BLIND SPEECH SEPARATION EMPLOYING DIRECTIONAL STATISTICS IN AN EXPECTATION MAXIMIZATION FRAMEWORK

Dang Hai Tran Vu and Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany
{tran,haeb}@nt.uni-paderborn.de

ABSTRACT

In this paper we propose to employ directional statistics in a complex vector space to approach the problem of blind speech separation in the presence of spatially correlated noise. We interpret the values of the short time Fourier transform of the microphone signals to be draws from a mixture of complex Watson distributions, a probabilistic model which naturally accounts for spatial aliasing. The parameters of the density are related to the a priori source probabilities, the power of the sources and the transfer function ratios from sources to sensors. Estimation formulas are derived for these parameters by employing the Expectation Maximization (EM) algorithm. The E-step corresponds to the estimation of the source presence probabilities for each time-frequency bin, while the M-step leads to a maximum signal-to-noise ratio (MaxSNR) beamformer in the presence of uncertainty about the source activity. Experimental results are reported for an implementation in a generalized sidelobe canceller (GSC) like spatial beamforming configuration for 3 speech sources with significant coherent noise in reverberant environments, demonstrating the usefulness of the novel modeling framework.

Index Terms— Noisy Source Separation, Sparse Signal Separation, EM-Algorithm, Directional Statistics, Speech Enhancement

1. INTRODUCTION

The objective of blind source separation (BSS) is to extract source signals from mixed signals at the sensors and, if possible, to estimate the unknown mixing channel. The BSS technique for speech dealt with in this paper can be used in many applications of speech enhancement including hands-free telecommunication and automatic meeting note taking.

Two main approaches have emerged for BSS. One is based on independent component analysis and the other relies on the sparseness of source signals. The proposed technique in this paper belongs to the latter approach. Building on the sparse source assumption each time frequency slot can be assigned to a single dominant source. A frequently used approach to identify the dominant source is by clustering amplitude and phase differences of closely spaced microphone pairs [1]. This approach has been refined in various ways, e.g. by using multiple microphones [2] and spatio-temporal filtering [3].

In this paper we propose a probabilistic framework similar to [4] for the detection of the dominant source and the mixing system identification with an arbitrarily arranged microphone array and in the presence of additive noise. As the source separation employs spatial information, we suggest to use the complex Watson distribution [5] to model the short time Fourier transform (STFT) coefficients of the microphone signals, a distribution frequently employed in directional statistics [6]. It is a bimodal distribution on the complex hypersphere with maxima at antipodal locations. The mixing system identification turns into the Maximum Likelihood (ML) estimation of the means and concentration parameters of a mixture of Watson distributions. The estimation is solved with the EM algorithm, where the hidden variable is the identity of the dominant source. In doing so, all transfer function ratios from sources to sensors can be determined even if multiple sources are simultaneously active and even if noise is present at all times.

This work was in part supported by Deutsche Forschungsgemeinschaft under contract number HA 3455/4-1.

2. PROPOSED METHOD

We are given an array of D microphones recording P speakers in a reverberant enclosure. Let $s_1(n), \ldots, s_P(n)$ be the discrete-time desired speech signal sources. The captured convolutive mixtures $x_1(n), \ldots, x_D(n)$ are given by

$$x_j(n) = \sum_{i=1}^{P} \sum_{l=0}^{L} h_{ij}(l) s_i(n-l) + n_j(n), \qquad (1)$$

where $j = 1, \ldots, D$ and $h_{ij}(l)$ is the unknown impulse response from source i to microphone j of length L and $n_j(n)$ is zero mean noise at sensor j with unknown spatial and spectral statistics.

