lambda,k)}. \tag{19}$$

This noise periodogram estimate equals the true noise periodogram $|N(\lambda,k)|^2$ when the speech presence detection is correctly detecting no-speech presence. When entering a region with speech presence, the noise periodogram estimate will take on the smooth shape of the minimum track, scaled with the bias compensation factor in (18) such that the power develops smoothly into the speech presence region.

5.2. Noise PSD estimation

The noise PSD estimate $\widetilde{P}_{\widehat{N}}(\lambda,k)$ is obtained exactly as the noise periodogram estimate but with (16) modified such that the noise PSD estimate is obtained directly as the power-scaled minimum tracks, that is,

$$\widetilde{P}_{\widehat{N}}(\lambda,k)=R_{\min}(\lambda)\mathcal{P}_{\min}(\lambda,k). \tag{20}$$

A smooth estimate of the noise magnitude spectrum can be obtained by taking the square root of the noise PSD estimates, that is,

$$|\widehat{N}(\lambda,k)|=\sqrt{\widetilde{P}_{\widehat{N}}(\lambda,k)}. \tag{21}$$

6. SPEECH ENHANCEMENT

We now describe the speech enhancement method for which the speech presence detection method has been developed. It is well known that methods that subtract a noise PSD estimate from a noisy speech periodogram, for example, using an attenuation rule, will introduce musical noise. This happens whenever the noisy speech periodogram exceeds the noise PSD estimate. If, on the other hand, the noise PSD estimate is too high, the attenuation will reduce more noise, but also will cause the speech estimate to be distorted. To mitigate these effects, we propose to distinguish between connected regions with speech presence and speech absence. In speech presence, we will use a traditional estimation technique, by means of generalized spectral subtraction, with the noise magnitude spectrum estimate, obtained using (21) from the noise PSD estimate. In speech absence, we will use a simple noise-scaling attenuation rule to preserve the naturalness in the residual noise. Note that this approach, but

Table 4: Noise estimation setup.

Variable	Value	Description
$D_{\min}$	150[i]	Minimum tracking window length
$\alpha_{\min}$	0.7	Scaling factor smoothing parameter

[i]Corresponds to a time duration of $D_{\min} \cdot R/F_s = 2.4$ seconds.

Table 5: Speech enhancement setup.

Variable	Value	Description
β_0	0.1	Noise scaling factor for no-speech presence
β_1	1.4	Noise overestimation factor for speech presence
a_1	0.8	Attenuation rule order for speech presence

with only a single speech presence decision covering all frequencies in each frame, has previously been proposed by Yang [2]. Moreover, Cohen and Berdugo [11] propose a binary detection of speech presence/absence (called the indicator function in their paper), which is similar to the one we propose in this paper. However, their decision includes noisy speech periodogram bins without smoothing, hence some decisions will not be regionally connected. In our experience, this leads to artifacts if the decisions are used directly in a speech enhancement scheme with two different attenuation rules for speech absence and speech presence. Cohen and Berdugo smooth their binary decisions to obtain estimated speech presence probabilities, which are used for a soft decision between two separate attenuation functions. Our approach, as opposed to this, is to obtain adequately time-frequency smoothed spectra from which connected speech presence regions can be obtained directly in a robust manner. As a consequence, we avoid distortion in speech absence regions, and thereby obtain a natural sounding background noise.

Let the generalized spectral subtraction variant be given similar to the one proposed by Berouti et al. [12], but with the decision of which attenuation rule to use given explicitly by the proposed speech presence decisions, instead of comparisons between the estimated speech power and an estimated noise floor. The immediate advantage of our approach is a higher degree of control with the properties of the enhancement algorithm. Our proposed method is given by

$$|\hat{S}(\lambda,k)| = \begin{cases} \left(|Y(\lambda,k)|^{a_1} - \beta_1 |\hat{N}(\lambda,k)|^{a_1}\right)^{1/a_1} & \text{if } D(\lambda,k) = 1, \\ \beta_0 |Y(\lambda,k)| & \text{if } D(\lambda,k) = 0, \end{cases} \tag{22}$$

where a_1 determines the power in which the subtraction is performed, β_1 is a noise overestimation factor that scales the estimated magnitude of the noise STFT coefficient $|\hat{N}(\lambda,k)|$, obtained from the noise PSD estimate by (21) in Section 5, raised to the a_1'th power. The factor β_0 scales the noisy speech STFT coefficient magnitude, which before this scaling equals the square root of the noise periodogram estimate for bins with $D(\lambda,k) = 0$. After the scaling, these noisy speech STFT magnitudes lead to the noise component that will be left, after STFT synthesis, in the speech estimate as artifact masking [1] and natural sounding attenuated background noise.

