CCIT Report #459

December 2003

A Noise Reduction Preprocessor for Mobile Voice Communication

Rainer Martin*, David Malah†, Richard V. Cox‡, Anthony J. Accardi§

November 18, 2003

Submitted to EURASIP Journal of Applied Signal Processing

Abstract

We describe a speech enhancement algorithm which leads to significant quality and intelligibility improvements when used as a preprocessor to a low bit rate speech coder. This algorithm was developed in conjunction with the Mixed Excitation Linear Prediction (MELP) coder which, by itself, is highly susceptible to environmental noise. The paper presents novel as well as known speech and noise estimation techniques and combines them into a highly effective speech enhancement system. The algorithm is based on short time spectral amplitude estimation, soft-decision gain modification, tracking of the *a priori* probability of speech absence, and Minimum Statistics noise power estimation. Special emphasis is placed on enhancing the performance of the preprocessor in non-stationary noise environments.

I. Introduction

With the advent and wide dissemination of mobile voice communication systems, telephone conversations are increasingly disturbed by environmental noise. This is especially true in hands-free

*Institute of Communication Acoustics, Ruhr-Universität Bochum, 44780 Bochum, Germany, E-mail: rainer.martin@rub.de.

†Dept. of Electrical Eng., Technion - Israel Institute of Technology, Haifa 32000, Israel.

‡AT&T Labs-Research, 180 Park Avenue, Florham Park, NJ 07932, U.S.A.

§Tellme Networks, 1310 Villa Avenue, Mountain View, CA 94040, U.S.A.

This work was sponsored by U.S. government contract MDA-904-97-C-0452. The authors collaborated on this project while at AT&T Labs-Research, Speech and Image Processing Services Research Lab, 180 Park Avenue, Florham Park, NJ 07932, U.S.A.

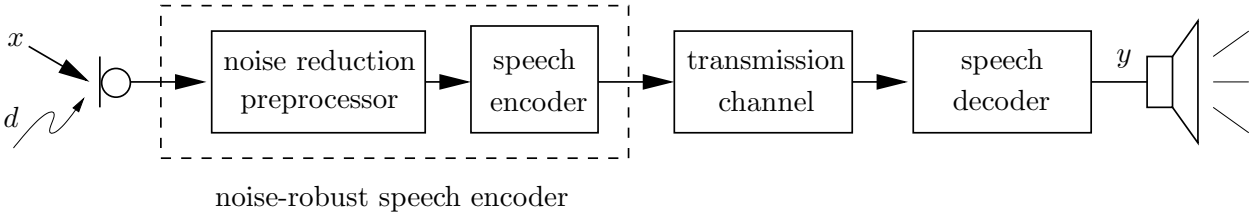

Figure 1: Speech communication system with noise reduction preprocessing.

environments where the microphone is far away from the speech source. As a result, the quality and intelligibility of the transmitted speech can be significantly degraded and fail to meet the expectations of mobile phone users. The environmental noise problem becomes even more pronounced when low bit rate coders are used in harsh acoustic environments. An example is the Mixed Excitation Linear Prediction (MELP) coder which operates at bit rates of 1.2 and 2.4 kbps. It is used for secure governmental communications and has been selected as the *Future NATO Narrowband Voice Coder* [1]. In contrast to waveform approximating coders, low bit rate coders transmit parameters of a speech production model instead of the quantized acoustic waveform itself. Thus, low bit rate coders are more susceptible to a mismatch of the input signal and the underlying signal model.

It is well known that single microphone speech enhancement algorithms improve the quality of noisy speech when the noise is fairly stationary. However, they typically do not improve the intelligibility when the enhanced signal is presented directly to a human listener. The loss of intelligibility is mostly a result of the distortions introduced into the speech signal by the noise reduction preprocessor. However, the picture changes when the enhanced speech signal is processed by a low bit rate speech coder as shown in Fig 1. In this case, a speech enhancement preprocessor can significantly improve quality as well as intelligibility [2]. Therefore, the noise reduction preprocessor should be an integral component of the low bit rate speech communication system.

Although many speech enhancement algorithms have been developed over the last two decades, such as Wiener and power-subtraction methods [3], maximum-likelihood (ML) [4], minimum mean-squared error (MMSE) [5, 6], and others [7, 8], improvements are still sought. In particular, since mobile voice communication systems frequently operate in non-stationary noise environments such as inside moving vehicles, effective suppression of non-stationary noise is of vital importance. While most existing enhancement algorithms assume that the spectral characteristics of the noise change very slowly compared to those of the speech, this may not be true when communicating from a

moving vehicle. Under such circumstances the noise may change appreciably during speech activity and so confining the noise spectrum updates to periods of speech absence may adversely affect the performance of the speech enhancement algorithm. To maximize enhancement performance, the noise characteristics should be tracked even during speech.

Most common enhancement techniques, including those cited above, operate in the frequency domain. These techniques apply a frequency-dependent gain function to the spectral components of the noisy signal, in an attempt to attenuate the noisier components to a greater degree. The gains applied are typically nonlinear functions of estimated signal and noise powers at each frequency. These functions are usually derived by either estimating the clean speech (e.g., the Wiener approach) or its spectral magnitude according to a specific optimization criterion (e.g., ML, MMSE). The noise suppression properties of these enhancement algorithms have been shown to improve when a *soft-decision* modification of the gain function, which takes speech-presence uncertainty into account, is introduced [4, 5, 9, 7]. To implement such a gain modification function, one must provide a value to the *a priori* probability of speech absence for each spectral component of the noisy signal. Therefore, we use the algorithm in [9] to estimate the *a priori* probability of speech absence as a function of frequency, on a frame-by-frame basis.

The objective of this paper is to describe a single microphone speech enhancement preprocessor which has been developed for voice communication in non-stationary noise environments with high quality *and* intelligibility requirements. Recently, this preprocessor has been proposed as an optional part of the *Future NATO Narrow Band Voice Coder* standard (also known as the MELPe coder [1]) and, in a slightly modified form, in conjunction with one of the ITU-T 4 kbps coder [10] proposals. The improvements we obtain with this system result from a synergy of several carefully designed system components. Significant contributions to the overall performance stem from a novel procedure for estimating the *a priori* probability of speech absence, and from a noise power spectral density estimation algorithm with small error variance and good tracking properties.

A block diagram of the algorithm is shown in Figure 2. Spectral analysis consists of applying a window and the DFT. Spectral synthesis inverts the analysis with the IDFT and overlap-adding consecutive frames. The algorithm includes an MMSE estimator for the spectral amplitudes, a procedure for estimating the noise power spectral density (PSD), the long term signal-to-noise ratio (SNR), and the *a priori* SNR, as well as a mechanism for the tracking of the *a priori* probability of

