

LIMITATIONS OF FIR MULTI-MICROPHONE SPEECH DEREVERBERATION IN THE LOW-DELAY CASE

Markus Hofbauer and Hans-Andrea Loeliger

Signal and Information Processing Laboratory
Swiss Federal Institute of Technology, Zürich, Switzerland

hofbauer@isi.ee.ethz.ch, loeliger@isi.ee.ethz.ch

ABSTRACT

In this paper multi-microphone dereverberation is considered under the constraint that no or little additional delay should be introduced by the FIR deconvolution filters. This is crucial for a number of applications such as hearing aids etc. Assuming that the acoustic impulse responses (AIRs) are known – e.g. by estimation, we determine the maximum degree of attainable dereverberation. Even though the AIRs are in general non-minimum phase, complete dereverberation can be accomplished in principle, using causal FIR filters of the same order as the AIRs, yielding no or only a little additional delay. We show that complete dereverberation with no or little delay will, however, reduce the SNR. For a given SNR gain and low delay, therefore, the achievable dereverberation is limited. We employ a time domain FIR multichannel Wiener filter with a delay constraint to find the MSE-sense optimal deconvolution filters. Dereverberation performance and SNR gain are demonstrated for typical AIRs with reverberation times of $T_{60} \approx 500\text{ms}$ and $N = 4000$ taps which have been measured in a conference room. Furthermore, we propose a new method utilizing a shaped desired total response, which is capable of selectively eliminating late reverberation while maintaining the SNR.

1. INTRODUCTION

Reverberation of speech in rooms with large reverberation time constants may decrease speech intelligibility and listening comfort. If the acoustic impulse response (AIR) from the source to the microphone is known, dereverberation can be performed by inverting the AIR. Due to the non-minimum phase nature of typical AIRs, an exact *single* channel inversion requires a noncausal – often slow decaying – IIR filter with a large delay [1]. However, it is well known that, using more than one microphone, a complete inversion can be performed by a set of causal FIR filters (with same order as the AIRs), which yield a zero total delay or any specified positive total delay [2]. Many applications do only accept a very small additional processing delay – in hearing aids, e.g. a maximum of about $t = 10\text{ms}$ is admissible. In this paper we determine the degree of dereverberation which can be expected, under the low delay constraint, for typical AIRs of medium sized rooms with reverberation times of $T_{60} \approx 500\text{ms}$ (using two microphones). We demonstrate that complete dereverberation can be performed in principle. However, the SNR with respect to ambient or coherent noise is decreased at the same time. For a given demanded SNR gain, therefore, the achievable dereverberation is limited.

There are several multi-microphone methods which accomplish some degree of dereverberation. In [3], beamforming is gen-

eralized to AIRs instead of simple propagation differences. Convolutional blind source separation techniques aim at blindly estimating the AIRs and to perform a separation and deconvolution of the sources [4],[5]. These methods – using filters mostly determined in the frequency domain – however generate non-causal filters which produce a large delay of about half of the AIR length. We assume the AIRs to be perfectly known (e.g. by estimation) and utilize a time domain FIR multichannel Wiener filter (WF), with a time domain delay constraint, to find the MSE-sense optimal deconvolution filters. The WF thus demonstrates the maximum degree of achievable dereverberation that can be expected.

To improve the low-delay dereverberation performance we propose a new method utilizing a shaped desired total response, which selectively eliminates late reverberation while maintaining the SNR.

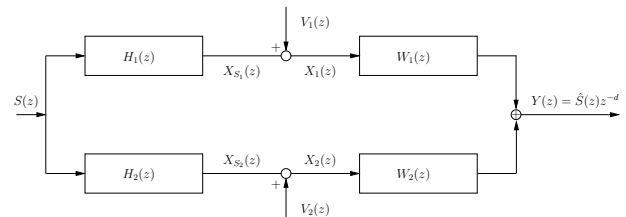


Fig. 1. Two channel deconvolution of AIRs h_i with total delay d .

