REAL-TIME TF-GSC IN NONSTATIONARY NOISE ENVIRONMENTS

Israel Cohen[1], *Sharon Gannot*[1] *and Baruch Berdugo*[2]

[1]Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel
[2]Lamar Signal Processing Ltd., Andrea Electronics Corp., P.O.Box 573, Yokneam Ilit 20692, Israel
{icohen,gannot}@ee.technion.ac.il; bberdugo@lamar.co.il

ABSTRACT

Adaptive beamforming techniques are inefficient for eliminating transient noise components that randomly arrive from unpredictable directions. In this paper, we present a real-time *transfer function generalized sidelobe canceller* (TF-GSC) for such nonstationary noise environments. Hypothesis testing in the spectral domain indicates either absence of transients, presence of an interfering transient, or presence of desired source components. The noise canceller branch of the TF-GSC is updated only during absence of transients, while the identification of the acoustical transfer function is carried out only when desired source components are present. Following the beamforming and the hypothesis testing, estimates for the signal presence probability, the noise power spectral density, and the desired speech log-spectral amplitude are derived. Experimental results demonstrate the usefulness of the proposed approach under nonstationary noise conditions.

1. INTRODUCTION

Adaptive beamforming techniques are inefficient for eliminating diffuse noise and nonstationary noise components that randomly arrive from unpredictable directions. Conventional postfiltering methods are useful for the former type of noise, but not for the latter. The time variation of the interfering signals is usually assumed to be sufficiently slow, such that the postfilter can track and adapt to the changes in the noise statistics. Unfortunately, transient interferences are often much too brief and abrupt for real-time tracking. In [1], a multichannel postfilter is combined with the *transfer function generalized sidelobe canceller* (TF-GSC) [2], and its performance is compared with that of a single-channel postfilter. The use of both the beamformer primary output and the reference noise signals (resulting from the blocking branch of the TF-GSC) for distinguishing between speech transients and noise transients, enables the algorithm to work in nonstationary noise environments, and allows handling of abrupt noise spectral variations. However, in former contributions the beamformer stage feeds the postfilter, but the adverse is not true. Taking into account the strong correlation of speech presence in the time-frequency domain, hypothesis testing made by the postfilter for distinguishing between speech, stationary noise and transient noise, can be used in the beamformer to enable real-time applications. This will also enable on-line tracking of time-varying *acoustical transfer functions* (ATFs) in case of moving sources.

In this paper, we present a real-time TF-GSC, which includes hypothesis testing as a feedback process to the adaptive beamformer. The beamformer is based on the TF-GSC, but the requirement for the stationarity of the noise is relaxed. The noise canceller branch of the TF-GSC is updated only during absence of transients, and the ATF identification is carried out only when desired source components are present. Following the beamforming and the hypothesis testing, estimates for the signal presence probability, the noise power spectral density, and the desired speech log-spectral amplitude are derived. Experimental evaluation of the proposed system, with comparison to an off-line system, demonstrates the performance in nonstationary noise environments.

2. TRANSFER FUNCTION GENERALIZED SIDELOBE CANCELLING

Let $x(t)$ denote a desired speech source signal that, subject to some acoustic propagation, is received by M microphones along with additive uncorrelated interfering signals. The interference at the ith sensor comprises a pseudo-stationary noise signal, $d_{is}(t)$, and a transient noise component, $d_{it}(t)$. The observed signals are given by

$$z_i(t) = a_i(t) * x(t) + d_{is}(t) + d_{it}(t), \quad i = 1, \ldots, M \qquad (1)$$

where $a_i(t)$ is the acoustical transfer function from the desired source to the ith sensor, and $*$ denotes convolution. Using the short-time Fourier transform (STFT), we have in the time-frequency domain

$$\mathbf{Z}(k,\ell) = \mathbf{A}(k,\ell)X(k,\ell) + \mathbf{D}_s(k,\ell) + \mathbf{D}_t(k,\ell) \qquad (2)$$

