ROBUSTNESS OF PARAMETRIC SOURCE DEMIXING IN ECHOIC ENVIRONMENTS

Radu Balan, Justinian Rosca, Scott Rickard

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

ABSTRACT

In this paper we present a robustness study of the two channel echoic parametric demixing problem. More specific, assume an oracle (or a perfect estimator) is providing a truncated estimate of the mixing room FIR filters for a source configuration. Based on this information, the unmixer is constructed with the adjoint of the (truncated) mixing matrix. For several degrees of truncation, we compute how fast the separation SNR is decaying with respect to a radius R, when the actual position of one source is uniformly distributed on a ball of radius R, around the assumed position. Numerical simulations of artificial echoic mixings show that despite the increasing of the demixing SNR gain with the demixing filter truncation order, the higher order filters are less robust to position uncertainties and the overall performance remains almost constant after the second order approximation.

1. INTRODUCTION

The Blind Source Separation problem has been the focus of many studies in recent years. Two international conferences (Aussois 1999, Helsinki 2000) have been dedicated to this (and Independent Component Analysis, that is a closely related subject) topic. Successful applications in images and medical signals have been presented. Yet, in audio signal processing, real-time blind source separation techniques proved only modest gains (see [1, 2]), maybe not completely unexpected would we take into account the results of multi-microphone signal enhancement techniques (see [3]). Several BSS methods have been proposed to separate and unmix more voices. In particular there are two classes of unmixing methods: one class uses parametric mixing models thus reducing the number of degrees of freedom of the identification problem, whereas the other class doesn't exploit the relative sparsity of the mixing model, but uses a full non-parametric (or at least, not explicitely parametric), unmixing scheme. We shall call the former class the *parametric BSS*, whereas the latter solution will be termed as *non-parametric BSS*. Parametric BSS solutions have first been studied in the context of unechoic mixtures (see [4, 5]). In such cases, only four parameters are needed: two delays and two attenuations. Moreover, if the microphones are close enough, the attenuations can be set to one, and only two parameters, namely the delays, have to be used. For echoic environments, the simple direct-path model can be used as an starting point for a more complex mixing (or unmixing) model (see [6]). Nonparametric mixing models are implementing either in time-domain or frequency domain. The time-domain approach considers long FIR or IIR filters and tries to adapt the filter coefficients so to obtain as independent outputs as possible (see [7, 8]). The frequency domain approach makes use of the following simple but very useful observation, namely, 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 requires long filters. On top of this, there is a permutation problem that 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 using some criterion, find the right permutation matrix (see [10, 11, 12, 13]).

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

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

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

where e is the number of echoes (path) the model has (is determined by the room reverberation time through the sampling frequency), $s_1(\cdot)$, $s_2(\cdot)$ are the source signals, $x_1(\cdot), x_2(\cdot)$ are the measured signals, a_{ij}^n is the n^{th} path attenuation coefficient from source j to microphone i, and τ_{ij}^n the corresponding delay. All the time variables (and delays) are measured in samples. For the delays, we assume the sampling frequency is sufficiently high, and the distance between microphones, respectively between sources, sufficiently large to be assumed integers. Let us denote by M the 2×2 matrix of mixing filter transfer functions:

$$M(z) = \left[\begin{array}{cc} M_{11}(z) & M_{12}(z) \\ M_{21}(z) & M_{22}(z) \end{array}\right] \quad (3)$$

$$M_{ij}(z) = \sum_{n=0}^{e} a_{ij}^n z^{-\tau_{ij}^n} \quad (4)$$

There are at least four techniques to enhance the sources s_1 and s_2 from the mixtures x_1, x_2. The first two methods aim toward source separation and use either the inverse of the mixing matrix, or its adjoint (see [5, 4]). The other two techniques aim mostly to signal enhancement and are the multiple delay-and-sum beamformer (when only information of arrival times is required), or matching filters (when the full mixing matrix is used) - see [14]. In this paper we discuss the use of the adjoint matrix as an unmixing solution.

