BLIND SOURCE EXTRACTION FOR A COMBINED FIXED AND WIRELESS SENSOR NETWORK

Brian Bloemendal[†] *Jakob van de Laar*[⋆] *Piet Sommen*[†]

b.bloemendal@tue.nl

[†]Eindhoven University of Technology,
Department of Electrical Engineering,
Eindhoven, The Netherlands

[⋆]Philips Research Laboratories,
Digital Signal Processing Group,
Eindhoven, The Netherlands

ABSTRACT

The emergence of wireless microphones in everyday life creates opportunities to exploit spatial diversity when using fixed microphone arrays combined with these wireless microphones. Traditional array signal processing (ASP) techniques are not suitable for such a scenario since the locations of the wireless sensors are unknown and probably vary over time.

In this paper we investigate the use of blind source extraction (BSE) techniques in such a combined acoustic sensor network to perform speech enhancement. We present strategies that apply traditional ASP techniques to the fixed microphone array, while simultaneously the spatial diversity provided by the wireless microphones is exploited. Our conclusion is that BSE techniques can be used in a combined wireless and fixed microphone network to perform speech enhancement.

Index Terms— Blind source extraction, second order statistics, wireless acoustic sensor network, post processing

1. INTRODUCTION

An important ASP task is the enhancement of a desired (speech) signal. Nowadays, the desired speech is more and more contaminated by interfering sources, e.g., speech or music. Using a wireless acoustic sensor network (WASN) consisting of a fixed microphone array combined with wireless microphones enables the use of the wireless microphones to increase the sound quality produced by the fixed microphone array. An example application is a consumer VoIP system. These systems typically use a relatively small microphone array consisting of two to four sensors and may be equipped with a wireless transceiver to obtain data from wireless microphones like mobile phones and tablets.

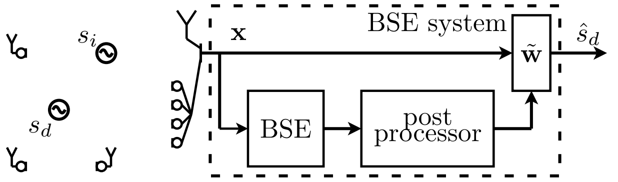


Fig. 1. Model of the blind source extraction (BSE) system for a combined fixed and wireless sensor network. As an example, one desired s_d and one interfering s_i source is depicted.

In a combined WASN, for which a model is depicted in Figure 1, it is impossible to use traditional beamforming techniques to enhance the target signal using all microphones due to unknown and varying sensor positions and a relatively large spacing. Therefore, more advanced ASP algorithms are required. In the literature, the multichannel Wiener filter [1], blind-LCMV (linear constraint minimum variance) beamforming [1] and a generalized sidelobe canceler implementation for the blind-LCMV beamforming algorithms [2] are proposed. These blind methods do not require prior knowledge about the locations of the sensors; however, they require exact knowledge about the activity of the desired source and the interferers. For a multiple speaker environment it is hard to build a reliable detector that indicates the noise-only and noise-and-interference-only periods. Additionally, the occurrences of these periods may become rare in practice, e.g., in a highly occupied living room.

In [3] a blind source extraction (BSE) method is proposed that extracts a desired signal from a complex mixture of spatially correlated signals. In order to identify the desired extraction filter, this method requires a rough guess of the mixing column of the desired source for a subset of the microphones. Applied to beamformer design, this type of prior information was shown to translate to a guess of the direction of arrival (DOA) of the desired source with respect to a few microphones. We expect to have this knowledge available in many practical situations. In the example VoIP application we may expect that the desired speaker is located more in front of the fixed microphone array than interfering speakers.

The BSE method in [3] has some nice properties that lead to a good performance for a combined fixed and wireless sensor network. First, a mismatch in the estimated DOA is compensated for by the algorithm and the extraction filter does not depend on this error. Second, the extraction filter is only data dependent, which means that not all sensor positions have to be known. Third, the method works for situations where no silent periods of the desired source are available.