where k represents the frequency bin index, ℓ the frame index, and $\mathbf{Z}(k,\ell)$, $\mathbf{A}(k,\ell)$, $\mathbf{D}_s(k,\ell)$ and $\mathbf{D}_t(k,\ell)$ are the corresponding M-dimensional vectors. The observed noisy signals are processed by the system shown in Fig. 1. The beamformer comprises three parts: 1) a transfer-function beamformer $\mathbf{W}(k,\ell) \triangleq \frac{\tilde{\mathbf{A}}(k,\ell)}{\|\tilde{\mathbf{A}}(k,\ell)\|^2}$, where $\tilde{\mathbf{A}}(k,\ell) \triangleq \frac{\mathbf{A}(k,\ell)}{A_1(k,\ell)}$ denotes ATF ratios, which aligns the desired signal components; 2) a blocking matrix $\mathbf{B}(k,\ell)$, which blocks the desired components thus yielding the reference noise signals $\mathbf{U}(k,\ell) = \mathbf{B}^H(k,\ell)\,[\mathbf{D}_s(k,\ell) + \mathbf{D}_t(k,\ell)]$; 3) a multichannel adaptive noise canceller $\mathbf{H}(k,\ell) = [H_2(k,\ell), \ldots, H_M(k,\ell)]^T$, which eliminates the stationary noise that leaks through the sidelobes of the fixed beamformer.

Let three hypotheses $\mathcal{H}_{0s}$, $\mathcal{H}_{0t}$ and $\mathcal{H}_1$ indicate respectively absence of transients, presence of an interfering transient, and presence of a desired source transient at the beamformer output. The optimal solution for the noise cancelling filter $\mathbf{H}(\ell)$ (for notational simplicity, we omit the argument k throughout the rest of the paper) is obtained by minimizing the power of the beamformer

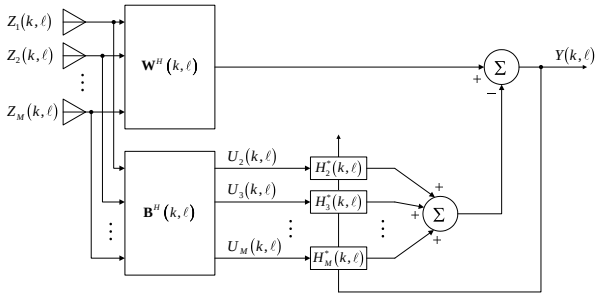

Fig. 1. Block diagram of the transfer function generalized sidelobe canceller (TF-GSC).

output during the stationary noise frames:

$$\min_{\mathbf{H}} \left\{ [\mathbf{W}(\ell) - \mathbf{B}(\ell)\mathbf{H}(\ell)]^H \, \mathbf{\Phi_{ZZ}}(\ell) \, [\mathbf{W}(\ell) - \mathbf{B}(\ell)\mathbf{H}(\ell)] \right\} \Big|_{\mathcal{H}_{0s}} \tag{3}$$

where $\mathbf{\Phi_{ZZ}}(\ell) = E\left\{\mathbf{Z}(\ell)\mathbf{Z}^H(\ell)\right\}$ denotes the PSD matrix of the observed signals. Using the normalized LMS algorithm [3] we have

$$\mathbf{H}(\ell+1) = \begin{cases} \mathbf{H}(\ell) + \frac{\mu_h}{P_{\text{est}}(\ell)} \mathbf{U}(\ell) Y^*(\ell)\,, & \text{if } \mathcal{H}_{0s} \text{ is true,} \\ \mathbf{H}(\ell)\,, & \text{otherwise,} \end{cases} \tag{4}$$

where μ_h is a step factor that regulates the convergence rate, and $P_{\text{est}}(\ell)$ is a proper normalization factor.

The ATF identification is carried out based on time-frequency bins that contain desired source components. Let $\mathcal{L}$ represent a set of past frames where $\mathcal{H}_1$ is true, and let $\left\{ \hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}(\ell) \,|\, \ell \in \mathcal{L} \right\}$ denote a corresponding set of past PSD estimates. Then, by exploiting the nonstationarity of the desired signal [4, 2], an estimate for the ATF ratios is recursively updated by

$$\tilde{\mathbf{A}}(\ell) = \frac{\left\langle \hat{\phi}_{Z_1 Z_1}(\ell) \hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}(\ell) \right\rangle - \left\langle \hat{\phi}_{Z_1 Z_1}(\ell) \right\rangle \left\langle \hat{\boldsymbol{\phi}}_{\mathbf{Z}Z_1}(\ell) \right\rangle}{\left\langle \hat{\phi}^2_{Z_1 Z_1}(\ell) \right\rangle - \left\langle \hat{\phi}_{Z_1 Z_1}(\ell) \right\rangle^2} \tag{5}$$

where the average operation $\langle \cdot \rangle$ is defined by

$$\langle f(\ell) \rangle \triangleq \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} f(\ell)\,.$$

