UNSUPERVISED SPATIAL DICTIONARY LEARNING FOR SPARSE UNDERDETERMINED MULTICHANNEL SOURCE SEPARATION

Francesco Nesta and Mahmoud Fakhry*

Center of Information Technology, Fondazione Bruno Kessler - Irst
via Sommarive 18, 38123 Trento, Italy
email:francesco.nesta@gmail.com,abdelraheem@fbk.eu

ABSTRACT

Multichannel sparse representation of acoustic sources has shown to provide an attractive framework for source separation. The multichannel sparse modeling assumes an ability to describe signals as linear combinations of few atoms from a pre-specified dictionary. The dictionary is built by simulating room impulse responses on a grid of locations, exploiting a prior knowledge on the room geometry and reflection coefficients. However, due to the simplified modeling, any mismatch between the simulated and true observed RIRs would generate a considerable distortion in the recovered output signals. In this work we propose an unsupervised adaptation of the dictionary through a semi-blind weighted Natural Gradient, assuming spatio-temporal source sparseness. The system continuously adapts the atoms with the incoming data, improving the match between the dictionary and the true mixing parameters. Results over simulated data show that the proposed framework is a promising solution to underdetermined convolutive source separation in difficult acoustic scenarios.

***Index Terms*—** multi-channel detection, source separation, sparse signal, matching pursuit, source localization.

1. INTRODUCTION

Underdetermined convolutive source separation is one of the most challenging problem related to acoustic source enhancement. Several algorithms have been proposed in the last year but still there is not a widely recognized solution. Many of them separate the sources with spectral filters, exploiting simplified spatial models for the acoustic propagation [1][2][3]. However, the need for modeling long separating filters, due to the presence of high reverberation, is one of the main factors limiting the performance of this class of algorithms. To mitigate the effect of reverberation, in some methods the separation problem is solved separately in different subbands [4] and later spatio-temporal models are used to cluster signal components related to the same source [5]. In alternative, other algorithms use temporal and spectral redundancies for factorizing different source components [6]. As discussed in [5] one of the key tasks that has to be solved is the estimation of the wide-band source mixing parameters. In fact, if the complete mixing system is available, spectral masking or L_p-norm minimization can be adopted for segregating the mixtures in their individual source components [7].

While unsupervised techniques are commonly adopted for source separation, nesting a prior knowledge in the estimation process is a new trend which promises to improve the robustness of blind methods. An emerging framework well fitting this view is multichannel sparse modeling (MSM). Sparse modeling of data assumes an ability to describe the signal using a pre-specified dictionary. The dictionary involves a proper definition of atoms in order that signals can be uniquely represented as a linear combination of them [8, 9]. In the multichannel scenario, i.e. when the source signals are recorded by a number of microphones $M \geq 2$, the dictionary can be defined through description of the source mixing parameters. In low reverberant environments this description can be approximated by anechoic models [1] while in more echoic environments, RIRs can be learned from the data off-line [10] when only one source at the time is active.

An alternative way to generate a spatial dictionary is by using a model-based definition of the room impulse responses. A dictionary of RIRs between the microphones and a set of points can be generated explicitly using information of the room geometry and of the acoustic parameters of the reflective surfaces [11][12]. While this approach has a high flexibility and requires only a limited knowledge to generate large dictionaries, any mismatch between the simulated atoms and the mixing parameters underlying the true mixing process is cause of high distortion on the recovered signals. In fact, simulated RIRs always differ from the true ones because they are built on a simplified geometrical description of the environment. Furthermore, while the early propagation paths can be approximatively modeled from the geometry, late reverberation cannot be easily represented deterministically. Moreover, mismatch arises also depending on the resolution of the sampling of the spatial locations, which should not be too high to limit the size of the dictionary. Although the effect of this mismatch can be mitigated by proper normalizations [12], still its effect is crucial and it should be reduced as much as possible. In response to this need, dictionary adaptation with the incoming data is a possible viable solution.

In this work we fuse the concept of model-based spatial dictionary and blind mixing system estimation in a single framework, through the effective combination of sparse modeling and ICA. A supervised ICA based on the weighted Natural Gradient is exploited to adapt the original dictionary with the incoming data. The ICA adaptation is based on the assumption of sparse spatio-temporal representation of the acoustic sources. Experimental results on simulated data have shown that the semi-blind learning can sensibly improve the match between the incoming data and the adapted dictionary with a consequent considerable improvement in the source separation performance.