For synthesis, we let the STFT spectrum of the estimated speech be given by the magnitude, obtained from (22), and the noisy phase $\angle Y(\lambda,k)$, that is,

$$\hat{S}(\lambda,k) = |\hat{S}(\lambda,k)| e^{j\angle Y(\lambda,k)}. \tag{23}$$

By applying the inverse STFT, we synthesize a time-domain frame, which we use in a WOLA scheme, as illustrated in Figure 1, to form the synthesized signal. Depending on the analysis window, a corresponding synthesis window $h_s(\mu)$ is applied before overlap add is performed.

7. EXPERIMENTAL SETUP

In the experiments, we use 6 speech recordings from the TIMIT database [13]. The speech is spoken by 3 different male and 3 different female speakers—all uttering different sentences of 2-3 seconds duration. These sentences are added with zero-mean highway noise and car interior noise in 0, 5, and 10 dB overall signal-to-noise ratios to form a test set of 36 noisy speech sequences. Spectrograms of time-domain signals are shown with time-frequency axes and always with the time-domain signals. When we plot intermediate coefficients, the figures are shown with axes of subsampled time index λ and frequency index k. For all illustrations in this paper, we use the noisy speech from one of the male speakers with additive highway noise in a 5 dB over all SNR. The spectrograms and time-domain signals of this particular case of noisy speech and the corresponding noise are shown in Figures 2a and 2b, respectively. The general setup in the experiments is listed in Table 3. The analysis window $h(\mu)$ is the square root of a Hanning window, scaled to unit energy. As the synthesis window $h_s(\mu)$, we also use the square root of a Hanning window, but scaled, such that an unmodified frame would be windowed by a Hanning window after both the analysis and synthesis window have been applied. It will therefore be ready for overlap add with 50% overlapping frames.

8. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed algorithm. We measure the performance of the algorithm by means of visual inspection of spectrograms, spectral distortion measures, and informal listening tests. To illustrate the properties of the proposed spectral-temporal smoothing method, we show the spectrogram of the smoothed noisy speech in Figure 3. By removing the power in speech absence regions and speech presence regions from the noisy speech periodogram, we see in Figures 4a and 4b, respectively, that most of the speech, that is detectable by visual inspection, has been detected by the proposed algorithm. Spectrograms

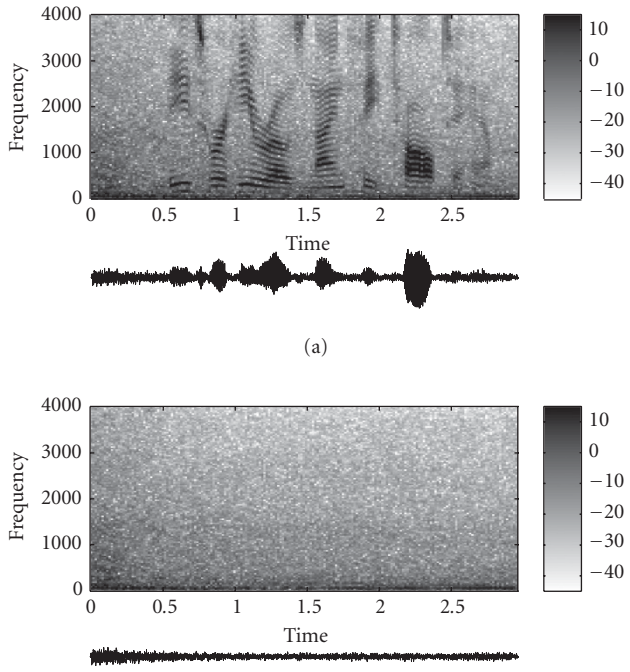

Figure 2: Spectrograms and time-domain signals of the illustrating speech recording with highway traffic noise (noisy speech) at 5 dB SNR (a) and the noise (b). The speech recording is of a male speaker uttering "*These were heroes, nine feet tall to him.*"

of the noise periodogram estimate and the noise PSD estimate, obtained using the methods we propose in Section 5, are shown in Figures 5a and 5b, respectively.

