CCIT Report # 426 June 2003

Robust System Identification Using Speech Signals

Israel Cohen

Abstract

An important component of a multichannel hands-free communication system is the identification of the coupling 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 includes the probability that the desired signal is present in the observed signals. An asymptotically unbiased estimate for the system's transfer function is derived, and a corresponding recursive on-line implementation is presented. We show that compared to a competing nonstationarity-based method, a significantly smaller error variance is achieved and generally shorter observation intervals are required. Furthermore, in case of a time-varying system, faster convergence and higher reliability of the system identification are obtained. 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

Array signal processing, system identification, signal detection, acoustic noise measurement, speech enhancement, spectral analysis, adaptive signal processing.

I. Introduction

An important component of a multichannel hands-free communication system is the identification of the coupling between sensors in response to a desired source signal [1], [2], [3]. This coupling, often referred to as the acoustical transfer function (ATF) ratio, represents the relation between the impulse responses of the sensors to the desired source. In reverberant and noisy environments, the coupling identification enables to construct an adaptive blocking channel, for an accurate derivation of a reference noise signal, and an adaptive noise canceller, for eliminating directional or coherent noise sources [4]. Furthermore, it also facilitates multichannel signal detection and postfiltering techniques, which employ the transient power ratio between the beamformer output and the reference signals [5], [6].

The author is with the Department of Electrical Engineering, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel (email: icohen@ee.technion.ac.il; tel.: +972-4-8294731; fax: +972-4-8323041).

Shalvi and Weinstein [1] have proposed to identify the coupling 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 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 system's transfer function and the (presumably stationary) cross-PSD of the primary sensor and a noise component. An asymptotically unbiased estimate for the system's transfer function 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 system 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 in higher SNR subintervals. The other requirement is minimizing the error variance of the noise estimate, which rather implies smaller weights in 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 quite 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 implies a very long observation interval. Furthermore, the system's transfer function 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], [8], [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 system identification in 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 system identification and noise spectra estimation. An estimate for the system's transfer function 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], [12]. Subsequently, minimum variance WLS estimate for the system's transfer function is derived, and a recursive on-line solution is obtained based on the least-mean-square (LMS) algorithm. We show that the error variance obtained by using the proposed method is significantly 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

system identification. Whereas the nonstationarity method requires a relatively long observation interval, regardless of the SNR, to retain the desired signal sufficiently nonstationary. Additionally, 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 case of a time-varying system, faster convergence and higher reliability of the system 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 the speech presence probability in the time-frequency domain. In Section V, we describe the system identification algorithm and its on-line implementation. Finally, in Section VI, we present experimental results, which demonstrate the improvement gained by the proposed approach.

II. Problem Formulation

Let $s(t)$ represent a desired source signal, denote by $u(t)$ and $w(t)$ additive interfering signals that are uncorrelated with the desired signal, and let $a(t)$ represent the coupling of the desired signal to the reference sensor. 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 $*$ denotes convolution. Our objective is to identify $a(t)$ in the general case where $u(t)$ is statistically correlated with $w(t)$.

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 statistically uncorrelated noise $u(t)$ as in (1). Furthermore, the desired signal is presumably uncorrelated with $v(t)$. It is easy to verify that the two above-mentioned problems are equal, with

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

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, and that the desired signal $s(t)$ is nonstationary.

Dividing the observation interval into M subintervals, and computing for each subinterval m ($m = 1, 2, \ldots, M$) the cross-PSD between y and x, we obtain from (3)

$$\phi_{y\,x}^{(m)}(\omega) = A(\omega)\,\phi_{x\,x}^{(m)}(\omega) + \phi_{v\,x}(\omega)\,, \tag{5}$$

where $A(\omega)$ is the Fourier transform of $a(t)$, and $\phi_{v\,x}(\omega)$ 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 $\hat{\phi}_{y\,x}^{(m)}(\omega)$, $\hat{\phi}_{x\,x}^{(m)}(\omega)$ and $\hat{\phi}_{v\,x}^{(m)}(\omega)$ be estimates for $\phi_{y\,x}^{(m)}(\omega)$, $\phi_{x\,x}^{(m)}(\omega)$ and $\phi_{v\,x}(\omega)$, respectively. Then

$$\begin{aligned} \hat{\phi}_{y\,x}^{(m)}(\omega) &= A(\omega)\,\hat{\phi}_{x\,x}^{(m)}(\omega) + \hat{\phi}_{v\,x}^{(m)}(\omega) \\ &= A(\omega)\,\hat{\phi}_{x\,x}^{(m)}(\omega) + \phi_{v\,x}(\omega) + \epsilon^{(m)}(w)\,, \end{aligned} \tag{6}$$

where

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

This can be written in a matrix form as

$$\begin{aligned} z &\triangleq \begin{bmatrix} \hat{\phi}_{y\,x}^{(1)}(\omega) \\ \hat{\phi}_{y\,x}^{(2)}(\omega) \\ \vdots \\ \hat{\phi}_{y\,x}^{(M)}(\omega) \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{x\,x}^{(1)}(\omega) & 1 \\ \hat{\phi}_{x\,x}^{(2)}(\omega) & 1 \\ \vdots & \vdots \\ \hat{\phi}_{x\,x}^{(M)}(\omega) & 1 \end{bmatrix} \begin{bmatrix} A(\omega) \\ \phi_{v\,x}(\omega) \end{bmatrix} + \begin{bmatrix} \epsilon^{(1)}(w) \\ \epsilon^{(2)}(w) \\ \vdots \\ \epsilon^{(M)}(w) \end{bmatrix} \\ &\triangleq \hat{G}\,\boldsymbol{\theta} + \boldsymbol{\epsilon}\,. \end{aligned} \tag{8}$$

