Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 96384, 9 pages
doi:10.1155/2007/96384

Research Article

Power Spectral Density Error Analysis of Spectral Subtraction Type of Speech Enhancement Methods

Peter Händel

Signal Processing Lab, School of Electrical Engineering, Royal Institute of Technology, SE-100 44 Stockholm, Sweden

Received 8 August 2005; Revised 28 April 2006; Accepted 16 July 2006

Recommended by Richard Heusdens

A theoretical framework for analysis of speech enhancement algorithms is introduced for performance assessment of spectral subtraction type of methods. The quality of the enhanced speech is related to physical quantities of the speech and noise (such as stationarity time and spectral flatness), as well as to design variables of the noise suppressor. The derived theoretical results are compared with the outcome of subjective listening tests as well as successful design strategies, performed by independent research groups.

1. INTRODUCTION

The human speech generates complex acoustic waves that sometimes are aimed at a nearby listener, and sometimes are aimed for being transmitted by technical systems such as radio broadcasting, fixed or wireless telephony, or services based on the Internet Protocol. A fundamental feature of digital transmission schemes is that the recorded speech samples are coded (compressed) in order to loosen the bandwidth requirement of the transmission channel. The compression, or speech coding, is typically model-based and optimized for compression of speech signals. Since the encoder is designed for speech, it is not suitable for compression of other sources such as environmental noise or music. Accordingly, reduction of noise from received noise-contaminated speech samples is a problem of great importance.

Digital noise reduction schemes are considered in several cellular systems. Early work includes the scheme employing a Kalman filter standardized for the pacific digital cellular system [1]. In the enhanced variable rate codec (EVRC) for CDMA mobile telephony systems, a frequency-domain noise reduction system is included [2]. In these telephony applications, it is important that the algorithms produce enhanced speech with marginal distortion.

Based on experimental studies and listener tests, several research groups have independently reported improvement in signal-to-noise ratio (SNR) of order 10 dB, without introducing audible artifacts and distortion. Yang reported 9 dB SNR improvement for a frequency-domain noise reduction algorithm [3], Gibson et al. reported figures near 7 dB [4], while Sörqvist et al. reported a figure of 10 dB [5]. The latter methods employ time-domain Kalman filters. One should notice the different SNR measures and speech material used in the cited works, and thus a direct comparison of SNR figures is not suitable.

Theoretical limits for speech enhancement were studied in [6], where it was shown that typical spectral subtraction methods are able to reduce the background noise by 10–20 dB during speech pauses. In [7], it was argued that 10 dB noise reduction can be achieved during speaker activity. Thus, the outcome of experimental studies (subjective listening tests) by independent research groups seems to be in agreement with the results predicted by mathematical modelling.

In this paper, we investigate predicted performance of noise reduction algorithms in terms of spectral subtraction. Spectral subtraction type of noise reduction can be found in a variety of applications, such as military voice communications [8], restoration of musical recordings [9, 10], speech recognition [11, 12], and mobile telephony [3, 5].

In this theoretical work, we will concentrate on the design of noise reduction algorithms for "high-quality" (HQ) speech enhancement, defined by (i) a nondistorted speech output, (ii) a sufficient reduction of the noise level, and (iii) a residual noise without annoying artifacts. The basic spectral subtraction methods are known to violate (i) above when

(ii) is fulfilled, or *vice versa*. In addition, in some cases (iii) is more or less violated since the methods may introduce, so-called, musical noise. The above drawbacks with the spectral subtraction methods have been known and, in the literature, several ad hoc modifications of the basic algorithms have appeared. However, the fundamental question how to design spectral subtraction methods that fulfill (i)–(iii) for general scenarios has remained unanswered. The reason for this is, of course, that the requirements (i)–(iii) are in conflict with each other. The correct question to be answered is what reduction of the noise level we can expect without distortion of the speech output and annoying artifacts in the residual noise.

We stress that HQ design of speech enhancement algorithms is not to be used in all applications mentioned above where other subjective criteria such as "crisp-and-clear" are favorable. HQ design has, for example, gained industrial impact in mobile telephony, in applications such as telephony in noisy acoustic cavities (car compartments, etc.).

The aim of this paper is to study speech enhancement by spectral subtraction in a theoretical framework. Clearly, simple models and basic mathematics cannot fully describe the complexity of the human hearing, and thus subjective listening tests are crucial for development of practical speech enhancement methods. Such development of practical speech enhancement methods is beyond the scope of this paper. Here, we concentrate on an understanding of the basic principles of speech enhancement from a statistical signal processing point of view.

Even though the objective criterion used in this paper only partially correlates with subjective human criteria, we will illustrate that the parameters (e.g., the subtraction factor) of the tuned methods coincide with the parameters values based on tuning by subjective tests.

2. NOISE REDUCTION BY LINEAR FILTERING

Consider the signal *model*

$$y(n) = x(n) + v(n), \tag{1}$$

where $y(n)$ denotes the observed discrete-time process, $x(n)$ models the speech, and $v(n)$ the additive noise. The index n is a running integer index, $n = \cdots, -1, 0, 1, \ldots$. The stochastic processes $x(n)$ and $v(n)$ are assumed wide-sense stationary, zero mean, and jointly uncorrelated. The autocorrelation function $r_Y(k) = E[y(n+k)y(n)]$ (where $E[\cdot]$ denotes statistical expectation) is then given by

$$r_Y(k) = r_X(k) + r_V(k), \tag{2}$$

where $r_X(k)$ and $r_V(k)$ denote the autocorrelation functions of $x(n)$ and $v(n)$, respectively, and k is an integer. The power spectral densities follow by taking the (time discrete) Fourier transform, that is,

$$R_Y(\nu) = R_X(\nu) + R_V(\nu). \tag{3}$$

In (3), ν denotes the normalized frequency, that is, $\nu = f/f_s$ with f being the absolute frequency in s^{-1} and f_s being the sampling frequency. When needed we will denote the discrete Fourier transform by $\mathcal{F}[\cdot]$, for example $R_X(\nu) = \mathcal{F}[r_X(k)]$.