*now at Conexant Systems, Inc., Newport Beach, CA (USA)

2. SPARSE MODELING FORMULATION

2.1. Signal model

N source signals are assumed to be recorded by an array of M elements. We refer to the discrete time-frequency representation of signals, obtained for example through the Short-time Fourier Transform (STFT). Let $S_n(k,l)$ and $X_m(k,l)$ be the *l-th* STFT frame coefficients obtained for the k-*th* frequency bin for the n-*th* source and m-*th* mixture signal, respectively. For convenience of notation we indicate the source signal vector with $\mathbf{S}(k,l) = [S_1(k,l)\cdots S_N(k,l)]^T$, and the mixtures $\mathbf{X}(k,l) = [X_1(k,l)\cdots X_M(k,l)]^T$ which can be then modeled as

$$\mathbf{X}(k,l) = \mathbf{H}(k)\mathbf{S}(k,l). \tag{1}$$

Equation (1) is a general representation of the mixing system but in real-world acoustic sources tend to have a highly sparse power representation in the STFT domain, i.e. only one source can be considered dominant in each time-frequency point. Therefore a convenient approximation is to consider the vector $\mathbf{S}(k,l)$ to be all zero other than for the element $n(k,l)$, which indicates the source dominating the point (k,l). Following this approximation, a better representation of the data which is independent on the source variance, is obtained by computing the ratios between the observed signals and a reference channel, e.g. the channel $m = 1$, and normalizing them by their magnitude

$$R_m(k,l) = \frac{X_m(k,l)X_1(k,l)^*}{|X_m(k,l)X_1(k,l)^*|} \simeq \frac{H_{m,n(k,l)}(k)H_{1,n(k,l)}(k)^*}{|H_{m,n(k,l)}(k)H_{1,n(k,l)}(k)^*|}. \tag{2}$$

This representation is convenient because the ratios $R_m(k,l)$ provide a redundant representation of the mixing systems of the sources $\forall k,l$. Note, the magnitude normalization is useful because it has the effect of binding the error due to the non perfect sparse assumption (see [12] for details). Therefore, if the sources are static, the signals can be modeled as linear combinations of few normalized mixing systems which can be considered atoms of a pre-built spatial dictionary[11, 12].

The exact modeling of the mixing system would require a geometrical description of all the reflective surfaces and of their sound absorption characteristic. A reasonable approximation can be obtained through the Image-source model (ISM) [13] if at least room and array geometry are available. A finite dictionary can be then modeled by selecting a finite set of points, e.g. on a two-dimensional grid. The channel $h_m^o(t)$ from the o-*th* location to the acoustic sensor m is simulated through the ISM method and the discrete Fourier transform is applied to obtain the frequency representation of the impulse response $H_m^o(k)$. A generic atom describing the normalized multichannel spatial propagation can be represented as

$$\mathbf{d}_m^o = \left[\frac{H_m^o(1)H_1^o(1)^*}{|H_m^o(1)H_1^o(1)^*|}, \cdots, \frac{H_m^o(N_{bins})H_1^o(N_{bins})^*}{|H_m^o(N_{bins})H_1^o(N_{bins})^*|}\right]^T \tag{3}$$

$$\mathbf{d}^o = [\mathbf{d}_2^o; \cdots; \mathbf{d}_M^o] \tag{4}$$

while the dictionary including all the simulated atom vectors is defined as $\mathbf{D} = [\mathbf{d}^1|\cdots|\mathbf{d}^{N_{atoms}}]$, where N_{atoms} indicates the number of simulated atoms, according to the used spatial resolution. N_{bins} indicates the number of non-symmetric frequency bins, i.e. $L/2+1$ where L is the window length of the STFT. Equivalently the observed data can be organized with the same atom representation as

$$\mathbf{R}_m^l = [R_m(1,l)\cdots R_m(N_{bins},l)]^T \tag{5}$$

$$\mathbf{R}^l = [\mathbf{R}_2^l; \cdots; \mathbf{R}_M^l] \tag{6}$$

3. MODIFIED ORTHOGONAL MATCHING PURSUIT

Searching for the set of atoms best representing our observations is an NP-hard problem for a redundant dictionary. Nevertheless greedy approximations such as those based on the matching pursuit (MP) allow us to reduce the complexity to a tractable level. In this work we adopt a modified orthogonal matching pursuit (OMP) [14], which was used in our former work [12] and has shown to perform well for the case of matching between atoms and observations. In short, in MP-based algorithms the atoms are initially matched with the observed data, i.e in our case $\mathbf{R}^l$, according to a predefined matching operator. The best matching atom is selected and at each *i-th* iteration the effect of it in the observed data is removed through the computation of a residual $\mathbf{Z}_i^l$. Iteratively, the matching procedure continues with the selection of new atoms matching the last computed residual till a predefined stopping criterion is satisfied, e.g. till N atoms are selected from the dictionary. Here, we define a matching operator between each atom of our dictionary and the observed residual as follows. First, assuming also temporal dominance, we compute the inner product of the columns of the current residual and the atoms of the sparse dictionary and select the one maximizing it inside each time frame l

