# THEORETICAL ANALYSIS OF THE GENERAL TRANSFER FUNCTION GSC

*Sharon Gannot*

K.U. Leuven, ESAT-SISTA
Kasteelpark Arenberg 10
B-3001 Heverlee (Leuven), Belgium
Sharon.Gannot@esat.kuleuven.ac.be

*David Burshtein and Ehud Weinstein*

Dept. of Elect. Eng. - Systems
Tel-Aviv University,
Tel-Aviv, 69978, Israel
{burstyn,udi}@eng.tau.ac.il

## ABSTRACT

In recent work we considered the use of a microphone array located in a reverberated room - where general acoustic transfer functions (ATFs) relate the source signal and the microphones - for enhancing a speech signal contaminated by interference. The resulting frequency-domain algorithm enables dealing with a complicated ATF in the same simple manner as Griffiths & Jim GSC algorithm deals with delay-only arrays. In this contribution a general expression of the enhancer output is derived. This expression is used for evaluating two figures of merit, i.e., noise reduction ability and the amount of distortion imposed. The performance is shown to be dependent on the ATFs involved, the noise field and the quality of estimation of the ATF ratios. Analytical performance evaluation of the method is obtained. It is shown that the proposed method maintains its good performance even in the general ATF case.

## 1. INTRODUCTION

The generalized sidelobe canceller, proposed by Griffiths & Jim [1], is widely used in the field of multi-microphone speech enhancement. Analytical calculations of the performance limitations have therefore attracted the attention of many researchers (e.g. [3],[4]). Recently, an extension of the GSC concept for the general ATF case was proposed by us [2]. This algorithm, nicknamed TF-GSC, have proven experimentally to outperform the classical GSC method for the more realistic room acoustics scenario. In this contribution we turn into analytical performance evaluation of this newly proposed method. In Section 2 we summarize the proposed method. In Section 3 we derive a general expression of the output power spectral density (PSD). This expression is used for evaluating the speech distortion and noise reduction in Section 3.1 and Section 3.2, respectively.

## 2. SUMMARY OF THE TF-GSC

Consider an array of sensors in a noisy and reverberant environment. Using short term frequency analysis notation (STFT) we have in vector form:

$$\mathbf{Z}(t,e^{j\omega}) = \mathbf{A}(e^{j\omega})S(t,e^{j\omega}) + \mathbf{N}(t,e^{j\omega}), \tag{1}$$

where

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

$Z_m(t,e^{j\omega})$ is the $m$-th sensor signal STFT, $S(t,e^{j\omega})$ is a desired non-stationary signal source (e.g. speech), $N_m(t,e^{j\omega})$ is a stationary interference signal at the $m$-th sensor (both coherent and ambient components) and $A_m(e^{j\omega})$ are slowly time-varying ATFs from the desired speech source to the $m$-th sensor. Figure 1 summarizes our suggested solution. The output of the algorithm is given by,

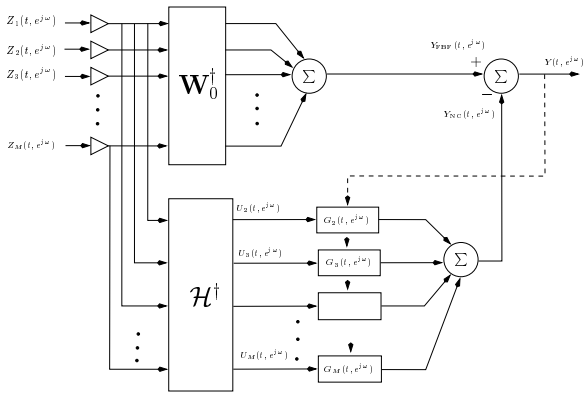


Figure 1: The general transfer function GSC. $\mathbf{W}_0$ is an ATF ratios matched filter. $\mathcal{H}$ is a blocking matrix.

$$\begin{aligned}&Y(t,e^{j\omega}) = Y_{\rm FBF}(t,e^{j\omega}) - Y_{\rm NC}(t,e^{j\omega}) = \\ &\mathbf{W}_0^\dagger(e^{j\omega})\mathbf{Z}(t,e^{j\omega}) - \mathbf{G}^\dagger(t,e^{j\omega})\mathbf{U}(t,e^{j\omega}) = \\ &\frac{\mathcal{F}^*(e^{j\omega})}{\|\widehat{\mathbf{H}}(e^{j\omega})\|^2}\widehat{\mathbf{H}}^\dagger(e^{j\omega})\mathbf{Z}(t,e^{j\omega}) - \mathbf{G}^\dagger(t,e^{j\omega})\widehat{\mathcal{H}}^\dagger(e^{j\omega})\mathbf{N}(t,e^{j\omega}).\end{aligned} \tag{2}$$