2. PROBLEM FORMULATION

In a reverberant and noisy environment the i th sensor of an array with m microphones receives the signal $x_i[k]$

$$x_i[k] = (h_i * s) \Big|_k + v_i[k], \quad (1)$$

where $i = \{1, \dots, m\}$, $h_i = \{h_i[k]\}$ is the acoustic impulse response from speech source s and $v_i[k]$ is the noise at the i th sensor.

We aim at finding a set of m causal FIR filters $w_i = \{w_i[k]\}$ of the same order as h_i such that

$$y[k] = \hat{s}[k - d] = \sum_{i=1}^m (w_i * x_i) \Big|_k. \quad (2)$$

Perfect deconvolution of the source component s is therefore obtained if

$$\sum_{i=1}^m (h_i * w_i) \Big|_k = t[k] = \delta[k - d], \quad (3)$$

where $t[k]$ is the total response from the speech source to the output y and d the total delay. The two channel case ($m = 2$) is shown in Fig. 1. The minimum possible delay d_{\min} corresponds to the propagation time delay from the source s to the microphones. For $d = d_{\min}$ no additional delay Δd is introduced and filters w_i are causal, with their main power concentrated at $k = 0$.

For $m = 2$, Eq. (3) defines a system of $n_h + n_w + 1$ equations for the $2n_w + 2$ coefficients of w , where n_w and n_h are the filter orders of w and h , respectively. If the filters h_i, h_j do not share common zeros in the frequency domain, and if $n_w = n_h - 1$, there exists a unique solution of filters w [2]. In [6] it is shown that for most AIRs h , orders of $n_w \geq \lceil (n_h - 1)/(m - 1) \rceil$ are sufficient to solve (3), i.e. in the two channel case $n_w \geq n_h - 1$.

The existence of a set of causal FIR filters w which solve (3), even for non-minimum phase AIRs h , follows from a well known identity for polynomials (Bezout's Theorem) [7],

$$\sum_{i=1}^m H_i(z)W_i(z) = 1 \quad (4)$$

where $H_i(z)$ and $W_i(z)$ are the z -transforms of h_i and w_i , respectively.

The noise components v_i will be conveyed to the output as

$$y_v[k] = \sum_{i=1}^m (w_i * v_i) \Big|_k. \quad (5)$$

Thus, when calculating optimal filters w_i , which fulfill (3), we also wish to minimize the noise output power.

3. OPTIMAL FIR DECONVOLUTION FILTERS

3.1. Distortionless response filter

Solving (3) gives a filter set w such that

$$y[k] = s[k - d], \quad (6)$$

i.e. the source is perfectly deconvolved and appears at the output without any distortion. At the same time one wishes to minimize the noise variance at the output. If no constraint on the delay and causality of the filters w is required, the minimum variance distortionless response filter (MVDR) can be calculated in the frequency domain [8]

$$\mathbf{W}(\omega) = \frac{\mathbf{H}(\omega)^H \Phi_{\mathbf{v}\mathbf{v}}^{-1}(\omega)}{\mathbf{H}(\omega)^H \Phi_{\mathbf{v}\mathbf{v}}^{-1}(\omega) \mathbf{H}(\omega)}, \quad (7)$$

with $\mathbf{H}(\omega) = [H_1(\omega) H_2(\omega) \cdots H_m(\omega)]^T$, $H_i(\omega)$ the Fourier transform of h_i , and $\Phi_{\mathbf{v}\mathbf{v}}(\omega)$ the noise correlation matrix. The filter $\mathbf{W}(\omega)$ minimizes the noise variance while maintaining

$$\mathbf{W}(\omega)^T \mathbf{H}(\omega) = 1, \quad (8)$$

i.e. perfect deconvolution of the source. If v stems from uncorrelated sensor noise, $\mathbf{W}(\omega)$ is a matched filter. If the AIRs h_i are simple propagation delays, $\mathbf{W}(\omega)$ reduces to a delay-and-sum beamformer. Typical filters $w_i[k]$ obtained via (7) exhibit non-causal taps and thus produce large delays in the order of $d \approx n_w/2$.