Now consider a linear time-invariant filter with frequency function $H(\nu)$. Then, the filtered observation (say, $\hat{x}(n)$) obeys

$$\hat{x}(n) = h(n) \star y(n), \tag{4}$$

where $h(n)$ is the pulse response, or the inverse Fourier transform of $H(\nu)$, and $\star$ denotes the convolution sum. The wide-sense stationary process $\hat{x}(n)$ models the enhanced speech signal.

There are different approaches for designing the filter $H(\nu)$ so that the filter output $\hat{x}(n)$ is a suitable estimate of $x(n)$. For the sake of completeness, two examples of speech enhancement filter designs are outlined below.

2.1. Power subtraction

One attempt to design a filter is known as *power subtraction* [8]. Consider the power spectral density $R_Y(\nu)$ for which an estimate is easily obtained from recorded samples $\{y(n)\}$. With knowledge about $R_V(\nu)$, an estimate of $R_X(\nu)$ (say $\hat{R}_X(\nu)$) is simply obtained by subtraction, that is, $\hat{R}_X(\nu) = \hat{R}_Y(\nu) - R_V(\nu)$, where $\hat{R}_Y(\nu)$ denotes the estimate formed by $\{y(n)\}$. Notice that we have no information of the phase of the signal when considering the power spectral density. However, as shown in [13] estimating the correct power spectral density is more important than trying to estimate the undistorted phase spectrum. Now, the basic power subtraction is given by the following.

(i) Collect the data samples $\{y(1)\cdots y(N)\}$ for a suitable value of N and calculate the discrete Fourier transform, that is, $Y(\nu) = |Y(\nu)| \cdot \exp[j\angle Y(\nu)]$.

(ii) Calculate an estimate of $N \cdot R_Y(\nu)$ by $|Y(\nu)|^2$, and subtract the (N times) power spectral density of the background noise from the obtained quantity. Denote the result by $|Z(\nu)|^2$, that is, $|Z(\nu)|^2 = |Y(\nu)|^2 - NR_V(\nu)$.

(iii) Now the Fourier transform of the enhanced speech is obtained as $\hat{X}(\nu) = |Z(\nu)| \cdot \exp[j\angle Y(\nu)]$, that is, the magnitude corresponding to the noise subtracted spectral density and the original phase of the collected samples. An inverse transformation back to the time domain gives the final result.

The resulting filter is linear and can be described by a pulse function $h(n)$ or equivalently a frequency function $H(\nu)$. Particulary, the frequency function $H(\nu)$ is zero phase or real-valued and is given by

$$\hat{X}(\nu) = H(\nu)Y(\nu) \Longrightarrow H(\nu) = \frac{|Z(\nu)| \cdot \exp[j\angle Y(\nu)]}{Y(\nu)}. \tag{5}$$

Replacing estimated quantities with the true (but unknown) counterparts, a straightforward calculation results in

$$H(\nu) = \sqrt{\frac{R_Y(\nu) - R_V(\nu)}{R_Y(\nu)}} = \sqrt{1 - \frac{R_V(\nu)}{R_Y(\nu)}}. \tag{6}$$

Equation (6) yields the frequency function of power subtraction.

2.2. *Wiener filter solution*

An alternative, well-known, filter design methodology is designing a filter $H(\nu)$ that minimizes the mean-square error $E[(\hat{x}(n) - x(n))^2]$, (cf. [14]). Different constraints on the pulse response such as finite number of nonzero coefficients (finite pulse response) or on causality are often used. Here, however we consider the case without constraints. The well-known solution is given by [14]:

$$H(\nu) = \frac{R_{XY}(\nu)}{R_Y(\nu)}, \tag{7}$$

where $R_{XY}(\nu)$ denotes the cross-spectral density between the observations $y(n)$ and the source process $x(n)$. The Wiener filter is the optimal estimator (in mean-square error) for Gaussian signals, and the optimal linear estimator in general. Under the additive signal model (1) where $x(n)$ and $v(n)$ are wide-sense stationary zero mean and jointly uncorrelated, it follows that the cross-correlation $E[X(n+k)y(n)]$ equals the autocorrelation $E[X(n+k)x(n)]$. Accordingly $R_{XY}(\nu) = R_X(\nu)$. Inserting this latter finding into (7) yields

$$H(\nu) = \frac{R_X(\nu)}{R_Y(\nu)} = 1 - \frac{R_V(\nu)}{R_Y(\nu)}. \tag{8}$$

One may note the structural similarity between (6) and (8).

The frequency function $H(\nu)$ depends on $R_Y(\nu)$ and $R_V(\nu)$, often as well as on user-chosen design variables. Typically, as illustrated by the two examples above, it is zero phase. The frequency function, or suppression rule, $H(\nu)$ can be derived from different error criteria, or it can be motivated from perceptual considerations. Different suppression rules are found in the literature on the topic, where power subtraction and Wiener filtering are two exemplary choices.

In order to avoid notational complications, we concentrate on the power subtraction suppression rule (6) and variants thereof [15]. We stress, however, that the introduced methodology for performance assessment of spectral subtraction speech enhancement algorithms can be applied to any frequency function $H(\nu)$ of $R_V(\nu)$ and $R_Y(\nu)$, including generalized methods [16, 17].

3. SPEECH ENHANCEMENT IN PRACTICE

Speech as well as the background noise are nonstationary processes, and in practice the digital signal processing is based on sample frames of a fixed length N. Accordingly, we consider the scenario in Figure 1 for the presented theoretical analysis, where τ noise-only frames containing N samples are followed by a frame including speech samples. It is further assumed that an ideal voice activity detector (VAD) is available in order to distinguish between the frames containing noisy speech and frames containing background noise only. In particular, we consider the output from the spectral subtraction method in frame $\ell + 1$.

