A Brief Survey of Speech Enhancement[1]

Yariv Ephraim, Hanoch Lev-Ari and William J.J. Roberts [2]

August 2, 2003

Abstract

We present a brief overview of the speech enhancement problem for wide-band noise sources that are not correlated with the speech signal. Our main focus is on the spectral subtraction approach and some of its derivatives in the forms of linear and non-linear minimum mean square error estimators. For the linear case, we review the signal subspace approach, and for the non-linear case, we review spectral magnitude and phase estimators. On line estimation of the second order statistics of speech signals using parametric and non-parametric models is also addressed.

1 Introduction

Speech enhancement aims at improving the performance of speech communication systems in noisy environments. Speech enhancement may be applied, for example, to a mobile radio communication system, a speech recognition system, a set of low quality recordings, or to improve the performance of aids for the hearing impaired. The interference source may be a wide-band noise in the form of a white or colored noise, a periodic signal such as in hum noise, room reverberations, or it can take the form of fading noise. The first two examples represent additive noise sources, while the other two examples represent convolutional and multiplicative noise sources, respectively. The speech signal may be simultaneously attacked by more than one noise source.

There are two principal perceptual criteria for measuring the performance of a speech enhancement system. The *quality* of the enhanced signal measures its clarity, distorted

[1] This paper is partially based on "Extension of the Signal Subspace Speech Enhancement Approach to Colored Noise" by Hanoch Lev-Ari and Yariv Ephraim which appeared in *IEEE Sig. Proc. Let.*, vol. 10, pp. 104-106, April 2003 ©2003 IEEE.

[2] Y. Ephraim is with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030. Email: yephraim@gmu.edu
H. Lev-Ari is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115. Email: levari@ECE.NEU.EDU
W. J. J. Roberts is with Atlantic Coast Technologies, Inc., 11499 Columbia Pike, Suite 200, Silver Spring, MD 20904. Email: robertsb@con2inc.com

nature, and the level of residual noise in that signal. The quality is a *subjective* measure that is indicative of the extent to which the listener is comfortable with the enhanced signal. The second criterion measures the *intelligibility* of the enhanced signal. This is an *objective* measure which provides the percentage of words that could be correctly identified by listeners. The words in this test need not be meaningful. The two performance measures are not correlated. A signal may be of good quality and poor intelligibility and vice versa. Most speech enhancement systems improve the quality of the signal at the expense of reducing its intelligibility. Listeners can usually extract more information from the noisy signal than from the enhanced signal by careful listening to that signal. This is obvious from the *data processing theorem* of information theory. Listeners, however, experience fatigue over extended listening sessions, a fact that results in reduced intelligibility of the noisy signal. Is such situations, the intelligibility of the enhanced signal may be higher than that of the noisy signal. Less effort would usually be required from the listener to decode portions of the enhanced signal that correspond to high signal to noise ratio segments of the noisy signal.

Both the quality and intelligibility are elaborate and expensive to measure, since they require listening sessions with live subjects. Thus, researchers often resort to less formal listening tests to assess the quality of an enhanced signal, and they use automatic speech recognition tests to assess the intelligibility of that signal. Quality and intelligibility are also hard to quantify and express in a closed form that is amenable to mathematical optimization. Thus, the design of speech enhancement systems is often based on mathematical measures that are somehow believed to be correlated with the quality and/or intelligibility of the speech signal. A popular example involves estimation of the clean signal by minimizing the mean square error (MSE) between the logarithms of the spectra of the original and estimated signals [5]. This criterion is believed to be more perceptually meaningful than the minimization of the MSE between the original and estimated signal waveforms [13].

Another difficulty in designing efficient speech enhancement systems is the lack of *explicit* statistical models for the speech signal and noise process. In addition, the speech signal, and possibly also the noise process, are not strictly stationary processes. Common parametric models for speech signals, such as an autoregressive process for short-term modeling of the signal, and a hidden Markov process (HMP) for long-term modeling of the signal, have not provided adequate models for speech enhancement applications. A variant of the *expectation-maximization* (EM) algorithm, for maximum likelihood (ML) estimation of the autoregressive parameter from a noisy signal, was developed by Lim and Oppenheim [12] and tested in speech enhancement. Several estimation schemes, which are based on hidden Markov modeling of the clean speech signal and of the noise process, were developed over the years, see, e.g., Ephraim [6]. In each case, the HMP's for the speech signal and noise process were designed from training sequences from the two processes, respectively. While autoregressive and hidden Markov models have proved extremely useful in coding and recognition of *clean* speech signals, respectively, they were not found to be sufficiently refined models for speech enhancement applications.