If a delay constraint is required, (3) can be used to find the distortionless response filters w . In order to avoid the computational demanding direct solution of the possibly ill-conditioned

equation system (3), we alternatively define the frequency domain cost function

$$J(\mathbf{W}(\omega)) = \left| \sum_i W_i(\omega) H_i(\omega) - T(\omega) \right|^2, \quad (9)$$

where $T(\omega)$ is the Fourier transform of the desired total response $t[k]$. We solve (9) by a gradient method, yielding the following update rule for the i th filter $W_i(\omega)$:

$$W_i^{l+1}(\omega) = W_i^l(\omega) - \mu * \left(\sum_i W_i^l(\omega) H_i(\omega) - T(\omega) \right) H_i^*(\omega) \quad (10)$$

For the specified desired total response $t[k] = \delta[k - d]$, (10) converges to the distortionless response filter w_i , fulfilling (3).

3.2. Multichannel Wiener filter (WF)

As will be shown in section 4, the filters w_i which achieve perfect deconvolution will produce an SNR loss at the output. In order to determine filters w_i which optimize the degree of deconvolution and the output SNR with respect to the mean squared error (MSE), we utilize the time domain FIR multichannel Wiener filter.

We define the following stacked data and filter vectors:

$$\mathbf{x}_i[k] = [x_i[k] \ x_i[k-1] \ \cdots \ x_i[k-n_w]]^T \quad (11)$$

$$\mathbf{x}[k] = [\mathbf{x}_1^T \ \mathbf{x}_2^T \ \cdots \ \mathbf{x}_m^T]^T \quad (12)$$

$$\mathbf{w}_i = [w_i[0] \ w_i[1] \ \cdots \ w_i[n_w]]^T \quad (13)$$

$$\mathbf{w} = [\mathbf{w}_1^T \ \mathbf{w}_2^T \ \cdots \ \mathbf{w}_m^T]^T \quad (14)$$

Equation (2) then can be written as:

$$y[k] = \mathbf{w}^T \mathbf{x}[k] \quad (15)$$

The multichannel Wiener filter (with m inputs and one output) is given by

$$\mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{x}\mathbf{x}}^{-1} \mathbf{r}_{\mathbf{x}s}, \quad (16)$$

where $\mathbf{R}_{\mathbf{x}\mathbf{x}}$ is the autocorrelation matrix of the microphone signals

$$\mathbf{R}_{\mathbf{x}\mathbf{x}} = \mathbb{E}\{\mathbf{x}[k]\mathbf{x}[k]^T\} = \mathbf{R}_{\mathbf{x}s\mathbf{x}s} + \mathbf{R}_{\mathbf{v}\mathbf{v}} \quad (17)$$

and $\mathbf{r}_{\mathbf{x}s}$ the cross-correlation vector of \mathbf{x} and s

$$\mathbf{r}_{\mathbf{x}s} = \mathbb{E}\{\mathbf{x}[k]s[k-d]\}. \quad (18)$$

In (18) the allowed delay d is specified, which imposes a constraint on the Wiener filter. We now calculate $\mathbf{R}_{\mathbf{x}\mathbf{x}}$

$$\mathbf{R}_{\mathbf{x}\mathbf{x}} = \begin{bmatrix} \mathbf{R}_{\mathbf{x}_1\mathbf{x}_1} & \mathbf{R}_{\mathbf{x}_1\mathbf{x}_2} & \cdots \\ \mathbf{R}_{\mathbf{x}_2\mathbf{x}_1} & \mathbf{R}_{\mathbf{x}_2\mathbf{x}_2} & \\ \vdots & & \ddots \end{bmatrix} \quad (19)$$