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

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

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

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

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

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

$$\hat{A}(\omega) = \frac{\left\langle \hat{\phi}_{y\,x}(\omega)\,\hat{\phi}_{x\,x}(\omega) \right\rangle - \left\langle \hat{\phi}_{y\,x}(\omega) \right\rangle \left\langle \hat{\phi}_{x\,x}(\omega) \right\rangle}{\left\langle \hat{\phi}^2_{x\,x}(\omega) \right\rangle - \left\langle \hat{\phi}_{x\,x}(\omega) \right\rangle^2} \tag{11}$$

with the average operation defined by

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

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

$$W_{m\,n} = \begin{cases} T_m \,/\, \hat{\phi}^{(m)}_{x\,x}(\omega)\,, & m = n \\ 0\,, & m \neq n \end{cases} \tag{13}$$

In which case, (9) yields

$$\hat{A}(\omega) = \frac{\left\langle 1 \,/\, \hat{\phi}_{x\,x}(\omega) \right\rangle \left\langle \hat{\phi}_{y\,x}(\omega) \right\rangle - \left\langle \hat{\phi}_{y\,x}(\omega) \,/\, \hat{\phi}_{x\,x}(\omega) \right\rangle}{\left\langle \hat{\phi}_{x\,x}(\omega) \right\rangle \left\langle 1 \,/\, \hat{\phi}_{x\,x}(\omega) \right\rangle - 1}\,, \tag{14}$$

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

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

where $T \stackrel{\triangle}{=} \sum_{m=1}^{M} T_m$, and $B = \frac{1}{\sum_\tau w^2(\tau)}$ is related to the window's bandwidth that is pre-selected 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 $\phi_{v\,x}(\omega)$ are carried out using the same weight matrix W. That is, each subinterval m is given the same weight, whether we are trying to find an estimate for $A(\omega)$ or for $\phi_{v\,x}(\omega)$. However, subintervals with higher SNRs are of greater importance when estimating $A(\omega)$, whereas the opposite is true when estimating $\phi_{v\,x}(\omega)$. 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 in higher SNR subintervals. The other requirement is minimizing the error variance of $\hat{\phi}_{v\,x}(\omega)$, which rather implies smaller weights in 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, including in the observation 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 $\hat{\phi}_{v\,x}(\omega)$ 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 system identification using nonstationarity is that the interfering signals are required to be stationary during the entire observation interval, and the observation interval should include quite 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 implies a very long observation interval, thus constraining the interfering signals to be stationary over long intervals. Furthermore, the system's transfer function $A(\omega)$ 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 $\phi_{v\,x}(\omega)$.

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 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 transfer function 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

$$\phi_{y\,x}(k,\ell) = A(k)\,\phi_{x\,x}(k,\ell) + \phi_{v\,x}(k,\ell)\,. \tag{17}$$

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

$$\phi_{y\,x}(k,\ell) = A(k)\,\phi_{s\,s}(k,\ell) + \phi_{w\,u}(k,\ell)\,. \tag{18}$$

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

$$\hat{\phi}_{y\,x}(k,\ell) - \hat{\phi}_{w\,u}(k,\ell) = A(k)\,\hat{\phi}_{s\,s}(k,\ell) + \varepsilon(k,\ell) \tag{19}$$

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

$$\begin{aligned}\hat{\boldsymbol{\psi}}(k) &\triangleq \begin{bmatrix} \hat{\phi}_{y\,x}(k,1) - \hat{\phi}_{w\,u}(k,1) \\ \hat{\phi}_{y\,x}(k,2) - \hat{\phi}_{w\,u}(k,2) \\ \vdots \\ \hat{\phi}_{y\,x}(k,L) - \hat{\phi}_{w\,u}(k,L) \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{s\,s}(k,1) \\ \hat{\phi}_{s\,s}(k,2) \\ \vdots \\ \hat{\phi}_{s\,s}(k,L) \end{bmatrix} A(k) + \begin{bmatrix} \varepsilon(k,1) \\ \varepsilon(k,2) \\ \vdots \\ \varepsilon(k,L) \end{bmatrix} \\ &\triangleq \hat{\boldsymbol{\phi}}_{s\,s}(k)\,A(k) + \boldsymbol{\varepsilon}\,. \end{aligned} \tag{20}$$

Since the transfer function $A(k)$ represents the coupling between the primary and reference sensor with regards to the desired source signal, the optimization criterion for the identification of $A(k)$ has to take into consideration the probability that the desired signal is present in the observed signals ($S(k,\ell) \neq 0$). Specifically, let $p(k,\ell) = \mathcal{P}\{S(k,\ell) \neq 0 \,|\, x(t), y(t)\}$ denote the conditional signal presence probability given the observed signals, and let P represent a diagonal matrix with the elements $[p(k,1), p(k,2), \ldots, p(k,L)]$ on its diagonal. Then the WLS estimate of $A(k)$ is obtained by

$$\begin{aligned}
\hat{A}(k) &= \arg\min_{A(k)} [P\,\boldsymbol{\varepsilon}]^H W\,[P\,\boldsymbol{\varepsilon}] \\
&= \arg\min_{A(k)} \left[\hat{\boldsymbol{\psi}}(k) - \hat{\boldsymbol{\phi}}_{s\,s}(k)\,A(k)\right]^H P\,W\,P\left[\hat{\boldsymbol{\psi}}(k) - \hat{\boldsymbol{\phi}}_{s\,s}(k)\,A(k)\right] \\
&= \left[\hat{\boldsymbol{\phi}}_{s\,s}^T(k) P\,W\,P\,\hat{\boldsymbol{\phi}}_{s\,s}(k)\right]^{-1} \hat{\boldsymbol{\phi}}_{s\,s}^T(k) P\,W\,P\,\hat{\boldsymbol{\psi}}(k)\,. \qquad (21)
\end{aligned}$$

