ROBUSTNESS OF PARAMETRIC SOURCE DEMIXING IN ECHOIC ENVIRONMENTS

Radu Balan, Justinian Rosca, Scott Rickard

Siemens Corporate Research
Princeton, NJ 08540
{radu.balan,justinian.rosca,scott.rickard}@scr.siemens.com

ABSTRACT

Blind source separation (BSS) of audio signals in echoic environments such as an office room is still a very challenging problem. Here we approach the problem from a practical perspective and shed light on how robust a two channel echoic parametric demixing can get. We assume that an oracle (i.e. a perfect estimator) provides a truncated estimate of the mixing FIR filters for a given source configuration. This way we can study the properties of a parametric demixer using the adjoint of the truncated mixing matrix. For several degrees of truncation, we compute how the separation SNR varies as a function of the uncertainty of the true source position. The true source position is uniformly distributed within a sphere of radius R around an assumed position, to reflect the fact that parameters of interest are imprecisely estimated. Simulations of artificial echoic mixings show that the higher order demixing filters have little robustness to position uncertainties (and therefore to errors of estimation) while the overall performance remains almost constant beyond the second order approximation. This should represent a guideline for what is practically achievable with a class of BSS techniques in echoic environments.

1. INTRODUCTION

The Blind Source Separation and Independent Component Analysis (ICA) problem has been under increased attention in recent years. Two international conferences (Aussois 1999, Helsinki 2000) have been dedicated to these topics, while many other signal processing related conferences have presented relevant research. Although a number of successful applications in image and medical signal processing have been presented, BSS techniques have proved only modest gains for audio signal processing [1, 2]. This may not come as a surprise to the array processing community, where results of signal enhancement techniques are modest in the case of a small number of sensors [3].

Several BSS methods are used to separate voices from acoustic mixtures, which we divide into two classes. The first class uses parametric mixing models and thus it reduces the number of degrees of freedom of the identification problem, whereas the second class uses a full non-parametric (or at least, not explicitly parametric) demixing scheme rather than exploit the relative sparseness of the mixing model. We call the former class *parametric BSS* and the latter *nonparametric BSS*. Parametric BSS solutions have first been studied in the context of anechoic mixtures [4, 5]. In such cases, only four parameters are needed: two delays and two attenuations. Moreover, if microphones sensors are close enough, the attenuations can be approximated to be unity, and only delay parameters are used. For echoic environments, the simple direct-path model can be used as a starting point for a more complex mixing (or demixing) model [6].

Nonparametric mixing models are implemented either in time-domain or frequency domain. Time-domain approaches considers long FIR or IIR filters and adapt the filter coefficients to obtain independent outputs [7, 8]. Frequency domain approaches use of a simple but useful observation: at each frequency a convolutive mixing becomes a simple multiplicative mixing. There is a caveat to this statement: the window size to perform FFT has to be sufficiently large compared to the room reverberation (see [9] for an analysis of the simple delay operator). This remark implies the need for long filters. Additionally, the permutation problem has to be solved. Several approaches have been proposed. They all use an ICA method to demix on each frequency independently from one another and then use some criterion to find the right permutation matrix [10, 11, 12, 13].

In this paper we approach the problem of echoic demixing from a practical perspective and analyze the robustness of the two channel echoic parametric demixing problem. We perform demixing using FIR filters truncated to various degrees of precision. For several degrees of truncation, we compute how the separation quality, measured in terms of the instantaneous SNR, varies as a function of the uncertainty of the estimated source position.

In the next section we discuss the parametric mixing and demixing models used in the present experiments. Section 3 defines the setup and robustness measures used. Section 4 presents experimental results. Section 5 concludes on what is practically feasible in echoic environments.