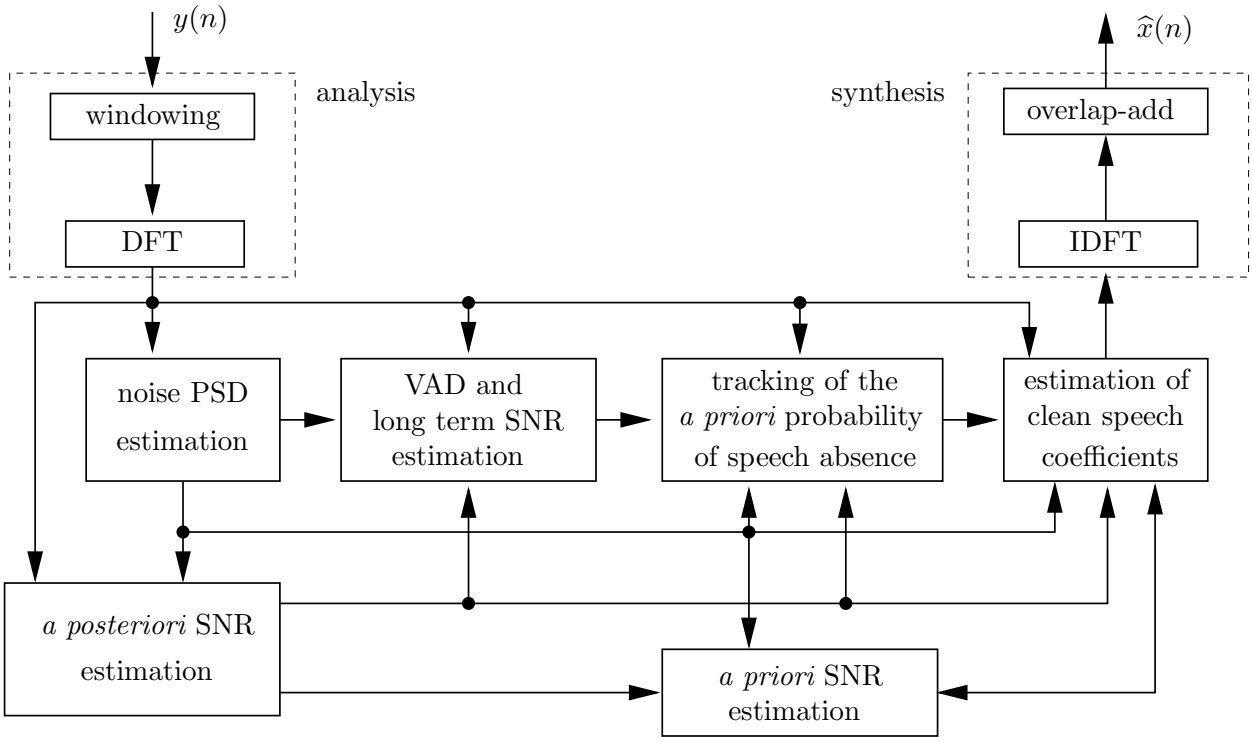

Figure 2: Block diagram of speech enhancement preprocessor.

speech absence. The spectral estimation procedure attenuates frequency components which contain primarily noise and passes those which contain mostly speech. As a result, the overall SNR of the processed speech signal is improved.

In the remainder of this paper we describe this algorithm in detail and evaluate its performance. In Section II we discuss windows for DFT-based spectral analysis and synthesis as well as the algorithmic delay of the joint enhancement and coding system. Sections III, IV, and V present estimation procedures for the spectral coefficients and the long term SNR. We outline the noise estimation algorithm [11] in Section VI, and summarize listening test results in Section VII. Section VIII concludes the paper. We reiterate that some components have been previously published [6, 9, 12, 11]. Our goal here is to tie all required components together, thereby providing a comprehensive description of the MELPe enhancement system.

II. Spectral Analysis and Synthesis

Assuming an additive, independent noise model, the noisy signal $y(n)$ is given by $x(n)+d(n)$, where $x(n)$ denotes the clean speech signal, and $d(n)$ the noise. All signals are sampled at a sampling rate of f_s. We apply a short-time Fourier analysis to the input signal by computing the DFT of each overlapping windowed frame,

$$Y(k,m) = \sum_{\ell=0}^{L-1} y(mM_E + \ell)h(\ell)e^{-j2\pi k\ell/L}\,. \tag{1}$$

Here, M_E denotes the frame shift, $m \in \mathbb{Z}$ is the frame index, $k \in \{0, 1..., L-1\}$ is the frequency bin index, which is related to the normalized center frequency $\Omega_k = k2\pi/L$, and $h(\ell)$ denotes the window function. Typical implementations of DFT-based noise reduction algorithms use a Hann window with a 50 % overlap ($M_E/L = 0.5$) or a Hamming window with a 75% overlap ($M_E/L = 0.25$) for spectral analysis, and a rectangular window for synthesis.

When no confusion is possible, we drop the frame index m and will write the frequency index k as a subscript. Thus, for a given frame m we have

$$Y(k,m) = X(k,m) + D(k,m) \quad \text{or} \quad Y_k = X_k + D_k, \tag{2}$$

where X_k and Y_k are characterized by their amplitudes A_k and R_k and their phases φ_k and θ_k, respectively,

$$\begin{aligned} X_k &= A_k \exp(j\varphi_k) \\ Y_k &= R_k \exp(j\theta_k). \end{aligned} \tag{3}$$

In the gain function derivations cited below, it is assumed that the DFT coefficients of both the speech and the noise are independent Gaussian random variables.

The segmentation of the input signal into frames and the selection of an analysis window is closely linked to the frame alignment of the speech coder [12] and the admissible algorithmic delay. The analysis/synthesis system must balance conflicting requirements of sufficient spectral resolution, little spectral leakage, smooth transitions between signal frames, low delay, and low complexity. Delay and complexity constraints limit the overlap of the signal frames. However, the frame advancement must not be too aggressive so as to degrade the enhanced signal's quality.

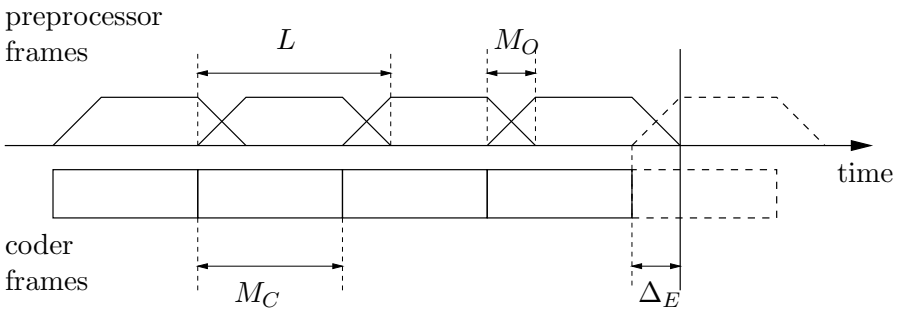

Figure 3: Frame alignment of enhancement preprocessor and speech coder with $M_E = M_C$.

When the frame overlap is less than 50%, we have obtained good results with a flat-top (Tukey) analysis window and a rectangular synthesis window.

The total algorithmic delay of the joint enhancement and coding system is minimized when the frame shift of the noise reduction preprocessor is adjusted such that $l(L - M_O) = lM_E = M_C$, with $l \in \mathbb{N}$ and where M_C and M_O denote the frame length of the speech coder and the length of the overlapping portions of the preprocessor frames, respectively. This situation is depicted in Figure 3.

The additional delay Δ_E, due to the enhancement preprocessor, is equal to M_O. For the MELP coder and its frame length of $M_C = 180$, we use an FFT length of $L = 256$ and have $M_O = 76$ overlapping samples between adjacent signal frames.

Reducing the number of overlapping samples M_O, and thus the delay of the joint system, has several effects. First, with a flat-top analysis window, this decreases the sidelobe attenuation during spectral analysis, which leads to increased crosstalk between frequency bins that might complicate the speech enhancement task. Most enhancement algorithms assume that adjacent frequency bins are independent and do not exploit correlation between bins. Second, as the overlap between frames is reduced, transitions between adjacent frames of the enhanced signal become less smooth. Discontinuities arise because the analysis window attenuates the input signal most at the ends of a frame, while estimation errors, which occur during the processing of the frame in the spectral domain, tend to spread evenly over the whole frame. This leads to larger relative estimation errors at the frame ends. The resulting discontinuities, which are most notable in low SNR conditions, may lead to pitch estimation errors and other speech coder artifacts.

These discontinuities are greatly reduced if we use a tapered window for spectral synthesis as

well as one for spectral analysis [12]. We found that a tapered synthesis window is beneficial when the overlap M_O is less than 40% of the DFT length L. In this case, the square root of the Tukey window

$$h(n) = \begin{cases} \sqrt{0.5(1-\cos(\frac{\pi n}{M_O}))} & 1 \le n \le M_O \\ 1 & M_O + 1 \le n \le L - M_O - 1 \\ \sqrt{0.5(1-\cos(\frac{\pi(L-n)}{M_O}))} & L - M_O \le n \le L \end{cases} \tag{4}$$

can be used as an analysis and synthesis window. It results in a perfect reconstruction system if the signal is not modified between analysis and synthesis. Note, that the use of a tapered synthesis window is also in line with the results of Griffin and Lim [13] for the MMSE reconstruction of modified short time spectra.