Using the STFT the mixture in (1) can be approximated by

$$X_j(m,k) = \sum_{i=1}^{P} H_{ij}(k) S_i(m,k) + N_j(m,k),\ j = 1, \ldots, D \qquad (2)$$

where $H_{ij}(k)$ is the transfer function (TF) from source i to microphone j. $S_i(m,k)$ and $N_j(m,k)$ are STFTs of the source and the noise signals, respectively, where m is the time frame (block) index and k denotes the frequency bin. In a more compact vector notation the set of equations (2) is given by

$$\mathbf{X}(m,k) = \sum_{i=1}^{P} \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k), \qquad (3)$$

where $\mathbf{X} = [X_1, \ldots, X_D]^T$ is the observation vector, $\mathbf{H}_i = [H_{i1}, \ldots, H_{iD}]^T$ is the vector of transfer functions (TF), and $\mathbf{N} = [N_1, \ldots, N_D]^T$ is the noise vector.

For sparse signals, such as speech, it can be assumed that at any time-frequency bin (m,k) only a single source is active (dominant). This allows us to approximate the observation model (3) as a hierarchical probabilistic model, where the first stochastic process corresponds to the selection of the dominant source, and the second to sampling a random vector according to the distribution of the chosen source. To be specific, let $Z(m,k) \in \{1, \ldots, P\}$ be a hidden random variable where $Z(m,k) = i$ indicates that source i is active

in time-frequency point (m, k). We can then approximate (3) by

$$Z(m,k) = i : \\ \mathbf{X}(m,k) = \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k). \quad (4)$$

In the following each bin k is independently processed and vectors of different frame indices m are assumed to be i.i.d. realizations of the same random variable. For the ease of notation we will drop the argument (m, k) in the next subsections wherever possible. Further, statistical expectations will be approximated by averages over the frame index m.

2.1. Spatial Whitening

To simplify subsequent statistical modeling we first carry out a spatial whitening step. Let $\mathbf{\Phi_{NN}} = \mathrm{E}\left[\mathbf{N}\mathbf{N}^H\right]$ be the power spectral density (PSD) matrix of the stationary noise vector which can be estimated in speech absence periods. Here, $(\cdot)^H$ is the conjugate transpose operator. The observation vector is spatially whitened by

$$\tilde{\mathbf{X}} = \mathbf{\Phi}_{\mathbf{NN}}^{-\frac{1}{2}} \mathbf{X}. \quad (5)$$

If noise and source signals are uncorrelated and zero mean, this prewhitening ensures that the conditional PSD matrix of the whitened observation vectors $\tilde{\mathbf{X}}|Z = i$ has the form

$$\mathrm{E}\left[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^H \middle| Z = i\right] = \sigma_i^2 \cdot \mathbf{W}_i \mathbf{W}_i^H + \mathbf{I} \quad (6)$$

where $\sigma_i^2 = \mathrm{E}\left[|S_i|^2\right] \cdot ||\mathbf{\Phi}_{\mathbf{NN}}^{-\frac{1}{2}} \mathbf{H}_i||^2$ and $\mathbf{I}$ is the identity matrix. The normalized complex vector $\mathbf{W}_i$ keeping spatial informations is given by

$$\mathbf{W}_i = \mathbf{\Phi}_{\mathbf{NN}}^{-\frac{1}{2}} \mathbf{H}_i \Big/ \left\| \mathbf{\Phi}_{\mathbf{NN}}^{-\frac{1}{2}} \mathbf{H}_i \right\|, \quad (7)$$

where $\|\cdot\|$ denotes the Euclidean norm.

2.2. Probabilistic Modeling and EM Algorithm

Now we are going to describe the new statistical model of the whitened and normalized observation vectors

$$\mathbf{Y} = \tilde{\mathbf{X}} \Big/ \left\| \tilde{\mathbf{X}} \right\|. \quad (8)$$

The normalization to unit length corresponds to a mapping of the data onto the unit hypersphere in D-dimensional complex vector space. By this, the important spatial information is kept, while variations due to the scalar signal source are disregarded.

Due to the spatial diversity of the sources, feature vectors form clusters on the hypersphere, where clusters in antipodal locations correspond to the same source, since the sign is due to the scalar source signal S_i. The spatial whitening of the previous section ensures that the density of the normalized observation vector for a given source is circular symmetric on the complex hypersphere.