2. PARAMETRIC MIXING MODEL

Assume a parametric mixing model with two sources and two microphones of the form:

$$x_1(t) = \sum_{k=0}^{L} a_{11}^k s_1(t - \tau_{11}^k) + a_{12}^k s_2(t - \tau_{12}^k) \quad (1)$$

$$x_2(t) = \sum_{k=0}^{L} a_{21}^k s_1(t - \tau_{21}^k) + a_{22}^k s_2(t - \tau_{22}^k) \quad (2)$$

where L is the number of echoic path the model uses. The choice of L should take into account the room reverberation time. Furthermore, $s_1(\cdot)$, $s_2(\cdot)$ are the source signals, $x_1(\cdot), x_2(\cdot)$ are the measured signals, a_{ij}^k is the k^{th} path attenuation coefficient from source j to microphone i, and τ_{ij}^k the corresponding delay. All the time variables are measured in samples. We assume the sampling frequency is sufficiently high and the distance between microphones and between sources is sufficiently large in order to induce integer delays. Let us denote by A the 2×2 matrix of mixing filter transfer functions:

$$A(z) = \begin{bmatrix} A_{11}(z) & A_{12}(z) \\ A_{21}(z) & A_{22}(z) \end{bmatrix} \quad (3)$$

$$A_{ij}(z) = \sum_{k=0}^{L} a_{ij}^k z^{-\tau_{ij}^k} \quad (4)$$

Techniques to extract the sources s_1 and s_2 from the mixtures x_1, x_2 range from methods that aim toward source separation and use either the inverse of the mixing matrix or its adjoint [4, 5], to techniques that aim mostly for signal enhancement such as the multiple delay-and-sum beamformer (when only information of arrival times is required) or matching filters (when the full mixing matrix is used) [14]. In this paper we discuss the use of the adjoint matrix as a demixing solution.

Let $W = adj(A)$, the adjoint of A, defined by:

$$adj(A) = \begin{bmatrix} A_{22}(z) & -A_{12}(z) \\ -A_{21}(z) & A_{11}(z) \end{bmatrix}$$

When applied on $(x_1(\cdot), x_2(\cdot))$, the outputs are:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = W \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (5)$$

$$\begin{array}{c} u_1(t) = \sum_{k=0}^{L} a_{22}^k x_1(t-n) - a_{12}^k x_2(t-n) \\ u_2(t) = \sum_{k=0}^{L} -a_{21}^k x_1(t-n) + a_{11}^k x_2(t-n) \end{array} \quad (6)$$

and combined with (1,2) we obtain:

$$u_1(t) = \sum_{n,m=0}^{L} (a_{22}^k a_{11}^m - a_{12}^k a_{21}^m) s_1(t-n-m) \quad (7)$$

$$u_2(t) = \sum_{n,m=0}^{L} (a_{22}^k a_{11}^m - a_{12}^k a_{21}^m) s_2(t-n-m) \quad (8)$$

Such a solution is good in practice in both quality of the output (i.e. little artifacts) and quantity of the cross-talk (ideally zero). However, it requires knowledge about the room impulse responses (i.e. mixing matrix A) and that is a daunting task when performed blindly. As we show next, a truncated approximation of the full mixing matrix yields good separation results. This suggest to use a lower dimensional parameterization of the mixing process. The issue then becomes, how robust is separation in the presence of uncertainties about the impulse response coefficients?

Next we analyze the question of robustness of parametric demixing solution in the case of echoic mixing. The problem can obviously be formulated in the case of more than two channels. However, we only consider a two microphone array due to the potential improvement over single microphone speech enhancement solutions and the economic potential.

3. APPROACH TO MEASURING ROBUSTNESS

Consider a mixing matrix of (sparse) FIR filters A as in (3), where the mixing coefficients a_{ij}^k are ordered according to their arrival time. We define the *truncation of order* q of this matrix as the 2×2 matrix of FIR filters obtained by truncating A_{ij} to its first $q + 1$ nontrivial (i.e. non-zero) terms. Thus:

$$trunc_q(A) = \begin{bmatrix} \sum_{k=0}^{q} a_{11}^k z^{-\tau_{11}^k} & \sum_{k=0}^{q} a_{12}^k z^{-\tau_{12}^k} \\ \sum_{k=0}^{q} a_{21}^k z^{-\tau_{21}^k} & \sum_{k=0}^{q} a_{22}^k z^{-\tau_{22}^k} \end{bmatrix} \quad (9)$$

The adjoint matrix of this truncated matrix, gives rise to an demixing filter denoted W_q. Thus $W_q = adj(trunc_q(A))$. Note that the two operation commute in this case:

$$trunc_q(adj(A)) = adj(trunc_q(A))$$

Thus W_q is the truncated matrix of the complete demixing matrix $W = adj(A)$.