This GSC solution is comprised of three components. A fixed beamformer (FBF) implementing a matched filter followed by an arbitrary filtering operation, $\mathcal{F}(e^{j\omega})$, a blocking matrix that constructs the *reference noise* signals, $\mathbf{U}(t,e^{j\omega})$, and a multi-channel noise canceller (NC). $\widehat{\mathbf{H}}$ is an estimate of the ATF ratios,

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

and $\widehat{\mathcal{H}}$ is an estimate of the $M \times (M-1)$ blocking matrix,

$$\mathcal{H}(e^{j\omega}) = \begin{bmatrix} -\frac{A_2^*(e^{j\omega})}{A_1^*(e^{j\omega})} & -\frac{A_3^*(e^{j\omega})}{A_1^*(e^{j\omega})} & \cdots & -\frac{A_M^*(e^{j\omega})}{A_1^*(e^{j\omega})} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \cdots & \ddots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{3}$$

Thus, knowledge of the ATF ratios is sufficient to implement the algorithm. Minimization of the output power can be implemented by adjusting the filters $\mathbf{G}(t,e^{j\omega})$ using the multi-channel Wiener filter (e.g. [3]),

$$\mathbf{G}(t,e^{j\omega}) = \Phi_{\mathbf{UU}}^{-1}(t,e^{j\omega})\Phi_{\mathbf{U}Y}(t,e^{j\omega}), \tag{4}$$

where,

$$\begin{aligned}\Phi_{\mathbf{U}Y}(t,e^{j\omega}) &= E\{\mathbf{U}(t,e^{j\omega})Y_{\mathrm{FBF}}^{*}(t,e^{j\omega})\} \\ \Phi_{\mathbf{UU}}(t,e^{j\omega}) &= E\{\mathbf{U}(t,e^{j\omega})\mathbf{U}^{\dagger}(t,e^{j\omega})\}.\end{aligned} \tag{5}$$

In actual scenarios the filters $\mathbf{H}(e^{j\omega})$ are not known in advance, and have to be estimated. This estimation can be stated as a problem of system identification with known input, $z_1(t)$, and known output, $z_m(t)$, as is evident from the reference noise signal definition,

$$Z_m(t,e^{j\omega}) = H_m(e^{j\omega})Z_1(t,e^{j\omega}) + U_m(t,e^{j\omega}). \tag{6}$$

Due to obvious correlation between the noise signal and the input signal, the use of a conventional identification procedure will yield a biased solution. Instead, an estimation procedure exploiting the desired signal nonstationarity, is used [2]. The idea is to divide the received signals into frames and estimate the PSD in each of them. Exploiting the facts that the desired signal is nonstationary while the noise term is stationary and the ATFs are fixed during the observation period - a set of equations in the same unknown ATF ratios is constructed. This set can be solved by virtue of the LS procedure.

## 3. ANALYTICAL PERFORMANCE EVALUATION

Conducting the required calculations using Eqs. 2,4,5 the output PSD, $\Phi_{oo}(t,e^{j\omega}) = E\left\{Y(t,e^{j\omega})Y^{*}(t,e^{j\omega})\right\}$, can be calculated for any given input signal $\mathbf{Z}(t,e^{j\omega}) = \mathbf{S}(t,e^{j\omega})$ having a PSD $\Phi_{\mathbf{SS}}(t,e^{j\omega})$. The result is given in Eq. 7 at the top of the next page. This complicated expression forms the basis of our analytical evaluation of the proposed algorithm. It depends on various parameters. The input signal PSD $[\Phi_{\mathbf{SS}}(t,e^{j\omega})]$, the noise field used for calculating the optimal filters $[\Phi_{\mathbf{NN}}(t,e^{j\omega})]$ and the ATF ratios estimate $\widehat{\mathbf{H}}(e^{j\omega})$ [which is also used for the blocking matrix $\widehat{\mathcal{H}}(e^{j\omega})$]. Note, that since we assume independence of the desired signal and the noise signal, we can use Eq. 7 to calculate the desired signal and the noise signal contributions separately. Thus, the noise reduction and the distortion imposed by the algorithm can be calculated.