In this paper we apply the method from [3] to a combined fixed and wireless acoustic sensor network. We present strategies to handle the following situations with a post processor. First, acoustic sources typically do not occupy the full frequency band, which also holds for the desired source. Because the desired source can be extracted only when it is active a detection mechanism is required for each frequency bin. Second, if the mixing system is ill-conditioned, constraints on the suppression of interference may lead to an undesired gain of spatially uncorrelated noise; therefore, it may be preferable in practice to apply a minimum variance distortionless response (MVDR) approach. We show that the conditioning of the mixing system can be measured and that both filters can be obtained from the BSE algorithm with a similar method. Finally, extraction filters can be identified up to an unknown complex scale per frequency bin. We present strategies to deal with this scaling problem by using only information that was already used by the BSE algorithm. To obtain insight in the BSE problem, we perform our analysis and simulations on mixtures that contain only a single delay and scaling.

The outline of this paper is as follows. In Section 2 we introduce our model and assumptions on the source and noise signals. In Section 3 we recap the procedure from [3]. In Section 4 we discuss the three problems addressed in this paper and present strategies for a post processor. Finally, in Section 5 we conclude this paper.

2. MODEL AND ASSUMPTIONS

We assume that mixtures of $S \geq 2$ source signals are observed by D sensors. The impulse responses from source to sensor are assumed to consist of a single weighted delay, i.e.,

$$x_i(t) = \sum_{j=1}^{S} a_i^j s_j(t - \tau_i^j) + \nu_i(t) \tag{1}$$

where $x_i(t)$ is the i'th sensor signal at time t, $s_j(t)$ is the signal of source j, $\nu_i(t)$ is noise at sensor i, and a_i^j and τ_i^j are the gain and delay between source j and sensor i. Additionally, we assume to have at least the same number of sensors as sources, i.e., $D \geq S$. Furthermore, we assume ideal wireless links and synchronized sensor nodes.

The transfer function from each source j to each sensor i is assumed to be constant in small frequency bands, which is a widely used and applied assumption [4, 5]. Therefore, by applying a DFT filterbank we obtain the following expression in the time-frequency domain:

$$x_i(n,m) = \sum_{j=1}^{S} h_i^j(m) \cdot s_j(n,m) + \nu_i(n,m) \tag{2}$$

where $n \in \mathbb{Z}$ is the discrete time index, $m \in \{0, M-1\}$ is the discrete frequency index, and $h_i^j(m) \approx a_i^j \, \mathrm{e}^{-\jmath 2\pi \frac{m}{M} \tau_i^j}$ is the transfer function from source j to sensor i for the m'th frequency band.

We can deal with noise in two ways. The first method is a well known and widely used method where the noise spectrum is measured in noise-only periods and subtracted from the observations. The second method exploits knowledge about a noise-free region of support (NF-ROS), which consists of a set of time-lag pairs (n,k) for frequency bin m where the sensor correlation data is noise-free. Each element is indicated by $\Omega_\kappa^m = \{(n,k)_\kappa, m\}$, where $\kappa \in [1, K_m]$ indicates the time-lag pair (n,k) number and $K_m \geq S$ is the number of time-lag pairs available for frequency bin m. In the NF-ROS the following conditions hold:

$$\begin{aligned} \mathbb{E}\{s_j[n,m]\bar{\nu}_i[n-k,m]\} &= 0 \quad \forall\, 1 \leq j \leq S, 1 \leq i \leq D \\ \mathbb{E}\{\nu_i[n,m]\bar{s}_j[n-k,m]\} &= 0 \quad \forall\, 1 \leq i \leq D, 1 \leq j \leq S \\ \mathbb{E}\{\nu_{i_1}[n,m]\bar{\nu}_{i_2}[n-k,m]\} &= 0 \quad \forall\, 1 \leq i_1, i_2 \leq D \end{aligned}$$

where $\mathbb{E}$ is the expectation operator and a bar denotes complex conjugation, e.g., $\bar{\nu}_i$.