The weight matrix W that minimizes the variance of $\hat{A}(k)$ is given by

$$W = [\mathrm{cov}(P\,\boldsymbol{\varepsilon})]^{-1} = P^{-}\,[\mathrm{cov}(\boldsymbol{\varepsilon})]^{-1}\,P^{-} \qquad (22)$$

where P^{-} is a generalized inverse of P, *i.e.*,

$$P^{-}_{\ell,\ell'} = \begin{cases} [p(k,\ell)]^{-1}\,, & \text{if } \ell = \ell' \text{ and } p(k,\ell) \neq 0 \\ 0\,, & \text{otherwise.} \end{cases}$$

This choice of W yields an asymptotically unbiased estimate for $A(k)$

$$\hat{A}(k) = \left(\hat{\boldsymbol{\phi}}_{s\,s}^T(k)\,[\mathrm{cov}(\boldsymbol{\varepsilon})]^{-1}\,\hat{\boldsymbol{\phi}}_{s\,s}(k)\right)^{-1} \hat{\boldsymbol{\phi}}_{s\,s}^T(k)\,[\mathrm{cov}(\boldsymbol{\varepsilon})]^{-1}\,\hat{\boldsymbol{\psi}}(k) \qquad (23)$$

whose variance is given by (see Appendix I)

$$\mathrm{var}\left\{\hat{A}(k)\right\} = \left(\boldsymbol{\phi}_{s\,s}^T(k)\,[\mathrm{cov}(\boldsymbol{\varepsilon})]^{-1}\,\boldsymbol{\phi}_{s\,s}(k)\right)^{-1}\,. \qquad (24)$$

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

$$\mathrm{cov}\left(\varepsilon(k,\ell), \varepsilon(k,\ell')\right) = \begin{cases} \frac{1-\alpha_s}{1+\alpha_s}\phi_{x\,x}(k,\ell)\,\phi_{v\,v}(k)\,, & \ell = \ell' \\ 0\,, & \ell \neq \ell' \end{cases} \qquad (25)$$

where α_s ($0 \leq \alpha_s < 1$) is a smoothing parameter used for the empirical cross-spectrum estimation by the Welch's method. Substituting (25) into (23) and (24) we obtain

$$\hat{A}(k) = \frac{\left\langle \hat{\phi}_{x\,x}^{-1}(k,\ell)\,\hat{\phi}_{s\,s}(k,\ell)\left[\hat{\phi}_{y\,x}(k,\ell) - \hat{\phi}_{w\,u}(k,\ell)\right]\right\rangle_\ell}{\left\langle \hat{\phi}_{x\,x}^{-1}(k,\ell)\,\hat{\phi}_{s\,s}^2(k,\ell)\right\rangle_\ell} \qquad (26)$$

$$\mathrm{var}\left\{\hat{A}(k)\right\} = \frac{1-\alpha_s}{(1+\alpha_s)L} \cdot \frac{\phi_{v\,v}(k)}{\left\langle \phi_{x\,x}^{-1}(k,\ell)\,\phi_{s\,s}^2(k,\ell)\right\rangle_\ell} \qquad (27)$$

where the average operation is defined by

$$\langle \varphi(k,\ell)\rangle_\ell \triangleq \frac{1}{L}\sum_{\ell=1}^{L} \varphi(k,\ell)\,. \qquad (28)$$

Note that the estimate $\hat{A}(k)$, as well as its variance, are independent of the speech presence probability, even though the error minimization in the first line of (21) is subject to the conditional probability that the desired signal is present in the observed signals. Furthermore, only frames that contain speech ($\hat{\phi}_{s\,s}(k,\ell) \neq 0$) influence the values of $\hat{A}(k)$ and $\mathrm{var}\left\{\hat{A}(k)\right\}$. Including in the observation interval additional segments that do not contain speech does not increase the variance of $\hat{A}(k)$.

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 $B\,T_0 = 1$ where T_0 is the frame's length. Accordingly, the variance of $\hat{A}(k)$ obtained using the nonstationarity method is

$$\mathrm{var}\left\{\hat{A}(k)\right\}\Big|_{\text{NS method}} = \frac{1}{L}\cdot\frac{\phi_{v\,v}(k)\left\langle\phi_{x\,x}^{-1}(k,\ell)\right\rangle_{\ell}}{\langle\phi_{x\,x}(k,\ell)\rangle_{\ell}\left\langle\phi_{x\,x}^{-1}(k,\ell)\right\rangle_{\ell}-1} \tag{29}$$

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

$$\rho \triangleq \frac{\mathrm{var}\left\{\hat{A}(k)\right\}\Big|_{\text{proposed method}}}{\mathrm{var}\left\{\hat{A}(k)\right\}\Big|_{\text{NS method}}} = \frac{1-\alpha_s}{1+\alpha_s}\cdot\frac{\langle\phi_{x\,x}(k,\ell)\rangle_{\ell}\left\langle\phi_{x\,x}^{-1}(k,\ell)\right\rangle_{\ell}-1}{\left\langle\phi_{x\,x}^{-1}(k,\ell)\right\rangle_{\ell}\left\langle\phi_{x\,x}^{-1}(k,\ell)\,\phi_{s\,s}^{2}(k,\ell)\right\rangle_{\ell}}\,. \tag{30}$$

Let $\xi(k,\ell) \triangleq \phi_{s\,s}(k,\ell)/\,\phi_{u\,u}(k)$ denote the *a priori* SNR at the primary sensor. Then substituting $\phi_{x\,x}(k,\ell) = \phi_{s\,s}(k,\ell) + \phi_{u\,u}(k)$ into (30) we obtain (see Appendix III)

$$\rho = \frac{1-\alpha_s}{1+\alpha_s}\cdot\frac{\langle\xi(k,\ell)+1\rangle_{\ell}\left\langle[\xi(k,\ell)+1]^{-1}\right\rangle_{\ell}-1}{\left\langle[\xi(k,\ell)+1]^{-1}\right\rangle_{\ell}\left\langle\xi^{2}(k,\ell)\,[\xi(k,\ell)+1]^{-1}\right\rangle_{\ell}} < \frac{1-\alpha_s}{1+\alpha_s} \tag{31}$$

