Recent Advancements in Speech Enhancement

Yariv Ephraim and Israel Cohen[1]

March 9, 2004

Abstract

Speech enhancement is a long standing problem with numerous applications ranging from hearing aids, to coding and automatic recognition of speech signals. In this survey paper we focus on enhancement from a single microphone, and assume that the noise is additive and statistically independent of the signal. We present the principles that guide researchers working in this area, and provide a detailed design example. The example focuses on minimum mean square error estimation of the clean signal's log-spectral magnitude. This approach has attracted significant attention in the past twenty years. We also describe the principles of a Monte-Carlo simulation approach for speech enhancement.

1 Introduction

Enhancement of speech signals is required in many situations in which the signal is to be communicated or stored. Speech enhancement is required when either the signal or its receiver is degraded. For example, hearing impaired individuals require enhancement of perfectly normal speech to fit their individual hearing capabilities. Speech signals produced in a room generate reverberations, which may be quite noticeable when a hands-free single channel telephone system is used and binaural listening is not possible. A speech coder may be designed for clean speech signals while its input signal may be noisy. Similarly, a speech recognition system may be operated in an environment different from that it was designed to work in. This short list of examples illustrates the extent and complexity of the speech enhancement problem.

In this survey paper, we focus on enhancement of noisy speech signals for improving their perception by human. We assume that the noise is additive and statistically independent of the signal. In addition, we assume that the noisy signal is the only signal available for

[1]Y. Ephraim is with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030. Email: yephraim@gmu.edu
I. Cohen is with the Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa 32000, ISRAEL. Email: icohen@ee.technion.ac.il

enhancement. Thus, no reference noise source is assumed available. This problem is of great interest, and has attracted significant research effort for over fifty years. A successful algorithm may be useful as a preprocessor for speech coding and speech recognition of noisy signals.

The perception of a speech signal is usually measured in terms of its quality and intelligibility. The *quality* is a subjective measure which reflects on individual preferences of listeners. *Intelligibility* is an objective measure which predicts the percentage of words that can be correctly identified by listeners. The two measures are not correlated. In fact, it is well known that intelligibility can be improved if one is willing to sacrifice quality. This can be achieved, for example, by emphasizing high frequencies of the noisy signal [35]. It is also well known that improving the quality of the noisy signal does not necessarily elevate its intelligibility. On the contrary, quality improvement is usually associated with loss of intelligibility relative to that of the noisy signal. This is due to the distortion that the clean signal undergoes in the process of suppressing the input noise. From a pure information theoretic point of view, such loss in "information" is predicted by the *data processing theorem* [10]. Loosely speaking, this theorem states that one can never learn from the enhanced signal, more than he can learn from the noisy signal, about the clean signal.

A speech enhancement system must perform well for all speech signals. Thus, from the speech enhancement system point of view, its input is a random process whose sample functions are randomly selected by the user. The noise is naturally a random process. Hence, the speech enhancement problem is a statistical estimation problem of one random process from the sum of that process and the noise. Estimation theory requires statistical models for the signal and noise, and a distortion measure which quantifies the similarity of the clean signal and its estimated version. These two essential ingredients of estimation theory are not explicitly available for speech signals. The difficulties are with the lack of a precise model for the speech signal and a perceptually meaningful distortion measure. In addition, speech signals are not strictly stationary. Hence, adaptive estimation techniques, which do not require explicit statistical model for the signal, often fail to track the changes in the underlying statistics of the signal.

In this paper we survey some of the main ideas in the area of speech enhancement from a single microphone. We begin in Section 2 by describing some of the most promising statistical models and distortion measures which have been used in designing speech enhancement systems. In Section 3 we present a detailed design example for a speech enhancement system which is based on minimum mean square error estimation of the speech spectral magnitude. This approach integrates several key ideas from Section 2, and has

attracted much attention in the past twenty years. In Section 4, we present the principles of a Monte-Carlo simulation approach to speech enhancement. Some concluding comments are given in Section 5.

2 Statistical Models and Estimation

Enhancement of noisy speech signals is essentially an estimation problem in which the clean signal is estimated from a given sample function of the noisy signal. The goal is to minimize the expected value of some distortion measure between the clean and estimated signals. For this approach to be successful, a perceptually meaningful distortion measure must be used, and a reliable statistical model for the signal and noise must be specified. At present, the best statistical model for the signal and noise, and the most perceptually meaningful distortion measure, are not known. Hence, a variety of speech enhancement approaches have been proposed. They differ in the statistical model, distortion measure, and in the manner in which the signal estimators are being implemented. In this section, we briefly survey the most commonly used statistical models, distortion measures, and the related estimation schemes.

2.1 Linear Estimation

Perhaps the simplest scenario is obtained when the signal and noise are assumed statistically independent Gaussian processes, and the mean squared error (MSE) distortion measure is used. For this case, the optimal estimator of the clean signal is obtained by the Wiener filter. Since speech signals are not strictly stationary, a sequence of Wiener filters is designed and applied to vectors of the noisy signal. Suppose that Y_t and W_t represent, respectively, l-dimensional vectors from the clean signal and the noise process where $t = 0, 1, 2, \ldots$. Let $Z_t = Y_t + W_t$ denote the corresponding noisy vector. Let R_{Y_t} and R_{W_t} denote the covariance matrices of Y_t and W_t, respectively. Then, the minimum mean squared error (MMSE) estimate of the signal Y_t is obtained by applying the Wiener filter to the noisy signal Z_t as follows

$$\hat{Y}_t = \left[R_{Y_t} (R_{Y_t} + R_{W_t})^{-1} \right] Z_t. \tag{2.1}$$

Remarkably, this simple approach is one of the most effective speech enhancement approaches known today. The key to its success is reliable estimation of the covariance matrices of the clean signal and of the noise process. Many variations on this approach have been developed and were nicely summarized by Lim and Oppenheim [26]. When R_{Y_t} is estimated by subtracting an estimate of the covariance matrix of the noise vector, say $\hat{R}_{W_t}$,

from an estimate of the covariance matrix of the noisy vector, say $\hat{R}_{Z_t}$, then the Wiener filter at time t becomes $(\hat{R}_{Z_t} - \hat{R}_{W_t})\hat{R}_{Z_t}^{-1}$. The subtraction is commonly performed in the frequency domain where it is simpler to control the positive definiteness of the estimate of R_Y. This approach results in the simplest form of the family of "spectral subtraction" speech enhancement approaches [26].