and $\mathbf{r}_{\mathbf{x}s}$

$$\mathbf{r}_{\mathbf{x}s} = [\mathbf{r}_{\mathbf{x}_1 s}^T \ \cdots \ \mathbf{r}_{\mathbf{x}_m s}^T]^T \quad (20)$$

for a given set of AIRs h_i , a given autocorrelation $r_s[k]$ of s and a given noise autocorrelation matrix $\mathbf{R}_{\mathbf{v}\mathbf{v}}$. Element $(\mathbf{R}_{\mathbf{x}_i\mathbf{x}_j})_{a,b}$ of matrix $\mathbf{R}_{\mathbf{x}_i\mathbf{x}_j}$ then amounts to

$$\begin{aligned} (\mathbf{R}_{\mathbf{x}_i\mathbf{x}_j})_{a,b} &= r_{x_i x_j}[a-b] = \mathbb{E}\{x_i[k-a+1]x_j[k-b+1]\} \\ &= (h_i[-\cdot] * h_j[\cdot] * r_s[\cdot]) \Big|_{k'=a-b} + (\mathbf{R}_{\mathbf{v}_i\mathbf{v}_j})_{a,b} \end{aligned} \quad (21)$$

where $h_i[-]$ denotes time reversion. Finally, element $\mathbf{r}_{x_i s}(a)$ of vector $\mathbf{r}_{x_i s}$ is given by

$$\begin{aligned} \mathbf{r}_{x_i s}(a) &= r_{x_i s}[a-1-d] = \mathbb{E}\{x_i[k-a+1]s[k-d]\} \\ &= (h_i[-] * r_s[\cdot]) \Big|_{k'=a-1-d}. \end{aligned} \quad (22)$$

With (21) and (22) the Wiener filter (16) can be obtained.

3.2.1. Noise fields

We consider two types of noise sources: a coherent point noise source and a completely incoherent source with spatially uncorrelated noise components at the sensors (e.g. sensor noise). The autocorrelation matrix of white sensor noise is given by $\mathbf{R}_{vv} = \sigma_v^2 \mathbf{I}$, whereas for the coherent point noise source \mathbf{R}_{vv} is calculated the same way as for the point speech source by (21), replacing h with the corresponding noise source AIRs h^v , and s with v , respectively. An ambient noise source with an almost completely diffuse sound field and a sinc-type coherence function may be considered as a superposition of the two above sources (i.e. coherent at low, incoherent at higher frequencies).

3.3. SER gain and SNR gain

In order to measure the degree of dereverberation we define the ratio of signal to echo power SER. If we assume a white source signal s , the SER at sensor 1 is calculated from h_1 as:

$$\text{SER}_{x_1} = 10 \log_{10} \left(\frac{\max(|h_1[k]|^2)}{\sum_k |h_1[k]|^2 - \max(|h_1[k]|^2)} \right) \quad (23)$$

The SER at the output y we obtain from the total response $t[k]$:

$$\text{SER}_y = 10 \log_{10} \left(\frac{\max(|t[k]|^2)}{\sum_k |t[k]|^2 - \max(|t[k]|^2)} \right) \quad (24)$$

The reduction of reverberation is then specified by the SER gain:

$$\text{SER}_{\text{gain}} = \text{SER}_y - \text{SER}_{x_1} \quad (25)$$

Signal and noise powers p at sensor 1 and at the output y are given as

$$p_{x_1 s} = \sum_k |h_1[k]|^2 \sigma_s^2 \quad (26)$$

$$p_{x_1 v} = \sigma_n^2 \quad (27)$$

$$p_{y v} = \sum_{i,k} |w_i[k]|^2 \sigma_v^2 \quad (28)$$

$$p_{y s} = \sum_k |t[k]|^2 \sigma_s^2 \quad (29)$$

for the case of white uncorrelated sensor noise. The white noise SNR gain thus amounts to:

$$\begin{aligned} \text{SNR}_{\text{gain}_{\text{WN}}} &= \text{SNR}_y - \text{SNR}_{x_1} \\ &= 10 \log_{10} \left(\frac{\sum_k |t[k]|^2}{\sum_{i,k} |w_i[k]|^2 \cdot \sum_k |h_1[k]|^2} \right) \end{aligned} \quad (30)$$