Note that in former work, the entire observation interval is used for the ATF identification, and $\tilde{\mathbf{A}}$ is estimated only once. Here $\tilde{\mathbf{A}}$ is updated on every frame based on past $\mathcal{H}_1$ frames, which facilitates real-time applications. Note also the tradeoff associated with the time interval used for the ATF identification. The time interval is defined by the frames in $\mathcal{L}$. On the one hand, it should be short for the ATF time-invariance assumption to hold, and therefore the maximal number of frames in $\mathcal{L}$ should be small. On the other hand, the number of frames in $\mathcal{L}$ should be large for stabilizing the solution. In frequency bins with low speech content, the interval required for obtaining an estimate for $\tilde{\mathbf{A}}(\ell)$ might be very long, since only frames for which $\mathcal{H}_1$ is true are considered. In practice, the time interval is chosen such that the estimation error due to variations in $\tilde{\mathbf{A}}(\ell)$ is comparable to the estimation error caused by restricting the size of $\mathcal{L}$.

3. HYPOTHESIS TESTING AND MULTICHANNEL POSTFILTERING

Generally, three different components are involved in the TF-GSC output: a non-stationary desired source component, a pseudo-stationary noise component, and a transient interference. Our objective is to determine which category a given time-frequency bin belongs to, based on the beamformer output $Y(\ell)$ and the reference signals $\{U_i(\ell) \,|\, 2 \le i \le M\}$. Clearly, if transients have not been detected at the beamformer output and the reference signals, we can accept the $\mathcal{H}_{0s}$ hypothesis. In case a transient is detected at the beamformer output, but not at the reference signals, the transient is likely a source component and therefore we determine that $\mathcal{H}_1$ is true. On the contrary, a transient that is detected at one of the reference signals but not at the beamformer output is likely an interfering component, which implies that $\mathcal{H}_{0t}$ is true. In case a transient is simultaneously detected at the beamformer output and at one of the reference signals, a further test is required, which involves the ratio between the transient power at beamformer output and the transient power at the reference signals.

Let $\mathcal{S}$ be a smoothing operator in the power spectral domain, and let $\mathcal{M}$ denote an estimator for the PSD of the background pseudo-stationary noise, derived using the *Minima Controlled Recursive Averaging* approach [5]. The decision rules for detecting transients at the TF-GSC output and reference signals are

$$\Lambda_Y(\ell) \triangleq \mathcal{S}Y(\ell) / \mathcal{M}Y(\ell) > \Lambda_0 \tag{6}$$

$$\Lambda_{\mathbf{U}}(\ell) \triangleq \max_{2 \le i \le M} \left\{ \frac{\mathcal{S}U_i(\ell)}{\mathcal{M}U_i(\ell)} \right\} > \Lambda_1\,, \tag{7}$$

respectively, where Λ_Y and Λ_U denote measures of the local non-stationarities (LNS), and Λ_0 and Λ_1 are the corresponding threshold values for detecting transients [6]. The *transient beam-to-reference ratio* (TBRR) is defined by the ratio between the transient power of the beamformer output and the transient power of the strongest reference signal:

$$\Omega(\ell) = \frac{\mathcal{S}Y(\ell) - \mathcal{M}Y(\ell)}{\max\limits_{2 \le i \le M} \{\mathcal{S}U_i(\ell) - \mathcal{M}U_i(\ell)\}}\,. \tag{8}$$