In this paper we review some common approaches to speech enhancement that were developed primarily for additive wide-band noise sources. Although some of these approaches have been applied to reduction of reverberation noise, we believe that the dereverberation problem requires a completely different approach that is beyond the scope of this paper. Our primary focus is on the spectral subtraction approach [13] and some of its derivatives such as the signal subspace approach [7], and the estimation of the short-term spectral magnitude [16], [4]-[5]. This choice is motivated by the fact that some derivatives of the spectral subtraction approach are still the best approaches available to date. These approaches are relatively simple to implement and they usually outperform more elaborate approaches which rely on parametric statistical models and training procedures.

The plan for this paper is as follows. In Section 2, we present the principles of the signal subspace approach which constitutes a finite-dimensional version of the spectral subtraction approach. In Section 3, we focus on short-time spectral estimation of the Fourier transform's magnitude and phase. In Section 4 we discuss some aspects of signal presence uncertainty in the noisy signal and its relation to the hidden Markov model based speech enhancement approach. In Section 5 we address estimation issues of the second-order statistics required for implementation of the various estimators described in this paper. A few concluding remarks are given in Section 6.

2 The Signal Subspace Approach

In this section we present the principles of the signal subspace approach and its relations to Wiener filtering and spectral subtraction. Our presentation follows [7] and [11]. This approach assumes that the signal and noise are non-correlated, and that their second-order statistics are available. It makes no assumptions about the distributions of the two processes.

Let Y and W be k-dimensional random vectors in a Euclidean space $\mathcal{R}^k$ representing the clean signal and noise, respectively. Assume that the expected value of each vector is zero in an appropriately defined probability space. Let $Z = Y + W$ denote the noisy vector. Let R_y and R_w denote the covariance matrices of the clean signal and noise process, respectively. Assume that R_w is positive definite. Let H denote a $k \times k$ real matrix in the linear space $\mathcal{R}^{k\times k}$, and let $\hat{Y} = HZ$ denote the linear estimator of Y given Z. The residual signal in this estimation is given by

$$Y - \hat{Y} = (I - H)Y - HW \tag{2.1}$$

where I denotes, as usual, the identity matrix. To simplify notation, we shall not explicitly indicate the dimensions of the identity matrix. These dimensions should be clear from the context. In (2.1), $D = (I - H)Y$ is the signal distortion and $N = HW$ is the residual noise in the linear estimation. Let $(\cdot)'$ denote the transpose of a real matrix or the conjugate transpose of a complex matrix. Let

$$\overline{\epsilon_d^2} = \frac{1}{k}\text{tr}E\{DD'\} = \frac{1}{k}\text{tr}\{(I-H)R_y(I-H)'\} \tag{2.2}$$

denote the average signal distortion power where tr$\{\cdot\}$ denotes the trace of a matrix. Similarly, let

$$\overline{\epsilon_n^2} = \frac{1}{k}\text{tr}E\{NN'\} = \frac{1}{k}\text{tr}\{HR_wH'\} \tag{2.3}$$

denote the average residual noise power.

The matrix H is estimated by minimizing the signal distortion $\overline{\epsilon_d^2}$ subject to a threshold on the residual noise power $\overline{\epsilon_n^2}$. It is obtained from

$$\begin{array}{c} \min_H \overline{\epsilon_d^2} \\ \text{subject to} : \overline{\epsilon_n^2} \leq \alpha \end{array} \tag{2.4}$$

for some given α. Let $\mu \geq 0$ denote the Lagrange multiplier of the inequality constraint. The optimal matrix, say $H = H_1$, is given by

$$H_1 = R_y(R_y + \mu R_w)^{-1}. \tag{2.5}$$

The matrix H_1 can be implemented as follows. Let $R_w^{1/2}$ denote the symmetric positive definite square root of R_w and let $R_w^{-1/2} = (R_w^{1/2})^{-1}$. Let U denote an orthogonal matrix of eigenvectors of the symmetric matrix $R_w^{-1/2}R_yR_w^{-1/2}$. Let $\Lambda = \text{diag}[\lambda_1, \ldots, \lambda_k]$ denote the diagonal matrix of non-negative eigenvalues of $R_w^{-1/2}R_yR_w^{-1/2}$. Then

$$H_1 = R_w^{1/2}U\Lambda(\Lambda + \mu I)^{-1}U'R_w^{-1/2}. \tag{2.6}$$

When H_1 in (2.6) is applied to Z, it first whitens the input noise by applying $R_w^{-1/2}$ to Z. Then, the orthogonal transformation U' corresponding to the covariance matrix of the whitened clean signal is applied, and the transformed signal is modified by a diagonal Wiener-type gain matrix.