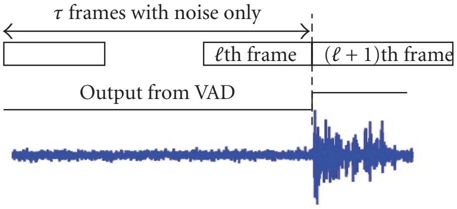

FIGURE 1: Setup for performance analysis of power subtraction. Each data frame is of length N, where N is assumed to (roughly) coincide with the short-time stationarity of the speech. The background noise is assumed long-time stationary over $\tau + 1$ frames. A perfect voice activity detector (VAD) is assumed, where its output speech presence probability (0% or 100%, resp.) is indicated in the figure.

In general, $R_Y(\nu)$ and $R_V(\nu)$ in the suppression rule $H(\nu)$ are unknown and have to be replaced by estimated quantities $\widehat{R}_Y(\nu)$ and $\widehat{R}_V(\nu)^\ell$. Due to the short-time stationarity of the speech, the estimate $\widehat{R}_Y(\nu)$ has to be calculated from the $(\ell+1)$th frame of noisy observations only, while $\widehat{R}_V(\nu)^\ell$ can (and should) be averaged from several past frames.

The actual spectral subtraction can be performed as follows. The digital audio samples $\{y(n)\}$ (for $n = 1, \ldots, N$) in frame $\ell + 1$ are transformed to form $\{Y(\nu)\}$. The frame is filtered by multiplication in the frequency-domain, that is, $\{\widehat{H}(\nu)Y(\nu)\}$ (which is the actual spectral subtraction). The suppression rule $\widehat{H}(\nu)$ is formed, for example, according to (6) with $R_Y(\nu)$ and $R_V(\nu)$ there replaced by estimated quantities, namely

$$\widehat{H}(\nu) = \sqrt{1 - \delta(\nu)\frac{\widehat{R}_V(\nu)^\ell}{\widehat{R}_Y(\nu)}}. \tag{9}$$

In (9), $\delta(\nu)$ is introduced as a user-chosen weighting function. The effects of the choice of $\delta(\nu)$ on the performance of the filtered output will be studied in detail in the sequel. The resulting frequency-domain signal is transformed back to the time domain by an inverse transformation. The result is a frame of samples in which the noise has been suppressed. By definition, $H(\nu)$ belongs to the interval $0 \le H(\nu) \le 1$, which does not necessarily hold true for (9). Historically, half-wave or full-wave rectification is used before the square root is calculated [8]. In other noise reduction systems, the suppression rule (9) is combined with a limiter so that $\widehat{H}(\nu)$ is ensured to be strictly larger than zero. Often a lower threshold or noise floor is used so that $\widehat{H}(\nu)$ always is larger than some threshold value. Such thresholding is known to reduce the so-called musical noise [8]. A minimum value of 0.1 corresponds to a noise floor of −20 dB.

Note that $y(n) = v(n)$ in (1) during speech pauses. Accordingly, an estimate of $R_V(\nu)$ is calculated using a running estimate, for example, using

$$\widehat{R}_V(\nu)^\ell = \rho\widehat{R}_V(\nu)^{\ell-1} + (1-\rho)\overline{R}_V(\nu). \tag{10}$$

In (10), $\widehat{R}_V(\nu)^\ell$ is the running averaged estimate of the spectral density based on data up to and including frame number

ℓ and $\overline{R}_V(\nu)$ is the estimate based on the ℓth frame. The scalar ρ is tuned based on the assumed stationarity of the background noise. An average over τ frames, that corresponds to the long-time stationarity of the background noise, roughly corresponds to ρ given by

$$\rho = 1 - \frac{2}{\tau}. \tag{11}$$

A solid proof of (11) in a somewhat different context can be found in [18].

With no prior assumptions on the spectral shape of the background noise, a suitable N-frame estimate of the power spectral density is given by

$$\overline{R}_V(\nu) = \frac{|V(\nu)|^2}{N}, \tag{12}$$

where $V(\nu) = \mathcal{F}[v(n)]$. With $\mathcal{F}[\cdot]$ being the discrete Fourier transform, $\overline{R}_V(\nu)$ is the periodogram spectral estimator and $\widehat{R}_V(\nu)^\ell$ is an exponential averaged periodogram. Both $\overline{R}_V(\nu)$ and $\widehat{R}_V(\nu)^\ell$ are leading to asymptotically unbiased estimates of $R_V(\nu)$, that is, (as $N \to \infty$) [19]

$$E[\overline{R}_V(\nu)] = R_V(\nu), \qquad E[\widehat{R}_V(\nu)^\ell] = R_V(\nu). \tag{13}$$

The asymptotic error variance of $\overline{R}_V(\nu)$ is

$$\operatorname{Var}[\overline{R}_V(\nu)] = R_V^2(\nu). \tag{14}$$

The variance term in (14) describes the accuracy of the estimate $\overline{R}_V(\nu)$. We introduce a (possibly frequency-dependent) quality factor $\gamma(\nu)$, so that for a general asymptotically unbiased spectral estimator $\widehat{R}(\nu)$ of $R(\nu)$, we have

$$\operatorname{Var}[\widehat{R}(\nu)] = \gamma(\nu)R^2(\nu). \tag{15}$$

That is, for $\overline{R}_V(\nu)$ in (12) we have $\gamma_V(\nu) = \gamma_V = 1$. Accordingly for $\widehat{R}_V(\nu)^\ell$ taken as an average of $\overline{R}_V(\nu)$ over τ frames, we obtain a reduced value proportional to $1/\tau$ (cf. [20]).

4. POWER SPECTRAL DENSITY ERROR ANALYSIS

It is obvious that the stationarity assumptions imposed on the speech as well as the background noise give rise to bounds on how accurate the estimate of the speech is in comparison with the clean speech. In this section, the analysis technique for spectral subtraction methods earlier introduced by the author in [7] is extended. It is based on first-order approximations of the power spectral density (PSD) estimates $\widehat{R}_Y(\nu)$ and $\widehat{R}_V(\nu)^\ell$, respectively, in combination with approximate (zero-order approximations) expressions for the accuracy of the introduced deviations. Explicitly, in the following an expression is derived for the frequency-domain error of the enhanced speech (the filtered output), due to the used suppression rule and due to the accuracy of the involved PSD estimators.