$$o_l^{\text{match}} = \arg\max_o |(\mathbf{d}^o)^*\, \mathbf{Z}_{i-1}^l|. \tag{7}$$

where $*$ indicates the conjugate transpose. Thus, the J most frequent atoms are selected and among of them the one leading to the largest inner product cumulated over all the frames is chosen:

$$\mathrm{j}^{\text{match}} = \arg\max_j \sum_l |(\mathbf{d}^{q_j})^*\, \mathbf{Z}_{i-1}^l|, \;\; j = 1,2,...,J \tag{8}$$

where q_j indicates the index of the J-atoms selected with (7).

We indicate with $\mathbf{D}_{\Gamma_i}$ the sub-dictionary of the selected matched atoms in the *i-th* iteration spanned by the atoms indexed in the subspace $\mathbf{\Gamma}_i$, $\mathbf{D}_{\Gamma_i}^\dagger$ is the pseudo-inverse of $\mathbf{D}_{\Gamma_i}$, and $\mathbf{D}_{\Gamma_i}^*$ is the conjugate transpose of $\mathbf{D}_{\Gamma_i}$. The OMP algorithm can be summarized as follows

---

Initialize: $\mathbf{Z}_0^l = \mathbf{R}^l,\ \Gamma_0 = \phi,\ \mathbf{D}_{\Gamma_0} = [\mathbf{0}]$.
For i = 1 to N
- find the index $\mathrm{j}^{\text{match}}$ of the best matching atom with $\mathbf{Z}_{i-1}^l$, $\forall l$ as in (8)
- update the sub-dictionary by the new atom $\mathbf{D}_{\Gamma_i} = [\mathbf{D}_{\Gamma_{i-1}}|\mathbf{d}^{\mathrm{j}^{\text{match}}}]$,
- update the sub-space by the new atom index $\mathbf{\Gamma}_i = \mathbf{\Gamma}_{i-1} \cup \mathrm{j}^{\text{match}}$,
- orthogonal projection : $\hat{\mathbf{P}}_i^l = \mathbf{D}_{\Gamma_i}^\dagger \mathbf{Z}_{i-1}^l, \forall l$,
- update the residual : $\mathbf{Z}_i^l = \mathbf{Z}_{i-1}^l - \mathbf{D}_{\Gamma_i}\hat{\mathbf{P}}_i^l, \forall l$,
- normalize each element of the residual vector $\mathbf{Z}_i^l$ to unit magnitude.

Return

---

4. DICTIONARY ADAPTATION THROUGH WEIGHTED NATURAL GRADIENT

As discussed before, the mismatch between the true mixing systems and the atoms in the dictionary is the main cause of poor performance of spatial sparse modeling. In contrast, blind techniques are able to estimate the mixing system without no specific geometrical knowledge and then better adapt to the observed data. However, their robustness is limited by low convergence, high estimation variance and signal conditions not well fitting the general hypothesis of independence in short-time. Here we effectively combine both the approaches in order to compensate their individual weak points, leading to a semi-blind estimation method.

We start with the hypothesis that there is only one source dominating a specific STFT frame. Therefore, each instant is used to adapt only the atom related to the dominating source. For this purpose we use a modification of the weighted Natural Gradient (wNG) proposed in [5]. The main idea behind wNG is to re-weight the gradient according to the likelihood of dominance of a source in a given frame in order to selectively estimate the mixing parameters related to different spatial locations. Following this idea, we select the atom in the dictionary best matching with the observed frame l

$$\tilde{o} = \arg\max_{o} \Pr(o,l), \quad \Pr(o,l) = |(\mathbf{d}^{o})^{*}\, \mathbf{R}^{l}|, \tag{9}$$

and normalize the respective projection as

$$\overline{\Pr}(\tilde{o},l) = \frac{\Pr(\tilde{o},l) - \Pr_{\tilde{o}}^{min}}{\Pr_{\tilde{o}}^{max} - \Pr_{\tilde{o}}^{min}} \tag{10}$$

where $\Pr_{\tilde{o}}^{min}$ and $\Pr_{\tilde{o}}^{max}$ are the minimum and maximum projection of the atom $\tilde{o}$ with all the previously observed data frames. The normalized projection is then a weight with values ranging from 0 to 1, indicating the dominance of the source at the location $\tilde{o}$ and at the frame l.

