Simultaneous Detection and Estimation Approach for Speech Enhancement

Ari Abramson, *Student Member, IEEE*, and Israel Cohen, *Senior Member, IEEE*

***Abstract*—In this paper, we present a simultaneous detection and estimation approach for speech enhancement. A detector for speech presence in the short-time Fourier transform domain is combined with an estimator, which jointly minimizes a cost function that takes into account both detection and estimation errors. Cost parameters control the tradeoff between speech distortion, caused by missed detection of speech components and residual musical noise resulting from false-detection. Furthermore, a modified decision-directed *a priori* signal-to-noise ratio (SNR) estimation is proposed for transient-noise environments. Experimental results demonstrate the advantage of using the proposed simultaneous detection and estimation approach with the proposed *a priori* SNR estimator, which facilitate suppression of transient noise with a controlled level of speech distortion.**

***Index Terms*—Acoustic noise, estimation, signal detection, siren noise, spectral analysis, speech enhancement.**

I. Introduction

OPTIMAL design of efficient speech enhancement algorithms has attracted significant research effort for several decades. Speech enhancement systems often operate in the short-time Fourier transform (STFT) domain, where the speech spectral coefficients are estimated from the spectral coefficients of the degraded signal. The spectral coefficients of the speech signal are generally sparse in the STFT domain in the sense that speech is present only in some of the frames, and in each frame only some of the frequency-bins contain the significant part of the signal energy. However, existing algorithms often focus on estimating the spectral coefficients rather than detecting their existence. The spectral-subtraction algorithm [1], [2] contains an elementary detector for speech activity in the time–frequency domain, but it generates musical noise caused by falsely detecting noise peaks as bins that contain speech, which are randomly scattered in the STFT domain. Subspace approaches for speech enhancement [3]–[6] decompose the vector of the noisy signal into a signal-plus-noise subspace and a noise subspace, and the speech spectral coefficients are estimated after removing the noise subspace. Accordingly, these algorithms are aimed at detecting the speech coefficients and subsequently estimating their values. McAulay and Malpass [7] were the first to propose a speech spectral estimator under a two-state model. They derived a maximum-likelihood (ML) estimator for the speech spectral amplitude under speech-presence uncertainty. Ephraim and Malah followed this approach of signal estimation under speech presence uncertainty and derived an estimator which minimizes the mean-square error (MSE) of the short-term spectral amplitude (STSA) [8]. In [9], speech presence probability is evaluated to improve the minimum MSE (MMSE) of the log-spectral amplitude (LSA) estimator, and in [10] a further improvement of the MMSE-LSA estimator is achieved based on a two-state model. Under speech absence hypothesis, Cohen and Berdugo [10] considered a constant attenuation factor to enable a more natural residual noise, characterized by reduced musicality.

Under slowly time-varying noise conditions, an estimator which minimizes the MSE of the STSA or the LSA under speech presence uncertainty may yield reasonable results [8], [10]. However, under quickly time-varying noise conditions, abrupt transients may not be sufficiently attenuated, since speech is falsely detected with some positive probability. Reliable detectors for speech activity and noise transients are necessary to further attenuate noise transients without much degrading the speech components [11], [12]. Despite the sparsity of speech coefficients in the time–frequency domain and the importance of signal detection for noise suppression performance, common speech enhancement algorithms deal with speech detection *independently* of speech estimation. Even when a voice activity detector is available in the STFT domain (e.g., [13]–[19]), it is not straightforward to consider the detection errors when designing the optimal speech estimator. High attenuation of speech spectral coefficients due to missed detection errors may significantly degrade speech quality and intelligibility, while falsely detecting noise transients as speech-contained bins, may produce annoying musical noise.

In this paper, we present a novel formulation of the speech enhancement problem, which incorporates simultaneous operations of detection and estimation. A detector for the speech coefficients is combined with an estimator, which jointly minimizes a cost function that takes into account both estimation and detection errors. Under speech-presence, the cost is proportional to a quadratic spectral amplitude (QSA) error [8], while under speech-absence, the distortion depends on a certain attenuation factor [2], [10], [20]. We derive a combined detector and estimator with cost parameters that enable to control the tradeoff between speech distortion, caused by missed detection of speech components and residual musical noise resulting from false-detection. The combined solution generalizes the well-known STSA algorithm, which involves merely estimation under signal

Manuscript received December 25, 2006; revised June 21, 2007. This work was supported by the Israel Science Foundation under Grant 1085/05. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rainer Martin.

The authors are with the Department of Electrical Engineering, The Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: aari@tx.technion.ac.il; icohen@ee.technion.ac.il).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.904231

presence uncertainty. In addition, we propose a modification of the decision-directed *a priori* signal-to-noise ratio (SNR) estimator, which is suitable for transient-noise environments. Experimental results show that the simultaneous detection and estimation yields better noise reduction than the STSA algorithm while not degrading the speech signal. The advantage of using a suitable indicator for transient noise is demonstrated in a nonstationary noise environment, where the proposed algorithm facilitates suppression of transient noise with a controlled level of speech distortion.

The paper is organized as follows. In Section II, we briefly review classical speech enhancement under signal presence uncertainty. In Section III, we reformulate the speech enhancement problem in the STFT domain as a simultaneous detection and estimation problem. In Section IV, we derive the combined solution for a QSA distortion function. In Section V, we relate our proposed approach to the spectral-subtraction approach. In Section VI, we present an *a priori* SNR estimator suitable for transient noise environments, and in Section VII we demonstrate the performance of the proposed approach compared to existing algorithms, both under stationary and transient-noise environments.

II. Classical Speech Enhancement

In this section, we present the classical approach for spectral speech enhancement in nonstationary noise environments, assuming that some indicator for transient noise activity is available.

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, and let $y(n) = x(n)+d(n)$ be the observed signal. Applying the STFT to the observed signal, we have

$$Y_{\ell k} = X_{\ell k} + D_{\ell k} \tag{1}$$

where $\ell = 0, 1, \ldots$ is the time frame index and $k = 0, 1, \ldots, K-1$ is the frequency-bin index. Let $H_1^{\ell k}$ and $H_0^{\ell k}$ denote, respectively, speech presence and absence hypotheses in the time–frequency bin (ℓ, k), i.e.,

$$\begin{aligned} H_1^{\ell k} &: Y_{\ell k} = X_{\ell k} + D_{\ell k} \\ H_0^{\ell k} &: Y_{\ell k} = D_{\ell k}. \end{aligned} \tag{2}$$

We assume that the noise expansion coefficients can be represented as the sum of two uncorrelated noise components $D_{\ell k} = D_{\ell k}^s + D_{\ell k}^t$, where $D_{\ell k}^s$ denotes a quasi-stationary noise component, and $D_{\ell k}^t$ denotes a highly nonstationary transient component. The transient components are generally rare, but they may be of high energy and thus cause significant degradation to speech quality and intelligibility. However, in many applications, a reliable indicator for the transient noise activity may be available in the system. For example, in an emergency vehicle (e.g., police or ambulance) the engine noise may be considered as quasi-stationary, but activating a siren results in a highly nonstationary noise which is perceptually very annoying. Since the sound generation in the siren is nonlinear, linear echo cancelers, e.g., [21], may be inappropriate. In a computer-based communication system, a transient noise such as a keyboard typing noise may be present in addition to quasi-stationary background office noise. Another example is a digital camera, where activating the lens-motor (zooming in/out) may result in high-energy transient noise components, which degrade the recorded audio. In the above examples, an indicator for the transient noise activity may be available, i.e., siren source signal, keyboard output signal and the lens-motor controller output. Furthermore, given that a transient noise source is active, a detector for the transient noise in the STFT domain may be designed and its spectrum can be estimated based on training data.

The objective of a speech enhancement system is to reconstruct the spectral coefficients of the speech signal such that under speech-presence a certain distortion measure between the spectral coefficient and its estimate, $d(X_{\ell k}, \hat{X}_{\ell k})$, is minimized, and under speech-absence a constant attenuation of the noisy coefficient would be desired to maintain a natural background noise [10], [20]. Although the speech expansion coefficients are not necessarily present, most classical speech enhancement algorithms try to estimate the spectral coefficients rather than detecting their existence, or try to independently design detectors and estimators. The well-known spectral subtraction algorithm estimates the speech spectrum by subtracting the estimated noise spectrum from the noisy squared absolute coefficients [1], [2], and thresholding the result by some desired residual noise level. Thresholding the spectral coefficients is in fact a detection operation in the time–frequency domain, in the sense that speech coefficients are assumed to be absent in the low-energy time–frequency bins and present in noisy coefficients whose energy is above the threshold.

