Semi-Blind Noise Extraction Using Partially Known Position of the Target Source

Zbyněk Koldovský, Jiří Málek, Petr Tichavský, and Francesco Nesta

***Abstract*—An extracted noise signal provides important information for subsequent enhancement of a target signal. When the target's position is fixed, the noise extractor could be a target-cancellation filter derived in a noise-free situation. In this paper we consider a situation when such cancellation filters are prepared for a set of several possible positions of the target in advance. The set of filters is interpreted as prior information available for the noise extraction when the target's exact position is unknown. Our novel method looks for a linear combination of the prepared filters via Independent Component Analysis. The method yields a filter that has a better cancellation performance than the individual filters or filters based on a minimum variance principle. The method is tested in a highly noisy and reverberant real-world environment with moving target source and interferers. A post-processing by Wiener filter using the noise signal extracted by the method is able to improve signal-to-noise ratio of the target by up to 8 dB.**

***Index Terms*—Independent component analysis (ICA), noise extraction, audio source separation, supervised localization, generalized sidelobe canceler (GSC).**

I. Introduction

SPEECH enhancement is a field that comprises a large number of methods designed to remove unwanted signals from speech [1]. Using multiple microphones became popular, because spatial information can be used to extract noise signals providing important information for subsequent enhancement of a target signal. For example, a popular beamformer called Generalized Sidelobe Canceler (GSC) consists of three building blocks, one of which is called the blocking matrix (BM) [2]. This block is designed to cancel the target and only pass through noise signals. The ability to extract noise signals is essential for the final performance of beamformers or other post-filtering approaches [3].

Manuscript received November 27, 2012; revised February 28, 2013 and May 06, 2013; accepted May 07, 2013. Date of publication May 22, 2013; date of current version July 22, 2013. This work was supported by Grant Agency of the Czech Republic through the project P103/11/1947. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emmanuel Vincent.

Z. Koldovský is with the Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, 461 17 Liberec, Czech Republic, and also with the Institute of Information Theory and Automation, 182 08 Prague 8, Czech Republic (e-mail: zbynek.koldovsky).

J. Málek is with the Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, 461 17 Liberec, Czech Republic (e-mail: jiri.malek@tul.cz).

P. Tichavský is with the Institute of Information Theory and Automation, 182 08 Prague 8, Czech Republic. E-mail: tichavsk@utia.cas.cz.

F. Nesta is with the Fondazione Bruno Kessler-Irst, 38123 Trent, Italy (e-mail: francesco.nesta@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2264674

In pioneering beamforming methods [4]–[6], the sound is assumed to propagate without any reflections, so only pure delays are taken into account. This model is useful in anechoic chambers where the reverberation time is very short, or when the target is sufficiently close to microphones so that the direct-to-reverberation ratio is high. In real-world environments such as a typical room in a house, the methods fail. The key problem is leakage of the target signal through the noise extractor, which is responsible for a distortion at the final output.

More recent methods take reverberation into account. For example, Gannot *et al.* [7] proposed a variant of GSC which aims to retrieve the responses (images) of the target on microphones (dereverberation is not the goal). The BM is constructed using a priori known transfer function ratios (TFRs) that are used to cancel the target at the BM output[1]. In other words, the BM is realized using target cancellation filters (CFs) defined through the known TFRs. The same or similar principles are also used in other methods; see e.g., [8]–[11].

Consequently, the key need is to acquire the CFs. For a fixed position of the target, they can be estimated from noise-free recordings of the target. The estimation can be done in the time-domain using the method of least squares. TFRs can be estimated in the frequency domain through estimating spectra and cross-spectra of signals [12]. In methods that operate in the short-time Fourier transform (STFT) domain, the latter approach is more desirable. Unbiased estimation of TFRs is possible in the presence of diffusive stationary noise [7], [13]. Other variants of such estimation were proposed in [8], [14], [15].

A problem arises when the noise is directive and nonstationary (e.g., there are other interfering speakers). CFs cannot then be updated from measured signals, and the cancellation relies on the position of the target remaining the same as the one for which the CFs were computed. However, in a simple experiment we show that even small movements of the target can cause target leakage through CFs, especially when the distance between the target and the microphones is far (say, more than 1.5 m) so that the direct-to-reverberation ratio becomes low. The need is to update the CF even under highly nonstationary conditions: for example, when there are moving interferers that are closer to microphones than the target.

To estimate CFs under general conditions, it is possible to use Blind Source Separation (BSS) methods [16]–[19], which do not need prior information about the scenario. However, there are two main drawbacks. First, the efficiency of BSS methods

[1]Hence, it is not necessary to know the transfer functions (the impulse responses) from the target to the microphones, which are hardly available in practice.

is limited [20], especially when sources are distant from microphones. Second, BSS methods have inherent permutation ambiguity for which any general solution does not exist. The higher complexity of BSS approaches also cannot be overlooked.

In this paper, we develop a simpler method that is able to cope with the aforementioned difficult situations. It can be categorized as *semi-blind*: It is based on the assumption that the location of a target is limited to a specific area and, for several points within this area, CFs are already known. Using this so-called *Cancellation Filter Bank* (CFB), the key task is to design a proper CF at any moment, that is, for any position of the target within the area, and in the presence of stationary as well as nonstationary and moving interferers.

In [21], Independent Component Analysis (ICA) is used to obtain the CF as a linear combination of CFs in the CFB such that its output is as independent of the target as possible. Here, we show that the linear combination of CFs cancels the target better than individual CFs, when the target's position is *not the same* as any of the positions for which the CFs are available (we will refer to these positions as to the *known positions*). We propose further improvements to this method, one of which is to have it detect whether the position of the target fits any of the known positions for which the particular CF from CFB is selected. As a byproduct, this will lead to a precise (supervised) localization of the target, which will proceed in a similar fashion as in [22]–[25] but also under noisy conditions.

This paper is organized as follows. Section II introduces some formalizations and describes the scenario and dataset considered throughout this paper. Section III is devoted to definitions of CFs and their estimation in time-domain using least squares. It is demonstrated that the CFs can be sensitive to small movements of the target. In Section IV, we propose two methods for the design of CFs using the CFB under general conditions. The first method is a straightforward approach based on the minimum output variance principle. The second method is the semi-blind approach using ICA. Section V shows results of several experiments with the dataset of real-world recordings. The enhancement of the target signal obtained by the proposed method is compared to the performance of a semi-blind frequency domain source separation method derived from [19].

II. The Problem Statement

A two-microphone recording of a target source, during which its position is fixed, is described by

$$\begin{aligned} x_{\mathrm{L}}(n) &= \{h_{\mathrm{L}} * s\}(n) + y_{\mathrm{L}}(n), \\ x_{\mathrm{R}}(n) &= \{h_{\mathrm{R}} * s\}(n) + y_{\mathrm{R}}(n) \end{aligned} \tag{1}$$

where n is the time index, $*$ denotes the convolution, $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$ are, respectively, the signals from the left and right microphones, $s(n)$ is the target signal, and $y_{\mathrm{L}}(n)$ and $y_{\mathrm{R}}(n)$ are noise signals (interferers). The noise signals can correspond to multiple sources but they are assumed to be independent of $s(n)$. $h_{\mathrm{L}}(n)$ and $h_{\mathrm{R}}(n)$ denote the microphone-target impulse responses that depend on the position of the target and on the acoustical properties of the environment. In this paper, we focus only on the two-microphone scenario due to its comparatively easy accessibility [26]. The concept, however, may be generalized to more microphones.

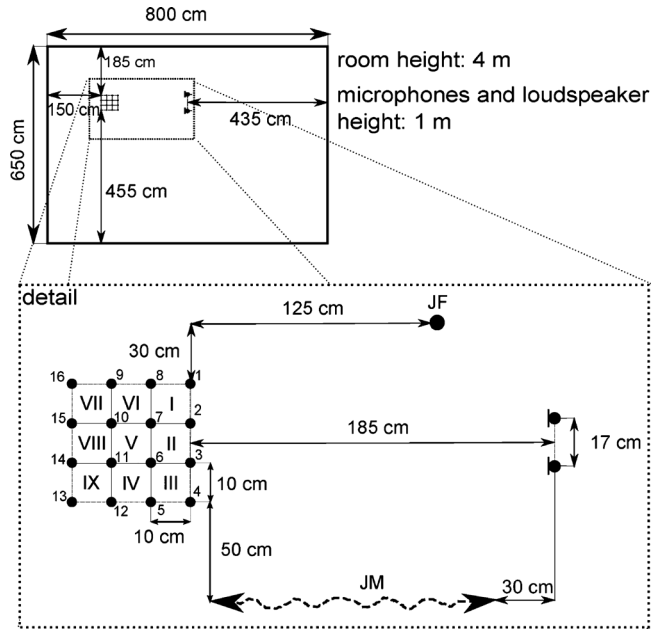

Fig. 1. Setup of the room which was utilized in our experiments. The known positions of the target source, located in a regular grid, are numbered from 1 to 16. The roman numerals correspond to gCFs for the groups of four adjacent positions. For example, the gCF IV is defined for positions 5, 6, 11 and 12.

A. Scenario

Throughout this paper we will consider a situation where a target speaker is recorded by two distant microphones, in an ordinary room that has natural acoustical properties (reverberation). The location of the speaker is limited to a small area, for example, such as in a meeting situation where the speaker, who is seated, makes limited movements with the head. In general, the goal is to enhance the noisy speech of the speaker when its position within the area is not exactly known and can even vary from one position to another.

We assume that there are I known positions within the speaker's area from which its noise-free recordings were obtained in advance. Each such recording can be described via (1) where $y_{\mathrm{L}} = y_{\mathrm{R}} = 0$. The recordings are used to estimate CFs for the known positions, as described in the next section.