III. Estimation of Speech Spectral Coefficients

Let C_k be some function of the short-time spectral amplitude A_k of the clean speech in the k-th bin (e.g., A_k, $\log A_k$, A_k^2). Taking the uncertainty of speech presence into account, the MMSE estimator $\widehat{C}_k$ of C_k is given by [4]:

$$\begin{aligned} \widehat{C}_k &= E\{C_k|Y_k, H_1^k\}P(H_1^k|Y_k) \\ &+ E\{C_k|Y_k, H_0^k\}P(H_0^k|Y_k) \end{aligned} \tag{5}$$

where H_0^k and H_1^k represent the hypotheses of speech being absent or present in the k-th frequency bin, i.e.,

$$\begin{aligned} H_0^k &: \quad \text{speech absent in } k\text{-th DFT bin,} \\ H_1^k &: \quad \text{speech present in } k\text{-th DFT bin,} \end{aligned}$$

and $E\{\cdot|\cdot\}$ and $P(\cdot|\cdot)$ denote conditional expectations and conditional probabilites, respectively. Since $E\{C_k|Y_k, H_0^k\} = 0$, we have:

$$\widehat{C}_k = E\{C_k|Y_k, H_1^k\}P(H_1^k|Y_k). \tag{6}$$

$P(H_1^k|Y_k)$ is thus the "soft-decision" modification of the optimal estimator under the signal presence hypothesis.

Applying Bayes' rule, one obtains [4] [5]:

$$P(H_1^k|Y_k) = \frac{p(Y_k|H_1^k)P(H_1^k)}{p(Y_k|H_0^k)P(H_0^k) + p(Y_k|H_1^k)P(H_1^k)} = \frac{\Lambda_k}{1 \ + \ \Lambda_k} \stackrel{\triangle}{=} G_M(k), \tag{7}$$

where $p(\cdot \mid \cdot)$ represents conditional probability densities, and

$$\Lambda_k \stackrel{\triangle}{=} \mu_k \frac{p(Y_k|H_1^k)}{p(Y_k|H_0^k)}; \quad \mu_k \stackrel{\triangle}{=} \frac{P(H_1^k)}{P(H_0^k)} = \frac{1-q_k}{q_k}.$$

Λ_k is a generalized likelihood ratio and q_k denotes the *a priori* probability of speech absence in the k-th bin.

$\widehat{C}_k$ is then used to find an estimate of the clean signal spectral amplitude A_k. If $C_k = A_k$, as for the MMSE amplitude estimator, one gets [5]:

$$\hat{A}_{SA}(k) = G_M(k) G_{SA}(k) R_k, \tag{8}$$

where, $\hat{A}_{SA}(k)$ is the MMSE estimator of A_k that takes into account speech presence uncertainty and, according to (6) and (7), $G_M(k)$ is the modification function of $G_{SA}(k) = E\{A_k|Y_k, H_1^k\}/R_k$. The derivation of $G_{SA}(k)$ can be found in [5].

A. MMSE-LSA and MM-LSA Estimators

Based on the results reported in [6], we prefer using the MMSE-LSA estimator (corresponding to $C_k = \log A_k$) over the MMSE-STSA ($C_k = A_k$) estimator [5], as the basic enhancement algorithm. In this case the amplitude estimator has the form:

$$\begin{aligned} \hat{A}_{LSA}(k) &= \exp[E\{\log A_k|Y_k, H_1^k\} G_M(k)] \\ &\stackrel{\triangle}{=} [G_{LSA}(k) R_k]^{G_M(k)}, \end{aligned} \tag{9}$$

where, again, $G_M(k)$ is the gain modification function defined in (7) and satisfies, of course, $0 \leq G_M(k) \leq 1$. Because the soft-decision modification of R_k in (9) is not multiplicative and does not result in a meaningful improvement over using $G_{LSA}(k)$ alone [6], we choose to use the following estimator, which is called the multiplicatively-modified LSA (MM-LSA) estimator [9]:

$$\hat{A}_L(k) = G_M(k) G_{LSA}(k) R_k \stackrel{\triangle}{=} G_L(k) R_k. \tag{10}$$

It should be mentioned that in [14, 15] the 2nd term in (5) is not zeroed out, as we did in arriving at (6), but is rather constrained in such a way that (9) can be replaced by: $[G_{LSA}(k)R_k]^{G_M(k)}[G_{min}R_k]^{1-G_M(k)}$,

where G_{min} is a threshold gain value [14, 15]. This way, one gets an exact multiplicative modification of R_k, by replacing the expression for $G_L(k)$ in (10) with $G_{LSA}(k)^{G_M(k)} G_{min}^{1-G_M(k)}$. Since the computation of $G_L(k)$ according to (10) is simpler, and gives close results for a wide range of practical SNR values [15], we prefer to continue with (10).

Under the above assumptions on speech and noise, the gain function $G_{LSA}(k)$ is derived in [6] to be:

$$G_{LSA}(\xi_k, \gamma_k) \;=\; \frac{\xi_k}{1+\xi_k} \exp\left(\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right), \tag{11}$$

where,

$$\begin{aligned} v_k &\triangleq \frac{\xi_k}{1+\xi_k}\gamma_k \;; & \gamma_k &\triangleq \frac{R_k^2}{\lambda_d(k)} \\ \xi_k &\triangleq \frac{\eta_k}{1-q_k} \;; & \eta_k &\triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \\ \lambda_x(k) &\triangleq E\{|X_k|^2\} = E\{A_k^2\} \;; & \lambda_d(k) &\triangleq E\{|D_k|^2\}. \end{aligned}$$

In [6], γ_k is called the *a posteriori* SNR for bin k, η_k is called the *a priori* SNR, and q_k is the prior probability of speech absence discussed earlier (see (7)).

With the above definitions, the expression for Λ_k in (7) is given by [5]:

$$\Lambda_k = \mu_k \frac{\exp(v_k)}{1+\xi_k}\bigg|_{\xi_k=\eta_k/(1-q_k)} \tag{12}$$

In order to evaluate these gain functions, one must first estimate the noise power spectrum λ_d. This is often done during periods of speech absence as determined by a Voice Activity Detector (VAD), or, as we will show below using the Minimum Statistics [11] approach. The estimated noise spectrum and the squared input amplitude R_k^2 provide an estimate for the *a posteriori* SNR. In [5] and [6], a decision-directed approach for estimating the *a priori* SNR is proposed:

$$\hat{\eta}_k(m) = \alpha_\eta \frac{\hat{A}^2(k,m)}{\lambda_d(k,m-1)} \;+\; (1 \;-\; \alpha_\eta)\max\{(\gamma(k,m) \;-\; 1), 0\}, \tag{13}$$

where $0 \le \alpha_\eta \le 1$.

An important property of both the MMSE-STSA [5] and the MMSE-LSA [6] enhancement algorithms is that they do not produce *musical noise* [16] that plagues many other frequency-domain algorithms. This can be attributed to the above decision-directed estimation method for the *a priori* SNR [16]. To improve the perceived performance of the estimator, [16] recommends

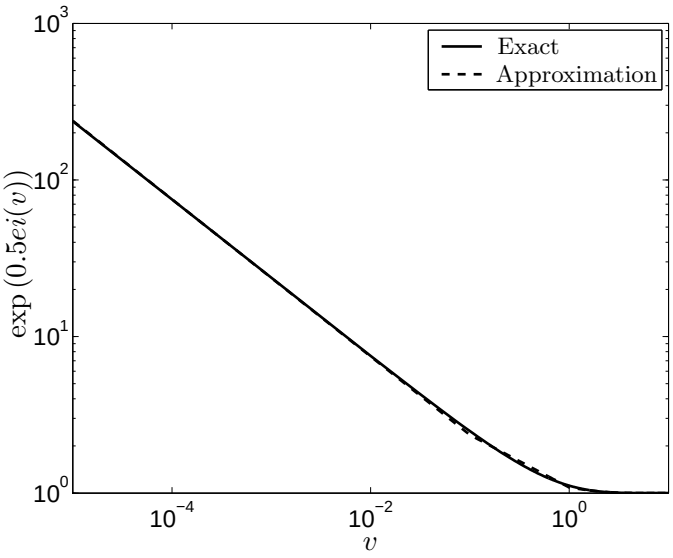

Figure 4: An approximation of $\exp(0.5ei(v))$ using the approximation for $ei(v)$ in (15).

imposing a lower limit η_{MIN} on the estimated η_k, analogous to the use of a "spectral floor" in [17]. This lower limit depends on the overall SNR of the noisy speech and may be adaptively adjusted as outlined in Section V. The parameter α_η in (13) provides a tradeoff between noise reduction and signal distortion. Typical values for α_η range between 0.90 and 0.99, where at the lower end one obtains less noise reduction but also less speech distortion.

