Relative Transfer Function Identification Using Speech Signals

Israel Cohen, *Senior Member, IEEE*

***Abstract*—An important component of a multichannel hands-free communication system is the identification of the relative transfer function between sensors in response to a desired source signal. In this paper, a robust system identification approach adapted to speech signals is proposed. A weighted least-squares optimization criterion is introduced, which considers the uncertainty of the desired signal presence in the observed signals. An asymptotically unbiased estimate for the system's transfer function is derived, and a corresponding recursive online implementation is presented. We show that compared to a competing nonstationarity-based method, a smaller error variance is achieved and generally shorter observation intervals are required. Furthermore, in the case of a time-varying system, faster convergence and higher reliability of the system identification are obtained by using the proposed method than by using the nonstationarity-based method. Evaluation of the proposed system identification approach is performed under various noise conditions, including simulated stationary and nonstationary white Gaussian noise, and car interior noise in real pseudo-stationary and nonstationary environments. The experimental results confirm the advantages of proposed approach.**

***Index Terms*—Acoustic noise measurement, adaptive signal processing, array signal processing, signal detection, spectral analysis, speech enhancement, system identification.**

I. Introduction

An IMPORTANT component of a multichannel hands-free communication system is the identification of the relative transfer function (RTF) between sensors in response to a desired source signal [1]–[3]. This transfer function, often referred to as the acoustical transfer function ratio [2], represents the coupling between sensors in response to a desired source. In reverberant and noisy environments, the RTF identification enables one to construct adaptive blocking channels and noise cancellers [4]. The blocking channel is used for blocking the desired signal and deriving a reference noise signal, and the noise canceller is used for eliminating directional or coherent noise sources. The RTF identification also facilitates multichannel signal detection and postfiltering techniques, which employ the transient power ratio between the beamformer output and the reference signals [5], [6].

Shalvi and Weinstein [1] have proposed to identify the RTF between sensors by using the nonstationarity of the desired signal. They assumed that the sensors contain additive interfering signals whose cross-correlation function is stationary, while the autocorrelation function of the desired signal is highly nonstationary. Then, dividing the observation interval into a sequence of subintervals, and computing for each subinterval the cross power spectral density (PSD) of the sensors, they obtained an overdetermined set of equations for the two unknown quantities: the RTF and the (presumably stationary) cross-PSD of the primary sensor and a noise component. An asymptotically unbiased estimate for the RTF was derived by using a weighted least-squares (WLS) approach for minimizing the error variance under certain assumptions.

A major limitation of the nonstationarity-based system identification is that both the RTF identification and noise estimation are carried out through the same WLS optimization criterion. The WLS optimization consists of two conflicting requirements: One is minimizing the error variance of the system's transfer function estimate, which pulls the weight up to higher values on higher SNR subintervals. The other requirement is minimizing the error variance of the noise estimate, which rather implies *smaller* weights on higher SNR subintervals. Another major limitation of this method is that the observation interval should be adequately long, so that for all frequency bands it includes a few subintervals that contain the desired signal. Unfortunately, in case the desired signal is speech, in some frequency bands the presence of speech may be sparse, which entails a very long observation interval. Furthermore, the RTF is assumed to be constant during the observation interval. Hence, very long observation intervals also restrict the capability of this technique to track time-varying systems (e.g., tracking moving talkers in hands-free communication scenarios [7]–[9]). Additionally, a fundamental assumption is that the interfering signals remain stationary during the entire observation interval. This is a very restrictive assumption, particularly in view of the generally long observation interval required for obtaining a reliable RTF identification in the case of speech signals.

In this paper, a robust system identification approach adapted to speech signals is proposed. The speech presence probability in the time-frequency domain is incorporated into the optimization criteria for RTF identification and noise spectra estimation. An RTF estimate is derived based on subintervals that contain speech, while subintervals that do not contain speech are of more significance when estimating the noise spectra. The estimate for the auto-PSD of the desired signal is obtained by applying a first-order recursive smoothing to its *optimally modified log-spectral amplitude* (OM-LSA) estimate [10]. The cross-PSD of the interfering signals is estimated by using the *minima controlled recursive averaging* (MCRA) approach [11],

Manuscript received June 18, 2003; revised March 22, 2004. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Walter Kellermann.

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

Digital Object Identifier 10.1109/TSA.2004.832975

1063-6676/04$20.00 © 2004 IEEE

[12]. Subsequently, minimum variance WLS estimate [13] for the system's transfer function is derived, and a recursive online solution is obtained based on the normalized least-mean-square (LMS) algorithm. We show that the error variance obtained by using the proposed method is generally smaller than that obtained by using the nonstationarity method. Furthermore, the contribution of a given time-frequency bin to the error-variance minimization depends on the relative power of the desired signal in that bin. The higher the SNR is, the shorter the observation interval required for obtaining a reliable RTF identification. Whereas the nonstationarity method requires a relatively long observation interval, regardless of the SNR, to retain the desired signal sufficiently nonstationary. Moreover, in contrast to the nonstationarity method, in the proposed method the statistical properties of the interfering signals are allowed to change during time-frequency windows that do not contain desired signal components. Accordingly, in the case of a time-varying system, faster convergence and higher reliability of the RTF identification are achieved by using the proposed method. Evaluation of the proposed method is performed under various noise conditions, including simulated stationary and nonstationary white Gaussian noise, and real car interior noise in pseudo-stationary and nonstationary environments. The experimental results confirm that the proposed algorithm is advantageous to the nonstationarity-based algorithm.

The paper is organized as follows. In Section II, we formulate the system identification problem. In Section III, we review the nonstationarity-based system identification technique, which heavily relies on the stationarity of the interfering signals and nonstationarity of the desired signal. In Section IV, we introduce a system identification approach that is more appropriate to speech signals. The optimal estimate for the system's transfer function is derived based on time-frequency bins which contain the desired signal components. In Section V, we describe the system identification algorithm and its online implementation. Finally, in Section VI, we present experimental results, which demonstrate the improvement gained by the proposed approach.

II. Problem Formulation

Let $s(t)$ denote a desired source signal, and let $u(t)$ and $w(t)$ denote additive interfering signals that are uncorrelated with the desired signal. The signals measured by a primary and reference sensors are given by

$$x(t) = s(t) + u(t) \tag{1}$$

$$y(t) = a(t) * s(t) + w(t) \tag{2}$$

where $a(t)$ represents the coupling of the desired signal to the reference sensor, and $*$ denotes convolution. Our objective is to identify $a(t)$ in the general case where $u(t)$ is statistically correlated with $w(t)$.

It worth noting that $s(t)$ is often a reverberated version of the source signal, i.e., $s(t) = a_0(t) * s_0(t)$, where $s_0(t)$ is the source signal and $a_0(t)$ is the impulse response of the primary sensor to the desired source. In that case, $a_1(t) \triangleq a(t) * a_0(t)$ represents the impulse response of the *reference* sensor to the desired source, and $a(t)$ represents the *relative* impulse response between the reference and primary sensors with respect to the desired source.