B. Dataset

A particular scenario where we recorded our data for demonstrations and experiments described in this paper is illustrated in Fig. 1. The situation is challenging as it takes place in a meeting room with the reverberation time T_{60} of about 650 ms. The target's position is limited to a 30 cm $\times$ 30 cm area whose center is at a 2 m distance from microphones. We consider static as well as moving interferer that is closer to the microphones than to the target. In this paper, we omit situations with close speakers and minor reverberation, because the a priori knowledge of the proposed method (the known CFB) is comparatively strong. Its applications in simple scenarios are therefore less interesting.

As the target's noise-free signals, three male and three female utterances each of 4 s length, were played over a loudspeaker from each of $I = 16$ positions that form a regular grid with 10

cm spacing within the target area. The responses of the signals were recorded by two microphones[2] directed towards the center of the grid. All recordings were sampled at the frequency of 44.1 kHz, but then the signals were downsampled to 16 kHz. The utterances were taken from the TIMIT database.

Selected utterances of two different speakers from TIMIT were used as noise signals. The utterances were recorded from the fixed position, marked as JF in Fig. 1, and from the dynamic position on the right-hand side of the target speaker, marked as JM. We achieved the latter position by moving the loudspeaker during the playback along the path sketched in the figure. In all cases, the loudspeaker was situated perpendicular to the wall behind the microphones.

Later, in Section V, we also add multisource background noise and white Gaussian noise to the mixtures of signals in order to approach the real-world environment as much as possible.

III. Cancellation Filters

As pointed out in our Introduction, the blocking matrix can be realized using an efficient cancellation filter (CF) selected specifically for the target's position. According to (1), an ideal CF for a fixed target's position generally consists of two SISO filters g_{L} and $-g_{\mathrm{R}}$ that satisfy

$$g_{\mathrm{L}} * h_{\mathrm{L}} = g_{\mathrm{R}} * h_{\mathrm{R}} \tag{2}$$

(we will omit the time index n if it is not necessary). Then, the filter output

$$\begin{aligned} z = g_{\mathrm{L}} * x_{\mathrm{L}} - g_{\mathrm{R}} * x_{\mathrm{R}} = g_{\mathrm{L}} * h_{\mathrm{L}} * s + g_{\mathrm{L}} * y_{\mathrm{L}} \\ - g_{\mathrm{R}} * h_{\mathrm{R}} * s - g_{\mathrm{R}} * y_{\mathrm{R}} = g_{\mathrm{L}} * y_{\mathrm{L}} - g_{\mathrm{R}} * y_{\mathrm{R}} \end{aligned} \tag{3}$$

does not contain the contribution of s, so the passed signal z provides information about the noise signals y_{L} and y_{R}. Its further exploitation is briefly discussed in Section V together with the AIC part of the GSC beamformer.

A special case is when g_{R} is put equal to the unit impulse δ (or a delayed δ due to the causality) and $g_{\mathrm{L}} = h_{\mathrm{L}}^{-1} * h_{\mathrm{R}}$ where h_{L}^{-1} denotes the inverse filter of h_{L}. In the frequency domain, $h_{\mathrm{L}}^{-1} * h_{\mathrm{R}}$ corresponds with TFR, that is, the ratio of Fourier transforms of h_{R} and h_{L}, hence the relation to [7]. The choice $g_{\mathrm{R}} = \delta$ is also related to the Equalization-Cancellation binaural hearing model [27]: the target signal on one microphone is equalized to have the same response as on the other microphone, and then the responses are subtracted [12]. The sources that propagate in ways different from the target are not canceled, therefore, the output of the CF provides a reference noise signal.

In this paper, we will follow the choice $g_{\mathrm{R}} = \delta$. Other options of (2) were studied, e.g., in [28].

A. Least-Squares Computation of CF From a Noise-Free Recording

For now, let x_{L} and x_{R} be the noise-free recordings of the target signal that is located in a fixed, known position. The filter g_{L}, denoted simply by g, can be designed through the minimization

$$g = \arg\min_{g_0} \sum_{n=1}^{N} |\{g_0 * x_{\mathrm{L}}\}(n) - x_{\mathrm{R}}(n-d)|^2 \tag{4}$$

where d is a short integer delay introduced for the case that the target signal reaches the right microphone earlier than the left one. We will call g the cancellation filter, although the CF is the MISO filter comprised of the two SISO filters g and $-\delta(n-d)$. The position associated with the cancellation filter will be called the *CF position*.

The least-squares problem in (4) leads to a Toeplitz system of L linear equations; L corresponds with the chosen length of g. The system can be solved effectively by the Levinson-Durbin algorithm in $\mathcal{O}(L^2)$ operations [34]. The approach is computationally more demanding than the frequency-domain estimation of g through TFRs [13]; nevertheless, our method computes the CFs in advance so the computational burden due to the solution of (4) is immaterial.

B. CF for Groups of Positions

Let x_{L}^i and x_{R}^i denote noise-free recordings of the target located at the ith position, $i \in \mathcal{I}$, where $\mathcal{I}$ is a set of indices. We define the so-called *group cancellation filter* (gCF) for the set of positions $\mathcal{I}$ as the one that minimizes

$$g = \arg\min_{g_0} \sum_{i\in\mathcal{I}} \sum_{n=1}^{N} \left|\left\{g_0 * x_{\mathrm{L}}^i\right\}(n) - x_{\mathrm{R}}^i(n-d)\right|^2. \tag{5}$$

Mathematically, the optimization problems (4) and (5) are equivalent. The difference between the gCF and an ordinary CF is that the former simultaneously cancels sources coming from all positions in $\mathcal{I}$, so it might be less sensitive to small movements of the target. On the other hand, its cancellation performance for a particular position in $\mathcal{I}$ cannot be higher than that of the CF for the same position.

C. Sensitivity to Small Movements

The CFs were derived for positions $1, \ldots, 16$ and gCFs for groups of positions $\mathrm{I}, \ldots, \mathrm{IX}$ defined in Fig. 1. Each filter, of length $L = 3000$, was computed using the recording of the first male utterance of 4 s in length. Then, we examined the dependence of their cancellation performance on the change of the target's position. The filters were applied to the female utterances played from each of 16 positions. We chose testing signals different from the learning data to avoid any overlearning effects. Figs. 2 and 3 show, respectively, the average residual variances of the CFs' and gCFs' outputs.

The residual variance reflects the degree of the target cancellation. The better the cancellation, the smaller the variance. Fig. 2 shows that the cancellation is significantly better when the target's position corresponds to the CF position (the main diagonal in Fig. 2), while it drops by about 10 dB when the position is different (even neighboring). For example, when taking the CF computed for position 1 but playing the signal from position 2, the CF cancels the target only by 3.1 dB (relative to the average variance of recordings which is -36.4 dB) while the

[2]We use RØ DE™ microphones NT55 with cardioid capsules. The audio sound card is EDIROL FA-101.

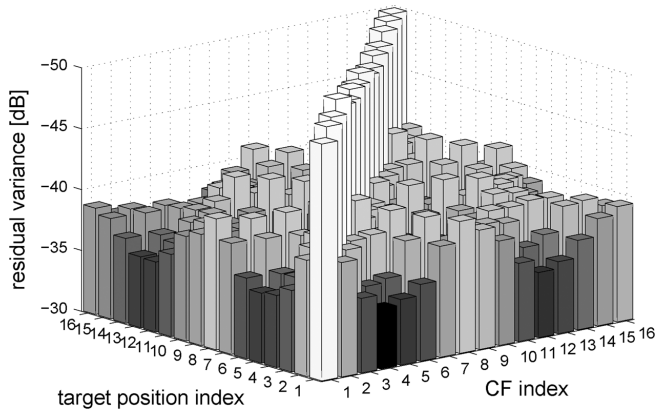

Fig. 2. Residual variances of CFs' outputs when the female target speaker utters from positions $1, \ldots, 16$ under noise-free conditions. The main diagonal corresponds to the case when the CF corresponds with the target position, which yields the best position.

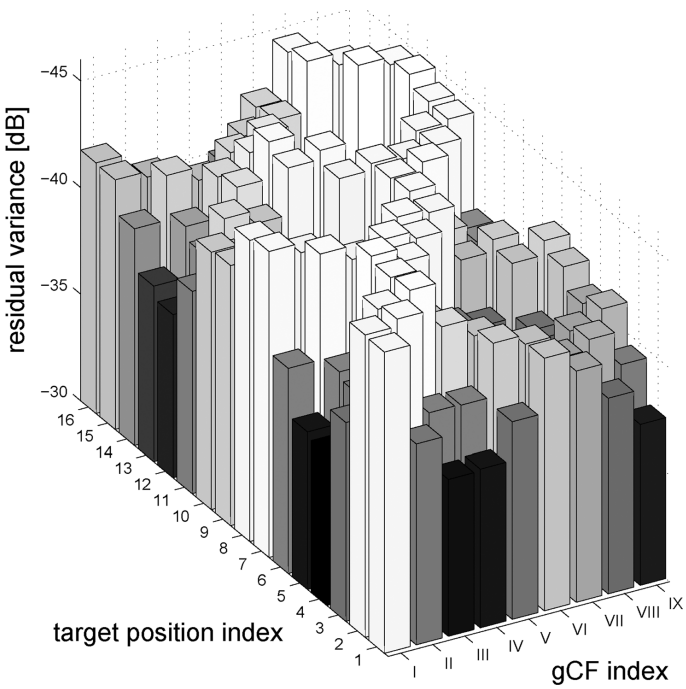

Fig. 3. Residual variances of gCFs' outputs when the female target speaker utters from positions $1, \ldots, 16$. For example, it is seen that the gCF I yields the best cancellation performance for positions 1, 2, 7, and 8, which corresponds to the filter definition in Fig. 1.

CF for position 2 cancels it by 13.8 dB. This illustrates the sensitivity of CFs to small movements of the target in a real-world environment.