For simplicity we keep using the symbol Ω to indicate noise-free correlation data, independent of the used method.

For mutually uncorrelated sources, the sensor correlation functions have a noise-free structure in the NF-ROS, i.e.,

$$r_{i_1 i_2}^x[\Omega_\kappa^m] = \sum_{j=1}^{S} h_{i_1}^j(m) \bar{h}_{i_2}^j(m) r_{jj}^s[\Omega_\kappa^m] \tag{3}$$

where $r_{i_1 i_2}^x[\Omega_\kappa^m] \triangleq \mathbb{E}\{x_{i_1}[n,m]\bar{x}_{i_2}[n-k,m]\}$ are sensor correlation functions and $r_{jj}^s[\Omega_\kappa^m] = \mathbb{E}\{s_j[n,m]\bar{s}_j[n-k,m]\}$ are source autocorrelation functions in the NF-ROS.

We exploit the structure in the sensor correlation data by constructing the following correlation matrices:

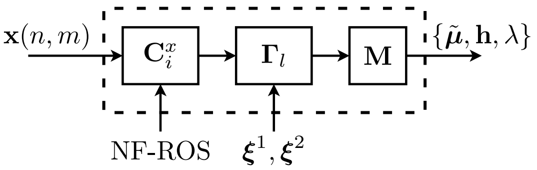
$$\mathbf{C}_i^x(m) \triangleq \begin{bmatrix} r_{i1}^x[\Omega_1^m] & \cdots & r_{i1}^x[\Omega_K^m] \\ r_{i2}^x[\Omega_1^m] & \cdots & r_{i2}^x[\Omega_K^m] \\ \vdots & \ddots & \ddots \\ r_{iD}^x[\Omega_1^m] & \ddots & r_{iD}^x[\Omega_K^m] \end{bmatrix} \quad \forall\, 1 \leq i \leq D \tag{4}$$

Using the following notation of a mixing matrix $\mathbf{H}(m)$:

$$\mathbf{H}(m) = \begin{bmatrix} \tilde{\mathbf{h}}_1(m) \\ \vdots \\ \tilde{\mathbf{h}}_D(m) \end{bmatrix} = \begin{bmatrix} h_1^1(m) & \cdots & h_1^S(m) \\ \vdots & \ddots & \ddots \\ h_D^1(m) & \ddots & h_D^S(m) \end{bmatrix} \tag{5}$$

where $\tilde{\mathbf{h}}_i(m)$ denotes the i'th row from matrix $\mathbf{H}(m)$, the structure of a sensor correlation matrix is given as follows:

$$\mathbf{C}_i^x(m) = \bar{\mathbf{H}}(m) \operatorname{diag}\left(\tilde{\mathbf{h}}_i(m)\right) \mathbf{C}^s(m) \tag{6}$$



$\mathbf{x}(n,m)$
$\mathbf{C}_i^x$
$\mathbf{\Gamma}_l$
$\mathbf{M}$
$\{\tilde{\boldsymbol{\mu}}, \mathbf{h}, \lambda\}$
NF-ROS
$\boldsymbol{\xi}^1, \boldsymbol{\xi}^2$

Fig. 2. BSE algorithm that uses the observations, NF-ROS, and vectors $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ to identify normalized extraction filter $\tilde{\boldsymbol{\mu}}$ and mixing column $\mathbf{h}$ as the left and right eigenvectors that correspond to the smallest eigenvalue λ of matrix $\mathbf{M}$.

where $\bar{\mathbf{H}}(m)$ is the conjugate of $\mathbf{H}(m)$, diag $(\cdot)$ puts the element of a vector on the diagonal of a matrix, and $\mathbf{C}^s(m)$ is an $S \times K$ source autocorrelation matrix with following structure:

$$\mathbf{C}^s(m) \triangleq \begin{bmatrix} r^s_{11}[\Omega_1^m] & \cdots & r^s_{11}[\Omega_K^m] \\ \vdots & \ddots & \ddots \\ r^s_{SS}[\Omega_1^m] & \ddots & r^s_{SS}[\Omega_K^m] \end{bmatrix} \quad (7)$$

We assume that this matrix is full rank, i.e., the source autocorrelation functions have to be linearly independent. Consequently, we are able to identify the number of active sources from the effective rank of the correlation matrices in (4) and apply subspace techniques to obtain $S \times K$ matrices $\hat{\mathbf{C}}_i^x$ [4,5].