An equivalent problem is to consider a linear time-invariant (LTI) system, whose input $x(t)$ and output $y(t)$ are related by

$$y(t) = a(t) * x(t) + v(t) \tag{3}$$

where $a(t)$ represents the impulse response of the system that we want to identify, and $v(t)$ denotes additive noise. The system input is assumed to be the sum of a desired signal $s(t)$ and a noise signal $u(t)$ as in (1). The signal $s(t)$ is assumed statistically uncorrelated with $u(t)$ and $v(t)$.

It is easy to verify that the two above-mentioned problems are equivalent, with the following relation between the interfering signals:

$$v(t) = w(t) - a(t) * u(t). \tag{4}$$

Equation (4) reveals that not only $v(t)$ is generally correlated with the system input $x(t)$ (both contain $u(t)$), but also depends on the impulse response of the system. Therefore, conventional system identification techniques, which assume that $v(t)$ is independent of $x(t)$ and $a(t)$, are inapplicable.

III. System Identification Using Nonstationarity

In this section, we review the system identification technique of Shalvi and Weinstein [1]. This method heavily relies on the assumption that $v(t)$ is stationary, the desired signal $s(t)$ is nonstationary, and that the support of $a(t)$ is finite.[1]

Dividing the observation interval into M subintervals, such that the support of each subinterval is sufficiently large compared with the duration of $a(t)$, and computing for each subinterval m $(m = 1, 2, \ldots, M)$ the cross-PSD between y and x, we obtain from (3)

$$\phi_{yx}^{(m)}(\omega) = A(\omega)\phi_{xx}^{(m)}(\omega) + \phi_{vx}(\omega) \tag{5}$$

where $A(\omega)$ is the Fourier transform of $a(t)$ (*i.e.,* the RTF of the system), and $\phi_{vx}(\omega)$ is independent of the subinterval index m due to the stationarity of $v(t)$ and $u(t)$, and the lack of correlation between $v(t)$ and $s(t)$. Let $\hat{\phi}_{yx}^{(m)}(\omega)$, $\hat{\phi}_{xx}^{(m)}(\omega)$, and $\hat{\phi}_{vx}^{(m)}(\omega)$ be estimates for $\phi_{yx}^{(m)}(\omega)$, $\phi_{xx}^{(m)}(\omega)$, and $\phi_{vx}(\omega)$, respectively. Then

$$\begin{aligned}\hat{\phi}_{yx}^{(m)}(\omega) &= A(\omega)\hat{\phi}_{xx}^{(m)}(\omega) + \hat{\phi}_{vx}^{(m)}(\omega)\\ &= A(\omega)\hat{\phi}_{xx}^{(m)}(\omega) + \phi_{vx}(\omega) + \epsilon^{(m)}(w)\end{aligned} \tag{6}$$

where

$$\epsilon^{(m)}(w) = \hat{\phi}_{vx}^{(m)}(\omega) - \phi_{vx}(\omega). \tag{7}$$

[1] Note that $a(t)$ is generally of infinite length, since it represents the impulse response of the ratio of room transfer functions. However, in real environments, the energy of $a(t)$ often decays exponentially for $t > T_D$, where T_D depends on the reverberation time [2]. Therefore, the finite support assumption is practically not very restrictive.

This can be written in a matrix form as

$$
\begin{aligned}
\mathbf{z} &\triangleq \begin{bmatrix} \hat{\phi}_{yx}^{(1)}(\omega) \\ \hat{\phi}_{yx}^{(2)}(\omega) \\ \vdots \\ \hat{\phi}_{yx}^{(M)}(\omega) \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{xx}^{(1)}(\omega) & 1 \\ \hat{\phi}_{xx}^{(2)}(\omega) & 1 \\ \vdots & \vdots \\ \hat{\phi}_{xx}^{(M)}(\omega) & 1 \end{bmatrix} \begin{bmatrix} A(\omega) \\ \phi_{vx}(\omega) \end{bmatrix} \\
&\quad + \begin{bmatrix} \epsilon^{(1)}(w) \\ \epsilon^{(2)}(w) \\ \vdots \\ \epsilon^{(M)}(w) \end{bmatrix} \\
&\triangleq \hat{\mathbf{G}}\boldsymbol{\theta} + \boldsymbol{\epsilon}.
\end{aligned} \tag{8}
$$

The WLS estimate of $\boldsymbol{\theta}$ is obtained by

$$
\begin{aligned}
\begin{bmatrix} \hat{A}(\omega) \\ \hat{\phi}_{vx}(\omega) \end{bmatrix} &= \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} (\mathbf{z} - \hat{\mathbf{G}}\boldsymbol{\theta})^H \mathbf{W} (\mathbf{z} - \hat{\mathbf{G}}\boldsymbol{\theta}) \\
&= (\hat{\mathbf{G}}^H \mathbf{W} \hat{\mathbf{G}})^{-1} \hat{\mathbf{G}}^H \mathbf{W} \mathbf{z}
\end{aligned} \tag{9}
$$

where $\mathbf{W}$ is a positive Hermitian weight matrix, ^{H} denotes conjugate-transpose, and $\hat{\mathbf{G}}^H \mathbf{W} \hat{\mathbf{G}}$ is required to be invertible.

Shalvi and Weinstein suggested two choices of a weight matrix. One choice is given by

$$
W_{mn} = \begin{cases} T_m, & m = n \\ 0, & m \neq n \end{cases} \tag{10}
$$

where T_m is the length of subinterval m, so that longer intervals obtain higher weights. In this case, (9) reduces to

$$
\hat{A}(\omega) = \frac{\langle \hat{\phi}_{yx}(\omega) \hat{\phi}_{xx}(\omega) \rangle - \langle \hat{\phi}_{yx}(\omega) \rangle \langle \hat{\phi}_{xx}(\omega) \rangle}{\langle \hat{\phi}_{xx}^2(\omega) \rangle - \langle \hat{\phi}_{xx}(\omega) \rangle^2} \tag{11}
$$

with the average operation defined by

$$
\langle \varphi(\omega) \rangle \triangleq \frac{\sum_{m=1}^{M} T_m \varphi^{(m)}(\omega)}{\sum_{m=1}^{M} T_m}. \tag{12}
$$

Another choice of $\mathbf{W}$ that minimizes the covariance of $\hat{\boldsymbol{\theta}}$ is given by

$$
W_{mn} = \begin{cases} \frac{T_m}{\hat{\phi}_{xx}^{(m)}(\omega)} & m = n \\ 0 & m \neq n. \end{cases} \tag{13}
$$

In which case, (9) yields

$$
\hat{A}(\omega) = \frac{\left\langle \frac{1}{\hat{\phi}_{xx}(\omega)} \right\rangle \langle \hat{\phi}_{yx}(\omega) \rangle - \left\langle \frac{\hat{\phi}_{yx}(\omega)}{\hat{\phi}_{xx}(\omega)} \right\rangle}{\langle \hat{\phi}_{xx}(\omega) \rangle \left\langle \frac{1}{\hat{\phi}_{xx}(\omega)} \right\rangle - 1} \tag{14}
$$

and the variance of $\hat{A}(\omega)$ is given by

$$
\mathrm{var}\{\hat{A}(\omega)\} = \frac{1}{BT} \cdot \frac{\phi_{vv}(\omega) \left\langle \frac{1}{\phi_{xx}(\omega)} \right\rangle}{\langle \phi_{xx}(\omega) \rangle \left\langle \frac{1}{\phi_{xx}(\omega)} \right\rangle - 1} \tag{15}
$$

where $T \triangleq \sum_{m=1}^{M} T_m$ is the total observation interval, and B is related to the window's bandwidth that is preselected for the empirical cross-spectrum estimation [1].