Set $W = adj(M)$, the adjoint of M. Recall the adjoint is defined by:

$$adj(M) = \left[\begin{array}{cc} M_{22}(z) & -M_{12}(z) \\ -M_{21}(z) & M_{11}(z) \end{array}\right]$$

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

$$\left[\begin{array}{c} u_1 \\ u_2 \end{array}\right] = W \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] \quad (5)$$

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

and combined with (1,2) we do obtain a separation between between the outputs:

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

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

Such a solution is very good in practice, in both the quality of the output (i.e. artifacts) and the quantity of the cross-talk (ideal is zero). However, it requires the knowledge of the room impulse responses (i.e. mixing matrix M) 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 parametrization of the mixing process. The issue then becomes, how robust the separation is in the presence of uncertainties on impulse response coefficients? In this paper we give an answer to this question, of the robustness of parametric demixing solution in the case of echoic source mixing. The problem can obviously be formulated in the case of more than 2 channels. Here we consider only the case of two microphones because of a practical constraint: We target applications into mobile communication area, and mobile phones, or PDA boards, are not big enough to make feasable the use of more than 2 microphones.

The organization of the paper is as follows: section 2 presents the setup and robustness measures; section 3 contains the numerical results; section 4 contains the conclusions and is followed by the bibliography.

2. MEASURES OF ROBUSTNESS

Consider a mixing matrix of (sparse) FIR filters M as in (3), where the mixing coefficients a_{ij}^n 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 M_{ij} to its first $q+1$ nontrivial (i.e. non-zero) terms. Thus:

$$trunc_q(M) = \left[\begin{array}{cc} \sum_{n=0}^{q} a_{11}^n z^{-\tau_{11}^n} & \sum_{n=0}^{q} a_{12}^n z^{-\tau_{12}^n} \\ \sum_{n=0}^{q} a_{21}^n z^{-\tau_{21}^n} & \sum_{n=0}^{q} a_{22}^n z^{-\tau_{22}^n} \end{array}\right] \quad (9)$$

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

$$trunc_q(adj(M)) = adj(trunc_q(M)).$$

Hence, we can equally say W_q is simply the truncated matrix of the complete unmixing matrix $W = adj(M)$.

Consider now the following setup. Into an echoic room ($4 \times 5 \times 2.5$ m) as in Figure 1, with reflection coefficients $(0.5, 0.5, 0.2)$ (floor, walls, ceil), we place two microphones at P_1 $(2.95, 2, 1)$ and P_2 $(3.05, 2, 1)$ and two independent sources of unit variance white noise at V_1 $(2, 2, 1.5)$, and V_2 (whose position will change). Assume the mixing filters are given for a nominal position of V_2, say $M(V_2 = V_{20})$, and an unmixing filter W_q is constructed according to (9). Fix this unmixing filter W_q. We want to evaluate its separation performance for the case when the actual position of the second source (V_2) differs from the assumed position V_{20}.

To do so we first introduce and explicitely compute the SNR gain of the overall scheme. Denote by M some generic mixing filters and by W some generic unmixing filters. Since we assumed the sources are unit variance white noise, the input SNRs are:

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

where the norms are given by:

$$\|M_{ij}\|^2 = \sum_{n=0}^{e} |a_{ij}^n|^2 \quad (11)$$

The output SNRs are given by:

$$\begin{aligned} SNR_1^o &= \frac{\|W_{11}M_{11}+W_{12}M_{21}\|^2}{\|W_{11}M_{12}+W_{12}M_{22}\|^2} \\ SNR_2^o &= \frac{\|W_{22}M_{22}+W_{21}M_{12}\|^2}{W_{21}M_{11}+W_{22}M_{21}\|^2} \end{aligned} \quad (12)$$

Hence the SNR gain is measured by:

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

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

Once we have established the robustness criterion, we define now the uncertainty model. Let us return to the setup presented before. For the nominal configuration of sources $(V_1, V_2 = V_{20})$ the mixing matrix is M_0. To such a mixing matrix there correspond a series of unmixing matrices defined via:

$$W_q = adj(trunc_q(M_0)) = W_q(V_{20}) \quad (15)$$

and indexed by the truncation order q. Asssume 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 unmixing scheme is characterized by the gains (13) computed for (M, W_q). Thus we obtain two position dependent 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_{20}, we want to estimate the average SNR gain of this demixing scheme. Then the quantities we are interested in are:

$$avG_1(q, R) = \frac{1}{Vol(B_R)}\int_{B_R(V_{20})} G_1(V_2)d^3V_2 \quad (16)$$

and $avG_2(q, R)$ defined similarly. In the next section we present the numerical results for the setup presented before. Since the behaviour of avG_2 is very much similar to that of avG_1, we concentrate only on the former criterion.

