Speech Enhancement Based on the General Transfer Function GSC and Postfiltering

Sharon Gannot, *Member, IEEE,* and Israel Cohen, *Senior Member, IEEE*

***Abstract*—In speech enhancement applications microphone array postfiltering allows additional reduction of noise components at a beamformer output. Among microphone array structures the recently proposed *general transfer function generalized sidelobe canceller* (TF-GSC) has shown impressive noise reduction abilities in a directional noise field, while still maintaining low speech distortion. However, in a diffused noise field less significant noise reduction is obtainable. The performance is even further degraded when the noise signal is nonstationary. In this contribution we propose three postfiltering methods for improving the performance of microphone arrays. Two of which are based on single-channel speech enhancers and making use of recently proposed algorithms concatenated to the beamformer output. The third is a multichannel speech enhancer which exploits noise-only components constructed within the TF-GSC structure. This work concentrates on the assessment of the proposed postfiltering structures. An extensive experimental study, which consists of both objective and subjective evaluation in various noise fields, demonstrates the advantage of the multichannel postfiltering compared to the single-channel techniques.**

***Index Terms*—Generalized sidelobe canceller, microphone arrays, nonstationarity, postfiltering, speech enhancement.**

I. Introduction

RECENTLY, an extension to the classical *generalized sidelobe canceller* (GSC), suggested by Griffiths and Jim [1], which deals with arbitrary transfer functions (TFs), was suggested by Gannot *et al.* [2], [3]. Although providing good results in the directional noise case, there is a significant degradation in the performance of the array, in nondirectional noise environments such as the *diffused noise* case [4], [5]. Furthermore, as the TF-GSC algorithm exploits the speech nonstationarity in concert with the noise stationarity, a significant performance degradation is expected in nonstationary noise environment.

The use of postfiltering is therefore called upon to improve the beamforming performance in nondirectional and nonstationary noise environments. Postfiltering for the simple *delay and sum* beamformer, based on the Wiener filter, has been suggested by Zelinski [6]. Later, postfiltering was incorporated into the Griffiths and Jim GSC beamformer [7], [8]. The authors suggest the use of two postfilters in succession. The first works on the fixed beamformer branch, and the second uses the GSC output. In directional noise source and the low frequency band of a diffused noise field, correlation between the noise components at each sensor exists. While the first postfilter is rendered useless in this case, the latter suppresses the noise. The low frequency band correlation in a diffused noise field is somewhat mitigated by using several harmonically nested subarrays in conjunction with the Wiener postfilter [9]. This structure is thoroughly analyzed by Marro *et al.* [10].

Note, that the beamformer output might be treated as a single channel containing speech signal and contaminated by (the residual) noise signal. This observation suggests the use of state-of-the-art single microphone speech enhancement algorithms. In [11], the use of the spectral subtraction algorithm [12] is suggested.

In this contribution, the use of two more modern algorithms is proposed and assessed. The first is the *mixture-maximum* (MIXMAX) algorithm [13], [14]. The second is the *optimally modified log spectral amplitude* estimator (OM-LSA) [15]. However, if the noise signal is both diffused and nonstationary, the single microphone postfilters fail to suppress it completely.

A method dealing with nonstationary noise sources was first suggested by Cohen and Berdugo [16]. This postfiltering method is working in conjunction with the classical Griffiths and Jim GSC beamformer and making use of both the beamformer output and noise reference signals resulting from the blocking branch, thus constituting multimicrophone postfiltering.

In this paper, we extend this method and incorporate it into the TF-GSC beamformer suggested by Gannot *et al.* [2]. The advantage of the TF-GSC is its ability to steer itself toward the desired speech signal, and to eliminate the desired signal leakage into the noise reference branch, even in a highly reverberated environment. The new multimicrophone postfilter method is assessed in various noise fields and compared with the single microphone postfilters.

The scenario of the problem is presented in Section II. The TF-GSC is briefly reviewed in Section III. The proposed multimicrophone postfilter is presented in Section IV. Section V is devoted to the assessment of the proposed method and to a comparison with the single microphone postfilters. Some conclusions are drawn in Section VI.

II. Problem Formulation

Consider an array of sensors in a noisy and reverberant environment. The received signal is comprised of three components. The first is a speech signal (The TF-GSC was originally suggested for enhancing an arbitrary nonstationary signal. In this

Manuscript received March 23, 2002; revised December 18, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dirk van Compernolle.

S. Gannot is with the School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel (e-mail: gannot@eng.biu.ac.il).

I. Cohen is with the Faculty of Electrical Engineering, Technion, Haifa 32000, Israel (e-mail: icohen@ee.technion.ac.il).

Digital Object Identifier 10.1109/TSA.2004.834599

contribution we limit the discussion to speech signals alone, as the postfiltering is relying on the specific speech characteristics). The second is some stationary interference signal and the third is some nonstationary (transient) noise component. Our goal is to reconstruct the speech component from the received signals. Thus, the received signals are given by

$$z_m(t) = a_m(t) * s(t) + n_m^s(t) + n_m^t(t); \quad m = 1, \ldots, M \quad (1)$$

where $z_m(t)$ is the mth sensor signal, $s(t)$ is the desired speech source, * denotes convolution operation. $n_m^s(t)$ and $n_m^t(t)$ are the stationary and transient noise components, respectively. Note, that both noise components might be comprised of coherent (directional) noise component and diffused noise component. $a_m(t)$ is the mth time-varying *acoustical transfer function* (ATF) from the speech source to the mth sensor. Using short term frequency analysis and assuming time-invariant ATFs we have in the time-frequency domain in a vector form

$$\boldsymbol{Z}(t, e^{j\omega}) = \boldsymbol{A}(e^{j\omega}) S(t, e^{j\omega}) + \boldsymbol{N}_s(t, e^{j\omega}) + \boldsymbol{N}_t(t, e^{j\omega}) \quad (2)$$

where

$$\begin{aligned}
\boldsymbol{Z}^T(t, e^{j\omega}) &= \left[Z_1(t, e^{j\omega}) \quad Z_2(t, e^{j\omega}) \quad \cdots \quad Z_M(t, e^{j\omega})\right] \\
\boldsymbol{A}^T(e^{j\omega}) &= \left[A_1(e^{j\omega}) \quad A_2(e^{j\omega}) \quad \cdots \quad A_M(e^{j\omega})\right] \\
\boldsymbol{N}_s^T(t, e^{j\omega}) &= \left[N_1^s(t, e^{j\omega}) \quad N_2^s(t, e^{j\omega}) \quad \cdots \quad N_M^s(t, e^{j\omega})\right] \\
\boldsymbol{N}_t^T(t, e^{j\omega}) &= \left[N_1^t(t, e^{j\omega}) \quad N_2^t(t, e^{j\omega}) \quad \cdots \quad N_M^t(t, e^{j\omega})\right]
\end{aligned}$$

and $Z_m(t, e^{j\omega})$, $S(t, e^{j\omega})$, $N_m^s(t, e^{j\omega})$, and $N_m^t(t, e^{j\omega})$ are the short time Fourier transforms (STFT) of the respective signals. $A_m(e^{j\omega})$ is the frequency response of the mth sensor ATF, assumed to be time invariant during the analysis period.

III. Summary of the TF-GSC Algorithm

An approach for signal enhancement based on the desired signal nonstationarity was suggested by Gannot *et al.* [2], [3]. The M microphone signals are filtered by a corresponding set of M filters, $W_m^*(t, e^{j\omega}); m = 1, \ldots, M$ (* denotes conjugation), and their outputs are summed to form the beamformer output

$$Y(t, e^{j\omega}) = \boldsymbol{W}^\dagger(t, e^{j\omega}) \boldsymbol{Z}(t, e^{j\omega}) \quad (3)$$

where † denotes conjugation transpose. $\boldsymbol{W}(t, e^{j\omega})$ is given by

$$\boldsymbol{W}^T(t, e^{j\omega}) = \left[W_1(t, e^{j\omega}) \quad W_2(t, e^{j\omega}) \quad \cdots \quad W_M(t, e^{j\omega})\right].$$

$\boldsymbol{W}(t, e^{j\omega})$ is determined by minimizing the output power subject to the constraint that the signal portion of the output is the desired signal, $S(t, e^{j\omega})$, up to some prespecified filter $\mathcal{F}^*(t, e^{j\omega})$ (usually a simple delay). This minimization can be efficiently implemented by constructing a GSC structure as depicted in Fig. 1.