A major limitation of the WLS optimization in (9) is that both the identification of $A(\omega)$ and the estimation of the cross-PSD $\phi_{vx}(\omega)$ are carried out using the same weight matrix $\mathbf{W}$. That is, each subinterval m is given the same weight, whether we are trying to find an estimate for $A(\omega)$ or for $\phi_{vx}(\omega)$. However, subintervals with higher SNRs are of greater importance when estimating $A(\omega)$, whereas the opposite is true when estimating $\phi_{vx}(\omega)$. Consequently, the optimization criterion in (9) consists of two conflicting requirements. One is minimizing the error variance of $\hat{A}(\omega)$, which pulls the weight up to higher values on higher SNR subintervals. The other requirement is minimizing the error variance of $\hat{\phi}_{vx}(\omega)$, which rather implies *smaller* weights on higher SNR subintervals. For instance, suppose we obtain observations on a relatively long low-SNR interval of length T_0, and on a relatively short high-SNR interval of length T_1 ($T_1 \ll T_0$). Then, the variance of $\hat{A}(\omega)$ in (15) is inversely proportional to the relative length of the high-SNR interval, $T_1/(T_0 + T_1)$. That is, the observation includes interval additional segments that do not contain speech (*i.e.,* increasing T_0) increases the variance of $\hat{A}(\omega)$. This unnatural consequence is a result of the desire to minimize the variance of $\hat{\phi}_{vx}(\omega)$ by using larger weights on the segments that do not contain speech, while increasing the weights on such subintervals degrades the estimate for $A(\omega)$.

Another major limitation of RTF identification using nonstationarity is that the interfering signals are required to be stationary during the entire observation interval. The observation interval should include a certain number of subintervals that contain the desired signal, such that $\phi_{xx}(\omega)$ is sufficiently nonstationary for all ω. Unfortunately, in case the desired signal is speech, the presence of the desired signal in the observed signals may be sparse in some frequency bands. This entails a very long observation interval, thus constraining the interfering signals to be stationary over long intervals. Furthermore, the RTF $A(\omega)$ is assumed to be constant during the observation interval. Hence, very long observation intervals also restrict the capability of the system identification technique to track varying $A(\omega)$ (e.g., tracking moving talkers in reverberant environments).

IV. System Identification Using Speech Signals

In this section, we propose a system identification approach that is adapted to speech signals. Specifically, we assume that the presence of the desired speech signal in the time-frequency domain is uncertain, and employ the speech presence probability to separate the tasks of system identification and cross-PSD estimation. An estimate for $A(\omega)$ is derived based on subintervals that contain speech, while subintervals that do not contain speech are of more significance when estimating the components of $\phi_{vx}(\omega)$.

Let the observed signals be divided in time into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). Assuming the support of $a(t)$ is finite, and that the support of the window function is sufficiently large compared with the duration of $a(t)$, (3) can be written in the time-frequency domain as

$$Y(k,\ell) = A(k)X(k,\ell) + V(k,\ell) \tag{16}$$

where $A(k)$ is the RTF of the system, k represents the frequency bin index ($k = 1,2,\ldots,K$), and ℓ is the frame index ($\ell = 1,2,\ldots,L$). Thus, similar to (5), we have

$$\phi_{yx}(k,\ell) = A(k)\phi_{xx}(k,\ell) + \phi_{vx}(k,\ell). \tag{17}$$

Since the desired signal $s(t)$ is uncorrelated with the interfering signals $u(t)$ and $w(t)$, (1) and (4) imply

$$\phi_{yx}(k,\ell) = A(k)\phi_{ss}(k,\ell) + \phi_{wu}(k,\ell). \tag{18}$$

Writing this equation in terms of the PSD estimates, we obtain

$$\hat{\phi}_{yx}(k,\ell) - \hat{\phi}_{wu}(k,\ell) = A(k)\hat{\phi}_{ss}(k,\ell) + \varepsilon(k,\ell) \tag{19}$$

where $\varepsilon(k,\ell)$ denotes an estimation error. This gives us L equations, which may be written in a matrix form as

$$\begin{aligned}\hat{\boldsymbol{\psi}}(k) &\triangleq \begin{bmatrix} \hat{\phi}_{yx}(k,1) - \hat{\phi}_{wu}(k,1) \\ \hat{\phi}_{yx}(k,2) - \hat{\phi}_{wu}(k,2) \\ \vdots \\ \hat{\phi}_{yx}(k,L) - \hat{\phi}_{wu}(k,L) \end{bmatrix} \\ &= \begin{bmatrix} \hat{\phi}_{ss}(k,1) \\ \hat{\phi}_{ss}(k,2) \\ \vdots \\ \hat{\phi}_{ss}(k,L) \end{bmatrix} A(k) + \begin{bmatrix} \varepsilon(k,1) \\ \varepsilon(k,2) \\ \vdots \\ \varepsilon(k,L) \end{bmatrix} \\ &\triangleq \hat{\boldsymbol{\phi}}_{ss}(k)A(k) + \boldsymbol{\varepsilon}(k). \end{aligned} \tag{20}$$

Since the RTF $A(k)$ represents the coupling between the primary and reference sensors with respect to the *desired* source signal, the optimization criterion for the identification of $A(k)$ has to take into account only short-time frames which contain desired signal components. Specifically, let $I(k,\ell)$ denote an indicator function for the signal presence (*i.e.,* $I(k,\ell) = 1$ if $\phi_{ss}(k,\ell) \neq 0$, and $I(k,\ell) = 0$ otherwise), and let $\mathbf{I}(k)$ represent a diagonal matrix with the elements $[I(k,1), I(k,2),\ldots,I(k,L)]$ on its diagonal. Then the WLS estimate of $A(k)$ is obtained by