Thus, the variance of $\hat{A}(k)$ obtained by using the proposed method is significantly smaller than that obtained by using the nonstationarity method. Additionally, 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, $\phi_{s\,s}(k,\ell)$. The higher the SNR is, the fewer 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 $\phi_{x\,x}(k,\ell)$. 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 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 $\phi_{x\,x}(k,\ell)$, $\phi_{y\,x}(k,\ell)$, $\phi_{s\,s}(k,\ell)$ and $\phi_{w\,u}(k,\ell)$. The first two estimates are obtained by applying a first-order recursive smoothing to the periodograms $|X(k,\ell)|^2$ and $Y(k,\ell)\,X^*(k,\ell)$ of the observed signals. Specifically,

$$\hat{\phi}_{x\,x}(k,\ell) = \alpha_s\,\hat{\phi}_{x\,x}(k,\ell-1) + (1-\alpha_s)\,|X(k,\ell)|^2 \tag{32}$$

$$\hat{\phi}_{y\,x}(k,\ell) = \alpha_s\,\hat{\phi}_{y\,x}(k,\ell-1) + (1-\alpha_s)Y(k,\ell)\,X^*(k,\ell) \tag{33}$$

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 [13]. 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

$$\hat{\phi}_{s\,s}(k,\ell) = \alpha_s\,\hat{\phi}_{s\,s}(k,\ell-1) + (1-\alpha_s)G^2(k,\ell)\,|X(k,\ell)|^2 \tag{34}$$

where $G(k,\ell)$ denotes the OM-LSA gain function.

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

$$\hat{\phi}_{w\,u}(k,\ell) = \tilde{\alpha}_u(k,\ell)\,\hat{\phi}_{w\,u}(k,\ell-1) + \beta\,[1-\tilde{\alpha}_u(k,\ell)]\,Y(k,\ell)\,X^*(k,\ell) \tag{35}$$

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. 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:

$$\tilde{\alpha}_u(k,\ell) = \alpha_u + (1-\alpha_u)p(k,\ell)\,. \tag{36}$$

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

Substituting the above spectral estimates into (26) we obtain an estimate for $A(k)$. Alternatively, a recursive on-line solution to (21) based on the LMS algorithm [14] is given by

$$\begin{aligned}\hat{A}(k,\ell) &= \hat{A}(k,\ell-1) - \mu\ \frac{\partial}{\partial A^*}\left[p^2(k,\ell)W_{\ell\ell}\left|\hat{\phi}_{y\,x}(k,\ell) - \hat{\phi}_{w\,u}(k,\ell) - A\,\hat{\phi}_{s\,s}(k,\ell)\right|^2\right]\Big|_{A=\hat{A}(k,\ell-1)} \\ &= \hat{A}(k,\ell-1) + \mu(k,\ell)\,\hat{\phi}_{s\,s}(k,\ell)\,\hat{\varepsilon}(k,\ell) \end{aligned} \tag{37}$$

where

$$\mu(k,\ell) = \frac{\tilde{\mu}}{\hat{\phi}_{x\,x}(k,\ell)\,\hat{\phi}_{v\,v}(k,\ell)} \tag{38}$$

is a time-varying frequency-dependent step-size parameter,

$$\hat{\varepsilon}(k,\ell) = \hat{\phi}_{y\,x}(k,\ell) - \hat{\phi}_{w\,u}(k,\ell) - \hat{A}(k,\ell-1)\hat{\phi}_{s\,s}(k,\ell) \tag{39}$$

is the estimation error, and by using the relation $V(k,\ell) = Y(k,\ell) - A(k)X(k,\ell)$ we obtain

$$\hat{\phi}_{v\,v}(k,\ell) = \hat{\phi}_{y\,y}(k,\ell) + \left|\hat{A}(k,\ell-1)\right|^2\hat{\phi}_{x\,x}(k,\ell) - 2\,\Re\left\{\hat{A}(k,\ell-1)\hat{\phi}^*_{y\,x}(k,\ell)\right\}. \tag{40}$$

The update of $\hat{A}(k,\ell)$ in (37) is carried out whenever the time-frequency bin (k,ℓ) contains some desired signal energy (*e.g.*, in the event that $10\,\log_{10}[\hat{\phi}_{s\,s}(k,\ell)/\hat{\phi}_{x\,x}(k,\ell)] > -10$ dB). The implementation of the proposed on-line system identification algorithm is summarized in Fig. 1.

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

$$\text{SBF} = 10\log_{10}\frac{E\left\{s^2(t)\right\}}{E\left\{r^2(t)\right\}} \quad [\text{dB}] \tag{41}$$

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

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

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 affect 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 a factor of

2.5 larger than the average power of the speech (SNR$= 4$ dB). The impulse response of the reference sensor to the desired signal is

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

where $T = 12.5$ ms is the sampling period. In addition, the reference sensor noise $w(t)$ is generated by

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

where

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

Figure 2 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. 3(a). The resultant SBF is 9.1 dB.

Figures 3(b) and (c) show the leakage signals obtained by using the proposed algorithms. Off-line speech-based system identification (see (26)) yields a SBF of 18.5 dB, whereas the on-line speech-based system identification (see (37)) 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 two seconds, and then decreasing it back to the original 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 (SNR$= -5.2$ dB at the primary sensor). The leakage signals produced by the above-mentioned algorithms are shown in Fig. 4. As in the stationary noise environment, the proposed speech-based algorithms achieve significantly higher SBF's 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 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 completely 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.

