

ROBUSTNESS OF SVD-BASED OPTIMAL FILTERING FOR NOISE REDUCTION IN MULTI-MICROPHONE SPEECH SIGNALS

Simon Doclo, Marc Moonen

K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA,
Kardinaal Mercierlaan 94, 3001 Heverlee, Belgium
{simon.doclo, marc.moonen}@esat.kuleuven.ac.be

ABSTRACT

This paper discusses an SVD-based signal enhancement procedure, applied to noise reduction in multi-microphone speech signals. The SVD-based signal enhancement procedure amounts to a specific optimal filtering problem when the so-called ‘desired response’ signal cannot be observed. The optimal filter can then be written as a function of the generalized singular vectors and singular values of a speech and noise data matrix.

It is shown that the SNR improvement provided by the SVD-based optimal filtering technique is better than the improvement obtained with standard beamforming techniques. Moreover most beamforming techniques assume the position of the speech source and the microphone array configuration to be known. Therefore the performance of these techniques is rather sensitive to deviations from these assumptions, *i.e.* incorrect estimation of the position of the speech source and uncalibrated microphone arrays. It is shown that the SVD-based optimal filtering technique is more robust to such deviations than standard beamforming techniques.

1. INTRODUCTION

In many speech communication applications, like hands-free mobile telephony and audio-conferencing, the recorded speech signals contain a considerable amount of acoustic noise. This is mainly due to the fact that the speaker is located at a certain distance from the microphones, which allows the microphones to record the noise sources too. Background noise causes a signal degradation which can lead to total unintelligibility of the speech.

Some techniques for noise reduction in speech have been proposed which are based on the singular value decomposition (SVD) [1][2]. Most of these techniques deal with the single microphone case and rely only on signal specific characteristics.

When using a microphone array, the spatial configuration of the speech and noise sources and the microphone array itself constitute an important aspect which should be incorporated into the

Simon Doclo is a Research Assistant supported by the I.W.T. (Flemish Institute for Scientific and Technological Research in Industry). Marc Moonen is a Research Associate with the F.W.O.-Vlaanderen (Fund for Scientific Research-Flanders). This research work was carried out in the frame of the Concerted Research Action *Model-based Information Processing Systems* (GOA/MIPS/95/99/3) of the Flemish Government, the F.W.O. Research Project nr. G.0295.97, *Design and implementation of adaptive digital signal processing algorithms for broadband applications* and the IT-project *Multimicrophone Signal Enhancement Techniques for handsfree telephony and voice controlled systems* (MUSETTE) of the I.W.T., and was partially sponsored by Philips-ITCL.

signal enhancement procedure. Therefore multi-microphone algorithms should exploit the characteristics of the channel between the sources and the microphone array. The SVD-based multi-microphone extensions proposed so far [3] do not exploit these channel characteristics.

Here, an optimal filtering approach is applied for the derivation of the SVD-based procedure. In section 2 the SVD-based optimal filtering technique is described and a number of simple symmetry properties of the optimal filter are derived. When the optimal filtering technique is applied to noise reduction in multi-microphone speech signals, it is shown that for simple scenarios this technique exhibits some kind of beamforming behavior. Section 3 compares the SNR improvement of this technique with standard beamforming techniques, showing improved performance. Section 4 finally compares the robustness of the SVD-based optimal filtering technique and standard beamforming techniques. Robustness is compared for incorrect estimation of the position of the speech source and for deviations in the configuration and characteristics of the microphone array.

2. SVD-BASED OPTIMAL FILTERING

2.1. Preliminaries

Consider the following optimal filtering problem (figure 1): $\mathbf{u}_k \in \mathbb{R}^N$ is the filter input vector at time k , \mathbf{y}_k is the filter output, $\mathbf{y}_k = \mathbf{u}_k^T \mathbf{W}$, with $\mathbf{W} \in \mathbb{R}^{N \times N}$ the optimal filter. The vector $\mathbf{d}_k \in \mathbb{R}^N$ is the desired response vector and $\mathbf{e}_k = \mathbf{d}_k - \mathbf{y}_k$ is the error vector. The MSE (mean square error) cost function for optimal filtering is

$$\mathbf{J}_{MSE}(\mathbf{W}) = \mathcal{E}\{\|\mathbf{e}_k\|_2^2\} = \mathcal{E}\{\mathbf{d}_k^T \mathbf{d}_k\} - 2\mathcal{E}\{\mathbf{u}_k^T \mathbf{W} \mathbf{d}_k\} + \mathcal{E}\{\mathbf{u}_k^T \mathbf{W} \mathbf{W}^T \mathbf{u}_k\}. \quad (1)$$