The output variances of gCFs shown in Fig. 3 behave differently. They are approximately the same for the group of four positions for which the gCF was computed. The drop of performance for the other positions is not so dramatic (by about 5 dB). On the other hand, the best cancellation performance is lower by about 5 dB than that of the CF for the CF position. For example, gCF I and II cancel the woman's voice from position 2, respectively, by 7.7 dB and 8.1 dB, while the CF cancels it by 13.8 dB. This is the price paid for the wider cancellation range[3].

[3]The phenomenon that gCFs cancel the target less than CFs can be used for detecting multiple active sources (e.g., cross-talk detection); see [29].

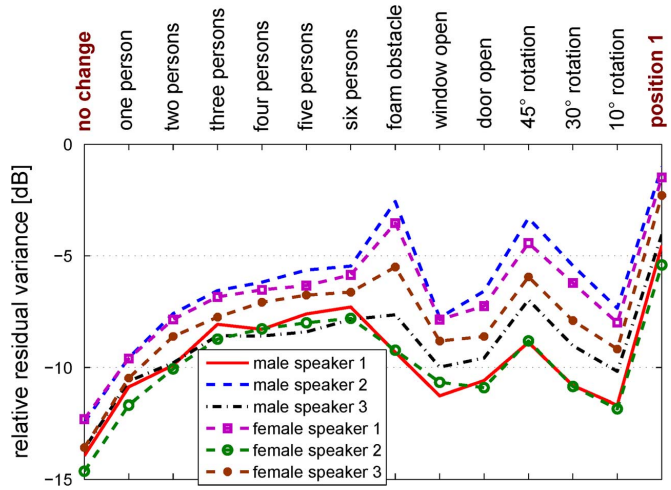

Fig. 4. Relative residual variances of the CF for position 2 and male speaker 1 (g_2) depending on the speaker and changes in the environment. Each residual variance is related to the variance of the input recording.

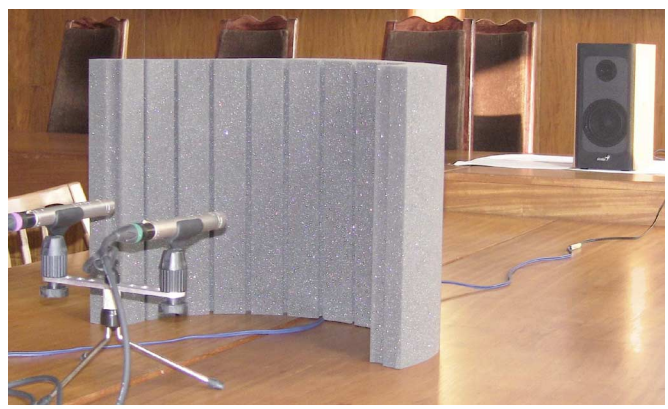

Fig. 5. Photo of the recording situation with the foam obstacle placed 40 cm in front of the microphones. The loudspeaker is located in position 2.

In this paper, we will consider both banks of the filters. Hence, the cancellation filter bank (CFB) contains 16 CFs denoted $g_1, \ldots, g_{16}$, while the group cancellation filter bank (gCFB) contains 9 gCFs denoted $g_{\mathrm{I}}, \ldots, g_{\mathrm{IX}}$. We use the filters computed from the first male utterance of length 4 s.

D. Sensitivity to Other Changes

The cancellation performance of a CF also depends on other changes in the environment. For illustration, we conducted an experiment where the loudspeaker was placed in position 2. Various changes in the room were made. Namely, several persons (1–6) were successively seated around the table with microphones, a door or window was opened, a foam obstacle was placed 40 cm in front of the microphones to attenuate the direct-path signal (see photo in Fig. 5), or the loudspeaker was rotated to the right. The utterances were recorded for each change, and the CF computed for the first male speaker in position 2 (without any changes, i.e., g_2) was applied to the recordings. The residual variances evaluated for different speakers are shown in Fig. 4.

Naturally, it holds that the greater the change, the greater the drop of the cancellation performance of g_2. For example, the performance declines gradually with the growing number of persons around the table. The change due to the foam obstacle blocking the direct-path signal is also significant. On the other hand, all the tested changes caused a smaller performance decrease than moving the target (loudspeaker) to position 1 did. We know already from Fig. 2 that the drop in performance when the target is in a different position than the CF is by about 10 dB. Position 1 is not exceptional. The results also indicate that the speaker's voice is less influential.

IV. Noise Extraction Using the Available CFB

In this section, we propose two approaches that are applicable as noise extractors, e.g., within the blocking matrix part of a GSC-type beamformer. Both approaches take advantage of the available bank of cancellation filters. The goal is to design a proper CF for any short interval when the position of the target is not precisely known (somewhere within the limited area) while noise is potentially present. It is assumed for simplicity that the target's position is fixed during that interval.

A. The Minimum Variance Approach

A straightforward approach selects the CF as the one filter from the CFB that provides the minimum variance on its output. We call this simple method the minimum variance approach (*MVA*). *MVA* is a competitor to the method proposed in the next section.

MVA relies on the assumption that a correctly canceled target source should have a minimum energy at the output even when noise is present. This assumption is reasonable provided that the energy of the target is significantly higher than the energy of noise.

1) Localization: On assumption that the CF is selected correctly, the CF position must agree with the true position of the target. *MVA* thus provides a simple method for localization for the target as a byproduct. The localization is *supervised* [25] in the sense that it relies on the a priori known CFB. Similar localization methods relying on prior knowledge of the acoustical environment are capable of localizing the source in 3D using only two microphones; see [22]–[24].

B. The Semi-Blind Approach

Now we describe the approach based on the use of ICA. Let $\mathbf{X}$ be a so-called *observation matrix* defined as

$$\mathbf{X} = \begin{bmatrix} \{g_1 * x_{\mathrm{L}}\}(N_1) & \dots & \{g_1 * x_{\mathrm{L}}\}(N_2) \\ \vdots & \vdots & \vdots \\ \{g_I * x_{\mathrm{L}}\}(N_1) & \dots & \{g_I * x_{\mathrm{L}}\}(N_2) \\ x_{\mathrm{R}}(N_1 - d) & \dots & x_{\mathrm{R}}(N_2 - d) \end{bmatrix} \tag{6}$$

where $\mathcal{I} = \{1, \dots, I\}$ is the set of indices of all CFs in the available CFB, and N_1 and N_2, respectively, denote the beginning and end of the interval of data.

The subspace spanned by rows of $\mathbf{X}$ contains outputs of all CFs in the CFB. For example, let

$$\mathbf{f}_k = [\underbrace{0, \dots, 0, 1}_{k}, 0, \dots, -1]^T, k = 1, \dots, I. \tag{7}$$

Then $\mathbf{f}_k^T \mathbf{X} = [\{g_k * x_{\mathrm{L}}\}(n) - x_{\mathrm{R}}(n - d)]_{n=N_1,\dots,N_2}$ is the output of the kth CF. It is therefore reasonable to scan the whole subspace to find the best linear combination of rows of $\mathbf{X}$ in terms of the target signal cancellation.

To find the linear combination or, equivalently, the corresponding CF or its output signal, we search for independent components (ICs) of $\mathbf{X}$, which is the idea first used in [21]. The key reason is that the noise is independent of the target. Hence, it can be expected that one such independent component corresponds to the noise or, in other words, to a residual signal in which the target is canceled as much as possible. A suitable ICA algorithm could be used by considering $\mathbf{X}$ as an instantaneous mixture $\mathbf{X} = \mathbf{AS}$ where $\mathbf{A}$ is a square mixing matrix and $\mathbf{S}$ is the matrix whose rows contains the ICs [35].

1) Noise Component Detection: Since the order of ICs is random, which is the inherent ambiguity of ICA, it must be determined which of them gives the best noise estimate. Let $\mathbf{W}$ be the estimated de-mixing matrix obtained by the ICA algorithm, that is, the estimate of $\mathbf{A}^{-1}$ up to the order of its rows. Let the scale of rows of $\mathbf{W}$ be such that all ICs have the same (unit) variance. We propose selecting the component according to the largest element (in absolute value) of the last column of $\mathbf{W}$.

To explain, note that the kth element of the last column of $\mathbf{W}$ determines to what extent the last row of $\mathbf{X}$ contributes to the kth component. The last row of $\mathbf{X}$ contains the signal from the right-hand microphone x_{R} while the other rows are filtered versions of x_{L}. The cancellation of the target signal is possible only if the diversity between x_{L} and x_{R} is exploited. Therefore, the last row of $\mathbf{X}$ must be "sufficiently" involved in the component in which the target is canceled.

2) Localization: Let $\mathbf{a}^T$ denote the row of $\mathbf{W}$ corresponding to the selected IC. In case the target is located in one of the known positions, $\mathbf{a}$ should be similar to $\mathbf{f}_k$ in (7) up to a scale factor where k is the index of the position. Therefore, the position of the target can be deduced according to the largest element (in absolute value) of $\mathbf{a}$ up to its last element.

Moreover, in a case where $\mathbf{a}$ is sufficiently similar to $\mathbf{f}_k$, it can be put equal to $\mathbf{f}_k$. The reason is that this situation is probable in case that the target is very close to the kth position for which the corresponding CF achieves an optimal performance. This arrangement helps us avoid the statistical error introduced by the ICA algorithm in the vicinity of known positions and improves the cancellation performance.

3) Selection of the ICA Algorithm: In this paper, we use a special case of the BARBI algorithm (BARBI(1)) from [36] instead of BGSEP, which was used in [21]. BARBI utilizes the nonstationarity of signals as well as their spectral diversity while BGSEP uses the former property only. Details of the algorithm are given in Appendix A. The complexity of BARBI(1) is about twice higher compared to BGSEP. Nevertheless, BARBI(1) is still very fast compared to many other ICA algorithms [37],

[38][4]. In the problem defined here, BARBI(1) performs better than BGSEP.

Finally, we summarize steps of the proposed semi-blind method, from now on denoted as *SBSS*. A period of data from microphones is processed as follows.