Figure 5 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 coupling between sensors in response to speech signals. The optimization criterion takes into account the probability that the desired speech is present in the received signals. Nevertheless, the estimate for the system's transfer function and its variance are independent of the speech presence probability, but require the auto-PSD of the desired signal and the cross-PSD of the interfering signals. 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 significantly 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 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 identification of the acoustical transfer function ratios, as part of the *transfer-function generalized sidelobe canceller* (TF-GSC) [2], [4], 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 [15], [6], [16].

APPENDIX I
ASYMPTOTIC VARIANCE OF $\hat{A}(k)$

Substituting (20) into (21), we obtain

$$\begin{aligned} \hat{A}(k) - A(k) &= \left[\hat{\boldsymbol{\phi}}_{s\,s}^{T}(k) P\, W\, P\, \hat{\boldsymbol{\phi}}_{s\,s}(k)\right]^{-1} \hat{\boldsymbol{\phi}}_{s\,s}^{T}(k) P\, W\, P\, \boldsymbol{\varepsilon} \\ &\approx \left[\boldsymbol{\phi}_{s\,s}^{T}(k) P\, W\, P\, \boldsymbol{\phi}_{s\,s}(k)\right]^{-1} \boldsymbol{\phi}_{s\,s}^{T}(k) P\, W\, P\, \boldsymbol{\varepsilon} \end{aligned} \tag{43}$$

where we have assumed that to a first order approximation $\hat{\boldsymbol{\phi}}_{s\,s}(k)$ is sufficiently close to $\boldsymbol{\phi}_{s\,s}(k)$. From (32)-(35), $\hat{\phi}_{y\,x}(k,\ell)$, $\hat{\phi}_{w\,u}(k,\ell)$ and $\hat{\phi}_{s\,s}(k,\ell)$ are unbiased estimates for $\phi_{y\,x}(k,\ell)$, $\phi_{w\,u}(k,\ell)$ and $\phi_{s\,s}(k,\ell)$, respectively. Hence,

$$\begin{aligned} E\left\{\varepsilon(k,\ell)\right\} &= E\left\{\hat{\phi}_{y\,x}(k,\ell) - \hat{\phi}_{w\,u}(k,\ell) - A(k)\,\hat{\phi}_{s\,s}(k,\ell)\right\} \\ &= \phi_{y\,x}(k,\ell) - \phi_{w\,u}(k,\ell) - A(k)\,\phi_{s\,s}(k,\ell) = 0\,. \end{aligned}$$

Accordingly $\boldsymbol{\varepsilon}$ is zero mean, which implies that $\hat{A}(k)$ is asymptotically an unbiased estimate for $A(k)$.

The choice of W that minimizes the variance of $\hat{A}(k)$ is given by

$$W = \left[\mathrm{cov}(P\,\boldsymbol{\varepsilon})\right]^{-1} = P^{-}\left[\mathrm{cov}(\boldsymbol{\varepsilon})\right]^{-1} P^{-} \tag{44}$$

where P^{-} is a generalized inverse of P, *i.e.*,

$$P^{-}_{\ell,\ell'} = \begin{cases} \left[p(k,\ell)\right]^{-1}\,, & \text{if } \ell = \ell' \text{ and } p(k,\ell) \neq 0 \\ 0\,, & \text{otherwise.} \end{cases}$$

The variance of $\hat{A}(k)$ is given by

$$\mathrm{var}\left\{\hat{A}(k)\right\} = \left[\boldsymbol{\phi}_{s\,s}^{T}(k) P\, W\, P\, \boldsymbol{\phi}_{s\,s}(k)\right]^{-1} \boldsymbol{\phi}_{s\,s}^{T}(k) P\, W\, P\, \mathrm{cov}(\boldsymbol{\varepsilon})\, P\, W^{H}\, P \boldsymbol{\phi}_{s\,s}(k) \left[\boldsymbol{\phi}_{s\,s}^{T}(k) P\, W^{H}\, P\, \boldsymbol{\phi}_{s\,s}(k)\right]^{-1} \tag{45}$$

Substituting (44) into (45), we obtain

$$\mathrm{var}\left\{\hat{A}(k)\right\} = \left(\boldsymbol{\phi}_{s\,s}^{T}(k)\left[\mathrm{cov}(\boldsymbol{\varepsilon})\right]^{-1}\boldsymbol{\phi}_{s\,s}(k)\right)^{-1} \tag{46}$$

where we used the identity $P^{-} P\, \boldsymbol{\phi}_{s\,s}(k) = \boldsymbol{\phi}_{s\,s}(k)$, since by definition $\phi_{s\,s}(k,\ell)$ reduces to zero whenever the speech presence probability $p(k,\ell)$ is zero.

APPENDIX II
ASYMPTOTIC COVARIANCE OF $\boldsymbol{\varepsilon}$

From (18) and (19), we have

$$\varepsilon(k,\ell) = \left[\hat{\phi}_{y\,x}(k,\ell) - \phi_{y\,x}(k,\ell)\right] - \left[\hat{\phi}_{w\,u}(k,\ell) - \phi_{w\,u}(k,\ell)\right] - A(k)\left[\hat{\phi}_{s\,s}(k,\ell) - \phi_{s\,s}(k,\ell)\right] \tag{47}$$

Using the relations

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

and

$$\phi_{x\,x}(k,\ell) = \phi_{s\,s}(k,\ell) + \phi_{u\,u}(k,\ell)$$

we obtain

$$\varepsilon(k,\ell) = \left[\hat{\phi}_{v\,x}(k,\ell) - \phi_{v\,x}(k,\ell)\right] - \left[\hat{\phi}_{v\,u}(k,\ell) - \phi_{v\,u}(k,\ell)\right] \tag{48}$$