We evaluate the performance of the noise estimation methods by means of their spectral distortion, which we measure as segmental noise-to-error ratios (SegNERs). We calculate the SegNERs in the time-frequency domain, as the ratio (in dB) between the noise energy and the noise estimation error energy. These values are upper and lower limited by 35 and 0 dB [15], respectively, that is,

$$\mathrm{SegNER}(\lambda) = \min\left(\max\left(\mathrm{NER}(\lambda), 0\right), 35\right), \tag{24}$$

where

$$\mathrm{NER}(\lambda) = 10\log_{10}\left(\frac{\sum_{k=0}^{K-1}\left|N(\lambda,k)\right|^2}{\sum_{k=0}^{K-1}\left(\left|N(\lambda,k)\right| - \left|\hat{N}(\lambda,k)\right|\right)^2}\right), \tag{25}$$

and averaged over all (M) frames, that is,

$$\mathrm{SegNER} = \frac{1}{M}\sum_{\lambda=0}^{M-1}\mathrm{SegNER}(\lambda). \tag{26}$$

In Table 6, we list the average SegNERs over the same 6 speakers that are used in the informal listening test of the

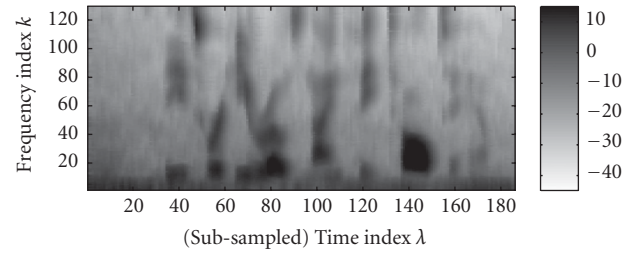

Figure 3: The noisy speech periodogram from Figure 2a after smoothing with the smoothing method from Section 3 (spectrally temporally smoothed noisy speech).

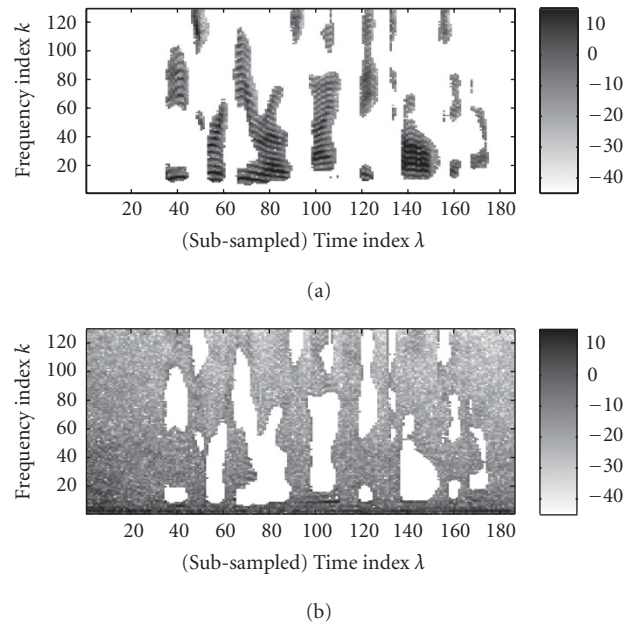

Figure 4: Noisy speech with speech absence regions removed, $D(\lambda,k) = 0$ bins removed (a); and with speech presence regions removed (b), noisy speech with $D(\lambda,k) = 1$ bins removed.

speech enhancement method. We list the average SegNERs for the noise periodogram estimation method, the noise PSD estimation method, our implementation of χ^2-based noise estimation [3], and minimum statistics (MS) noise estimation [5]. Our implementation of the χ^2-based noise estimation uses the MS noise estimate [5] in the calculation of the optimum smoothing parameters, as suggested by Martin and Lotter [3]. The spectral averaging in our implementation of the χ^2-based noise estimation is performed in sliding spectral windows of the same size as used by the two proposed noise estimation methods. We see that the noise PSD estimate has less spectral distortion than both our implementation of the χ^2-based noise estimation [3] and MS noise estimation [5]. This can be explained by a more accurate bias compensation factor, which uses speech presence information. Note that in many scenarios, the proposed smooth and low-biased noise PSD estimate is preferable over the noise periodogram estimate.

Table 6: Segmental noise-to-error ratios in dB.