Before we consider the estimation of the prior probabilities, we mention that in order to reduce computational complexity, the Exponential Integral in (11) may be evaluated using the functional approximation below instead of iterative solutions or tables. Thus, to approximate

$$ei(v) \triangleq \int_v^\infty \frac{e^{-t}}{t} dt \tag{14}$$

we use:

$$\tilde{ei}(v) = \begin{cases} -2.31\log_{10}(v) - 0.6 & \text{for } v < 0.1 \\ -1.544\log_{10}(v) + 0.166 & \text{for } 0.1 \le v \le 1 \\ 10^{-(0.52v+0.26)} & \text{for } v > 1. \end{cases} \tag{15}$$

Since in (11) we need $\exp(0.5ei(v))$, we show this function (solid line) alongside its approximation (dashed line) in Fig.4. For the present purpose this approximation is more than adequate.

B. Estimation of Prior Probabilities

A key feature of our speech enhancement algorithm is the estimation of the set of prior probabilities $\{q_k\}$ required in (11) and (12), where k is the frequency bin index. Our first objective is to estimate a fixed q (i.e., a frequency independent value) for each frame that contains speech. The basic idea is to estimate the relative number of frequency bins that do not contain speech and use a short time average of this statistic as an estimate for q. Due to this averaging, the estimated q will vary in time and will serve as a control parameter in the above gain expressions.

The absence of speech energy in the k-th bin clearly corresponds to $\eta_k = 0$. However, since the analysis is done with a finite length window, we can expect some leakage of energy from other bins. In addition, the human ear is unable to detect signal presence in a bin if the SNR is below a certain level η_{min}. In general, η_{min} can vary in frequency and should be chosen in accordance with a perceptual masking model. Here we choose a constant η_{min} for all the frequency bins, and set its value to the minimum level, $\eta_{\rm MIN}$, that the estimate $\hat{\eta}$ in (13) is allowed to attain. The values used in our work ranged between 0.1 and 0.2. It is interesting to note that the use of a lower threshold on the *a priori* SNR has a similar effect to constraining the gain, when speech is absent, to some G_{min}, which is the basis for the derivation of the gain function in [14, 15].

Due to the nonlinearity of the estimator for η_k in (13), there is a "locking" phenomenon to $\eta_{\rm MIN}$ when the speech signal level is low. Hence, one could consider using $\eta_{\rm MIN}$ as a threshold value to which $\hat{\eta}_k$ is compared in order to decide whether or not speech is present in bin k. However, our attempt to use this threshold resulted in excessively high counts of noise-only bins, leading to high values of q (i.e., closer to one). This is easily noticed in the enhanced signal which suffers from an over-aggressive attenuation by the gain modification function $G_M(k)$.

We therefore turn our attention to the *a posteriori* SNR, γ_k, defined in (11) and determined directly from the squared amplitude R_k^2, once an estimate for noise spectrum $\lambda_d(k)$ is given. Assuming that the DFT coefficients of the speech and noise are independent Gaussian random variables, the pdf of γ_k for a given value of the *a priori* SNR, η_k, is given by [5]:

$$p(\gamma_k) \;=\; \frac{1}{1+\eta_k}\exp\left(-\frac{\gamma_k}{1+\eta_k}\right) \;\;;\;\; \gamma_k \geq 0. \tag{16}$$

To decide whether speech is present in the k-th bin (in the sense that the true η_k has a value larger

or equal to $\eta_{\min}$), we consider the following composite hypotheses:

$$\begin{aligned} \mathcal{H}_0 \quad &: \qquad \eta_k \geq \eta_{\min} \quad \text{(speech present in } k\text{-th bin)} \\ \mathcal{H}_A \quad &: \qquad \eta_k < \eta_{\min} \quad \text{(speech absent in } k\text{-th bin)} \end{aligned}$$

We have chosen the *null hypothesis* $\mathcal{H}_0$ as stated above since its rejection when true is more grave than the alternative error of accepting when false. This is because the first type of error corresponds to deciding that speech is absent in the bin when it is actually present. Making this error would increase the estimated value of q, which would have a worse effect on the enhanced speech than if the value of q is under-estimated. Since η_k parameterizes the pdf of γ_k, as shown in (16), γ_k can be used as a test statistic. In particular, since the likelihood ratios that correspond to simple alternatives to the above two hypotheses

$$\frac{p(\gamma_k \mid \eta_k = \eta_{\min})}{p(\gamma_k \mid \eta_k = \eta_k^a)} \ , \tag{17}$$

for any $\eta_k^a < \eta_{\min}$, are monotonic functions in γ_k (for $\gamma_k > 0$ and any chosen $\eta_{\min} > 0$), it can be shown [18] that the likelihood ratio test for the following decision between two simple hypotheses is a *uniformly most powerful test* for our original problem:

$$\begin{aligned} \mathcal{H}_0' \quad &: \quad \eta_k = \eta_{\min} \\ \mathcal{H}_A' \quad &: \quad \eta_k = \eta_k^a \ ; \qquad \eta_k^a < \eta_{\min} \end{aligned} \tag{18}$$

This gives the test:

$$\gamma_k \underset{\mathcal{H}_A}{\overset{\mathcal{H}_0}{\gtrless}} \gamma_{\text{TH}}, \tag{19}$$

where γ_{TH} is set to satisfy a desired significance level [19] (or size [18]) α_0 of the test. I.e., α_0 is the probability of rejecting $\mathcal{H}_0$ when true, and is therefore:

$$\alpha_0 = \int_0^{\gamma_{\text{TH}}} p(\gamma_k | \eta_k = \eta_{\min}) d\gamma_k \tag{20}$$

Substituting the pdf of γ_k from (16), we obtain:

$$\gamma_{\text{TH}} = (1 + \eta_{\min}) \log \left(\frac{1}{1 - \alpha_0} \right). \tag{21}$$

Let M be the number of positive frequency bins to consider. Typically, $M = (L/2) + 1$, where L is the DFT transform size. However, if the input speech is limited to a narrower band, M should be chosen accordingly. Let $N_q(m)$ be the number of bins out of the M examined bins in frame m for which the test in (19) results in the rejection of hypothesis $\mathcal{H}_0$. With $r_q(m) \stackrel{\triangle}{=} N_q(m)/M$, the proposed estimate for $q(m)$ is formed by recursively smoothing $r_q(m)$ in time:

$$\hat{q}(m) \;=\; \alpha_q \hat{q}(m-1) \;+\; (1 \;-\; \alpha_q) r_q(m) \tag{22}$$

The smoothing in (22) is performed only for frames which contain speech (as determined from a VAD). We selected the parameters based on informal listening tests. We noticed improved performance with $\alpha_0 = 0.5$ (giving $\gamma_{\text{TH}} = 0.8$ in (21)) and $\alpha_q = 0.95$ in (22).

Yet, as discussed earlier, a better gain-modification could be expected if we allow different q's in different bins. Let $I(k,m)$ be an index function that denotes the result of the test in (19), in the k-th bin of frame m. That is, $I(k,m) = 1$ if $\mathcal{H}_0$ is rejected, and $I(k,m) = 0$ if it is accepted. We suggest the following estimator for $q(k,m)$:

$$\hat{q}(k,m) \;=\; \alpha_q \hat{q}(k,m-1) \;+\; (1 \;-\; \alpha_q) I(k,m) \tag{23}$$

The same settings for γ_{TH} and α_q above are appropriate here also. This way, averaging $\hat{q}(k,m)$ over k in frame m results in the $\hat{q}(m)$ of (22).

IV. Voice Activity Detection and Long Term SNR Estimation

The noise power estimation algorithm described in Section VI does not rely on a voice activity detector and therefore need not deal with detection errors. Nevertheless, it is beneficial to have a VAD available for controlling certain aspects of the preprocessor. In our algorithm we use VAD decisions to control estimates of the *a priori* probability of speech absence and of the long term signal-to-noise ratio. We briefly describe our *delayed decision* VAD and the long term SNR estimation.