We consider the PSD error [7]

$$\widetilde{R}_X(\nu) = R_{\widehat{X}}(\nu) - R_X(\nu), \tag{16}$$

where $R_{\widehat{X}}(\nu)$ is the spectral density of the enhanced speech, given by

$$R_{\widehat{X}}(\nu) = |\widehat{H}(\nu)|^2 R_Y(\nu) = \widehat{H}(\nu)^2 R_Y(\nu). \tag{17}$$

A similar frequency-domain error is used as an error criterion in [16] when the parameters of a generalized spectral subtraction method are derived. Note that $\widetilde{R}_X(\nu)$ by construction is an error term describing the difference in the frequency-domain between the magnitude-squared filtered noisy observation and the power spectral density of the clean speech. Therefor, $\widetilde{R}_X(\nu)$ can take both positive and negative values and is not the power spectral density of any time-domain signal. In (17), the suppression rule $\widehat{H}(\nu)$ is not restricted to power subtraction, but other choices can be analyzed as well. In particular, one can note that a comparison between magnitude subtraction and power subtraction was performed in [7].

In order to perform the analysis, we assume a small error so that $\widehat{R}_Y(\nu)$ and $\widehat{R}_V(\nu)^\ell$ used to form $\widehat{H}(\nu)$ are close to the underlying spectral densities. Technically, it is required that consistent spectral estimators are employed and that the frame length N is sufficiently large ($N \gg 1$). Introduce the first-order deviations

$$\begin{aligned} \widehat{R}_Y(\nu) &= R_Y(\nu) + \Delta_Y(\nu), \\ \widehat{R}_V(\nu) &= R_V(\nu) + \Delta_V(\nu), \end{aligned} \tag{18}$$

where $\Delta_Y(\nu)$ and $\Delta_V(\nu)$ are zero-mean stochastic variables such that we have the quality factors

$$\gamma_Y(\nu) = \frac{E[\Delta_Y^2(\nu)]}{R_Y^2(\nu)} \ll 1, \qquad \gamma_V(\nu) = \frac{E[\Delta_V^2(\nu)]}{R_V^2(\nu)} \ll 1, \tag{19}$$

for all ν. Variance reduction by introducing bias in the estimates is commonly applied, but using asymptotically biased estimators of the spectral density, a similar analysis holds true replacing (18) with

$$\begin{aligned} \widehat{R}_Y(\nu) &= R_Y(\nu) + \Delta_Y(\nu) + B_Y(\nu), \\ \widehat{R}_V(\nu) &= R_V(\nu) + \Delta_V(\nu) + B_V(\nu), \end{aligned} \tag{20}$$

where $B_Y(\nu)$ and $B_V(\nu)$ are deterministic terms describing the asymptotic bias in the employed estimators. Such analysis is straightforward to perform, but beyond the scope of this paper. Small quality factors (19) are essential for the analysis to hold true, although the forthcoming results indicate a wider applicability of the theory. Large-quality factors may appear for (at a first glance) reasonable settings. For example for a white Gaussian input, the background noise periodogram (12) is known to be chi-square with two degrees of freedom with asymptotically $E[\Delta_V^2(\nu)] = R_V^2(\nu)$, and thus $\gamma_V(\nu) = \gamma_V$ where $\gamma_V = 1$! In practice, exponential smoothing of periodograms according to (10) is employed, and thus the exponential averaged estimate of the power spectral density approaches the Gaussian shape corresponding to $\gamma_V = 1/(\tau - 1)$ [20], where typical settings for τ is 50 or more with τ related to ρ according to (11). Accordingly, the taken approach is easily justified for the background noise spectral estimate. Due to the nonstationarity of speech, interframe smoothing is not appropriate when estimating $R_Y(\nu)$. Accordingly, in order to secure that $\gamma_Y \ll 1$, intraframe approaches have to be employed, for example by model-based approaches that will be studied in the sequel.

Further, the correlation time of the noise is assumed to be short compared to the frame length N, meaning that $E[\{\hat{R}_V(\nu)^\ell - R_V(\nu)\}\{\hat{R}_V(\nu)^k - R_V(\nu)\}] \approx 0$ for $k \neq \ell$. This, in turn, implies that $\Delta_Y(\nu)$ and $\Delta_V(\nu)$ are approximately uncorrelated. This latter assumption on the background noise may seem restrictive. If the noise is strongly correlated, we can, on the other hand, assume that $R_V(\nu)$ has a limited small number n (i.e., $n \ll N$) of (strong) peaks located at frequencies $\nu_1, \dots, \nu_n$. Then, $E[\{\hat{R}_V(\nu)^\ell - R_V(\nu)\}\{\hat{R}_V(\nu)^k - R_V(\nu)\}] \approx 0$ for $\nu \neq \nu_j$, for $j = 1, \dots, n$, and for $k \neq \ell$ and the analysis still holds true for $\nu \neq \nu_j$, $j = 1, \dots, n$.

Starting with the basic suppression rule for power subtraction (6), the PSD error (16)-(17) for known spectral densities makes some sense. That is, inserting (6) into (16)-(17) gives

$$\tilde{R}_X(\nu) = \left(1 - \frac{R_V(\nu)}{R_Y(\nu)}\right)R_Y(\nu) - R_X(\nu) = 0. \qquad (21)$$

Thus, if the spectral densities of the speech and noise are perfectly known, power subtraction is optimal in the sense of minimizing the squared PSD error. Perfectly known PSDs are characterized by zero-valued quality factors. We emphasize that if the PSD error is zero as in (21) does not mean that the restoration is perfect. In order to measure the SNR improvement (in dBs), we make use of

$$\begin{aligned} \text{SNR}_{\text{Improvement}} &= 10\log_{10} E[v(n)^2] \\ &\quad - 10\log_{10} E[(\hat{x}(n) - x(n))^2]. \end{aligned} \qquad (22)$$

For example, if $x(n)$ and $v(n)$ are uncorrelated white noises of equal power, the power subtraction rule results in $\hat{x}(n) = y(n)/\sqrt{2}$ which imply that $\text{SNR}_{\text{Improvement}} = 2.32\,\text{dB}$, although the PSD error is null.