Consider the setup of an echoic environment as in Figure 1. We assume specific reflection coefficients on floor, walls, and ceiling, two microphones placed at P_1 and P_2, and two independent sources of unit variance white noise positioned at V_1 and V_2 (whose position will change). Assume mixing filters are given for a *nominal* position of V_2, say $A(V_2^0)$, and a demixing filter W_q is constructed according to (9). We evaluate separation performance for the case when the actual position of the second source (V_2) differs from the assumed position V_2^0.

To do so we first introduce and explicitly compute the SNR gain of the overall scheme. Since we assumed the sources are unit variance white noises, the input SNRs are:

$$SNR_1^i = \frac{\|A_{11}\|^2}{\|A_{12}\|^2} \;,\; SNR_2^i = \frac{\|A_{22}\|^2}{\|A_{21}\|^2} \quad (10)$$

where the norms are given by:

$$||A_{ij}||^2 = \sum_{k=0}^{L} |a_{ij}^k|^2 \tag{11}$$

The output SNRs are given by:

$$\begin{aligned} SNR_1^o &= \frac{||W_{11}A_{11}+W_{12}A_{21}||^2}{||W_{11}A_{12}+W_{12}A_{22}||^2} \\ SNR_2^o &= \frac{||W_{22}A_{22}+W_{21}A_{12}||^2}{W_{21}A_{11}+W_{22}A_{21}||^2} \end{aligned} \tag{12}$$

Hence the SNR gain is measured by:

$$G_1 = \quad 10\,log_{10}\left(\frac{||W_{11}A_{11}+W_{12}A_{21}||^2}{||W_{11}A_{12}+W_{12}A_{22}||^2}\frac{||A_{12}||^2}{||A_{11}||^2}\right) \tag{13}$$

$$G_2 = \quad 10\,log_{10}\left(\frac{||W_{21}A_{12}+W_{22}A_{22}||^2}{||W_{21}A_{11}+W_{22}A_{21}||^2}\frac{||A_{21}||^2}{||A_{22}||^2}\right) \tag{14}$$

Having established the robustness criterion, now we define how we represent *uncertainty* in the estimates of the demixing parameters. For the nominal configuration of sources (V_1, V_2^0) the mixing matrix is A_0. To it there correspond a series of demixing matrices defined via:

$$W_q = adj(trunc_q(A_0)) = W_q(V_2^0) \tag{15}$$

and indexed by the truncation order q. Assume now that one of the sources (which in our setup will be source number two) is in fact located in a different position, say V_2. Then, the true mixing matrix is $M = M(V_2)$ and the overall performance of the demixing scheme is characterized by the gains (13) computed for (M, W_q). Thus we obtain two position dependent gain functions $G_1^q(V_2)$, $G_2^q(V_2)$, indexed by the truncation order q. Assuming the position V_2 is uniformly distributed in a ball of radius r around the nominal position V_2^0, we want to estimate the average SNR gain of this demixing scheme. Then the quantities we are interested in are:

$$\overline{G_1}(q, R) = \frac{1}{Vol(B_R)} \int_{B_R(V_2^0)} G_1(V_2) d^3 V_2 \tag{16}$$

and $\overline{G_2}(q, R)$ defined similarly.

Next we present experimental separation results for the setup presented before. Since the behavior of $\overline{G_1}$ and $\overline{G_2}$ is very similar we concentrate only on the former.

4. EXPERIMENTAL RESULTS

An echoic room has been simulated as in Figure 1 with reflection coefficients $(0.5, 0.5, 0.2)$ on floor, walls, and ceiling. This roughly corresponds to a reverberation time of about 100msec. The microphone distance was $10cm$ and the distances between the sources and mid-point between microphones were 1m and 1.5m respectively. The first source was fixed on the line connecting the microphones while the second source was rotated in increments of 30 degrees from -120^o to $+120^o$. Each such position was a nominal position for robustness measurement. Impulse responses were computed by taking into account sound bouncing of the wall up to order 5 at a sampling frequency of 16kHz, in a ray-tracing model. On average, we obtained about 200 coefficients per channel (see Figure 2). The truncation order ranged from 0 (direct path) to 10 (direct path plus 10 echoes). The ball radius varied from 5cm to 1m in increments of 5cm. On each spherical corona we computed the gain for 288 points and then averaged out the result to obtain an estimate of $\overline{G_1}$ as in (16). The average SNR gains are presented in Table 1 for $\theta = 30^o$ and $\theta = 60^o$.