As in [7] (see also [20]), we found that the mean value $\bar{\gamma}$ of γ_k (averaged over all frequency bins in a given frame), is useful for indicating voice activity in each frame. For stationary noise and assuming independent DFT coefficients, $\bar{\gamma}$ is approximately normal with mean 1 and standard deviation $\sigma_{\bar{\gamma}} = \sqrt{1/M}$ (for sufficiently large M, which is usually the case). Thus, by comparing

$\bar{\gamma}$ to a suitable fixed threshold, one can obtain a reliable VAD — as long as the short-time noise spectrum does not change too fast. Typically, we use threshold values $\bar{\gamma}_{\text{th}}$ in the range between 1.35 and 2, where the lower value, which we denote by $\bar{\gamma}_{\text{th}}^{\text{min}}$, corresponds to $1+4\sigma_{\bar{\gamma}}$ for $M = L/2+1$ with a transform size of $L = 256$ (32 msec window). We found this value suitable for stationary noise at input SNR values down to 3 dB. The higher threshold value allows for larger fluctuations of $\bar{\gamma}$ (as expected if the noise is non-stationarity) without causing a decision error in noise-only frames, but may result in misclassification of weak speech signals as noise, particularly at SNR values below 10 dB. We may further improve the VAD decision by considering the maximum of γ_k, $k = 0 \ldots M$, and the average frame SNR. We declare a speech pause if $\bar{\gamma} < \bar{\gamma}_{\text{th}}$, $\max\limits_{k}(\gamma_k) < \gamma_{\text{max-th}}$, and $\text{mean}(\widehat{\eta}(k,m)) < 2\bar{\gamma}_{\text{th}}$, where $\gamma_{\text{max-th}} \approx 25\bar{\gamma}_{\text{th}}$. Finally, we require a consistent VAD decision for at least two consecutive frames before taking action.

The long term signal-to-noise ratio $SNR_{LT}(m)$ characterizes the SNR of the noisy input speech averaged over periods of one to two seconds. It is used for the adaptive limiting of the *a priori* SNR and the adaptive smoothing of the signal power, as outlined below. The computation of $SNR_{LT}(m)$ requires a VAD since the average speech power can be updated only if speech is present. The signal power is computed using a first order recursive system update on the average frame power with time constant T_{LT}:

$$\overline{\lambda_y}(m) = \alpha_{LT}\overline{\lambda_y}(m-1) + (1-\alpha_{LT})\frac{1}{M+1}\sum_{k=0}^{M} R^2(k,m)\,, \tag{24}$$

where $\alpha_{LT} \approx 1 - M_E/(T_{LT} f_s)$. $SNR_{LT}(m)$ is then given by

$$SNR_{LT}(m) = \frac{(M+1)\overline{\lambda_y}(m)}{\sum_{k=0}^{M}\lambda_d(k,m)} - 1. \tag{25}$$

If $SNR_{LT}(m)$ is smaller than zero it is set equal to $SNR_{LT}(m-1)$, the estimated long term SNR of the previous frame.

V. Adaptive Limiting of the *A Priori* SNR

After applying the noise reduction preprocessor described so far to the MELP coder, we found that most of the degradations in quality and intelligibility that we witnessed were due to errors in estimating the spectral parameters in the coder. In this section, we present a modified spectral

weighting rule which allows for better spectral parameter reproduction in the MELP coder, where *Linear Predictive Coefficients* (LPC) are transformed into *Line Spectral Frequencies* (LSF). We use an adaptive limiting procedure on the spectral gain factors applied to each DFT coefficient. We note that while spectral valleys in between formant frequencies are not important for speech *perception* (and thus can be filled with noise to give a better auditory impression), they are important for LPC *estimation*.

It was stressed in [16, 9] that in order to avoid structured "musical" residual noise and to achieve good audio quality, the *a priori* SNR estimate $\hat{\eta}_k$ should be limited to values between 0.1 and 0.2. This means that less signal attenuation is applied to bins with low SNR in the spectral valleys between formants. By limiting the attenuation, we largely avoid the annoying "musical" distortions and the residual noise appears very natural. However, this attenuation distorts the overall spectral shape of speech sounds, which impacts the spectral parameter estimation. One solution to this problem is the adaptive limiting scheme we outline below.

We utilize the VAD to distinguish between speech+noise and noise only signal frames. Whenever we detect pauses in speech, we set a preliminary lower limit for the *a priori* SNR estimate in the m-th frame to $\eta_{\mathrm{MIN1}}(m) = \eta_{minP}$ (typically, $\eta_{minP} = 0.15$) in order to achieve a smooth residual noise. During speech activity, the lower limit $\eta_{\mathrm{MIN1}}(m)$ is set to

$$\eta_{\mathrm{MIN1}}(m) = \eta_{minP} 0.0067 (0.5 + SNR_{LT}(m))^{0.65} \tag{26}$$

and is limited to a maximum of 0.25. We obtained (26) by fitting a function to data from listening tests using several long term SNR values. We then smooth this result using a first order recursive system,

$$\eta_{\mathrm{MIN}}(m) = 0.9 \eta_{\mathrm{MIN}}(m-1) + 0.1 \eta_{\mathrm{MIN1}}(m) \,, \tag{27}$$

to obtain smooth transitions between active and pause segments. We use the resulting η_{MIN} as a lower limit for $\hat{\eta}_k$. The enhanced speech sounds appear to be less noisy when using the adaptive limiting procedure, while at the same time the background noise during speech pauses is very smooth and natural. This method was also found to be effective in conjunction with other speech coders. A slightly different dynamic lower limit optimized for the 3GPP AMR coder [21] is given in [22].

VI. Noise Power Spectral Density Estimation

The importance of an accurate noise power spectral density estimate can be easily demonstrated in a computer simulation by estimating it directly from the isolated noise source. In fact, it turns out that many of the annoying artifacts in the processed signal are due to errors in the noise PSD estimate. It is therefore of paramount importance both to estimate the noise power spectral density with a small error variance and to effectively track non-stationary noise. This requires a careful balance between the degree of smoothing and the noise tracking rate.

A common approach is to use a VAD and to update the estimated noise PSD during speech pauses. Since the noise PSD might also fluctuate during speech activity, VAD-based methods do not work satisfactorily when the noise is non-stationary or when the SNR is low. Soft-decision update strategies which take the probability of speech presence in each frequency bin into account [9, 20] allow us to also update the noise PSD during speech activity, e.g., in between the formants of the speech spectrum or in between the pitch peaks during voiced speech.

The approach we present here is based on the "Minimum Statistics" method [23, 11] which is very robust, even for low SNR conditions. The Minimum Statistics approach assumes that speech and noise are statistically independent and that the spectral characteristics of speech vary faster in time than those of the noise. During both speech pauses and speech activity, the PSD of the noisy signal frequently decays to the level of the noise. The noise floor can therefore be estimated by tracking spectral minima within a finite time window without relying on a VAD decision. The noise PSD can be updated during speech activity, just as with soft-decision methods. An important feature of the Minimum Statistics method is its use of an optimally smoothed power estimate which provides a balance between the error variance and effective tracking properties.