In practice, neither $R_Y(\nu)$ nor $R_V(\nu)$ is known, but are replaced by estimated quantities resulting in positive (nonzero) quality factors. The deviations of the estimated spectral densities from the underlying true ones are described by the stochastic quantities $\Delta_Y(\nu)$ and $\Delta_V(\nu)$ introduced in (18), respectively. For performance optimization, a natural distortion criterion is the averaged mean-squared PSD error

$$\text{MSE} = \int_{-1/2}^{1/2} E[\tilde{R}_X^2(\nu)]d\nu, \qquad (23)$$

where the expectation is over the stochastic quantities $\Delta_Y(\nu)$ and $\Delta_V(\nu)$, respectively.

4.1. *Analysis of power subtraction*

With the above-mentioned tools for performance analysis, the power subtraction suppression rule in (9) can be analyzed. In (9), $\delta(\nu)$ is a possibly frequency-dependent user-chosen design variable. In particular with a constant $\delta > 1$, the method is often referred to as power subtraction with *oversubtraction*. This ad hoc modification significantly decreases the noise level and reduces the audible artifacts. In addition, it significantly distorts the output speech, which makes this modification (more or less) useless for high-quality speech enhancement. This fact is easily seen from (9) when $\delta \gg 1$. Thus for moderate and low speech-to-noise ratios (in the ν-domain), the expression under the root sign is very often negative and the rectifying device will therefore set it to zero (or any other predetermined small value), which in turn implies that only frequency bands where the local signal-to-noise ratio is high appear in the output. Due to the nonlinear rectifying device, the present analysis technique is not directly applicable in this case.

An interesting case is when $\delta(\nu) < 1$, which is seen from the following heuristic discussion. As stated previously, when $R_Y(\nu)$ and $R_V(\nu)$ are exactly known, (9) with $\delta(\nu) \equiv 1$ is optimal in the sense of minimizing the MSE (23). When no (whatsoever) information about $R_Y(\nu)$ and $R_V(\nu)$ can be observed, on the other hand, the best we can do is to let the filter output equals the noisy speech. This case corresponds to the use of (9) with $\delta(\nu) \equiv 0$. Due to the above two extremes, one can expect that when estimates $\hat{R}_Y(\nu)$ and $\hat{R}_V(\nu)$ are used to form the suppression rule, the minimum MSE is obtained for some (possibly frequency-dependent function) $\delta(\nu)$ in the interval $0 \leq \delta(\nu) \leq 1$.

The PSD error for the case $0 \leq \delta(\nu) \leq 1$ is given by

$$\tilde{R}_X(\nu) = \left(1 - \delta(\nu)\frac{\hat{R}_V(\nu)}{\hat{R}_Y(\nu)}\right)R_Y(\nu) - R_X(\nu). \qquad (24)$$

In the appendix, it is shown that (24) can be rewritten as

$$\tilde{R}_X(\nu) = (1 - \delta(\nu))R_V(\nu) + \delta(\nu)\left(\frac{R_V(\nu)}{R_Y(\nu)}\Delta_Y(\nu) - \Delta_V(\nu)\right). \qquad (25)$$

Under the given assumptions, we have

$$\begin{aligned} E[\tilde{R}_X^2(\nu)] &= (1 - \delta(\nu))^2 R_V^2(\nu) \\ &\quad + \delta(\nu)^2\left(\frac{R_V^2(\nu)}{R_Y^2(\nu)}E[\Delta_Y^2(\nu)] + E[\Delta_V^2(\nu)]\right). \end{aligned} \qquad (26)$$

The quantity $E[\tilde{R}_X^2(\nu)]$ is quadratic in $\delta(\nu)$ and can be analytically minimized for all ν. Denoting the optimal function by $\overline{\delta}(\nu)$, the result reads

$$\overline{\delta}(\nu) = \frac{1}{1 + \gamma_Y(\nu) + \gamma_V(\nu)}, \qquad (27)$$

where $\gamma_Y(\nu)$ is the quality factor of the method used for estimating the instantaneous power spectral density in frame $\ell+1$, and $\gamma_V(\nu)$ is the quality factor for $\hat{R}_V(\nu)^\ell$. Since the quality factors are positive quantities, it follows that $0 \leq \overline{\delta}(\nu) \leq 1$. Inserting (27) into (26) yields

$$\begin{aligned} E[\tilde{R}_X^2(\nu)]_{\delta(\nu)=\overline{\delta}(\nu)} &= R_V^2(\nu)\frac{\gamma_Y(\nu) + \gamma_V(\nu)}{1 + \gamma_Y(\nu) + \gamma_V(\nu)} \\ &= R_V^2(\nu)(1 - \overline{\delta}(\nu)). \end{aligned} \qquad (28)$$

In order to influence the performance of power subtraction, we have to decrease the quality factors $\gamma_Y(\nu)$ and $\gamma_V(\nu)$ as much as possible. It means that we have to select an appropriate estimator of the instantaneous spectral density in the $(\ell+1)$th frame, as well as an estimator for the long-time stationary background noise. Further, we can influence the performance by proper selection of the suppression rule, see [16].

Using fast-Fourier-transform-(FFT-) based spectral estimators, we have that $\gamma_Y = 1$ (as earlier discussed, for the analysis to hold true, the quality factors should be much smaller than unity) and γ_V is proportional to $1/\tau$. Thus, for $\tau \gg 1$, the dominant term in $\gamma = \gamma_Y + \gamma_V$ is γ_Y, and thus the main error source is the single-frame spectral estimation of the noisy speech.