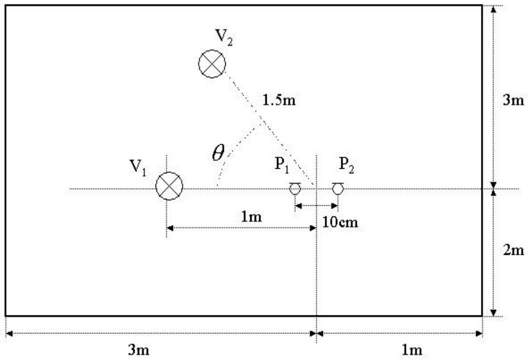

Fig. 1. Setup Configuration.

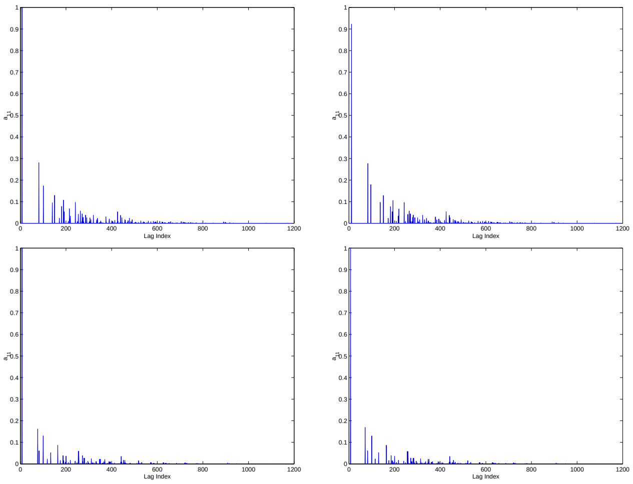

Fig. 2. Impulse Responses: A_{11} (top-left), A_{12}(top-right), A_{21} (bottom left) and A_{22} (bottom right) for $\theta = 90$.

Figures 3-11 represent the variations of SNRs with re-

$q\backslash r[m]$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	6.05	5.94	5.67	5.36	5.19	4.90	4.69	4.46	4.22	3.98	3.72
1	6.80	6.10	5.77	5.52	5.37	5.09	4.89	4.65	4.41	4.16	3.89
2	8.01	6.02	5.61	5.34	5.17	4.89	4.68	4.44	4.21	3.97	3.71
3	8.93	5.83	5.31	4.97	4.78	4.49	4.29	4.05	3.83	3.60	3.35
4	9.29	5.89	5.35	5.03	4.83	4.54	4.33	4.10	3.87	3.64	3.39
5	9.54	5.88	5.32	4.99	4.79	4.50	4.29	4.06	3.83	3.61	3.36
6	9.75	5.92	5.36	5.02	4.82	4.53	4.32	4.09	3.86	3.63	3.38
7	10.62	5.80	5.25	4.92	4.71	4.42	4.22	3.99	3.77	3.54	3.30
8	12.01	5.75	5.20	4.87	4.67	4.38	4.18	3.95	3.74	3.51	3.27
9	12.00	5.73	5.19	4.86	4.66	4.37	4.17	3.94	3.73	3.50	3.26
10	12.10	5.75	5.20	4.87	4.67	4.38	4.18	3.95	3.73	3.51	3.27