Transient signal components are relatively strong at the beamformer output, whereas transient noise components are relatively strong at one of the reference signals. Hence, we expect $\Omega(\ell)$ to be large for signal transients, and small for noise transients. Assuming there exist thresholds Ω_{high} and Ω_{low} such that

$$\Omega(\ell)|_{\mathcal{H}_{0t}} \le \Omega_{\text{low}} \le \Omega_{\text{high}} \le \Omega(\ell)|_{\mathcal{H}_1} \tag{9}$$

the decision rule for differentiating desired signal components from the transient interference components is

$$\begin{aligned} \mathcal{H}_{0t} &: \gamma_s(\ell) \le 1 \text{ or } \Omega(\ell) \le \Omega_{\text{low}} \\ \mathcal{H}_1 &: \gamma_s(\ell) \ge \gamma_0 \text{ and } \Omega(\ell) \ge \Omega_{\text{high}} \\ \mathcal{H}_r &: \text{otherwise} \end{aligned} \tag{10}$$

where

$$\gamma_s(\ell) \triangleq \frac{|Y(\ell)|^2}{\mathcal{M}Y(\ell)} \tag{11}$$

represents the *a posteriori* SNR at the beamformer output with respect to the pseudo-stationary noise, γ_0 denotes a constant satisfying $\mathcal{P}\left(\gamma_s(\ell) \ge \gamma_0 \mid \mathcal{H}_{0s}\right) < \epsilon$ for a certain significance level

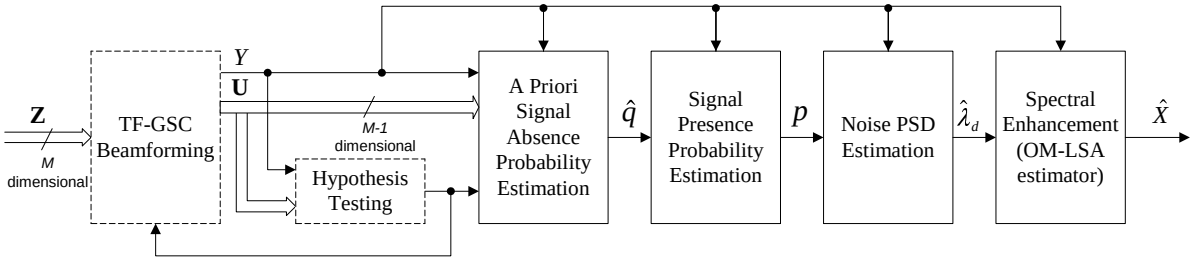

Fig. 2. Block diagram of the multichannel postfiltering.

ϵ, and $\mathcal{H}_r$ designates a *reject* option where the conditional error of making a decision between $\mathcal{H}_{0t}$ and $\mathcal{H}_1$ is high.

Following the beamforming and the hypothesis testing, a multichannel postfiltering is applied as depicted in Fig. 2. The *a priori* signal absence probability $\hat{q}(\ell)$ is set to 1 if signal absence hypotheses ($\mathcal{H}_{0s}$ or $\mathcal{H}_{0t}$) are accepted, and is set to 0 if signal presence hypothesis ($\mathcal{H}_1$) is accepted. In case of the reject hypothesis $\mathcal{H}_r$, a soft signal detection is accomplished by letting $\hat{q}(\ell)$ be inversely proportional to $\Omega(\ell)$ and $\gamma_s(\ell)$:

$$\hat{q}(\ell) = \max\left\{\frac{\gamma_0 - \gamma_s(\ell)}{\gamma_0 - 1}, \frac{\Omega_{\text{high}} - \Omega(\ell)}{\Omega_{\text{high}} - \Omega_{\text{low}}}\right\}. \qquad (12)$$

Based on a Gaussian statistical model [7], the signal presence probability is given by

$$p(\ell) = \left\{1 + \frac{q(\ell)}{1 - q(\ell)}(1 + \xi(\ell))\exp(-\upsilon(\ell))\right\}^{-1} \qquad (13)$$

where $\xi(\ell) \triangleq \lambda_x(\ell)/\lambda_d(\ell)$ is the *a priori* SNR, $\lambda_d(\ell)$ is the noise PSD at the beamformer output, $\upsilon(\ell) \triangleq \gamma(\ell)\,\xi(\ell)/(1 + \xi(\ell))$, and $\gamma(\ell) \triangleq |Y(\ell)|^2/\lambda_d(\ell)$ is the *a posteriori* SNR.