### 3.1. Desired Signal distortion

#### *3.1.1. Effects of ATF ratios estimation error*

Signal distortion is caused by errors in estimating the ATFs. This estimation error has twofold influence. First, as the FBF is not accurate, it can degrade the alignment of the signal, causing noncoherent addition. Second, the blocking matrix, which terms depend on $\mathbf{H}(e^{j\omega})$ estimate, would not block the desired signal completely, causing self-cancellation.

The distortion imposed by the algorithm can be calculated using Eq. 7 with a signal impinging the array from "direction" $\mathbf{A}(e^{j\omega})$. As the filter $\mathcal{F}(e^{j\omega})$ is an arbitrary predetermined filter and the ATF $A_1(e^{j\omega})$ can not be eliminated by the algorithm, we will define the distortion as,

$$DIS(t,e^{j\omega}) = \frac{\Phi_{oo}^{s}(t,e^{j\omega})}{|\mathcal{F}(e^{j\omega})|^2|A_1(e^{j\omega})|^2}. \tag{8}$$

This expression depends on the desired signal's ATFs, its estimation accuracy and the noise field.

The simpler situation, when only delay relates the sources and the sensors (i.e. the *direction of arrival* (DoA) of the sources completely determines the ATFs) is depicted in Figure 2. In this case the array is optimize to cancel noise

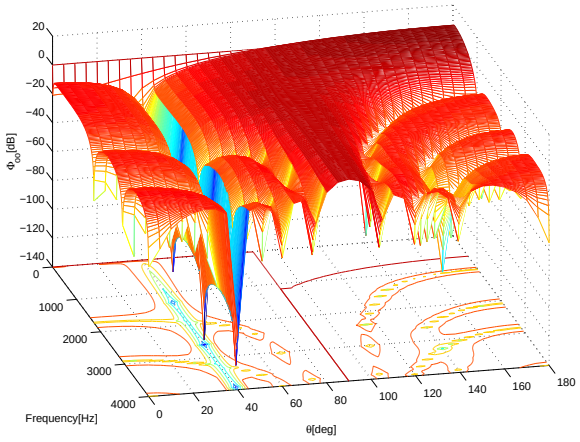


Figure 2: Output PSD of linear array with $M = 5$ sensors. Delay-only ATFs for both speech and noise.

source from $\theta = 40^{o}$ (by optimization of the array we refer to designing the optimal Wiener filter in the noise cancellation branch), and the desired signal impinges the array from DoA $\theta = 90^{o}$. It can be seen, that signals from $\theta = 90^{o}$ direction bare no loss while low distortion is caused by steering errors around the correct angle (regardless of the ATFs' identification method). The performance in diffused or incoherent noise fields is not significantly different.

#### *3.1.2. Analysis of ATF ratios estimation error*

The estimation error depends on the method used. Main results of the nonstationarity method are cited here. Let, the estimated ATF ratios vector be,

$$\widehat{\mathbf{H}}(e^{j\omega}) = \mathbf{H}(e^{j\omega}) + \mathbf{E}(e^{j\omega}).$$

Then the mean of the error term is zero, and its variance is given by,

$$\mathrm{var}\{E^{m}(e^{j\omega})\} = \frac{1}{BT}\frac{1}{SNR_{ave}^{m}(e^{j\omega})}\Xi(e^{j\omega}), \tag{9}$$

where, we defined the *Non-Stationarity Index* as

$$\Xi(e^{j\omega}) = \frac{<\Phi_{z_1z_1}(t,e^{j\omega})><1/\Phi_{z_1z_1}(t,e^{j\omega})>}{<\Phi_{z_1z_1}(t,e^{j\omega})><1/\Phi_{z_1z_1}(t,e^{j\omega})> - 1},$$

and the averaged signal to noise ratio as,

$$SNR_{ave}^{m}(e^{j\omega}) = \frac{<\Phi_{z_1z_1}(t,e^{j\omega})>}{\Phi_{u_mu_m}(t,e^{j\omega})}.$$