We note that in this scenario, $\overline{\delta}(\nu)$ is (at least, approximately) independent of frequency, that is, $\overline{\delta}(\nu) = \overline{\delta}$. We also note that $\overline{\delta}$ is smaller than unity, that is, for $\gamma_Y = 1$ and $\gamma_V = 1/\tau$, then $\overline{\delta} < 0.5$ for all τ. The fact that $\overline{\delta} \ll 1$ indicates that the statistical accuracy of the spectral estimators, and in particular the statistical accuracy of $\widehat{R}_Y(\nu)$, have a large impact on the quality of the output enhanced speech. Here, the MSE introduced in (23) reduces to

$$\text{MSE}\,|_{\delta(\nu)=\overline{\delta}} = (1-\overline{\delta})\int_{-1/2}^{1/2} R_V^2(\nu)d\nu. \tag{29}$$

The value of the quality factor $\gamma_Y(\nu)$ may be decreased by using averaging techniques, such as blocking data into subframes and using an averaged periodogram. Such an approach is included in the IS-127 standard where 16 frequency bands are used [2]. Another appealing approach is to reduce $\gamma_Y(\nu)$ by parametric modelling. This is the topic below.

5. PARAMETRIC MODELLING FOR QUALITY IMPROVEMENTS

The key observation in the above section was that the quality factors have a major impact on the achievable level of noise reduction. Another observation was that proper averaging of the background noise spectral density estimate has high accuracy, that is, a quality factor close to zero. Here, we concentrate on the weakest part of the chain, that is, the estimate of the instantaneous spectral density at the most present data frame.

A standard technique to model speech is to use autoregressive (AR) modelling, that is, $x(n)$ in (1) can be accurately described by an AR model of order p, that is,

$$x(n) = -a_1x(n-1) - \cdots - a_px(n-p) + w(n), \tag{30}$$

where $w(n)$ is white zero-mean noise with power σ_W^2. Typically, p is $p \approx 10$. At a first glance, it may seem restrictive to consider AR models only. However, the use of AR models for speech modelling is not only motivated by physical modelling of the vocal tract, but more importantly by physical limitations from the noisy speech on the accuracy of the estimated models.

With the AR structure (30) imposed on $x(n)$, the spectral density of the noisy observations is

$$R_Y(\nu) = \frac{\sigma_W^2}{|A(e^{j2\pi\nu})|^2} + R_V(\nu), \tag{31}$$

where $A(e^{j2\pi\nu}) = 1 + a_1e^{j\nu} + \cdots + a_pe^{j\nu p}$. For the sake of the discussion, $R_V(\nu)$ may be described by a parametric ARMA model

$$R_V(\nu) = \frac{\sigma_U^2|B(e^{j2\pi\nu})|^2}{|C(e^{j2\pi\nu})|^2}, \tag{32}$$

where $B(e^{j2\pi\nu})$ and $C(e^{j2\pi\nu})$ are qth- and rth-order polynomials, defined similarly to $A(e^{j2\pi\nu})$. Then, the noisy observation $y(n)$ has a spectral density given by

$$R_Y(\nu) = \frac{\sigma_W^2|C(e^{j2\pi\nu})|^2 + \sigma_U^2|A(e^{j2\pi\nu})B(e^{j2\pi\nu})|^2}{|A(e^{j2\pi\nu})C(e^{j2\pi\nu})|^2}. \tag{33}$$

Estimating the speech AR parameters in (30) is straightforward when no additional noise is present. However, estimating the parameters in (33) is a stand-alone research problem, especially in the case above when $R_V(\nu)$ is partially known through $\widehat{R}_V(\nu)^{\ell}$. Here, a more pedestrian approach is taken and a method based on the autocorrelation method is sought. The motivation for this is fourfold, that is, (i) the autocorrelation method is well-known. In particular, the estimated parameters are minimum-phase, ensuring the stability of the resulting filter, (ii) using the Levinson algorithm, the method is easily implemented and has a low numerical complexity, (iii) an optimal procedure includes a nonlinear optimization, explicitly requiring some initialization procedure. The autocorrelation method requires none, and (iv) from a practical point of view, it is favorable if the same estimation procedure can be used for degraded speech and, respectively, the clean speech when it is available. In other words, the estimation method should be independent of the actual scenario of operation.

It is well-known that the ARMA process in (33) can be modelled by an infinite-order AR process. When a finite set of data is available for parameter estimation, the infinite-order AR model has to be truncated. Here the model used is

$$y(n) = -f_1y(n-1) - \cdots - f_{\overline{p}}y(n-\overline{p}) + e(n), \tag{34}$$

where $e(n)$ is the residual noise, with power σ_E^2. An appropriate model order follows from the discussion below. The approximate model (34) is close to the speech in noise process if their spectral densities are approximately equal, that is,

$$\frac{\sigma_W^2|C(e^{j2\pi\nu})|^2 + \sigma_U^2|A(e^{j2\pi\nu})B(e^{j2\pi\nu})|^2}{|A(e^{j2\pi\nu})C(e^{j2\pi\nu})|^2} \approx \frac{\sigma_E^2}{|F(e^{j2\pi\nu})|^2}. \tag{35}$$

Based on the physical modelling of the vocal tract, we have $p \approx 10$. From (35), it is also clear that $\overline{p}$ has to be (much)

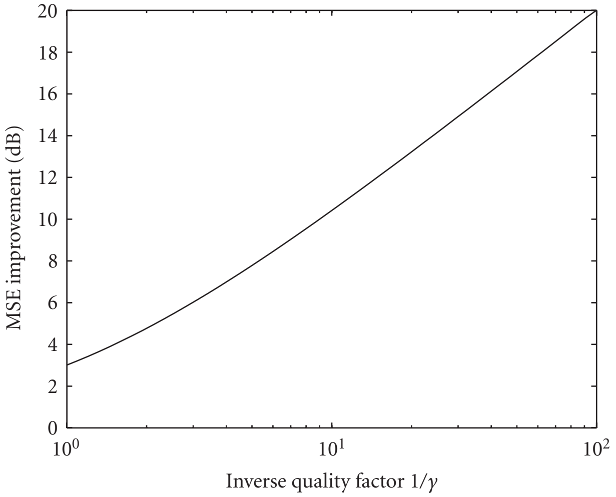

FIGURE 2: MSE improvement as function of quality factor for optimized power subtraction.

greater than the order of the denominator, that is, $\overline{p} \gg p + r$ where $p + r$ roughly equals twice the number of peaks in $R_Y(\nu)$. On the other hand, modelling noisy narrowband processes using AR modelling requires $\overline{p} \ll N$ in order to ensure reliable estimates of the spectral density and a rule of thumb is given by $\overline{p} \approx \sqrt{N}$ [21].