McAulay and Malpass were the first to propose a two-state model for the speech signal in the time–frequency domain [7]. Accordingly, the MMSE estimator follows: [22]

$$\begin{aligned} \hat{X}_{\ell k} &= E\{X_{\ell k}|Y_{\ell k}\} \\ &= E\left\{X_{\ell k}|Y_{\ell k}, H_1^{\ell k}\right\} p\left(H_1^{\ell k}|Y_{\ell k}\right). \end{aligned} \tag{3}$$

The resulting estimator does not detect speech components, but rather, a soft-decision is performed to further attenuate the signal estimate by the *a posteriori* speech presence probability. Ephraim and Malah followed the same approach and derived an estimator which minimizes the MSE of the STSA under signal presence uncertainty [8]. Accordingly

$$|\hat{X}_{\ell k}| = E\left\{|X_{\ell k}||Y_{\ell k}, H_1^{\ell k}\right\} p\left(H_1^{\ell k}|Y_{\ell k}\right). \tag{4}$$

Both in [7] and [8], under $H_0^{\ell k}$ the speech components are assumed zero and the *a priori* probability of speech presence is both time and frequency invariant, i.e., $p(H_1^{\ell k}) = p(H_1)$. In [9] and [10], the speech presence probability is evaluated for each frequency-bin and time-frame to improve the performance of the MMSE-LSA estimator [23]. Further improvement of the MMSE-LSA suppression rule can be achieved by considering under $H_0^{\ell k}$ a constant attenuation factor $G_f \ll 1$, which is determined by subjective criteria for residual noise naturalness, see also [20]. The OM-LSA estimator [10] is given by

$$\begin{aligned} |\hat{X}_{\ell k}| = \left(\exp\left[E\left\{\log|X_{\ell k}||Y_{\ell k}, H_1^{\ell k}\right\}\right]\right)^{p\left(H_1^{\ell k}|Y_{\ell k}\right)} \\ \times \left(G_f|Y_{\ell k}|\right)^{1-p\left(H_1^{\ell k}|Y_{\ell k}\right)}. \end{aligned} \tag{5}$$

Suppose that an indicator for the presence of transient noise components is available in a highly nonstationary noise environment, then high-energy transients may be attenuated by using one of the aforementioned estimators (3)–(5) and heuristically

setting the *a priori* speech presence probability $p(H_1^{\ell k})$ to a sufficiently small value. Unfortunately, this also results in suppression of desired speech components and intolerable degradation of speech quality. In general, an estimation-only approach under signal presence uncertainty produces larger speech degradation for small $p(H_1^{\ell k})$, since the optimal estimate is attenuated by the *a posteriori* speech presence probability. On the other hand, increasing $p(H_1^{\ell k})$ prevents the estimator from sufficiently attenuating noise components. Integrating a jointly optimal detector and estimator into the speech enhancement system may significantly improve the speech enhancement performance under highly nonstationary noise conditions and may allow further reduction of transient components without much degradation of the desired signal.

III. Reformulation of the Speech Enhancement Problem

In this section, we reformulate the speech enhancement as a simultaneous detection and estimation problem.

Middleton and Esposito [22] were the first to propose simultaneous signal detection and estimation within the framework of statistical decision theory. A decision space $\{\eta_0^{\ell k}, \eta_1^{\ell k}\}$ is assumed for the detection operation where under the decision $\eta_j^{\ell k}$, signal hypothesis $H_j^{\ell k}$ is accepted and a corresponding estimate $\hat{X}_{\ell k} = \hat{X}_{\ell k,j}$ is considered. The detection and estimation are strongly coupled so that the detector is optimized with the knowledge of the specific structure of the estimator, and the estimator is optimized in the sense of minimizing a Bayesian risk associated with the combined operations. For notation simplification, we omit the time–frequency indices (ℓ, k). Let $C_j(X, \hat{X}) \geq 0$ denote the cost of making a decision η_j (and choosing an estimator $\hat{X}_j$) where X is the desired signal. Then, the Bayes risk of the two operations associated with simultaneous detection and estimation is defined by [22] and [24]

$$R = \sum_{j=0}^{1} \int_{\Omega_y} \int_{\Omega_x} C_j(X, \hat{X}) p(\eta_j|Y) p(Y|X) p(X) dX dY \quad (6)$$

where Ω_x and Ω_y are the spaces of the speech and noisy signals, respectively. The simultaneous detection and estimation approach is aimed at jointly minimizing the Bayes risk over both the decision rule and the corresponding signal estimate. Let $q \triangleq p(H_1)$ denote the *a priori* speech presence probability and let X_R and X_I denote the real and imaginary parts of the expansion coefficient X. Then, the *a priori* distribution of the speech expansion coefficient follows:

$$p(X) = qp(X|H_1) + (1-q)p(X|H_0) \quad (7)$$

where $p(X|H_0) = \delta(X)$ and $\delta(X) \triangleq \delta(X_R, X_I)$ denotes the Dirac-delta function. The cost function $C_j(X, \hat{X})$ may be defined differently whether H_1 or H_0 is true. Therefore, we let $C_{ij}(X, \hat{X}) \triangleq C_j(X, \hat{X}|H_i)$ denote the cost which is conditioned on the true hypothesis.[1] The cost function $C_{ij}(X, \hat{X})$ depends on both the true signal value and its estimate under the decision η_j and therefore couples the operations of detection and estimation. By substituting (7) into (6), we obtain

$$\begin{aligned} R = \int_{\Omega_y} \int_{\Omega_x} p(Y|X) \{ & p(\eta_0|Y) \\ & \times \Big[qp(X|H_1) C_{10}(X, \hat{X}) + (1-q) \\ & \quad \times p(X|H_0) C_{00}(X, \hat{X}) \Big] + p(\eta_1|Y) \\ & \times \Big[qp(X|H_1) C_{11}(X, \hat{X}) \\ & \quad + (1-q) p(X|H_0) C_{01}(X, \hat{X}) \Big] \Big\} \\ & \times dX dY. \end{aligned} \quad (8)$$

Let

$$r_{ij}(Y) = \int_{\Omega_x} C_{ij}(X, \hat{X}) p(X|H_i) p(Y|X) dX \quad (9)$$

denote a risk associated with the pair $\{H_i, \eta_j\}$ and the observation Y. Then, the combined Bayes risk follows:

$$\begin{aligned} R = \int_{\Omega_y} & p(\eta_0|Y) \left[qr_{10}(Y) + (1-q) r_{00}(Y) \right] \\ & + p(\eta_1|Y) \left[qr_{11}(Y) + (1-q) r_{01}(Y) \right] dY. \end{aligned} \quad (10)$$

Since the detector's decision under a given observation is binary, i.e., $p(\eta_j|Y) \in \{0, 1\}$, for minimizing the combined risk we first evaluate the optimal estimator under each of the decisions, then the optimal decision rule is derived based on the optimal estimators $\hat{X}_0$, $\hat{X}_1$ to further minimize the combined risk. The two-stage minimization guaranties minimum combined risk [24]. The optimal *nonrandom* decision rule which minimizes the combined risk (10) is given by

Decide η_1 (i.e., $p(\eta_1|Y) = 1$) if

$$q \left[r_{10}(Y) - r_{11}(Y) \right] \geq (1-q) \left[r_{01}(Y) - r_{00}(Y) \right] \quad (11)$$

otherwise, decide η_0.

The optimal estimator under a decision η_j is obtained from (10) by

$$\arg\min_{\hat{X}_j} \left\{ qr_{1j}(Y) + (1-q) r_{0j}(Y) \right\}. \quad (12)$$

Note that $r_{ij}(Y)$ depends on the estimate $\hat{X}_j$ through the cost function. Fig. 1 shows a block diagram of the simultaneous detection and estimation scheme compared with an independent detection and estimation system. The standard, noncoupled detection and estimation system (a) consists of an estimator and a detector which independently chooses to accept or reject the estimator output. In the simultaneous detection and estimation scheme, the estimator is obtained by (12) and the interrelated decision rule (11) chooses the appropriate estimator, $\hat{X}_0$ or $\hat{X}_1$, for minimizing the combined Bayes risk. Since the risk $r_{ij}(Y)$ is a function of the signal estimate $\hat{X}_j$, the decision rule (11) requires knowledge of the estimator under any of its own decisions. Therefore, the arrow between the estimation and the detection blocks is unidirectional. It is important to note that the optimal estimator (12) minimizes the Bayes risk under any

[1] Note that $X = 0$ implies that H_0 is true, and $X \neq 0$ implies H_1 so the subindex i may seem to be redundant. However, this notation simplifies the subsequent formulations.

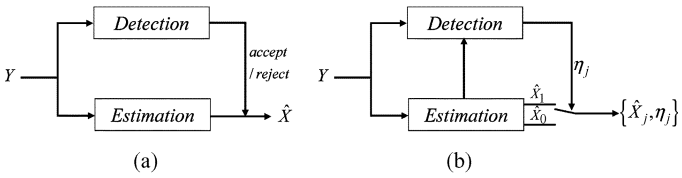

Fig. 1. (a) Independent detection and estimation system. (b) Strongly coupled detection and estimation system.

given decision rule, even if the detector is not optimal and/or is unknown to the estimator.