The symbol $\langle\cdot\rangle$ denotes frame averaging, $T$ is the total observation time and $B$ is the bandwidth of the window

$$\Phi^s_{oo}(t,e^{j\omega}) = \frac{|\mathcal{F}(e^{j\omega})|^2}{\|\widehat{\mathbf{H}}(e^{j\omega})\|^4} \times \Big\{ \widehat{\mathbf{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{SS}}(t,e^{j\omega})\widehat{\mathbf{H}}(e^{j\omega}) - \widehat{\mathbf{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\widehat{\mathcal{H}}(e^{j\omega}) \left(\widehat{\mathcal{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\widehat{\mathcal{H}}(e^{j\omega})\right)^{-1} -$$
$$\widehat{\mathcal{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{SS}}(t,e^{j\omega})\widehat{\mathbf{H}}(e^{j\omega})\widehat{\mathbf{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{SS}}(t,e^{j\omega})\widehat{\mathcal{H}}(e^{j\omega}) \left(\widehat{\mathcal{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\widehat{\mathcal{H}}(e^{j\omega})\right)^{-1} \widehat{\mathcal{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\widehat{\mathbf{H}}(e^{j\omega}) +$$
$$\widehat{\mathbf{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\widehat{\mathcal{H}}(e^{j\omega}) \left(\widehat{\mathcal{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\widehat{\mathcal{H}}(e^{j\omega})\right)^{-1} \widehat{\mathcal{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{SS}}(t,e^{j\omega})\widehat{\mathcal{H}}(e^{j\omega})$$
$$\times \left(\widehat{\mathcal{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\widehat{\mathcal{H}}(e^{j\omega})\right)^{-1} \widehat{\mathcal{H}}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\widehat{\mathbf{H}}(e^{j\omega}) \Big\} \quad (7)$$

used by the Blackman-Tukey PSD estimation procedure. $\Phi_{z_1z_1}(t,e^{j\omega})$ is the first microphone PSD. $\Phi_{u_mu_m}(t,e^{j\omega})$ is the $m$-th reference noise signal PSD. $1 < \Xi(e^{j\omega}) < \infty$, tends to infinity as the signal $z_1(t)$ is more stationary. According to Eqs. 6,9 the influence of the noise signal is twofold. High levels of noise reduce the averaged SNR level [i.e. reduce $SNR^m_{ave}(e^{j\omega})$] and increase the amount of $z_1(t)$ stationarity [i.e., increase $\Xi(e^{j\omega})$], thus causing an increase in the error variance. Neglecting second order effects we will analyze the distortion caused by errors in estimating ATFs recorded in a real room. The test scenario shown in Figure 3 is studied. The enclosure is a conference room with

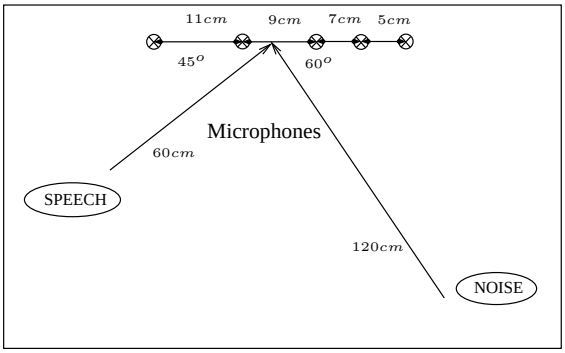


Figure 3: Test scenario: an array of five microphones in a noisy conference room.

dimensions $5m \times 4m \times 2.8m$. A linear array is placed on a table at the center of the room. Two loudspeakers are used. The left one for speech source and the right one for the noise source. The locations are marked in the Figure. Accurate ATFs estimates was obtained for each signal separately. These ATFs were used in Eqs. 8,9 to evaluate the predicted distortion. It is shown by the simulation, that even for input SNR as low as $-5$dB, the predicted distortion is no more than 6dB in the interesting frequency band. This result is with good agreement with the algorithm performance presented in [2]. Only weak dependency of the amount of distortion on the noise field was encountered.