3. NUMERICAL RESULTS

The echoic environment presented before was simulated on a Pentium III machine. The microphone distance was $10cm$, and the distances between the sources and mid-point between microphones were 1m, respectively 1.5m. The first source was fixed on the line connecting the microphones (as in Figure 1), whereas source 2 was rotated in increments of 30 degrees between -120 degree and +120 degree. Each such position was a nominal position for robustness measurement. The impulse responses were computed by taking into account all sound bouncing of the wall up to order 5 at

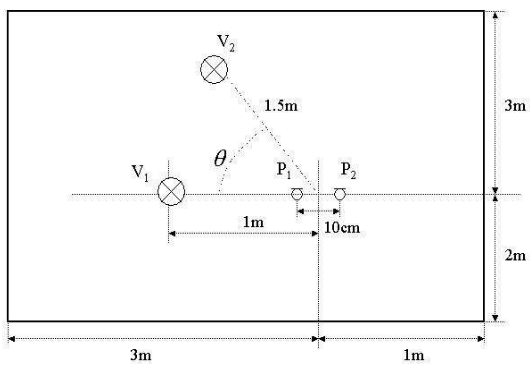

Fig. 1. Setup Configuration.

a sampling frequency of 16KHz. In average, we obtained about 200 coefficients per channel (see Figure 2). The truncation order ranged from 0 (direct path) to 10 (direct path + 10 echoes). The ball radius varied from 5cm up to 1m, in increments of 5cm. On each spherical corona we computed the gain for 288 points, and then averaged out the result thus obtaining an estimate of avG_1 of (16). For $\theta = 30^o$ and $\theta = 60^o$ the average SNR gains are presented in Table 1.

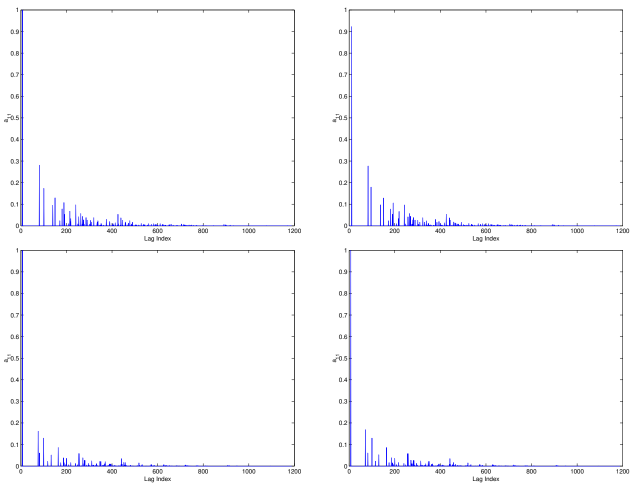

Fig. 2. Impulse Responses: M_{11} (top-left), M_{12}(top-right), M_{21} (bottom left) and M_{22} (bottom right) for $\theta = 90$.

Next we plot the variations of SNRs with respect to the approximation degree q, for 11 values of r (from 0 to $1.0m$ in increments of $10cm$: $r = 0, 0.1, 0.2, \ldots, 1.0$) - left plots - and the variation of SNRs with respect to the distance r, for 11 values of q (from 0 to 10) - right plots (Figures 3-11).

$q\backslash r[m]$	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).