MMSE estimation under Gaussian assumptions leads to linear estimation in the form of Wiener filtering given in (2.1). The same filter could be obtained if the Gaussian assumptions are relaxed, and the best *linear* estimator in the MMSE sense is sought. If we denote the linear filter for Y_t by the $l \times l$ matrix H_t, then the optimal H_t is obtained by minimizing the MSE given by $E\{||Y_t - H_t Z_t||^2\}$. Here $E\{\cdot\}$ denotes expected value, and $||\cdot||$ denotes the usual Euclidean norm. Note that when the filter H_t is applied to the noisy signal Z_t, it provides a residual signal given by

$$Y_t - \hat{Y}_t = Y_t - H_t Z_t = (I - H_t)Y_t + H_t W_t. \tag{2.2}$$

The term $(I - H_t)Y_t$ represents the distortion caused by the filter, and the term $H_t W_t$ represents the residual noise at the output of the filter. Since the signal and noise are statistically independent, the MSE error is the sum of two terms, the distortion energy $\overline{\epsilon_d^2} = E\{||(I - H_t)Y_t||^2\}$ and the residual noise energy $\overline{\epsilon_n^2} = E\{||H_t W_t||^2\}$. The Wiener filter minimizes $\overline{\epsilon_d^2} + \overline{\epsilon_n^2}$ over all possible filters H_t. An alternative approach proposed by Ephraim and Van-Trees [18] was to design the filter H_t by minimizing the distortion energy $\overline{\epsilon_d^2}$ for a given level of acceptable residual noise energy $\overline{\epsilon_n^2}$. This approach allows the design of a filter which controls the contributions of the two competing components $\overline{\epsilon_d^2}$ and $\overline{\epsilon_n^2}$ to the MSE. The resulting filter is similar to that in (2.1) except that R_{W_t} is replaced by $\mu_t R_{W_t}$ where μ_t is the Lagrange multiplier of the constrained optimization problem. The idea was extended to filter design which minimizes the distortion energy for a given desired spectrum of the residual noise. This interesting optimization problem was solved by Lev-Ari and Ephraim in [25]. The estimation criterion was motivated by the desire to adjust the spectrum of the residual noise so that it is least audible.

In [18], the two estimation criteria outlined above were applied to enhancement of noisy speech signals. It was noted that there is strong empirical evidence that supports the notion that covariance matrices of many speech vectors are not full rank matrices. This notion is also supported by the popular sinusoidal model for speech signals, in which a speech vector with $l = 200 - 400$ samples at 8kHz sampling rate, is spanned by fewer than l sinusoidal components. As such, some of the eigenvalues of R_{Y_t} are practically zero, and the vector Y_t occupies a subspace of the Euclidean space $\mathcal{R}^l$. A white noise, however, occupies the entire space $\mathcal{R}^l$. Thus, the Euclidean space $\mathcal{R}^l$ may be decomposed into a "signal subspace"

containing signal plus noise, and a complementary "noise subspace" containing noise only. Thus, in enhancing a noisy vector Z_t, one can first null out the component of Z_t in the noise subspace and filter the noisy signal in the signal subspace. The decomposition of Z_t into its signal subspace component and noise subspace component can be performed by applying the Karhunen-Loève transform to Z_t.

2.2 Spectral Magnitude Estimation

In Section 2.1 we focused on MMSE estimation of the waveform of the speech signal. This estimation may be cast in the frequency domain as follows. We use $(\cdot)'$ to denote conjugate transpose. Let D' denote the discrete Fourier transform (DFT) matrix. Let $\mathbf{Z}_t = \frac{1}{\sqrt{l}} D' Z_t$ denote the vector of spectral components of the noisy vector Z_t. For convenience, we have chosen to use normalized DFT. We denote the kth spectral component of the noisy vector Z_t by $\mathbf{Z}_{tk}$. Let $\Lambda_{\mathbf{Z}_t}$ be a diagonal matrix with the variances of the spectral components $\{\mathbf{Z}_{tk}, k = 0, 1, \ldots, l-1\}$ on its main diagonal. Assume, for simplicity, that R_{Y_t} and R_{W_t} are circulant matrices [24]. This means that $R_{Y_t} = \frac{1}{l} D \Lambda_{\mathbf{Y}_t} D'$ and $R_{W_t} = \frac{1}{l} D \Lambda_{\mathbf{W}_t} D'$. Let $\hat{\mathbf{Y}}_t = \frac{1}{\sqrt{l}} D' \hat{Y}_t$ be the normalized DFT of the MMSE estimate $\hat{Y}_t$. Under these assumptions, (2.1) becomes

$$\hat{\mathbf{Y}}_t = \left[\Lambda_{\mathbf{Y}_t} (\Lambda_{\mathbf{Y}_t} + \Lambda_{\mathbf{W}_t})^{-1} \right] \mathbf{Z}_t. \tag{2.3}$$

This filter performs MMSE estimation of the spectral components $\{\ \mathbf{Y}_{tk}\}$ of the clean vector Y_t. It is commonly believed, however, that the human auditory system is more sensitive to the short-term spectral magnitude $\{|\ \mathbf{Y}_{tk}|, k = 0, 1, \ldots, l-1\}$ of the speech signal than to its short-term phase $\{\arg(\ \mathbf{Y}_{tk}), k = 0, 1, \ldots, l-1\}$. This has been demonstrated by Wang and Lim [37] in a sequence of experiments. They have synthesized speech signals using short-term spectral magnitude and phase derived from two noisy versions of the same speech signal at different signal to noise ratios (SNR's). Thus, they could control the amount of noise in the spectral magnitude and in the phase. Hence, it was suggested that better enhancement results could be obtained if the spectral magnitude of a speech signal rather than its waveform is directly estimated. In this situation, the phase of the noisy signal is combined with the spectral magnitude estimator in constructing the enhanced signal. Maximum likelihood estimates of the short-term spectral magnitude of the clean signal were developed by McAulay and Malpass [32] for additive Gaussian noise. An MMSE estimator of the short-term spectral magnitude of speech signal was developed by Ephraim and Malah [14]. The spectral components of the clean signal and of the noise process were assumed statistically independent Gaussian random variables. Under the same assumptions, the MMSE estimator of the short-term complex exponential of the clean signal,

$\exp(j \arg(\mathbf{Y}_{tk}))$, which does not affect the spectral magnitude estimator (i.e., has a unity modulus), was shown in [14] to be equal to the complex exponential of the noisy signal. This confirmed the intuitive use of the noisy phase in systems which capitalize on spectral magnitude estimation.

It is further believed that the human auditory system compresses the signal's short-term spectral magnitude in the process of its decoding. It was suggested that a form of logarithmic compression is actually taking place. Hence, better enhancement of the noisy signal should be expected if the logarithm of the short-term spectral magnitude is directly estimated. An MMSE estimator of the log-spectral magnitude of speech signal was developed by Ephraim and Malah [15] under the same Gaussian assumptions described above. This approach has attracted much interest in recent years and will be presented in more details in Section 3.