The GSC solution is comprised of three components: A fixed beamformer (FBF) implemented by $\boldsymbol{W}_0^\dagger(t, e^{j\omega})$, a blocking matrix (BM) implemented by $\mathcal{H}^\dagger(e^{j\omega})$ that constructs the noise reference signals (both stationary and transient components) and

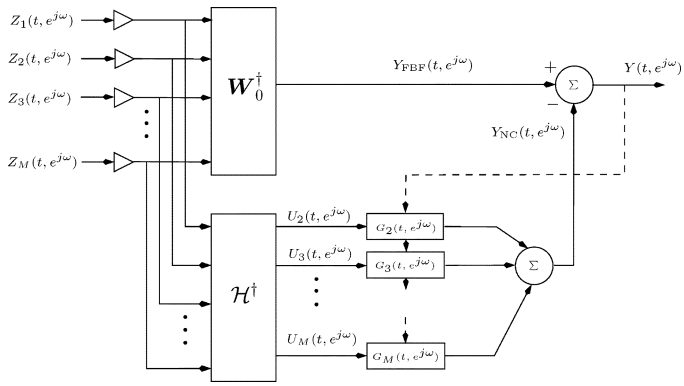


Fig. 1. GSC solution for the general TFs case (TF-GSC).

1. **TF-s ratios**: $\boldsymbol{H}(e^{j\omega}) = \frac{\boldsymbol{A}(e^{j\omega})}{A_1(e^{j\omega})}$.
2. Construct a **blocking matrix**, $\mathcal{H}^\dagger(e^{j\omega})\boldsymbol{A}(e^{j\omega}) = 0$.
3. Fixed beamformer (FBF) $\boldsymbol{W}_0(t, e^{j\omega}) = \frac{\boldsymbol{H}(e^{j\omega})}{\|\boldsymbol{H}(e^{j\omega})\|^2}\mathcal{F}(e^{j\omega})$.
   FBF output: $Y_{\mathrm{FBF}}(t, e^{j\omega}) = \boldsymbol{W}_0^\dagger(e^{j\omega})\boldsymbol{Z}(t, e^{j\omega})$.
4. **Noise reference signals**:
   $\boldsymbol{U}(t, e^{j\omega}) = \mathcal{H}^\dagger(e^{j\omega})\boldsymbol{Z}(t, e^{j\omega}) = \mathcal{H}^\dagger(e^{j\omega})\boldsymbol{N}(t, e^{j\omega})$
   (or $U_m(e^{j\omega}) = Z_m(t, e^{j\omega}) - \frac{A_m(e^{j\omega})}{A_1(e^{j\omega})} Z_1(t, e^{j\omega})\,; m = 2, \ldots, M$).
5. **Output signal**: $Y(t, e^{j\omega}) = Y_{\mathrm{FBF}}(t, e^{j\omega}) - \boldsymbol{G}^\dagger(t, e^{j\omega})\boldsymbol{U}(t, e^{j\omega})$.
6. **Filters update**. For $m = 1, \ldots, M-1$:
   $\tilde{G}_m(t+1, e^{j\omega}) = G_m(t, e^{j\omega}) + \mu \frac{U_m(t, e^{j\omega}) Y^*(t, e^{j\omega})}{P_{\mathrm{est}}(t, e^{j\omega})}$
   $G_m(t+1, e^{j\omega}) \overset{\mathrm{FIR}}{\longleftarrow} \tilde{G}_m(t+1, e^{j\omega})$
   where, $P_{\mathrm{est}}(t, e^{j\omega}) = \rho P_{\mathrm{est}}(t-1, e^{j\omega}) + (1-\rho)\sum_m |Z_m(t, e^{j\omega})|^2$.
7. Keep only non-aliased samples, according to the
   overlap & save method [18].

Fig. 2. Summary of the TF-GSC algorithm.

a multichannel noise canceller (NC) implemented by the filters $\boldsymbol{G}(t, e^{j\omega})$. The filters $\boldsymbol{G}(t, e^{j\omega})$ are adjusted to minimize the power at the output, $Y(t, e^{j\omega})$, exactly as in the classical Widrow problem [17]. The filters are usually constrained to an FIR structure for stabilizing the update algorithm.

Although an exact knowledge of the ATFs $\boldsymbol{A}(e^{j\omega})$ would yield distortionless reconstruction of the desired speech signal, it has been shown that the ATFs ratio, $\boldsymbol{H}(e^{j\omega})$, alone is sufficient in practice. Using the following definition for the ATFs ratio

$$\boldsymbol{H}(e^{j\omega}) = \frac{\boldsymbol{A}(e^{j\omega})}{A_1(e^{j\omega})} = \left[1 \quad \frac{A_2(t, e^{j\omega})}{A_1(t, e^{j\omega})} \quad \cdots \quad \frac{A_M(t, e^{j\omega})}{A_1(t, e^{j\omega})}\right]$$

a suboptimal FBF block becomes $\boldsymbol{W}_0(t, e^{j\omega}) = (\boldsymbol{H}(e^{j\omega})/\|\boldsymbol{H}(e^{j\omega})\|^2)\mathcal{F}(e^{j\omega})$. The blocking matrix $\mathcal{H}(e^{j\omega})$ can also be determined by using the ATFs ratio alone [2]. The algorithm is summarized in Fig. 2, where, the ATFs ratio vector is assumed to be known. However, in practice $\boldsymbol{H}(e^{j\omega})$ is unknown and should be estimated. We use an estimation method which is based on the nonstationarity of the desired signal. The analysis interval is split into frames, such that the desired signal may be considered stationary during each frame (quasistationarity assumption for speech signals), while $H_m(e^{j\omega})$ is still considered fixed during

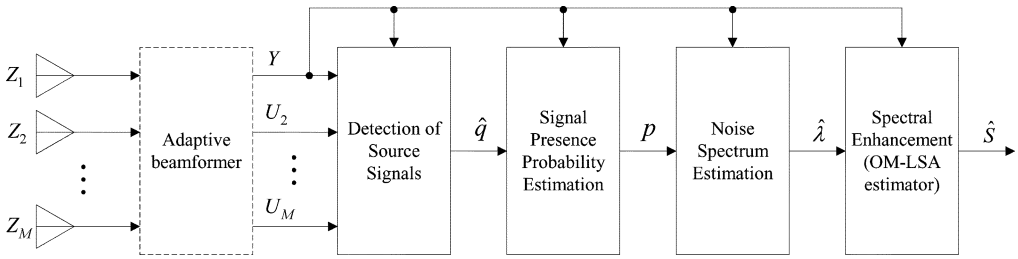


Fig. 3. Block diagram of the multimicrophone postfiltering.

the entire analysis interval. Define, $\Phi_{z_i z_j}^{(k)}(e^{j\omega})$ to be the cross-PSD (*power spectral density*) between z_i and z_j (ith and jth noisy signal observations, respectively) during the kth frame ($k = 1, \ldots, K$). Further define $\Phi_{u_m z_1}(e^{j\omega})$ to be the cross-PSD between $u_m(t)$ (mth noise reference signal) and $z_1(t)$. Let $\hat{\Phi}_{z_i z_j}^{(k)}(e^{j\omega})$ and $\hat{\Phi}_{u_m z_1}^{(k)}(e^{j\omega})$ represent the corresponding estimates. An unbiased estimate for $H_m(e^{j\omega})$ is obtained by applying *least squares* fit to the following set of over-determined equations

$$\begin{bmatrix} \hat{\Phi}_{z_m z_1}^{(1)}(e^{j\omega}) \\ \hat{\Phi}_{z_m z_1}^{(2)}(e^{j\omega}) \\ \vdots \\ \hat{\Phi}_{z_m z_1}^{(K)}(e^{j\omega}) \end{bmatrix} = \begin{bmatrix} \hat{\Phi}_{z_1 z_1}^{(1)}(e^{j\omega}) & 1 \\ \hat{\Phi}_{z_1 z_1}^{(2)}(e^{j\omega}) & 1 \\ \vdots & \\ \hat{\Phi}_{z_1 z_1}^{(K)}(e^{j\omega}) & 1 \end{bmatrix} \times \begin{bmatrix} H_m(e^{j\omega}) \\ \Phi_{u_m z_1}(e^{j\omega}) \end{bmatrix} + \begin{bmatrix} \varepsilon_m^{(1)}(e^{j\omega}) \\ \varepsilon_m^{(2)}(e^{j\omega}) \\ \vdots \\ \varepsilon_m^{(K)}(e^{j\omega}) \end{bmatrix} \quad (4)$$

where a separate set of equations is used for each microphone signal ($m = 2, \ldots, M$) and frequency index ($e^{j\omega}$), and K is the number of frames within the analysis interval. The error term to be minimized is defined by $\varepsilon_m^{(k)}(e^{j\omega}) = \Phi_{u_m z_1}(e^{j\omega}) - \hat{\Phi}_{u_m z_1}(e^{j\omega})$; $k = 1, \ldots, K$.

