SPEECH ENHANCEMENT BASED ON THE GENERAL TRANSFER FUNCTION GSC AND POSTFILTERING

Sharon Gannot and Israel Cohen

Department of Electrical Engineering
Technion - Israel Institute of Technology, Haifa 32000, Israel
gannot@siglab.technion.ac.il;icohen@ee.technion.ac.il

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 is nonstationary . In this contribution we present 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 multi–channel speech enhancer which exploits noise–only components constructed within the TF–GSC structure. An experimental study, which consists of both objective and subjective evaluation in various noise fields, demonstrates the advantage of the multi–channel postfiltering compared to the single–channel techniques.

1. INTRODUCTION

Recently, an extension to the classical Griffiths & Jim *Generalized Sidelobe Canceller* (GSC), which deals with arbitrary transfer functions (TFs), was suggested by Gannot et al. [1]. 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. Furthermore, as noise statistics might change over time (nonstationary noise framework), the expected performance is even lower. 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 was suggested by Zelinski [2]. Later, postfiltering was incorporated into the Griffiths & Jim GSC beamformer [3].

A method dealing with nonstationary noise sources was first suggested by Cohen and Berdugo [4]. 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 multi-microphone postfiltering.

In this paper we extend this method and incorporate it into the TF–GSC beamformer suggested by Gannot et al. [1]. This method is assessed in various noise fields and compared with the single microphone postfilters. Furthermore, the use of two modern algorithms is proposed and assessed. The first is the *Mixture-Maximum* (MIXMAX) algorithm [5]. The second is the *optimally modified log spectral amplitude* estimator (OM–LSA) [6].

The scenario of the problem is presented in Section 2. The TF–GSC is briefly reviewed in Section 3. The proposed multi-microphone postfilter is presented in Section 4. Section 5 is devoted to the assessment of the proposed method and to a comparison with the single microphone postfilters.

2. 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 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. Let, $z_m(t)$ be the m-th sensor signal, $s(t)$ be the desired speech source, $n_m^s(t)$ and $n_m^t(t)$ be the stationary and transient noise components, respectively. Note, that both noise components might be comprised of coherent (directional) noise component and diffused noise component. $Z_m(t,e^{j\omega}), S(t,e^{j\omega})$, $N_m^s(t,e^{j\omega})$ 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 m-th *acoustical transfer function* (ATF) from the speech source to the m-th sensor, assumed to be time invariant during the analysis period. We have in the time–frequency domain in a vector form,

$$\boldsymbol{Z}(t,e^{j\omega}) = \boldsymbol{A}(e^{j\omega})S(t,e^{j\omega}) + \boldsymbol{N}_s(t,e^{j\omega}) + \boldsymbol{N}_t(t,e^{j\omega}) \quad (1)$$

where

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

3. SUMMARY OF THE TF–GSC ALGORITHM

An approach for signal enhancement based on the desired signal nonstationarity was suggested by Gannot et al. [1]. The M microphone signals are filtered by a corresponding set of M filters, $W_m^*(t,e^{j\omega})\ ; m = 1,\ldots,M$, and their outputs are summed to form the beamformer output, $Y(t,e^{j\omega}) = \boldsymbol{W}^\dagger(t,e^{j\omega})\boldsymbol{Z}(t,e^{j\omega})$. $\boldsymbol{W}^\dagger(t,e^{j\omega}) = \left[\, W_1^*(t,e^{j\omega})\ W_2^*(t,e^{j\omega})\ \cdots\ W_M^*(t,e^{j\omega}) \,\right]$,

* denotes conjugation and † denotes conjugation transpose. $\boldsymbol{W}(t, e^{j\omega})$ 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 pre-specified 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 Figure 1. The GSC solution is comprised of three components: A

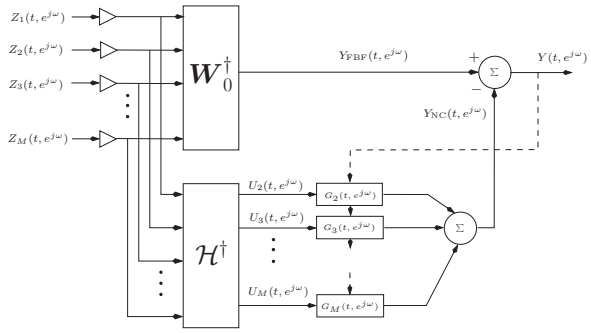

Fig. 1. GSC solution for the general TFs case (TF–GSC).