2.3 The Gaussian Model

The assumption that spectral components of the speech signal at any given frame are statistically independent Gaussian random variables, underlies the design of many speech enhancement systems. In this model, the real and imaginary parts of each spectral component are also assumed statistically independent identically distributed random variables. We have mentioned here the Wiener filter for MMSE estimation of the spectral components of the speech signal, and the MMSE estimators for the spectral magnitude and for the logarithm of the spectral magnitude of the clean signal. The Gaussian assumption is mathematically tractable, and it is often justified by a version of the central limit theorem for correlated signals [4, Theorem 4.4.2]. The Gaussian assumption for the real and imaginary parts of a speech spectral component has been challenged by some authors, see, e.g., [33], [30]. In [33], for example, the spectral magnitude was claimed to have a Gamma distribution. In [30], the real and imaginary parts of a spectral component were assumed statistically independent Laplace random variables. We now show that the Gaussian and other models are not necessarily contradictory.

The assumption that a spectral component is Gaussian is always conditioned on knowledge of the variance of that component. Thus, the Gaussian assumption is attributed to the conditional probability density function (pdf) of a spectral component given its variance. A conditionally Gaussian spectral component may have many different marginal pdf's. To demonstrate this point, consider the spectral component $\mathbf{Y}_{tk}$ and its variance $\sigma^2_{\mathbf{Y}_{tk}}$. Let the real part of $\mathbf{Y}_{tk}$ be denoted by Y. Let the variance $\sigma^2_{\mathbf{Y}_{tk}}/2$ of the real part of $\mathbf{Y}_{tk}$ be denoted by V. Assume that the conditional pdf of Y given V is Gaussian. Denote this pdf

by $p(y|v)$. Assume that the variance V has a pdf $p(v)$. Then the marginal pdf of Y is given by

$$p(y) = \int p(y|v)p(v)dv. \tag{2.4}$$

The pdf of Y is thus a continuous mixture of Gaussian densities. This pdf may take many different forms which are determined by the specific prior pdf assumed for V. For example, suppose that V is exponentially distributed with expected value $2\lambda^2$, i.e., assume that

$$p(y|v) = \frac{e^{-\frac{y^2}{2v}}}{\sqrt{2\pi v}} \quad \text{and} \quad p(v) = \frac{e^{-\frac{v}{2\lambda^2}}}{2\lambda^2}u(v) \tag{2.5}$$

where $u(\sigma)$ is a unit step-function. Substituting (2.5) into (2.4) and using [23, eq. (3.325)] shows that

$$p(y) = \frac{1}{2\lambda}e^{-\frac{|y|}{\lambda}} \tag{2.6}$$

or that Y has a Laplace pdf just as it was assumed in [30]. This argument shows that estimators for a spectral component of speech signal obtained under non-Gaussian models may be derived using the conditional Gaussian pdf and an appropriately chosen pdf for the variance of the spectral component. In our opinion, using the conditional Gaussian model is preferable, since it is much better understood, and it is significantly easier to work with.

The variance of a spectral component must be assumed a random variable, since speech signals are not strictly stationary. Thus, the variance sequence $\{\sigma^2_{\mathbf{Y}_{tk}}, t = 1, 2, \ldots\}$ corresponding to the sequence of spectral components $\{\ \mathbf{Y}_{tk}, t = 1, 2, \ldots\}$ at a given frequency k, is not known in advance and is best described as a random sequence. In [14], [15], the variance of each spectral component of the clean signal was estimated and updated from the noisy signal using the decision-directed estimator. In [13], the variance sequence was assumed a Markov chain and it was estimated online from the noisy signal. In [8], a recursive formulation of the variance estimator is developed following the rational of Kalman filtering.

A closely related statistical model for speech enhancement is obtained by modeling the clean speech signal as a hidden Markov process (HMP). An overview of HMP's may be found in [19]. Speech enhancement systems using this model were first introduced by Ephraim, Malah and Juang [16]. An HMP is a bivariate process of state and observation sequences. The state sequence is a homogeneous Markov chain with a given number of states, say M. The observation sequence is conditionally independent given the sequence of states. This means that the distribution of each observation depends only on the state at the same time and not on any other state or observation. Let $S^n = \{S_1, \ldots, S_n\}$ denote the state sequence where we may assume without loss of generality that $S_t \in \{1, \ldots, M\}$. Let

$Y^n = \{Y_1, \ldots, Y_n\}$ denote the observation sequence where each Y_t is a vector in a Euclidean space $\mathcal{R}^l$. The joint density of (S^n, Y^n) is given by

$$p(s^n, y^n) = \prod_{t=1}^{n} p(s_t|s_{t-1})p(y_t|s_t) \tag{2.7}$$

where $p(s_1|s_0) = p(s_1)$. When $S_t = j$, we replace $p(y_t|s_t)$ by $p(y_t|j)$. In [16], [17], $p(y_t|j)$ was assumed to be the pdf of a vector from a zero mean Gaussian autoregressive process. The parameter of the process, i.e., the autoregressive coefficients and gain, depends on the state j. This parameter characterizes the power spectral density of the signal in the given vector. Thus, $p(y_t|j)$ was assumed in [16], [17] to be conditionally Gaussian given the power spectral density of the signal. There are M power spectral density prototypes for all vectors of the speech signal. The HMP assumes that each vector of the speech signal is drawn with some probability from one of the M autoregressive processes. The identity of the autoregressive process producing a particular vector is not known, and hence the pdf of each vector is a finite mixture of Gaussian autoregressive pdf's. In contrast, (2.4) represents a mixture of countably infinite Gaussian pdf's. In the HMP model, spectral components of each vector of the speech signal are assumed correlated since each vector is assumed autoregressive, and consecutive speech vectors are weakly dependent since they inherit the memory of the Markov chain.

2.4 Signal Presence Uncertainty

In all models presented thus far in this section, the clean signal was assumed to be present in the noisy signal. Thus we have always viewed the noisy signal vector at time t as $Z_t = Y_t + W_t$. In reality, however, speech contains many pauses while the noise may be continuously present. Thus the noisy signal vector at time t may be more realistically described as resulting from two possible hypotheses: H_1 indicating signal presence and H_0 indicating signal absence. We have

$$Z_t = \begin{cases} Y_t + W_t & \text{under } H_1 \\ W_t & \text{under } H_0 \end{cases} \tag{2.8}$$

This insightful observation was first made by McAulay and Malpass [32] who have modified their speech signal estimators accordingly. For MMSE estimation, let $E\{Y_t|Z_t, H_1\}$ denote the conditional mean estimate of Y_t when the signal is assumed present in Z_t. Let $P(H_1|Z_t)$ denote the probability of signal presence given the noisy vector. The MMSE of Y_t given Z_t is given by

$$E\{Y_t|Z_t\} = P(H_1|Z_t)E\{Y_t|Z_t, H_1\}. \tag{2.9}$$

The model of speech presence uncertainty may be refined and attributed to spectral components of the vector Z_t [14]. This aspect will be dealt with more details in Section 3.