$$\begin{aligned}\hat{A} &= \arg\min_{A}\{[\mathbf{I}\boldsymbol{\varepsilon}]^H \mathbf{W}\,[\mathbf{I}\boldsymbol{\varepsilon}]\} \\ &= \arg\min_{A}\left\{\left[\hat{\boldsymbol{\psi}} - \hat{\boldsymbol{\phi}}_{ss}A\right]^H \mathbf{IWI}\left[\hat{\boldsymbol{\psi}} - \hat{\boldsymbol{\phi}}_{ss}A\right]\right\} \\ &= \left[\hat{\boldsymbol{\phi}}_{ss}^T \mathbf{IWI}\hat{\boldsymbol{\phi}}_{ss}\right]^{-1}\hat{\boldsymbol{\phi}}_{ss}^T\mathbf{IWI}\hat{\boldsymbol{\psi}} \end{aligned} \tag{21}$$

where the argument k has been omitted for notational simplicity. Recognizing the product $\mathbf{IWI}$ as the equivalent weight matrix, the variance of $\hat{A}$ is given by [14, p. 405]

$$\begin{aligned}\mathrm{var}\left\{\hat{A}\right\} = \left(\boldsymbol{\phi}_{ss}^T\mathbf{IWI}\boldsymbol{\phi}_{ss}\right)^{-1}\boldsymbol{\phi}_{ss}^T\mathbf{IWI}\mathrm{cov}(\boldsymbol{\varepsilon}) \\ \times\mathbf{IWI}\boldsymbol{\phi}_{ss}\left(\boldsymbol{\phi}_{ss}^T\mathbf{IWI}\boldsymbol{\phi}_{ss}\right)^{-1}\end{aligned} \tag{22}$$

where $\mathrm{cov}(\boldsymbol{\varepsilon})$ is the covariance matrix of $\boldsymbol{\varepsilon}$. The matrix $\mathbf{W}$ that minimizes the variance of $\hat{A}$ therefore, satisfies [14, prop. 8.2.4]

$$\mathbf{IWI} = \mathbf{I}\,[\mathrm{cov}(\boldsymbol{\varepsilon})]^{-1}\,\mathbf{I}. \tag{23}$$

This choice of $\mathbf{W}$ yields an asymptotically unbiased estimator

$$\hat{A} = \left(\hat{\boldsymbol{\phi}}_{ss}^T\mathbf{I}\,[\mathrm{cov}(\boldsymbol{\varepsilon})]^{-1}\,\mathbf{I}\hat{\boldsymbol{\phi}}_{ss}\right)^{-1}\hat{\boldsymbol{\phi}}_{ss}^T\mathbf{I}\,[\mathrm{cov}(\boldsymbol{\varepsilon})]^{-1}\,\mathbf{I}\hat{\boldsymbol{\psi}} \tag{24}$$

which is known as the *minimum variance* or *Markov estimator*. Substituting (23) into (22), we obtain the variance of the resulting estimator

$$\mathrm{var}\left\{\hat{A}\right\} = \left(\boldsymbol{\phi}_{ss}^T\mathbf{I}[\mathrm{cov}(\boldsymbol{\varepsilon})]^{-1}\mathbf{I}\boldsymbol{\phi}_{ss}\right)^{-1}. \tag{25}$$

The elements of $\mathrm{cov}(\boldsymbol{\varepsilon})$ are asymptotically given by (see Appendix I)

$$\mathrm{cov}\left(\varepsilon(k,\ell),\varepsilon(k,\ell')\right) = \begin{cases}\phi_{ss}(k,\ell)\phi_{vv}(k,\ell), & \text{if } \ell = \ell' \\ 0 & \text{otherwise.}\end{cases} \tag{26}$$

Assuming that the interfering signals $w(t)$ and $u(t)$ are stationary, (4) implies that $\phi_{vv}(k,\ell)$ is independent of the frame index ℓ (in practice, as demonstrated in Section VI, it suffices that the statistics of the interfering signals is slowly changing compared with the statistics of the desired signal). Denoting by $\langle\cdot\rangle_\ell$ an average operation over the frame index ℓ

$$\langle\varphi(k,\ell)\rangle_\ell \triangleq \frac{1}{L}\sum_{\ell=1}^{L}\varphi(k,\ell) \tag{27}$$

and substituting (26) into (24) and (25), we obtain

$$\hat{A}(k) = \frac{\langle I(k,\ell)[\hat{\phi}_{yx}(k,\ell) - \hat{\phi}_{wu}(k,\ell)]\rangle_\ell}{\langle I(k,\ell)\hat{\phi}_{ss}(k,\ell)\rangle_\ell} \tag{28}$$

$$\mathrm{var}\{\hat{A}(k)\} = \frac{\phi_{vv}(k)}{L\langle I(k,\ell)\phi_{ss}(k,\ell)\rangle_\ell}. \tag{29}$$

Note that for a given frequency-bin index k, only frames that contain speech ($I(k,\ell) \neq 0$) influence the values of $\hat{A}(k)$ and $\mathrm{var}\{\hat{A}(k)\}$. In contrast with the nonstationarity method, including in the observation interval additional segments that do not contain speech does not increase the variance of $\hat{A}(k)$ for any k. However, the proposed identification approach requires an estimate for $I(k,\ell)$, *i.e.,* identifying which time-frequency bins (k,ℓ) contain the desired signal. In practice, the speech presence probability $p(k,\ell)$ is estimated from the noisy signals [10], and an estimate for the indicator function is obtained by

$$\hat{I}(k,\ell) = \begin{cases}1, & \text{if } p(k,\ell) \geq p_0 \\ 0, & \text{otherwise}\end{cases} \tag{30}$$

where p_0 ($0 \leq p_0 < 1$) is a predetermined threshold. The parameter p_0 controls the trade-off between the detection and false alarm probabilities, which are defined by $P_D \triangleq \mathcal{P}\{p(k,\ell) \geq p_0 | I(k,\ell) = 1\}$ and $P_{FA} \triangleq \mathcal{P}\{p(k,\ell) \geq p_0 | I(k,\ell) = 0\}$. A smaller value of p_0 increases the detection probability and allows for more short-time frames to be involved in the estimation of A. However, a smaller value of p_0 also increases the false alarm probability, which may cause a mis-modification of $\hat{A}$ due to frames that do not contain desired speech components.