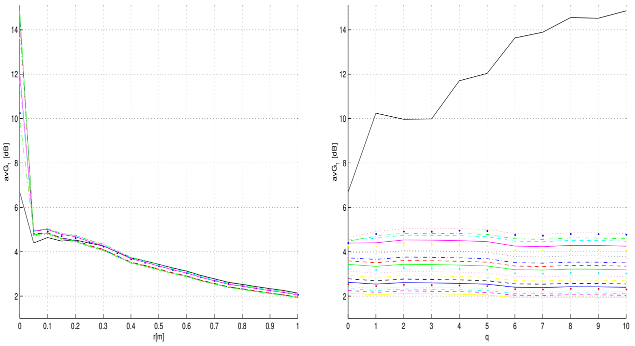

Fig. 3. SNR Gain for $\theta = -120^o$

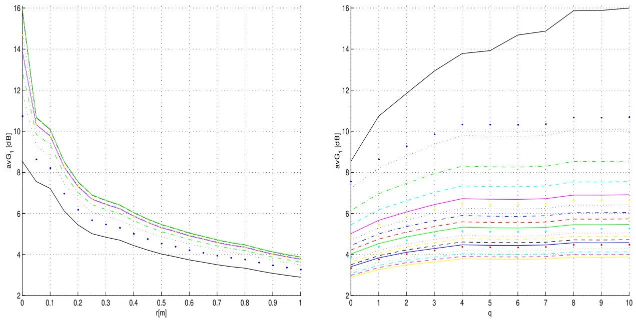

Fig. 4. SNR Gain for $\theta = -90^o$

These plots show that a significant SNR impovement is obtained by higher order demixing schemes, when the source positions are known precisely (zero error). However, in the presence of uncertainties, the performance degrades very fast. Thus, as little as 5 cm makes the performance insensitive to the modeling degree (see the angles $\theta = -120$, $\theta = -30$, $\theta = 30$ and $\theta = 120$), whereas at $\theta = 0$, the performance downgrades with the increasing of the model order. On the other hand, for uncertainty as little as 10cm, the SNR gains 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.