Since the estimate for $\phi_{v\,u}(k,\ell)$ is derived based on frames that do not contain speech ($\hat{\phi}_{v\,u}(k,\ell)$ is not updated during speech presence, *i.e.*, when $p(k,\ell) \neq 0$), we have

$$\begin{aligned}
\mathrm{cov}\left\{p(k,\ell)\varepsilon(k,\ell), p(k,\ell')\varepsilon(k,\ell')\right\} &= p(k,\ell)p(k,\ell')\mathrm{cov}\left\{\varepsilon(k,\ell),\varepsilon(k,\ell')\right\} \\
&= p(k,\ell)p(k,\ell')\mathrm{cov}\left\{\hat{\phi}_{v\,x}(k,\ell),\hat{\phi}_{v\,x}(k,\ell')\right\} \\
&= \mathrm{cov}\left\{p(k,\ell)\hat{\phi}_{v\,x}(k,\ell), p(k,\ell')\hat{\phi}_{v\,x}(k,\ell')\right\}
\end{aligned} \tag{49}$$

Then, for the purpose of WLS optimization (*i.e.*, minimization of $[P\,\boldsymbol{\varepsilon}]^H\,W\,[P\,\boldsymbol{\varepsilon}]$), the elements of $\mathrm{cov}(\boldsymbol{\varepsilon})$ can be substituted with $\mathrm{cov}\left\{\hat{\phi}_{v\,x}(k,\ell),\hat{\phi}_{v\,x}(k,\ell')\right\}$.

Cross-spectrum estimation by using Welch's method [17] implies

$$\mathrm{var}\left\{\hat{\phi}_{v\,x}(k,\ell)\right\} \approx \frac{1}{N_\ell}\phi_{x\,x}(k,\ell)\,\phi_{v\,v}(k,\ell) \tag{50}$$

where N_ℓ is the number of cross-periodograms that are averaged. Applying a first-order smoothing with a smoothing parameter α_s ($0 \leq \alpha_s < 1$) for the empirical cross-spectrum estimation

$$\hat{\phi}_{v\,x}(k,\ell) = \alpha_s\,\hat{\phi}_{v\,x}(k,\ell-1) + (1-\alpha_s)V(k,\ell)X^*(k,\ell)\,,$$

and assuming that observations in the time-frequency domain associated with different frames are statistically independent, we have

$$\mathrm{cov}\left(\varepsilon(k,\ell),\varepsilon(k,\ell')\right) = \begin{cases} \frac{1-\alpha_s}{1+\alpha_s}\phi_{x\,x}(k,\ell)\,\phi_{v\,v}(k)\,, & \ell = \ell' \\ 0\,, & \ell \neq \ell' \end{cases} \tag{51}$$

where we have used $N_\ell \approx \frac{1+\alpha_s}{1-\alpha_s}$.

APPENDIX III
DERIVATION OF (31)

By (30),

$$\frac{1+\alpha_s}{1-\alpha_s}\cdot\rho = \frac{\langle\phi_{x\,x}\rangle_\ell\left\langle\phi_{x\,x}^{-1}\right\rangle_\ell - 1}{\left\langle\phi_{x\,x}^{-1}\right\rangle_\ell\left\langle\phi_{x\,x}^{-1}\,\phi_{s\,s}^2\right\rangle_\ell}\,. \tag{52}$$

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

$$\begin{aligned}
\frac{1+\alpha_s}{1-\alpha_s}\cdot\rho &= \frac{\langle\xi+1\rangle_\ell \left\langle (\xi+1)^{-1}\right\rangle_\ell - 1}{\left\langle (\xi+1)^{-1}\right\rangle_\ell \left\langle \xi^2\,(\xi+1)^{-1}\right\rangle_\ell} \\
&= \frac{\left\langle (\xi+1)^{-1}\right\rangle_\ell \left\langle (2\,\xi+1)\,(\xi+1)^{-1}\right\rangle_\ell - 1 + \left\langle (\xi+1)^{-1}\right\rangle_\ell \left\langle \xi^2\,(\xi+1)^{-1}\right\rangle_\ell}{\left\langle (\xi+1)^{-1}\right\rangle_\ell \left\langle \xi^2\,(\xi+1)^{-1}\right\rangle_\ell} \\
&= \frac{\left\langle (\xi+1)^{-1}\right\rangle_\ell \left\langle \xi\,(\xi+1)^{-1}\right\rangle_\ell - \left\langle \xi\,(\xi+1)^{-1}\right\rangle_\ell}{\left\langle (\xi+1)^{-1}\right\rangle_\ell \left\langle \xi^2\,(\xi+1)^{-1}\right\rangle_\ell} + 1 \\
&= 1 - \frac{\left\langle \xi\,(\xi+1)^{-1}\right\rangle_\ell^2}{\left\langle (\xi+1)^{-1}\right\rangle_\ell \left\langle \xi^2\,(\xi+1)^{-1}\right\rangle_\ell} < 1\,. \qquad (53)
\end{aligned}$$

REFERENCES

[1] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2055–2063, 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, no. 8, pp. 1614–1626, August 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, no. 10, pp. 2677–2684, October 1999.

[4] S. Gannot, D. Burshtein, and E. Weinstein, "Theoretical performance analysis of the general transfer function GSC," Technion - Israel Institute of Technology, Haifa, Israel, CCIT Technical Report 381, May 2002.

[5] I. Cohen, "Multi-channel post-filtering in non-stationary noise environments," Technion - Israel Institute of Technology, Haifa, Israel, Technical Report, EE PUB 1314, April 2002.

[6] I. Cohen, S. Gannot, and B. Berdugo, "An integrated real-time beamforming and postfiltering system for non-stationary noise environments," *to appear in special issue of EURASIP JASP on Signal Processing for Acoustic Communication System*, 2003.