Noise type	Highway traffic			Car interior		
Noisy speech SNR (dB)	0	5	10	0	5	10
Noise periodogram estimation	19.3	17.0	14.7	18.3	16.6	15.0
Noise PSD estimation	4.6	4.6	4.4	3.0	3.1	3.2
χ^2-based noise estimation [3]	3.6	3.1	2.6	2.7	2.3	2.0
MS noise estimation [5]	1.0	1.8	2.4	1.9	2.1	2.6

Table 7: Opinion score scale.

Score	Description
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

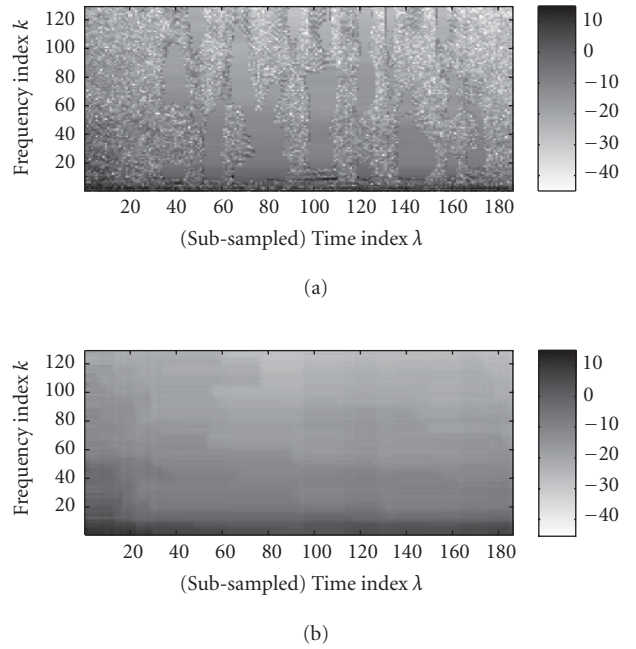

Figure 5: Spectrograms of the noise periodogram estimate (a) and the noise PSD estimate (b). In regions with speech presence, the noise periodogram estimate equals the noise PSD estimate.

As an objective measure of time-domain waveform similarity, we list the signal-to-noise ratios, and as a subjective measure of speech quality, we conduct an informal listening test. In this test, test subjects give scores from the scale in Table 7 ranging from 1 to 5, in steps of 0.1, to three different speech enhancement methods, with the noisy speech as a reference signal. Higher score is given to the preferred speech enhancement method. The test subjects are asked to take parameters, such as the naturalness of the enhanced speech, the quality of the speech, and the degree of noise reduction into account, when assigning a score to an estimate. The presentation order of estimates from individual methods is blinded, randomized, and vary in each test set and for each test subject. A total of 8 listeners, all working within the field of speech signal processing, participated in the test. The proposed speech enhancement method was compared with our implementation of two reference methods.

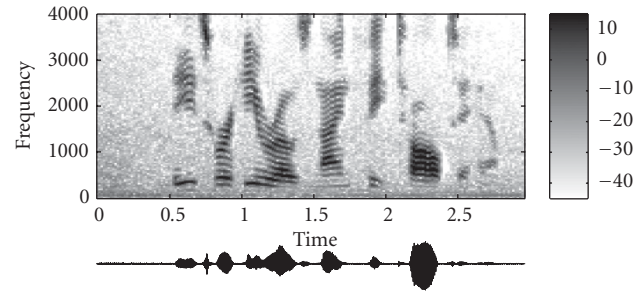

Figure 6: Spectrogram and time-domain plot of the enhanced speech from the enhancement method proposed in this paper. The noisy speech is shown in Figure 2a. The naturalness is preserved by the enhancement method and, in particular, the enhanced speech does not contain any audible musical noise.

(i) MMSE-LSA. Minimum mean-square error log-spectral amplitude estimation, as proposed by Ephraim and Malah [7].

(ii) MMSE-LSA-DD. Decision-directed MMSE-LSA, which is the MMSE-LSA estimation in combination with a smoothing mechanism [7]. Constants are as proposed by Ephraim and Malah.

All three methods in the test use the proposed noise PSD estimate, as shown in Figure 5b. Also, they all use the analysis/synthesis setup described in Section 7. The enhanced speech obtained from the noisy speech signal in Figure 2a is shown in Figure 6.