2.5 Multi-State Speech Model

The signal presence uncertainty model may be seen as a two-state model for the noisy signal. A five-state model for the clean signal was proposed earlier by Drucker [12]. The states in his model represent fricative, stop, vowel, glide, and nasal speech sounds. For enhancing a noisy signal, he proposed to first classify each vector of the noisy signal as originating from one of the five possible class sounds, and then to apply a class-specific filter to the noisy vector.

The HMP model for the clean signal described in Section 2.3 is a multi-class model. When HMP's are used, the classes are not a-priori defined, but they are rather created in a learning process from some training data of clean speech signals. The learning process is essentially a clustering process that may be performed using vector quantization techniques [22]. For example, each class may contain spectrally similar vectors of the signal. Thus, each class may be characterized by a prototype power spectral density which may be parameterized as an autoregressive process. Transitions from one spectral prototype to another are probabilistic and are performed in a Markovian manner. The noise process may be similarly represented. If there are M speech classes and N noise classes, then $M \times N$ estimators must be designed for enhancing noisy speech signals. Suppose that we are interested in estimating the speech vector Y_t given a sequence of noisy speech vectors $z^t = \{z_1, \ldots, z_t\}$. Let $p((i,j)|z^t)$ denote the probability of the signal being in state i and the noise being in state j given z^t. Then, the MMSE estimator of Y_t from z^t is given by [17]

$$E\{Y_t|z^t\} = \sum_{i=1}^{M}\sum_{j=1}^{N} p((i,j)|z^t) E\{Y_t|z^t,(i,j)\}. \tag{2.10}$$

3 MMSE Spectral Magnitude Estimation

In this section we focus on MMSE estimation of the logarithm of the short-term spectral magnitude of the clean signal. We provide a design example of a speech enhancement system which relies on conditional Gaussian modeling of spectral components and on speech presence uncertainty. Recall that the kth spectral component of the clean speech vector Y_t is denoted by $\mathbf{Y}_{tk}$. The variance of $\mathbf{Y}_{tk}$ is denoted by $\sigma^2_{\mathbf{Y}_{tk}}$. It is assumed that spectral components $\{\mathbf{Y}_{tk}\}$ with given variances $\{\sigma^2_{\mathbf{Y}_{tk}} > 0\}$ are statistically independent Gaussian

random variables. Similar assumptions are made for the spectral components of the noise process $\{\mathbf{W}_{tk}\}$.

The spectral component $\mathbf{Z}_{tk}$ of the noisy signal is given by

$$\mathbf{Z}_{tk} = \mathbf{Y}_{tk} + \mathbf{W}_{tk}. \tag{3.1}$$

Let H_1^{tk} and H_0^{tk} denote the hypotheses of speech presence and speech absence in the noisy spectral component $\mathbf{Z}_{tk}$, respectively. Let q_{tk} denote the probability of H_1^{tk}. The spectral components of the noisy signal $\{\mathbf{Z}_{tk}\}$ are statistically independent Gaussian random variables given their variances $\{\sigma^2_{\mathbf{Z}_{tk}}\}$.

We are interested in estimating the logarithm of the spectral magnitude of each component of the clean signal from all available spectral components of the noisy signal. Under the statistical model assumed here, given the variances of the spectral components and the probabilities of speech presence, estimation of $\log|\mathbf{Y}_{tk}|$ is performed from $\mathbf{Z}_{tk}$ only. Since the variances of the spectral components and the probabilities of speech presence are not available, however, these quantities are estimated for each frequency k from the noisy spectral components observed up to time t, and the estimates are plugged in the signal estimate. We use $\widehat{\sigma^2_{\mathbf{Y}_{tk}}}$ and $\widehat{\sigma^2_{\mathbf{W}_{tk}}}$ to denote estimates of the variances of $\mathbf{Y}_{tk}$ and $\mathbf{W}_{tk}$, respectively, and $\hat{q}_{tk}$ to denote an estimate of q_{tk}. We next present estimation of the signal and its assumed known parameter.

3.1 Signal Estimation

The signal estimator is conveniently expressed in terms of the a-priori and a-posteriori SNR's. These quantities are defined as

$$\xi_{tk} = \frac{\sigma^2_{\mathbf{Y}_{tk}}}{\sigma^2_{\mathbf{W}_{tk}}} \quad \text{and} \quad \gamma_{tk} = \frac{|\mathbf{Z}_{tk}|^2}{\sigma^2_{\mathbf{W}_{tk}}} \tag{3.2}$$

respectively, where l denotes the frame length. We also define

$$\vartheta_{tk} = \frac{\xi_{tk}}{\xi_{tk}+1}\gamma_{tk}. \tag{3.3}$$

The estimates of ξ_{tk} and γ_{tk} used here are $\hat{\xi}_{tk} = \widehat{\sigma^2_{\mathbf{Y}_{tk}}}/\widehat{\sigma^2_{\mathbf{W}_{tk}}}$ and $\hat{\gamma}_{tk} = |\mathbf{Z}_{tk}|^2/\widehat{\sigma^2_{\mathbf{W}_{tk}}}$. To prevent estimation of the logarithm of negligibly small spectral magnitudes under the hypothesis that speech is absent in $\mathbf{Z}_{tk}$, Cohen and Berdugo [6] proposed to estimate the conditional mean of the following function of $\mathbf{Y}_{tk}$

$$f(\mathbf{Y}_{tk}) = \begin{cases} \log|\mathbf{Y}_{tk}|, & \text{under } H_1^{tk} \\ \log\nu_{tk}, & \text{under } H_0^{tk} \end{cases} \tag{3.4}$$

where ν_{tk} is a spectral threshold. They showed that

$$\begin{aligned} |\,\widehat{\mathbf{Y}_{tk}}| &= \exp\left\{E\left\{f(\,\mathbf{Y}_{tk})\,|\,\mathbf{Z}_{tk};\,\widehat{\sigma^2_{\mathbf{Y}_{tk}}},\,\hat{\xi}_{tk},\,\hat{q}_{tk}\right\}\right\} \\ &= \left[G(\hat{\xi}_{tk},\hat{\gamma}_{tk})\,|\mathbf{Z}_{tk}|\right]^{\hat{q}_{tk}}\,\nu_{tk}^{1-\hat{q}_{tk}} \end{aligned} \tag{3.5}$$

where

$$G(\xi,\gamma) = \frac{\xi}{\xi+1}\exp\left(\frac{1}{2}\int_{\vartheta}^{\infty}\frac{e^{-x}}{x}dx\right) \tag{3.6}$$

represents the spectral gain function derived by Ephraim and Malah [15] under H_1^{tk}. Note that this gain function depends on $\mathbf{Z}_{tk}$ and hence the estimator in (3.5) is nonlinear even when the parameter of the statistical model is known. It was further proposed in [6] to replace ν_{tk} in (3.5) by $G_{\min}|\mathbf{Z}_{tk}|$ where $G_{\min} \ll 1$. This substitution provides a constant attenuation of $|\mathbf{Z}_{tk}|$ under H_0^{tk} rather than using a constant term that is independent of $|\mathbf{Z}_{tk}|$. This practice is closely related to the "spectral floor" modification of the spectral subtraction method proposed by Berouti, Schwartz and Makhoul [3]. The constant attenuation retains the naturalness of the residual noise when the signal is absent. Substituting this constant attenuation in (3.5) gives

$$|\,\widehat{\mathbf{Y}_{tk}}| = [G(\hat{\xi}_{tk},\hat{\gamma}_{tk})]^{\hat{q}_{tk}}\,G_{\min}^{1-\hat{q}_{tk}}\,|\mathbf{Z}_{tk}|. \tag{3.7}$$