In the case of a coherent point noise source the SNR gain is

$$\begin{aligned} \text{SNR}_{\text{gain}_{\text{PS}}} &= \text{SNR}_y - \text{SNR}_{x_1} \\ &= 10 \log_{10} \left(\frac{\sum_k |t[k]|^2 \cdot \sum_k |h_1^v[k]|^2}{\sum_k |t^v[k]|^2 \cdot \sum_k |h_1[k]|^2} \right) \end{aligned} \quad (31)$$

where $t^v[k] = \sum_{i=1}^m (h_i^v * w_i)|_k$ is the total response for the point noise source v .

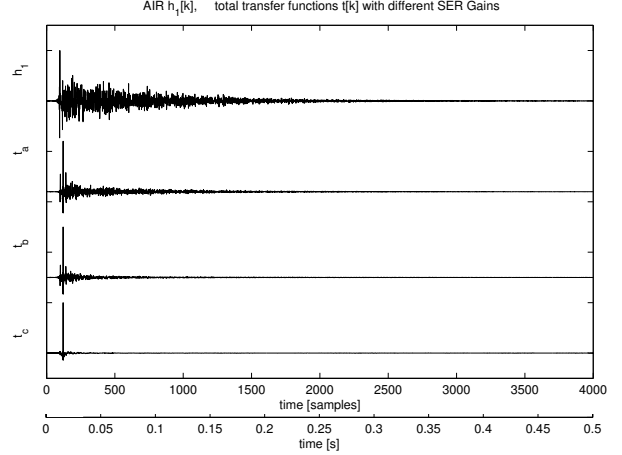


Fig. 2. Three examples of SER and SNR gains: AIR $h_1[k]$ of room2 and total responses t ; t_a $\text{SER}_{\text{gain}} = 8.4$ dB and $\text{SNR}_{\text{gain}} = 1.0$ dB; t_b $\text{SER}_{\text{gain}} = 12.9$ dB and $\text{SNR}_{\text{gain}} = -5.1$ dB; t_c $\text{SER}_{\text{gain}} = 21.7$ dB and $\text{SNR}_{\text{gain}} = -19.6$ dB.

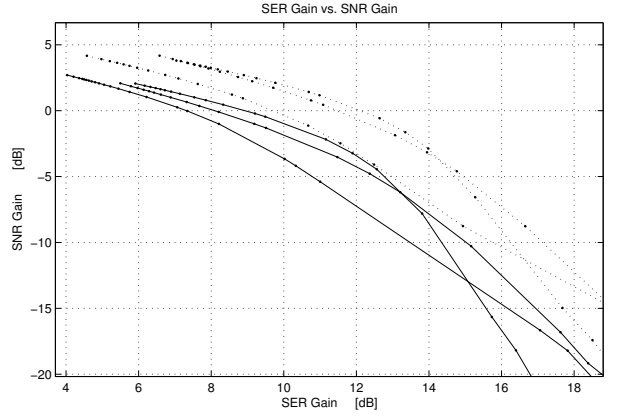


Fig. 3. SER gain vs. SNR gain of room1 (three sets) with allowed additional delays $\Delta d = 10$ samples (solid) and $\Delta d = 80$ (dotted).