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 a multi-channel 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. The filters are usually constrained to an FIR structure for stabilizing the update algorithm. Note that, the role of minimization [by adjusting $\boldsymbol{G}(t, e^{j\omega})$] and constraining [by applying $\boldsymbol{W}_0(t, e^{j\omega})$] operations are decoupled by this structure.

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 alone, $\boldsymbol{H}(e^{j\omega})$ may be sufficient in practice. A sub-optimal FBF block, which aligns the desired signal components but does not eliminate the reverberation term $A_1(e^{j\omega})$ was used. The following $M \times (M-1)$ matrix $\mathcal{H}(e^{j\omega})$ can serve as a blocking matrix,

$$\mathcal{H}(e^{j\omega}) = \begin{bmatrix} -\frac{A_2^*(e^{j\omega})}{A_1^*(e^{j\omega})} & -\frac{A_3^*(e^{j\omega})}{A_1^*(e^{j\omega})} & \cdots & -\frac{A_M^*(e^{j\omega})}{A_1^*(e^{j\omega})} \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & \dots & \ddots & \\ 0 & 0 & \dots & 1 \end{bmatrix}. \quad (2)$$

where, the ATFs ratio vector, $\boldsymbol{H}(e^{j\omega})$, is assumed to be known. However, in practice $\boldsymbol{H}(e^{j\omega})$ is not known and should be estimated. An estimation method based on the desired signal nonstationarity was suggested in [1]. This estimation method is based on two assumptions. First, it is assumed that the ATFs ratios are slowly changing in time compared to the time variations of the desired speech signal. Second, it is assumed that no transient noise component is active during the analysis interval, i.e. the noise statistics is assumed to be fixed. These assumptions are exploited for deriving a set of equations for the same unknown ATFs ratio.

4. MULTI-MICROPHONE POSTFILTER

In this section, we address the problem of estimating the noise PSD at the beamformer output, and present the multi-microphone postfiltering technique. Desired speech components are detected at the beamformer output, and an estimate $\hat{q}(t, e^{j\omega})$ for the *a priori* speech absence probability is produced. Based on a Gaussian statistical model [7], and a decision-directed estimator for the *a priori* SNR under signal presence uncertainty [6], we derive an estimator $p(t, e^{j\omega})$ for the speech presence probability. This estimator controls the components that are introduced as noise into the PSD estimator. 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 [6].

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

$$\mathcal{S}Y(t, e^{j\omega}) = \alpha_s \cdot \mathcal{S}Y(t-1, e^{j\omega}) + (1-\alpha_s) \sum_{\omega'=-\Omega}^{\Omega} b(e^{j\omega'}) |Y(t, e^{j(\omega-\omega')})|^2 \quad (3)$$

where α_s ($0 \le \alpha_s \le 1$) 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 [8]. Then, we define a *transient beam-to-reference ratio* (TBRR) [4] by

$$\psi(t, e^{j\omega}) = \frac{\max\left\{\mathcal{S}Y(t, e^{j\omega}) - \mathcal{M}Y(t, e^{j\omega}), 0\right\}}{\max\left\{\{\mathcal{S}U_m(t, e^{j\omega}) - \mathcal{M}U_m(t, e^{j\omega})\}_{m=2}^{M}, \varepsilon\, \mathcal{M}Y(t, e^{j\omega})\right\}} \quad (4)$$

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. For desired source components, the transient power of the beamformer output is significantly larger than that of the reference signals. Hence, the nominator in (4) is much larger than the denominator. On the other hand, for interfering transients, the TBRR is smaller than 1, since the 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 [6]).

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,

$$\hat{q}(t,e^{j\omega}) = \begin{cases} 1, & \text{if } \gamma_s(t,e^{j\omega}) \le \gamma_{low} \text{ or } \psi(t,e^{j\omega}) \le \psi_{low} \\ \max\left\{\frac{\gamma_{high}-\gamma_s(t,e^{j\omega})}{\gamma_{high}-\gamma_{low}}, \frac{\psi_{high}-\psi(t,e^{j\omega})}{\psi_{high}-\psi_{low}}, 0\right\}, & \text{otherwise,} \end{cases} \quad (5)$$