### 3.2. Noise reduction

Starting again from the general expression in Eq. 7, substituting the input signal with the same noise signal used for calculating the Wiener filter and assuming perfect knowledge of the ATF ratios, gives the output noise PSD.

$$\Phi^n_{oo}(t,e^{j\omega}) = \Phi^n_{fbf}(t,e^{j\omega}) - \quad (10)$$
$$\frac{|\mathcal{F}(e^{j\omega})|^2}{\|\mathbf{H}(e^{j\omega})\|^4}\mathbf{H}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\mathcal{H}(e^{j\omega}) \times$$
$$\left(\mathcal{H}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\mathcal{H}(e^{j\omega})\right)^{-1} \mathcal{H}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\mathbf{H}(e^{j\omega}),$$

where, $\Phi^n_{fbf}(t,e^{j\omega})$ is given by,

$$\begin{aligned}\Phi^n_{fbf}(t,e^{j\omega}) &= E\{Y^n_{\mathrm{FBF}}(t,e^{j\omega})Y^{n*}_{\mathrm{FBF}}(t,e^{j\omega})\} \\ &= \frac{|\mathcal{F}(e^{j\omega})|^2}{\|\mathbf{H}(e^{j\omega})\|^4}\mathbf{H}^\dagger(e^{j\omega})\Phi_{\mathbf{NN}}(t,e^{j\omega})\mathbf{H}(e^{j\omega}).\end{aligned}$$

An interesting figure of merit is the extra noise reduction achieved by the noise cancelling branch (see also [4]),

$$\mathrm{NR}_{\mathrm{nc}}(t,e^{j\omega}) = \frac{\Phi^n_{fbf}(t,e^{j\omega})}{\Phi^n_{oo}(t,e^{j\omega})}. \quad (11)$$

#### *3.2.1. Dependency on Noise Field*

The resulting expression for the output noise PSD depends on the sensors noise PSD. Three important noise fields are addressed. Coherent (point source), diffused (spatially extended) and incoherent (noise signals generated at the sensors, e.g., amplifier noise, are assumed to be uncorrelated).

For a single point source noise signal with general ATFs $\mathbf{B}(e^{j\omega})$ the sensors noise spectral matrix is given by

$$\Phi_{\mathbf{NN}}(t,e^{j\omega}) = \Phi nn(t,e^{j\omega})\mathbf{B}(e^{j\omega})\mathbf{B}^\dagger(e^{j\omega}),$$

where, $\Phi nn(t,e^{j\omega})$ is the noise source PSD. The output noise PSD turns out to be $\Phi^n_{oo}(t,e^{j\omega}) = 0$, provided that $\mathbf{B}(e^{j\omega}) \neq \mathbf{A}(e^{j\omega})$, i.e. perfect noise cancellation is achieved (due to the noise cancelling branch). In the delay-only case this result is manifested by the deep notch at $\theta = 40^o$ in Figure 2. It was shown by Bitzer *et al.* [4] that even the classical Griffiths & Jim beamformer may achieve the same amount of noise reduction in the delay-only case. This property is generalized by the proposed method even for the more complicated general ATF case, as shown in Figure 4 (for the same test scenario shown in Figure 3). It is evident that a noise reduction of up to 70dB at the interesting frequency band (there is almost no signal in the lower frequencies ) can be achieved for this coherent noise field.

In highly reverberating acoustical environment, such as a car enclosure, the noise field tends to be diffused (e.g., see [4]), i.e. the cross-coherence function between signals received by two sensors $(i,j)$ at distance $d_{ij}$ is,

$$\Gamma_{Z_iZ_j}(e^{j\omega}) = \frac{\Phi_{Z_iZ_j}(e^{j\omega})}{\sqrt{\Phi_{Z_iZ_i}(e^{j\omega})\Phi_{Z_jZ_j}(e^{j\omega})}} = \frac{\sin(\omega d_{ij}/c)}{\omega d_{ij}/c},$$

where $c$ is the speed of sound. The noise PSD at the sensors input is thus,

$$\Phi_{\mathbf{NN}}(t,e^{j\omega}) = \Phi_{nn}(t,e^{j\omega})\Gamma(e^{j\omega}).$$

$\Gamma(e^{j\omega})$ is the coherence matrix, which components are given above. The amount of extra noise reduction achieved by the noise cancelling branch depends on the ATF ratios $\mathbf{H}(e^{j\omega})$

$$\mathrm{NR}_{\mathrm{nc}}(t, e^{j\omega}) = 1 \Bigg/ \left(1 - \frac{\mathbf{H}^{\dagger}(e^{j\omega})\Gamma(e^{j\omega})\mathcal{H}(e^{j\omega})\left(\mathcal{H}^{\dagger}(e^{j\omega})\Gamma(e^{j\omega})\mathcal{H}(e^{j\omega})\right)^{-1}\mathcal{H}^{\dagger}(e^{j\omega})\Gamma(e^{j\omega})\mathbf{H}(e^{j\omega})}{\mathbf{H}^{\dagger}(e^{j\omega})\Gamma(e^{j\omega})\mathbf{H}(e^{j\omega})}\right) \quad (12)$$