IV. Multimicrophone Postfilter

In this section, we address the problem of estimating the noise PSD at the beamformer output, and present the multimicrophone postfiltering technique. Fig. 3 describes the block diagram of the proposed postfiltering approach. Desired speech components are detected at the beamformer output, using the ratio between the transient power at the beamformer output Y and the transient power at the reference signals $\{U_k\}_{k=2}^{M}$. Then an estimate $\hat{q}(t, e^{j\omega})$ for the *a priori* speech absence probability is derived, and the speech presence probability $p(t, e^{j\omega})$ is estimated based on a Gaussian statistical model. Subsequently, the noise PSD is estimated by recursively smoothing the periodogram of the beamformer output, where the speech presence probability controls the time-varying frequency-dependent smoothing parameter to prevent the noise estimate from increasing as a result of speech components. Finally, spectral enhancement of the beamformer output is achieved by applying an OM-LSA gain function, which minimizes the mean-square error of the log-spectra [15].

Let $\mathcal{S}$ be a smoothing operator in the power spectral domain, defined by

$$\mathcal{S}Y(t, e^{j\omega}) = \alpha_s \cdot \mathcal{S}Y(t-1, e^{j\omega}) + (1-\alpha_s) \sum_{\omega'=-\Omega}^{\Omega} b\left(e^{j\omega'}\right) \left|Y\left(t, e^{j(\omega-\omega')}\right)\right|^2 \quad (5)$$

where $\alpha_s (0 \leq \alpha_s \leq 1)$ is a forgetting factor for the smoothing in time, and b is a normalized window function ($\sum_{\omega'=-\Omega}^{\Omega} b(e^{j\omega'}) = 1$) that determines the order of smoothing in frequency (2Ω is the frequency bandwidth). Let $\mathcal{M}$ denote a *minima controlled recursive averaging* (MCRA) estimator for the PSD of the background pseudo-stationary noise [19], [20]. Then, we define a *transient beam-to-reference ratio* (TBRR) [16] as shown at the bottom of the page, where ε is a constant (typically $\varepsilon = 0.01$), preventing the denominator from decreasing to zero in the absence of a transient power at the reference signals. This gives a ratio between the transient power at the beamformer output and the transient power at the reference signals, which indicates whether a transient component is more likely derived from speech or from environmental noise. Assuming that the steering error of the beamformer is relatively low, and that the interfering noise is uncorrelated with the desired speech, the TBRR is generally higher if transients are related to desired sources [21]. For desired source components, the transient power of the beamformer output is significantly larger than that of the reference signals. Hence, the nominator in (6) is much larger than the denominator. On the other hand, for interfering transients, the TBRR is smaller than 1, since the

$$\psi(t, e^{j\omega}) = \frac{\max\left\{\mathcal{S}Y(t, e^{j\omega}) - \mathcal{M}Y(t, e^{j\omega}), 0\right\}}{\max\left\{\left\{\mathcal{S}U_m(t, e^{j\omega}) - \mathcal{M}U_m(t, e^{j\omega})\right\}_{m=2}^{M}, \varepsilon\mathcal{M}Y(t, e^{j\omega})\right\}} \quad (6)$$

transient power of at least one of the reference signals is larger than that of the beamformer output. By modifying the speech presence probability based on the TBRR, we can generate a double mechanism for nonstationary noise reduction: First, through a fast update of the noise estimate (an increase in the noise estimate essentially results in lower spectral gain). Second, through the spectral gain computation (the spectral gain is exponentially modified by the speech presence probability [15]).

Let $\gamma_s(t,e^{j\omega}) \triangleq |Y(t,e^{j\omega})|^2/\mathcal{M}Y(t,e^{j\omega})$ denote a posteriori SNR at the beamformer output with respect to the pseudo-stationary noise. Then, the likelihood of speech presence is high only if both $\gamma_s(t,e^{j\omega})$ and $\psi(t,e^{j\omega})$ are large. A large value of $\gamma_s(t,e^{j\omega})$ implies that the beamformer output contains a transient, while the TBRR indicates whether such a transient is desired or interfering. Therefore

$$\hat{q}(t,e^{j\omega}) = \begin{cases} 1, & \text{if } \gamma_s(t,e^{j\omega}) \le \gamma_{\text{low}} \\ & \text{or } \psi(t,e^{j\omega}) \le \psi_{\text{low}} \\ \max\left\{\frac{\gamma_{\text{high}}-\gamma_s(t,e^{j\omega})}{\gamma_{\text{high}}-\gamma_{\text{low}}}, \right. & \\ \left. \frac{\psi_{\text{high}}-\psi(t,e^{j\omega})}{\psi_{\text{high}}-\psi_{\text{low}}}, 0\right\}, & \text{otherwise} \end{cases} \tag{7}$$

can be used as a heuristic expression for estimating the *a priori* speech absence probability. It assumes that speech is surely absent if either $\gamma_s(t,e^{j\omega}) \le \gamma_{\text{low}}$ or $\psi(t,e^{j\omega}) \le \psi_{\text{low}}$. Speech presence is assumed if $\gamma_s(t,e^{j\omega}) \ge \gamma_{\text{high}}$ and $\psi(t,e^{j\omega}) \ge \psi_{\text{high}}$. The constants ψ_{low} and ψ_{high} represent the uncertainty in $\psi(t,e^{j\omega})$ during speech activity, and γ_{low} and γ_{high} represent the uncertainty associated with $\gamma_s(t,e^{j\omega})$. In the regions $\gamma_s \in [\gamma_{\text{low}},\gamma_{\text{high}}]$ and $\psi \in [\psi_{\text{low}},\psi_{\text{high}}]$ we assume that $\hat{q}(t,e^{j\omega})$ is a smooth bilinear function of $\gamma_s(t,e^{j\omega})$ and $\psi(t,e^{j\omega})$.

Based on a Gaussian statistical model [22], the speech presence probability is given by

$$p(t,e^{j\omega}) = \left\{1+\frac{q(t,e^{j\omega})}{1-q(t,e^{j\omega})}\left(1+\xi(t,e^{j\omega})\right)\exp\left(-\upsilon(t,e^{j\omega})\right)\right\}^{-1} \tag{8}$$

where $\xi(t,e^{j\omega}) \triangleq E\{|S(t,e^{j\omega})|^2\}/\lambda(t,e^{j\omega})$ is the *a priori* SNR, $\lambda(t,e^{j\omega})$ is the noise PSD at the beamformer output (including the stationary as well as the nonstationary noise components), $\upsilon(t,e^{j\omega}) \triangleq \gamma(t,e^{j\omega})\xi(t,e^{j\omega})/(1+\xi(t,e^{j\omega}))$, and $\gamma(t,e^{j\omega}) \triangleq |Y(t,e^{j\omega})|^2/\lambda(t,e^{j\omega})$ is the a posteriori total SNR. The *a priori* SNR is estimated using a "decision-directed" method[1] [15]

$$\hat{\xi}(t,e^{j\omega}) = \alpha G_{H_1}^2(t-1,e^{j\omega})\gamma(t-1,e^{j\omega}) + (1-\alpha)\max\left\{\gamma(t,e^{j\omega})-1,0\right\} \tag{9}$$

where α is a weighting factor that controls the tradeoff between noise reduction and signal distortion, and

$$G_{H_1}(t,e^{j\omega}) \triangleq \frac{\xi(t,e^{j\omega})}{1+\xi(t,e^{j\omega})} exp\left(\frac{1}{2}\int_{\upsilon(t,e^{j\omega})}^{\infty}\frac{e^{-x}}{x}dx\right) \tag{10}$$

[1]This is a modified version of the "decision-directed" estimator of Ephraim and Malah [22].