In (2.6), components of the whitened noisy signal that contain noise only are nulled. The indices of these components are given by $\{j : \lambda_j = 0\}$. When the noise is white, $R_w = \sigma_w^2 I$, and U and Λ are the matrices of eigenvectors and eigenvalues of R_y/σ_w^2, respectively. The existence of null components $\{j : \lambda_j = 0\}$ for the signal means that the signal lies in a subspace of the Euclidean space $\mathcal{R}^k$. At the same time, the eigenvalues of the noise are all equal to σ_w^2 and the noise occupies the entire space $\mathcal{R}^k$. Thus, the signal subspace approach first eliminates the noise components outside the signal subspace and then modifies the signal components inside the signal subspace in accordance with the criterion (2.4).

When the signal and noise are *wide-sense stationary*, the matrices R_y and R_w are Toeplitz with associated power spectral densities of $f_y(\theta)$ and $f_w(\theta)$, respectively. The angular frequency θ lies in $[0, 2\pi)$. When the signal and noise are *asymptotically weakly stationary*, the matrices R_y and R_w are asymptotically Toeplitz and have the associated power spectral densities $f_y(\theta)$ and $f_w(\theta)$, respectively [10]. Since the latter represents a somewhat more general situation, we proceed with asymptotically weakly stationary signal and noise. The filter H_1 in (2.5) is then asymptotically Toeplitz with associated power spectral density

$$h_1(\theta) = \frac{f_y(\theta)}{f_y(\theta) + \mu f_w(\theta)}. \tag{2.7}$$

This is the noncausal Wiener filter for the clean signal with an adjustable noise level determined by the constraint α in (2.4). This filter is commonly implemented using estimates of the two power spectral densities. Let $\hat{f}_y(\theta)$ and $\hat{f}_w(\theta)$ denote the estimates of $f_y(\theta)$ and $f_w(\theta)$, respectively. These estimates could, for example, be obtained from the periodogram or the smoothed periodogram. In that case, the filter is implemented as

$$\hat{h}_1(\theta) = \frac{\hat{f}_y(\theta)}{\hat{f}_y(\theta) + \mu \hat{f}_w(\theta)}. \tag{2.8}$$

When $\hat{f}_y(\theta)$ is implemented as

$$\hat{f}_y(\theta) = \begin{cases} \hat{f}_z(\theta) - \hat{f}_w(\theta) & \text{if non-negative} \\ \varepsilon & \text{otherwise} \end{cases} \tag{2.9}$$

then a *spectral subtraction* estimator for the clean signal results. The constant $\varepsilon \geq 0$ is often referred to as a "spectral floor." Usually $\mu \geq 2$ is chosen for this estimator.

The enhancement filter H could also be designed by imposing constraints on the spectrum of the residual noise. This approach enables shaping of the spectrum of the residual noise to minimize its perceptual effect. Suppose that a set $\{v_i, i = 1, \ldots, m\}$, $m \leq k$, of

k-dimensional real or complex orthonormal vectors, and a set $\{\alpha_i, i = 1, \ldots, m\}$ of nonnegative constants, are chosen. The vectors $\{v_i\}$ are used to transform the residual noise into the spectral domain, and the constants $\{\alpha_i\}$ are used as upper bounds on the variances of these spectral components. The matrix H is obtained from

$$\begin{array}{c} \min_H \overline{\epsilon_d^2} \\ \text{subject to}: \;\; E\{|v_i' N|^2\} \;\; \leq \;\; \alpha_i, \quad i = 1, \ldots, m. \end{array} \tag{2.10}$$

When the noise is white, the set $\{v_i\}$ could be the set of eigenvectors of R_y and the variances of the residual noise along these coordinate vectors are constrained. Alternatively, the set $\{v_i\}$ could be the set of orthonormal vectors related to the DFT. These vectors are given by $v_i' = k^{-1/2}(1, e^{-j\frac{2\pi}{k}(i-1)\cdot 1}, \ldots, e^{-j\frac{2\pi}{k}(i-1)\cdot(k-1)})$. Here we must choose $\alpha_i = \alpha_{k-i+2}, i = 2, \ldots, k/2$, assuming k is even, for the residual noise power spectrum to be symmetric. This implies that at most $k/2 + 1$ constraints can be imposed. The DFT-related $\{v_i\}$ enable the use of constraints that are consistent with auditory perception of the residual noise.