$q\backslash r[m]$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	7.48	5.71	5.57	5.25	4.71	4.42	4.12	3.77	3.52	3.34	3.12
1	8.59	6.22	6.00	5.69	5.12	4.81	4.51	4.13	3.86	3.66	3.43
2	10.17	6.83	6.44	6.06	5.44	5.10	4.77	4.39	4.11	3.90	3.66
3	11.62	6.97	6.44	6.02	5.37	5.02	4.69	4.30	4.03	3.82	3.58
4	12.20	7.16	6.61	6.16	5.50	5.14	4.80	4.40	4.12	3.90	3.65
5	12.42	7.14	6.58	6.13	5.47	5.11	4.77	4.38	4.09	3.88	3.63
6	12.70	7.20	6.62	6.16	5.49	5.13	4.78	4.39	4.10	3.89	3.64
7	13.50	7.18	6.56	6.09	5.43	5.06	4.72	4.33	4.05	3.84	3.59
8	14.02	7.31	6.67	6.19	5.52	5.14	4.79	4.40	4.11	3.89	3.64
9	14.04	7.32	6.67	6.19	5.52	5.14	4.79	4.40	4.11	3.89	3.64
10	14.69	7.21	6.57	6.09	5.43	5.06	4.72	4.32	4.04	3.83	3.59

Table 1. SNR gains in [dB] for $\theta = 30^o$ (top) and $\theta = 60^o$ (bottom).

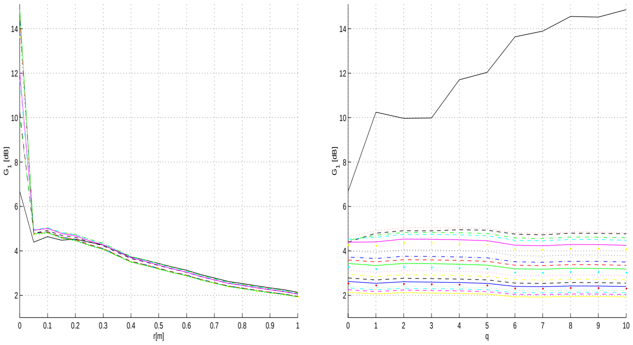

Fig. 3. SNR Gain for $\theta = -120^o$. Text describes the family of plots in the left and right figure.

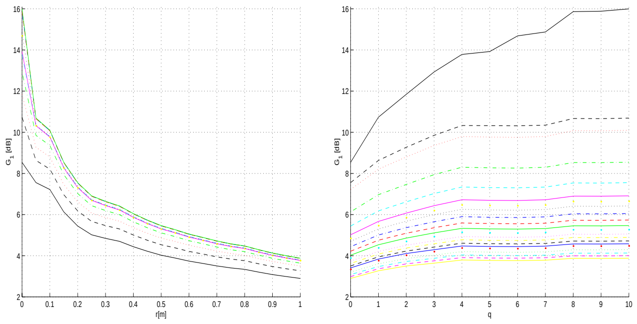

Fig. 4. SNR Gain for $\theta = -90^o$

spect to the approximation degree q, for 11 values of r (from 0 to $1.0m$ in increments of $10cm$: $r = 0, 0.1, 0.2, \ldots, 1.0$) and the variation of SNRs with respect to the distance r, for 11 values of q (from 0 to 10), in the left and right positions respectively. The left family of 11 plots is parameterized by q, where $q = 0$ is given by continuous line, $q = 1$ by dashed line, etc. In general, the higher is q the higher the average gain, but not always. The right family of 11 plots is parameterized by r, where $r = 0$ is given by continuous line, $r = 0.1$ by dashed line, etc. The higher is r the lower is the average gain.

These plots show that a significant SNR improvement is obtained by higher order demixing schemes, when the source positions are known precisely (zero error). However, in the presence of uncertainties performance degrades fast. Thus, as little as 5 cm makes the performance insensitive to the modeling degree the angles $\theta = -120$, $\theta = -30$, $\theta = 30$ and $\theta = 120$), whereas at $\theta = 0$, the performance degrades when increasing the model order. On the other hand, for an uncertainty as little as 10cm, the SNR gain increases by only 1-3dB when going from the lowest order model (direct path) to the highest complexity model considered here (direct path + 10 echoes). This shows that higher-order-model based demixing behaves almost as well as the direct-path-only demixer in the presence of position uncertainties.