Initialize variables at the first frame ($t=0$) for all frequency bins ω:

$\mathcal{S}Y(0,e^{j\omega}) = \mathcal{M}Y(0,e^{j\omega}) = \hat{\lambda}(0,e^{j\omega}) = |Y(0,e^{j\omega})|^2$

$G_{H_1}(0,e^{j\omega})1; \gamma(0,e^{j\omega}) = 1.$

For all frames $t > 0$

For all frequency bins ω

Compute the recursively averaged spectrum of the beamformer output $\mathcal{S}Y(t,e^{j\omega})$ using Eq. (5), and update the MCRA estimate of the background pseudo-stationary noise $\mathcal{M}Y(t,e^{j\omega})$ using [19].

Compute the transient beam-to-reference ratio $\psi(t,e^{j\omega})$ using Eq. (6), and compute the *a priori* speech absence probability $\hat{q}(t,e^{j\omega})$ using Eq. (7).

Compute the *a priori* SNR $\hat{\xi}(t,e^{j\omega})$ using Eq. (9), the conditional gain $G_{H_1}(t,e^{j\omega})$ using Eq. (10), and the speech presence probability $p(t,e^{j\omega})$ using Eq. (8).

Compute the time-varying frequency-dependent smoothing parameter $\tilde{\alpha}_\lambda(t,e^{j\omega})$ using Eq. (12), and update the noise spectrum estimate $\hat{\lambda}(t+1,e^{j\omega})$ using Eq. (11).

Compute an estimate for the clean signal $\hat{S}(t,e^{j\omega})$ using Eqs. (13) and (14).

Fig. 4. Multimicrophone postfiltering algorithm.

is the spectral gain function of the *log-spectral amplitude* (LSA) estimator when speech is surely present [23].

The noise estimate at the beamformer output is obtained by recursively averaging past spectral power values of the noisy measurement. The speech presence probability controls the rate of the recursive averaging. Specifically, the noise PSD estimate is given by

$$\hat{\lambda}(t+1,e^{j\omega}) = \tilde{\alpha}_\lambda(t,e^{j\omega})\hat{\lambda}(t,e^{j\omega}) + \beta\cdot\left[1-\tilde{\alpha}_\lambda(t,e^{j\omega})\right]\left|Y(t,e^{j\omega})\right|^2 \tag{11}$$

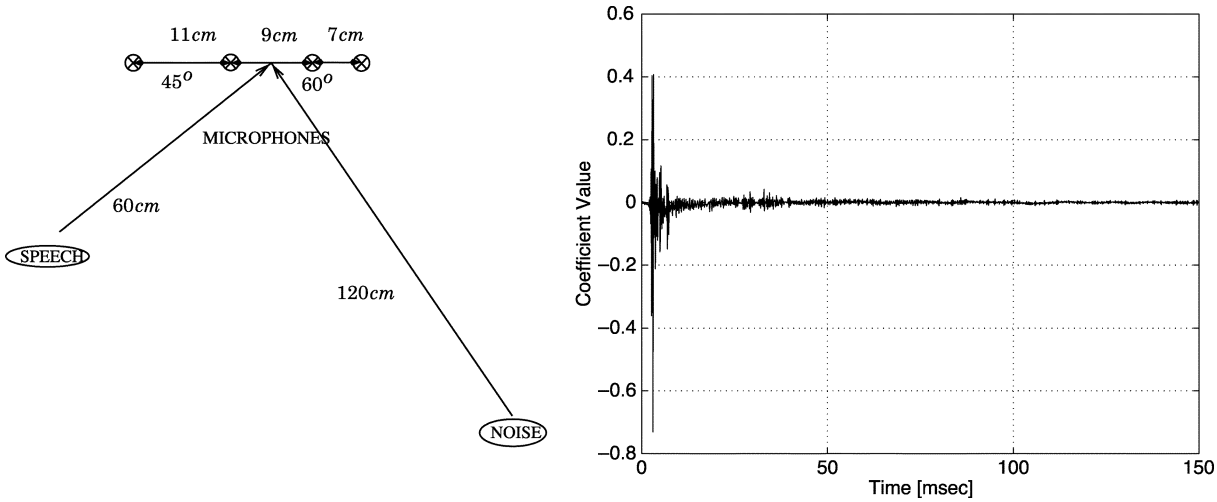
where $\tilde{\alpha}_\lambda(t,e^{j\omega})$ is a time-varying frequency-dependent smoothing parameter, and β is a factor that compensates the bias when speech is absent [19]. The smoothing parameter is determined by the speech presence probability $p(t,e^{j\omega})$, and a constant $\alpha_\lambda (0<\alpha_\lambda<1)$ that represents its minimal value

$$\tilde{\alpha}_\lambda(t,e^{j\omega}) \triangleq \alpha_\lambda + (1-\alpha_\lambda)p(t,e^{j\omega}). \tag{12}$$

When speech is present, $\tilde{\alpha}_\lambda(t,e^{j\omega})$ is close to 1, thus preventing the noise estimate from increasing as a result of speech components. In case of speech absence and stationary background noise or interfering transients, the TBRR as defined in (6) is relatively small (compared to ψ_{low}). Accordingly, the *a priori* speech absence probability (7) increases to 1, and the speech presence probability (8) decreases to 0. As the probability of speech presence decreases, the smoothing parameter gets smaller, facilitating a faster update of the noise estimate. In particular, the noise estimate in (11) is able to manage transient as well as stationary noise components. It differentiates between transient interferences and desired speech components by using the power ratio between the beamformer output and the reference signals.

An estimate for the clean signal STFT is finally given by

$$\hat{S}(t,e^{j\omega}) = G(t,e^{j\omega})Y(t,e^{j\omega}) \tag{13}$$



(a) (b)

Fig. 5. Test scenario. (a) Array of four microphones in a noisy conference room. (b) Impulse response from speech source to microphone #1.

where

$$G(t,e^{j\omega}) = \left\{G_{H_1}(t,e^{j\omega})\right\}^{p(t,e^{j\omega})} \cdot G_{\min}^{1-p(t,e^{j\omega})} \qquad (14)$$

is the OM-LSA gain function and $G_{\min}$ denotes a lower bound constraint for the gain when speech is absent. The implementation of the multichannel postfiltering algorithm is summarized in Fig. 4. Typical values of the respective parameters, for a sampling rate of 8 kHz, are given in Table II.

V. Experimental Study

In this section we apply the proposed postfiltering algorithms to the speech enhancement problem and evaluate their performance. We assess the algorithms' performance both in a conference room scenario and in a car environment and compare the simpler single microphone postfilters (MIXMAX and OM-LSA) with the more complex multimicrophone algorithm.

A. Test Scenario

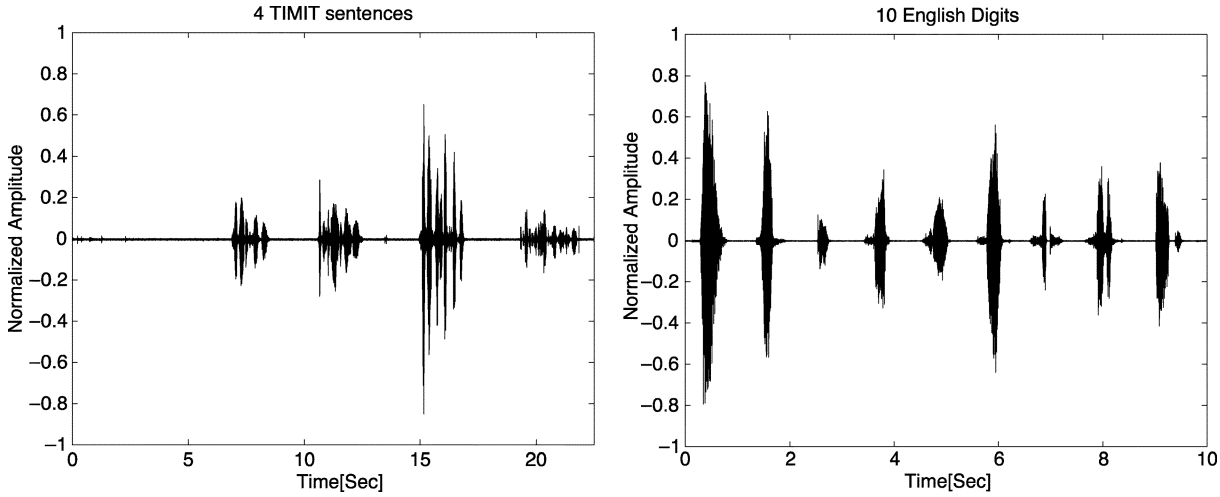
For the conference room the scenario shown in Fig. 5 was studied. The enclosure is a conference room with dimensions $5\ \mathrm{m} \times 4\ \mathrm{m} \times 2.8\ \mathrm{m}$. A linear array was placed on a table at the center of the room. Two loudspeakers were used. One for the speech source and the other for the noise source. Their locations and the locations of the four microphones are depicted in the left-hand side of Fig. 5. The impulse response from the speech source to the first microphone is depicted in the right-hand side of the figure. This response was obtained using a least squares fit between the input signal source and the received microphone signal (the response includes the loudspeaker). We note that in all our experiments we used the actual recordings and did not use the estimated impulse responses.