To present the optimal filter, let e_l denote a unit vector in $\mathcal{R}^k$ for which the lth component is one and all other components are zero. Let $V = [v_1, \ldots, v_k]$. Define $Q = R_w^{-1/2}U$ and $T = Q'V$. Let $M = \text{diag}[k\mu_1, \ldots, k\mu_k]$ denote the matrix of k times the Lagrange multipliers which are assumed nonnegative. The optimal estimation matrix, say $H = H_2$, is given by [11]

$$H_2 = R_w^{1/2} U \tilde{H}_2 U' R_w^{-1/2} \tag{2.11}$$

where the columns of $\tilde{H}_2$ are given by

$$\tilde{h}_l = T\lambda_l(M + \lambda_l I)^{-1} T^{-1} e_l, \qquad l = 1, \ldots, k. \tag{2.12}$$

The optimal estimator first whitens the noise, then applies the orthogonal transformation U' obtained from eigendecomposition of the covariance matrix of the whitened signal, and then modifies the resulting components using the matrix $\tilde{H}_2$. This is analogous to the operation of the estimator H_1 in (2.6). The matrix $\tilde{H}_2$, however, is not diagonal when the input noise is colored.

When the noise is white and $m = k$ is chosen, the optimization problem (2.10) becomes trivial since knowledge of input and output noise variances uniquely determines the filter H. This filter is given by $H = UAU'$ where $A = \text{diag}[\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_k}]$ [7]. For this case, the heuristic choice of

$$\alpha_i = \exp\{-\nu/\lambda_i\}, \tag{2.13}$$

where $\nu \geq 1$ is an experimental constant, was found useful in practice [7]. This choice is motivated by the observation that for $\nu = 2$, the first order Taylor expansion of $\alpha_i^{-1/2}$ leads to an estimator $H = UAU'$ which coincides with the Wiener estimation matrix (2.6) with $\mu = 1$. The estimation matrix using (2.13) performs significantly better than the Wiener filter in practice.

3 Short-Term Spectral Estimation

In another earlier approach for speech enhancement, the short-time spectral magnitude of the clean signal is estimated from the noisy signal. The speech signal and noise process are assumed statistically independent, and spectral components of each of these two processes are assumed zero mean statistically independent Gaussian random variables. Let $A_y e^{j\theta_y}$ denote a spectral component of the clean signal Y in a given frame. Let $A_z e^{j\theta_z}$ denote the corresponding spectral component of the noisy signal. Let $\sigma_y^2 = E\{A_y^2\}$ and $\sigma_z^2 = E\{A_z^2\}$ denote, respectively, the variances of the clean and noisy spectral components. If the variance of the corresponding spectral component of the noise process in that frame is denoted by σ_w^2, then we have $\sigma_z^2 = \sigma_y^2 + \sigma_w^2$. Let

$$\xi = \frac{\sigma_y^2}{\sigma_w^2}; \qquad \gamma = \frac{A_z^2}{\sigma_w^2}; \qquad \vartheta = \frac{\xi}{\xi + 1}\gamma \tag{3.1}$$

The MMSE estimation of A_y from $A_z e^{j\theta_z}$ is given by [4]

$$\hat{A}_y = \Gamma(1.5)\frac{\sqrt{\vartheta}}{\gamma}\exp\left(-\frac{\vartheta}{2}\right)\left[(1+\vartheta)I_0\left(\frac{\vartheta}{2}\right) + \vartheta I_1\left(\frac{\vartheta}{2}\right)\right]A_z \tag{3.2}$$

where $\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$, and $I_0(\cdot)$ and $I_1(\cdot)$ denote, respectively, the modified Bessel functions of the zeroth and first order. Similarly to the Wiener filter (2.5), this estimator requires knowledge of second order statistics of each signal and noise spectral components, σ_y^2 and σ_w^2, respectively.

To form an estimator for the spectral component of the clean signal, the spectral magnitude estimator (3.2) is combined with an estimator of the complex exponential of that component. Let $\widehat{e^{j\theta_y}}$ be an estimator of $e^{j\theta_y}$. This estimator is a function of the noisy spectral component $A_z e^{j\theta_z}$. MMSE estimation of the complex exponential $e^{j\theta_y}$ is obtained from

$$\begin{array}{c} \min_{\widehat{e^{j\theta_y}}} E\{|e^{j\theta_y} - \widehat{e^{j\theta_y}}|^2\} \\ \text{subject to} : |\widehat{e^{j\theta_y}}| = 1. \end{array} \tag{3.3}$$

The constraint in (3.3) ensures that the estimator $\widehat{e^{j\theta_y}}$ does not effect the optimality of the estimator $\hat{A}_y$ when the two are combined. The constrained minimization problem in (3.3) results in the estimator $\widehat{e^{j\theta_y}} = e^{j\theta_z}$ which is simply the complex exponential of the noisy signal.