A weighting matrix $\mathbf{P}^{\tilde{o}}(l)$ is defined as a diagonal matrix with the first element equal to $\overline{\Pr}(\tilde{o},l)$ and the remaining elements set to $1-\overline{\Pr}(\tilde{o},l)$. A squared $M \times M$ mixing matrix, describing the source propagating from the location related to the atom o at the frequency bin k, is initialized as

$$\widehat{\mathbf{H}}^{o}(k) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \mathbf{d}_2^{o}(k) & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{d}_M^{o}(k) & 0 & \cdots & 1 \end{bmatrix}, \quad \forall o \tag{11}$$

where $\mathbf{d}_m^{o}(k)$ indicates the k-th element of the vector $\mathbf{d}_m^{o}$ Then, according to the weighted NG, for each frame l, the atom selected in (9) and its corresponding mixing system is updated as follows

$$\mathbf{Y}(k,l) = [\widehat{\mathbf{H}}^{\tilde{o}}(k)]^{-1}\mathbf{X}(k,l) \tag{12}$$

$$\Delta\mathbf{H}(k) = [\widehat{\mathbf{H}}^{\tilde{o}}(k)(\mathbf{I} - \Phi(\mathbf{Y}(k,l))\mathbf{Y}(k,l)^{H})]\mathbf{P}^{\tilde{o}}(l) \tag{13}$$

$$\widehat{\mathbf{H}}^{\tilde{o}}(k) = \widehat{\mathbf{H}}^{\tilde{o}}(k) - \eta\Delta\mathbf{H}(k) \tag{14}$$

$$\mathbf{d}_m^{\tilde{o}} = \left[\frac{\widehat{H}_{m1}^{\tilde{o}}(1)\widehat{H}_{11}^{\tilde{o}}(1)^{*}}{|\widehat{H}_{m1}^{\tilde{o}}(1)\widehat{H}_{11}^{\tilde{o}}(1)^{*}|}, \cdots, \frac{\widehat{H}_{m1}^{\tilde{o}}(N_{bins})\widehat{H}_{11}^{\tilde{o}}(N_{bins})^{*}}{|\widehat{H}_{m1}^{\tilde{o}}(N_{bins})\widehat{H}_{11}^{\tilde{o}}(N_{bins})^{*}|}\right]^{T} \tag{15}$$

$$\mathbf{d}^{\tilde{o}} = [\mathbf{d}_2^{\tilde{o}}; \cdots; \mathbf{d}_M^{\tilde{o}}] \tag{16}$$

where η is the adaptation step-size, $\mathbf{I}$ the identity matrix and $\Phi(\cdot)$ is a non-linearity. In practice, the weighting matrix induces the gradient to update the first column of $\widehat{\mathbf{H}}_{\tilde{o}}(k)$ when the source located in $\tilde{o}$ is dominant.

Interestingly, the above adaptation structure differs from that of traditional on-line determined BSS which updates a single mixing/demixing matrix, in order to split the observed mixtures in their individual components. In contrast the proposed algorithm realizes a semi-blind spatio-temporal learning, i.e. the learning proceeds not only in time but also in the spatial domain, according to the prior knowledge given by the geometry. Therefore, the learning can continue even when the source of interest is silent but some localized noise sources are active, so that a *learning from noise* becomes possible. This is an attractive property which can considerably increase speed and robustness of separation when compared to any blind method.

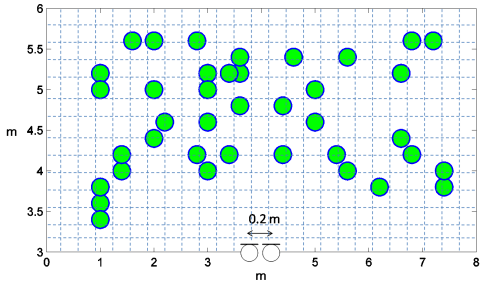

Fig. 1. Simulation setup: dots indicate the true locations of the sources in the mixtures while cross points in the grid indicate the spatial locations modeled by the original dictionary.