From the above discussion, we can expect that a parametric model approach based on AR modelling is fruitful when $N \gg 100$. We can also conclude from (33) that the flatter the noise spectra is, the smaller the values of N are allowed. Even if $\overline{p}$ is not large enough, the parametric approach is expected to give reasonable results. A reason for this is that the parametric approach produces smooth estimates of the spectral densities, which reduce artifacts such as audible residual noise. The quality of parametric spectral estimates based on all-pole modelling is investigated below.

An attempt to analyze the quality of the parametric estimator in the above section is as follows. Decompose $y(n)$ into its spectral components by aid of a Fourier series expansion of $x(n)$ and assume that the noise is spectrally flat. Then, the asymptotic (both in the frame length and in the model order) variance of $\widehat{R}_Y(\nu)$ is given by the quality factor $\gamma_Y = 2\overline{p}/N$ [22], an expression that also holds true for a pure high-order AR process. With the given rule of thumb, we have $\gamma_Y = 2/\sqrt{N}$ that should be compared with $\gamma_Y = 1$ for the periodogram-based spectral estimator.

6. DISCUSSION

The improvement in MSE for the optimized power subtraction is given by

$$-10\log_{10}(1-\overline{\delta}) = -10\log_{10}\left(\frac{\gamma}{1+\gamma}\right), \tag{36}$$

where γ is the total quality factor $\gamma = \gamma_Y + \gamma_V$. In Figure 2, the improvement in MSE is depicted as function of $1/\gamma$.

Below, the outcome of the presented analysis is compared with independent work presented in the literature. The performance measure employed in the current work is not related in a simple way to well-known objective performance measures used in this field, such as the segmental SNR, spectral distortion, log-likelihood ratio. Accordingly, the below comparison with the outcome of experimental studies is illustrative rather than conclusive.

6.1. The SS method of [5]

In a mobile telephony hands-free environment, it is reasonable to assume that the background noise is stationary for about 0.5 second (at 8 kHz sampling rate and frame length of $N = 256$) that gives $\tau = 15$, and thus $\gamma_V = 0.067$. This is the settings used in the SS (i.e., spectral subtraction) method of [5]. Further, with an AR model of order $\overline{p} = \sqrt{256} = 16$, we have $\gamma_Y = 0.062$ resulting in an improvement in MSE of 9.5 dB. For a completely stationary background noise (i.e., $\tau \to \infty$), the improvement is limited by 12 dB.

6.2. Results from subjective listening tests [23]

In [23], an empirical quantity, the averaged spectral distortion improvement, was experimentally studied with respect to a scalar subtraction factor for magnitude subtraction. Based on several experiments, Kushner et al. conclude that the optimal subtraction factor preferably should be in the interval that spans from 0.6 to 0.7 at a signal-to-noise ratio of 15 dB. In [23], a three-second averaging at 10 kHz sampling rate is reported, $N = 512$ samples and 50% overlap. Accordingly, $\gamma_Y = 0.5$ (50% frame overlap). The averaging is over 58 frames (with overlap), so $\gamma_V = 0.009$. Inserting the numerical values into (27) results in the subtraction factor $\overline{\delta} = 0.66$. The outcome of the theoretical small error analysis presented here is in good agreement with the outcome based on subjective listening test in [23].

6.3. Relations to standard IS-127

A state-of-the-art noise suppressor is the one included in IS-127, developed by Motorola [2]. In the TIA/EIA standard IS-127 for the EVRC speech codec, $N = 104$ noisy speech samples are used (the transformation $\mathcal{F}[\cdot]$ is performed by aid of a 128-point FFT). Further, the suppression rule is calculated in 16 different bands, and the transformation back to the time domain is performed with some overlap between adjacent frames. A rough estimate of the quality factor γ_Y is as follows. The decomposition into subbands reduces γ_Y by a factor of 16, whereas the averaging between frames reduces it further by a factor of 1.5 (an overlap-and-add with 48 samples is used), that is, $\gamma_Y = 0.04$. This figure results in an MSE improvement (36) up to 14 dB.

We cannot figure out a suitable value for γ_V from [2]. However, in [2] the noise floor is reported to be −13 dB. Assuming a smooth transition from frequency function to noise floor for power subtraction (6), the noise floor is given by

$$-20\log\sqrt{1-\overline{\delta}} = -13. \tag{37}$$

Solving for γ_V (with $\gamma_Y = 0.04$ and $\overline{\delta}$ given by (27)) results in $\gamma_V = 0.013$. Here, $\gamma_Y \approx 3\gamma_V$, which seems to be a reasonable tradeoff between the short-time spectral estimation and noise averaging.

6.4. Conclusions

In order to conclude, an analysis technique for spectral subtraction type of methods has been discussed. The power spectral density error was introduced, and its mean square error was optimized for power subtraction. We have compared the theoretical findings with independent work and we have noted an agreement between our predictions and the reported work. For example, we may note the slightly different design strategies employed in [2, 5], where in former work employed $\gamma_Y \approx \gamma_V$ for maximal adaption for changes in the noise characteristics, whereas the latter work employed $\gamma_Y \approx 3\gamma_V$ for increased noise suppression. In both cases, intraframe averaging is used to ensure accurate short-time spectral estimates.