1) Define $\mathbf{X}$ according to (6).
2) Decompose $\mathbf{X}$ by BARBI(1) and obtain the de-mixing matrix $\mathbf{W}$.
3) Let the last column of $\mathbf{W}$ be denoted by $\mathbf{w}$. Find the largest element of $\mathbf{w}$ in absolute value and denote its index by ℓ_1.
4) Let $\mathbf{a}^T$ denote the ℓ_1th row of $\mathbf{W}$ divided by the negative value of its last element, that is by $-\mathbf{W}_{\ell_1,I+1}$. Hence, the last element of $\mathbf{a}^T$ is -1 (this resolves the scaling ambiguity introduced by the ICA algorithm).
5) If $\mathbf{a}$ is close enough to $\mathbf{f}_k$ for some $k = 1, \ldots, I$, report that the target is in position k and select g_k as the CF. Otherwise, use the CF defined through $\mathbf{a}^T$.

We use an experimentally verified criterion for $\mathbf{a}$ being close to some $\mathbf{f}_k$, which is $s - a < \epsilon$ where s is the sum of all positive elements in $\mathbf{a}$ and a is the largest element of $\mathbf{a}$ whose index is ℓ_2. To explain, note that ideally $\mathbf{a} = \mathbf{f}_{\ell_2}$ and $s = a$. Our choice for ϵ is 1.

C. Example

The motivation for the proposed method is illustrated by the following example that helps in understanding.

Let the target signal be the three female utterances played from position 6 (see Fig. 1). The signal is mixed with the speech of the man moving along the path JM at the signal-to-noise ratio (SNR) of 0 dB. Here, the SNR is evaluated over all three female utterances of 12 s in total length.

In this example, we aim at examining two cases: the CF g_6 either is or is not available in the CFB. The latter case corresponds to the interesting situation when the target occurs in a new position (e.g., between two known positions) for which the CF is not in the CFB. Let $\mathbf{X}_{-6}$ denote the matrix $\mathbf{X}$ without the 6th row.

We apply BARBI(1) to the first half second of data, that is, $\mathbf{X}$ and $\mathbf{X}_{-6}$ are defined with $N_1 = 1$ and $N_2 = 8000$. The number of blocks in BARBI(1) is 40. For comparison, we examine also BGSEP used in [21] with the same number of blocks. The number of samples used for ICA is limited to 8000 in order to show that the ICA methods are able to operate on short intervals.

Fig. 6(a) evaluates outputs of all CFs (including g_6) in terms of the SNR and the output variance. Naturally, the best SNR of -13.8 dB is achieved by g_6 that corresponds to the true position of the target. The filter also yields the minimum output variance as assumed by *MVA* (nevertheless, the difference from the variance outputs of the other filters is at most 0.7 dB).

The SNR of ICs of $\mathbf{X}$ obtained by BARBI(1) and BGSEP are shown in Fig. 6(b). The best independent component by BARBI(1) is the 8th one and achieves -12.7 dB of SNR, which is only slightly worse than the SNR of g_6. The best component of BGSEP is the first one and yields -10.0 dB of SNR.

[4]A Matlab implementation of BARBI is available at `http://si.utia.cas.cz/downloadPT.htm`.

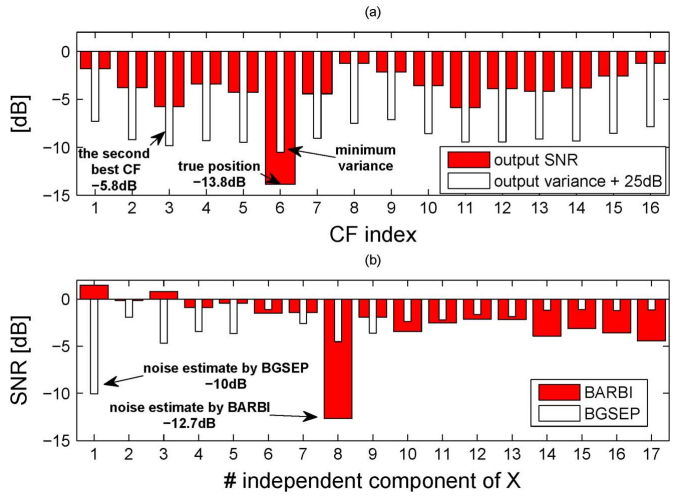

Fig. 6. (a) SNR and output variance of all CFs $g_1, \ldots, g_{16}$ and (b) SNR of independent components of $\mathbf{X}$. The target position is 6.

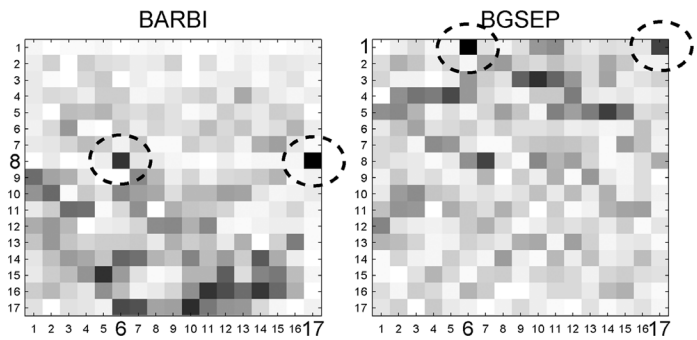

Fig. 7. Absolute values of elements of the de-mixing matrices estimated by BARBI and BGSEP. The last columns and rows with the largest last elements are marked by ovals. The largest element in those rows (up to the last one) point to the position of the target, which is 6. Note that the information provided by BARBI is clearer, which demonstrates the better performance of BARBI compared to BGSEP in this task.

Fig. 7 shows absolute values of elements of the de-mixing matrices obtained, respectively, by BARBI(1) and BGSEP. The figure shows clearly what elements reveal the best ICs and the position of the target. Specifically, the maximum element in the last column of $\mathbf{W}$ points to the best IC. The maximum element of the corresponding row (up to the last element) points to the target's position. Both phenomena are explained by the fact that the linear combination of rows of $\mathbf{X}$ yielding a good noise estimate should be similar to $\mathbf{f}_6$. It is worth emphasizing that the ICA algorithms discover this blindly.

Now we consider the situation when g_6 is not available in the CFB. Here, *MVA* selects g_3 since its output variance is the second smallest after g_6. The achieved SNR of the *MVA* output is -5.8 dB, which is simultaneously the best SNR among all the other CFs in the CFB. Fig. 8(a) shows the SNRs of ICs of $\mathbf{X}_{-6}$ obtained by BARBI(1) and BGSEP. The best SNR of -7.8 dB yields the 8th component by BARBI(1) and -5.7 dB yields the 4th component by BGSEP. Consequently, the CF obtained via BARBI(1) cancels the target signal better than that via BGSEP and *MVA*. Naturally, the localization fails here because the CF for the true position is not available (the selected rows in Fig. 8(b) are not similar to any $\mathbf{f}_k$).

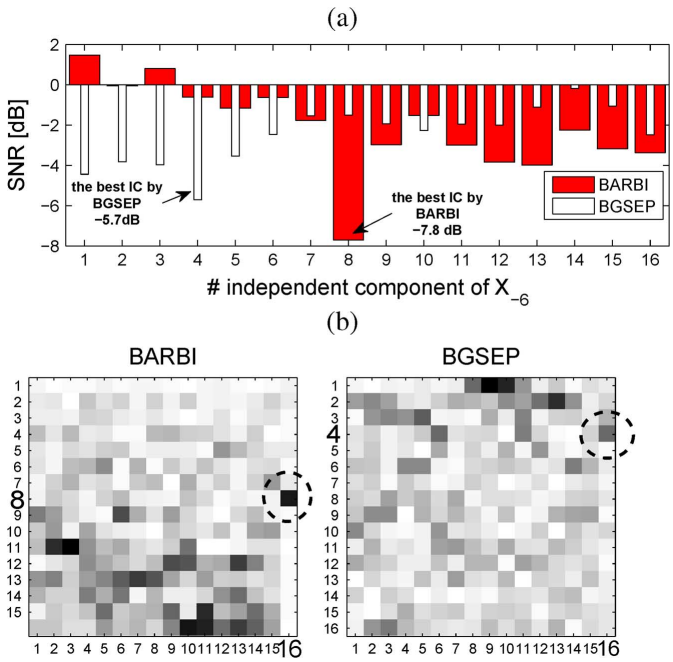

Fig. 8. (a) SNR of independent components of $\mathbf{X}_{-6}$, and (b) absolute values of elements of the estimated de-mixing matrices. The circles denote the maximum elements of last columns of the matrices. They correspond to components yielding the minimum SNR.

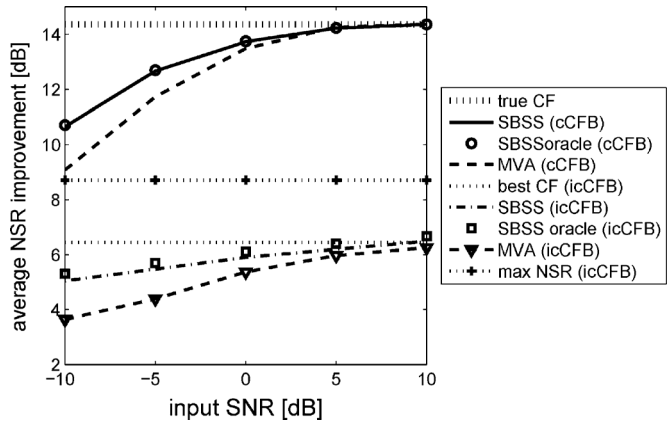

Fig. 9. NSR improvement averaged over all positions and processed blocks of data when complete CFB is available (*cCFB*) and when the CF for the target position is missing (*icCFB*).