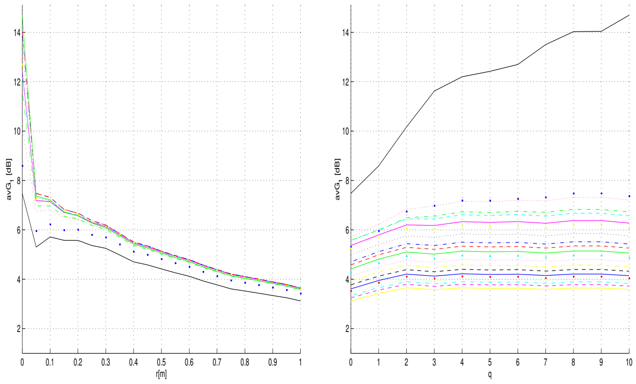

Fig. 5. SNR Gains for $\theta = -60^o$

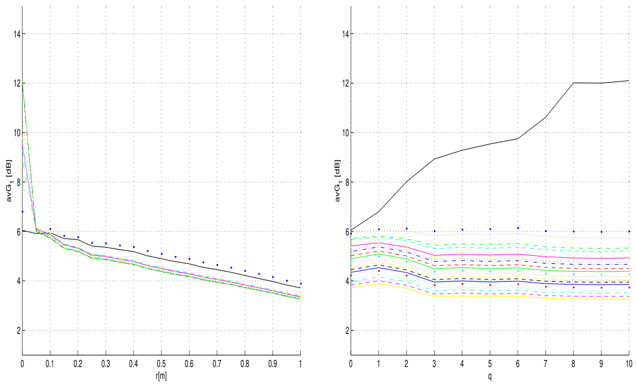

Fig. 6. SNR Gain for $\theta = -30^o$

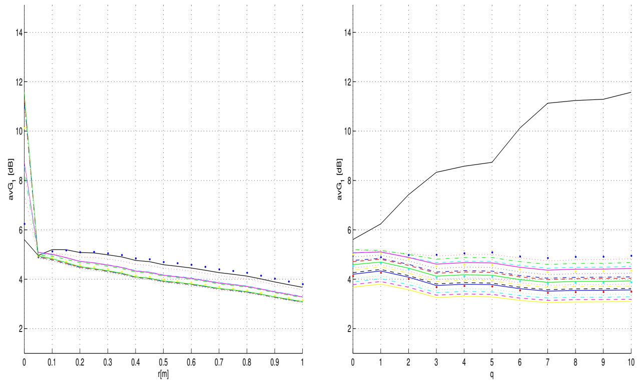

Fig. 7. SNR Gain for $\theta = 0^o$

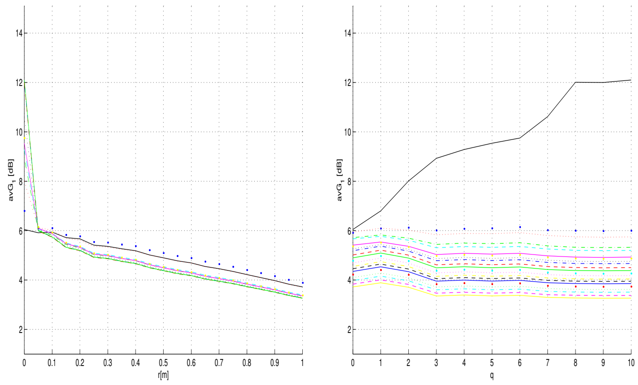

Fig. 8. SNR Gain for $\theta = 30^o$

4. CONCLUSIONS

In this paper we studied the behaviour of room modeling based demixing schemes under the presence of uncertainties. We have computed analytically the SNR gain of the two-microphone demixing scheme based on the adjoint of the mixing matrix. Assuming that an oracle (or through a precalibration) would tell us the mixing matrix for a specified position of the sources, we analyzed the influence of the position uncrtainty to the SNR gain for several degrees of approximation. In particular we varied the demixing filter order by considering up to 10 multipaths, and the position uncertainty from 0 to 1m, in increments of 5cm.

The results show a dramatic degrading of the performance for as little as 5cm uncertainty in the position of the sources. They also show that higher order models do not sensibly improve, compared to the direct path only or other lower order demixing schemes. In fact, for some configuration, the performance degrades by increasing the demixing model order.

Since a higher order parametric model identification algorithm is very expensive, and the corresponding demixing scheme would improve by very little in the presence of uncertainties, it seems reasonable that further research should avoid increasing the mixing model complexity (by parametrizing several multipaths), instead it should concentrate in lower order mixing models (direct path only, or direct path plus one or two paths).

5. 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] Justinian Rosca, Joseph Ó Ruanaidh, Alexander Jourjine, and Scott Rickard, "Broadband direction-of-

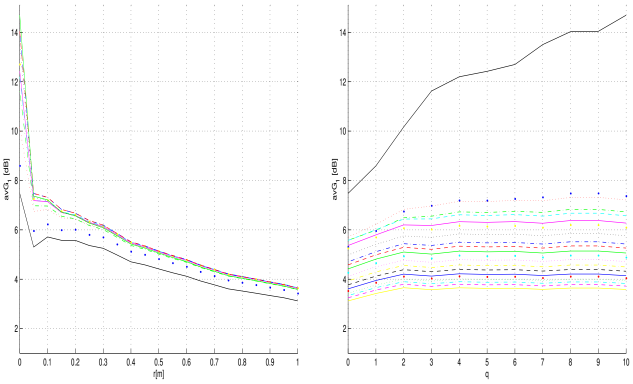

Fig. 9. SNR Gain for $\theta = 60^o$

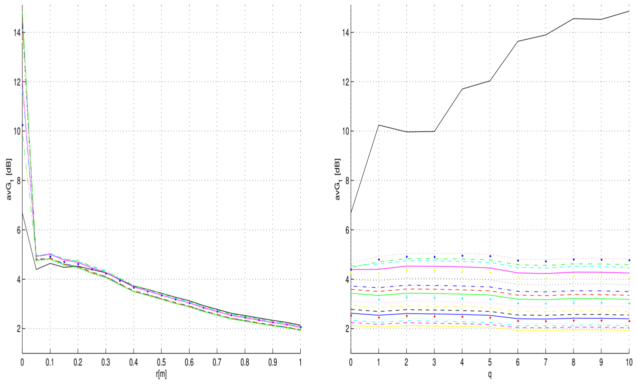

Fig. 11. SNR Gain for $\theta = 120^o$

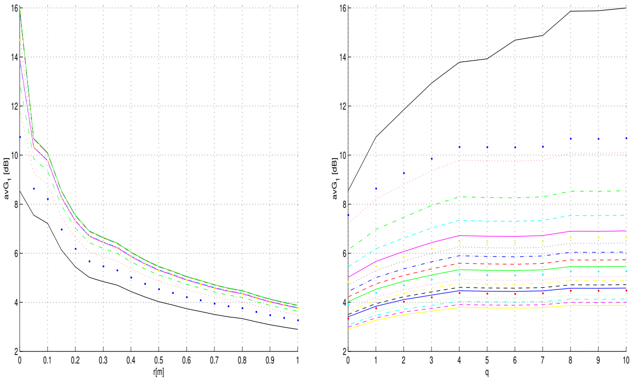

Fig. 10. SNR Gain for $\theta = 90^o$

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.

[5] 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.

[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.