The speech source was comprised of four sentences drawn from the *Texas Instruments and Massachusetts Institute of Technology* (TIMIT) database [24] with various gain levels, as depicted in the left-hand side of Fig. 6. The microphone signals' input were generated by mixing speech and noise components, that were created separately, at various SNR levels, measured at the microphones. We considered three noise sources. The first was a point noise source. The second was a diffused noise source and the third was a nonstationary diffused noise source. In order to generate the point noise source, we transmitted an actual recording of fan noise (low-pass PSD) through a loudspeaker. The diffused noise source was generated by simulating an omnidirectional emittance of a flat PSD bandpass filtered noise signal based on Dal-Degan and Prati [25] method. The third was the same diffused noise source but with alternating amplitude to demonstrate the ability of the algorithm to cope with transients in the noise signals.

The car scenario was tested by actual (separate) recordings of a speech signal comprised of the ten English digits, as depicted in the right-hand side of Fig. 6, and the car noise signal. The windows of the car were slightly open. Transient noise is received as a result of passing cars and wind blows. The stationary component of the noise results from the constant hum of the road. Four microphones were mounted onto the visor in broadside steering configuration. The microphone signals were generated by mixing the speech and noise signals with various SNR levels.

B. Algorithms' Parameters

The sampling rate for the entire system was 8-kHz. In the TF-GSC algorithm the following parameters were used. The blocking filters $H_m(e^{j\omega})$ were modeled by noncausal FIR-s with 180 coefficients in the interval $[-90, 89]$. The cancelling filters $G_m(e^{j\omega})$ were modeled by noncausal FIR-s with 250 coefficients in the interval $[-125, 124]$. In order to implement the overlap & save procedure, segments with 512 samples were used. For the conference room environment, the system identification procedure utilized 13 segments, 1000-samples long each. For the car environment eight segments, 500 segments long, was proven to be sufficient. We note that system identification was



(a) (b)

Fig. 6. Clean speech signals. (a) Four TIMIT sentences in conference and (b) ten English digits in car.

applied only during active speech periods, while the noise maintains stationary characteristics. However an accurate *voice activity detector* (VAD) is not necessary for this purpose.

Three types of postfiltering procedures were applied, namely, MIXMAX, OM-LSA and the multmicrophone.

For the MIXMAX algorithm [13], [14] the frame length was set to $L = 256$ (with 50% overlapping), which corresponds to $K = 129$ relevant frequency bins. Threshold levels for limiting the noise canceller gain were set to $\delta_k = 0.35$ for $0 \leq k \leq 36$, and $\delta_k = 0.18$ for $37 \leq k \leq 128$, i.e. the algorithm gain was limited by the given values of δ_k in each frequency bin.

For the OM-LSA algorithm the STFT is implemented with Hamming windows of 256 samples length (32 ms) and 64 samples frame update step (75% overlapping frames). The *a priori* SNR is estimated using the modified decision-directed approach with $\alpha = 0.92$. The spectral gain is restricted to a minimum of -20 dB, and the noise PSD is estimated using the *improved* MCRA technique [19]. Values of parameters used for the estimation of the *a priori* speech absence probability are summarized in Table I (the estimator and its parameters are described in [15]).

The multimicrophone postfilter parameters are shown in Table II.

TABLE I
VALUES OF PARAMETERS USED IN THE OM-LSA ALGORITHM FOR THE ESTIMATION OF THE *A PRIORI* SPEECH ABSENCE PROBABILITY

$\beta = 0.7$	$\zeta_{min} = -10\text{dB}$	$\zeta_{p\,\min} = 0\text{dB}$
$w_{local} = 1$	$\zeta_{max} = -5\text{dB}$	$\zeta_{p\,\max} = 10\text{dB}$
$w_{global} = 15$	$q_{max} = 0.95$	h_λ: Hann windows

TABLE II
VALUES OF PARAMETERS USED IN THE IMPLEMENTATION OF THE PROPOSED MULTIMICROPHONE POSTFILTERING

$\alpha = 0.92$	$\alpha_s = 0.9$	$\alpha_\lambda = 0.85$	$\beta = 1.47$
$\psi_{low} = 1$	$\psi_{high} = 3$	$\gamma_{low} = 1$	$\gamma_{high} = 4.6$
$b = [\ 0.25 \quad 0.5 \quad 0.25\ ]$		$\epsilon = 0.01$	$G_{min} = -20\text{dB}$

C. Objective Evaluation

Three objective quality measures were used to asses the algorithms' performance.

The first objective quality measure is the *noise level* (NL) during nonactive speech periods, defined as

$$\mathrm{NL} = \mathrm{Mean_t}\left\{10\log_{10}\left(\mathrm{E(t)}, \mathrm{t} \in \text{Speech Nonactive}\right)\right\}$$

where $E(t) = \sum_{\tau \in T_t} y^2(\tau)$, $y(t)$ is the signal to be assessed (noisy signal or algorithm's output) and T_t are the time instances corresponding to segment number t. Note, that the lower the NL figures are the better the result obtained by the respective algorithm is.

The second figure of merit is the *weighted segmental SNR* (W-SNR). This measure applies weights to the segmental SNR within frequency bands. The frequency bands are spaced proportionally to the ear's critical bands, and the weights are constructed according to the perceptual quality of speech.

Let, $z_{1,s}(t) = a_1(t) * s(t)$ be the speech-only part in the first microphone and $y(t)$ the signal to be assessed. Further define, $Z_{1,s}(t, B_k)$ and $Y(t, B_k)$ to be the corresponding signals at frequency band B_k. Now, define $\mathrm{SNR(t, B_k)} = (\sum_{\tau \in \mathrm{T_t}} \mathrm{Y^2}(\tau, \mathrm{B_k}))/(\sum_{\tau \in \mathrm{T_t}} (\mathrm{Y}(\tau, \mathrm{B_k}) - \mathrm{Z_{1,s}}(\tau, \mathrm{B_k}))^2)$ the SNR in segment number t and frequency band B_k. W-SNR is defined as

$$\mathrm{W\text{–}SNR} = \mathrm{Mean_t}\left\{10\log_{10}\left(\sum_{\mathrm{k}} \mathrm{W(B_k)SNR(t,B_k)}, \mathrm{t} \in \text{Speech Active}\right)\right\}.$$

The frequency bands B_k and their corresponding *importance weights* $W(B_k)$ are according to the ANSI standard [26]. Studies have shown that the W-SNR measure is more closely related to a listener's perceived notion of quality than the classical SNR or segmental SNR.