D. Frequency-Domain Implementation

The proposed method as well as *MVA* can be implemented in the frequency domain which leads to computational savings. The CFs can be transformed to the Fourier domain and stored in memory in advance. The signals from microphones can be processed block-by-block, applying the short-time Fast Fourier transform (FFT) at the beginning of the process. The CFs are then applied in parallel by multiplying the Fourier images with the transformed blocks of signals.

Next, the *MVA* as well as the *SBSS* can proceed without the need to transform the filtered signals back to the time-domain. In case of *MVA*, the output variances of CFs can be evaluated in the frequency-domain due to the Parseval equality. In *SBSS*, the ICA algorithm can be applied to $\tilde{\mathbf{X}} = [\Re\{\hat{\mathbf{X}}\}\, \Im\{\hat{\mathbf{X}}\}]$, where $\hat{\mathbf{X}}$ denotes the row-wise Fourier transform's counterpart of $\mathbf{X}$ (only one half since the input signals are real) and $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denote, respectively, the real- and imaginary-part operators. The reason for this is that the model $\mathbf{X} = \mathbf{AS}$ holds equivalently for $\tilde{\mathbf{X}}$.

Finally, the inverse FFT and the overlap-add procedure are needed only to obtain the time-domain output signal. The frequency-domain implementation is used in Section V-C.

V. Experiments

All experiments of this section were conducted using the data that were recorded in the scenario described in Section II.

A. Fixed Target

The example of Section IV-C is now repeated for each of the 16 fixed positions of the female speakers. Signals are processed in a batch on-line processing regime, that is block-by-block, where the length of each block is 8000 samples (half a second) with 50% overlaps. The performance of the noise extraction (target cancellation) is assessed by measuring the output Noise-to-Signal Ratio (NSR) in each block. The final criterion is the average taken over the blocks and over all positions of the target and is related to the input SNR (*NSR improvement*).

The mixture of noise signals is created in the following way. As interfering speakers, we use the woman's speech played from position JF and the man's speech played from the dynamic position JM. After six seconds the speakers are interchanged so the next six seconds is the man in position JF and the woman moves along the path JM. These signals are mixed with a non-stationary background (two-channel) noise used in CHiME [32] and with a stationary white Gaussian noise, respectively, in the ratio of -30 dB and -50 dB. The resulting mixture of the noise signals is then added to the target signal at a selected *input SNR* (evaluated over the whole recordings).

We examine the two situations, respectively, denoted by *cCFB* and *icCFB*, when the prior CFB is complete and incomplete (the CF for the current target's position is missing). Besides the proposed *MVA* and the semi-blind method denoted as *SBSS*, four supervised approaches are considered:

- *True CF* selects the CF from CFB for the true position of the target,
- *best CF* selects the one CF from CFB that gives the maximum NSR,
- *max NSR* finds linear combination of rows of $\mathbf{X}$ such that the NSR is maximal, and
- *SBSS oracle* is the proposed semi-blind approach where the independent component is selected based on the maximum NSR.

Note that *true CF*, *best CF*, and *max NSR* perform equally in the *cCFB* case. For the *icCFB* case, *True CF* is not available.

The performance values of the compared approaches are shown in Fig. 9. For *cCFB*, the best NSR improvement is naturally yielded by *True CF*, whose cancellation performance is independent of the input SNR. *True CF* provides a performance bound for *MVA* and *SBSS*. The bound is approached for input SNR$\geq$ 5 dB. For lower SNRs, the performance of *MVA* and *SBSS* decreases, which is mainly caused by the worsened ability to detect the correct CF or, equivalently, the target's position.

TABLE I
POSITION CLASSIFICATION OF FIXED TARGET [%]

input SNR	-10 dB	-5 dB	0 dB	5 dB	10 dB
SBSS	44.3	71.6	86.4	95.9	99.7
MVA	26.0	51.2	77.3	95.1	100.0

TABLE II
NOISE-TO-SIGNAL RATIO IMPROVEMENT OF EXTRACTED NOISE FROM DATA WITH MOVING TARGET (IN DECIBELS)

(a)

method	gender of speakers	input SNR -10 dB	-5 dB	0 dB	5 dB	10 dB
Best CF	male	8.7	8.7	8.7	8.7	8.7
	female	7.7	7.7	7.7	7.7	7.7
SBSS oracle	male	6.6	7.2	7.6	8.3	8.6
	female	5.5	6.3	7.3	7.5	7.6
SBSS	male	6.4	7.0	7.5	8.3	8.6
	female	5.2	6.1	7.2	7.4	7.6
MVA	male	3.7	6.1	7.5	8.1	8.5
	female	3.0	4.2	6.5	7.3	7.7

(b)

method	gender of speakers	input SNR -10 dB	-5 dB	0 dB	5 dB	10 dB
Best CF	male	7.9	7.9	7.9	7.9	7.9
	female	7.0	7.0	7.0	7.0	7.0
SBSS oracle	male	6.2	6.8	7.3	7.9	8.2
	female	5.0	6.2	6.7	7.0	7.4
SBSS	male	6.2	6.7	7.3	7.9	8.2
	female	5.0	6.2	6.7	7.0	7.4
MVA	male	4.2	5.8	7.0	7.5	7.9
	female	4.0	4.8	5.8	6.8	6.9

Table I shows the accuracy of position classification of both approaches evaluated over blocks of signals and all positions of the target. The localization is less accurate when the SNR value on input goes below 5 dB. *SBSS* outperforms *MVA* for low input SNR, that is, in situations when the minimum variance assumption is not properly satisfied while the independence principle is still reliable. Finally, *SBSS oracle* performs slightly better than *SBSS* but not by too much. This proves that the procedure for the IC selection proposed in Section IV-B.1 is efficient.

In the *icCFB* situation, the NSR improvement of all approaches drops by 6–8 dB. There are two different performance bounds provided by *max NSR* and *best CF*. The bound given by *max NSR* is higher by about 2 dB than the bound given by *best CF*. While *MVA* is limited by *best CF*, *SBSS* is limited by *max NSR*. Unfortunately, neither *SBSS* nor *MVA* achieve either of these two bounds. Similarly to *cCFB*, *SBSS* performs better than *MVA* for lower input SNR (≤ 5 dB).

B. Moving Target

In this experiment, we were moving[5] the target source continuously from position 1 to position 16. The male utterances and the female utterances were played in sequence during the movement, so the total length of the recording is 2×12 s. The recorded signals were mixed with the noise signals from the previous example (played twice). The signals are processed block-by-block as in the previous experiment.

[5]A video of the recording is available at `http://itakura.ite.tul.cz/zbynek/dwnld/semiBSSdemo/moving.avi`.

The noise extraction techniques were applied to cancel the target signal at different SNR levels. Two a priori banks of cancellation filters were considered: the CFB with 16 CFs and the gCFB with 9 gCFs introduced in Section III. In both cases, the whole banks were used (i.e., with no missing filters). The results of this experiment in terms of the NSR improvement are listed in Table II.

The moving scenario can be seen as a combination of the *cCFB* and *icCFB* situations, because the target occurs more or less nearby the known positions. The average NSR improvement is between 3 and 8.7 dB, which is in agreement with the results of the previous experiment. The best performance is achieved by *best CF*, *oracle SBSS* performs slightly better than *SBSS*, and *SBSS* outperforms *MVA*, especially when the input SNR is low.

By comparing the results achieved for different genders of the target speakers, the results for the male speakers are better by about 1 dB. This is explained by the fact that the cancellation filters were derived from signals of the first man, so they are better adapted to male voices.

Results achieved with the gCFB are comparable with those attained using the CFB. Here, they are slightly worse (by no more than 0.4 dB) but in other trials of the experiment, not shown due to limited space, we also observed small improvements. Based on this, we conclude that there are two advantages of the gCFB prior to CFB. First, *SBSS* and *MVA* do not always recognize the CF giving the maximum NSR (unlike *best CF*), so they can profit from the wider cancellation range of a selected gCF. Second, the gCFB covers the same area as the CFB while containing a smaller number of filters. This leads to considerable computational savings since outputs of all filters in the bank must be computed. Note that the speed of the ICA algorithm within *SBSS* mostly depends on the dimension of (6), which is equal to the number of filters in the bank plus one.

Fig. 10 shows the estimated positions of the target within blocks of signals by *SBSS* and *MVA* (using CFB). The accuracy cannot be evaluated here, because the correct position of the target is not uniquely determined. Nevertheless, the position index should be gradually growing from 1 to 16. The positions determined by *best CF* provide certain reference localization, which is more or less approximated by *SBSS* as well as by *MVA*, especially for high input SNR (10 dB).

C. Enhancement of the Target's Speech

Now we consider the problem of enhancing the signals of the moving target from the previous section. The enhancer has a simple structure similar to the GSC beamformer: The noise is extracted using one of the proposed noise extraction methods. Its output is used to control a frequency-domain adaptive Wiener filter with an adjustable gain parameter. The Wiener filter performs post-filtering on signals from microphones. A diagram of the enhancer working in the frequency domain is shown in Fig. 11.

Let $Z(k,\ell)$ and $X_i(k,\ell)$ denote, respectively, the short-time Fourier transform (STFT) of the extracted noise signal and that of $x_i(n)$, $i = \mathrm{L}, \mathrm{R}$; k is the frequency index and ℓ is the time-

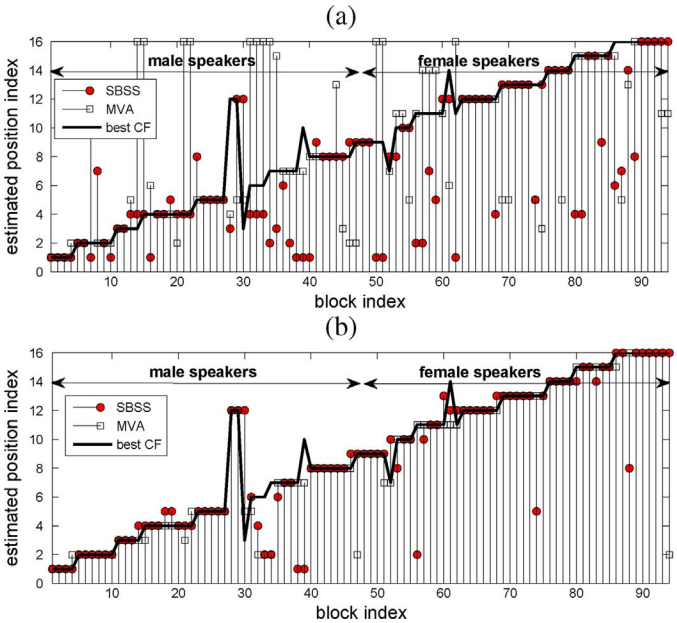

Fig. 10. Estimated positions of the moving target for (a) SNR $= 0$ dB and (b) SNR $= 10$ dB.