For the comparison with the nonstationarity method, we replace the subinterval index m in (15) with the frame index ℓ, and normalize the window function so that $BT_0 = 1$ where T_0 is the frame's length. Accordingly, the variance of $\hat{A}$ obtained by using the nonstationarity method is

$$\mathrm{var}\{\hat{A}(k)\}\Big|_{\text{NS method}} = \frac{1}{L} \cdot \frac{\phi_{vv}(k)\left\langle \phi_{xx}^{-1}(k,\ell)\right\rangle_\ell}{\left\langle \phi_{xx}(k,\ell)\right\rangle_\ell \left\langle \phi_{xx}^{-1}(k,\ell)\right\rangle_\ell - 1}. \quad (31)$$

Consequently, the ratio between the variance obtained by the proposed method and that obtained by the nonstationarity method is given by

$$\rho \triangleq \frac{\mathrm{var}\{\hat{A}(k)\}\Big|_{\text{proposed method}}}{\mathrm{var}\{\hat{A}(k)\}\Big|_{\text{NS method}}} = \frac{\langle\phi_{xx}(k,\ell)\rangle_\ell\langle\phi_{xx}^{-1}(k,\ell)\rangle_\ell - 1}{\langle\phi_{xx}^{-1}(k,\ell)\rangle_\ell\langle\hat{I}(k,\ell)\phi_{ss}(k,\ell)\rangle_\ell}. \quad (32)$$

Let $\xi(k,\ell) \triangleq \phi_{ss}(k,\ell)/\phi_{uu}(k)$ denote the *a priori* SNR at the primary sensor. Then approximating

$$\left\langle \hat{I}(k,\ell)\phi_{ss}(k,\ell)\right\rangle \approx \langle I(k,\ell)\phi_{ss}(k,\ell)\rangle = \langle \phi_{ss}(k,\ell)\rangle \quad (33)$$

and substituting $\phi_{xx}(k,\ell) = \phi_{ss}(k,\ell) + \phi_{uu}(k)$ into (32), we obtain (see Appendix II)

$$\rho = \frac{\langle\xi(k,\ell)+1\rangle_\ell\langle[\xi(k,\ell)+1]^{-1}\rangle_\ell - 1}{\langle[\xi(k,\ell)+1]^{-1}\rangle_\ell\langle\xi(k,\ell)\rangle_\ell} < 1. \quad (34)$$

Thus, as long as (33) is satisfied (*i.e.,* desired speech components are sufficiently detected), the variance of $\hat{A}(k)$ obtained by using the proposed method is smaller than that obtained by using the nonstationarity method. Additionally, (29) implies that the contribution of a given time-frequency bin (k,ℓ) to the quality (error variance minimization) of the proposed estimator depends on the desired signal power contained in that bin, $\phi_{ss}(k,\ell)$. The higher the SNR is, the fewer the number of frames required for setting a certain upper limit to the error variance. Whereas with the nonstationarity method, regardless of the SNR, a large number of frames is necessary to account for the nonstationarity of $\phi_{xx}(k,\ell)$. Furthermore, in the nonstationarity method, a fundamental assumption is that the interfering signals remain stationary during the entire observation interval. This is a very restrictive assumption, particularly in view of the generally long observation interval required for obtaining a reliable $A(k)$ estimate by using the nonstationarity method. On the other hand in the proposed method, not only a shorter observation interval suffices, but also the statistical properties of the interfering signals are not required to be time-invariant during time-frequency windows that do not contain desired signal components. Accordingly, in the case of a time-varying system, a faster convergence and higher reliability of the system identification is achieved by using the proposed method.

V. Implementation

Our algorithm requires estimates for $\phi_{yx}(k,\ell)$, $\phi_{ss}(k,\ell)$ and $\phi_{wu}(k,\ell)$. An estimate for $\phi_{yx}(k,\ell)$ is obtained by applying a first-order recursive smoothing to the cross-periodogram of the observed signals, $Y(k,\ell)X^*(k,\ell)$. Specifically

$$\hat{\phi}_{yx}(k,\ell) = \alpha_s\hat{\phi}_{yx}(k,\ell-1) + (1-\alpha_s)Y(k,\ell)X^*(k,\ell) \quad (35)$$

where the smoothing parameter α_s $(0 \leq \alpha_s < 1)$ determines the equivalent number of cross-periodograms that are averaged, $N_\ell \approx (1+\alpha_s)/(1-\alpha_s)$. Typically, speech periodograms are recursively smoothed with an equivalent rectangular window of $T_s = 0.2$ seconds length, which represents a good compromise between smoothing the noise and tracking the speech spectral variations [15]. Therefore, for a sampling rate of 8 kHz, a STFT window length of 256 samples and a frame update step of 128 samples, we use $\alpha_s = (T_s{\cdot}8000/128-1)/(T_s{\cdot}8000/128+1) \approx 0.85$.

To obtain an estimate for the PSD of the desired signal, we first estimate the STFT of the desired signal by using the *optimally modified log-spectral amplitude* (OM-LSA) estimation technique [10]. Subsequently, the periodogram of the desired signal is recursively smoothed

$$\hat{\phi}_{ss}(k,\ell) = \alpha_s\hat{\phi}_{ss}(k,\ell-1) + (1-\alpha_s)\left[G^{p(k,\ell)}(k,\ell)G_{\min}^{1-p(k,\ell)}\left|X(k,\ell)\right|\right]^2 \quad (36)$$

where $G(k,\ell)$ denotes the log-spectral amplitude gain function [16], $p(k,\ell)$ is the speech presence probability [10], and $G_{\min}$ is the minimal spectral gain. The cross-PSD of the interfering signals, $w(t)$ and $u(t)$, is estimated by using the *minima controlled recursive averaging* (MCRA) approach [11], [12]. Specifically, past spectral cross-power values of the noisy observed signals are recursively averaged with a time-varying frequency-dependent smoothing parameter

$$\hat{\phi}_{wu}(k,\ell) = \tilde{\alpha}_u(k,\ell)\hat{\phi}_{wu}(k,\ell-1) + \beta\left[1-\tilde{\alpha}_u(k,\ell)\right]Y(k,\ell)X^*(k,\ell) \quad (37)$$

where $\tilde{\alpha}_u(k,\ell)$ is the smoothing parameter $(0 < \tilde{\alpha}_u(k,\ell) \leq 1)$, and β $(\beta \geq 1)$ is a factor that compensates the bias when the desired signal is absent [12]. The smoothing parameter is determined by the signal presence probability, $p(k,\ell)$, and a constant α_u $(0 < \alpha_u < 1)$ that represents its minimal value

$$\tilde{\alpha}_u(k,\ell) = \alpha_u + (1-\alpha_u)p(k,\ell). \quad (38)$$

The value of $\tilde{\alpha}_u$ is close to 1 when the desired signal is present to prevent the noise cross-PSD estimate from increasing as a result of signal components. It decreases linearly with the probability of signal presence to allow a faster update of the noise estimate. The value of α_u compromises between the tracking rate (response rate to abrupt changes in the noise statistics) and the variance of the noise estimate. Typically, in case of high

TABLE I
SUMMARY OF ONLINE SPEECH-BASED SYSTEM IDENTIFICATION ALGORITHM

```
Initialize variables on the first frame for all frequency bins $k$:
    $\hat{\phi}_{yx}(k,0) = \hat{\phi}_{wu}(k,0) = Y(k,0)\,X^*(k,0)$ ;
    $\hat{\phi}_{ss}(k,0) = 0$ ;   $\hat{A}(k,0) = 1$ .
For all time frames $\ell$
    For all frequency bins $k$
        Compute the recursively averaged cross-periodogram $\hat{\phi}_{yx}(k,\ell)$ by using (35).
        Compute the signal presence probability $p(k,\ell)$ by using [10], the time-varying smoothing parameter $\tilde{\alpha}_u(k,\ell)$ by using (38), and the cross-PSD of the interfering signals $\hat{\phi}_{wu}(k,\ell)$ by using (37).
        Compute the log-spectral amplitude gain function $G(k,\ell)$ by using [16], and the recursively averaged periodogram of the desired signal $\hat{\phi}_{ss}(k,\ell)$ by using (36).
        Compute the estimation error $\hat{\varepsilon}(k,\ell)$ by using (40), and the estimate for $I(k,\ell)$ by using (30).
        Update the estimate for the system's transfer function $\hat{A}(k,\ell)$ by using (39).
```

levels of nonstationary noise, a good compromise is obtained by $\alpha_u = 0.85$ [12].