The optimal filter \mathbf{W}_{WF} is the well-known Wiener filter,

$$\mathbf{W}_{WF} = \mathcal{E}\{\mathbf{u}_k \cdot \mathbf{u}_k^T\}^{-1} \cdot \mathcal{E}\{\mathbf{u}_k \cdot \mathbf{d}_k^T\}. \quad (2)$$

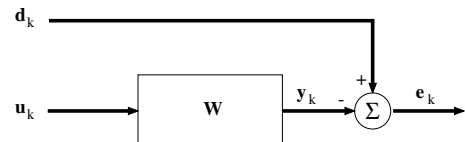


Figure 1: Optimal filtering problem with desired response \mathbf{d}_k

In the following, we consider problems where only observations of \mathbf{u}_k are available, and the observed signal \mathbf{u}_k contains a signal-of-interest \mathbf{s}_k (e.g. a speech signal) plus additive noise \mathbf{n}_k ,

$$\mathbf{u}_k = \mathbf{s}_k + \mathbf{n}_k. \quad (3)$$

If we consider speech applications and use a robust speech/noise detection algorithm [4], noise-only observations can be made during speech pauses (time k'), $\mathbf{u}_{k'} = \mathbf{n}_{k'}$, which allows to estimate the spatial and temporal color of the noise. Our goal is to reconstruct the signal-of-interest \mathbf{s}_k from \mathbf{u}_k by means of a linear filter \mathbf{W} . In the optimal filtering context this means that the desired signal is in fact equal to the signal-of-interest, $\mathbf{d}_k = \mathbf{s}_k$, but that now the desired signal \mathbf{d}_k is an unobservable signal.

We make two assumptions: short-term stationarity of the noise, $\mathcal{E}\{\mathbf{n}_k \cdot \mathbf{n}_k^T\} = \mathcal{E}\{\mathbf{n}_{k'} \cdot \mathbf{n}_{k'}^T\}$, and statistical independence of the speech and noise signals, $\mathcal{E}\{\mathbf{s}_k \cdot \mathbf{n}_k^T\} = \mathbf{0}$. Using these assumptions the optimal filter \mathbf{W}_{WF} becomes

$$\mathbf{W}_{WF} = \mathcal{E}\{\mathbf{u}_k \cdot \mathbf{u}_k^T\}^{-1} \left(\mathcal{E}\{\mathbf{u}_k \cdot \mathbf{u}_k^T\} - \mathcal{E}\{\mathbf{n}_k \cdot \mathbf{n}_k^T\} \right). \quad (4)$$

An interesting simplification is derived from the joint diagonalization [5] of the symmetric correlation matrices $\mathcal{E}\{\mathbf{u}_k \cdot \mathbf{u}_k^T\}$ and $\mathcal{E}\{\mathbf{n}_k \cdot \mathbf{n}_k^T\}$,

$$\begin{cases} \mathcal{E}\{\mathbf{u}_k \cdot \mathbf{u}_k^T\} &= X \cdot \text{diag}\{\sigma_i^2\} \cdot X^T \\ \mathcal{E}\{\mathbf{n}_k \cdot \mathbf{n}_k^T\} &= X \cdot \text{diag}\{\eta_i^2\} \cdot X^T \end{cases} \quad (5)$$

In practice, X , σ_i^2 and η_i^2 are estimated by means of a generalized singular value decomposition (GSVD) of the speech data matrix $\mathbf{U}_k \in \mathbb{R}^{p \times N}$ and the noise data matrix $\mathbf{N}_k \in \mathbb{R}^{q \times N}$ (with p and q typically larger than N),

$$\mathbf{U}_k = \begin{bmatrix} \mathbf{u}_k^T \\ \mathbf{u}_{k+1}^T \\ \vdots \\ \mathbf{u}_{k+p-1}^T \end{bmatrix} \quad \mathbf{N}_k = \begin{bmatrix} \mathbf{n}_k^T \\ \mathbf{n}_{k+1}^T \\ \vdots \\ \mathbf{n}_{k+q-1}^T \end{bmatrix} \quad (6)$$