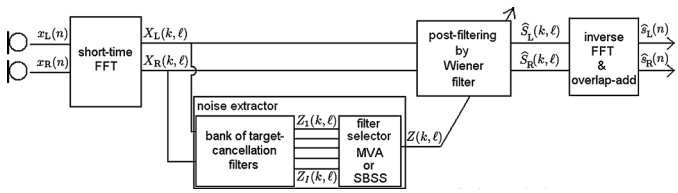

Fig. 11. Diagram of the enhancement method for a moving target source. Variables written by upper case letters denote time-frequency domain transforms of their counterparts.

frame index. The frequency-domain Wiener filter with the gain parameter τ is defined through

$$W_i(k,\ell) = \frac{|X_i(k,\ell)|^2}{|X_i(k,\ell)|^2 + \tau |Z(k,\ell)|^2}, \qquad i = \mathrm{L}, \mathrm{R}. \tag{8}$$

The output of the filter is given by $\hat{S}_i(k,\ell) = W_i(k,\ell) X_i(k,\ell)$, in the time-domain denoted by $\hat{s}_i(n)$, $i = \mathrm{L}, \mathrm{R}$.

An important fact that should be taken into account is that the spectra of the extracted noise signals are colored by the cancellation filters; see (3). Therefore, $|Z(k,\ell)|$ needs to be reconstructed before they are used in (8). An approach to the reconstruction is described, e.g., in [39]. The spectra are corrected with the aid of the mean-square minimization of error between the extracted noise and the original signal from the microphone. We call this *the noise spectrum normalization* and apply it in association with the *MVA*, *SBSS* and *best CF* approaches.

The second method to be compared is based on an on-line implementation of the weighted Natural Gradient in frequency-domain (*FD-BSS*) [11]. This algorithm estimates the mixing parameters of the target and noise sources, which is the counterpart of the CFs, in order to cancel the target signal and extract the responses of the noise. As for the proposed *SBSS*, the method in [11] is not fully blind in the sense that it requires a priori knowledge that the target source lies in a predefined angular range (see Appendix B for more details). In order to remove the typical scaling ambiguity of frequency-domain separation, the Minimal Distortion Principle (MDP) [45] is applied and therefore no further noise spectrum normalization is required. Finally, as for the SBSS the target speech is enhanced by (8) with the estimated noise signal.

To assess the quality of the enhancement [40], we express $\hat{s}_i(n)$ as a sum $\tilde{s}_i(n) + \tilde{y}_i(n)$ where $\tilde{s}_i(n)$ is the contribution of the target's speech and $\tilde{y}_i(n)$ is the contribution of the noise signals. The evaluation is based on three complementary criteria: signal-to-noise (SNR), signal-to-distortion (SDR) and signal-to-distortion-plus-noise ratio (SDNR). They are defined, respectively, by:

$$\mathrm{SNR}_i = \frac{\hat{\mathrm{E}}\left[\tilde{s}_i^2(n)\right]}{\hat{\mathrm{E}}\left[\tilde{y}_i^2(n)\right]} \tag{9}$$

$$\mathrm{SDR}_i = \frac{\hat{\mathrm{E}}\left[\tilde{s}_i^2(n)\right]}{\min_\alpha \hat{\mathrm{E}}[(\tilde{s}_i(n) - \alpha s_i(n))^2]} \tag{10}$$

$$\mathrm{SDNR}_i = \frac{\hat{\mathrm{E}}\left[\tilde{s}_i^2(n)\right]}{\min_\alpha \hat{\mathrm{E}}[(\tilde{s}_i(n) - \alpha s_i(n))^2] + \hat{\mathrm{E}}\left[\tilde{y}_i^2(n)\right]} \tag{11}$$

where $\hat{\mathrm{E}}[\cdot]$ denotes the sample-mean operator, and $i = \mathrm{L}, \mathrm{R}$ is the channel index. The ideal output of the adaptive filter is denoted by $s_i(n)$, which is the contribution of the target signal in $x_i(n)$, that is, $\{h_i * s\}(n)$. SNR measures the residual noise in the enhanced signal while SDR reflects the damage of the target signal in it. SDNR reflects both features, so it serves as an overall criterion. The criteria are evaluated over the complete recordings.

The influence of the parameter τ in (8) on the criteria is significant. With growing τ, SNR usually increases while SDR decreases. To avoid the influence of τ on our comparison, we select the value from $[0, 10]$ for which SDNR achieves its maximum; the interval is limited by 10 since it is experimentally verified that results for $\tau < 10$ are perceptually good. The value of τ is optimized by means of the function `fminbnd` in Matlab.

Fig. 12 shows results[6] in terms of SNR improvement and relative SDR (related to the SDR achieved by *best CF*). The order of the achieved performance values reflect the results of the previous experiment (Table II). *Best CF* gives the best results in terms of SNR, and *SBSS* is better by 0.5–1.5 dB than *MVA* for input SNR lower than 0 dB. Otherwise, the performance values of *SBSS* and *MVA* are similar but *SBSS* achieves a slightly better SDR (0–0.5 dB) for input SNR lower than 10 dB. *SBSS* performs slightly better with gCFB for input SNR lower than 0 dB; *best CF* is uniformly better with CFB.

The results achieved by the *FD-BSS* algorithm are principally different. The algorithm results in a significantly lower SNR value than *SBSS*, namely, by 2–3.5 dB. The relative SDR is better by 0–2.5 dB (up to the input SNR 0 dB), which is achieved thanks to the minimum distortion principle [45]. Nevertheless, the small differences in SDR have little influence on a perceptual quality of the enhanced signals, so the better SNR achieved

[6]Demonstrative samples are available at `http://itakura.ite.tul.cz/zbynek/semiBSSdemo.htm`

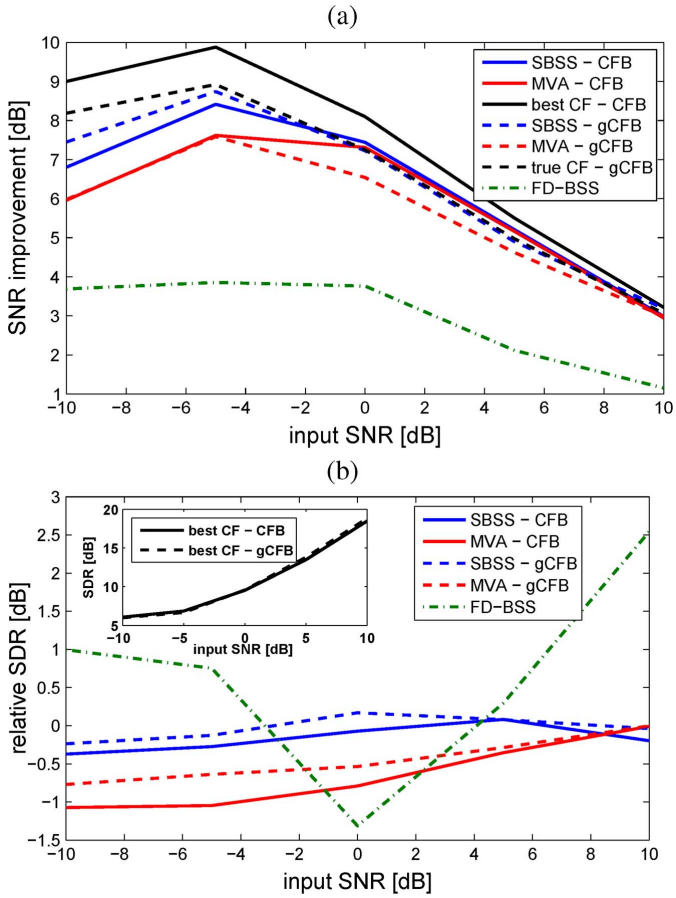

Fig. 12. Results in terms of SNR improvement and relative SDR that were averaged over both channels of the enhanced target signal. The average is also taken for the man and woman target speaker.

by *SBSS* is more important. Note also that *FD-BSS* is much more complex than *SBSS*.

VI. Conclusions and Future Works

We have proposed a method that extracts noise (or, equivalently, estimates the CF) from a noisy recording of a target source whose position is known only roughly in terms of a range of possible locations. The noise is extracted using a bank of pre-measured CFs for several positions of the target and ICA, which is a combination of prior knowledge and of a blind method. In many cases, the extracted noise was shown to have a better NSR than the outputs of individual CFs in the CFB or the output of the filter derived by *MVA*. The proposed method was shown to be useful to enhance the target source. Compared to fully blind algorithms, the number of parameters necessary to estimate the CF is much smaller (it is limited by the size of the CFB), which leads to a considerable simplification, and the method is able to extract the noise even in very difficult conditions.