A. Adaptive Optimal Short Term Smoothing

To derive an optimal smoothing procedure for the PSD of the noisy signal, we assume a pause in speech and consider a first order smoothing recursion for the short term power of the DFT coefficients $Y(k,m)$ of the m-th frame (1), using a time and frequency dependent smoothing parameter $\alpha(k,m)$:

$$\widehat{\lambda}_y(k,m+1) = \alpha(k,m)\widehat{\lambda}_y(k,m) + (1-\alpha(k,m))|Y(k,m)|^2 \,. \tag{28}$$

Since we want $\widehat{\lambda}_y(k,m)$ to be as close as possible to the true noise PSD $\lambda_d(k,m)$, our objective is to minimize the conditional mean squared error:

$$E\{(\widehat{\lambda}_y(k,m+1) - \lambda_d(k,m))^2 \mid \widehat{\lambda}_y(k,m)\} \tag{29}$$

from one frame to the next. After substituting (28) for $\widehat{\lambda}_y(k,m+1)$ in (29) and using $E\{|Y(k,m)|^2\} = \lambda_d(k,m)$ and $E\{|Y(k,m)|^4\} = 2\lambda_d^2(k,m)$, the mean squared error is given by:

$$\begin{aligned} &E\{(\widehat{\lambda}_y(k,m+1) - \lambda_d(k,m))^2 \mid \widehat{\lambda}_y(k,m)\} \\ &\quad = \alpha^2(k,m)(\widehat{\lambda}_y(k,m) - \lambda_d(k,m))^2 \\ &\qquad + \lambda_d^2(k,m)(1-\alpha(k,m))^2 \,, \end{aligned} \tag{30}$$

where we also assumed the statistical independence of successive signal frames. Setting the first derivative with respect to $\alpha(k,m)$ to zero yields

$$\alpha_{opt}(k,m) = \frac{1}{1 + (\widehat{\lambda}_y(k,m)/\lambda_d(k,m) - 1)^2} \,, \tag{31}$$

and the second derivative, being non-negative, reveals that this is indeed a minimum. The term $\widehat{\lambda}_y(k,m)/\lambda_d(k,m) = \overline{\gamma}(k,m)$ on the right hand side of (31) is a smoothed version of the *a posteriori* SNR. Figure 5 plots the optimal smoothing parameter α_{opt} for $0 \leq \overline{\gamma} \leq 10$. This parameter is between zero and one, thus guaranteeing a stable and non-negative noise power estimate $\widehat{\lambda}_y(k,m)$.

Assuming a pause in speech in the above derivation does not pose any major problems. The optimal smoothing procedure reacts to speech activity in the same way as to highly non-stationary noise. During speech activity, the smoothing parameter is small, allowing the PSD estimate to closely follow the time varying PSD of the noisy speech signal.

To compute the optimal smoothing parameter in (31), we replace the true noise PSD $\lambda_d(k,m)$ with an estimate $\widehat{\lambda}_d(k,m)$. However, since the estimated noise PSD may be either too small or too large we have to take special precautions. If the computed smoothing parameter is smaller than the optimal value, the smoothed PSD estimate $\widehat{\lambda}_y(k,m)$ will have an increased variance. This is not a problem if the noise estimator is unbiased, since the smoothed PSD will still track the true signal PSD and the estimated noise PSD will eventually converge to the true noise PSD. However, if the computed smoothing parameter is too large, the smoothed power will not accurately track the true signal PSD, leading to noise PSD estimation errors. We therefore introduce an additional

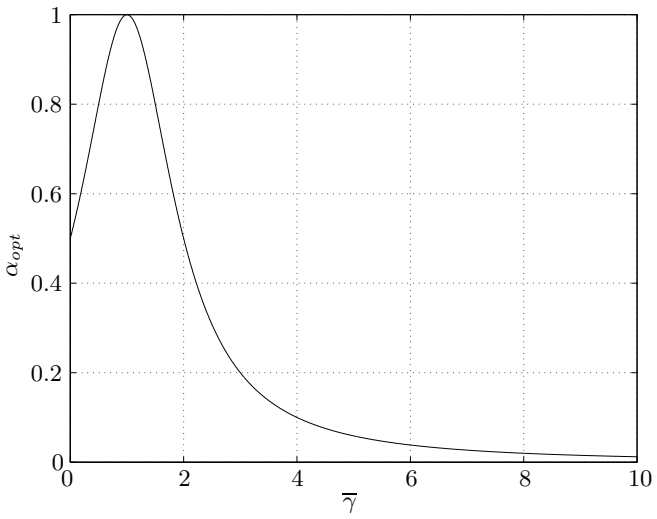

Figure 5: Optimal smoothing parameter α_{opt} as a function of the smoothed *a posteriori* SNR $\overline{\gamma}(k,m)$.

factor $\alpha_c(m)$ in the numerator of the smoothing parameter which decreases whenever deviations between the average smoothed PSD estimate and the average signal power are detected. Now the smoothing parameter has the form

$$\alpha(k,m) = \frac{\alpha_c(m)}{1 + (\widehat{\lambda}_y(k,m)/\widehat{\lambda}_d(k,m) - 1)^2} \tag{32}$$

where

$$\alpha_c(m) = c_{max}\alpha_c(m-1) + (1 - c_{max})\max(\widetilde{\alpha}_c(m), 0.7) \tag{33}$$

and

$$\widetilde{\alpha}_c(m) = \frac{\alpha_{max}}{1 + (\sum_{k=0}^{L-1}\widehat{\lambda}_y(k,m)/\sum_{k=0}^{L-1}|Y(k,m)|^2 - 1)^2}\,. \tag{34}$$

α_{max} is a constant smaller than but close to 1 and prevents the freezing of the PSD estimator. c_{max} does not appear to be a sensitive parameter and was set to 0.7. Equation (34) ensures that the average smoothed power of the noisy signal cannot deviate by a large factor from the power of the current frame. The ratio of powers $\Xi = \sum_{k=0}^{L-1}\widehat{\lambda}_y(k,m)/\sum_{k=0}^{L-1}|Y(k,m)|^2$ in (34) is evaluated in terms of the soft weighting function $\alpha_{max}/(1 + (\Xi - 1)^2)$, which we found very suitable for this purpose [11].

To improve the performance of the noise estimator in non-stationary noise environments we found it necessary to also apply a lower limit α_{min} to $\alpha(k,m)$. Since α_{min} limits the rise and decay

times of $\widehat{\lambda}_y(k,m)$, this lower limit is a function of the overall signal-to-noise ratio of the speech sample. To avoid attenuating weak consonants at the end of a word we require $\widehat{\lambda}_y(k,m)$ to decay from its peak values to the noise level in about $\Delta T = 64$ ms. Therefore, α_{min} can be computed as

$$\alpha_{min} = SNR_{LT}^{-\frac{M_E}{\Delta T f_s}} . \tag{35}$$

B. The Minimum Tracking Algorithm

If $\widehat{\lambda}_{min}(k,m)$ denotes the minimum of D consecutive PSD estimates $\widehat{\lambda}_y(k,\ell)$, $\ell = m-D+1,\ldots,m$, an unbiased estimator of the noise power spectral density $\lambda_d(k,m)$ is given by

$$\widehat{\lambda}_d(k,m) = B_{min}(D,Q(k,m))\lambda_{min}(k,m) \tag{36}$$

where the bias compensation factor $B_{min}(D,Q(k,m))$ can be approximated by [11, 23]

$$B_{min}(k,m) \approx 1 + (D-1)\frac{2(1-M(D))}{Q(k,m)-2M(D)} . \tag{37}$$

$M(D)$ is approximated by

$$M(D) = 0.025 + 0.23(1+\log(D)^{0.8}) + 2.7\cdot 10^{-6}D^2 - 1.14\cdot 10^{-3}D - 7\cdot 10^{-2} . \tag{38}$$

The unbiased estimator requires the knowledge of the degrees of freedom $Q(k,m)$ of the smoothed PSD estimate $\widehat{\lambda}_y(k,m)$ at any given time and frequency index. In our context, $Q(k,m)$ can attain non-integer values since the PSD is obtained via recursive smoothing and consecutive signal frames might be correlated. Since the variance of the smoothed power spectral density estimate $\widehat{\lambda}_y(k,m)$ is inversely proportional to $Q(k,m)$, we compute $1/Q(k,m)$ as

$$\frac{1}{Q(k,m)} = \frac{var(\widehat{\lambda}_y(k,m))}{2\lambda_d^2(k,m)} , \tag{39}$$

which then allows us to approximate $B_{min}(D,Q(k,m))$ via (37).

To compute the variance of the smoothed PSD estimate $\widehat{\lambda}_y(k,m)$, we estimate the first and the second moments, $E\{\widehat{\lambda}_y(k,m)\}$ and $E\{\widehat{\lambda}_y^2(k,m)\}$, of $\widehat{\lambda}_y(k,m)$ by means of first order recursive systems,

$$\overline{P}(k,m+1) = \beta(k,m)\overline{P}(k,m) + (1-\beta(k,m))\widehat{\lambda}_y(k,m+1) \tag{40}$$

$$\overline{P^2}(k,m+1) = \beta(k,m)\overline{P^2}(k,m) + (1-\beta(k,m))\widehat{\lambda}_y^2(k,m+1) \tag{41}$$

$$\widehat{var}\{\widehat{\lambda}_y(k,m)\} = \overline{P^2}(k,m) - \overline{P}^2(k,m) . \tag{42}$$

We choose $\beta(k,m) = \alpha^2(k,m)$ and limit $\beta(k,m)$ below 0.8.