such that $\mathcal{E}\{\mathbf{u}_k \cdot \mathbf{u}_k^T\} \simeq \mathbf{U}_k^T \cdot \mathbf{U}_k$ and $\mathcal{E}\{\mathbf{n}_k \cdot \mathbf{n}_k^T\} \simeq \mathbf{N}_k^T \cdot \mathbf{N}_k$. The GSVD of the matrices \mathbf{U}_k and \mathbf{N}_k is defined as

$$\begin{cases} \mathbf{U}_k &= U \cdot \text{diag}\{\tilde{\sigma}_i\} \cdot \tilde{X}^T \\ \mathbf{N}_k &= V \cdot \text{diag}\{\tilde{\eta}_i\} \cdot \tilde{X}^T, \end{cases} \quad (7)$$

with U and V orthogonal matrices, \tilde{X} an invertible (but not necessarily orthogonal) matrix and $\frac{\tilde{\sigma}_i}{\tilde{\eta}_i}$ the generalized singular values. By substituting these formulas into (4), one obtains

$$\mathbf{W}_{WF} = \tilde{X}^{-T} \cdot \text{diag}\left\{\frac{\tilde{\sigma}_i^2 - \tilde{\eta}_i^2}{\tilde{\sigma}_i^2}\right\} \cdot \tilde{X}^T. \quad (8)$$

In fact, the filter \mathbf{W}_{WF} belongs to a more general class of estimators, which can be described by

$$\mathbf{W} = \tilde{X}^{-T} \cdot \text{diag}\{f(\tilde{\sigma}_i^2, \tilde{\eta}_i^2)\} \cdot \tilde{X}^T. \quad (9)$$

The estimation error \mathbf{e}_k is defined as $\mathbf{e}_k = \mathbf{s}_k - \mathbf{y}_k = \mathbf{s}_k - \mathbf{W}_{WF}^T \mathbf{u}_k$, such that error covariance matrix can be written as

$$\mathcal{E}\{\mathbf{e}_k \cdot \mathbf{e}_k^T\} = \mathcal{E}\{\mathbf{n}_k \cdot \mathbf{n}_k^T\} \cdot \mathbf{W}_{WF}. \quad (10)$$

In particular, we are interested in the diagonal elements of the matrix $\{\mathcal{E}\{\mathbf{n}_k \cdot \mathbf{n}_k^T\} \cdot \mathbf{W}_{WF}\}_{ii}$, since these elements indicate how well $\{\mathbf{s}_k\}_i$ (the i^{th} component of \mathbf{s}_k) is estimated in the time series filtering context (see section 2.2). The smallest element on the main diagonal of the error covariance matrix corresponds to the best estimator. The best estimator, which is the corresponding column of \mathbf{W}_{WF} , will be denoted as $\mathbf{w}_{WF}^{min} \in \mathbb{R}^N$, and the corresponding column index will be denoted by i_{min} .

2.2. Time series filtering

When applying the optimal filtering technique to single microphone noise reduction, the vector \mathbf{u}_k is taken from a time series $u(k)$, i.e.

$$\mathbf{u}_k = [u(k) \quad u(k-1) \quad u(k-2) \quad \dots \quad u(k-N+1)]^T. \quad (11)$$

The vectors \mathbf{s}_k and \mathbf{n}_k are similarly defined. The data matrices \mathbf{U}_k and \mathbf{N}_k , as defined in equation (6), now are Toeplitz matrices and the correlation matrices $\mathcal{E}\{\mathbf{u}_k \cdot \mathbf{u}_k^T\}$ and $\mathcal{E}\{\mathbf{s}_k \cdot \mathbf{s}_k^T\}$ are symmetric Toeplitz matrices. Symmetric Toeplitz matrices belong to the class of double symmetric matrices, which are symmetric about both the main diagonal and the secondary diagonal. The eigenvectors of such matrices are known to have special symmetry properties [6]. Using these properties, one can prove the following symmetry property for the optimal filter \mathbf{W}_{WF} [7].

Theorem 1 *If \mathbf{W}_{WF} is constructed according to equation (9), then \mathbf{W}_{WF} satisfies the symmetry properties*

$$\mathbf{W}_{WF} = J \cdot \mathbf{W}_{WF} \cdot J \quad (12)$$

$$\mathbf{W}_{WF}^T = J \cdot \mathbf{W}_{WF}^T \cdot J, \quad (13)$$

with J the reverse identity matrix. These properties hold in the white noise case as well as in the colored noise case.