Note that the Wiener filter (2.8) has zero phase and hence it effectively uses the complex exponential of the noisy signal $e^{j\theta_z}$ in estimating the clean signal spectral component. Thus, both the Wiener estimator (2.8) and the MMSE spectral magnitude estimator (3.2) use the complex exponential of the noisy phase. The spectral magnitude estimate of the clean signal obtained by the Wiener filter, however, is not optimal in the MMSE sense.

Other criteria for estimating A_y could also be used. For example, A_y could be estimated from

$$\min_{\hat{A}_y} E\{(\log A_y - \log \hat{A}_y)^2\}. \tag{3.4}$$

This criterion aims at producing an estimate of A_y whose logarithm is as close as possible to the logarithm of A_y in the MMSE sense. This perceptually motivated criterion results in the estimator given by [5]

$$\hat{A}_y = \frac{\sigma_y^2}{\sigma_y^2 + \sigma_w^2} \exp\left(\frac{1}{2} \int_{\vartheta}^{\infty} \frac{e^{-t}}{t} dt \right) A_z. \tag{3.5}$$

The integral in (3.5) is the well know exponential integral of ϑ and it can be numerically evaluated.

In another example, A_y^2 could be estimated from

$$\min_{\hat{A}_y} E\{(A_y^2 - \widehat{A_y^2})^2\} \tag{3.6}$$

and an estimate of A_y could be obtained from $\sqrt{\widehat{A_y^2}}$. The criterion in (3.6) aims at estimating the magnitude squared of the spectral component of the clean signal in the MMSE sense. This estimator is particularly useful when subsequent processing of the enhanced signal is performed, for example, in autoregressive analysis for low bit rate signal coding applications [13]. In that case, an estimator of the autocorrelation function of the clean signal can be obtained from the estimator $\widehat{A_y^2}$. The optimal estimator in the sense of (3.6) is well known and is given by

$$\widehat{A_y^2} = \frac{\sigma_y^2 \sigma_w^2}{\sigma_y^2 + \sigma_w^2} + \left| \frac{\sigma_y^2}{\sigma_y^2 + \sigma_w^2} A_z \right|^2. \tag{3.7}$$

4 Multiple State Speech Model

All estimators presented in Sections 2 and 3 make the implicit assumption that the speech signal is always present in the noisy signal. In the notation of Section 2, $Z = Y + W$. Since the length of a frame is relatively short, in the order of $30 - 50$ msec, it is more realistic to assume that speech may be present in the noisy signal with some probability, say η, and may be absent from the noisy signal with one minus that probability. Thus we have two hypotheses, one of speech presence, say H_1, and the other of speech absence, say H_0, that occur with probabilities η and $1 - \eta$, respectively. We have

$$Z = \begin{cases} Y + W & \text{under } H_1 \\ W & \text{under } H_0 \end{cases} \tag{4.1}$$

The MMSE estimator of A_y under the uncertainty model can be shown to be

$$\begin{aligned} \tilde{A}_y &= \Pr(H_1|Z) \cdot E\{A_y|Z, H_1\} + \Pr(H_0|Z) \cdot E\{A_y|Z, H_0\} \\ &= \Pr(H_1|Z) \cdot \hat{A}_y \end{aligned} \tag{4.2}$$

since $E\{A_y|Z, H_0\} = 0$ and $E\{A_y|Z, H_1\} = \hat{A}_y$ as given by (3.2). The more realistic estimator (4.2) was found useful in practice as it improved the performance of the estimator (3.2). Other estimators may be derived under this model. Note that the model (4.1) is not applicable to the estimator (3.5) since A_y must be positive for the criterion (3.4) to be meaningful. Speech enhancement under the speech presence uncertainty model was first proposed and applied by McAulay and Malpass in their pioneering work [16].

An extension of this model leads to the assumption that speech vectors may be in different states at different time instants. The speech presence uncertainty model assumes two states representing speech presence and speech absence. In another model of Drucker [3], five states were proposed representing fricative, stop, vowel, glide, and nasal speech sounds. A speech absence state could be added to that model as a sixth state. This model requires an estimator for each state just as in (4.2).