The cost function associated with the pair $\{H_i, \eta_j\}$ is generally defined by

$$C_{ij}(X,\hat{X}) = b_{ij} d_{ij}(X,\hat{X}) \tag{13}$$

where $d_{ij}(X,\hat{X})$ is an appropriate distortion measure and the cost parameters b_{ij} control the tradeoff between the costs associated with the pairs $\{H_i, \eta_j\}$. That is, a high-valued b_{01} raises the cost of a false alarm, (i.e., decision of speech presence when speech is actually absent) which may result in residual musical noise. Similarly, b_{10} is associated with the cost of missed detection of a signal component, which may cause perceptual signal distortion. Under a correct classification, normalized cost parameters are generally used, $b_{00} = b_{11} = 1$. However, $d_{ii}(\cdot,\cdot)$ is not necessarily zero since estimation errors are still possible even when there is no detection error.

Contrary to the approach in [22], [24], and [25], we do not reject the signal estimator when a decision η_0 is made. Instead, we allow the estimator $\hat{X}_0 \neq 0$ to compensate for any detection errors and to reduce potential musical noise and audible distortions. Furthermore, when speech is indeed absent, the distortion function is defined to allow some natural background noise level such that under H_0, the attenuation factor will be lower bounded by a constant gain floor $G_f \ll 1$ as proposed in [2], [10], [20], and [26].

IV. Quadratic Spectral Amplitude Cost Function

In this section, we derive a speech simultaneous detection and estimation scheme for a QSA cost function.

The distortion measure of the QSA cost function is defined by

$$d_{ij}(X,\hat{X}) = \begin{cases} \left(|X| - |\hat{X}_j|\right)^2, & i = 1 \\ \left(G_f|Y| - |\hat{X}_j|\right)^2, & i = 0 \end{cases} \tag{14}$$

and is related to the STSA suppression rule of Ephraim and Malah [8]. We assume that both X and D are statistically independent, zero-mean, complex-valued Gaussian random variables with variances λ_x and λ_d, respectively. Let $\xi \triangleq \lambda_x/\lambda_d$ denote the *a priori* SNR under hypothesis H_1, let $\gamma \triangleq |Y|^2/\lambda_d$ denote the *a posteriori* SNR and let $\upsilon \triangleq \gamma\xi/(1+\xi)$. For evaluating the optimal detector and estimator under the QSA cost function we denote by $X \triangleq ae^{j\alpha}$ and $Y \triangleq Re^{j\theta}$ the clean and noisy spectral coefficients, respectively, where $a = |X|$ and $R = |Y|$. Accordingly, the pdf of the speech expansion coefficient under H_1 satisfies

$$p(a,\alpha|H_1) = \frac{a}{\pi\lambda_x}\exp\left(-\frac{a^2}{\lambda_x}\right). \tag{15}$$

The combined risk under the QSA cost function is independent of the signal phase nor the estimation phase. Therefore, we define $\hat{a}_j = |\hat{X}_j|$ as the estimated amplitude under η_j. Substituting the QSA cost function into (12) we have

$$\hat{a}_j = \arg\min_{\hat{a}} \left\{ q b_{1j} \int_0^\infty \int_0^{2\pi} (a-\hat{a})^2 p(a,\alpha|H_1) p(Y|a,\alpha) d\alpha da + (1-q) b_{0j} (G_f R - \hat{a})^2 p(Y|H_0) \right\} \tag{16}$$

and by constraining the derivative according to $\hat{a}$ to equal zero, we obtain

$$\hat{a}_j \left[b_{1j}\Lambda(Y) + b_{0j}\right] = b_{1j}\Lambda(Y) \int_0^\infty \int_0^{2\pi} a p(a,\alpha|H_1) \times p(Y|a,\alpha) d\alpha\, da / p(Y|H_1) + b_{0j} G_f R \tag{17}$$

where $\Lambda(Y)$ is the generalized likelihood ratio defined by [8]

$$\Lambda(Y) \triangleq \frac{q}{(1-q)} \frac{p(Y|H_1)}{p(Y|H_0)} = \frac{q}{(1-q)} \frac{e^{\upsilon}}{1+\xi}. \tag{18}$$

Note that given the *a priori* speech presence probability, the generalized likelihood ratio is a function of the *a priori* and *a posteriori* SNRs $\Lambda(\xi,\gamma)$. Using [8] we observe that

$$\int_0^\infty \int_0^{2\pi} a p(a,\alpha|H_1) p(Y|a,\alpha) d\alpha da / p(Y|H_1) = \frac{\sqrt{\pi\upsilon}}{2\gamma}\exp\left(-\frac{\upsilon}{2}\right)\left[(1+\upsilon) I_0\left(\frac{\upsilon}{2}\right) + \upsilon I_1\left(\frac{\upsilon}{2}\right)\right] R \triangleq G_{\mathrm{STSA}}(\xi,\gamma) R \tag{19}$$

where $I_\nu(\cdot)$ denotes the modified Bessel function of order ν. Let $\phi_j(\xi,\gamma) \triangleq b_{1j}\Lambda(\xi,\gamma) + b_{0j}$. Then, by using the phase of the noisy signal [8], we obtain from (17) and (19) the optimal estimation under the decision η_j, $j \in \{0,1\}$:

$$\hat{X}_j = \left[b_{1j}\Lambda(\xi,\gamma) G_{\mathrm{STSA}}(\xi,\gamma) + b_{0j} G_f\right] \phi_j(\xi,\gamma)^{-1} Y \triangleq G_j(\xi,\gamma) Y. \tag{20}$$

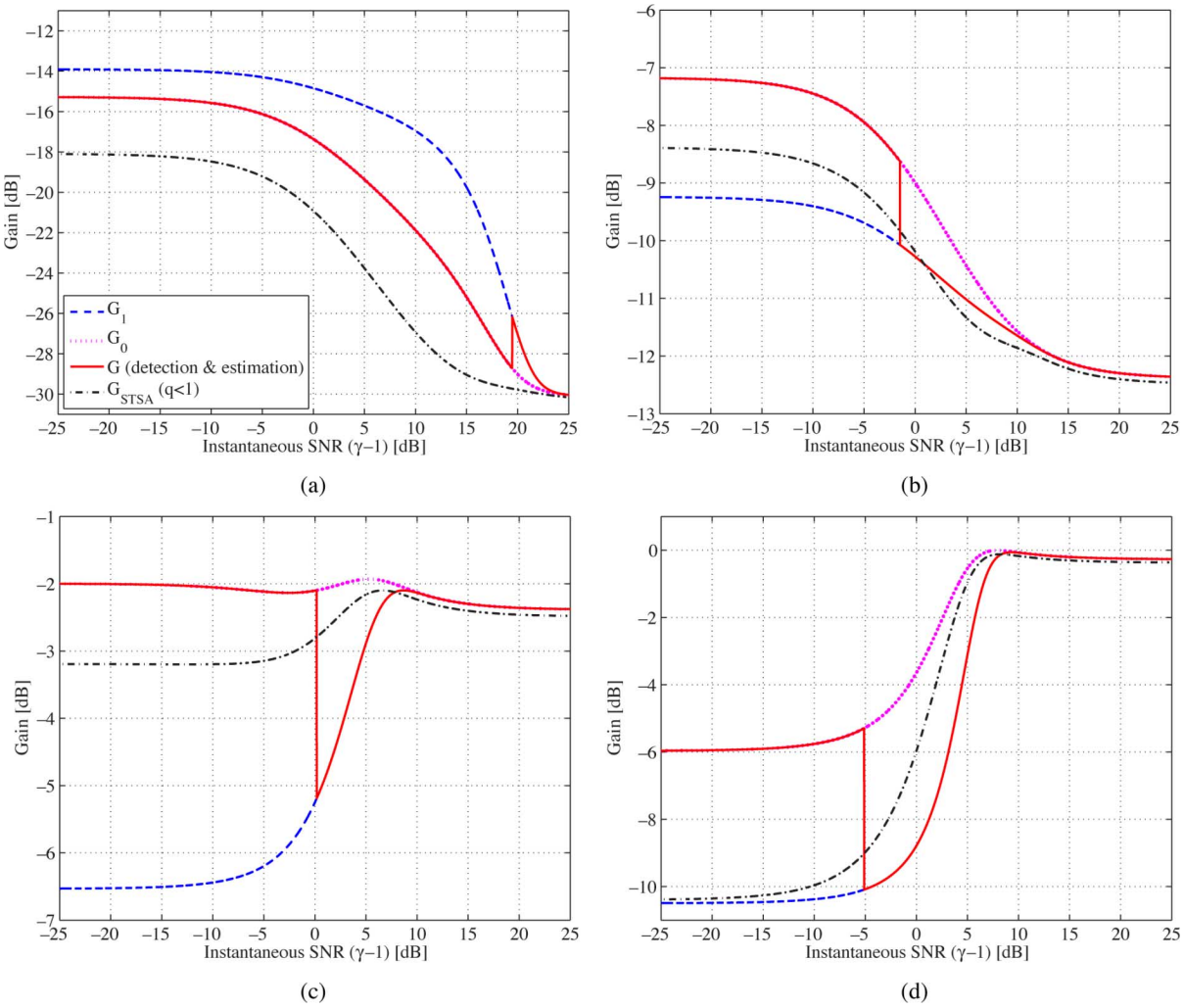

Fig. 2. Gain curves of G_1 (dashed line), G_0 (dotted line), and the total detection and estimation system gain curve (solid line), compared with the STSA gain under signal presence uncertainty (dashed-dotted line). The *a priori* SNRs are (a) $\xi = -15$ dB, (b) $\xi = -5$ dB, (c) $\xi = 5$ dB, and (d) $\xi = 15$ dB.