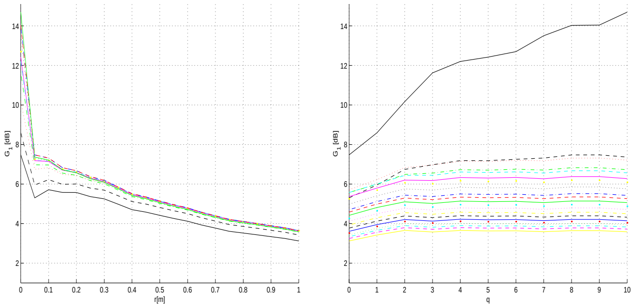

Fig. 5. SNR Gains for $\theta = -60^o$

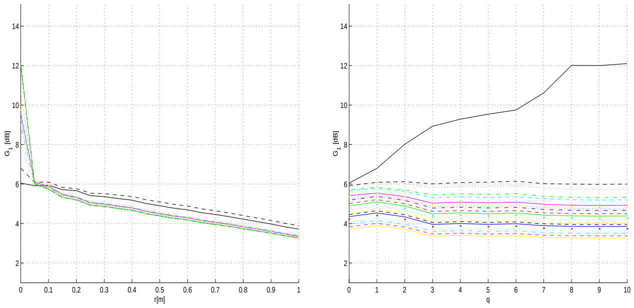

Fig. 6. SNR Gain for $\theta = -30^o$

5. CONCLUSIONS

We studied the behavior of a class of two-channel parametric demixing schemes based on room modeling under parameter estimate uncertainties. Assuming that an oracle (e.g. a precalibration) provides the FIR filter mixing matrix for a specified position of the sources, we analyzed the influence of the position uncertainty to the SNR gain for several degrees of approximation. In particular we varied the demixing filter order by considering up to ten multi-paths, and the position uncertainty from 0 to 1m, in increments of 5cm. We computed analytically the SNR gain for the two-microphone demixing scheme based on the adjoint of the mixing matrix.