A further extension of these ideas, in which multiple states that evolve in time are possible, is obtained when one models the speech signal by a hidden Markov process (HMP) [8]. An HMP is a bivariate random process of states and observations sequences. The state process $\{S_t, t = 1, 2, \ldots\}$ is a finite-state homogeneous Markov chain that is not directly observed. The observation process $\{Y_t, t = 1, 2, \ldots\}$ is conditionally independent given the state process. Thus, each observation depends statistically only on the state of the Markov chain at the same time and not on any other states or observations. Consider, for example, an HMP observed in an additive white noise process $\{W_t, t = 1, 2, \ldots\}$. For each t, let $Z_t = Y_t + W_t$ denote the noisy signal. Let $Z^t = \{Z_1, \ldots, Z_t\}$. Let J denote the number of states of the Markov chain. The causal MMSE estimator of Y_t given $\{Z^t\}$ is given by [6]

$$\hat{Y}_t = E\{Y_t|Z^t\}$$

$$= \sum_{j=1}^{J} \Pr(S_t = j|Z^t) E\{Y_t|S_t = j, Z_t\}. \tag{4.3}$$

The estimator (4.3) reduces to (4.2) when $J = 2$ and $\Pr(S_t = j|Z^t) = \Pr(S_t = j|Z_t)$, or when the states are statistically independent of each other.

An HMP is a parametric process that depends on the initial distribution and transition matrix of the Markov chain and on the parameter of the conditional distributions of observations given states. The parameter of an HMP can be estimated off-line from training data and then used in constructing the estimator (4.3). This approach has a great theoretical appeal since it provides a solid statistical model for speech signals. It also enjoys a great intuitive appeal since speech signals do cluster into sound groups of distinct nature, and dedication of a filter for each group is appropriate. The difficulty in implementing this approach is in achieving low probability of error in mapping vectors of the noisy speech onto states of the HMP. Decoding errors result in wrong association of speech vectors with the set of predesigned estimators $\{E\{Y_t|S_t = j, Z_t\}, j = 1, \ldots, J\}$, and thus in poorly filtered speech vectors. In addition, the complexity of the approach grows with the number of states, since each vector of the signal must be processed by all J filters.

The approach outlined above could be applied based on other models for the speech signal. For example, in [20], a harmonic process based on an estimated pitch period was used to model the speech signal.

5 Second-Order Statistics Estimation

Each of the estimators presented in Sections 2 and 3 depends on some statistics of the clean signal and noise process which are assumed known a-priori. The signal subspace estimators (2.5) and (2.11) require knowledge of the covariance matrices of the signal and noise. The spectral magnitude estimators (3.2), (3.5) and (3.7) require knowledge of the variance of each spectral component of the speech signal and of the noise process. In the absence of explicit knowledge of the second-order statistics of the clean signal and noise process, these statistics must be estimated either from training sequences or directly from the noisy signal.

We note that when estimates of the second-order statistics of the signal and noise replace the true second-order statistics in a given estimation scheme, the optimality of that scheme can no longer be guaranteed. The quality of the estimates of these second-order statistics is key for the overall performance of the speech enhancement system. Estimation of the second-order statistics can be performed in various ways usually outlined in the theory of spectral estimation [19]. Some of these approaches are reviewed in this section.

Estimation of speech signals second-order statistics from training data has proven successful in coding and recognition of clean speech signals. This is commonly done in coding applications using vector quantization and in recognition applications using hidden Markov modeling. When only noisy signals are available, however, matching of a given speech frame to a codeword of a vector quantizer or to a state of an HMP is susceptible to decoding errors. The significance of these errors is the creation of a mismatch between the true and estimated second order statistics of the signal. This mismatch results in application of wrong filters to frames of the noisy speech signal, and in unacceptable quality of the processed noisy signal.

On-line estimation of the second-order statistics of a speech signal from a sample function of the noisy signal has proven to be a better choice. Since the analysis frame length is usually relatively small, the covariance matrices of the speech signal may be assumed Toeplitz. Thus, the autocorrelation function of the clean signal in each analysis frame must be estimated. The Fourier transform of a windowed autocorrelation function estimate provides estimates of the variances of the clean signal spectral components.

For a wide-sense stationary noise process, the noise autocorrelation function can be estimated from an initial segment of the noisy signal that contains noise only. If the noise process is not wide-sense stationary, frames of the noisy signal must be classified as in (4.1), and the autocorrelation function of the noise process must be updated whenever a new noise frame is detected.