SNRs and mean opinion scores (MOSs) from the informal subjective listening test are listed in Tables 8 and 9. All results are averaged over both speakers and listeners. The best obtained results are emphasized using bold letters. To identify if the proposed method is significantly better, that is, has a higher MOS, than MMSE-LSA-DD, we use the matched sample design [16], where the absolute values of the opinion scores are eliminated as a source of variation. Let μ_d be the mean of the opinion score difference between the proposed method and the MMSE-LSA-DD. Using this formulation, we

Table 8: Highway traffic noise speech enhancement results.

	SNR (dB)	MOS	SNR (dB)	MOS	SNR (dB)	MOS
Proposed method	7.7	3.50	10.3	3.56	13.0	3.74
MMSE-LSA-DD	7.4	2.75	11.1	2.85	15.0	3.07
MMSE-LSA	4.6	1.63	9.3	1.92	14.0	2.04
Noisy speech	0.0	—	5.0	—	10.0	—

Table 9: Car interior noise speech enhancement results.

	SNR (dB)	MOS	SNR (dB)	MOS	SNR (dB)	MOS
Proposed method	10.5	3.53	13.4	3.82	16.5	3.95
MMSE-LSA-DD	7.3	2.54	10.9	2.99	15.4	3.29
MMSE-LSA	3.1	1.89	7.7	2.07	12.6	2.37
Noisy speech	0.0	—	5.0	—	10.0	—

Table 10: Highway traffic noise statistics at a 99% level of confidence.

SNR (dB)	Test statistics	Test result	Interval estimate
0	$z = 10.3$	Reject H_0	0.75 ± 0.19
5	$z = 10.2$	Reject H_0	0.72 ± 0.18
10	$z = 10.1$	Reject H_0	0.67 ± 0.17

Table 11: Car interior noise statistics at a 99% level of confidence.

SNR (dB)	Test statistics	Test result	Interval estimate
0	$z = 11.4$	Reject H_0	1.00 ± 0.23
5	$z = 9.4$	Reject H_0	0.83 ± 0.23
10	$z = 6.7$	Reject H_0	0.66 ± 0.25

write the null and alternative hypotheses as

$$H_0 : \mu_d \leq 0,$$
$$H_A : \mu_d > 0, \tag{27}$$

respectively. The null hypothesis H_0 in this context should not be mistaken for the hypothesis H_0 in the speech presence detection method. With 48 experiments at each combination of SNR and noise type, we are in the large sample case, and we therefore assume that the differences are normally distributed. The rejection rule, at 1% level of significance, is

$$\text{Reject } H_0 \quad \text{if } z > z_{.01}, \tag{28}$$

with $z_{.01} = 2.33$. Tables 10 and 11 list the test statistic z, and the corresponding test result. Also, the two-tailed 99% confidence interval [16], of the difference between the MOS of the proposed method and MMSE-LSA-DD, for highway traffic and car interior noise, respectively, is listed. From our results we can therefore state with a confidence level of 99% that the proposed method has a higher perceptual quality than MMSE-LSA-DD. Furthermore, the difference corresponds generally to more than 0.5 MOS, which generally change the ratings from somewhere between Poor and Fair to somewhere between Fair and Good on the MOS scale.

9. DISCUSSION

We have in this paper presented new noise estimation and speech enhancement methods that utilize a proposed connected region speech presence detection method. Despite the simplicity, the proposed methods are shown to have superior performance when compared to our implementation of state-of-the-art reference methods in the case of both noise estimation and speech enhancement.

In the first proposed noise estimation method, the connected speech presence regions are used to achieve noise periodogram estimates in the regions where speech is absent. In the remaining regions, where speech is present, minimum tracks of the smoothed noisy speech periodograms are bias compensated with a factor that is updated in regions with speech absence. A second proposed noise estimation method provides a noise PSD estimate by means of the same power-scaled minimum tracks that are used by the noise periodogram estimation method when speech is present. It is shown that the noise PSD estimate has less spectral distortion than both our implementation of χ^2-based noise estimation [3] and MS noise estimation [5]. This can be explained by a more accurate bias compensation factor, which uses speech presence information. The noise periodogram estimate is by far the less spectrally distorted noise estimate of the tested noise estimation methods. This verifies the connected region speech presence principle which is fundamental for the proposed speech enhancement method.