For evaluating the optimal decision rule, we need to compute the risk $r_{ij}(Y)$. Under H_1, we obtain

$$\begin{aligned} r_{1j}(Y) &= \frac{b_{1j}}{\pi}\frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{1+\xi} \\ &\quad\times\left\{G_j^2\gamma+\frac{\xi}{1+\xi}(1+\upsilon)-G_j\sqrt{\pi\upsilon}\exp\left(-\frac{\upsilon}{2}\right)\right. \\ &\qquad\left.\times\left[(1+\upsilon)I_0\left(\frac{\upsilon}{2}\right)+\upsilon I_1\left(\frac{\upsilon}{2}\right)\right]\right\} \\ &= \frac{b_{1j}}{\pi}\frac{\exp\left(-\frac{\gamma}{1+\xi}\right)}{1+\xi} \\ &\quad\times\left\{G_j^2\gamma+\frac{\xi}{1+\xi}(1+\upsilon)-2\gamma G_j G_{\mathrm{STSA}}\right\} \end{aligned} \tag{21}$$

(see proof in the Appendix) where G_j holds for $G_j(\xi,\gamma)$, the gain function under the QSA cost function and the decision η_j which is defined in (20), and G_{STSA} holds for $G_{\mathrm{STSA}}(\xi,\gamma)$ which is defined in (19).

For deriving the risk under H_0, $r_{0j}(Y)$, we observe $p(X_R,X_I|H_0)=\delta(X_R,X_I)$. Consequently

$$\begin{aligned} r_{0j}(Y) &= b_{0j}\int_{-\infty}^{\infty}\int\left\{[G_j(\xi,\gamma)-G_f]^2|Y|^2\right\} \\ &\quad\times p(X_R,X_I|H_0)p(Y|X_R,X_I)dX_R dX_I \\ &= \frac{b_{0j}}{\pi}[G_j(\xi,\gamma)-G_f]^2\gamma e^{-\gamma}. \end{aligned} \tag{22}$$

Substituting (21) and (22) into (11), we obtain the optimal decision rule under the QSA cost function

$$\begin{aligned} &\Lambda(\xi,\gamma)\left\{b_{10}G_0^2-G_1^2+\frac{\xi}{(1+\xi)\gamma}(1+\upsilon)(b_{10}-1)+\right. \\ &\left.2(G_1-b_{10}G_0)G_{\mathrm{STSA}}\right\}\underset{\eta_0}{\overset{\eta_1}{\gtrless}} b_{01}(G_1-G_f)^2-(G_0-G_f)^2. \end{aligned} \tag{23}$$

To conclude the above results, simultaneous detection and estimation from noisy observations requires 1) calculating the gain factor under any of the decisions using (20) and 2) finding the

optimal decision η_j using (23). The corresponding signal estimate is obtained by applying the gain G_j to the noisy observation.

Fig. 2 demonstrates attenuation curves under QSA cost function as a function of the *instantaneous* SNR defined by $\gamma - 1$, for several *a priori* SNRs, using the parameters $q = 0.8$, (as proposed in [8]) $G_f = -25$ dB and cost parameters $b_{01} = 5$ and $b_{10} = 1.1$. The gains G_1 (dashed line), G_0 (dotted line), and the total detection and estimation system gain (solid line) are compared to the STSA gain under signal presence uncertainty of Ephraim and Malah [8] (dashed-dotted line). The *a priori* SNRs range from -15 to 15 dB. Not only that the cost parameters shape the STSA gain curve, when combined with the detector the proposed method provides a significant noncontinuous modification of the standard STSA estimator. For example, for *a priori* SNRs of $\xi = -5$ and $\xi = 15$ dB, as shown in Fig. 2(b) and (d), respectively, as long as the instantaneous SNR is higher than about -2 dB (for $\xi = -5$ dB) or -5 dB (for $\xi = 15$ dB), the detector decision is η_1, while for lower instantaneous SNRs, the detector decision is η_0. Note that if an ideal detector for the speech coefficients would be available, a more significantly noncontinuous gain would be desired to block the noise-only coefficients. However, in the proposed simultaneous detection and estimation approach, the detector is not ideal but optimized to minimize the combined risk and the noncontinuity of the system gain depends on the chosen cost parameters as well as on the gain floor. As shown in our experimental results, this noncontinues gain function may yield greater noise reduction with slightly higher level of musicality, while not degrading speech quality.

It is of interest to examine the asymptotic behavior of the estimator (20) under each of the decisions. When the cost parameter associated with false alarm is much smaller than the generalized likelihood ratio, i.e., $b_{01} \ll \Lambda(\xi,\gamma)$, the spectral gain of the estimator under the decision η_1 is $G_1(\xi,\gamma) \cong G_{\mathrm{STSA}}(\xi,\gamma)$, which is optimal when the signal is surely present. However, if $b_{01} \gg \Lambda(\xi,\gamma)$, the spectral gain under η_1 needs to compensate the possibility of a high-cost false-decision made by the detector, and thus $G_1(\xi,\gamma) \cong G_f$. On the other hand, if the cost parameter associated with missed detection is small and we have $b_{10} \ll \Lambda(\xi,\gamma)^{-1}$, then $G_0(\xi,\gamma) \cong G_f$ (i.e., estimation where speech is surely absent) but under $b_{10} \gg \Lambda(\xi,\gamma)^{-1}$, in order to overcome the high cost related to missed detection, we have $G_0(\xi,\gamma) \cong G_{\mathrm{STSA}}(\xi)$.

Recall that

$$\frac{\Lambda(\xi,\gamma)}{1+\Lambda(\xi,\gamma)} = p(H_1|Y) \tag{24}$$

is the *a posteriori* probability for speech presence [8], it can be shown that the proposed estimator (20) generalizes the well-known STSA estimator. For the case of $b_{ij} = 1\ \forall i,j$ we have

$$\begin{aligned}\hat{X}_0 &= \left[p(H_1|Y)G_{\mathrm{STSA}}(\xi,\gamma) + (1-p(H_1|Y))\,G_f\right]Y \\ &= \hat{X}_1.\end{aligned} \tag{25}$$

In that case, the detection operation is not required since the estimation is independent of the decision rule. If we also set G_f to zero, the estimation reduces to the STSA suppression rule under signal presence uncertainty [8].

The simultaneous detection and estimation approach requires the calculation of two gain functions, $G_0(\xi,\gamma)$ and $G_1(\xi,\gamma)$, and the decision rule. However, as can be seen from (20), both $G_0(\xi,\gamma)$ and $G_1(\xi,\gamma)$ are linear functions of $G_{\mathrm{STSA}}(\xi,\gamma)$ and the generalized likelihood ratio $\Lambda(\xi,\gamma)$. In addition, the decision rule (23) requires the calculation of a second-order polynomial. Therefor, the additional complexity of the simultaneous detection and estimation approach is insignificant compared to the STSA estimator [8], which also requires the calculation of the gain function $G_{\mathrm{STSA}}(\xi,\gamma)$ (19) and the generalized likelihood function (28).

V. Relation to Spectral Subtraction

The general formulation of the spectral subtraction approach assumes a spectral estimator which can be written as [1], [2]

$$\hat{X}_{\ell k} = \max\left\{\left(|Y_{\ell k}|^{\tau} - \mu E\left[|D_{\ell k}|^{\tau}\right]\right)^{\frac{1}{\tau}}, \beta E\left[|D_{\ell k}|^{\tau}\right]^{\frac{1}{\tau}}\right\}\frac{Y_{\ell k}}{|Y_{\ell k}|} \tag{26}$$

where $E[|D_{\ell k}|^{\tau}]$ is the τ-order moment of the noise spectral coefficient, $\mu \geq 1$ represents an over-subtraction factor, and $0 < \beta \ll 1$ represents spectral floor factor. Boll [1] considered $\tau = 1$ while Berouti *et al.* [2] used $\tau = 2$. McAulay and Malpass [7] showed that under a Gaussian statistical model, spectral subtraction with $\tau = 2$, $\mu = 1$, and $\beta = 0$ yields a maximum-likelihood estimator for the speech spectral variance.

The spectral subtraction scheme (26) classifies high-energy time–frequency bins as active speech bins, and only in these bins the signal is estimated. Low-energy bins below a given threshold are classified as noise-only bins, and set to some background noise level for reducing the residual musical noise. Consequently, low-energy bins that contain the speech signal are not detected, while noise peaks are detected as speech bins. When the over-subtraction factor μ is increased, fewer noise peaks are detected as speech and therefore the residual musical noise is reduced at the expanse of deterioration of speech quality. The spectral floor $\beta E[|D_{\ell k}|^{\tau}]^{1/\tau}$ "fills-in" the valleys of the residual noise, which yields a more natural noise with less annoying musicality [2]. However, a large β reduces the background noise suppression. Further reduction of the musical noise may be achieved by local smoothing of the noisy spectral values prior to noise subtraction. As a result, noise peaks are attenuated and the spectral estimation error can be reduced [1]. However, as the speech signal is highly nonstationary, its intelligibility may be dramatically decreased when the smoothing parameter increases.