4. ACHIEVABLE SER AND SNR GAINS FOR REAL ENVIRONMENTS

We have measured a series of sets of AIRs h_1 and h_2 (two microphones), with lengths of $n_h = 4000$ taps ($f_s = 8$ kHz), in a conference room (room1, 12m x 5m x 2.5m) with a reverberation time of $T_{60} \approx 500$ ms, and in an office room (room2, 5m x 3.5m x 2.5m, $T_{60} \approx 350$ ms). The speaker microphone distance was larger than 4m and the microphones were 14cm apart. Positions of speaker and microphones were changed for each set. Assuming white uncorrelated sensor noise and varying the input SNR_{x_1} , we then calculated the Wiener filter. The allowed additional delay was restricted to a few samples ($\Delta d < 80$). Fig. 2 shows the AIR h_1 and the resulting total responses t_a , t_b and t_c for three different SER gains. Case t_c indicates a high level dereverberation ($\text{SER}_{\text{gain}} = 21.7$ dB), which however yields a severe SNR loss ($\text{SNR}_{\text{gain}} = -19.6$ dB). Case t_a yields some dereverberation ($\text{SER}_{\text{gain}} = 8.4$ dB), while maintaining a positive SNR gain ($\text{SNR}_{\text{gain}} = 1.0$ dB).

Fig. 3 shows the SER- and SNR gains which can be expected for a typical room ($T_{60} \approx 500\text{ms}$) and a low delay constraint. If a positive SNR gain is desired, the maximum achievable SER gain is about 8 – 12 dB. Listening tests still reveal a significantly reduced reverberation for these SER gains. Perfect dereverberation can be obtained in principle with values $\text{SER}_{\text{gain}} = 50\text{ dB}$ and $\text{SNR}_{\text{gain}} = -30\text{ dB}$ (not depicted).

Fig. 4 illustrates the SER gain when the allowed delay is increased and a fixed SNR gain of 0 dB is demanded. Larger delays result in higher SER gains. In general, a coherent point noise source allows higher SER gains than uncorrelated sensor noise or an ambient noise source.

5. SHAPED DESIRED TOTAL RESPONSE TO SELECTIVELY ELIMINATE LATE REVERBERATION

Eq. (3) has a causal FIR solution for any desired total response $t_{\text{des}}[k]$ as long as $t_{\text{des}}[k] = 0$ for $k < d_{\text{min}}$ and $k > n_w + n_h + 1$. In order to selectively eliminate late reverberation, while maintaining positive SNR gains, we propose the usage of a shaped desired total response $t_{\text{des}}[k]$

$$t_{\text{des}}[k] = \text{env}[k] \cdot \sum_{i=1}^m h_i[k], \quad (32)$$

where $\text{env}[k]$ is an envelope to shape $t_{\text{des}}[k]$. We choose, for instance, an exponentially decaying envelope with time constant τ_{env} :

$$\text{env}[k] = \exp[-\tau_{\text{env}}k]. \quad (33)$$

A variety of envelope shapes are possible. We use $t_{\text{des}}[k]$ directly in (3) or in a slightly modified form of eq. (22), namely

$$\mathbf{r}_{\mathbf{x}_i \mathbf{s}}(a) = \left(h_i[-\cdot] * r_s[\cdot] * t_{\text{des}}[\cdot] \right) \Big|_{k'=a-1-d} \quad (34)$$

to calculate the Wiener filter. Fig. 5 shows the resulting total responses t_a and t_b of the Wiener filter for two envelope shapes. Late reverberation is clearly more reduced compared to the unshaped case t_c with the desired total response $t_{\text{des}}[k] = \delta[k - d]$. Note that in all cases the SNR gain is the same ($\text{SNR}_{\text{gain}} \approx 1\text{ dB}$). Listening tests confirm these results.

6. CONCLUSION

Given the AIRs from the speech source to the microphones, a complete dereverberation can be performed, in principle, using causal FIR filters of the same length as the AIRs. In this case the total response from speech source to the output is a delayed pulse $t[k] = \delta[t - d]$. However, for a small allowed delay d of a few samples, complete dereverberation will result in a severe SNR loss. If a positive SNR gain is required, the degree of achievable dereverberation is limited to SER gains of about 8 – 12 dB for medium sized rooms with reverberation times of $T_{60} \approx 500\text{ms}$ (2 mic. case). Listening tests still reveal a significantly reduced reverberation for these SER gains.