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

[8] Y. Huang, J. Benesty, and G. W. Elko, *Microphone Arrays for Video Camera Steering*, ch. 11, pp. 239–259.

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

[10] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, October 2001.

[11] ——, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, January 2002.

[12] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Signal Processing*, May 2003.

[13] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[14] B. Widrow and S. D. Stearns, Eds., *Adaptive Signal Processing*. Prentice-Hall, 1985.

[15] I. Cohen and B. Berdugo, “Microphone array post-filtering for non-stationary noise suppression,” in *Proc. 27th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2002*, Orlando, Florida, 13–17 May 2002, pp. 901–904.

[16] ——, “Two-channel signal detection and speech enhancement based on the transient beam-to-reference ratio,” in *Proc. 28th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2003*, Hong Kong, 6–10 April 2003, pp. V 233–236.

[17] P. D. Welch, “The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short modified periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. AU-15, no. 2, pp. 70–73, June 1967.

Initialize variables on the first frame for all frequency bins k:

$$\hat{\phi}_{x\,x}(k,0) = |X(k,0)|^2; \qquad \hat{\phi}_{y\,x}(k,0) = \hat{\phi}_{w\,u}(k,0) = Y(k,0)\,X^*(k,0).$$
$$\hat{\phi}_{s\,s}(k,0) = P_s(k,0) = 0; \quad \hat{A}(k,0) = 1$$

For all time frames ℓ

For all frequency bins k

Compute the recursively averaged periodograms $\hat{\phi}_{x\,x}(k,\ell)$ and $\hat{\phi}_{y\,x}(k,\ell)$ using (32) and (33).

Compute the signal presence probability $p(k,\ell)$ using [10], the time-varying smoothing parameter $\tilde{\alpha}_u(k,\ell)$ using (36), and the cross-PSD of the interfering signals $\hat{\phi}_{w\,u}(k,\ell)$ using (35).

Compute the OM-LSA gain function $G(k,\ell)$ using [10], and the recursively averaged periodograms of the desired signal $\hat{\phi}_{s\,s}(k,\ell)$ using (34).

Compute the step-size parameter $\mu(k,\ell)$ and the estimation error $\hat{\varepsilon}(k,\ell)$ using (38) and (39).

If the time-frequency bin contains some desired signal energy (*e.g.,* in the event that $10\,\log_{10}[\hat{\phi}_{s\,s}(k,\ell)/\hat{\phi}_{x\,x}(k,\ell)] > -10\,\text{dB}$), then update the estimate for the system's transfer function $\hat{A}(k,\ell)$ using (37).

Fig. 1. On-line speech-based system identification algorithm.

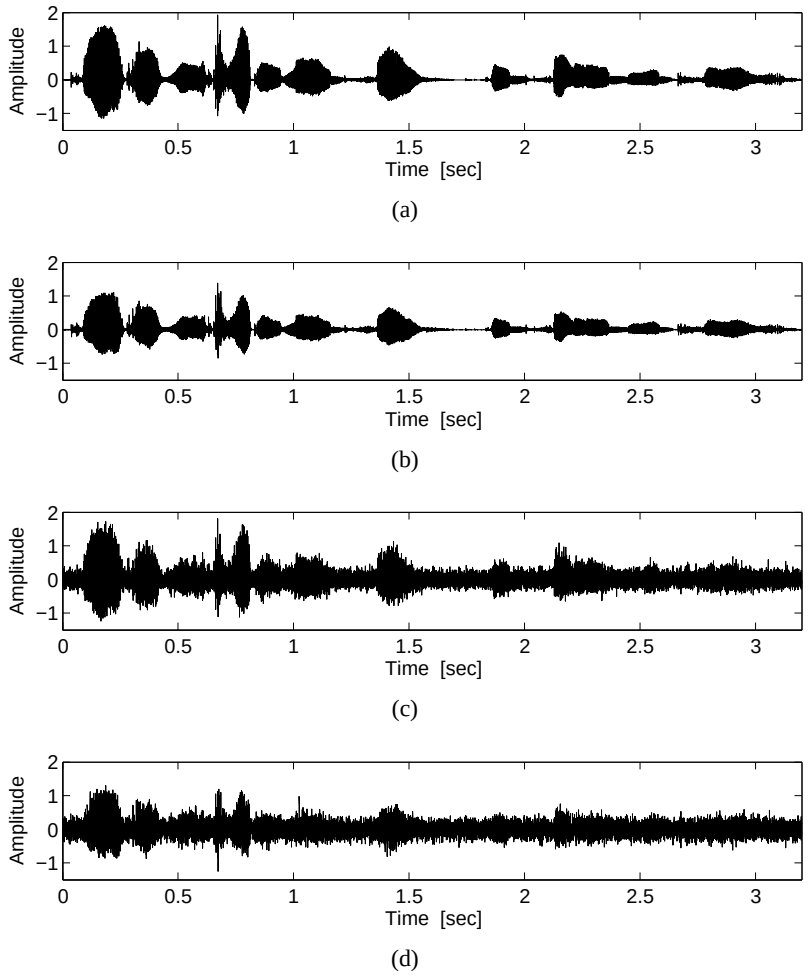
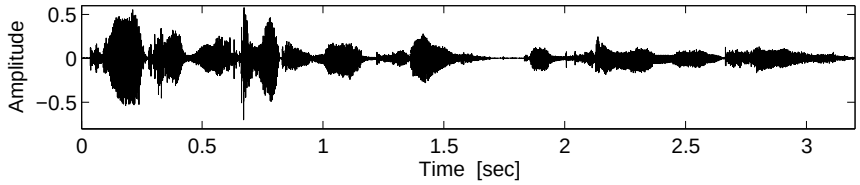
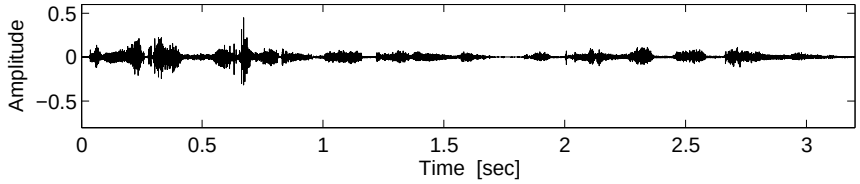