We propose to model the conditional statistics of $\mathbf{Y}$ with the complex Watson distribution [5]

$$p\left(\mathbf{Y}|Z = i; \mathbf{W}_i, \kappa_i\right) = \frac{(D-1)!}{2\pi^D \; M\left(1, D, \kappa_i\right)} \, e^{\kappa_i \, |\mathbf{W}_i^H \mathbf{Y}|^2}, \quad (9)$$

where the mean orientation $\mathbf{W}_i$ is given by (7) and the concentration parameter $\kappa_i \geq 0$ is a function of σ_i^2 and $M(a, b, z)$ is the confluent hypergeometric function of the first kind. The distribution is bipolar for $\kappa_i > 0$. The greater the value of κ_i, the more the observations $\mathbf{Y}$ are concentrated around the mean orientation $\pm\mathbf{W}_i$.

Using the hierarchical model suggested in (4) the statistics of the feature vectors are characterized by a finite mixture of Watson distributions

$$p\left(\mathbf{Y}; \mathbf{\Theta}\right) = \sum\nolimits_{i=1}^{P} \alpha_i \; p\left(\mathbf{Y}|Z = i; \mathbf{W}_i, \kappa_i\right), \quad (10)$$

where $\mathbf{\Theta} = \{\alpha_1, \ldots, \alpha_P, \mathbf{W}_1, \ldots, \mathbf{W}_P, \kappa_1, \ldots, \kappa_P\}$ is the unknown parameter set and $\alpha_i = P(Z = i)$ are non-negative mixture weights which sum to one.

Modeling the statistics of the feature vectors with (9) has two advantages. Firstly, the inner product $\mathbf{W}_i^H \mathbf{Y}$ is equivalent to a spatial correlation. Taking the absolute square of this implies that we are using the response power of the spatial matched beamformer as a distance measure to separate the sources. This concept naturally fits to the beamforming point of view. Secondly, the morphology of the complex hypersphere perfectly reflects spatial aliasing occurring in high frequencies, hence the cyclic nature of the phase differences are implicitly regarded.

Now we apply the EM algorithm to derive a ML estimator for the unknown parameter set. Let $\hat{\mathbf{\Theta}}^{(\nu)}$ denote the estimate at iteration ν. Further let $\mathcal{Y} = \{\mathbf{Y}(1), \ldots, \mathbf{Y}(T)\}$ be the data set consisting of T independently drawn feature vectors following the mixture model (10) and let $\mathcal{Z} = \{Z(1), \ldots, Z(T)\}$ be the set of corresponding hidden random variables indicating the dominant source in the respective time frequency slot.

The conditional expectation over of the complete data log-likelihood is then given by

$$J = \mathrm{E}\left[\log p\left(\mathcal{Y}, \mathcal{Z}; \mathbf{\Theta}\right) \middle| \mathcal{Y}; \hat{\mathbf{\Theta}}^{(\nu)}\right] \quad (11)$$

$$= \sum_{m=1}^{T} \sum_{i=1}^{P} \gamma_i(m) \;\; \log \hat{\alpha}_i^{(\nu)} + \\ \sum_{m=1}^{T} \sum_{i=1}^{P} \gamma_i(m) \;\; \log p\left(\mathbf{Y}(m) \middle| Z(m) = i; \hat{\mathbf{W}}_i^{(\nu)}, \hat{\kappa}_i^{(\nu)}\right). \quad (12)$$

Here, $\gamma_i(m)$ is the posterior probability of the i-th source being dominant, which is computed in the E-step according to

$$\gamma_i(m) = P\left(Z(m) = i \middle| \mathbf{Y}(m); \hat{\mathbf{\Theta}}^{(\nu)}\right) \\ = \frac{p\left(\mathbf{Y}(m) \middle| Z(m) = i; \hat{\mathbf{W}}_i^{(\nu)}, \hat{\kappa}_i^{(\nu)}\right) \;\; \hat{\alpha}_i^{(\nu)}}{\sum_{l=1}^{P} p\left(\mathbf{Y}(m) \middle| Z(m) = l; \hat{\mathbf{W}}_l^{(\nu)}, \hat{\kappa}_l^{(\nu)}\right) \;\; \hat{\alpha}_l^{(\nu)}}. \quad (13)$$