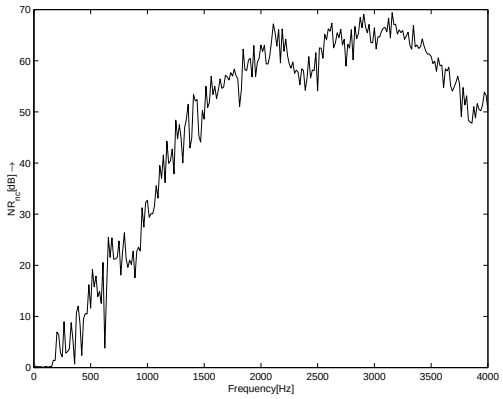


Figure 4: Expected noise reduction for both speech and noise general ATF case.

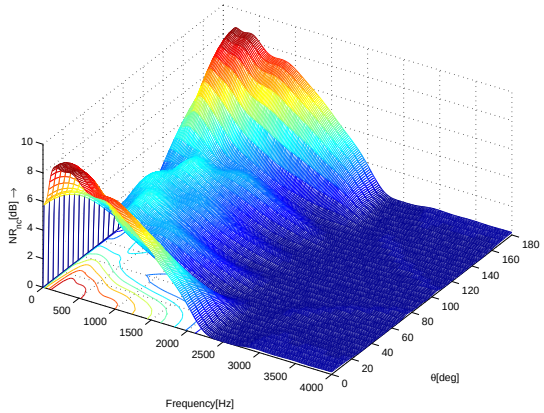


Figure 5: Linear array with $M = 5$ sensors for delay-only ATFs. Extra noise reduction of noise cancelling branch for diffused noise field.

and is given in Eq. 12 at the top of the page. This extra noise reduction, as a function of the DoA and the frequency, is given in Figure 5 for a delay-only desired signal ATFs in a diffused noise field. The amount of noise reduction achieved by the noise cancelling branch of the algorithm in the higher frequencies is shown to be almost zero. The general ATFs case is no better as shown in Figure 6. In the Incoherent Noise field no noise reduction is achieved by the noise cancelling branch.

## 4. CONCLUSIONS

While it is commonly known that the performance of the classical Griffiths & Jim GSC algorithm severely degrades in the general ATFs case, it was shown that the recently proposed TF-GSC algorithm still maintains its good performance in terms of both noise reduction and signal distortion

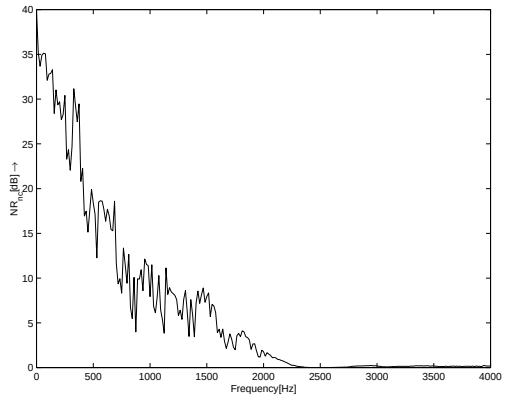


Figure 6: Expected noise reduction for general ATF speech signal and Diffused noise field.

figures of merit. The analytical performance evaluation presented supports the previously achieved experimental study.

## 5. REFERENCES

[1] L. J. Griffiths and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Trans. on Antennas and Propagation*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[2] S. Gannot, D. Burshtein and E. Weinstein, "Signal Enhancement Using Beamforming and Non-Stationarity with application to Speech," To appear in IEEE Trans. on Sig. Proc., Aug. 2001.

[3] S. Nordholm, I. Claesson and P. Eriksson, "The Broadband Wiener solution for Griffiths-Jim Beamformers," *IEEE trans. on Signal Proc.*, vol. 40, no. 2, pp. 474–478, Feb. 1992.

[4] J. Bitzer, K.U. Simmer and K.D. Kammeyer, "Theoretical Noise Reduction Limits of the Generalized Sidelobe Canceller (GSC) for Speech Enhancement," in *Int. Conf. on Acoustics, Speech and Signal Proc.*, Phoenix, Arizona, USA, May 1999.