To form an estimator $\hat{\mathbf{Y}}_{tk}$ for the clean spectral component $\mathbf{Y}_{tk}$, the spectral magnitude estimator $|\,\widehat{\mathbf{Y}_{tk}}|$ is combined with an estimator of the phase of $\mathbf{Y}_{tk}$. Ephraim and Malah [14] proposed to use the MMSE estimator of the complex exponential of that phase. The modulus of the estimator was constrained to a unity so that it does not affect the optimality of the spectral magnitude estimator $|\,\widehat{\mathbf{Y}_{tk}}|$. They showed that the constrained MMSE estimator is given by the complex exponential of the noisy phase.

The integral in (3.6) is the well known Exponential Integral of ϑ, and it can be numerically evaluated, e.g., using the *expint* function in MATLAB. Alternatively, it may be evaluated by using the following computationally efficient approximation, which was developed by Martin *et al.* [31]

$$\text{expint}(\vartheta) = \int_{\vartheta}^{\infty}\frac{e^{-x}}{x}dx \approx \begin{cases} -2.31\,\log_{10}(\vartheta) - 0.6\,, & \text{for } \vartheta < 0.1 \\ -1.544\,\log_{10}(\vartheta) + 0.166\,, & \text{for } 0.1 \le \vartheta \le 1 \\ 10^{-0.52\,\vartheta - 0.26}\,, & \text{for } \vartheta > 1\,. \end{cases} \tag{3.8}$$

3.2 Signal Presence Probability Estimation

In this section we address the problem of estimating the speech presence probability q_{tk}. Define a binary random variable V_{tk} which indicates whether or not speech is present in

the spectral component $\mathbf{Z}_{tk}$.

$$V_{tk} = \begin{cases} 1 & \text{under} \quad H_1^{tk} \\ 0 & \text{under} \quad H_0^{tk} \end{cases} \tag{3.9}$$

Cohen and Berdugo [6] proposed to estimate q_{tk} as the conditional mean of V_{tk} given $\mathbf{Z}_{tk}$ and an estimate of the parameter of the statistical model. Specifically,

$$\hat{q}_{tk} = E\{V_{tk}|\mathbf{z}_{tk}; \widehat{\sigma^2_{\mathbf{W}_{tk}}}, \hat{\xi}_t\} = P(H_1^{tk}|\mathbf{z}_{tk}; \widehat{\sigma^2_{\mathbf{W}_{tk}}}, \hat{\xi}_t). \tag{3.10}$$

Using Bayes' rule, they expressed the conditional probability of H_1^{kt} in (3.10) in terms of the Gaussian densities of $\mathbf{Z}_{tk}$ under the two hypotheses and some estimate of the prior probability of H_1^{tk}. They provided a scheme for estimating the prior probability from spectral components observed up to time $t-1$. Let the prior probability estimate be denoted by $\hat{q}_{tk|t-1}$. Following this approach they showed that [6]

$$\hat{q}_{tk} = \left[1 + \frac{1 - \hat{q}_{tk|t-1}}{\hat{q}_{tk|t-1}}(1 + \hat{\xi}_{tk}) \exp(-\hat{\vartheta}_{tk})\right]^{-1} \tag{3.11}$$

where $\hat{\vartheta}_{tk}$ is the estimate of ϑ_{tk} defined in (3.3).

The estimator $\hat{q}_{tk|t-1}$ is based on the distribution of the a priori SNR, and the relation between the likelihood of speech absence in the time-frequency domain and the local and global averages of the a priori SNR. The speech absence probability is estimated for each frequency bin and each frame by a soft-decision approach, which exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames.

3.3 A Priori SNR Estimation

Reliable estimation of the speech spectral component variances is crucial for successful implementation of the signal estimator (3.7). Ephraim and Malah [14] proposed a decision-directed variance estimator for their MMSE spectral magnitude estimator. The variance estimator at a given frame uses the signal spectral magnitude estimate from the previous frame along with the current noisy spectral component. Let $\hat{A}_{tk} = |\widehat{\mathbf{Y}_{tk}}|$ denote the MMSE signal spectral magnitude estimate from $\mathbf{Z}_{tk}$. The decision-directed estimate of the variance of $\mathbf{Y}_{tk}$ is given by

$$\widehat{\sigma^2_{\mathbf{Y}_{tk}}} = \frac{1}{\hat{q}_{tk}} \left[\alpha \hat{A}^2_{t-1,k} + (1-\alpha) \max\left\{|\mathbf{Z}_{tk}|^2 - \widehat{\sigma^2_{\mathbf{W}_{tk}}}, 0\right\}\right] \tag{3.12}$$

where $0 \leq \alpha \leq 1$ is an experimental constant. The estimator was also found useful when $\hat{A}_{tk}$ is the MMSE log-spectral magnitude estimator [15]. In the latter case, the estimator was used with $\hat{q}_{tk} = 1$ since the signal was assumed zero under the null hypothesis. While this

estimator was found useful in practice, the division by $\hat{q}_{tk}$ may deteriorate the performance of the speech enhancement system [34]. In some cases, it introduces interaction between the estimated $\hat{q}_{tk}$ and the a priori SNR, resulting in unnaturally structured residual noise [28].