In [3] a procedure is presented that uses the correlation matrix structure in (6) to identify the following desired extraction filter $\tilde{\boldsymbol{\mu}}(m)$:

$$\tilde{\boldsymbol{\mu}}(m)\mathbf{H}(m) = \alpha \tilde{\mathbf{e}}_d \quad (8)$$

where the vectors $\alpha \in \mathbb{C}$ is a non-zero scaling and $\tilde{\mathbf{e}}_d$ has a one at the index of the desired source and zeros elsewhere. From now on, we omit frequency index m for convenience.

3. SUMMARY OF THE BSE ALGORITHM

An overview of the BSE algorithm from [3] is depicted in Figure 2. First, two linear combinations $\boldsymbol{\xi}^1 \in \mathbb{C}^D$ and $\boldsymbol{\xi}^2 \in \mathbb{C}^D$ of the reduced size, noise-free correlation matrices are taken, which leads to the following matrices:

$$\mathbf{\Gamma}_l = \sum_{i=1}^{D} \xi_i^l \hat{\mathbf{C}}_i^x \quad \forall\, 1 \le l \le 2 \quad (9)$$

where $\boldsymbol{\xi}^l = \left[\xi_1^l, \cdots, \xi_D^l\right]^T$ are designed later. From these linear combinations the following matrix $\mathbf{M}$, which has a very specific eigenstructure, is constructed:

$$\mathbf{M} = \bar{\mathbf{\Gamma}}_2 \left(\mathbf{\Gamma}_1\right)^{\dagger} \mathbf{\Gamma}_2 \left(\bar{\mathbf{\Gamma}}_1\right)^{\dagger} \equiv \hat{\mathbf{H}} \mathbf{\Lambda} \hat{\mathbf{H}}^{-1} \quad (10)$$

where $(\cdot)^{\dagger}$ is a pseudo-inverse, $\mathbf{\Lambda}$ is a diagonal matrix, and $\hat{\mathbf{H}}$ is the reduced mixing matrix.

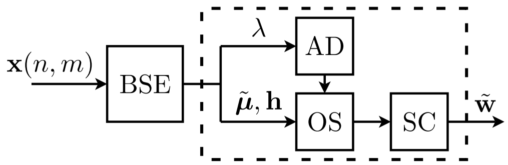


Fig. 3. Overview of the post processor where the eigenvalues from the BSE stage are used for desired source activity detection (AD). The AD result and the left and right eigenvectors are used by the objective selection (OS) algorithm. Finally, the scaling (SC) module scales the desired filter.

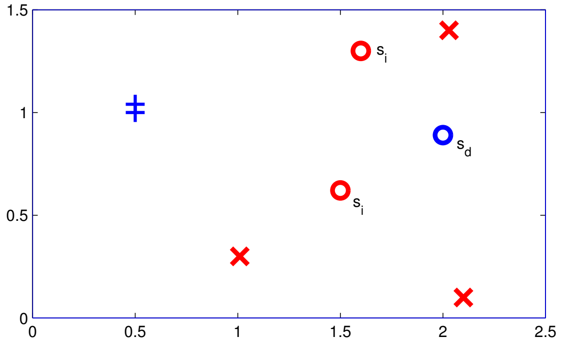


Fig. 4. Overview of the reflection-free room of size 2.5 by 1.5 meters. Desired source s_d and interferences s_i are indicated with circles, fixed microphones with plus signs, and wireless microphones with cross signs. The fixed microphones have a mutual spacing of 0.04 meters.