Substituting the above spectral estimates into (28) we obtain an estimate for $A(k)$. Alternatively, a recursive online solution to (20) based on the normalized LMS algorithm [17] is given by

$$\begin{aligned}\hat{A}(k,\ell) &= \hat{A}(k,\ell-1) - \frac{\mu}{\hat{\phi}^2_{ss}(k,\ell)}\frac{\partial}{\partial A^*} \\ &\quad \times [I(k,\ell)|\hat{\phi}_{yx}(k,\ell) \\ &\qquad - \hat{\phi}_{wu}(k,\ell) - A\hat{\phi}_{ss}(k,\ell)|^2]|_{A=\hat{A}(k,\ell-1)} \\ &= \hat{A}(k,\ell-1) + \mu\hat{I}(k,\ell)\hat{\phi}^{-1}_{ss}(k,\ell)\hat{\varepsilon}(k,\ell) \qquad (39)\end{aligned}$$

where

$$\hat{\varepsilon}(k,\ell) = \hat{\phi}_{yx}(k,\ell) - \hat{\phi}_{wu}(k,\ell) - \hat{A}(k,\ell-1)\hat{\phi}_{ss}(k,\ell) \quad (40)$$

represents the estimation error. Accordingly, the update of $\hat{A}(k,\ell)$ in (39) is carried out only when the time-frequency bin (k,ℓ) contains some desired signal energy (*i.e.,* when $\hat{I}(k,\ell) \neq 0$). The implementation of the online system identification algorithm is summarized in Table I.

VI. EXPERIMENTAL RESULTS

In this section, the proposed system identification approach is compared to the nonstationarity method in various noise environments. The performance evaluation includes simulated stationary and nonstationary white Gaussian noise (WGN), as well as pseudo-stationary and nonstationary noise signals recorded in a car environment. A quantitative comparison between the system identification methods is obtained by evaluating the signal blocking factor (SBF), defined by

$$\mathrm{SBF} = 10\log_{10}\frac{E\{s^2(t)\}}{E\{r^2(t)\}} \quad [\mathrm{dB}] \qquad (41)$$

where $E\{s^2(t)\}$ is the energy contained in the clean speech signal, and $E\{r^2(t)\}$ is the energy contained in the leakage signal

$$r(t) = a(t) * s(t) - \hat{a}(t) * s(t). \qquad (42)$$

The leakage signal represents the difference between the reverberated clean signal at the reference sensor and its estimate $\hat{a}(t) * s(t)$ given the desired signal at the primary sensor. It has a major effect on the amount of distortion introduced by the transfer function GSC [4]. The SBF measure is associated with the capability to block the desired signal and produce a noise-only signal by computing $\hat{v}(t) = y(t) - \hat{a}(t) * x(t)$.

The first experiment was performed on a speech signal (female speaker) sampled at 8 kHz. Similar to the experiment in [1], the noise $u(t)$ is a stationary zero-mean Gaussian process whose average power is lower than the average power of the speech by a factor of 2.5 ($\mathrm{SNR} = 4\ \mathrm{dB}$). The impulse response of the reference sensor to the desired signal is

$$a(t) = \delta(t-6T) - 0.5\delta(t-7T) + 0.25\delta(t-8T)$$

where $T = 125\ \mu\mathrm{s}$ is the sampling period. In addition, the reference sensor noise $w(t)$ is generated by

$$w(t) = g(t) * u(t)$$

where

$$g(t) = -\delta(t) - 0.5\delta(t-T) + 0.1\delta(t-2T).$$

Fig. 1 shows the clean speech signals at the primary and reference sensors, and the observed noisy signals.

We have applied the nonstationarity-based system identification algorithm (14) to a 4-s observation interval (32 000 samples) that was arbitrarily divided into disjoint subintervals of 128 samples length. As is suggested in [2], only subintervals in which speech is active (SNR in the subinterval is greater than 0 dB) were taken into account. The leakage signal $r(t)$ is plotted in Fig. 2(a). The resultant SBF is 9.1 dB.

Fig. 2(b) and (c) show the leakage signals obtained by using the proposed algorithms. Offline speech-based system identification [see (28)] yields a SBF of 18.5 dB, whereas the online speech-based system identification [see (39)] yields a SBF of 13.9 dB. Both algorithms achieve a significantly higher SBF than the nonstationarity-based algorithm.

In the second experiment, a nonstationary WGN $u(t)$ was simulated by increasing the stationary WGN at a rate of 6 dB/s for a period of 2 s, and then decreasing it back to the original

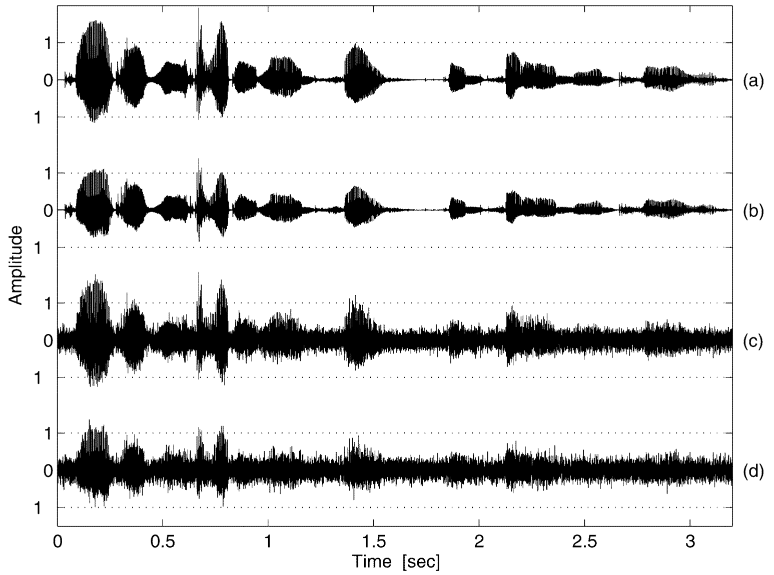

Fig. 1. Speech waveforms. (a) Clean signal $s(t)$ at the primary sensor: "draw every outer line first, then fill in the interior." (b) Reverberated clean signal $a(t)*s(t)$ at the reference sensor. (c) The observed noisy signal at the primary sensor ($\mathrm{SNR} = 4.0\ \mathrm{dB}$); (d) the observed noisy signal at the reference sensor ($\mathrm{SNR} = -0.1\ \mathrm{dB}$).