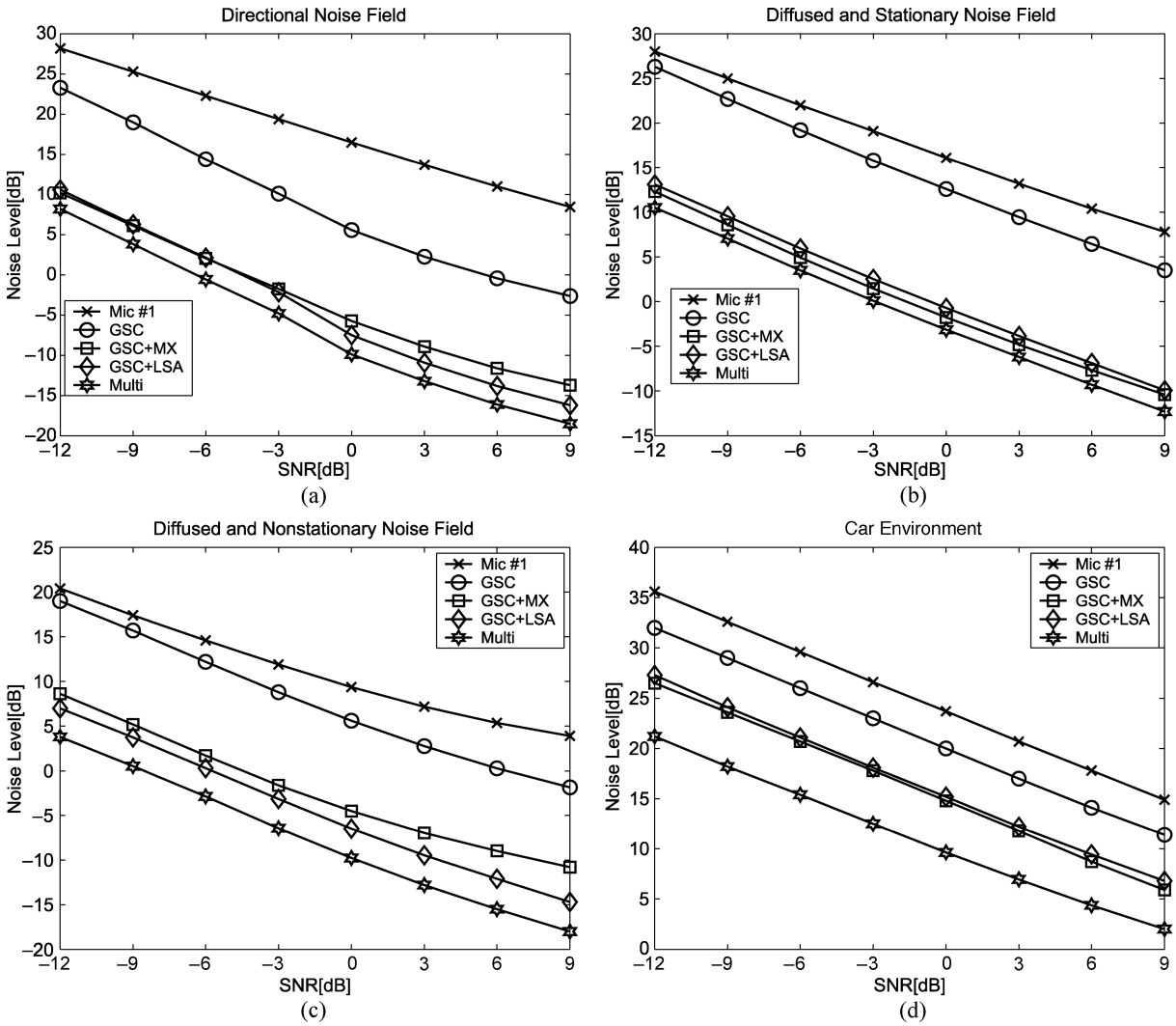


Fig. 7. Mean noise level (NL) during nonactive speech periods.

The third objective speech quality measure which is with better correlation with *mean opinion score* (MOS) is the *log spectral distance* (LSD) defined by

$$\text{LSD} = \text{Mean}_{\text{t}}\left\{\sqrt{\text{Mean}_{\omega}\left\{[20\log_{10}|\text{S}(\text{t},\text{e}^{\text{j}\omega})| - 20\log_{10}|\text{Y}(\text{t},\text{e}^{\text{j}\omega})|]^2\right\}}\right. \\ \left. t \in \text{Speech Active}\right\}.$$

Recall that $S(t, e^{j\omega})$ and $Y(t, e^{j\omega})$ are the STFT of the input and assessed signals, respectively. Note, that a lower LSD level corresponds to better performance.

The NL figure of merit is shown in Fig. 7 for the four noise conditions. It is evident from Fig. 7 that the residual noise level obtains its lowest level by using the multimicrophone postfilter for each of the noise sources. In the stationary noise cases the performance of the two single-channel postfilters (MIXMAX and OM-LSA) is comparable although somewhat degraded related to the multimicrophone postfilter. Thus, the advantage of using the multimicrophone postfilter instead of the single-microphone postfilters is less significant. The TF-GSC beamformer obtains better results in the directional noise source, and accordingly, the role of all postfilters is not as crucial as in the diffused noise field case.

In Fig. 8 results for the W-SNR are presented. Again, generally speaking, the best performance (highest W-SNR) is obtained with the multimicrophone postfilter. Its importance is more evident in the nonstationary noise cases (nonstationary diffused and car noise). In the directional (and stationary) noise field the performance of the MIXMAX postfilter and the multimicrophone postfilter is almost identical. However, the TF-GSC obtains quite good results without any postfilter. The LSD results are depicted in Fig. 9. It is evident that the results manifested by the LSD quality measure are in accordance with the previous discussion.

It is also interesting to trace the changes over time of the LSD and W-SNR figures of merits. In Fig. 10 traces for both quality measures for the car noise case is given. For convenience, the VAD decisions are also depicted in the figure. It shows that the

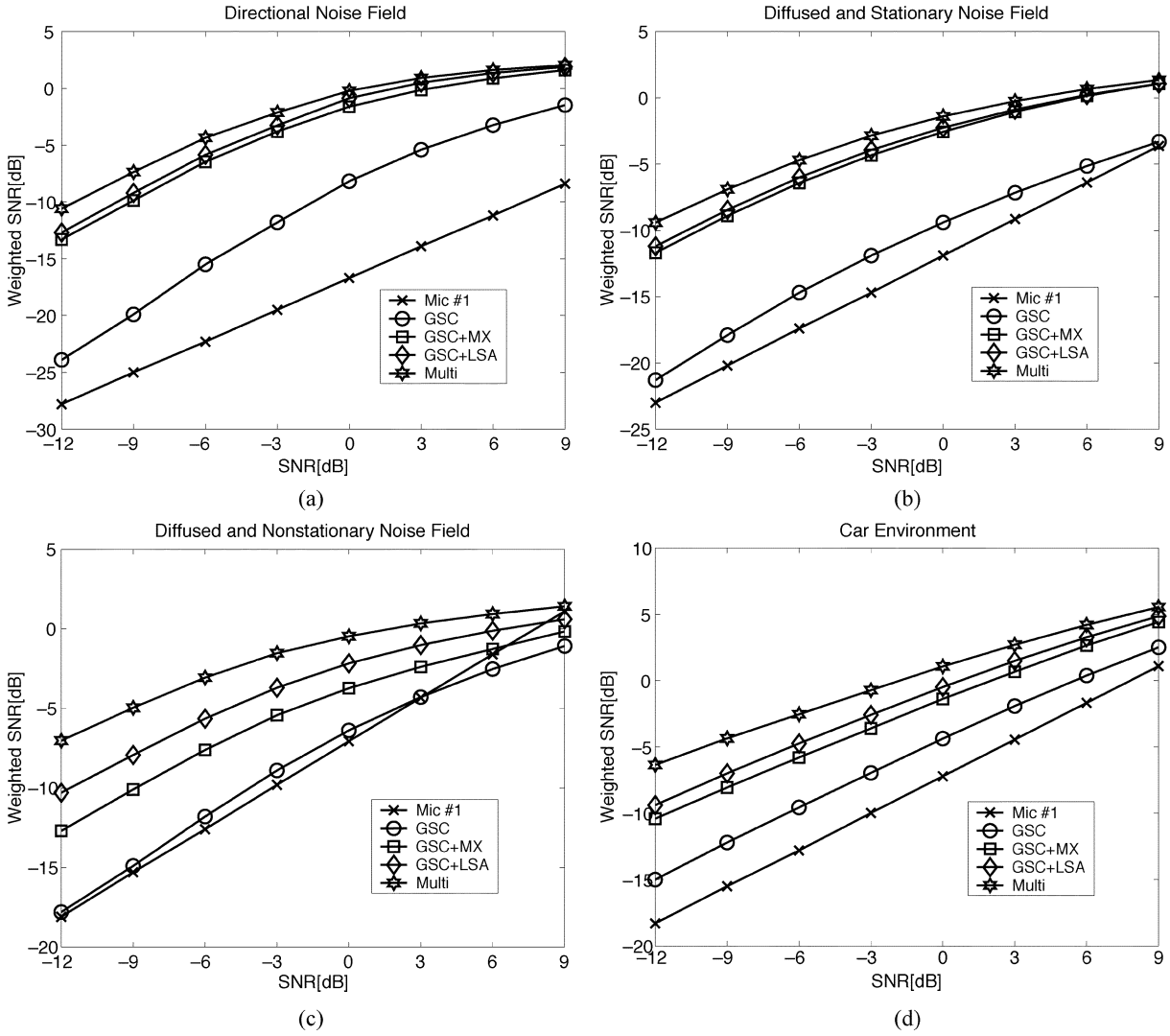


Fig. 8. Mean weighted SNR during active speech periods.

use of the multimicrophone postfilter at the TF-GSC output improves the performance. The improvement in both quality measures is particularly impressive during nonactive speech periods.