The classical spectral subtraction approach heuristically combines a detector and an estimator for the speech spectral coefficients while the parameters μ, β, and the smoothing length control the tradeoff between the residual musical noise and the speech quality. In the proposed simultaneous detection and estimation approach, the detector is optimally designed jointly with the estimator. The residual noise musicality is controlled by both the spectral gain floor G_f which bounds the attenuation and the false-alarm cost parameter b_{01}. A high-valued false-alarm cost parameter (with relation to the generalized likelihood ratio) reduces the estimation gain under η_1, which compensates for a false-detection. The amount of speech distortion is affected by

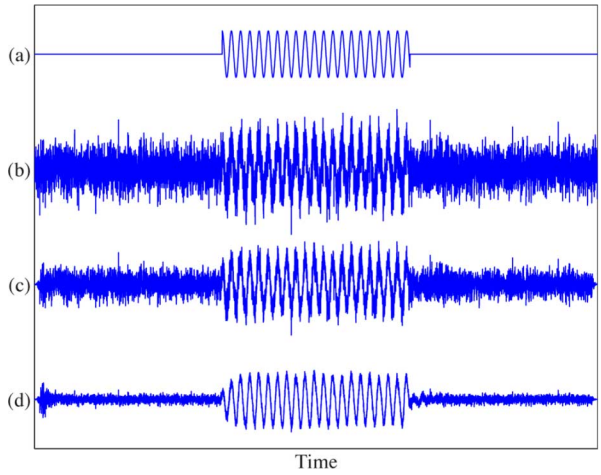

Fig. 3. Signals in the time domain. (a) Clean sinusoidal signal.(b) Noisy signal. (c) Enhanced signal obtained by using the spectral-subtraction estimator. (d) Enhanced signal obtained by using the detection and estimation approach.

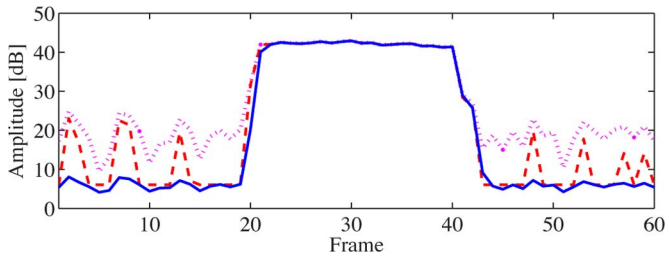

Fig. 4. Amplitudes of the STFT coefficients along the time-trajectory corresponding to the frequency of the sinusoidal signal: noisy signal (dotted line), spectral subtraction (dashed line), and simultaneous detection and estimation (solid line).

the missed detection parameter b_{10}, which increases the estimation gain under η_0. Since the decision rule depends on both parameters as well as on the gain floor, it is the combination of the three parameters that control the tradeoffs between noise reduction and speech distortion.

The different behaviors of the spectral subtraction and the simultaneous detection and estimation approach are illustrated in Figs. 3 and 4. The signals in the time domain are shown in Fig. 3. The clean signal is a sinusoidal wave which is active only in a specific time interval and the noisy signal contains white Gaussian noise with an SNR of 5 dB. The noisy signal is transformed into the STFT domain using half-overlapping Hamming windows of 256 taps. The signal enhanced by spectral subtraction with $\tau = 2$, $\mu = 1$, and $\beta = 0.2$ is shown in Fig. 3(c), and the signal enhanced by using the proposed algorithm is shown in Fig. 3(d) with $b_{01} = 3$, $b_{10} = 5$, $G_f = -20$ dB, and $q = 0.8$. The *a priori* SNR needed for the simultaneous detection and estimation approach is estimated using the decision-directed approach as will be defined in (27), with a weighting factor $\alpha = 0.92$ and $\xi_{min} = -20$ dB as the lower bound for the *a priori* SNR, while the variance of the background noise coefficients is evaluated from the noise signal (for both algorithms). The amplitudes of the signals in the STFT domain (at the specific frequency band of the desired signal's frequency) are shown in Fig. 4. It can be seen that when the desired signal is absent, high-energy noise components are falsely detected by the spectral subtraction algorithm which potentially results in an annoying musical noise. The detection and estimation algorithm results in a higher attenuation of the noise peaks and smoother and more natural background noise while not increasing the audible distortion in the enhanced signal. Furthermore, it may seem from Fig. 4 that when the desired signal is active and the instantaneous SNR is high, both algorithms imply similar results. However, in time frames where the desired signal is present, the spectral subtraction approach results in higher residual noise in frequencies where the signal is absent or of low SNR. Therefore, the enhanced signal using the spectral subtraction approach is inferior to the enhanced signal using the detection and estimation approach even in time intervals where the signal is present, as can be seen from Figs. 3(c) and (d).

VI. *A Priori* SNR Estimation

Speech enhancement in the STFT domain generally relies on an estimation-only approach under signal presence uncertainty, e.g., [7], [8], and [10]. The *a priori* SNR is often estimated by using the decision-directed approach [8]. Accordingly, in each time–frequency bin we compute

$$\hat{\xi}_{\ell k} = \max\Big\{\alpha G^2(\hat{\xi}_{\ell-1,k}, \gamma_{\ell-1,k})\gamma_{\ell-1,k} + (1-\alpha)(\gamma_{\ell k} - 1), \xi_{\min}\Big\} \quad (27)$$

where α $(0 \leq \alpha \leq 1)$ is a weighting factor that controls the tradeoff between noise reduction and transient distortion introduced into the signal, and $\xi_{\min}$ is a lower bound for the *a priori* SNR which is necessary for reducing the residual musical noise in the enhanced signal [8], [20]. Since the *a priori* SNR is defined under the assumption that $H_1^{\ell k}$ is true, it is proposed in [10] to replace the gain G in (27) by G_{H_1} which represents the spectral gain when the signal is surely present (i.e., $q = 1$). Increasing the value of α results in a greater reduction of the musical noise phenomena, at the expense of further attenuation of transient speech components (e.g., speech onsets) [20]. By using the proposed approach with high cost for false speech detection, the musical noise can be reduced without increasing the value of α, which enables rapid changes in the *a priori* SNR estimate. The lower bound for the *a priori* SNR is related to the spectral gain floor G_f since both imply a lower bound on the spectral gain. The latter parameter is used to evaluate both the optimal detector and estimator while taking into account the desired residual noise level.

The decision-directed estimator is widely used, but is not suitable for transient noise environments, since a high-energy noise burst may yield an instantaneous increase in the *a posteriori* SNR and a corresponding increase in $\hat{\xi}_{\ell k}$ as can be seen from (27). The spectral gain would then be higher than the desired value, and the transient noise component would not be sufficiently attenuated. Let $\hat{\lambda}^s_{d_{\ell k}}$ denote the estimated spectral variance of the stationary noise component and let $\hat{\lambda}^t_{d_{\ell k}}$ denote the estimated spectral variance of the transient component. The former may be practically estimated by using the improved minima-controlled recursive averaging (IMCRA) algorithm [10], [27] or by using the minimum-statistics approach [28], while $\lambda^t_{d_{\ell k}}$ may be evaluated based on a training phase as

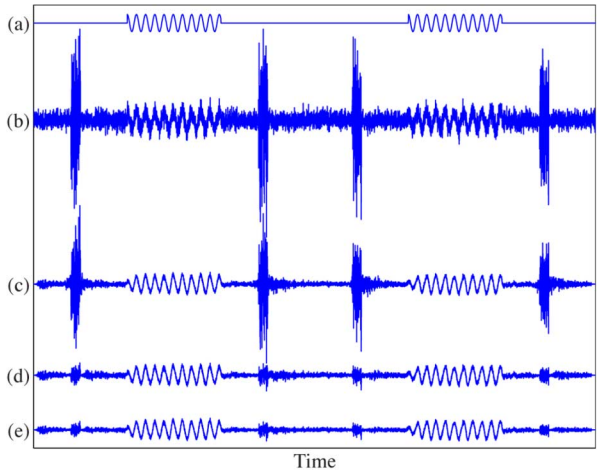

Fig. 5. Signals in the time domain. (a) Clean sinusoidal signal. (b) Noisy signal with both stationary and transient components. (c) Enhanced signal obtained by using the STSA and the decision-directed estimators. (d) Enhanced signal obtained by using the STSA and the modified *a priori* SNR estimators. (e) Enhanced signal obtained by using the detection and estimation approach and the modified *a priori* SNR estimator.