Since acoustical environments are highly variable, we could not address all possible variants within one paper. There are several emerging problems that may be subject to future research or development, some of which we list now:

- Neither the target's area nor the grid of known positions need be regular. Rotations of the speaker's head, the motion within the 3-D space and changes in the environment should be taken into account. This may lead to an excessive increase of the number of CFs in the CFB. A reduction of this number by use of the group cancellation filters may be a reasonable solution.
- The CFB may be better adapted to different speakers by using longer training data that are sufficiently nonstationary [47].
- The CFB need not be fixed and could be adaptively modified. If a reliable detector of target-only periods is available, novel CFs can be computed from noise-free recordings and compared with the existing ones in the CFB. On the other hand, rarely used CFs can be removed from the CFB.
- The position of the microphones should be chosen depending on the environment and application. On the one hand, it is advantageous to place the microphones as close to the target as possible. On the other hand, the microphones may also be used to target other persons and then their position should be somewhat strategic. The appropriate spacing of microphones is not trivial and should be investigated as well.
- *MVA* provides a simplistic method applicable in devices demanding low-cost solutions, e.g., in mobile phones [48]. In particular, *MVA* performs almost equally well as *SBSS* when signal-to-noise ratio is greater than 0 dB.

Appendix A
Details of BARBI(1)

BARBI(1) is a blind source separation method that relies on signal nonstationarity and on spectral diversity [36]. It assumes that the data matrix $\mathbf{X}$ of the size $d \times N$ can be partitioned to $M > 1$ submatrices $\mathbf{X}_m$, $m = 1, \ldots, M$, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_M]$, where $\mathbf{X}_m$ have, for simplicity, equal size $d \times N_1$, assuming that N is an integer multiple of M, i.e., $N = MN_1$.

The mixing model assumes that $\mathbf{X}_m = \mathbf{A}\mathbf{S}_m$, where each row of $\mathbf{S}_m$ represents an independent Gaussian autoregressive process of the order 1. (Similarly, BARBI(k) assumes AR models of order k).

Estimation of mixing matrix $\mathbf{A}$ relies on sample covariance matrices of lag 0,

$$\hat{\mathbf{R}}_m = \frac{1}{N_1}\mathbf{X}_m\mathbf{X}_m^T = \frac{1}{N_1}\sum_{n=1}^{N_1}\mathbf{x}_{mn}\mathbf{x}_{mn}^T \qquad (12)$$

where $\mathbf{x}_{mn}$ is the nth column of $\mathbf{X}_m$, $m = 1, \ldots, M$, $n = 1, \ldots, N_1$, and on sample covariance matrices of lag 1,

$$\hat{\mathbf{Q}}_m = \frac{1}{2N_1}\sum_{n=1}^{N_1}\left(\mathbf{x}_{mn}\mathbf{x}_{m,n-1}^T + \mathbf{x}_{m,n-1}\mathbf{x}_{mn}^T\right). \qquad (13)$$

To estimate $\mathbf{A}$, BARBI(1) does a weighted approximate joint diagonalization (AJD) of $\mathbf{R}_m$ and $\mathbf{Q}_m$, $m = 1, \ldots, M$. BGSEP utilizes only the matrices $\mathbf{R}_m$ and is therefore about twice as fast but, potentially, less accurate.

The AJD in BARBI(1) proceeds by seeking a demixing matrix $\hat{\mathbf{V}}$ such that the matrices $\hat{\mathbf{V}}\hat{\mathbf{R}}_m\hat{\mathbf{V}}^T$ and $\hat{\mathbf{V}}\hat{\mathbf{Q}}_m\hat{\mathbf{V}}^T$, $m = 1, \ldots, M$, are all approximately diagonal. For future use, let $\mathbf{r}_{k\ell}(\hat{\mathbf{V}})$ and $\mathbf{q}_{k\ell}(\hat{\mathbf{V}})$ be $M \times 1$ vectors composed of $r_{k\ell m}$ and

$q_{k\ell m}$ that are the (k,ℓ)th elements of $\hat{\mathbf{V}}\hat{\mathbf{R}}_m\hat{\mathbf{V}}^T$ and $\hat{\mathbf{V}}\hat{\mathbf{Q}}_m\hat{\mathbf{V}}^T$, respectively.

Next, put

$$a_{km} = -\mathrm{sign}\left(\frac{q_{kkm}}{r_{kkm}}\right)\min\left(\left|\frac{q_{kkm}}{r_{kkm}}\right|, r_{\max}\right), \quad (14)$$

$$\sigma^2_{km} = r_{kkm}\left(1 - a^2_{km}\right), \quad (15)$$

where $0 < r_{\max} \leq 1$ is a constant close to 1, say $r_{\max} = 0.99$. Then, $(1, a_{km})$ can be interpreted as estimated AR coefficients of the kth partially separated signal in the mth block, and σ^2_{km} is an estimate of the variance of the innovation sequence for $k = 1, \ldots, d$ and $m = 1, \ldots, M$. The bound on the maximum allowed radius of poles $r_{\max}$ is used as a constraint on stable AR models.

The initial demixing matrix $\hat{\mathbf{V}}^{[0]}$ is obtained by applying the AJD algorithm UWEDGE [42] to the set of matrices $\{\mathbf{R}_m, \mathbf{Q}_m, m = 1, \ldots, M\}$. Then, BARBI proceeds by iterating

$$\hat{\mathbf{V}}^{[i+1]} = (\hat{\mathbf{A}}^{[i]})^{-1}\hat{\mathbf{V}}^{[i]}$$

where $\hat{\mathbf{A}}^{[i]}$ has ones on its main diagonal, and the (k,ℓ)th and (ℓ,k)th elements are obtained by solving the 2×2 systems

$$\begin{bmatrix} \hat{A}^{[i]}_{k\ell} \\ \hat{A}^{[i]}_{\ell k} \end{bmatrix} = \begin{bmatrix} \mathbf{r}^T_{\ell\ell}\boldsymbol{\phi}_k + \mathbf{q}^T_{\ell\ell}\boldsymbol{\psi}_k & \mathbf{r}^T_{kk}\boldsymbol{\phi}_k + \mathbf{q}^T_{kk}\boldsymbol{\psi}_k \\ \mathbf{r}^T_{kk}\boldsymbol{\phi}_k + \mathbf{q}^T_{kk}\boldsymbol{\psi}_k & \mathbf{r}^T_{kk}\boldsymbol{\phi}_\ell + \mathbf{q}^T_{kk}\boldsymbol{\psi}_\ell \end{bmatrix}^{-1} \cdot \begin{bmatrix} \boldsymbol{\phi}^T_k\mathbf{r}_{k\ell} + \boldsymbol{\psi}^T_k\mathbf{q}_{k\ell} \\ \boldsymbol{\phi}^T_\ell\mathbf{r}_{k\ell} + \boldsymbol{\psi}^T_\ell\mathbf{q}_{k\ell} \end{bmatrix} \quad (16)$$

where

$$\boldsymbol{\phi}_k = \left(\frac{1+a^2_{k1}}{2\sigma^2_{k1}}, \ldots, \frac{1+a^2_{kM}}{2\sigma^2_{kM}}\right)^T \quad (17)$$

$$\boldsymbol{\psi}_k = \left(\frac{a_{k1}}{\sigma^2_{k1}}, \ldots, \frac{a_{kM}}{\sigma^2_{kM}}\right)^T \quad (18)$$

for $k, \ell = 1, \ldots, d$. The form (16)–(18) can be derived from the general expressions for $\mathrm{BARBI}(n)$ in [36] for $n = 1$.

APPENDIX B
DETAILS OF FD-BSS

Here, we describe details of the compared *FD-BSS* method, which is a semi-blind variant of [19]. For a compact notation, let $\mathbf{X}(k,\ell) = [X_{\mathrm{L}}(k,\ell), X_{\mathrm{R}}(k,\ell)]^T$. Here, the length of STFT frames is 2048 samples with time-shift of 128 samples. An on-line weighted Natural Gradient algorithm [43] is applied to $\mathbf{X}(k,\ell)$ to estimate a 2×2 mixing matrix $\mathbf{H}(k,\ell)$ whose inverse is able to split at the outputs the target signal from the remaining noise components [19].

The mixing matrix and the corresponding output signals are estimated as follows:

$$\mathbf{Y}(k,\ell) = [\mathbf{H}(k,\ell)]^{-1}\mathbf{X}(k,\ell) \quad (19)$$

$$\Delta\mathbf{H}(k,\ell) = \mathbf{H}(k,\ell)(\mathbf{I} - \Phi(\mathbf{Y}(k,\ell))\mathbf{Y}(k,\ell)^H)]\boldsymbol{\Psi}(k,\ell) \quad (20)$$

$$\mathbf{H}(k,\ell+1) = \mathbf{H}(k,\ell) - \eta\Delta\mathbf{H}(k,\ell) \quad (21)$$

where η is the step-size, $\Phi(\cdot)$ is a non-linearity, and $\boldsymbol{\Psi}(k,\ell)$ is a diagonal matrix with diagonal elements $\psi_i(k,\ell)$ being weights with values ranging from 0 to 1. The weight $\psi_1(k,\ell)$ is set to the posterior probability of observing the target source in the time-frequency point (k,ℓ), while $\psi_2(k,\ell)$ is set to $1 - \psi_1(k,\ell)$, which indicates the probability of absence of the target source. The probabilities $\psi_1(k,\ell)$ are approximated by spatial binary masks computed from the conjugate projection between the observed normalized cross-power spectrum, $r(k,\ell) = \frac{X_{\mathrm{R}}(k,\ell)X_{\mathrm{L}}(k,\ell)^*}{|X_{\mathrm{R}}(k,\ell)X_{\mathrm{L}}(k,\ell)^*|}$, and the approximated anechoic propagation model $e^{-j2\pi f_k\tau_i}$ as

$$\psi_1(k,l) = \begin{cases} 1, & \text{if } \mathrm{argmax}_i\ \Re\{r(k,l)e^{j2\pi f_k\tau_i}\} = 1 \\ 0, & \text{otherwise} \end{cases}, \quad (22)$$

where f_k is the frequency corresponding to the kth bin and τ_i represents the Time-Difference Of Arrivals (TDOAs) of the acoustic waves impinging the array and propagating from the ith source. In practice, τ_i is estimated by selecting N maxima of a spatial-coherence function computed from the observed STFT frames, through the GCC-PHAT [46] or other enhanced multisource versions [44]. Here, τ_1 should always correspond to a TDOA which in turn corresponds to a location in the admissible range for the target source ($\pm 15^\circ$).