D. Subjective Evaluation

A useful subjective quality measure is the assessment of sonograms. Several observations can be drawn from the sonograms depicted in Fig. 11. Noise signal with wide frequency content is present between $t = 2.5$ [s] and $t = 4$ [s] (due to a passing car). The beamformer can not cope alone with this nonstationary noise. Although the single-microphone postfilters reduce the noise level, only the multimicrophone postfilter gives satisfactory results. Wind blows (low frequency content) are present between $t = 4.2$ [s] and $t = 5.5$ [s]. This disturbance is not completely eliminated by the multimicrophone postfilter, but it performs better than the other algorithms. The low distortion manifested by the algorithm is also evident from the sonograms.

Informal listening tests validates these conclusions. Examples of the processed speech signals can be found at [27].

VI. Conclusions

Multimicrophone arrays are often used in speech enhancement applications. It is known that the expected performance of these arrays is somewhat limited, especially when the noise field tends to become more diffused. Diffused noise field is usually assumed in car compartments. Several postfiltering methods are proposed in this work to further reduce the noise at the beamformer output. Two of the methods use modern single-microphone speech enhancers at the output of the TF-GSC beamformer. Namely, the previously proposed MIXMAX and OM-LSA algorithm are used. As an alternative, a novel multimicrophone postfilter is incorporated into the TF-GSC. The latter method improves the noise estimation by making use of the noise reference signals which are constructed within the TF-GSC. All postfiltering methods are assessed by virtue of objective (noise reduction, weighted segmental SNR and log spectral distance) and subjective quality measures (sonograms and informal listening tests). All postfilters improve the noise reduction of the combined system, especially

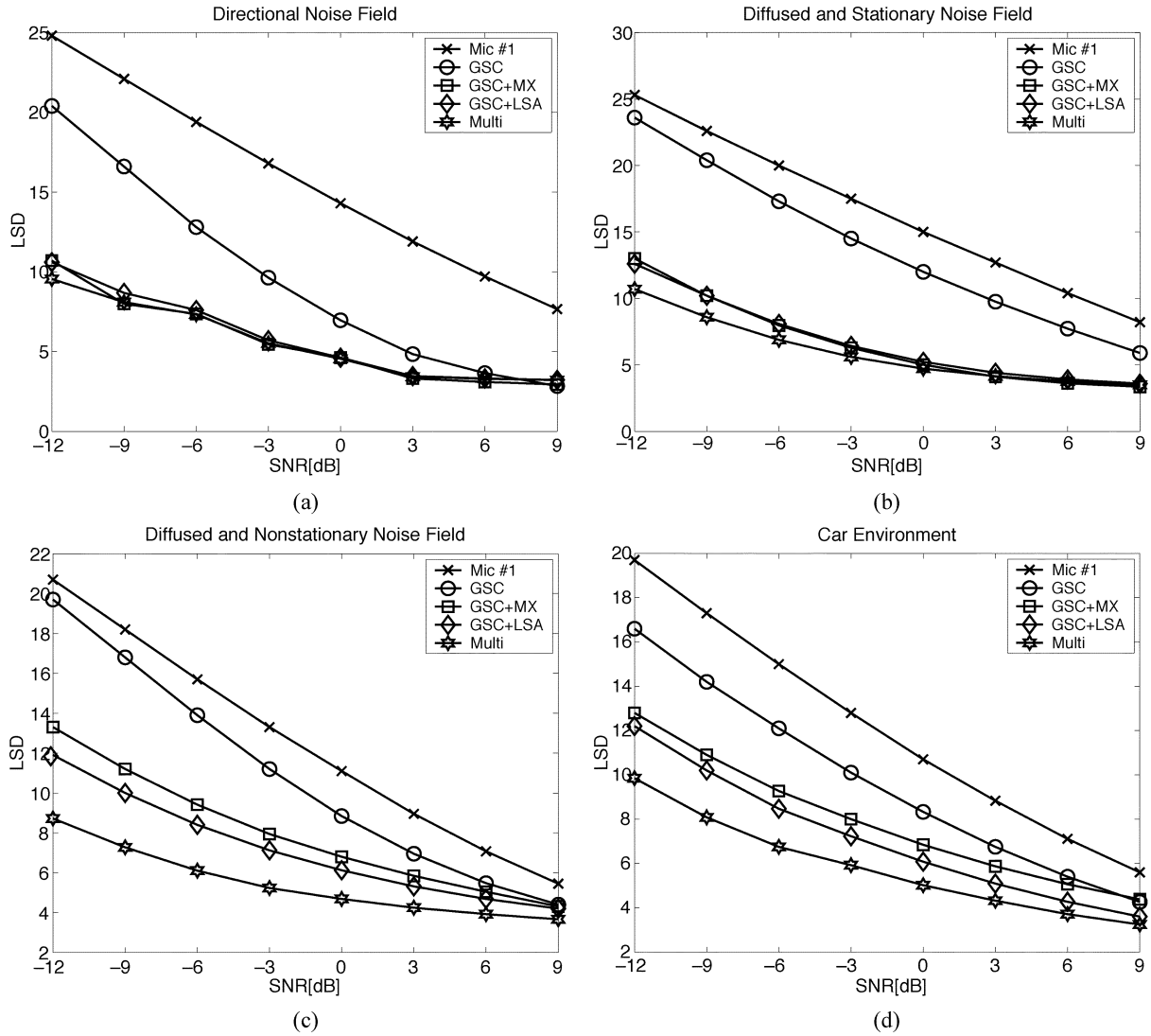


Fig. 9. Mean LSD during active speech periods.

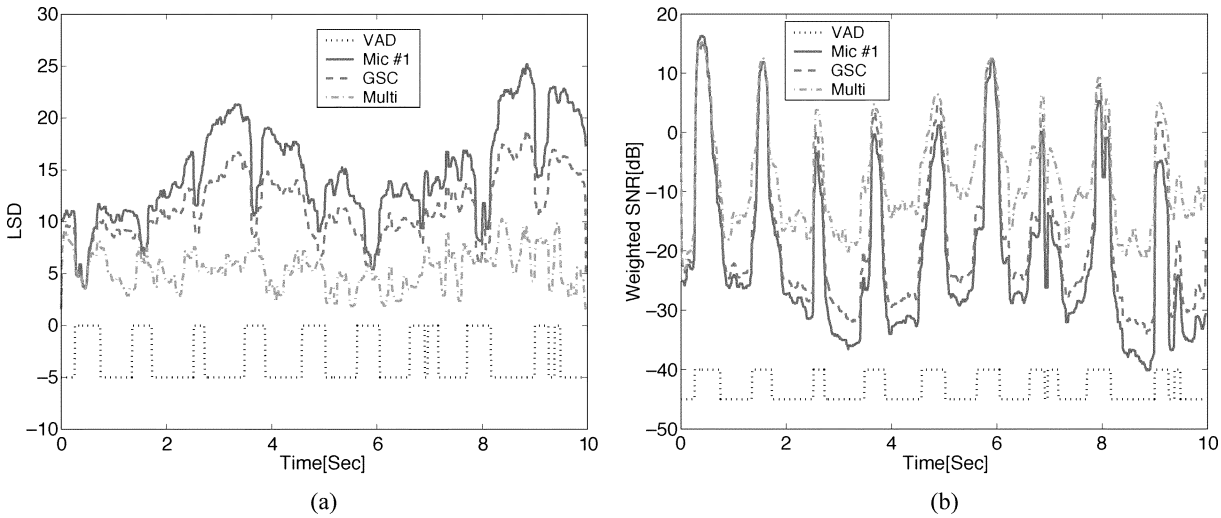


Fig. 10. Traces of LSD and W-SNR for car noise.

in the diffused noise field. However, the multimicrophone postfilter achieves the best noise reduction ability while still maintaining the low speech distortion obtained at the TF-GSC main output. This advantage is emphasized in the nonstationary noise environment, where the improved noise estimation can be more strongly manifested.