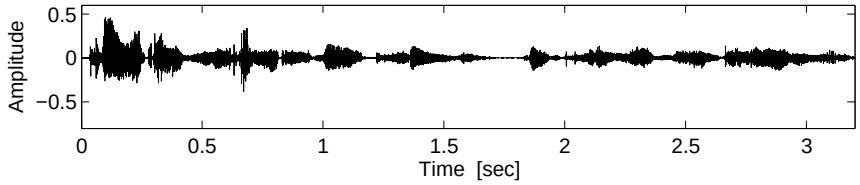
Fig. 2. Speech waveforms. (a) Clean signal $s(t)$ at the primary sensor: "Draw every outer line first, then fill in the interior."; (b) Reverberated clean signal $a(t) * s(t)$ at the reference sensor; (c) The observed noisy signal at the primary sensor (SNR $= 4.0$ dB); (d) The observed noisy signal at the reference sensor (SNR $= -0.1$ dB).



(a)



(b)



(c)

Fig. 3. Signal leakage $r(t)$ in stationary noise environment: (a) Nonstationarity-based system identification (SBF $= 9.1$ dB); (b) Speech-based system identification (SBF $= 18.5$ dB); (c) On-line speech-based system identification (SBF $= 13.9$ dB).

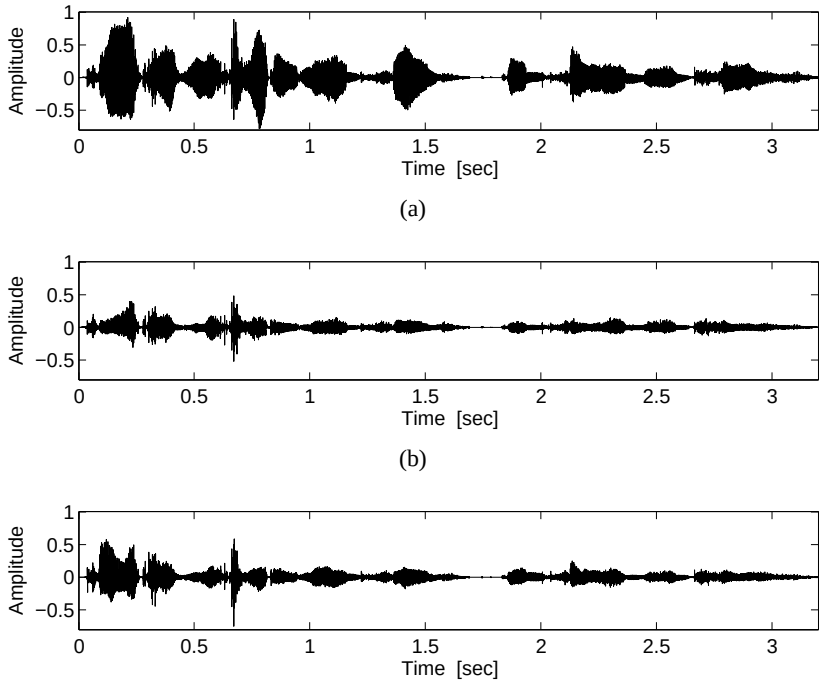


Fig. 4. Signal leakage $r(t)$ in nonstationary noise environment: (a) Nonstationarity-based system identification (SBF $= 4.9$ dB); (b) Speech-based system identification (SBF $= 13.8$ dB); (c) On-line speech-based system identification (SBF $= 11.5$ dB).

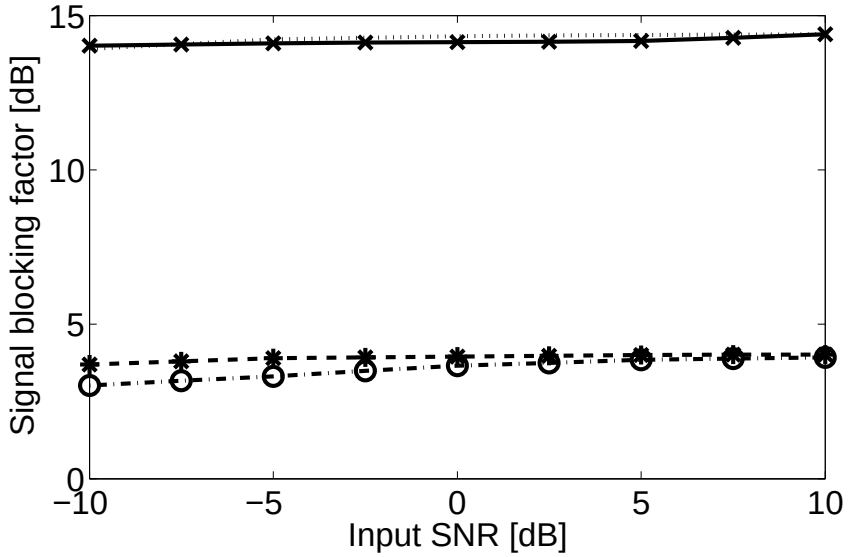


Fig. 5. Average signal blocking factor (SBF) under various car noise conditions. Nonstationarity-based system identification in pseudo-stationary (dashed, $*$) and nonstationary (dash-dot, $\circ$) car noise environments; Speech-based system identification in pseudo-stationary (solid, $\times$) and nonstationary (dotted) car noise environments.