The M-step is taken by maximization of the objective function (11) with the constraints $||\hat{\mathbf{W}}_i^{(\nu+1)}|| = 1$, $i = 1, \ldots, P$ to ensure unit normalization of the mean orientations and $\sum_{i=1}^{P} \hat{\alpha}_i^{(\nu+1)} = 1$. We obtain the following update equations for $\hat{\mathbf{\Theta}}^{(\nu+1)}$:

$$\mathbf{\Phi}_{\mathbf{YY},i} \; \hat{\mathbf{W}}_i^{(\nu+1)} = \lambda_i \; \hat{\mathbf{W}}_i^{(\nu+1)} \quad (14)$$

$$\hat{\alpha}_i^{(\nu+1)} = \frac{1}{T} \sum\nolimits_{m=1}^{T} \gamma_i(m) \quad (15)$$

$$\frac{M\left(2, D+1, \kappa_i^{(\nu+1)}\right)}{D \cdot M\left(1, D, \kappa_i^{(\nu+1)}\right)} = \frac{\hat{\mathbf{W}}_i^{(\nu+1),H} \mathbf{\Phi}_{\mathbf{YY},i} \hat{\mathbf{W}}_i^{(\nu+1)}}{\hat{\alpha}_i^{(\nu+1)}}, \quad (16)$$

for $i = 1, \ldots, P$. Here, λ_i is a complex constant and the power spectral density matrix $\mathbf{\Phi}_{\mathbf{YY},i}$ is obtained by

$$\mathbf{\Phi}_{\mathbf{YY},i} = \frac{1}{T} \sum\nolimits_{m=1}^{T} \gamma_i(m) \;\; \mathbf{Y}(m) \mathbf{Y}^H(m). \quad (17)$$

Observe that (14) is an eigenvalue equation. Thus, the ML estimate of the mean orientation is the eigenvector corresponding to the largest eigenvalue of the power spectral density matrix $\mathbf{\Phi}_{\mathbf{YY},i}$.

The reestimation formula (16) for the concentration parameter is more complicated, as it requires the solution of an implicit equation involving the ratio of confluent hypergeometric functions. As it is not possible to obtain a closed form solution, one has to resort to numerical approximations (see [5]).

Starting with an initial guess, the E-step (13) and the M-step (14) - (16) are iterated until no significant changes of the estimates occur. We denote the estimates of the parameter set in steady state by $\hat{\mathbf{\Theta}}^{(\infty)}$.

Note that the solution of (14), i.e. the principal eigenvector $\hat{\mathbf{W}}_i^{(\infty)}$, is only unique up to an arbitrary complex scalar. Thus, also the true source-to-sensor transfer functions $\mathbf{H}_i(k)$ cannot be determined uniquely, but rather

$$\hat{\mathbf{H}}_i(k) := \mathbf{\Phi}_{\mathbf{NN}}^{\frac{1}{2}} \hat{\mathbf{W}}_i^{(\infty)}(k) = \xi_i(k)\mathbf{H}_i(k) \tag{18}$$

where $\xi_i(k)$ is an arbitrary complex scaling constant. In other words only transfer function ratios can be determined. The argument k has been included in (18) to highlight the fact that the scaling constant $\xi_i(k)$ may be frequency dependent, resulting in an arbitrary filtering operation. This is the well-known scaling problem in BSS literature.