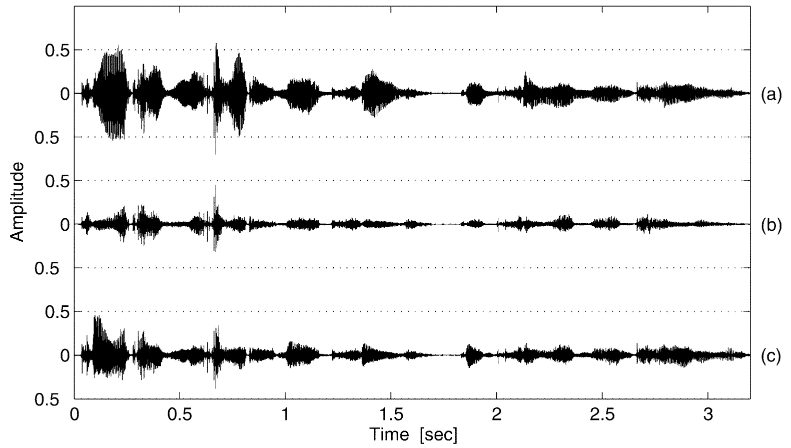

Fig. 2. Signal leakage $r(t)$ in stationary noise environment. (a) Nonstationarity-based system identification ($\mathrm{SBF} = 9.1\ \mathrm{dB}$). (b) Speech-based system identification ($\mathrm{SBF} = 18.5\ \mathrm{dB}$). (c) Online speech-based system identification ($\mathrm{SBF} = 13.9\ \mathrm{dB}$).

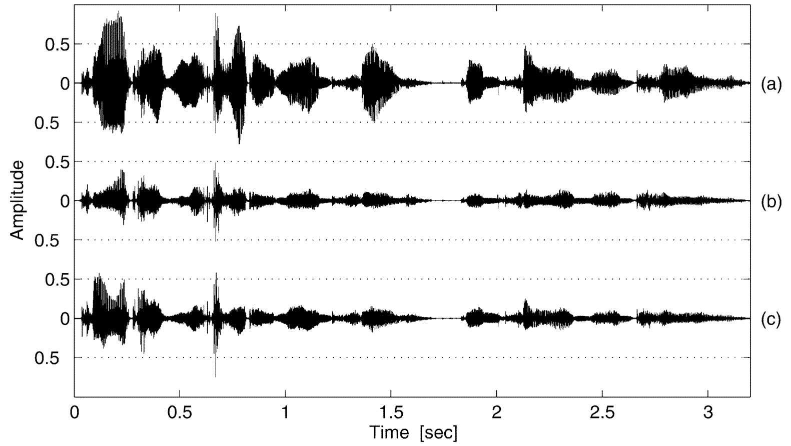

Fig. 3. Signal leakage $r(t)$ in nonstationary noise environment. (a) Nonstationarity-based system identification ($\mathrm{SBF} = 4.9\ \mathrm{dB}$). (b) Speech-based system identification ($\mathrm{SBF} = 13.8\ \mathrm{db}$). (c) Online speech-based system identification ($\mathrm{SBF} = 11.5\ \mathrm{db}$).

level at the same rate. We used again the same speech signal, and the same impulse responses, $a(t)$ and $g(t)$, of the reference sensor to the desired signal and the primary sensor noise ($\mathrm{SNR} = -5.2\ \mathrm{dB}$ at the primary sensor). The leakage signals produced by the above-mentioned algorithms are shown in Fig. 3. As in the stationary noise environment, the proposed speech-based algorithms achieve significantly higher SBFs than the nonstationarity-based algorithm. Furthermore, the performance degradation of the proposed algorithms, when compared to the stationary noise case, is less substantial than that of the

TABLE II
AVERAGE SIGNAL BLOCKING FACTOR (SBF) UNDER VARIOUS CAR NOISE CONDITIONS, OBTAINED BY USING THE NONSTATIONARITY-BASED METHOD AND THE SPEECH-BASED (PROPOSED) METHOD

Input SNR	Pseudo-stationary car noise		Nonstationary car noise	
[dB]	NS	Speech-Based	NS	Speech-Based
-10	3.70	14.02	3.02	13.92
-5	3.91	14.10	3.32	14.23
0	3.95	14.14	3.66	14.33
5	4.01	14.17	3.84	14.37
10	4.02	14.40	3.93	14.39

nonstationarity-based algorithm. This is due to the fact that in the proposed algorithms the noise cross-PSD estimate is continuously updated during speech presence and absence, whereas in the nonstationarity-based algorithm the noise is assumed stationary and the system identification is entirely based on the nonstationarity of the desired signal alone.

In the third experiment, two microphones with 10 cm spacing are mounted in a car on the visor. Clean speech signals are recorded at a sampling rate of 8 kHz in the absence of background noise (standing car, silent environment). Car noise signals are recorded while the car speed is about 60 km/h, and the window next to the driver is either closed or slightly open (about 5 cm; the other windows remain closed). The noise PSD is pseudo-stationary in the former case, while varies substantially in the latter case due to wind blows and passing cars. The input microphone signals are generated by mixing the speech and noise signals at various SNR levels in the range $[-10, 10]$ dB.

Table II shows experimental results of the average SBF obtained under various car noise conditions using the competing system identification algorithms. Clearly, the proposed system identification method is considerably more efficient than the nonstationarity-based method even in the pseudo-stationary noise environment. The rationale is that subintervals with low SNR are more useful for noise estimation, whereas subintervals with high SNR are more useful for system identification. Therefore, by weighting the subintervals for noise estimation differently than the weighting for system identification, improved performance is achieved. Moreover, the proposed algorithm is less sensitive to variations in the noise statistics in case the noise is nonstationary. For a given input SNR, the performance of the proposed algorithm in a *nonstationary* noise environment might be even slightly better than that obtained in a stationary noise environment. This is related to the fact that for a given input SNR and nonstationary noise, there are necessarily subintervals where the instantaneous noise power is lower than its average, and these subintervals are given higher weights in the system identification process. On the contrary, the performance of the nonstationarity-based algorithm, which is based on the nonstationarity of the desired signal alone, essentially is impaired in nonstationary noise environments.

VII. CONCLUSION

We have proposed a robust system identification approach for the relative transfer function between sensors in response to speech signals. The optimization criterion takes into account only time-frequency bins which contain the desired speech components. The auto-PSD of the desired signal is estimated by recursively smoothing the log-spectral amplitude estimate of the signal. The cross-PSD of the interfering signals is estimated by applying a time-varying frequency-dependent recursive smoothing to the cross-PSD of the observed signals, and compensating the bias in accordance with the MCRA method. We showed that the proposed minimum variance WLS estimate for the system's transfer function yields a smaller error variance than that obtained by the nonstationarity method. Generally shorter observation intervals are required for obtaining a reliable system identification, and also the interfering signals are not required to be stationary during absence of the desired signal. In the case of a time-varying system, *e.g.,* moving talkers in hands-free communication scenarios, the proposed method allows to faster and more reliably track the variations. Using the proposed method for the RTF identification, as part of the transfer-function generalized sidelobe canceller (TF-GSC) [2], [4], essentially leads to improved adaptation of the blocking matrix and the noise canceller, and facilitates multichannel signal detection and postfiltering techniques, which employ the transient power ratio between the beamformer output and the reference signals [6], [18], and [19].