An estimate for noise PSD $\hat{\lambda}_d(\ell)$ is obtained by recursively averaging past spectral power values of the noisy measurement, using a time-varying frequency-dependent smoothing parameter [5]. Subsequently, spectral enhancement of the beamformer output is achieved by applying the *Optimally-Modified Log-Spectral Amplitude* (OM-LSA) gain function [5], which minimizes the mean-square error of the log-spectral amplitude under signal presence uncertainty.

4. EXPERIMENTAL RESULTS

In this section, the performance of the proposed real-time system is evaluated under non-stationary noise conditions, and compared to an off-line system consisting of a TF-GSC and a single-channel postfilter. The evaluation includes objective quality measures, a subjective study of speech spectrograms and informal listening tests.

A linear array, consisting of four microphones with 5 cm spacing, is 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). An interfering speaker and car noise signals are recorded while the car speed is about 60 km/h, and the window next to the driver is slightly open (about 5 cm; the other windows are closed). The input microphone signals are generated by mixing the speech and noise signals at various SNR levels in the range $[-5, 10]$ dB. Off-line TF-GSC beamforming is applied to the noisy multichannel signals, and its output is enhanced using the OM-LSA estimator. The result is referred to as single-channel postfiltering output. Alternatively, the proposed real-time integrated TF-GSC and multichannel postfiltering is applied to the noisy signals. Its output is referred to as multichannel postfiltering output.

Figure 3 shows experimental results obtained for various noise levels. The two quality measures, *segmental SNR* (SegSNR) and *log spectral distance* (LSD) [8], are evaluated at the first microphone, the off-line TF-GSC output, and the postfiltering outputs. A theoretical limit postfiltering, achievable by calculating the noise PSD from the noise itself, is also considered. It can be readily seen that TF-GSC alone does not provide sufficient noise reduction in a car environment, owing to its limited ability to reduce diffuse noise [2]. Furthermore, multichannel postfiltering is considerably better than single-channel postfiltering.

A subjective comparison between multichannel and single-channel postfiltering was conducted using speech spectrograms and validated by informal listening tests. Typical examples of speech spectrograms are presented in Fig. 4. The noise PSD at the beamformer output varies substantially due to the residual interfering components of speech, wind blows, and passing cars. The TF-GSC output is characterized by a high level of noise. Single-channel postfiltering suppresses pseudo-stationary noise components, but is inefficient at attenuating the transient noise components. By contrast, the proposed system achieves superior noise attenuation, while preserving the desired source components. This is verified by subjective informal listening tests.

5. CONCLUSION

We have described a real-time beamformer that is particularly advantageous in non-stationary noise environments. The TF-GSC primary output and the reference noise signals are exploited for deciding between speech, stationary noise and transient noise hypotheses. The decisions are used for deriving estimators for the signal presence probability and for the noise PSD. The signal presence probability modifies the spectral gain function for estimating the clean signal spectral amplitude. It is worth mentioning that the postfilter is designed for suppressing the stationary noise, as well as transient noise components that do not overlap with desired signal components in the time-frequency domain. The overlapping part between desired and undesired transients is less attenuated by the postfilter, to reduce signal distortion, particularly since such noise components are perceptually masked by the desired speech [9]. We note that the computational complexity and practical simplifications of the proposed system were not addresses.

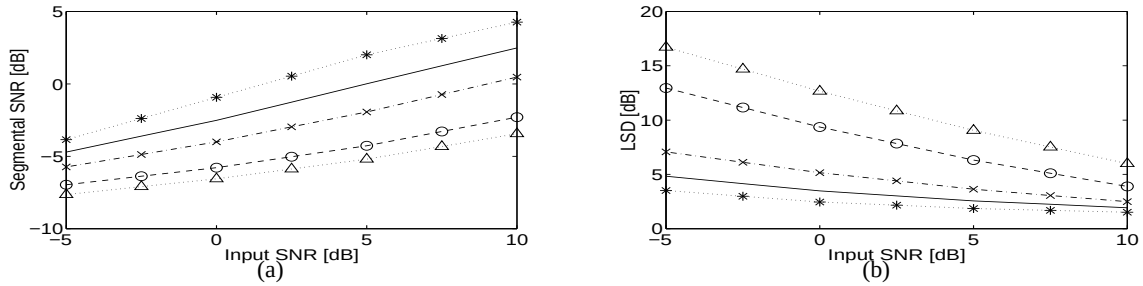

Fig. 3. (a) Average segmental SNR, and (b) average log-spectral distance, at ($\triangle$) microphone #1, ($\circ$) TF-GSC output, ($\times$) single-channel postfiltering output, (solid line) multichannel postfiltering output, and ($*$) theoretical limit postfiltering output.