2.3. Multi-channel filtering and beamforming behavior

Consider M microphones where each microphone signal $m_j(k)$, $j = 1 \dots M$, consists of a filtered version of the desired signal $s(k)$ and an additive noise term, $m_j(k) = h_j(k) \otimes s(k) + n_j(k)$. The vector $\mathbf{u}_k \in \mathbb{R}^{MN}$ now takes the form

$$\mathbf{u}_k = [\mathbf{m}_{1k}^T \quad \mathbf{m}_{2k}^T \quad \dots \quad \mathbf{m}_{Mk}^T]^T, \quad (14)$$

$$\mathbf{m}_{jk} = [m_j(k) \quad m_j(k-1) \quad \dots \quad m_j(k-N+1)]^T. \quad (15)$$

Using the same formulas as for the single channel case, the optimal filter \mathbf{W}_{WF} and the best (MN) -taps estimator \mathbf{w}_{WF}^{min} can be computed. The enhanced signal $\hat{s}(k)$ can be computed as

$$\hat{s}(k) = [\hat{s}(k) \quad \hat{s}(k+1) \quad \dots \quad \hat{s}(k+p-1)]^T = \mathbf{U}_k \cdot \mathbf{w}_{WF}^{min}, \quad (16)$$

where $\hat{s}(k)$ is an estimate for $h_{i_{min}}(k) \otimes s(k)$. This operation can be considered as multi-channel filtering (see figure 2), where each of the M channels is filtered with an N -taps filter A_j , with $\mathbf{w}_{WF}^{min} = [A_1^T \quad A_2^T \quad \dots \quad A_M^T]^T$.

In [7] the frequency and spatial filtering properties of the SVD-based optimal filtering technique are discussed, showing that this technique exhibits the desired beamforming behavior when we consider localized sources and no multi-path propagation.

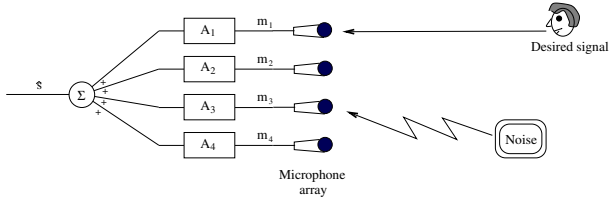


Figure 2: Multi-channel filtering

3. COMPARISON WITH STANDARD BEAMFORMING

In this section we compare the performance (SNR of the enhanced signal $\hat{s}(k)$) of the SVD-based optimal filtering technique and standard beamforming techniques (delay-and-sum beamforming and Generalized Sidelobe Canceler (GSC)) [8]. The simulated room configuration is depicted in figure 3, with dimensions $7m \times 3.5m \times 3m$. It consists of a microphone array, a speech source and a noise source. The linear equi-spaced microphone array has 5 microphones and the distance between two adjacent microphones is $5cm$. The speech source is located in front of the microphone array. The signals used are an 8 kHz clean speech signal and a temporally white noise source.

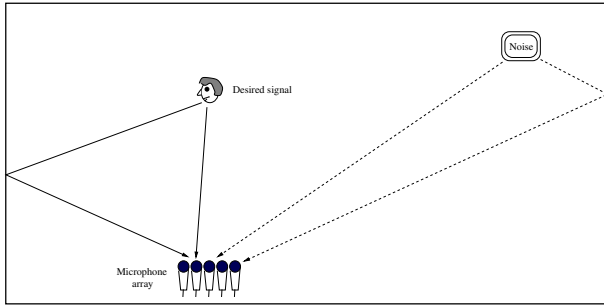


Figure 3: Room configuration

The comparison in SNR improvement is done for different reverberation times T_{60} of the room. Low reverberation times correspond to highly correlated noise, while high reverberation times correspond to highly uncorrelated (diffuse) noise. The reverberation time T_{60} can be expressed as a function of the reflection coefficient α of the walls, according to Eyring's formula,

$$T_{60} = \frac{0.163V}{-S \log(1 - \alpha)}, \quad (17)$$

with V the volume of the room and S the total surface of the room. The reflection coefficient is a necessary parameter for calculating the impulse response through the image method described in [9]. Figure 4 compares the performance of the delay-and-sum beamformer and the GSC-beamformer with the SVD-based optimal filtering technique (filterlength $N = 10, 20, 50$). As can be seen, for small T_{60} the GSC-beamformer performs much better than for high T_{60} . This is obvious because the GSC-beamformer is designed for correlated noise, not for diffuse noise. Unlike the GSC-beamformer, the SVD-based optimal filtering technique still performs well for high T_{60} . As can be seen, for all reverberation times, the SVD-based optimal filtering technique performs better than the GSC-beamformer, if the filterlength N is large enough.