APPENDIX I
ASYMPTOTIC COVARIANCE OF ε

From (18) and (19), we have

$$\varepsilon(k,\ell) = \left[\hat{\phi}_{yx}(k,\ell) - \phi_{yx}(k,\ell)\right] - \left[\hat{\phi}_{wu}(k,\ell) - \phi_{wu}(k,\ell)\right] - A(k)\left[\hat{\phi}_{ss}(k,\ell) - \phi_{ss}(k,\ell)\right]. \quad (43)$$

Using the relations

$$V(k,\ell) = Y(k,\ell) - A(k)X(k,\ell) = W(k,\ell) - A(k)U(k,\ell)$$

and

$$\phi_{xx}(k,\ell) = \phi_{ss}(k,\ell) + \phi_{uu}(k,\ell)$$

we obtain

$$\begin{aligned}\varepsilon(k,\ell) &= [\hat{\phi}_{vx}(k,\ell) - \phi_{vx}(k,\ell)] - [\hat{\phi}_{vu}(k,\ell) - \phi_{vu}(k,\ell)] \\ &= \hat{\phi}_{vs}(k,\ell) - \phi_{vs}(k,\ell).\end{aligned} \quad (44)$$

Cross-spectrum estimation by using the cross-periodogram (*e.g.,* [14, sec.5.4]) implies

$$\mathrm{var}\{\hat{\phi}_{vs}(k,\ell)\} = \phi_{vv}(k,\ell)\phi_{ss}(k,\ell). \quad (45)$$

Since we assume that the overlap between successive windows of the short-time Fourier transform is small enough, such that observations in the time-frequency domain associated with different frames can be regarded as statistically independent, we have

$$\begin{aligned}\mathrm{cov}\left(\varepsilon(k,\ell), \varepsilon(k,\ell')\right) &= \mathrm{cov}\left(\hat{\phi}_{vs}(k,\ell), \hat{\phi}_{vs}(k,\ell')\right) \\ &= \begin{cases}\phi_{ss}(k,\ell)\phi_{vv}(k), & \ell = \ell' \\ 0, & \ell \neq \ell'.\end{cases}\end{aligned} \quad (46)$$

APPENDIX II
DERIVATION OF (34)

By (32),

$$\rho = \frac{\langle \phi_{xx} \rangle_\ell \left\langle \phi_{xx}^{-1} \right\rangle_\ell - 1}{\left\langle \phi_{xx}^{-1} \right\rangle_\ell \langle \hat{I} \phi_{ss} \rangle_\ell}. \tag{47}$$

where, for notational simplicity, the arguments k and ℓ are omitted. Denoting by $\xi = \phi_{ss}/\phi_{uu}$ the *a priori* SNR at the primary sensor, and using $\phi_{xx} = \phi_{ss} + \phi_{uu}$ and $\langle \hat{I} \phi_{ss} \rangle \approx \langle I \phi_{ss} \rangle = \langle \phi_{ss} \rangle$, together with the assumption that $u(t)$ is stationary (ϕ_{uu} is independent of the frame index ℓ), we have

$$\begin{aligned} \rho &= \frac{\langle \xi + 1 \rangle_\ell \langle (\xi + 1)^{-1} \rangle_\ell - 1}{\langle (\xi + 1)^{-1} \rangle_\ell \langle \xi \rangle_\ell} \\ &= \frac{\langle (\xi + 1)^{-1} \rangle_\ell - 1 + \langle (\xi + 1)^{-1} \rangle_\ell \langle \xi \rangle_\ell}{(\xi + 1)^{-1} \rangle_\ell \langle \xi \rangle_\ell} \\ &= \frac{\langle (\xi + 1)^{-1} - 1 \rangle_\ell}{\langle (\xi + 1)^{-1} \rangle_\ell \langle \xi \rangle_\ell} + 1 \\ &= 1 - \frac{\langle \xi (\xi + 1)^{-1} \rangle_\ell}{\langle (\xi + 1)^{-1} \rangle_\ell \langle \xi \rangle_\ell} < 1. \end{aligned} \tag{48}$$

ACKNOWLEDGMENT

The author wishes to thank the anonymous reviewers for their constructive comments and helpful suggestions.

REFERENCES

[1] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, pp. 2055–2063, Aug. 1996.

[2] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing*, vol. 49, pp. 1614–1626, Aug. 2001.

[3] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, pp. 2677–2684, Oct. 1999.

[4] S. Gannot, D. Burshtein, and E. Weinstein, "Theoretical Performance Analysis of the General Transfer Function GSC," Technion—Israel Inst. Technol., Haifa, CCIT Tech. Rep. 381, 2002.

[5] I. Cohen, "Multi-Channel Post-Filtering in Non-Stationary Noise Environments," *IEEE Trans. Signal Process.*, vol. 52, pp. 1149–1160, May 2002.

[6] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for nonstationary noise environments," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1064–1073, Oct. 2003.

[7] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput., Speech, Lang.*, vol. 11, no. 2, pp. 91–126, Apr. 1997.

[8] Y. Huang, J. Benesty, and G. W. Elko, *Microphone Arrays for Video Camera Steering*. Norwell, MA: Kluwer, 2000, pp. 239–259.

[9] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Robust Localization in Reverberant Rooms*. New York: Springer-Verlag, ch. 8, pp. 157–179.

[10] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[11] ——, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Lett.*, vol. 9, pp. 12–15, Jan. 2002.

[12] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 466–475, Sept. 2003.

[13] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall, 1974.

[14] D. G. Manolakis, V. K. Ingle, and S. M. Kogan, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. New York: McGraw-Hill, 2000.

[15] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 504–512, July 2001.

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

[17] B. Widrow and S. D. Stearns, Eds., *Adaptive Signal Processing*. New York: Prentice-Hall, 1985.

[18] I. Cohen and B. Berdugo, "Microphone array post-filtering for nonstationary noise suppression," in *Proc. 27th IEEE Int. Conf. Acoustics Speech Signal Processing*, Orlando, FL, May 13–17, 2002, pp. 901–904.

[19] ——, "Two-channel signal detection and speech enhancement based on the transient beam-to-reference ratio," in *Proc. 28th IEEE Int. Conf. Acoustics Speech Signal Processing*, Hong Kong, Apr. 6–10, 2003, pp. V 233–236.

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

From 1990 to 1998, he was a Research Scientist at RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate in the Computer Science Department, Yale University, New Haven, CT. Since 2001, he has been a Senior Lecturer with the Electrical Engineering Department, Technion. His research interests are 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 is Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.