The relation of the proposed method to the MaxSNR beamformer presented in [7] is revealed by inserting (8), (5) and (17) in equation (14). After some algebra we obtain the equivalent equation

$$\mathbf{\Phi}_{\mathbf{XX},i}\hat{\mathbf{F}}_i = \lambda \mathbf{\Phi}_{\mathbf{NN}}\hat{\mathbf{F}}_i, \tag{19}$$

where $\hat{\mathbf{F}}_i = \mathbf{\Phi}_{\mathbf{NN}}^{-1}\hat{\mathbf{H}}_i$ and

$$\mathbf{\Phi}_{\mathbf{XX},i} = \frac{1}{T}\sum_{m=1}^{T} \gamma_i(m)\, \frac{\mathbf{X}(m)\mathbf{X}^H(m)}{\mathbf{X}^H(m)\mathbf{\Phi}_{\mathbf{NN}}^{-1}\mathbf{X}(m)}. \tag{20}$$

This is almost identical to the generalized eigenvalue problem to be solved in a beamformer which attempts to maximize the signal-to-noise ratio at the beamformer output (MaxSNR beamformer). The only difference is the presence of the posterior $\gamma_i(m)$, which is the probability that the i-th source is the dominant source for the observation $\mathbf{X}(m)$. In the single source beamforming scenario studied in [7] this term always equals one.

2.3. Permutation Alignment, Separation and Noise Suppression

Since the separation is carried out in each frequency bin separately, we have to solve the permutation problem, which is typical for frequency domain BSS approaches. We use the correlation among the posteriors $\gamma_i(m,k)$ (13) in each frequency bin to solve the permutation problem. We expect that the correlation of posteriors in different bins is high if the two bins are excited by the same source. This is in particular valid for adjacent bins or frequency bins in harmonic relation. Permutation alignment is accomplished by finding mappings which minimize the inter frequency correlation of different outputs. Since a detailed algorithm derivation for permutation alignment is not the focus of this paper we refer to known methods, e.g. [4] and assume in the following that the permutation problem is solved.

One approach to reconstruct the source signals, which can also be applied in underdetermined BSS problems ($P > D$), is to use the posteriors $\gamma_i(m,k)$ as soft masks, replacing binary masks, which were e.g. used in [1]. This approach, however, may suffer from musical tones. Since in our applications we are usually concerned with overdetermined BSS ($P < D$), we describe in the following a spatio-temporal filtering approach to recover the source signals. As this approach has been described in more detail in [8] and [3] we will only give a brief outline.

The approach is based on a GSC-like beamforming structure as depicted in figure 1.

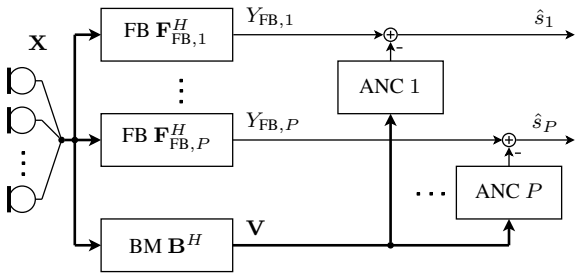

Fig. 1. GSC-like structure consisting of fixed beamformer (FB), blocking matrix (BM) and adaptive noise canceller (ANC)

The fixed beamformer (FB) outputs $Y_{\text{FB},i}(m,k)$, $i = 1,\ldots,P$ and the noise reference signals $\mathbf{V}(m,k)$ are computed by

$$Y_{\text{FB},i}(m,k) = \mathbf{F}_{\text{FB},i}^H(k)\mathbf{X}(m,k), \tag{21}$$

$$\mathbf{V}(m,k) = \mathbf{B}^H(k)\mathbf{X}(m,k), \tag{22}$$

where $\mathbf{F}_{\text{FB},i}$ and $\mathbf{B}$ are obtained from a Gram-Schmidt orthogonalization of the identified mixing system. To simplify notation we illustrate this process for the computation of the output corresponding to the P-th source. First, the intermediate coefficients $\mathbf{U}_i(k)$, $i = 1,\ldots,P$ are computed recursively. The recursion is initialized by

$$\mathbf{U}_1(k) = \hat{\mathbf{H}}_1(k) \Big/ \left\| \hat{\mathbf{H}}_1(k) \right\|. \tag{23}$$