Cohen and Berdugo [6] showed that a preferable variance estimator is obtained if $\hat{A}_{t-1,k}$ in (3.12) is replaced by the estimator $\hat{A}_{t-1,k|H_1^{tk}}$ for the magnitude of $\mathbf{Y}_{t-1,k}$ obtained under the signal presence hypothesis, and the division by $\hat{q}_{tk}$ is not performed. The resulting estimator is given by

$$\widehat{\sigma^2_{\mathbf{Y}_{tk}}} = \alpha\, \hat{A}^2_{t-1,k|H_1^{tk}} + (1-\alpha) \max\left\{ |\mathbf{Z}_{tk}|^2 - \widehat{\sigma^2_{\mathbf{W}_{tk}}}\, ,\, 0 \right\} \tag{3.13}$$

Expressing $\hat{A}_{t-1,k|H_1^{tk}}$ in terms of the gain function form (3.7), dividing by $\widehat{\sigma^2_{\mathbf{W}_{tk}}}$, and imposing a lower bound $\xi_{\min} > 0$ on the a-priori SNR estimate as proposed by Cappé [5], they obtained the following recursive estimator for ξ_{tk}

$$\hat{\xi}_{tk} = \max\left\{ \alpha\, G^2\left(\hat{\xi}_{t-1,k}, \hat{\gamma}_{t-1,k}\right)\, \hat{\gamma}_{t-1,k} + (1-\alpha)\,(\hat{\gamma}_{tk} - 1)\, ,\, \xi_{\min} \right\}. \tag{3.14}$$

The parameters α and $\xi_{\min}$ control the trade-off between the noise reduction and the transient distortion introduced into the signal [14], [5]. Greater reduction of the musical noise phenomena is obtained by using a larger α and a smaller $\xi_{\min}$, at the expense of attenuated speech onsets and audible modifications of transient speech components. Typical values for α range between 0.9 and 0.99, and typical values for $\xi_{\min}$ range between -10 and -25 dB.

3.4 Noise Spectrum Estimation

In stationary noise environments, the noise variance of each spectral component is time invariant, i.e., $\sigma^2_{\mathbf{W}_{tk}} = \sigma^2_{\mathbf{W}_k}$ for all t. An estimator for $\sigma^2_{\mathbf{W}_k}$ may be obtained from recursive averaging of $\{|\mathbf{Z}_{tk}|^2\}$ for all spectral components classified as containing noise only.

In non-stationary noise environments, an alternative approach, known as the *minimum statistics*, was proposed by Martin [27], [29]. In this approach, minima values of a smoothed power spectral density estimate of the noisy signal are tracked, and multiplied by a constant that compensates the estimate for possible bias. We present here a recent algorithm, developed by Cohen and Berdugo [7], [9], which is based on *minima controlled recursive averaging*. This noise variance estimator is capable of fast adaptation to abrupt changes in the noise spectrum.

Recall that H_0^{tk} and H_1^{tk} denote, respectively, speech absence and presence hypotheses in the noisy spectral component $\mathbf{Z}_{tk}$. A recursive estimate for the noise spectral variance

can be obtained as follows.

$$\widehat{\sigma^2_{\mathbf{W}_{t+1,k}}} = \begin{cases} \mu\ \widehat{\sigma^2_{\mathbf{W}_{tk}}} + (1-\mu)\,\beta\,|\mathbf{Z}_{tk}|^2 & \text{under } H_0^{tk} \\ \widehat{\sigma^2_{\mathbf{W}_{tk}}} & \text{under } H_1^{tk} \end{cases} \tag{3.15}$$

where $0 < \mu < 1$ is a smoothing parameter and $\beta \geq 1$ is a bias compensation factor [9]. The probability of H_1^{tk} is estimated here independently of $\hat{q}_{tk}$ in Section 3.2, since the penalty in misclassification of the two hypotheses has different consequences when estimating the signal than when estimating the noise spectral variance. Generally, here we tend to decide H_0^{tk} with higher confidence than in Section 3.2. Let $\tilde{q}_{tk}$ denote the estimate of the probability of H_1^{tk} in this section. A soft-decision recursive estimator can be obtained from (3.15) by

$$\begin{aligned} \widehat{\sigma^2_{\mathbf{W}_{t+1,k}}} &= \tilde{q}_{tk}\,\widehat{\sigma^2_{\mathbf{W}_{tk}}} + (1-\tilde{q}_{tk})\left[\mu\ \widehat{\sigma^2_{\mathbf{W}_{tk}}} + (1-\mu)\,\beta\,|\mathbf{Z}_{tk}|^2\right] \\ &= \tilde{\mu}_{tk}\ \widehat{\sigma^2_{\mathbf{W}_{tk}}} + (1-\tilde{\mu}_{tk})\,\beta\,|\mathbf{Z}_{tk}|^2 \end{aligned} \tag{3.16}$$

where $\tilde{\mu}_{tk} = \mu + (1-\mu)\,\tilde{q}_{tk}$ is a time-varying smoothing parameter.

The probability $\tilde{q}_{tk}$ is estimated using (3.11) when $\hat{q}_{tk|t-1}$ is substituted by a properly designed estimate $\tilde{q}_{tk|t-1}$. Cohen [9] proposed an estimator $\tilde{q}_{tk|t-1}$ which is controlled by the minima values of a smoothed power spectrum of the noisy signal. The estimation procedure comprises two iterations of smoothing and minimum tracking. The first iteration provides a rough voice activity detection in each frequency. Smoothing during the second iteration excludes relatively strong speech components, which makes the minimum tracking during speech activity more robust.

3.5 Summary of Algorithm

i) For $t = 0$ and all k's, set $\widehat{\sigma^2_{\mathbf{W}_{tk}}} = |\mathbf{Z}_{0k}|^2$, $\hat{\gamma}_{-1,k} = 1$, $\hat{\xi}_{-1,k} = \xi_{\min}$. Set $t = 1$.

ii) For each k

- Calculate $\hat{\gamma}_{tk}$ from (3.2), and $\hat{\xi}_{tk}$ from (3.14).
- Calculate $\hat{q}_{tk|t-1}$ from [6, eq. (29)], and $\hat{q}_{tk}$ from (3.11).
- Calculate $G\left(\hat{\xi}_{tk}, \hat{\gamma}_{tk}\right)$ from (3.6), and $|\ \widehat{\mathbf{Y}_{tk}}|$ by using (3.7).
- Calculate $\tilde{q}_{tk|t-1}$ from [9, eq. (28)], and $\tilde{q}_{tk}$ from the analog of (3.11).
- Update $\widehat{\sigma^2_{\mathbf{W}_{tk}}}$ by using (3.16).

iii) Set $t \rightarrow t+1$ and go to step ii) for enhancement of the next frame.

4 Monte-Carlo Simulation

The Monte-Carlo simulation approach for audio signal enhancement has been promoted by Vermaak, Andrieu, Doucet, Godsill, Fong and West [20], [36]. In this section we present the principles of this approach. The clean and noisy speech signals are represented by the sequences of scalar random variables $\{Y_t, t = 0, 1, \ldots\}$ and $\{Z_t, t = 1, 2, \ldots\}$, respectively. These signals are assumed to satisfy some time-varying state-space equations. The time-varying parameter of the system is denoted by $\{\theta_t, t = 1, 2, \ldots\}$. The system is characterized by three deterministically known non-linear transition functions which we denote here by f, g and h. The explicit dependence of f on t, and of g and h on θ_t, is expressed by writing these functions as f_t, g_{θ_t} and h_{θ_t}, respectively. The innovation processes of the dynamical system are denoted by $\{U_t, t = 1, 2, \ldots\}$, $\{V_t, t = 1, 2, \ldots\}$ and $\{W_t, t = 1, 2, \ldots\}$. These three processes are assumed statistically independent iid processes. The state-space equations are given by

$$\begin{aligned} \theta_t &= f_t(\theta_{t-1}, U_t) \\ Y_t &= g_{\theta_t}(Y_{t-1}, V_t) \\ Z_t &= h_{\theta_t}(Y_t, W_t) \end{aligned} \tag{4.1}$$

for $t = 1, 2, \ldots$.