APPENDIX

A. PROOF OF (25)

In order to prove (25), note that inserting (18) into (24) gives

$$\widetilde{R}_X(\nu) = \left(1-\delta(\nu)\frac{R_V(\nu)+\Delta_V(\nu)}{R_Y(\nu)+\Delta_Y(\nu)}\right)R_Y(\nu) - R_X(\nu). \tag{A.1}$$

By using the Taylor series expansion $(1+x)^{-1} = 1 - x + \cdots$ and neglecting higher than first-order deviations, we have

$$\left(R_Y(\nu)+\Delta_Y(\nu)\right)^{-1} = \frac{1}{R_Y(\nu)}\left(1-\frac{\Delta_Y(\nu)}{R_Y(\nu)}\right). \tag{A.2}$$

Inserting (A.2) into (A.1) yields

$$\begin{aligned}\widetilde{R}_X(\nu) &= R_Y(\nu) - \delta(\nu)\left(R_V(\nu)+\Delta_V(\nu)\right)\left(1-\frac{\Delta_Y(\nu)}{R_Y(\nu)}\right) - R_X(\nu)\\ &= R_V(\nu) - \delta(\nu)\left(R_V(\nu)+\Delta_V(\nu)\right)\left(1-\frac{\Delta_Y(\nu)}{R_Y(\nu)}\right).\end{aligned} \tag{A.3}$$

A rearrangement of terms and neglecting the second-order term $-\delta(\nu)\Delta_V(\nu)\Delta_Y(\nu)/R_Y(\nu)$ result in (25).

REFERENCES

[1] T. Ohya, H. Suda, and T. Miki, “5.6 kbits/s PSI-CELP of the half-rate PDC speech coding standard,” in *IEEE 44th Vehicular Technology Conference*, vol. 3, pp. 1680–1684, Stockholm, Sweden, June 1994.

[2] T. V. Ramabadran, J. P. Ashley, and M. J. McLaughlin, “Background noise suppression for speech enhancement and coding,” in *Proceedings of the IEEE Workshop on Speech Coding for Telecommunications Proceeding*, pp. 43–44, Pocono Manor, Pa, USA, September 1997.

[3] J. Yang, “Frequency domain noise suppression approaches in mobile telephone systems,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’93)*, vol. 2, pp. 363–366, Minneapolis, Minn, USA, April 1993.

[4] J. D. Gibson, B. Koo, and S. D. Gray, “Filtering of colored noise for speech enhancement and coding,” *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, 1991.

[5] P. Sörqvist, P. Händel, and B. Ottersten, “Kalman filtering for low distortion speech enhancement in mobile communication,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’97)*, vol. 2, pp. 1219–1222, Munich, Germany, April 1997.

[6] P. Vary, “Noise suppression by spectral magnitude estimation—mechanism and theoretical limits,” *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.

[7] P. Händel, “Low-distortion spectral subtraction for speech enhancement,” in *Proceedings of the 4th European Conference on Speech Communication and Technology*, vol. 2, pp. 1549–1552, Madrid, Spain, September 1995.

[8] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[9] O. Cappe, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[10] O. Cappe and J. Laroche, “Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 84–93, 1995.

[11] J. H. L. Hansen and M. A. Clements, “Constrained iterative speech enhancement with application to speech recognition,” *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 795–805, 1991.

[12] P. Lockwood, J. Boudy, and M. Blanchet, “Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments,” in *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP ’92)*, vol. I, pp. 265–268, San Francisco, Calif, USA, March 1992.

[13] D. L. Wang and J. S. Lim, “The unimportance of phase in speech enhancement,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[14] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*, Principles of Electrical Engineering Series, MIT Press, Cambridge, Mass, USA, 1949.

[15] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[16] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, “A parametric formulation of the generalized spectral subtraction method,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–336, 1998.

[17] P. Sovka, P. Pollak, and J. Kybic, “Extended spectral subtraction,” in *Proceedings of the 5th European Conference on Speech Communication and Technology*, pp. 963–966, Trieste, Italy, September 1995.

[18] P. Stoica and A. Nehorai, "On the asymptotic distribution of exponentially weighted prediction error estimators," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 1, pp. 136–139, 1988.

[19] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Prentice Hall, Upper Saddle River, NJ, USA, 1997.

[20] T. T. J. M. Peeters and Ö. Ciftcioglu, "Statistics on exponential averaging of periodograms," *IEEE Transactions on Signal Processing*, vol. 43, no. 7, pp. 1631–1636, 1995.

[21] J. Angeby, P. Stoica, and T. Söderström, "Asymptotic statistical analysis of autoregressive frequency estimates," *Signal Processing*, vol. 39, no. 3, pp. 277–292, 1994.

[22] P. Händel, P. Stoica, and T. Söderström, "Asymptotic variance of the AR spectral estimator for noisy sinusoidal data," *Signal Processing*, vol. 35, no. 2, pp. 131–139, 1994.

[23] W. M. Kushner, V. Goncharoff, C. Wu, V. Nguyen, and J. N. Damoulakis, "The effects of subtractive-type speech enhancement/noise reduction algorithms on parameter estimation for improved recognition and coding in high noise environments," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '89)*, vol. 1, pp. 211–214, Glasgow, Scotland, May 1989.

Peter Händel received the M.S. degree in engineering physics and the Ph.D. degree in automatic control, all from the Department of Technology, Uppsala University, Uppsala, Sweden, in 1987 and 1993, respectively. During 1987–1988, he was at The Svedberg Laboratory, Uppsala University. Between 1988 and 1993, he was a Teaching and Research Assistant at the Systems and Control Group, Uppsala University. During 1993–1997, he was with Ericsson Radio Systems AB, Kista, Sweden. During the academic year 1996/1997, he was a Visiting Scholar at the Signal Processing Laboratory, Tampere University of Technology, Finland. Since August 1997, he has been with the School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden, where he currently is Professor in signal processing. He has conducted research in a wide area including design and analysis of digital and adaptive filters, measurement and estimation theory, system identification, speech processing. Recent research interests include characterization and modelling of analog-to-digital converters and power amplifiers as well as signal processing for navigation. He is a Member of the editorial board of the EURASIP Journal on Advances in Signal Processing. He is Associate Editor of the IEEE Transactions on Signal Processing. He is a Member of IMEKO Technical Committee TC-4.