Then for $i = 1,\ldots,P-1$:

$$\tilde{\mathbf{U}}_{i+1}(k) = \hat{\mathbf{H}}_{i+1}(k) - \sum\nolimits_1^i \left(\mathbf{U}_i^H(k)\hat{\mathbf{H}}_{i+1}(k) \right) \mathbf{U}_i(k) \tag{24}$$

$$\mathbf{U}_{i+1}(k) = \tilde{\mathbf{U}}_{i+1}(k) \Big/ \left\| \tilde{\mathbf{U}}_{i+1}(k) \right\|. \tag{25}$$

Unfortunately, the resulting filter coefficients $\mathbf{U}_P(k)$ have no constraint for target direction gain, since the scaling problem in equation (18) still remains. An approximative solution to this issue is achieved by the normalization

$$\mathbf{F}_{\text{FB},P}(k) = \mathbf{U}_P(k) \Big/ \left(\mathbf{D}_P^H(k)\mathbf{U}_P(k) \right), \tag{26}$$

where $\mathbf{D}_P(k) = [1, e^{-j\omega_k \tau_{P,2}}, ..., e^{-j\omega_k \tau_{P,D}}]^T$ is the steering vector and $\tau_{P,j}$ is the time difference of arrival (TDOA) between the the first and j-th sensor when source P is dominant. Estimates of $\tau_{P,j}$ are obtained by searching for the maximum of the cross correlation of the impulse responses corresponding to the first and the j-th estimated room transfer function [3].

Using equation (21) with the coefficients given by (26) amounts to placing spatial nulls at the interferers' directions, while preserving a distortionless response to the P-th source signal. Alternatives to the gain normalization (26) have been studied in [7].

Finally, the coefficients of the blocking matrix (BM) are obtained by

$$\mathbf{B}^H(k) = \mathbf{I} - \sum\nolimits_{i=1}^{P} \mathbf{U}_i(k)\mathbf{U}_i^H(k). \tag{27}$$

It should be noted that $\mathbf{B}^H(k)$ is equal for all P beamformers, hence a single BM provides the noise-only references for all adaptive noise cancellers, which in turn are updated in speech absence periods only.

3. SIMULATION RESULTS

We performed experiments for a simulated meeting situation with $P = 3$ sources and $D = 8$ microphones in an enclosure with varying reverberation times T_{60}. Sources are placed at a distance of 1 m around a uniform circular array of 0.1 m radius. The physical conditions are similar to the setup in [9]. The source signals were recordings of 25 s length obtained from concatenating sentences randomly drawn from the TIMIT database at 16 KHz sampling rate. To simulate coherent noise, fan noise recordings of a video projector were placed as an additional source in the room. Furthermore, white noise at the level of 30 dB below the source signal power was added to all sensors. A minimum statistics based voice activity detection (VAD) was used. Initial mean orientations are set to random values and initial concentration parameters are set to $\kappa_i^{(0)} = 20$. The STFT frame size corresponds to 64 ms with 64/4 ms frame shift. The system performance was evaluated in terms of signal-to-interference-ratio (SIR) gain, signal-to-noise-ratio (SNR) gain and signal-to-distortion-ratio (SDR)

$$\text{SIR} := 10\log_{10}\left(\text{E}\left[\hat{s}^2(t)\right]/\text{E}\left[\tilde{s}^2(t)\right]\right) \quad (28)$$

$$\text{SNR} := 10\log_{10}\left(\text{E}\left[\hat{s}^2(t)\right]/\text{E}\left[\tilde{n}^2(t)\right]\right) \quad (29)$$

$$\text{SDR} := 10\log_{10}\left(\text{E}\left[\hat{s}^2(t)\right]/\left(\text{E}\left[(\hat{s}(t) - a\acute{s}(t-\delta))^2\right]\right)\right) \quad (30)$$

where $\hat{s}(t)$ is the target signal component, $\tilde{s}(t)$ is the interferers' component and $\tilde{n}(t)$ is the noise component at the systems output. The reference for gain measurement was the input signal at the first sensor. The reference $\acute{s}(t)$ for the speech distortion measurement was the output of a perfectly steered delay-and-sum beamformer. The parameters a and δ were chosen to compensate the amplitude and delay differences.