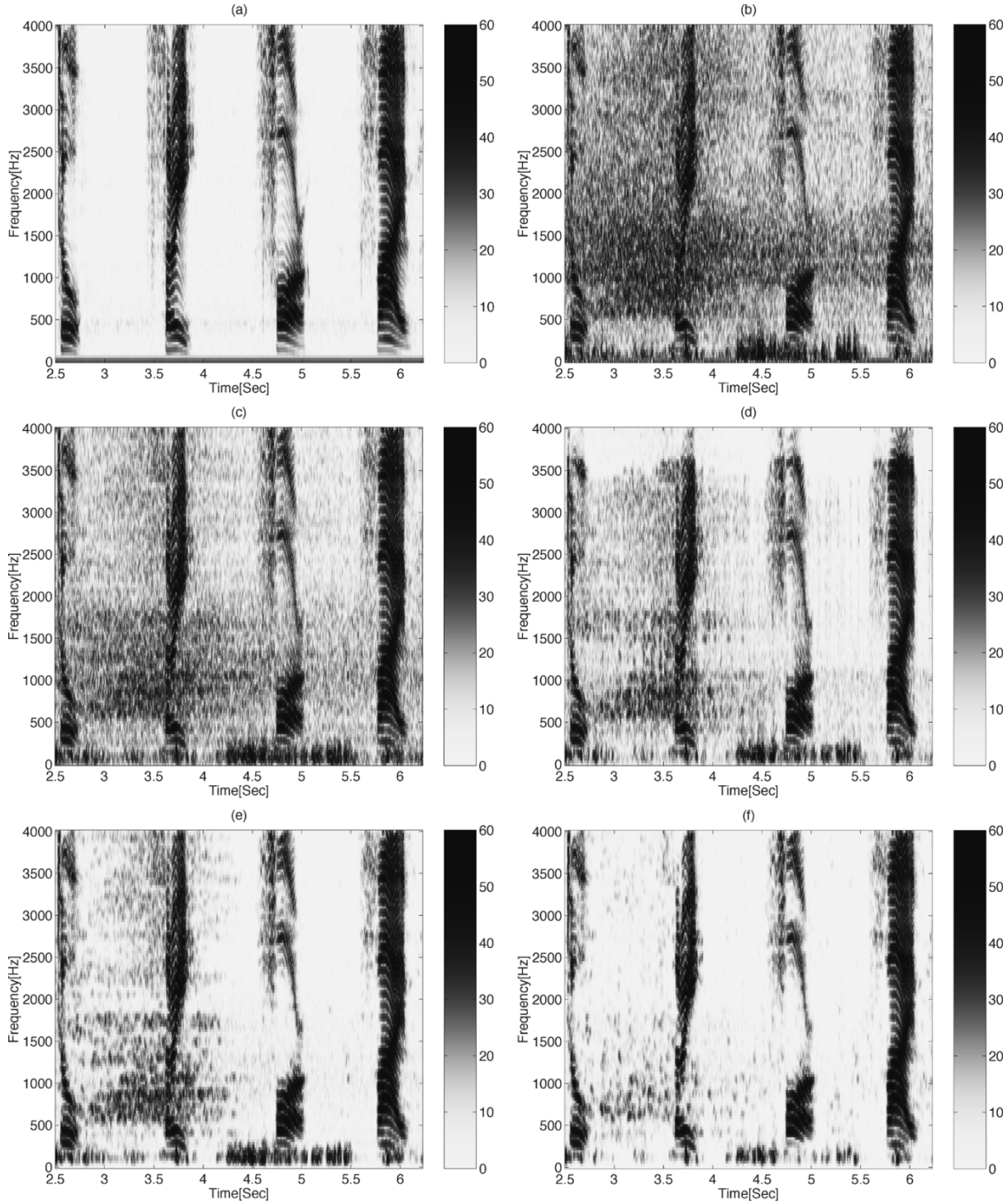


Fig. 11. (a) Sonograms of clean car signal. (b) Noisy signal at Microphone #1. (c) TF-GSC. (d) TF-GSC+MIXMAX. (e) TF-GSC+OM-LSA. (f) Multimicrophone postfilter.

REFERENCES

[1] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, pp. 27–34, Jan. 1982.

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

[3] ——, "Beamforming methods for multi-channel speech enhancement," in *Proc. Int. Workshop Acoustic Echo Noise Control*, Pocono Mannor, PA, Sept. 1999, pp. 96–99.

[4] ——, "Theoretical analysis of the general transfer function GSC," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC01)*, Darmstadt, Germany, Sept. 2001.

[5] ——, "Analysis of the Power Spectral Deviation of the General Transfer Function GSC," *IEEE Trans. Signal Processing*, vol. 52, pp. 1115–1121, Apr. 2004.

[6] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. Int. Conf. Acoustics, Speech Signal Proc.*, 1988, pp. 2578–2581.

[7] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction by post-filter and superdirective beamformer," in *Proc. Int. Workshop Acoustic Echo Noise Control*, Pocono Manor, PA, Sept. 1999, pp. 100–103.
[8] ——, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Commun.*, vol. 34, pp. 3–12, 2001.
[9] S. Fischer and K.-D. Kammeyer, "Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environment," in *Proc. Int. Conf. Acoustics, Speech Signal Proc.*, vol. 1, Munich, Germany, 1997, pp. 359–362.
[10] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 240–259, May 1998.
[11] J. Meyer and K. U. Simmer, "Multichannel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. Int. Conf. Acoustics, Speech Signal Proc.*, Munich, Germany, Apr. 1997.
[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *Speech Enhancement*, J. S. Lim, Ed. Englewood Cliffs, NJ: Prentice-hall, 1983, pp. 61–68.
[13] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," in *Proc. 6th Eur. Conf. Speech Communication Tech.—EUROSPEECH*, vol. 6, Budapest, Hungary, Sept. 1999, pp. 2591–2594.
[14] ——, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 341–351, Sept. 2002.
[15] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.
[16] I. Cohen and B. Bedugo, "Microphone array post-filtering for nonstationary noise suppression," in *Proc. Int. Conf. Acoustics, Speech Signal Proc. (ICASSP)*, Orlando, FL, May 2002, pp. 901–904.
[17] B. Widrow, J. R. Glover Jr., J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeider, E. Dong Jr., and R. C. Goodlin, "Adaptive noise cancelling: principals and applications," *Proc. IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.
[18] R. E. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 99–102, Feb. 1980.
[19] *"Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging,"*, vol. 11, pp. 466–475, Sept. 2003.
[20] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Lett.*, vol. 9, pp. 12–15, Jan. 2002.
[21] *Multi-Channel Post-Filtering in Non-Stationary Noise Environments*, vol. 52, pp. 1149–1160, May 2004.
[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.
[23] ——, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 443–445, Apr. 1985.
[24] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, Nat. Inst. Standards Technology. (1991, Oct.). *NIST Speech Disc 1-1.1* [CD-ROM]
[25] N. Dal-Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio application," *Signal Processing*, vol. 15, no. 4, pp. 43–56, Jul. 1988.
[26] ANSI, Specifications for Octave-Band and Fractional-Octave-Band Analog and Digital Filters, S1.1-1986 (ASA 65-1986), 1993.
[27] S. Gannot and I. Cohen. (2002) Audio Sample Files. [Online] http://www.eng.biu.ac.il/~gannot/examples1.html

Sharon Gannot (S'92–M'01) received the B.Sc. degree (4) from the Technion-Israeli Institute of Technology, Haifa, Israel, in 1986, and the M.Sc. (*cum laude*), and Ph.D. degrees from Tel-Aviv University, Tel-Aviv, Israel, in 1995 and 2000 respectively, all in electrical engineering.

From 1986 to 1993, he was Head of research and development for the Israeli Defense Forces. In the year 2001 he held a post-doctoral position in the Department of Electrical Engineering (SISTA) at Katholieke Universiteit (K.U.), Leuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Signal and Image Processing Lab (SIPL), Faculty of Electrical Engineering, Technion-Israeli Institute of Technology, Israel. Currently, he is a Lecturer in the School of Engineering, Bar-Ilan University, Bar-Ilan, Israel. His research interests include parameter estimation, statistical signal processing, and speech processing, using either single or multimicrophone arrays. He serves as an Associate Editor for the *Eurasip Journal of Applied Signal Processing.*

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

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 at the Computer Science Department, Yale University, New Haven, CT. Since 2001, he has been a Senior Lecturer with the Electrical Engineering department, Technion, Israel. 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 serves as Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING and IEEE SIGNAL PROCESSING LETTERS.