5. SIMULATION

In order to have a better control of the evaluation, the proposed method was tested on simulated data[1]. Room impulse responses related to a room with size $8 \times 6 \times 3$ meters and an array of 2 omni directional microphones spaced of 0.2 m were considered. Microphones were located in the middle of the room with the same elevation as for the sources (0.5 m height). Synthetic RIRs were simulated through ISM between multiple locations in the room and the microphones, over a grid of two-dimensional points with a spatial resolution of 0.25 m (i.e. a total of N_{atoms}= 341 atoms), with a sampling frequency of f_s=16kHz and transformed to the discrete frequency-domain with a DFT of $L = 4096$ points. For the evaluation, we simulated a mismatched set of RIRs for the generation of the dictionary and for the generation of the mixtures. The first set was obtained simulating RIRs assuming uniform absorption coefficients over all the room surfaces (walls, ceiling and floor) and with a reverberation time of about $T_{60} = 50$ ms. The generation was obtained through the simulator provided by [16]. The second set was simulated in a similar way but sampling the spatial locations with a random offset (between 0 and 5 cm) with respect to the atom locations and using a larger reverberation time $T_{60} = 250$ ms. In this way we generated a double mismatch between the RIRs in the dictionary and those underlying the mixtures. Indeed, this is a realistic condition that one would observe in real-world because the source location cannot be exactly restricted to the points sampled in the dictionary and the accuracy of the modeled RIRs is always limited by the used geometrical model.

For the generation of the mixtures three different datasets were considered: a set for adaptation used for updating the atoms in the dictionary; two sets for evaluating the separation performance.

The adaptation set was generated by creating mixtures of acoustic sources using domestic noise signal samples in the Freesound[2] and the Logic Pro libraries, and added in order to generate a time-varying degree of overlap (for a maximum of three sources overlapping in time). It consists of two hundred mixtures of 12 seconds each for a total of about 40 minutes. Time-domain mixtures were generated by individually convolving simulated RIRs for a given set of locations (see Figure 1), with the original source signals and adding the source image contributions at each microphone.

The mixtures for the first evaluation dataset were generated by using a speech signal selected from the TIMIT database and three random domestic noise signals. The second evaluation dataset was

[1] An example of real-world application for the spatial dictionary learning is presented in [15].

[2] http://www.freesound.org/

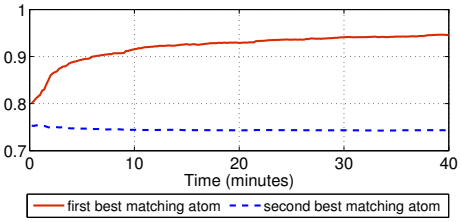

Fig. 2. Average inner product between the true mixing systems with the first best matching atom (solid line) and the second best matching atom (dotted line)

generated by using only speech signals, for a total of 4 overlapping speakers. Both test sets consist in 20 mixtures of about 15s. In all the datasets source locations were randomly modified for each mixture. The discrete time-frequency representation of the mixture $\mathbf{X}(k,l)$ was obtained through STFT with Hanning windows with length L, shifted of 512 samples. In the weighted Natural Gradient the adaptation step-size was set to $\eta = 0.02$ and the non-linearity to $\Phi(x) = \tanh(10 \cdot |x|)\frac{x}{|x|}$.

6. PERFORMANCE EVALUATION

Two different criteria are considered for the evaluation: system identification performance and benchmarking of the signal separation.

6.1. System identification

Figure 2 shows the average projection obtained after having adapted the dictionary with a certain amount of data and showing the projection when considering the first and second best matching atom. At the instant 0 the average projection corresponds to the performance evaluated with the original unadapted dictionary. It can be noted that as the learning proceeds over time the average projection of the first atom approaches the unity, which means that each true mixing system will eventually have a close match with one of the adapted atom. On the other hand, the second best matching atom remains unaltered during the learning which means that the discrimination between the atoms increases with the learning, which is a desirable feature for MP-based detection algorithms.