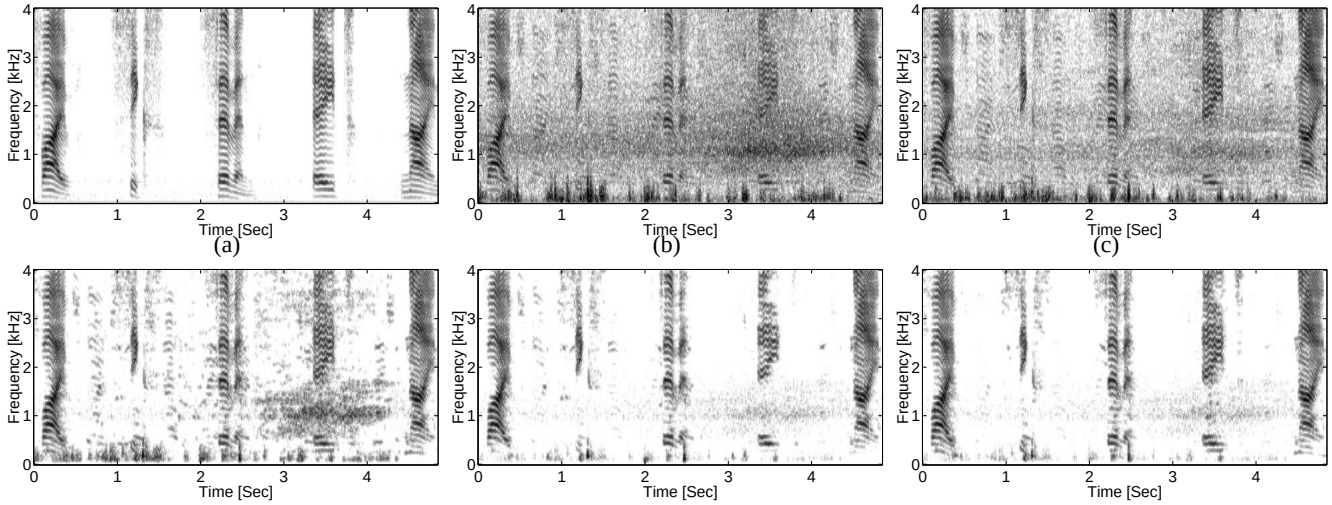

Fig. 4. Speech spectrograms. (a) Original clean speech signal at microphone #1: "Five six seven eight nine."; (b) Noisy signal at microphone #1 (SNR $= -0.9$ dB, SegSNR $= -6.2$ dB, LSD $= 15.4$ dB); (c) TF-GSC output (SegSNR $= -5.3$ dB, LSD $= 12.2$ dB); (d) Single-channel postfiltering output (SegSNR $= -3.8$ dB, LSD $= 7.4$ dB); (e) Multichannel postfiltering output (SegSNR $= -1.3$ dB, LSD $= 4.6$ dB); (f) Theoretical limit (SegSNR $= -0.4$ dB, LSD $= 4.0$ dB).

Here, the main contribution is the incorporation of the hypothesis test results into the beamformer stage. The hypotheses control the noise canceller branch of the beamformer, as well as the ATF identification, thus enabling real-time tracking of moving talkers.

6. REFERENCES

[1] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," submitted to *IEEE Trans. Speech and Audio Processing* (also CCIT Report 380, EE Pub. 1318, Technion - IIT, Haifa, Israel, May 2002).

[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] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1985.

[4] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2055–2063, August 1996.

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

[6] I. Cohen, "Multi-channel post-filtering in non-stationary noise environments," to appear in *IEEE Trans. Signal Processing* (also CCIT Report 376, EE Pub. 1314, Technion - IIT, Haifa, Israel, April 2002).

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[8] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[9] T. F. Quatieri and R.B. Dunn, "Speech enhancement based on auditory spectral chance," in *Proc. 27th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2002*, Orlando, Florida, 13-17 May 2002, pp. 257–260.