can be used as a heuristic expression for estimating the *a priori* speech absence probability. It assumes that speech is surely absent if either $\gamma_s(t,e^{j\omega}) \le \gamma_{low}$ or $\psi(t,e^{j\omega}) \le \psi_{low}$. Speech presence is assumed if $\gamma_s(t,e^{j\omega}) \ge \gamma_{high}$ and $\psi(t,e^{j\omega}) \ge \psi_{high}$. 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 $\gamma_s \in [\gamma_{low}, \gamma_{high}]$ and $\psi \in [\psi_{low}, \psi_{high}]$ 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 [7], the speech presence probability is given by

$$p(t,e^{j\omega}) = \left\{1 + \frac{q(t,e^{j\omega})}{1-q(t,e^{j\omega})}(1+\xi(t,e^{j\omega}))\exp(-\upsilon(t,e^{j\omega}))\right\}^{-1} \quad (6)$$

where $\xi(t,e^{j\omega}) \triangleq E\left\{|S(t,e^{j\omega})|^2\right\}/\lambda(t,e^{j\omega})$ 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 $\gamma(t,e^{j\omega}) \triangleq \left|Y(t,e^{j\omega})\right|^2/\lambda(t,e^{j\omega})$ is the *a posteriori* total SNR. The *a priori* SNR is estimated using a "decision-directed" method [6] :

$$\hat{\xi}(t,e^{j\omega}) = \alpha\, G^2_{H_1}(t-1,e^{j\omega})\,\gamma(t-1,e^{j\omega}) + (1-\alpha)\max\left\{\gamma(t,e^{j\omega})-1, 0\right\} \quad (7)$$

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

$$G_{H_1}(t,e^{j\omega}) \triangleq \frac{\xi(t,e^{j\omega})}{1+\xi(t,e^{j\omega})} exp\left(\frac{1}{2}\int_{\upsilon(t,e^{j\omega})}^{\infty}\frac{e^{-x}}{x}\,dx\right) \quad (8)$$

is the spectral gain function of the *Log-Spectral Amplitude* (LSA) estimator when speech is surely present [9].

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

$$\hat{\lambda}(t+1,e^{j\omega}) = \tilde{\alpha}_\lambda(t,e^{j\omega})\hat{\lambda}(t,e^{j\omega}) + \beta\cdot[1-\tilde{\alpha}_\lambda(t,e^{j\omega})]|Y(t,e^{j\omega})|^2 \quad (9)$$

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. The smoothing parameter is determined by the speech presence probability $p(t,e^{j\omega})$, and a constant α_λ ($0 < \alpha_\lambda < 1$) that represents its minimal value:

$$\tilde{\alpha}_\lambda(t,e^{j\omega}) \triangleq \alpha_\lambda + (1-\alpha_\lambda)\,p(t,e^{j\omega}). \quad (10)$$

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 (4) is relatively small (compared to ψ_{low}). Accordingly, the *a priori* speech absence probability (5) increases to 1, and the speech presence probability (7) 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 Eq. (10) 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

$$\hat{S}(t,e^{j\omega}) = G(t,e^{j\omega})Y(t,e^{j\omega}), \quad (11)$$

where

$$G(t,e^{j\omega}) = \left\{G_{H_1}(t,e^{j\omega})\right\}^{p(t,e^{j\omega})} \cdot G_{min}^{1-p(t,e^{j\omega})} \quad (12)$$

is the OM-LSA gain function and G_{min} denotes a lower bound constraint for the gain when speech is absent.

5. 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 single microphone postfilters (MIXMAX and OM–LSA) with the Multi-Microphone algorithm.

The enclosure is a conference room with dimensions $5m \times 4m \times 2.8m$. A linear array comprised of four microphones $27cm$ long was placed on a table at the center of the room. Two loudspeakers were used. One, at the left of the array ($0.6m$ from its center), for the speech source and the other, at the right of the array ($1.2m$ from its center), for the noise source. The speech source was comprised of four TIMIT sentences with various levels. The microphone inputs were generated by mixing speech and noise components, that were created separately at various SNR levels. We considered three noise sources: a point source, a diffused source, and a nonstationary diffused 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 omni–directional emittance of a flat PSD bandpass filtered noise signal. 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 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. The microphone signals were generated by mixing the speech and noise signals with various SNR levels.

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,

$$\mathrm{NL} = \mathrm{Mean}_{t\in\text{speech nonactive}}\left\{10\log_{10}(E(t))\right\}$$

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

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

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 third figure of merit is the well–known *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.