The deconvolving filters were calculated by a time domain multichannel Wiener filter, which is the optimal linear estimator with respect to the MSE cost function. An imposed delay constraint ensures the specified total delay d .

In order to selectively eliminate late reverberation, while maintaining positive SNR gains, we have proposed to use the given AIRs shaped by an envelope as desired total response, instead of a delayed dirac-pulse. This concept demonstrated an effective elimination of late reverberation with positive SNR gains, even for a low total delay.

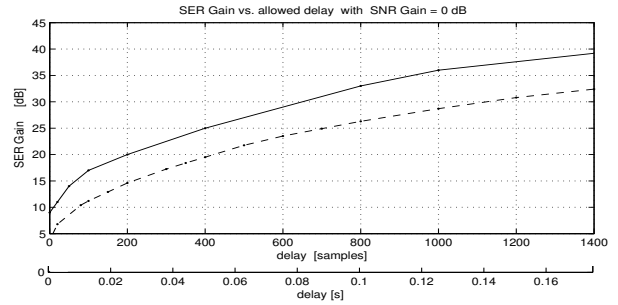


Fig. 4. SER gain vs. allowed delay d with fixed $\text{SNR}_{\text{gain}} = 0\text{ dB}$: uncorrelated sensor noise/ambient noise source (dashed); coherent point noise source (solid); AIRs were simulated with $n_h = 4000$.

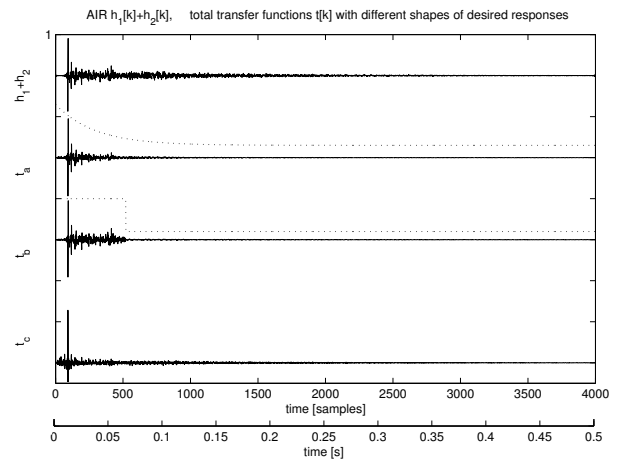


Fig. 5. Late reverberation cancellation: h_1+h_2 of room1; t_a : WF total response utilizing exponential window shaped desired total response t_{des} ; t_b : step window shaped t_{des} ; t_c no shaping ($t_{\text{des}}[k] = \delta[k - d]$). Note that $\text{SNR}_{\text{gain}} \approx 1\text{ dB}$ for t_a , t_b and t_c .

7. ACKNOWLEDGMENTS

The authors would like to thank Marcel Joho for helpful discussions and comments. This work was supported by CTI.

8. REFERENCES

- [1] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, 1979.
- [2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE T. Acoust. Speech and SP.*, 1988.
- [3] S. Gannot, D. Burshtein, and E. Weinstein, "Theoretical analysis of the general transfer function GSC," *IWAENC*, 2001.
- [4] K. Rahbar, J.P. Reilly, and J. H. Manton, "Blind identification of MIMO FIR systems driven by quasi-stationary sources using second order statistics: A frequency domain approach," *IEEE T. Signal Processing*, 2002.
- [5] M. Guerelli and C. L. Nikias, "Evam: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," *IEEE T. Signal Processing*, 1995.
- [6] G. Harikumar and Y. Bresler, "FIR perfect signal reconstruction from multiple convolutions: Minimum deconvolver orders," *IEEE T. Signal Processing*, 1998.
- [7] C. A. Berenstein and A. Yger, *Residue Currents and Bezout Identities*, Birkhaeuser, Switzerland, 1993.
- [8] M. Brandstein and D. Ward, *Microphone Arrays*, Springer, 2001.