In the signal subspace approach [7], the power spectral densities of the noise and noisy processes are first estimated using windowed autocorrelation function estimates. Then, the power spectral density of the clean signal is estimated using (2.9). That estimate is inverse Fourier transformed to produce an estimate of the desired autocorrelation function of the clean signal. In implementing the MMSE spectral estimator (3.2), a recursive estimator for the variance of each spectral component of the clean signal developed in [4] is often used, see, e.g., [1], [2]. Let $\hat{A}_y(t)$ denote an estimator of the magnitude of a spectral component of the clean signal in a frame at time t. This estimator may be the MMSE estimator (3.2). Let $A_z(t)$ denote the magnitude of the corresponding spectral component of the noisy signal. Let $\widehat{\sigma_w^2}(t)$ denote the estimated variance of the spectral component of the noise process in that frame. Let $\widehat{\sigma_y^2}(t)$ denote the estimated variance of the spectral component of the clean signal in that frame. This estimator is given by

$$\widehat{\sigma_y^2}(t) = \beta \hat{A}_y^2(t-1) + (1-\beta)\max\{A_z^2(t) - \sigma_w^2(t), 0\} \tag{5.1}$$

where $0 \leq \beta \leq 1$ is an experimental constant. We note that while this estimator was found useful in practice, it is heuristically motivated and its analytical performance is not known.

A parametric approach for estimating the power spectral density of the speech signal from a given sample function of the noisy signal was developed by Musicus and Lim [18]. The clean signal in a given frame is assumed a zero mean Gaussian autoregressive process. The noise is assumed a zero mean white Gaussian process that is independent of the signal. The parameter of the noisy signal comprises the autoregressive coefficients, the gain of the autoregressive process, and the variance of the noise process. This parameter is estimated in the ML sense using the EM algorithm. The EM algorithm and conditions for its convergence were originally developed for this problem by Musicus [17]. The approach starts with an initial estimate of the parameter. This estimate is used to calculate the conditional mean estimates of the clean signal and of the sample covariance matrix of the clean signal given the noisy signal. These estimates are then used to produce a new parameter, and the process is repeated until a fixed point is reached or a stopping criterion is met.

This EM procedure is summarized as follows. Consider a scalar autoregressive process $\{Y_t, t = 1, 2, \ldots\}$ of order r and coefficients $a = (a_1, \ldots, a_r)'$. Let $\{V_t\}$ denote an independent identically distributed (iid) sequence of zero mean unit variance Gaussian random variables. Let σ_v denote a gain factor. A sample function of the autoregressive process is given by

$$y_t = -\sum_{i=1}^{r} a_i y_{t-i} + \sigma_v v_t. \tag{5.2}$$

Assume that the initial conditions of (5.2) are zero, i.e., $y_t = 0$ for $t < 1$. Let $\{W_t, t = 1, 2, \ldots\}$ denote the noise process which comprises an iid sequence of zero mean unit variance Gaussian random variables. Consider a k-dimensional vector of the noisy autoregressive process. Let $Y = (Y_1, \ldots, Y_k)'$, and define similarly the vectors V and W corresponding to the processes $\{V_t\}$ and $\{W_t\}$, respectively. Let A denote a lower triangular Toeplitz matrix with the first column given by $(1, a_1, \ldots, a_r, 0, \ldots, 0)'$. Then

$$Z = \sigma_v A^{-1} V + \sigma_w W. \tag{5.3}$$

Suppose $\phi_m = (a(m), \sigma_v(m), \sigma_w(m))$ denotes an estimate of the parameter of Z at the end of the mth iteration. Let $A(m)$ denote the matrix A constructed from the vector $a(m)$. Let $R_y(m) = \sigma_v^2(m) A(m)^{-1} (A(m))^{-1})'$ denote the covariance matrix of the autoregressive process (5.2) as obtained from ϕ_m. Let $R_z(m) = R_y(m) + \sigma_w^2(m) I$ denote the covariance matrix of the noisy signal (5.3) based on ϕ_m. Let $\hat{Y}$ and $\hat{R}$ denote, respectively, the conditional mean estimates of the clean signal Y and of the sample covariance of Y given Z and the current estimate ϕ_m. Under the Gaussian assumptions of this problem, we have similarly to (2.5) and (3.7)

$$\hat{Y} \;=\; E\{Y|Z; \phi_m\}$$

$$= \quad R_y(m)R_z^{-1}(m)Z \tag{5.4}$$

$$\begin{aligned} \hat{R} &= E\{YY'|Z;\phi_m\} \\ &= R_y(m) - R_y(m)R_z^{-1}(m)R_y(m) + \hat{Y}\hat{Y}'. \end{aligned} \tag{5.5}$$

The estimation of the statistics of the clean signal in (5.4) and (5.5) comprise the E-step of the EM algorithm.