Figure 2 shows the simulation results as a function of room reverberation time T_{60}. We achieved a good suppression of the interfering sources in low reverberation conditions and at moderate noise levels. At high noise levels (0 dB) we observe a significant drop in SIR gain. A detailed analysis of this behavior uncovers that this is caused by wrong VAD decisions. As expected, separation performance decreases for higher reverberation times where the multiplicative transfer function approximation (2) and the sparse source assumption (4) become less accurate. Noise suppression is very large at low reverberation times and surprisingly good even at high reverberation times.

Speech quality evaluation shows good results in low reverberation conditions. At high reverberation time the SDR measurement has to be viewed with caution since a fair quantitative comparison especially in reverberant environment is difficult. In hearing tests speech quality was considered to be good.

4. CONCLUSION

We introduced a new statistical model for blind speech separation, from which dominant source detection in each time frequency bin and mixing system identification algorithms are obtained as an instance of the Expectation Maximization algorithm. We confirmed by simulations that the approach works well in low to medium reverberation environments. The good results encourage us to further explore the potentials of directional statistics for approaching source separation and beamforming problems.

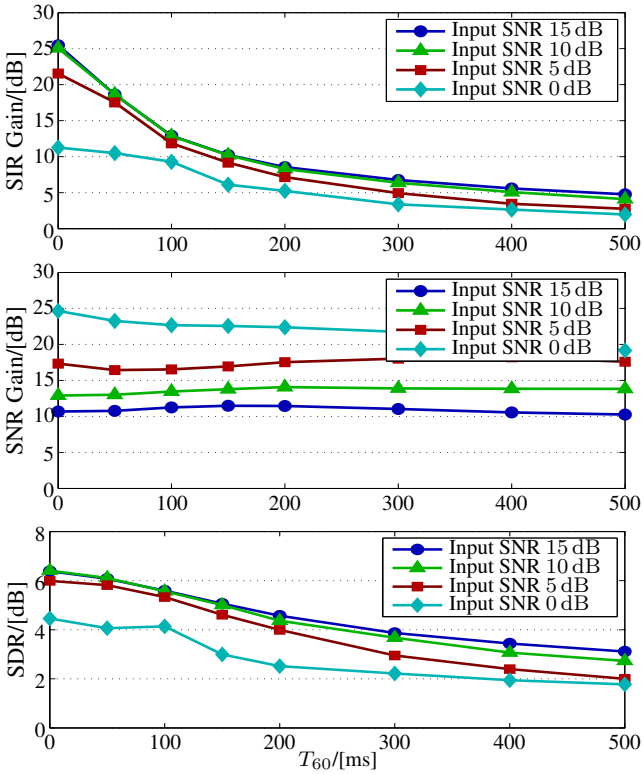

Fig. 2. *SIR Gain, SNR Gain and SDR for 3 sources and 8 sensors at different reverberation times and input coherent noise levels.*

5. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, 2007.

[3] D. H. Tran Vu, A. Krueger, and R. Haeb-Umbach, "Generalized eigenvector blind speech separation under coherent noise in a gsc configuration," in *Proc. IWAENC*, 2008.

[4] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[5] K. V. Mardia and I. L. Dryden, "The complex watson distribution and shape analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 4, 1999.

[6] K. V. Mardia and P. E. Jupp, *Directional Statistics*, Wiley, 1999.

[7] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 5, 2007.

[8] E. Warsitz, A. Krueger, and R. Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," in *Proc. ICASSP*, 2008.

[9] M. Lincoln, "Speech separation challenge part ii," *http://homepages.inf.ed.ac.uk/mlincol1/SSC2/*.