The equivalence in (10) follows from (6) and (9) and from the equivalence it follows that the left and right eigenvectors of the matrix $\mathbf{M}$ form the rows from the inverse of the mixing system and the columns of the mixing system, respectively. The corresponding eigenvalues, i.e., the elements on the diagonal of $\mathbf{\Lambda}$, have the following structure:

$$\lambda^j = \left| \frac{\left\langle \boldsymbol{\xi}^2, \mathbf{h}^j \right\rangle}{\left\langle \boldsymbol{\xi}^1, \mathbf{h}^j \right\rangle} \right|^2 \quad \forall\, 1 \le j \le S \quad (11)$$

with $\langle \cdot, \cdot \rangle$ the Euclidean inner product. From (11) it follows that each eigenvalue depends only on the vectors $\boldsymbol{\xi}^1$, $\boldsymbol{\xi}^2$, and the mixing column of a single source. From [3] we know that $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ should be designed as smooth beamformers, using a guess of the direction of arrival (DOA) of the desired source.

4. BSE POST PROCESSING STRATEGIES

Before the extraction filters $\tilde{\boldsymbol{\mu}}(m)$ can be used to enhance the desired signal, a post processor, as in Figure 3, is required. The scenario from Figure 4 is used to present our strategies.

Three equal power sources were generated by filtering white Gaussian signals by a feedback comb filter of order 128 with feedback coefficients -0.9, -0.3, and -0.7, respectively.

The observations were generated by filtering the source signals with the respective weighted delays for a sampling frequency of 8 kHz. We added white Gaussian noise such that the desired signal to noise ratio at the fixed sensors is 14 dB. We used these signals in order to be independent to estimation errors and to visualize the concept.

The 1,280,000 observed samples were filtered by a 128-band DFT filterbank, which lead to 20,000 samples per band. The BSE algorithm was initialized with a NF-ROS of lag 1 up to 10 for each band. The DOA of the desired source s_d was estimated as zero degrees, which is slightly (5°) wrong. Notice that the eigenvalues of $\mathbf{M}$ are the ratios of the beampatters of the beamformers formed by $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$, per DOA of each source [3]. Therefore, $\boldsymbol{\xi}^1$ is chosen as a delay and sum beamformer for the estimated DOA w.r.t. the fixed sensors, i.e., $\boldsymbol{\xi}^1 = \begin{bmatrix}1 & 1 & 0 & 0 & 0\end{bmatrix}^T$. Similarly, $\boldsymbol{\xi}^2$ is chosen as a delay and subtract beamformer, i.e., $\boldsymbol{\xi}^2 = \begin{bmatrix}1 & -1 & 0 & 0 & 0\end{bmatrix}^T$.

4.1. Desired source activity detection

Prior to the BSE algorithm we apply an algorithm that estimates the number of active sources from the effective rank of the correlation data [5]. The desired extraction filter is identified even if certain interferers are active in only a few frequency bands. However, if the desired source is not active in every frequency bin, then the extraction filter for an interferer is identified. Therefore, we require an activity detector (AD) that classifies if the desired source is active.

The eigenvalues of the matrix $\mathbf{M}$ are a function of the mixing columns of the active sources and the vectors $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ as in (11) and can be used for activity detection. The eigenvalues measure the Hermitian angle θ^j [6] between mixing column j and the vector $\boldsymbol{\xi}^1$ according to $|\lambda^j| = (\tan\theta^j)^2$. In Figure 5 we depict estimated Hermitian angles $\hat{\theta}^j$ using the following transformation of all eigenvalues of matrix $\mathbf{M}$: $\hat{\theta}^j = \arctan(\sqrt{|\lambda^j|})$. Additionally, we depict the expected Hermitian angle for a DOA of 10 degrees.

We observe that there is a linear trend over frequencies when starting from frequency zero. The outliers at the highest frequencies are due to a limited bandwidth of the delay filters in the mixing system. Selecting the smallest eigenvalue would lead to a selection of the desired filter if the desired source is active. Alternatively, if the desired source is not active in a frequency bin, the estimated number of sources in that bin reduces and the smallest eigenvalue has a relatively large value with respect to its neighbors.