assumed in [29]. The total variance of the noise component is $\hat{\lambda}_{d_{\ell k}} = \hat{\lambda}^s_{d_{\ell k}} + \hat{\lambda}^t_{d_{\ell k}}$. Note that $\lambda^t_{d_{\ell k}} = 0$ in time–frequency bins where the transient noise is inactive. Since the *a priori* SNR is highly dependent on the noise variance, we first estimate the speech spectral variance by

$$\hat{\lambda}_{x_{\ell k}} = \max\left\{\alpha G^2_{H_1}(\hat{\xi}_{\ell-1,k}, \gamma_{\ell-1,k})|Y_{\ell-1,k}|^2 + (1-\alpha)\left(|Y_{\ell k}|^2 - \hat{\lambda}_{d_{\ell k}}\right), \lambda_{\min}\right\} \quad (28)$$

where $\lambda_{\min} = \xi_{\min}\hat{\lambda}^s_{d_{\ell k}}$. Then, the *a priori* SNR is evaluated by $\hat{\xi}_{\ell k} = \hat{\lambda}_{x_{\ell k}}/\hat{\lambda}_{d_{\ell k}}$. It is straightforward to show that in a stationary noise environment, the proposed *a priori* SNR estimator reduces to the decision-directed estimator (27), with G_{H_1} substituting G. However, under the presence of a transient noise component, the proposed method yields a lower *a priori* SNR estimate, which enables higher attenuation of the high-energy transient noisy component. Furthermore, to allow further reduction of the transient noise component to the level of the residual stationary noise, we modify the gain floor by $\tilde{G}_f = G_f\hat{\lambda}^s_{d_{\ell k}}/\hat{\lambda}_{d_{\ell k}}$ as proposed in [30].

The different behaviors under transient noise conditions of the proposed modified decision-directed *a priori* SNR estimator and the decision-directed estimator as proposed in [10] are illustrated in Figs. 5 and 6. Fig. 5 shows the signals in the time domain: the analyzed signal contains a sinusoidal wave which is active in only two specific segments. The noisy signal contains both additive white Gaussian noise with 5-dB SNR and high-energy transient noise components. The signal enhanced by using the decision-directed estimator and the STSA suppression rule is shown in Fig. 5(c). The signal enhanced by using the modified *a priori* SNR estimator and the STSA suppression rule is shown in Fig. 5(d), and the result obtained by using the proposed modified *a priori* SNR estimation with the detection and estimation approach is shown in Fig. 5(d) (using the same parameters as in the previous section). Both the decision-directed estimator and the modified *a priori* SNR estimator are applied with $\alpha = 0.98$ and $\xi_{\min} = -20$ dB. Clearly, in stationary noise intervals, and where the SNR is high, similar results are obtained by both *a priori* SNR estimators. However, the proposed modified *a priori* SNR estimator obtain higher attenuation of the transient noise, whether it is incorporated with the STSA or the simultaneous detection and estimation approach. Fig. 6 shows the amplitudes of the STFT coefficients of the noisy and enhanced signals at the frequency band which contains the desired sinusoidal component. Accordingly, the modified *a priori* SNR estimator enables a greater reduction of the background noise, particularly transient noise components. Moreover, it can be seen that using the simultaneous detection and estimation yields better attenuation of both the stationary and background noise compared to the STSA estimator, even while using the same *a priori* SNR estimator.

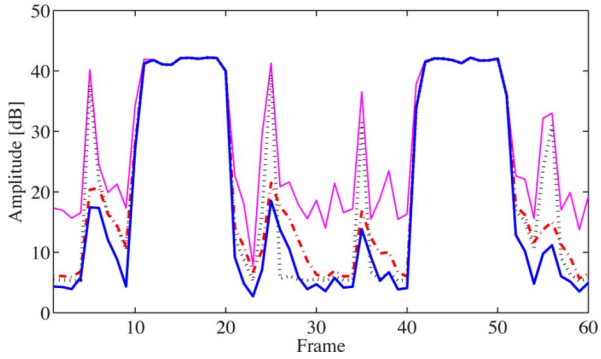

Fig. 6. Amplitudes of the STFT coefficients along time-trajectory corresponding to the frequency of the sinusoidal signal: noisy signal (light solid line), STSA with decision-directed estimation (dotted line), STSA with the modified *a priori* SNR estimator (dashed-dotted line), and simultaneous detection and estimation with the modified *a priori* SNR estimator (dark solid line).

VII. Experimental Results

In our experimental study, we first evaluate the detection and estimation approach compared with the STSA suppression rule under a stationary noise environment. Then, we consider the problem of hands-free communication in an emergency vehicle and demonstrate the advantage of the modified *a priori* SNR estimator together with the simultaneous detection and estimation approach under transient noise environment. Speech signals are taken from the TIMIT database [31], sampled at 16 kHz and degraded by additive noise. The test signals include 16 speech utterances from 16 different speakers, half male half female. The noisy signals are transformed into the STFT domain using half-overlapping Hamming windows of 32-ms length, and the background-noise spectrum is estimated by using the IMCRA algorithm (for all the considered enhancement algorithms) [10], [27]. The performance evaluation in our study includes objective quality measures, a subjective study of spectrograms, and informal listening tests. The first quality measure is the segmental SNR defined, in dB, by [32]

$$\text{SegSNR} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \times\mathcal{T}\left\{10\log_{10}\frac{\sum_{n=0}^{K-1} x^2(n+\ell K/2)}{\sum_{n=0}^{K-1}\left[x(n+\ell K/2)-\hat{x}(n+\ell K/2)\right]^2}\right\} \quad (29)$$

TABLE I
SEGMENTAL SNR AND LOG SPECTRAL DISTORTION OBTAINED BY USING EITHER THE SIMULTANEOUS DETECTION AND ESTIMATION APPROACH OR THE STSA ESTIMATOR IN STATIONARY NOISE ENVIRONMENT

Input SNR [dB]	Input Signal		Detection and Estimation		STSA ($\alpha = 0.98$)		STSA ($\alpha = 0.92$)	
	SegSNR	LSD	SegSNR	LSD	SegSNR	LSD	SegSNR	LSD
-5	-6.801	20.897	1.255	7.462	0.085	9.556	-0.684	10.875
0	-3.797	16.405	4.136	5.242	3.169	6.386	2.692	7.391
5	0.013	12.130	5.98	3.887	5.266	4.238	5.110	4.747
10	4.380	8.194	6.27	3.143	5.93	3.167	6.014	3.157

where $\mathcal{L}$ represents the set of frames which contain speech, $|\mathcal{L}|$ denotes the number of elements in $\mathcal{L}$, $K = 512$ is the number of samples per frame, and the operator $\mathcal{T}$ confines the SNR at each frame to a perceptually meaningful range between -10 and 35 dB. The second quality measure is log-spectral distortion (LSD) which is defined, in dB, by

$$\mathrm{LSD} = \frac{1}{L}\sum_{\ell=0}^{L-1}\left\{\frac{1}{K/2+1}\times\sum_{k=0}^{K/2}[10\log_{10}\mathcal{C}X_{\ell k} - 10\log_{10}\mathcal{C}\hat{X}_{\ell k}]^2\right\}^{\frac{1}{2}} \tag{30}$$

where $\mathcal{C}X \triangleq \max\{|X|^2, \epsilon\}$ is a spectral power clipped such that the log-spectrum dynamic range is confined to about 50 dB, that is, $\epsilon = 10^{-50/10}\cdot\max_{\ell,k}\{|X_{\ell k}|^2\}$. The third quality measure (used in Section VII-B) is the perceptual evaluation of speech quality (PESQ) score [33].

A. Comparison With the STSA Estimator

In this section, the suppression rule results from the proposed simultaneous detection and estimation approach is compared to the STSA estimation [8] for stationary white Gaussian noise with SNRs in the range [−5, 10] dB. For both algorithms the *a priori* SNR is estimated by the decision-directed approach (27) with $\xi_{min} = -15$ dB, and the *a priori* speech presence probability is $q_{\ell k} = 0.8$, as proposed in [8]. For the STSA estimator a decision-directed estimation [10] with $\alpha = 0.98$ reduces the residual musical noise but generally implies transient distortion of the speech signal [8], [20]. However, the inherent detector obtained by the simultaneous detection and estimation approach may improve the residual noise reduction, and therefore a lower weighting factor α may be used to allow lower speech distortion. Indeed, we have found out that for the simultaneous detection and estimation approach $\alpha = 0.92$ implies better results, while for the STSA algorithm, better results are achieved with $\alpha = 0.98$. The cost parameters for the simultaneous detection and estimation should be chosen according to the system specification, i.e., whether the quality of the speech signal or the amount of noise reduction is of higher importance. Table I summarizes the average segmental SNR and LSD for these two enhancement algorithms, with cost parameters $b_{01} = 10$ and $b_{10} = 2$, and $G_f = -15$ dB for the simultaneous detection and estimation algorithm. The results for the STSA algorithm are presented for $\alpha = 0.98$ as well as for $\alpha = 0.92$ (note that for the STSA estimator $G_f = 0$ is considered as originally proposed). It shows that the simultaneous detection and estimation yields improved segmental SNR and LSD, while a greater improvement is achieved for lower input SNR. Informal subjective listening tests and inspection of spectrograms demonstrate improved speech quality with higher attenuation of the background noise. However, since the weighting factor used for the *a priori* SNR estimate is lower, and the gain function is discontinuous, the residual noise resulting from the simultaneous detection and estimation algorithm is slightly more musical than that resulting from the STSA algorithm (examples are available online).[2]

B. Speech Enhancement Under Nonstationary Noise Environment

In this section, we demonstrate the potential advantage of the simultaneous detection and estimation approach with the proposed *a priori* SNR estimator under transient noise. We consider a hands-free communication in an emergency vehicle (police car, ambulance etc.) where the engine noise is assumed quasi-stationary. However, activating the emergency siren significantly degrades the perceptual quality and intelligibility of the speech signal, since its energy is much higher than that of the speech signal. The sound generation in a siren is nonlinear, which produces harmonics not present in the original signal (siren source signal), as can be seen in Fig. 7(b). However, using the available siren source signal, a reliable indicator in the time–frequency domain for the presence of siren noise, and an estimate for the variance of the transient noise, $\lambda^t_{d_{\ell k}}$, may be designed in a training phase. Note that standard echo-cancellation algorithms are not suitable for eliminating noise generated by nonlinear systems and nonlinear algorithms may be required (e.g., [34], [35]).