Assume first that the sample path of $\{\theta_t\}$ is known. For this case, the signal $\{Y_t\}$ can be recursively estimated from $\{Z_t\}$. To simplify notation, we present these recursions without explicitly showing the dependence of the various pdf's on the assumed known parameter path. We use lower case letters to denote realizations of the random variables in (4.1). We also denote $z^t = \{z_1, \ldots, z_t\}$. The filtering and prediction recursions result from Markov properties of the signals in (4.1) and from Bayes' rule. These recursions are, respectively, given by

$$p(y_t|z^t) = \frac{p(y_t|z^{t-1})p(z_t|y_t)}{\int p(y_t|z^{t-1})p(z_t|y_t)dy_t}, \quad t = 1, \ldots, n \tag{4.2}$$

where $p(y_1|z_1^0) = p(y_1)$, and by

$$p(y_t|z^{t-1}) = \int p(y_t|y_{t-1})p(y_{t-1}|z^{t-1})dy_{t-1}, \quad t = 2, \ldots, n. \tag{4.3}$$

The smoothing recursion was derived by Askar and Derin [2, Theorem 1] and it is given by

$$p(y_t|z^n) = p(y_t|z^t) \int \frac{p(y_{t+1}|z_t)p(y_{t+1}|z^n)}{p(y_{t+1}|z^t)} dy_{t+1} \tag{4.4}$$

for $t = n-1, n-2, \ldots, 1$, where $p(y_n|z^n)$ is given by (4.2).

When the sample path of $\{\theta_t\}$ is given, or when the parameter is time-invariant and known ($\theta_t = \theta_0$ for all t), these recursions can be implemented with reasonable complexity

for two well known cases. First, when g and h are linear functions, $\{V_t\}$ and $\{W_t\}$ are Gaussian processes, and the initial distribution of Y_0 is Gaussian. In that case, $\{Y_t\}$ can be estimated using the Kalman filter or smoother. Second, when $\{Y_t\}$ takes finitely many values, then the integrals become summations and the recursions coincide with a version of the forward-backward recursions for hidden Markov processes, see, e.g., [19, eqs. (5.14)-(5.16)]. For all other systems, the estimation problem is highly non-linear and requires multidimensional integrations. No simple solution exists for these situations. Approximate solutions are often obtained using the extended Kalman filter. The latter applies Kalman filtering to locally linearized versions of the state space equations.

When the sample path of $\{\theta_t\}$ is not known, but the three transition functions are linear and the innovation processes are Gaussian, maximum a-posteriori estimation of $\{\theta_t\}$ is possible using the expectation-maximization (EM) algorithm. This was shown by Dembo and Zeitouni [11] who developed an EM algorithm for estimating $\{\theta_t\}$ when the signal $\{Y_t\}$ is a time-varying autoregressive process. The parameter estimator relies on Kalman smoothers for the clean signal $\{Y_t\}$ and its covariance at each EM iteration. Thus, an estimate of the clean signal is obtained as a by product in this algorithm. A similar approach for maximum likelihood estimation of deterministically unknown parameter was implemented and tested for speech enhancement by Gannot, Burshtein and Weinstein [21].

The computational difficulties in estimating the parameter or the clean signal in (4.1) have stimulated the use of Monte-Carlo simulations. A good tutorial on the subject was written by Arulampalam, Maskell, Gordon and Clapp [1]. In this approach, probability distributions are sampled and replaced by empirical distributions. Thus integrals involving the sampled pdf's can be straightforwardly evaluated using sums. Recursive sampling is often desirable to facilitate the approach. The filters or smoothers designed in this way are often referred to as *particle filters*. The "particles" refer to the point masses obtained from sampling the distribution which is of interest in the given problem. There is more than one way to simulate the filtering or smoothing recursions presented earlier. We focus here on the work in [20], [36] where the approach has been applied to speech and audio signals and compared with the extended Kalman filter. In [20], Monte-Carlo approaches for filtering as well as smoothing were developed. We shall demonstrate here only the principles of the filtering approach.

Similarly to the work of Dembo and Zeitouni [11], the signal in [20] was assumed a Gaussian time-varying autoregressive process, and the additive noise was assumed Gaussian. In fact, the reflection coefficients of the time-varying autoregressive process were assumed a Gaussian random walk process, which was constrained to the interval of $(-1, 1)$, but the

nonlinear transformation from the reflection coefficients to the autoregressive coefficients was ignored. The logarithm of the gain of the autoregressive process was also modeled as a Gaussian random walk. The pdf $p(\theta_t|z^t)$ of θ_t given z^t was shown to be proportional to

$$p(\theta_t|z^t) \propto \int p(z_t|\theta^t, z^{t-1})p(\theta_t|\theta_{t-1})p(\theta^{t-1}|z^{t-1})d\theta^{t-1}. \tag{4.5}$$

This equation can be derived similarly to (4.2). The goal now is to recursively sample $p(\theta_t|z^t)$ and estimate the signal using an efficient algorithm such as the Kalman filter. Suppose that at time t we have an estimate of $p(\theta^{t-1}|z^{t-1})$. This pdf can be sampled N times to produce N sample paths of θ^{t-1}. Let these sample pathes be denoted by $\{\theta^{t-1}(i), i = 1, \ldots, N\}$. Next, for each $i = 1, \ldots, N$, the pdf $p(\theta_t|\theta_{t-1}(i))$ can be sampled N times to provide $\{\theta_t(1), \ldots, \theta_t(N)\}$. Augmenting the former and latter samples, we obtain N sample paths of θ^t given z^{t-1}. We denote these sample paths by $\{\theta^t(i), i = 1 \ldots, N\}$. The empirical distribution of θ^t given z^{t-1} is given by

$$q(\theta^t|z^{t-1}) = \sum_{i=1}^{N} \delta(\theta^t - \theta^t(i)) \tag{4.6}$$

where $\delta(\cdot)$ denotes the Dirac function. Substituting (4.6) for $p(\theta_t|\theta_{t-1})p(\theta^{t-1}|z^{t-1})$ in (4.5) gives

$$p(\theta_t|z^t) \propto \sum_{i=1}^{N} p(z_t|\theta^t(i), z^{t-1})\delta(\theta_t - \theta_t(i)). \tag{4.7}$$

Next, it was observed that Z_t given $\theta^t(i)$ and z^{t-1} is Gaussian with conditional mean and covariance that can be calculated using the Kalman filter for estimating Y_t given $\theta^t(i)$ and z^{t-1}. Following this procedure, we now have an estimate of $p(\theta_t|z^t)$ which can be resampled to obtain a new estimate of θ^{t+1} and of Y^{t+1} at time $t+1$, and so on. Note that the estimate of the signal is obtained as a by-product in this procedure.