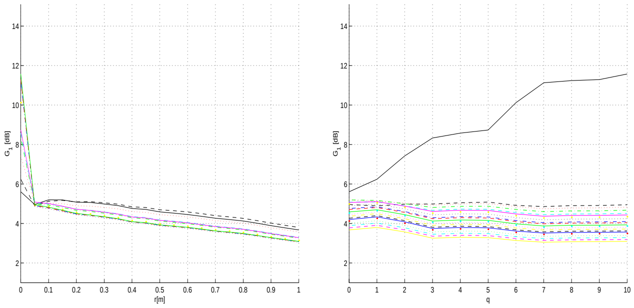

Fig. 7. SNR Gain for $\theta = 0^o$

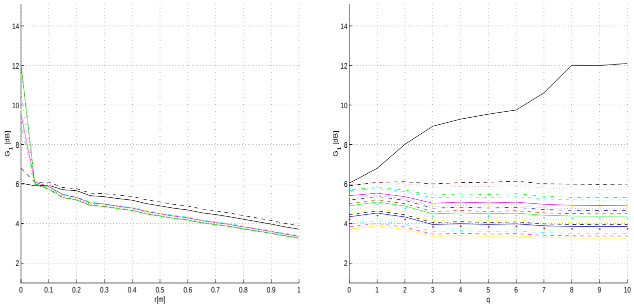

Fig. 8. SNR Gain for $\theta = 30^o$

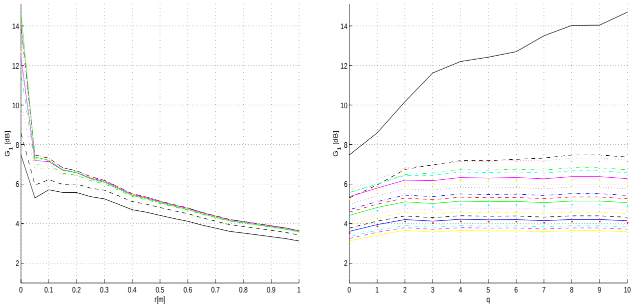

Fig. 9. SNR Gain for $\theta = 60^o$

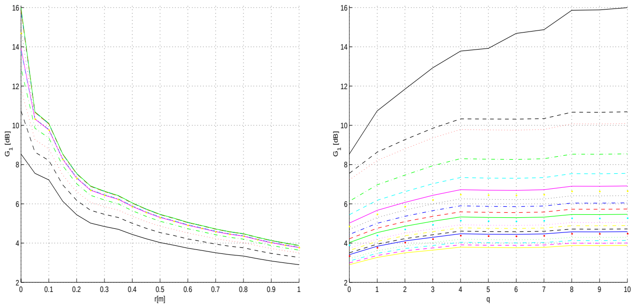

Fig. 10. SNR Gain for $\theta = 90^o$

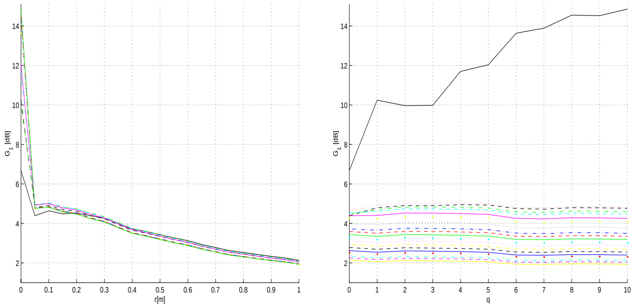

Fig. 11. SNR Gain for $\theta = 120^o$

The results showed a dramatic degradation in SNR performance for as little as 5cm uncertainty in source position. They also showed that higher order models do not sensibly improve compared to the pure direct path or lower order demixing schemes. Moreover, performance degrades when increasing the demixing model order in some cases.

A higher order parametric model identification algorithm would be expensive while its demixing scheme offers only marginal improvements, if any, under the reasonable assumption that parameters are not estimated perfectly (as modeled by our uncertainty in source position). Therefore we suggest that further research avoid increasing the mixing model complexity (e.g. by complex parameterization) and instead concentrate on lower order mixing models.

6. REFERENCES

[1] K. Torkolla, "Blind separation for audio signals: Are we there yet?," in *First International Workshop on Independent component analysis and blind source separation*, Aussois, France, Jan. 1999, pp. 239–244.

[2] F. Asano and S. Ikeda, "Evaluation and real-time implementation of blind source separation system using time-delayed decorrelation," in *Proceedings of the Second International Workshop on ICA and BSS*, P. Pajunen and J. Karhunen, Eds. 2000, Otamedia.

[3] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[4] T. J. Ngo and N.A. Bhadkamkar, "Adaptive blind separation of audio sources by a physically compact device using second order statistics," in *First International Workshop on ICA and BSS*, Aussois, France, Jan. 1999, pp. 257–260.

[5] Justinian Rosca, Joseph Ó Ruanaidh, Alexander Jourjine, and Scott Rickard, "Broadband direction-of-arrival estimation based on second order statistics," in *Advances in Neural Information Processing Systems 12*, S.A. Solla, T.K. Leen, and K.-R. Müller, Eds. 2000, pp. 775–781, MIT Press.

[6] Y. Xiang, Y. Hua, S. An, and A. Acero, "Experimental investigation of delayed instantaneous demixer for speech enhancement," in *Proceedings ICASSP*. 2001, IEEE Press.

[7] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, 1996.

[8] Lucas Parra, Clay Spence, and Bert De Vries, "Convolutive blind source separation based on multiple decorrelation," in *NNSP98*, 1988.

[9] Radu Balan, Justinian Rosca, Scott Rickard, and Joseph Ó Ruanaidh, "The influence of windowing on time delay estimates," in *Proceedings CISS 2000, Princeton, NJ*, 2000, Princeton.

[10] S. Ikeda and N. Murata, "A method of ica in time-frequency domain," in *Proceedings of the 1st ICA Conference, Aussois France*, 1999, pp. 365–370.

[11] Jrn Anemller and Birger Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceedings of the second international workshop on independent component analysis and blind signal separation*, Petteri Pajunen and Juha Karhunen, Eds., Helsinki, Finland, June 19–22 2000, pp. 215–220.

[12] R. Balan and J Rosca, "Statistical properties of stft ratios for two channel systems and applications to blind source separation," in *Proceedings ICA 2000, Helsinki*, Petteri Pajunen and Juha Karhunen, Eds. 2000, pp. 429–434, Otamedia, Helsinki, Finland, June 2000.

[13] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency domain ica and beamforming," in *Proceedings ICASSP*. 2001, IEEE Press.

[14] J.L. Flanagan, A.C. Surendran, and E.E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, pp. 207–222, 1993.