Based on this result we propose two strategies to detect if the desired source is active. First, if identification of only the smallest eigenvalue of the matrix $\mathbf{M}$ is desired we could use a threshold to detect if the desired source is active. Alternatively, if all eigenvalues of the matrix $\mathbf{M}$ are identified, then we could train a classifier to detect if the desired source is active in a certain frequency bin. In both strategies knowledge about the estimated number of active sources could be included as well.

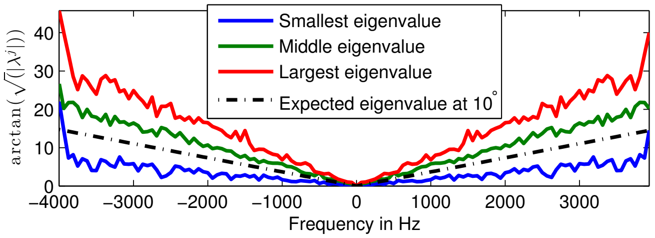


Fig. 5. Plot of the ordered and transformed eigenvalues in solid lines and a threshold on a ±10 degree angle with respect to the vector $\boldsymbol{\xi}^1$ in a black dashed line.

4.2. Objective selection

From the structure in (10) it follows that the left eigenvectors are rows from the inverse of the mixing system and the right eigenvectors are columns from the mixing system. The left eigenvectors can be used as extraction filters and extended to the LCMV solution that requires zeros for undesired sources, while the right eigenvectors can be used to extend to the MVDR, which has no specific constraints on the interfering sources, i.e., we can minimize the output power in the space orthogonal to the right eigenvector.

In practice, the preferred objective, i.e., MVDR or LCMV, depends on the application. One of the considerations is the sound quality that is produced by the LCMV filter, which depends on the conditioning of the mixing system. If the projection of the mixing column of the desired source onto the extraction filter is very small, then the output of this extraction filter becomes very noise sensitive and the MVDR filter may be a more appropriate choice. This conditioning can be measured by the Hermitian angle θ_H [6] between the extraction filter and the mixing column of the desired source, i.e.,

$$\cos\theta_H = \frac{|\langle\tilde{\boldsymbol{\mu}},\mathbf{h}\rangle|}{||\tilde{\boldsymbol{\mu}}||\,||\mathbf{h}||} \quad \text{where} \quad \langle\tilde{\boldsymbol{\mu}},\mathbf{h}\rangle = \tilde{\boldsymbol{\mu}}\mathbf{h} \tag{12}$$

If this angle is small, the separation of the interference and desired source is well conditioned and the LCMV objective can be used. Otherwise, the conditioning is bad and the MVDR objective may be preferred. Choosing one of these objectives should be performed per frequency bin and leads to a system with a frequency dependent number of constraints.

In Figure 6 we compare the observed and actual Hermitian angle between the mixing column and extraction filter of the desired source for the scenario from Figure 4. We observe that the estimated angle follows the Hermitian angle of the actual vectors. The oscillating behavior of the angle, i.e., the conditioning of the mixing matrix, per frequency bin is due to spatial aliasing from the relative large spacing between the microphones. If an array of only fixed microphones with relative small spacing were used, a more fluent conditioning

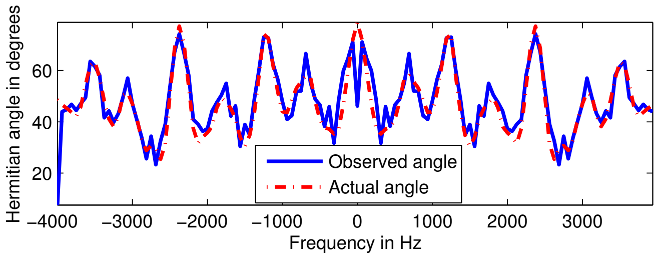


Fig. 6. Observed and actual Hermitian angle between the mixing column and the extraction filter. This parameter per frequency bin can be used to select the desired filter.

would be observed; however, the overal conditioning of the system would be poorer, especially for lower frequencies.