The NL figure of merit is shown in Figure 2 for the four noise conditions. It is evident from Figure 2 that the residual noise level

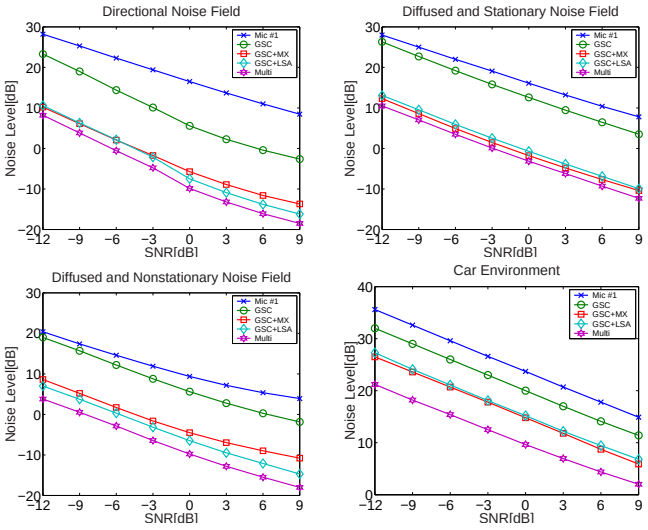

Fig. 2. Mean Noise Level (NL) during nonactive speech periods.

obtains its lowest level by using the multi-microphone 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 multi-microphone postfilter. Thus, the advantage of using the multi-microphone 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.

The LSD results are depicted in Figure 3. Generally speaking, the best performance (lowest LSD) is obtained with the Multi-Microphone postfilter. Its importance is more evident in the non-stationary noise cases (nonstationary diffused and car noise). In the directional (and stationary) noise field the performance of the MIXMAX postfilter and the multi-microphone postfilter is almost identical. However, the TF–GSC obtains quite good results without any postfilter. The results manifested by the W-SNR quality measure are in accordance with the previous discussion. Subjective evaluation of sonograms and non–formal listening tests validates these conclusions. Examples of the processed speech signals can be found at [10].

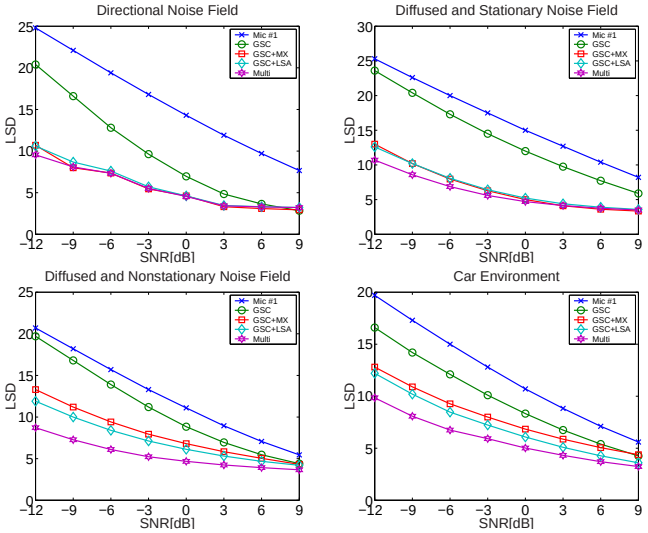

Fig. 3. Mean LSD during active speech periods.

6. REFERENCES

[1] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Non-Stationarity with application to Speech," *IEEE Trans. on Sig. Proc.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[2] R. Zelinski, "A Microphone Array With Adaptive Post-Filtering For Noise Reduction In Reverberant Rooms," in *Int. Conf. on Acoustics, Speech and Signal Proc.*, 1988, pp. 2578–2581.

[3] J. Bitzer, K.U. Simmer, and K.-D. Kammeyer, "Multi-Microphone Noise Reduction by Post-Filter and Superdirective Beamformer," in *Int. Workshop on Acoustic Echo and Noise Control*, Pocono Manor, Pennsylvania, USA, Sep. 1999, pp. 100–103.

[4] I. Cohen and B. Bedugo, "Microphone Array Posr-Filtering for Non-Stationary Noise Suppression," in *Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP'02)*, Orlando, Florida, USA, May. 2002.

[5] D. Burshtein and S. Gannot, "Speech Enhancement Using a Mixture-Maximum Model," Accepted for publication in the IEEE Trans. on Speech and Audio Proc., Jan. 2002.

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

[7] 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, 1984.

[8] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, January 2002.

[9] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 443–445, 1985.

[10] S. Gannot and I. Cohen, "Audio sample files," `http://www-sipl.technion.ac.il/~gannot`, Apr. 2002.