5 Comments

We have reviewed traditional as well as more recent research approaches to enhancement of noisy speech signals. The paper was not intended to be comprehensive but rather to provide a general overview of the area. We have emphasized the methodology and principles of the various approaches, and presented in some more details one design example of a speech enhancement system.

6 Further Reading

The following is a non-comprehensive list of references for further reading on the subject. The edited book by Lim [R1] provides a collection of key papers in the area of speech enhancement. The book by Quatieri [R4] provides extensive background for speech processing including speech enhancement. The National Academy Press Report [R2] details the state of the art of speech enhancement at the time of publication. It also addresses evaluation of speech enhancement systems.

[R1] J. S. Lim, ed., *Speech Enhancement.* Prentice-Hall, Inc, New Jersey, 1983.

[R2] J. Makhoul, T. H. Crystal, D. M. Green, D. Hogan, R. J. McAulay, D. B. Pisoni, R. D. Sorkin, and T. G. Stockham, *Removal of Noise From Noise-Degraded Speech Signals.* Panel on removal of noise from a speech/noise National Research Council, National Academy Press, Washington, D.C., 1989.

[R3] Y. Ephraim, "Statistical model based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526-1555, Oct. 1992.

[R4] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* Prentice Hall, 2001.

[R5] Y. Ephraim, H. Lev-Ari, W. J. J. Roberts, "A Brief Survey of Speech Enhancement," to appear in CRC Electronic Engineering Handbook, 200?.

Defining Terms:

Speech Enhancement: A subject dealing with processing of speech signals, in particular of noisy speech signals, aiming at improving their perception by human or their correct decoding by machines.

Quality: A subjective measure of speech perception reflecting individual preferences of listeners.

Intelligibility: An objective measure which predicts the percentage of spoken words (often meaningless) that can be correctly transcribed.

Statistical model: A set of assumptions, formulated in mathematical terms, on the behavior of many examples of signal and noise samples.

Distortion measure: A mathematical function that quantifies the dissimilarity of two speech signals such as the clean and processed signal.

Signal estimator: A function of the observed noisy signal which approximates the clean signal by minimizing a distortion measure based on a given statistical model.

Wiener filter: An optimal linear signal estimator in the minimum mean squared error sense.

Monte-Carlo Simulation: A statistical approach to develop signal estimators by sampling their statistical model.

Hidden Markov Process: A Markov chain observed through a noisy communication channel.

References

[1] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp,"A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking Signal Processing," *IEEE Trans. Signal Proc.*, vol. 50, pp. 174 -188, Feb. 2002

[2] M. Askar and H. Derin, "A recursive algorithm for the Bayes solution of the smoothing problem," *IEEE Trans. Automatic Control*, vol. 26, pp. 558-561, 1981.

[3] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, pp. 208-211, 1979.

[4] D. R. Brillinger, *Time Series: Data Analysis and Theory.* SIAM, Philadelphia, 2001.

[5] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 345 -349, April 1994.

[6] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403-2418, 2001.

[7] I. Cohen and B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Sig. Proc. Let.*, vol. 9, pp. 12-15, Jan. 2002.

[8] I. Cohen,"Relaxed Statistical Model for Speech Enhancement and *A Priori* SNR Estimation," Technion - Israel Institute of Technology, Technical Report, CCIT No. 443, Oct. 2003.

[9] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 466-475, Sep. 2003.

[10] T. M. Cover and J. A. Thomas, *Elements of information Theory.* John Wiley & Sons, Inc., New york, 1991.

[11] A. Dembo and O. Zeitouni, "Maximum a posteriori estimation of time-varying ARMA processes from noisy observations," *IEEE Trans. on Acoustics, Speech, and Signal Processing* vol. 36, pp. 471 -476, Apr. 1988.

[12] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 165-168, Jun. 1968.

[13] Y. Ephraim and D. Malah, "Signal to noise ratio estimation for enhancing speech using the Viterbi algorithm," Technion, EE Pub. No. 489, Mar. 1984

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error Log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.

[16] Y. Ephraim, D. Malah and B.-H. Juang "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1846-1856, Dec. 1989.

[17] Y. Ephraim, "A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models," *IEEE Trans. Signal Processing*, vol. 40, pp. 725-735, Apr. 1992.

[18] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Proc.*, vol. 3, pp. 251-266, July 1995.

[19] Y. Ephraim and N. Merhav, "Hidden Markov Processes," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1518-1569, June 2002.

[20] W. Fong, S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing with application to audio signal enhancement," *IEEE Trans. Signal Processing*, vol. 50, pp. 438 -449, Feb. 2002

[21] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms," *IEEE Trans. Speech and Audio Proc.*, vol. 6, pp. 373-385, July 1998.

[22] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1991

[23] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, Inc., New York, 2000.

[24] R. M. Gray, *Toeplitz and Circulant Matrices: II.* Stanford Electron. Lab., Tech. Rep. 6504-1, Apr. 1977.

[25] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Sig. Proc. Let.*, vol. 10, pp. 104-106, April 2003.

[26] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586-1604, Dec. 1979.

[27] R. Martin, "Spectral Subtraction Based on Minimum Statistics," *Proc. 7th European Signal Processing Conf., EUSIPCO-94*, pp. 1182-1185, Sept. 1994.

[28] R. Martin, I. Wittke and P. Jax, "Optimized Estimation of Spectral Parameters for the Coding of Noisy Speech," *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, vol. 9, pp. 1479-1482, Jul. 2001.

[29] 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, Jul. 2001.

[30] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," *Proc. 8th Internat. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 8.11, pp. 87.90., Sept. 2003.

[31] R. Martin and D. Malah and R. V. Cox and A. J. Accardi, "A Noise Reduction Preprocessor for Mobile Voice Communication," Technion - Israel Institute of Technology, Technical Report, CCIT No. 459, Dec. 2003.

[32] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing,* ASSP-28, pp. 137-145, Apr. 1980.

[33] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," *IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, vol. 9, pp. 53-56, Mar 1984.

[34] I. Y. Soon, S. N. Koh and C. K. Yeo, "Improved Noise Suppression Filter Using Self-Adaptive Estimator of Probability of Speech Absence, *Signal Processing*, vol. 75, no. 2, pp. 151-159, Jun. 1999.

[35] I. B. Thomas and A. Ravindran, "Intelligibility enhancement of already noisy speech signals," *J. Audio Eng. Soc.*, vol. 22, pp. 234-236, May 1974.

[36] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 173 -185, Mar. 2002

[37] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 679-681, Aug. 1982.