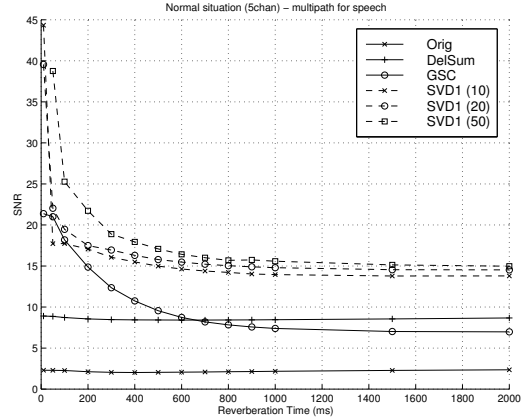


Figure 4: SNR of noisy signal $m_1(k)$ and SNR of enhanced signal $\hat{s}(k)$ for delay-and-sum beamformer, GSC beamformer and SVD-based optimal filtering technique ($N = 10, 20, 50$)

4. ROBUSTNESS ISSUES

Using the same room configuration, we compare the sensitivity of the optimal filtering technique and the GSC-beamformer for deviations from the nominal situation. In this case the nominal situation can be described as follows :

- speech source located in front of microphone array
- linear equi-spaced microphone array
- all microphones have the same characteristics (gain, spatial directivity, frequency behavior, ...)

The GSC-beamformer starts from the a priori assumption that the position of the speech source and the microphone array configuration are known. The SVD-based optimal filtering technique does not make any assumptions of this kind. Therefore we can already expect the GSC-beamformer to be more sensitive to deviations from the nominal situation. We will compare the robustness for three kinds of deviations :

1. incorrect estimation of the position of the speech source (we assume that the speech source is located at $\theta^{nom} = 90^\circ$, while in fact it is located at a different angle θ)
2. microphone displacement (we assume a linear equi-spaced microphone array, while in fact the second microphone is not at its nominal position $x_{mic2}^{nom} = 2.05$)
3. microphone gain (we assume that all microphones have the same gain $\gamma^{nom} = 1$, while in fact the second microphone has a different gain γ)

4.1. Incorrect estimation of the position of the speech source

Figure 5 shows the difference in performance between the SVD-based optimal filtering technique and the GSC-beamformer for different angles θ . The SVD-based optimal filtering technique is more robust than the GSC-beamformer if the difference in performance increases the more the actual situation deviates from the nominal situation. In figure 5 it can be observed that the difference in performance increases the more the angle θ deviates from the nominal angle $\theta^{nom} = 90^\circ$.

4.2. Microphone displacement

Figure 6 shows the difference in performance between the SVD-based optimal filtering technique and the GSC-beamformer for dif-

ferent microphone positions x_{mic2} . Because the difference in performance increases the more the microphone position x_{mic2} deviates from the nominal position $x_{mic2}^{nom} = 2.05$, we can conclude that for microphone displacement, the SVD-based optimal filtering technique is more robust than the GSC-beamformer.

4.3. Microphone gain

Figure 7 shows the difference in performance between the SVD-based optimal filtering technique and the GSC-beamformer for a different gain γ of the second microphone. Because the difference in performance increases the more the gain γ deviates from the nominal gain $\gamma^{nom} = 1$, we can conclude that for this kind of deviation, the SVD-based optimal filtering technique is more robust than the GSC-beamformer. Theoretically it can be proven that the SVD-based optimal filtering technique is in fact independent of different gains for the different microphones.

5. CONCLUSION

In this paper we have shown that the described SVD-based optimal filtering technique outperforms the standard GSC-beamformer. For all reverberation times the SVD-based optimal filtering technique results in a larger SNR improvement than the GSC-beamformer. Since in practice a combination of all three described deviations from the nominal situation occur, we can expect the SVD-based filtering technique to be considerably less sensitive than the GSC-beamformer.