Finally, we estimate $1/Q(k,m)$ by

$$\frac{1}{Q(k,m)} \approx \frac{\widehat{var}(\widehat{\lambda}_y(k,m))}{2\widehat{\lambda}_d^2(k,m)} \tag{43}$$

and limit this estimate below 0.5. This limit corresponds to the minimum degrees of freedom, $Q = 2$, which we obtain when no smoothing is in effect ($\alpha(k,m) = 0$). Furthermore, since the error variance of the Minimum Statistics noise estimator is larger than the error variance of an ideal moving average estimator [11], we increase the inverse bias $B_{min}(k,m)$ by a factor $B_c(m) = 1 + a_v\sqrt{\overline{Q^{-1}}(m)}$ with $\overline{Q^{-1}}(m) = \frac{1}{L}\sum_{k=0}^{L-1}\frac{1}{Q(k,m)}$ and a_v typically set to $a_v = 1.5$.

C. Tracking Non-Stationary Noise

The Minimum Statistics method searches for the bias-compensated minimum $\lambda_{min}(k,m)$ of D consecutive PSD estimates $\widehat{\lambda}_y(k,l)$, $l = m-D+1,\ldots,m$. For each frequency bin k, the D samples are selected by sliding a rectangular window over the smoothed power data $\widehat{\lambda}_y(k,l)$. Furthermore, we divide the window of D samples into U sub-windows of V samples each ($UV = D$). This allows us to update the minimum of $\widehat{\lambda}_y(k,m)$ every V samples while keeping the computational complexity low. For every V samples read, we compute the minimum of the current sub-window and store it for later use. We obtain an overall minimum after considering all such sub-window minima. Also, we achieve better tracking of non-stationary noise when we take local minima in the vicinity of the overall minimum $\lambda_{min}(k,m)$ into account. For our purposes, we ignore sub-window minima where the minimum value is attained in the first or the last frame of a sub-window. Since (36) is a function of the window length, computing power estimates on the sub-window level requires a bias compensation for the minima obtained from sub-windows as well (i.e., put $D = V$ in (36)). A local (sub-window) minimum may then override the overall minimum $\lambda_{min}(k,m)$ when it is close to the overall minimum $\lambda_{min}(k,m)$ of the D consecutive power estimates. This procedure uses the spectral minima of the shorter sub-windows for improved tracking. To reduce the likelihood of large estimation errors when using sub-window minima, we apply a threshold *noise_slope_max* to the difference between the sub-window minima and the overall minimum. This threshold depends on the normalized averaged variance $\overline{Q^{-1}}(m)$ of $\widehat{\lambda}_y(k,m)$ according to the procedure shown in Fig. 6. A large update is only possible, when the normalized averaged variance $\overline{Q^{-1}}(m)$ is small and

- computation of $noise\_slope\_max$
  - if $\overline{Q^{-1}}(m) < 0.03$, $noise\_slope\_max = 8$
  - elseif $\overline{Q^{-1}}(m) < 0.05$, $noise\_slope\_max = 4$
  - elseif $\overline{Q^{-1}}(m) < 0.06$, $noise\_slope\_max = 2$
  - else $noise\_slope\_max = 1.2$

Figure 6: Computation of $noise\_slope\_max$.

hence when speech is most likely absent. Thus, we update the noise PSD estimate when a local minimum is found, and when the difference between the sub-window minimum and the overall minimum does not exceed the threshold $noise\_slope\_max$. A pseudocode program of the complete noise estimation algorithm is shown in Fig. 7.

We point out that the tracking of non-stationary noise is significantly influenced by this mechanism and may be improved (at the expense of speech signal distortion) by increasing the $noise\_slope\_max$ threshold. We also note that it is important to use an adaptive smoothing parameter $\alpha(k, m)$ as in (32). Otherwise, for a high SNR and a fixed smoothing parameter close to 1, the estimated signal power will decay too slowly after a period of speech activity. Hence, the minimum search window might then be too small to track the noise floor without being biased by the speech.

Although the Minimum Statistics approach [23, 11] was originally developed for a sampling rate of $f_s = 8000$ Hz and a frame advance of 128 samples, it can be easily adapted to other sampling rates and frame advance schemes. The length D of the minimum search window must be set proportional to the frame rate. For a given sampling rate f_s and frame advance M_E, the duration of the time window for minimum search, $D \cdot M_E/f_s$, should be equal to approximately 1.5 seconds. For $U = 8$ sub-windows we therefore use $V = \lceil 0.1875 f_s/M_E \rceil$, where $\lceil x \rceil$ denotes the smallest integer larger than or equal to x. When a constant smoothing parameter [23] is used in (28) the length D of the window for minimum search must be at least 50% larger than that for the adaptive smoothing algorithm.

- compute smoothing parameter $\alpha(k,m)$, (32)
- compute smoothed power $\widehat{\lambda}_y(k,m)$, (28)
- compute $\overline{Q^{-1}}(m) = \sum_k 1/Q(k,m)$
- compute bias correction $B_{min}(k,m)$ and $B_{min\_sub}(k,m)$, (37, 38, 43), and $B_c(m)$
- set update-flag $k\_mod(k) = 0$ for all k
- if $\widehat{\lambda}_y(k,m)B_{min}(k,m)B_c(m) < actmin(k,m)$
  - $actmin(k,m) = \widehat{\lambda}_y(k,m)B_{min}(k,m)B_c(m)$
  - $actmin\_sub(k,m) = \widehat{\lambda}_y(k,m)B_{min\_sub}(k,m)B_c(m)$
  - set $k\_mod(k) = 1$;
- if $subwc == V$
  - if $k\_mod(k) == 1$
    - $lmin\_flag(k,m) = 0$
  - store $actmin(k,m)$
  - find $\lambda_{min}(k,m)$, the minimum of the last U stored values of $actmin$
  - compute $noise\_slope\_max$
  - if $lmin\_flag(k,m)$ & $(actmin\_sub(k,m) < noise\_slope\_max\lambda_{min}(k,m))$ & $(actmin\_sub(k,m) > \lambda_{min}(k,m))$
    - $\lambda_{min}(k,m) = actmin\_sub(k,m)$
    - replace all previously stored values of $actmin(k,\ell)$ by $actmin\_sub(k,m)$
  - $lmin\_flag(k,m) = 0$;
  - set $subwc = 1$ and $actmin(k,m)$ to its maximum initial value
- else
  - if $subwc > 1$
    - if $k\_mod(k) == 1$
      - set $lmin\_flag(k,m) = 1$
    - compute $\widehat{\lambda}_d(k,m) = \min(actmin\_sub(k,m), \lambda_{min}(k,m))$
    - set $\lambda_{min}(k,m) = \widehat{\lambda}_d(k,m)$
  - set $subwc = subwc + 1$

Figure 7: The minimum statistics noise estimation algorithm [11]. All computations are embedded into loops over all frequency indices k and all frame indices m. Sub-window quantities are subscripted by *sub*. *subwc* is a sub-window counter which is initialized to $subwc = V$ at the start of the program. $actmin(k,m)$ and $actmin\_sub(k,m)$ are the spectral minima of the current window and sub-window up to frame m, respectively.

VII. Experimental Results

The evaluation of noise reduction algorithms using instrumental ("objective") measures is an ongoing research topic [24, 25]. Frequently, quality improvements are evaluated in terms of (segmental) SNR and the achieved noise attenuation. These measures, however, can be misleading as speech signal distortions and unnatural-sounding residual noise are not properly reflected. Also, as long as the reduction of noise power is larger than the reduction of speech power the performance with respect to these metrics may be improved by applying more attenuation to the noisy signal at the expense of speech quality. The basic noise attenuation versus speech distortion tradeoff is application and possibly listener dependent. Even listening tests do not always lead to conclusive results, as was experienced during the standardization process of a noise reduction preprocessor for the ETSI/3GPP AMR coder [26, 27]. Specifically, the outcome of these tests depends on whether an *Absolute Category Rating* (ACR) or a *Comparison Category Rating* (CCR) method is favored.