The proposed approach is compared with the STSA algorithm [8] and the OM-LSA algorithm [10]. The speech presence probability required for the OM-LSA estimator as well as for the simultaneous detection and estimation approach is estimated as proposed in [10], while for the STSA estimator $\hat{q}_{\ell k} = 0.8$ is used as originally proposed in [8]. However, since the *a priori* SNR estimate has a major importance under transient noise, the proposed modified decision-directed estimator is applied both for the simultaneous detection and estimation approach and for the STSA algorithm with $\xi_{min} = -20$ dB. For the simultaneous detection and estimation algorithm $\alpha = 0.92$ is used while for the STSA algorithm $\alpha = 0.98$ (as shown in Section VII-A to be more appropriate for the STSA estimator). For the OM-LSA algorithm, the decision-directed estimator with $\alpha = 0.92$ is implemented as specified in [10], and the gain floor is $G_f = -20$ dB. Fig. 7 shows waveforms and spectrograms of a clean signal, noisy signal, and enhanced signals. The noisy signal contains engine car noise with 0-dB

[2]A. Abramson homepage. [Online]. Available: http://siglab.technion.ac.il/~ari_a

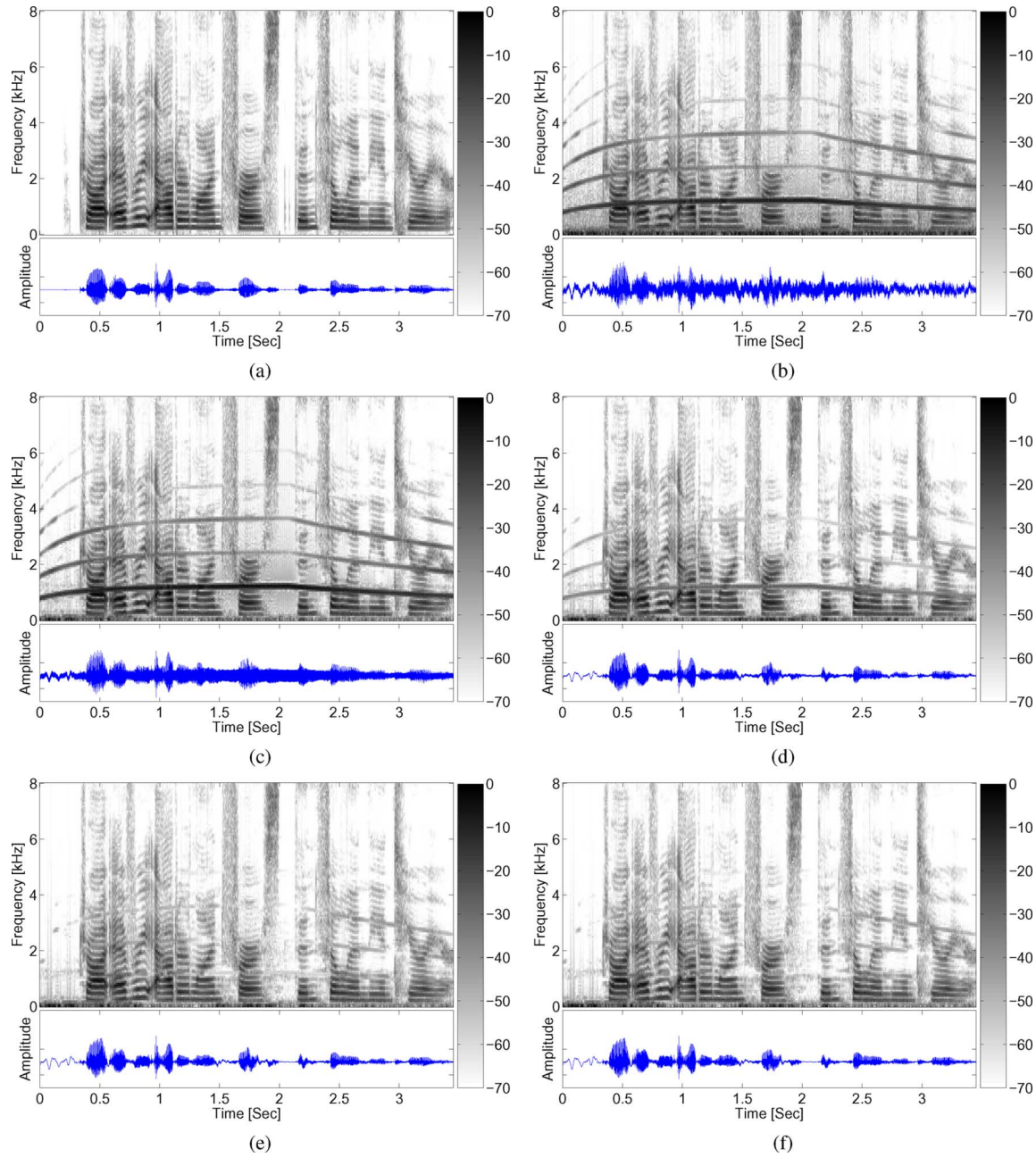

Fig. 7. Speech spectrograms (in dB) and waveforms. (a) Clean speech signal: "Draw every outer line first, then fill in the interior." (b) Speech degraded by engine car noise and siren noise with SNR of -3 dB. (c) Speech enhanced by using the OM-LSA estimator. (d) Speech enhanced by using the STSA estimator (together with the modified *a priori* SNR estimator). (e) Speech enhanced by using the simultaneous detection and estimation approach with $b_{01} = b_{10} = 1.5$. (f) Speech enhanced by using the simultaneous detection and estimation approach with $b_{01} = b_{10} = 5$.

SNR and additional siren noise with -1-dB SNR, such that the total SNR is about -3 dB. The speech enhanced by using the OM-LSA algorithm and the STSA algorithm are shown in Fig. 7(c) and (d), respectively. The signal enhanced by using the simultaneous detection and estimation approach is shown in Fig. 7(e) and (f) with $b_{01} = b_{10} = 1.5$ and $b_{01} = b_{10} = 5$, respectively, and a gain floor of $G_f = -20$ dB. It can be seen that compared with the decision-directed-based OM-LSA algorithm, the modified *a priori* SNR estimator substantially contributes to the transient noise reduction, whether it is integrated with the simultaneous detection and estimation approach or with the STSA algorithm. However, the simultaneous detection and estimation approach which is combined with adapted speech presence probability and gain floor yields greater reduction of

TABLE II
SEGMENTAL SNR, LOG SPECTRAL DISTORTION, AND PESQ SCORE UNDER TRANSIENT NOISE

	SegSNR	LSD	PESQ
Input Signal	-6.703	6.587	2.017
OM-LSA	-4.94	5.338	2.141
STSA	4.502	3.580	2.839
Detection and estimation $b_{01} = b_{10} = 1.5$	5.761	3.236	3.072
Detection and estimation $b_{01} = b_{10} = 5$	6.506	3.141	3.071

transient noise without affecting the quality of the enhanced speech signal. Averaged quality measures for the whole set of tested utterances are summarized in Table II, for the same noise conditions. The results demonstrate improved speech quality obtained by using the modified *a priori* SNR estimator either while combined with the STSA or the simultaneous detection and estimation approach, applying the detection and estimation approach introduced additional improvement to the enhanced signal. Subjective listening tests confirm that the speech quality improvement achieved by using the proposed method is perceptually substantial (audio files are available online).

VIII. Conclusion

We have presented a novel formulation of the single-channel speech enhancement problem in the time–frequency domain. Our formulation relies on coupled operations of detection and estimation in the STFT domain and a cost function that combines both the estimation and detection errors. A detector for the speech coefficients and a corresponding estimator for their values are jointly designed to minimize a combined Bayes risk. In addition, cost parameters enable to control the tradeoff between speech quality, noise reduction, and residual musical noise. The proposed method generalizes the traditional spectral enhancement approach which considers estimation-only under signal presence uncertainty. In addition we propose a modified decision-directed *a priori* SNR estimator which is adapted to transient noise environment. Experimental results show greater noise reduction with improved speech quality when compared with the STSA suppression rules under stationary noise. Furthermore, it is demonstrated that under transient noise environment, greater reduction of transient noise components may be achieved by exploiting reliable information for the *a priori* SNR estimation with simultaneous detection and estimation approach.