Our proposed enhancement method uses different attenuation rules for each of the two types of speech presence regions. When no speech is present, the noisy speech is downscaled and left in the speech estimate as natural sounding masking noise, and when speech is present, a noise PSD estimate is used in a traditional generalized spectral subtraction. In addition to enhancing the speech, the most distinct feature of the proposed speech enhancement method is that it

leaves natural sounding background noise matching the actual surroundings of the person wearing the hearing aid. The proposed method performs well at SNRs equal to or higher than 0 dB for noise types with slowly changing and spectrally smooth periodograms. Rapid, and speech-like, changes in the noise will be treated as speech, and will therefore be enhanced, causing a decrease in the naturalness of the background noise. At very low SNRs, the detection of speech presence will begin to fail. In this case, we suggest the implementation of the proposed method in a scheme, where low SNR is detected and causes a change to an approach with only a single and very conservative attenuation rule. Strong tonal interferences will affect the speech presence decisions as well as the noise estimation and enhancement method and should be detected and removed by preprocessing of the noisy signal immediately after the STFT analysis. Otherwise, a sufficiently strong tonal interference with duration longer than the minimum search window will cause the signal to be treated as if speech is absent and the speech enhancement algorithm will downscale the entire noisy speech by multiplication with β_0.

Our approach generalizes to other noise reduction schemes. As an example, the proposed binary scheme can also be used with MMSE-LSA-DD for the speech presence regions. For such a combination, we expect performance similar to, or better than, what we have shown in this paper for the generalized spectral subtraction. This is supported by the findings of Cohen and Berdugo [11] that have shown that a soft-decision approach improves MMSE-LSA-DD.

The informal listening test confirms that listeners prefer the downscaled background noise with fully preserved naturalness over the less realistic whitened residual noise from, for example, MMSE-LSA-DD. From our experiments, we can conclude, with a confidence level of 99%, that the proposed speech enhancement method receives significantly higher MOS than MMSE-LSA-DD at all tested combinations of SNR and noise type.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for many constructive comments and suggestions to the previous versions of the manuscript, which have largely improved the presentation of this work. This work was supported by The Danish National Centre for IT Research, Grant no. 329, and Microsound A/S.

REFERENCES

[1] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[2] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '93)*, vol. 2, pp. 363–366, Minneapolis, Minn, USA, April 1993.

[3] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms ," in *Proc. International Workshop on Acoustic Echo Control and Noise Reduction (IWAENC '01)*, pp. 43–46, Darmstadt, Germany, September 2001.

[4] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. 7th European Signal Processing Conference (EUSIPCO '94)*, pp. 1182–1185, Edinburgh, Scotland, September 1994.

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

[6] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Lett.*, vol. 9, no. 1, pp. 12–15, 2002.

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

[8] K. S. Shanmugan and A. M. Breipohl, *Random Signals - Detection, Estimation, and Data Analysis*, John Wiley & Sons, New York, NY, USA, 1988.

[9] K. V. Sørensen and S. V. Andersen, "Speech presence detection in the time-frequency domain using minimum statistics," in *Proc. 6th Nordic Signal Processing Symposium (NORSIG '04)*, pp. 340–343, Espoo, Finland, June 2004.

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

[11] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[12] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '79)*, vol. 4, pp. 208–211, Washington, DC, USA, April 1979.

[13] *DARPA TIMIT Acoustic-Phonetic Speech Database*, National Institute of Standards and Technology (NIST), Gaithersburg, Md, USA, CD-ROM.

[14] D. Brillinger, *Time Series: Data Analysis and Theory*, Holden-Day, San Francisco, Calif, USA, 1981.

[15] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Wiley-Interscience, Hoboken, NJ, USA, 2000.

[16] D. R. Anderson, D. J. Sweeney, and T. A. Williams, *Statistics for Business and Economics*, South-Western, Mason, Ohio, USA, 1990.

Karsten Vandborg Sørensen received his M.S. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 2002. Since 2003, he has been a Ph.D. student with the Digital Communications (DICOM) Group at Aalborg University. His research areas are within noise reduction in speech signals: noise estimation, speech presence detection, and enhancement.

Søren Vang Andersen received his M.S. and Ph.D. degrees in electrical engineering from Aalborg University, Aalborg, Denmark, in 1995 and 1999, respectively. Between 1999 and 2002, he was with the Department of Speech, Music and Hearing at the Royal Institute of Technology, Stockholm, Sweden, and Global IP Sound AB, Stockholm, Sweden. Since 2002, he has been an Associate Professor with the Digital Communications (DICOM) Group at Aalborg University. His research interests are within multimedia signal processing: coding, transmission, and enhancement.