To capture the possible degradations of both the speech signal *and* the background noise, a multi-faceted approach such as the well-established Diagnostic Acceptability Measure (DAM) is useful. The DAM evaluates a large number of quality characteristics, including the nature of the residual background noise in the enhanced signal. Intelligibility tests are more conclusive and reproducible despite being rarely used. In our investigation, we evaluated intelligibility using the standard Diagnostic Rhyme Test (DRT). For both tests, higher scores are an indication of better quality. More information about the DAM and the DRT may be found in [28].

While preliminary results for a floating point implementation of the preprocessor were presented in [2], we summarize our results here for a 16 bit fixed-point implementation, used in conjunction with the MELP coder. We evaluate quality and intelligibility respectively, all using DAM and DRT scores obtained via formal listening tests. To provide an additional reference, we compare the 2.4 kbps MELP coder using our enhancement preprocessor (denoted in [1] by MELPe) with the toll quality 8 kbps ITU-T coder, G.729a (without a preprocessor). Compared to the results reported for the floating-point implementation [2], the fixed-point implementation scores about 2 points less on both the DAM and the DRT scales. Table 1 presents DAM scores for the MELPe and the G.729a coders without environmental noise. Clearly, the G.729a coder, operating at a much higher rate than the MELP coder, delivers significantly better quality. In the presence of vehicular noise with an average SNR of about 6 dB (Table 2), the MELPe scores significantly higher than the

coder	DAM	S. Error
MELPe	68.6	0.90
G.729a	80.9	1.80

Table 1: DAM scores and standard error without environmental noise.

coder	DAM	S. Error
unprocessed	45.0	1.2
MELP	38.9	1.1
MELPe	50.3	0.80
G.729a	46.3	0.90

Table 2: DAM scores and standard error with vehicular noise (average SNR $\approx$ 6 dB).

standalone MELP coder, the unprocessed signal, and the G.729a coder. Note that, the G.729a achieves approximately the same DAM score as the unprocessed signal.

Tables 3 and 4 show intelligibility results for the clean and noisy conditions. For the clean condition, the higher bit rate G.729a coder is clearly more transparent, but the intelligibility of the MELPe is surprisingly close. This reinforces the frequently made observation that high intelligibility can be achieved with low bit rate coders. For the noisy environment (Table 4) we find that the unprocessed (and unencoded) signal achieves the best intelligibility. The MELPe coder, containing the noise reduction preprocessor, results in a significant intelligibility improvement. These intelligibility improvements are mostly due to the conservative noise estimation algorithm which is unbiased for stationary noise but underestimates the noise floor for non-stationary noise [11]. More detailed results for different noise environments may be found in [29].

coder	DRT	S. Error
MELPe	93.9	0.53
G.729a	94.7	0.25

Table 3: DRT scores and standard error without environmental noise.

coder	DRT	S. Error
unprocessed	91.1	0.37
MELP	67.3	0.8
MELPe	72.5	0.58
G.729a	77.8	0.58

Table 4: DRT scores and standard error with vehicular noise (average SNR $\approx$ 6 dB).

VIII. Conclusion

We have presented a noise reduction preprocessor based on MMSE estimation techniques and the Minimum Statistics noise estimation approach. The combination of these algorithms and the careful selection of parameters lead to a noise reduction preprocessor that achieves improvements both in quality *and* intelligibility when used with the 2.4 kbps MELP coder. Thus, in the context of low bit rate coding, single microphone enhancement algorithms can result in intelligibility improvements. The loss of intelligibility is not as severe for high bit rate coders as for low bit rate coders like the MELP coder.

We believe that the potential for further improving speech transmission in noisy conditions has not yet been fully exploited. Further improvements might be obtained by using optimal enhancement algorithms for the various parameters found in speech coders, such as the LPC coefficients, the pitch, and the representation of the prediction residual signal. Such an approach is proposed in [30]. Novel noise PSD and *a priori* SNR estimation procedures [14, 15], as well as more realistic assumptions for the probability density functions of the speech and noise spectral coefficients [31, 32], could also lead to improved performance.

Acknowledgements

This work was generously supported by AT&T Labs Research and the U.S. government. The authors would like to thank Dr. John Collura for many stimulating discussions and for providing the subjective listening test results.

References

[1] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J. Collura, "A 1200/2400 BPS Coding Suite Based on MELP," in *IEEE Workshop on Speech Coding*, pp. 90–92, 2002.

[2] J. Collura, "Speech Enhancement and Coding in Harsh Acoustic Noise Environments," in *IEEE Workshop on Speech Coding*, pp. 162–164, 1999.

[3] S. Lim and A. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, 1979.

[4] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, pp. 137–145, December 1980.

[5] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, December 1984.

[6] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, April 1985.

[7] J. Yang, "Frequency Domain Noise Suppression Approaches in Mobile Telephone Systems," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 363–366, 1993.

[8] P. Scalart and J. Vieira Filho, "Speech Enhancement Based on *A Priori* Signal to Noise Estimation," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 629–632, 1996.

[9] D. Malah, R. Cox, and A. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 2, pp. 789–792, 1999.

[10] J. Thyssen, Y. Gao, A. Benyassine, E. Shlomot, C. Murgia, H.-Y. Su, K. Mano, Y. Hiwasaki, H. Ehara, K. Yasunaga, C. Lamblin, B. Kovesi, J. Stegmann, and H.-G. Kang, "A Candidate

for the ITU-T 4 KBIT/S Speech Coding Standard," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 2, pp. 681–684, 2001.

[11] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 504–512, July 2001.

[12] R. Martin and R. Cox, "New Speech Enhancement Techniques for Low Bit Rate Speech Coding," in *Proc. IEEE Workshop on Speech Coding*, (Porvoo,Finland), pp. 165–167, 1999.

[13] D. Griffin and J. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 236–243, Apr. 1984.

[14] I. Cohen and B. Berdugo, "Speech Enhancement for Non-stationary Noise Environments," *Signal Processing, Elsevier*, vol. 81, pp. 2403–2418, 2001.

[15] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *Signal Proc. Letters*, vol. 9, pp. 112–116, 2002.

[16] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, April 1994.

[17] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 208–211, 1979.

[18] T. Ferguson, *Mathematical Statistics - A Decision Theoretic Approach*. Academic Press, 1967.

[19] J. A. Rice, *Mathematical Statistics and Data Analysis*. Duxbury Press, Wadworth Pub. Co., 1995.

[20] J. Sohn and W. Sung, "A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, pp. 365–368, 1998.

[21] E. Ekudden, R. Hagen, I. Johansson, and J. Svedberg, "The Adaptive Multi-Rate Speech Coder," in *IEEE Workshop on Speech Coding*, 1999.

[22] R. Martin, J. Wittke, and P. Jax, "Optimized Estimation of Spectral Parameters for the Coding of Noisy Speech," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 3, pp. 1479–1482, 2000.

[23] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, pp. 1182–1185, 1994.

[24] E. Paajanen and V.-V. Mattila, "Improved Objective Measures for Characterization of Noise Suppression Algorithms," in *IEEE Workshop on Speech Coding*, 2002.

[25] P. Dreiseitel, "Hybrid Quality Measures for Single-Channel Speech Enhancement Algorithms," *European Trans. Telecommunications*, vol. 13, no. 2, pp. 159–165, 2002.

[26] ETSI, *TS 122 076, V5.0.0: Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Noise Suppression for the AMR Codec; Service Description*, June 2002.

[27] ETSI, *TR 126 978, V4.0.0: Universal Mobile Telecommunications System (UMTS); Results of the AMR noise suppression selection phase (3GPP TR 26.978 version 4.0.0 Release 4)*, March 2001.

[28] S. Quackenbush, T. Barnwell III, and M. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.

[29] M. Street, "STANAG 4591 Results," in *NC3A Workshop on STANAG 4591*, 2002.

[30] A. Accardi and R. Cox, "A Modular Approach to Speech Enhancement with an Application to Speech Coding," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, pp. 201–204, 1999.

[31] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, pp. 253–256, 2002.

[32] R. Martin, "Speech Enhancement based on Minimum Mean Square Error Estimation and Supergaussian Priors," *IEEE Trans. Speech and Audio Processing*, 2003 (accepted).

List of Figures