Appendix

In this appendix, we derive the risk $r_{1j}(Y)$. Under $\{H_1, \eta_j\}$ we obtain

$$r_{1j}(Y) = b_{1j} \int_0^\infty \int_0^{2\pi} (a - G_j R)^2 p(a, \alpha | H_1) p(Y | a, \alpha) d\alpha da \tag{31}$$

and the multiplication of the two pdf's implies

$$p(a, \alpha | H_1) p(Y | a, \alpha) = \frac{a}{\pi^2 \lambda_x \lambda_d} \times \exp\left\{ -\left(\gamma + \frac{a^2}{\lambda} - \frac{2Ra\cos(\alpha - \theta)}{\lambda_d} \right) \right\} \tag{32}$$

where $\lambda \triangleq (1/\lambda_x + 1/\lambda_d)^{-1}$. Integrating (31) with regard to the phase variable we obtain [36, eq. 3.339, 8.406.3]

$$\int_0^{2\pi} \exp\left\{ \frac{2Ra\cos(\alpha - \theta)}{\lambda_d} \right\} d\alpha = 2\pi J_0\left(i \frac{2R}{\lambda_d} a \right) \tag{33}$$

where $J_0(\cdot)$ denotes the Bessel function of order zero. Note that in this appendix $i \triangleq \sqrt{-1}$. Using [37, eq. 13.3.1, 2] we have

$$\int_0^\infty a \exp\left(-\frac{a^2}{\lambda} \right) J_0\left(i \frac{2R}{\lambda_d} a \right) da = \frac{\lambda}{2} e^{\upsilon} \tag{34}$$

and

$$\int_0^\infty a^2 \exp\left(-\frac{a^2}{\lambda} \right) J_0\left(i \frac{2R}{\lambda_d} a \right) da = \frac{\lambda^{1.5} \Gamma(1.5)}{2\Gamma(1)} {}_1F_1(1.5; 1; \upsilon) \tag{35}$$

where $\Gamma(\cdot)$ denotes the Gamma function with $\Gamma(1) = 1$ and $\Gamma(1.5) = \sqrt{\pi}/2$, and ${}_1F_1(a; b; x)$ is the confluent hypergeometric function [38, eq. A.1.31.c]

$${}_1F_1(1.5; 1; \upsilon) = e^{\frac{\upsilon}{2}} \left[(1 + \upsilon) I_0\left(\frac{\upsilon}{2} \right) + \upsilon I_1\left(\frac{\upsilon}{2} \right) \right]. \tag{36}$$

Using [37, eq. 13.3.2], [38, eq. A.1.19.c], we obtain

$$\int_0^\infty a^3 \exp\left(-\frac{a^2}{\lambda} \right) J_0\left(i \frac{2R}{\lambda_d} a \right) da = \frac{\lambda^2 \Gamma(2)}{2\Gamma(1)} {}_1F_1(2; 1; \upsilon) = \frac{\lambda^2}{2} (1 + \upsilon) e^{\upsilon}. \tag{37}$$

Substituting (32)–(37) into (31) yields

$$\begin{aligned} r_{1j}(Y) &= \frac{b_{1j}}{\pi} \frac{\exp\left(-\frac{\gamma}{1+\xi} \right)}{1 + \xi} \\ &\quad \times \left\{ G_j^2 \gamma + \frac{\xi}{1 + \xi} (1 + \upsilon) - G_j \sqrt{\pi \upsilon} \exp\left(-\frac{\upsilon}{2} \right) \right. \\ &\quad \left. \times \left[(1 + \upsilon) I_0\left(\frac{\upsilon}{2} \right) + \upsilon I_1\left(\frac{\upsilon}{2} \right) \right] \right\} \\ &= \frac{b_{1j}}{\pi} \frac{\exp\left(-\frac{\gamma}{1+\xi} \right)}{1 + \xi} \\ &\quad \times \left\{ G_j^2 \gamma + \frac{\xi}{1 + \xi} (1 + \upsilon) - 2\gamma G_j G_{\mathrm{STSA}} \right\}. \end{aligned} \tag{38}$$

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions.

References

[1] S. F. Boll, "Suppression of acousting noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP'79*, Apr. 1979, vol. 4, pp. 208–211.

[3] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[4] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.

[5] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.
[6] F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," in *Proc. 27th IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'02*, Orlando, FL, May 2002, pp. 569–572.
[7] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.
[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[9] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments," in *Proc. 24th IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'99*, Phoenix, AZ, Mar. 1999, pp. 789–792.
[10] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary environments," *Signal Process.*, vol. 81, pp. 2403–2418, Nov. 2001.
[11] I. Cohen and B. Berdugo, "Multichannel signal detection based on transient beam-to-reference ratio," *IEEE Signal Process. Lett.*, vol. 10, no. 9, pp. 259–262, Sep. 2003.
[12] A. Subramanya, M. L. Seltzer, and A. Acero, "Automatic removal of typed keystrokes from speech signals," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 363–366, May 2007.
[13] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. 23rd IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP'98*, Seattle, WA, May 1998, vol. 1, pp. 365–368.
[14] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[15] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett.*, vol. 8, no. 10, pp. 276–278, Oct. 2001.
[16] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498–505, Sep. 2003.
[17] A. Davis, S. Nordholm, and R. Tongneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 412–423, Mar. 2006.
[18] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.
[19] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Process. Lett.*, vol. 11, no. 12, pp. 956–959, Dec. 2004.
[20] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
[21] W. A. Harrison, J. S. Lim, and E. Singer, "A new application of adaptive noise cancellation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 1, pp. 21–27, Feb. 1986.
[22] D. Middleton and F. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 434–444, May 1968.
[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.
[24] A. Fredriksen, D. Middleton, and D. Vandelinde, "Simultaneous signal detection and estimation under multiple hypotheses," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 5, pp. 607–614, 1972.
[25] A. G. Jaffer and S. C. Gupta, "Coupled detection-estimation of Gaussian processes in gaussian noise," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 106–110, Jan. 1972.
[26] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 68, no. 4, pp. 698–709, Apr. 2006.
[27] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[28] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[29] A. Abramson and I. Cohen, "Enhancement of speech signals under multiple hypotheses using an indicator for transient noise presence," in *Proc. 32nd IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP'07*, Honolulu, HI, Apr. 2007, pp. IV-553–IV-556.
[30] E. Habets, I. Cohen, and S. Gannot, "MMSE log-spectral amplitude estimator for multiple interferences," in *Proc. Int. Workshop Acoust. Echo Noise Control., IWAENC'06*, Paris, France, Sep. 2006.
[31] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Inst. Stand. Technol. (NIST), Gaithersburg, MD, Tech. Rep., (prototype as of December 1988).
[32] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[33] *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, ITU-T Rec. P.862, Int. Telecomm. Union, Geneva, Switzerland, Feb. 2001.
[34] A. Guérin, G. Faucon, and R. L. Bouquin-Jeannès, "Nonlinear acoustic echo cancellation based on voltera filters," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 672–683, Nov. 2003.
[35] E. Haensler and G. Schmidt, Eds., *Topics in Acoustic Echo and Noise Control*. New York: Springer-Verlag, 2006.
[36] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, A. Jefferey and D. Zwillinger, Eds., 6th ed. New York: Academic, 2000.
[37] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1962.
[38] D. Middleton, *An Introduction to Statistical Comunication Theory*, 2nd ed. Piscataway, NJ: IEEE Press, 1996.

Ari Abramson (S'06) received the B.Sc. degree in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 2002. He is currently pursuing the Ph.D. degree in electrical engineering in the direct-tract doctoral program at The Technion—Israel Institute of Technology, Haifa.

From 1993 to 2004, he served as a combat copilot in the Israeli Air Force, and since 2004 he has been a flight-test engineer on reserve duty. His research interests are statistical signal processing, speech enhancement and detection, and estimation theory.

Mr. Abramson received the Wolf Foundation Excellence Award in 2005 and the Best Student Paper Award at the International Workshop on Acoustic, Echo, and Noise Control in 2006.

Israel Cohen (M'01–SM'03) received the B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees in electrical engineering from The Technion—Israel Institute of Technology, Haifa, in 1990, 1993, and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT. In 2001, he joined the Electrical Engineering Department, The Technion, where he is currently an Associate Professor. His research interests are in statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification, and adaptive filtering.

Dr. Cohen received the Technion Excellent Lecturer Awards in 2005 and 2006. He serves as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS, and as Guest Editor of a special issue of the *EURASIP Journal on Applied Signal Processing* on advances in multimicrophone speech processing and a special issue of the *EURASIP Speech Communication Journal* on speech enhancement. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2007).