Define the $r+1 \times r+1$ covariance matrix S with entries

$$S(i,j) = \sum_{l=\max(i,j)}^{k-1} \hat{R}(l-i, l-j), \qquad i,j = 0,\ldots,r. \tag{5.6}$$

Define the $r \times 1$ vector $q = (S(1,0),\ldots,S(r,0))'$ and the $r \times r$ covariance matrix $Q = \{S(i,j), i,j = 1,\ldots,r\}$. The new estimate of parameter ϕ at the end of the $m+1$ iteration is given by

$$a(m+1) = -Q^{-1}q \tag{5.7}$$

$$\sigma_v^2(m+1) = [S(0,0) + 2a'(m)q + a'(m)Qa(m)]/k \tag{5.8}$$

$$\begin{aligned} \sigma_w^2(m+1) &= \text{tr } [E\{(Z-Y)(Z-Y)'|Z,\phi_m\}]/k \\ &= \text{tr } [ZZ' - 2\hat{Y}Z' + \hat{R}]/k. \end{aligned} \tag{5.9}$$

The calculation of the parameter of the noisy signal in (5.7)-(5.9) comprise the M-step of the EM algorithm.

Note that the enhanced signal can be obtained as a by product of this algorithm using the estimator (5.4) at the last iteration of the algorithm. Formulation of the above parameter estimation problem in a state space assuming colored noise in the form of another autoregressive process, and implementation of the estimators (5.4) and (5.5) using Kalman filtering and smoothing was done in [9].

The EM approach presented above differs from an earlier approach pioneered by Lim and Oppenheim [12]. Assume for this comparison that the noise variance is known. In the EM approach, the goal is to derive the ML estimate of the true parameter of the autoregressive process. This is done by local maximization of the likelihood function of the noisy signal over the parameter set of the process. The EM approach results in alternate estimation of the clean signal Y and its sample correlation matrix YY', and of the parameter of the autoregressive process (a, σ_v). The approach taken in [12] aims at estimating the parameter of the autoregressive process that maximizes the joint likelihood function of the clean and noisy signals. Using alternate maximization of the joint likelihood function, the approach resulted in alternate estimation of the clean signal Y and of the parameter of the autoregressive process (a, σ_v). Thus, the main difference between the two approaches is in the estimated statistics from which the autoregressive parameter is estimated in each iteration. This difference impacts the convergence properties of the algorithm [12] which is known to produce an inconsistent parameter estimate. The algorithm [12] is simpler to implement than the EM approach and it is popular among some authors. To overcome the inconsistency, they developed a set of constraints for the parameter estimate in each iteration.

6 Concluding Remarks

We have surveyed some aspects of the speech enhancement problem and presented state of the art solutions to this problem. In particular, we have identified the difficulties inherent to speech enhancement, and presented statistical models and distortion measures commonly used in designing estimators for the clean signal. We have mainly focused on speech signals degraded by additive non-correlated wide-band noise. As we have noted earlier, even for this case, a universally accepted solution is not available and more research is required to refine current approaches or alternatively develop new ones. Other noise sources, such as room reverberation noise, present a more significant challenge as the noise is a non-stationary process that is correlated with the signal and it can not be easily modelled. The speech enhancement problem is expected to attract significant research effort in the future due to challenges that this problem poses, the numerous potential applications, and future advances in computing devices.

References

[1] 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.

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

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

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

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

[11] 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.

[12] J. S. Lim and A. V. Oppenheim, “All-pole modeling of degraded speech,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197-210, June 1978.

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

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

[15] 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.

[16] 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.

[17] B. R. Musicus, *An Iterative Technique for Maximum Likelihood Parameter Estimation on Noisy Data.* S.M. Thesis, M.I.T., Cambridge, Massachusetts, 1979.

[18] B. R. Musicus and J. S. Lim, “Maximum likelihood parameter estimation of noisy data,” *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, pp. 224-227, 1979.

[19] M. B. Priestley, *Spectral Analysis and Time Series*, Academic Press, London, 1989.

[20] T. F. Quatieri and R. J. McAulay, “Noise reduction using a soft-decision sin-wave vector quantizer,” *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 821-824, 1990.

Defining Terms:

Speech Enhancement: The action of improving perceptual aspects of a given sample of speech signals.

Quality: A subjective measure of the way a speech signal is perceived.

Intelligibility: An objective measure which indicates the percentage of words from a given text that are expected to be correctly understood by listeners.

Signal estimator: A function of the observed noisy signal which approximates the clean signal.

Expectation-maximization: An iterative approach for parameter estimation using alternate estimation and maximization steps.

Autoregressive process: A random process obtained by passing white noise through an all-pole filter.

Wide-sense stationarity: A property of a random process whose second-order statistics do not change with time.

Asymptotically weakly stationarity: A property of a random process indicating eventual wide-sense stationarity.

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