6. REFERENCES

- [1] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of Broad-Band Noise in Speech by Truncated QSVD," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 6, pp. 439–448, Nov. 1995.
- [2] P. S. K. Hansen, *Signal Subspace Methods for Speech Enhancement*, Ph.D. thesis, Technical University of Denmark, Lyngby, Denmark, Sept. 1997.
- [3] I. Dologlou, J.-C. Pesquet, and J. Skowronski, "Projection-based rank reduction algorithms for multichannel modelling and image compression," *Signal Processing*, vol. 48, no. 2, pp. 97–109, Jan. 1996.
- [4] F. Xie and S. Van Gerven, "Comparative study of 3 speech detection methods," Tech. Rep. MI2-SPCH-95-8, ESAT, K.U.Leuven, Belgium, Oct. 1995.
- [5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, MD : John Hopkins University Press, Baltimore, 3rd edition, 1996.
- [6] P. Butler and A. Cantoni, "Eigenvalues and eigenvectors of symmetric centrosymmetric matrices," *Linear Algebra and its Applications*, vol. 13, pp. 275–288, Mar. 1976.
- [7] S. Doclo and M. Moonen, "SVD-based optimal filtering with applications to noise reduction in speech signals," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'99)*, New Paltz, New York, USA, Oct. 1999, pp. 143–146.
- [8] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, pp. 4–24, Apr. 1988.
- [9] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.

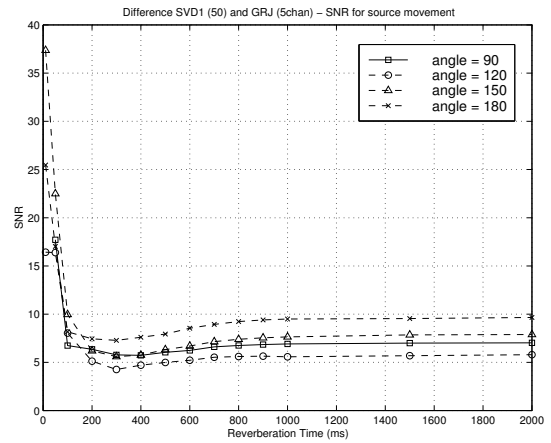


Figure 5: SNR-difference between SVD-based optimal filtering and GSC-beamformer for different angles θ of the speech source

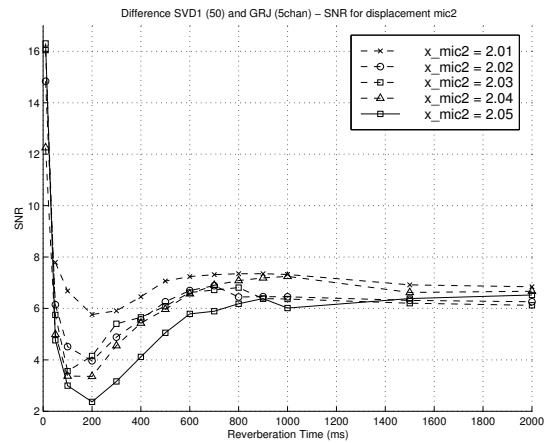


Figure 6: SNR-difference between SVD-based optimal filtering and GSC-beamformer for different microphone positions x_{mic2}

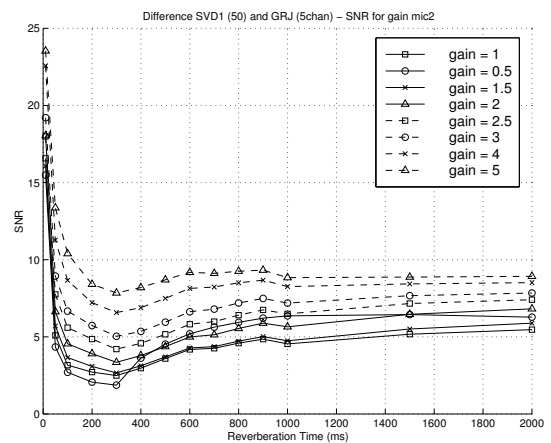


Figure 7: SNR-difference between SVD-based optimal filtering and GSC-beamformer for different gains γ