4.3. Scaling

The extraction filter and mixing column vector can be identified up to an unknown complex scaling per frequency bin, which leads to an undesired and unknown filtering of the desired source. This scaling problem can be decomposed into a local and global scaling problem. The global scaling means that we don't know the overall gain of the system, which can be easily solved by normalizing the output power. The local scaling problem, i.e., scaling per frequency bin, can be solved using different strategies. If the processing is completely blind, two straightforward strategies are to fix one of the extraction filter coefficients for each frequency bin or to restrict the norm of each extraction filter per frequency bin. Both these strategies typically result in a filtered version of the desired signal.

Alternatively, if additional knowledge is available we can use it to improve the scaling. In the scenario from Figure 4, the array response vector of the fixed sensor array for the estimated DOA, which is the vector $\boldsymbol{\xi}^1$, can be used to scale the extraction filter. This procedure leads to the following scaled mixing columns $\hat{\mathbf{h}}$ and extraction filters $\hat{\tilde{\boldsymbol{\mu}}}$:

$$\hat{\mathbf{h}} = \frac{\mathbf{h}}{(\boldsymbol{\xi}^1)^H \mathbf{h}} \quad \text{and} \quad \hat{\tilde{\boldsymbol{\mu}}} = \frac{(\boldsymbol{\xi}^1)^H \mathbf{h}}{\tilde{\boldsymbol{\mu}}\mathbf{h}} \tilde{\boldsymbol{\mu}} \tag{13}$$

In Figure 7 we depict the spectra of the impulse responses from the sources to the output of the extraction system for extraction filter $\hat{\tilde{\boldsymbol{\mu}}}$. We observe that the desired source has the highest gain with respect to the interferences with a relatively flat response for all frequencies, which means that the desired source is extracted successfully.

5. CONCLUSIONS

We presented strategies to enhance a desired signal for a combined fixed and wireless acoustic sensor network. Future research topics include a performance analysis of the estimated parameters, application of the algorithm to real speech signals, and the development of classification algorithms.

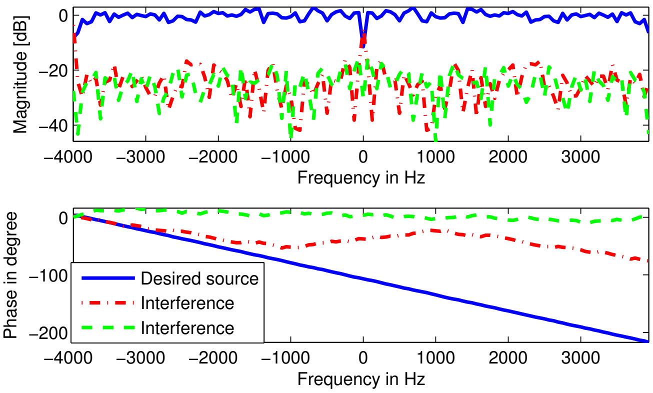


Fig. 7. Magnitude and phase of the transfer functions from the sources to the output of the BSE system. The scaled extraction filter with the LCMV objective is used for all frequencies and the global scaling is normalized.

6. REFERENCES

[1] Alexander Bertrand, *Signal Processing Algorithms for Wireless Acoustic Sensor Networks*, Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium, May 2011.

[2] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1071 –1086, aug. 2009.

[3] B.B.A.J. Bloemendal, J. van de Laar, and P.C.W. Sommen, "Beamformer design exploiting blind source extraction techniques," in *Proceedings of ICASSP 2012*, 2012.

[4] Andrzej Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley & Sons, Inc., New York, NY, USA, 2002.

[5] Pierre Comon and Christian Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.

[6] K. Scharnhorst, "Angles in complex vector spaces," *Acta Applicandae Mathematicae*, vol. 69, pp. 95–103, 2001, 10.1023/A:1012692601098.