ACKNOWLEDGMENT

We thank Prof. Sharon Gannot for fruitful and helpful discussions.

REFERENCES

[1] *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., 1st ed. Heidelberg, Germany: Springer-Verlag, 2005.
[2] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.
[3] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, Nov. 2004.
[4] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[5] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Jan. 1972.
[6] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Audio, Speech, Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
[7] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
[8] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 206–219, Jan. 2011.
[9] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.
[10] J. Chen, J. Benesty, and Y. Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 481–493, Mar. 2008.
[11] F. Nesta and M. Matassoni, "Blind source extraction for robust speech recognition in multisource noisy environments," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 703–725, May 2013.
[12] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Commun.*, vol. 53, no. 5, pp. 677–689, Jun. 2011.

[13] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.

[14] I. Cohen, "Relative transfer function identification using speech signals," *Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[15] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[16] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. New York, NY, USA: Springer, Sep. 2007.

[17] Y. Zheng, K. Reindl, and W. Kellermann, "BSS for improved interference estimation for blind speech signal extraction with two microphones," in *Proc. Int. Workshop Comp. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Aruba, Dutch Antilles, Dec. 2009, pp. 253–256.

[18] J. Even, C. Ishi, H. Saruwatari, and N. Hagita, "Close speaker cancellation for suppression of non-stationary background noise for hands-free speech interface," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech '10)*, Makuhari, Chiba, Japan, Sep. 26–30, 2010, pp. 977–980.

[19] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal source correlation," in *Proc. 10th Int. Conf. Latent Variable Anal. Source Separat. (LVA/ICA 2012)*, Tel-Aviv, Israel, Mar. 12–15, 2012, pp. 222–230.

[20] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Real-time blind extraction of dominant target sources from many background interference sources," in *Proc. IWAENC '05*, Sep. 2005, pp. 73–76.

[21] J. Málek, Z. Koldovský, and P. Tichavský, "Semi-blind source separation based on ICA and oOverlapped speech detection," in *Proc. 10th Int. Conf. Latent Variable Anal. Source Separat. (LVA/ICA '12)*, Tel-Aviv, Israel, Mar. 12–15, 2012, pp. 462–469.

[22] P. Smaragdis, "Position and trajectory learning for microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 358–368, Jan. 2007.

[23] J.-S. Hu and W.-H. Liu, "Location classification of nonstationary sound using binaural room distribution patterns," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 682–692, May 2009.

[24] S. Vesa, "Binaural sound source distance learning in rooms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1498–1507, Nov. 2009.

[25] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2011, pp. 245–248.

[26] M. Jeub, Ch. Herglotz, Ch. Nelke, Ch. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *Proc. ICASSP '12*, Kyoto, Japan, Mar. 25–30, 2012, pp. 1693–1696.

[27] N. I. Durlach, "Equalization and cancellation theory of binaural masking level differences," *J. Acoust. Soc. Amer.*, vol. 35, no. 8, pp. 1206–1218, 1963.

[28] Y. Lin, J. Chen, Y. Kim, and D. Lee, "Blind channel identification for speech dereverberation using ℓ_1 norm sparse learning," in *Proc. 21st Annu. Conf. Neural Inf. Process. Syst., Advances in Neural Inf. Process. Syst. 20*, Vancouver, BC, Canada, Dec. 3–6, 2007, MIT Press.

[29] M. Yu and J. Xin, "Exploring off time nature for speech enhancement," in *Proc. 13th Annual Conf. Int. Speech Commun. Assoc. (Interspeech '12)*, Portland, OR, Sep. 9–13, 2012.

[30] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[31] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[32] J. P. Barker, E. Vincent, N. Ma, H. Christensen, and P. D. Green, "The PASCAL CHiME speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 621–633, May 2013.

[33] Z. Koldovský, J. Málek, M. Balík, and J. Nouza, "CHiME data separation based on target signal cancellation and noise masking," in *Proc. Int. Workshop Mach. Listening in Multisource Environ.*, Florence, Italy, Aug. 2011, pp. 47–50.

[34] N. Levinson, "The Wiener RMS error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278, 1947.

[35] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 90, no. 8, pp. 2009–2026, Oct. 1998.

[36] P. Tichavský, A. Yeredor, and Z. Koldovský, "A fast asymptotically efficient algorithm for blind separation of a linear mixture of block-wise stationary autoregressive processes," in *Proc. ICASSP '09*, Taipei, Taiwan, Apr. 2009, pp. 3133–3136.

[37] Z. Koldovský and P. Tichavský, "A comparison of independent component and independent subspace analysis algorithms," in *Proc. EUSIPCO '09*, Glasgow, U.K., Aug. 24–28, 2009, pp. 1447–1451.

[38] P. Tichavský and Z. Koldovský, "Fast and accurate methods of independent component analysis: A survey," *Kybernetika*, vol. 47, no. 3, pp. 426–438, Jun. 2011.

[39] I. Tashev, *Sound Capture and Processing: Practical Approaches*. New York, NY, USA: Wiley, 2009.

[40] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. Int. Workshop Ind. Compon. Anal. Signal Separat. (ICA'99)*, Aussois, France, Jan. 1999, pp. 261–266.

[41] Z. Koldovský and P. Tichavský, "Time-domain blind separation of audio sources on the basis of a complete ICA decomposition of an observation space," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 406–416, Feb. 2011.

[42] P. Tichavský and A. Yeredor, "Fast approximate joint diagonalization incorporating weight matrices," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 878–891, Mar. 2009.

[43] A. Cichocki and S.-I. Amari, *Adaptive Signal and Image Processing: Learning Algorithms and Applications*. New York, NY, USA: Wiley, 2002.

[44] F. Nesta and M. Omologo, "Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources," in *Proc. ICASSP '12*, Kyoto, Japan, Mar. 25–30, 2012, pp. 213–216.

[45] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. 3rd Int. Conf. Ind. Compon. Anal. Blind Source Separat. (ICA'01)*, San Diego, CA, USA, Dec. 2001, pp. 722–727.

[46] Ch. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Signal Process.*, vol. SP-24, no. 4, pp. 320–327, Apr. 1976.

[47] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: A time domain approach," *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 340–349, Mar. 1994.

[48] Z. Koldovský, P. Tichavský, and D. Botka, "Noise reduction in dual-microphone mobile phones using a bank of pre-measured target-cancellation filters," in *Proc. ICASSP '13*, Vancouver, BC, Canada, May 2013, pp. 679–683.

Zbyněk Koldovský (S'03–M'04) was born in Jablonec nad Nisou, Czech Republic, in 1979. He received the M.S. degree and Ph.D. degree in mathematical modeling from Faculty of Nuclear Sciences and Physical Engineering at the Czech Technical University in Prague in 2002 and 2006, respectively.

He is currently an associate professor at the Institute of Information Technology and Electronics, Technical University of Liberec. He has also been with the Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic since 2002. His main research interests are focused on audio signal processing, blind source separation, statistical signal processing and multilinear algebra.

Jiří Málek received his master and Ph.D. degrees from Technical University in Liberec (TUL, Czech Republic) in 2006 and 2011, respectively, in technical cybernetics. Currently, he holds a postdoctoral position at the Institute of Information Technology and Electronics, TUL. His research interests include blind source separation and speech enhancement.

Petr Tichavský (M'98–SM'04) received the M.S. degree in mathematics in 1987 from the Czech Technical University, Prague, Czechoslovakia and the Ph.D. degree in theoretical cybernetics from the Czechoslovak Academy of Sciences in 1992. Since that time he has been with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic in Prague. In 1994 he received the *Fulbright grant* for a 10 month fellowship at Yale University, Department of Electrical Engineering, in New Haven, CT, U.S.A. In 2002 he received the *Otto Wichterle Award* from Academy of Sciences of the Czech Republic.

He is author and co-author of research papers in the area of sinusoidal frequency/frequency-rate estimation, adaptive filtering and tracking of time varying signal parameters, algorithm-independent bounds on achievable performance, sensor array processing, independent component analysis and blind source separation.

Petr Tichavský served as associate editor of the IEEE SIGNAL PROCESSING LETTERS from 2002 to 2004, and as associate editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2005 to 2009 and from 2011 to now. Since 2009 he is a member of the IEEE Signal Processing Society's Signal Processing Theory and Methods (SPTM) Technical Committee. Petr Tichavský has also served as a general co-chair of the 36th IEEE Int. Conference on Acoustics, Speech and Signal Processing ICASSP 2011 in Prague, Czech Republic.

Francesco Nesta received the Laurea degree in computer engineering from Politecnico di Bari, Bari, Italy, in September 2005 and the Ph.D. degree in information and communication technology from University of Trento, Trento, Italy, in April 2010, with research on blind source separation and localization in adverse environments.

He has been conducting his research at Bruno Kessler Foundation IRST, Povo di Trento, from 2006 to 2012. He was a Visiting Researcher from September 2008 to April 2009 with the Center for Signal and Image Processing Department, Georgia Institute of Technology, Atlanta. His major interests include statistical signal processing, blind source separation, speech enhancement, adaptive filtering, acoustic echo cancellation, semi-blind source separation and multiple acoustic source localization. He is currently working at Conexant System, Irvine (CA, USA) on the development of audio enhancement algorithms for far-field applications.

Dr. Nesta serves as a reviewer for several journals such as the IEEE TRANSACTION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, *Elsevier Signal Processing Journal*, *Elsevier Computer Speech and Language*, and in several conferences and workshops in the field of acoustic signal processing. He has served as Co-Chair in the third community-based Signal Separation Evaluation Campaign (SiSEC 2011) and as organizer of the 2nd CHIME challenge.