6.2. Signal separation

To complete the analysis we report the signal separation performance in terms of Signal-to-Distortion ratio (SDR) and Delta Signal-to-Interference-Ratio (ΔSIR), as defined in [17]. In the evaluation both the original dictionary and the updated dictionary were considered. The signals were separated using the L_0-norm minimization [7], applied to each time-frequency point independently by defining the full estimated mixing system as

$$\widetilde{\mathbf{H}}(k) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \mathbf{d}_2^{o_1}(k) & \mathbf{d}_2^{o_2}(k) & \cdots & \mathbf{d}_2^{o_N}(k) \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{d}_M^{o_1}(k) & \mathbf{d}_M^{o_2}(k) & \cdots & \mathbf{d}_M^{o_N}(k) \end{bmatrix} \quad (17)$$

where $\mathbf{d}_M^{o_i}(k)$ indicates the k-th element of the atom selected at the i-th iteration by the OMP algorithm. After separation in each frequency, the Minimal Distortion Principle (MDP) [18] was used to estimate the multichannel image of each source. Finally, the STFT signals where reconstructed back to time-domain through a weighted Overlap-and-add (WOLA) using the Griffin and Lim's method [19].

Metric	Adapted dictionary	Original Dictionary
SDR	**8.2(2)**	**2.9(4.5)**
Δ SIR	**2.9(3.5)**	**-4(7)**

Table 1. Mean (standard deviation) performance in dB for separated signals with and without dictionary adaptation for test dataset with 1 speech + 3 noise random signals. Performance only refer to the target speech signal.

Metric	S1	S2	S3	S4	Avg
SDR	6.4(1.6)	6.4(2.7)	4(2.9)	3.8(3.4)	**5.1(2.9)**
Δ SIR	14.6(2.4)	3.3(3)	7.2(3.2)	13(3.3)	**9.5(5.4)**

Adapted dictionary

Metric	S1	S2	S3	S4	Avg
SDR	3.2(3.2)	2.4(4.4)	4.7(5.4)	2.3(4.4)	**3.1(4.4)**
Δ SIR	10.9(4.6)	-1.7(5.3)	6.9(6.6)	11.2(4.4)	**6.8(7.3)**

Original dictionary

Table 2. Mean (standard deviation) performance in dB for separated signals with and without dictionary adaptation for test dataset with 4 speech signals. S1, S2, S3 and S4 indicate the performance averaged over multiple locations but for the same source.

Tables 1 and 2 show the performance with and without adaptation when the separation algorithm is applied to both the test datasets, reporting mean and standard deviation. In the dataset with a single speech plus multiple noise sources, performance refers to the speech signal only, while for the other dataset the performance for each speaker is reported. In the first dataset a sensible average improvement in SDR can be observed compared with the original dictionary and the low deviation also indicates a much stable separation result. It is also worth noting that the SIR improvement is not so large even with the adaptation because the average SIR is already high at the input, although it may become very low in some instants where an impulse noise source becomes active. However, if the adaptation is not applied a degradation of SIR is also observed (see the negative value), because separation with wrong demixing systems tends to cancel the signal of the target speech. With the second dataset the ΔSIR improvement becomes more clear. In fact, performance are averaged over all the speech sources with signals of comparable average power. It is also important to mention that the acoustic conditions of this dataset are much more difficult than those observed in other public datasets such as the underdetermined speech dataset of SiSEC2010 [20]. Indeed sources are at considerable distance from the microphones, averagely around 2.5 meters for a maximum of about 4 meters, and therefore the direct-to-reverberant ratio is low making the estimation of the full mixing system very difficult. Furthermore, since for each mixture the source locations were randomly chosen, sources may be very close to each other.

7. CONCLUSIONS

This article discusses a novel framework for semi-blind source separation based on a tight combination of sparse learning and blind system identification. A modified matching pursuit algorithm is used to select from a predefined dictionary the mixing parameters of multiple sources best matching with the observed data. The dictionary is initialized according to prior environmental geometrical knowledge and adapted on-line with the incoming data through a weighted Natural Gradient. It is shown that the spatio-temporal adaptation mitigates the mismatch between the true mixing systems and the simulated geometrical models, which is cause of high distortion in the separated signals.

Future investigations will focus on improvements in the used geometrical model and on alternative dictionary adaptation strategies.

8. REFERENCES

[1] B. Loesch and B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," in *Proceedings of LVA/ICA*. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 1–8.

[2] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined convolutive blind source separation using spatial covariance models," in *Proc. ICASSP*, 2010.

[3] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proc. ICA*, 2009.

[4] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2165–2173, 2006.

[5] F. Nesta and M. Omologo, "Convolutive underdetermined sources separation through weighted interleaved ica and spatio-temporal correlation," in *Proceeding of LVA/ICA*, March 2012.

[6] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[7] E. Vincent, "Complex nonconvex l_p norm minimization for underdetermined source separation," in *Proc. ICA*, 2007, pp. 430–437.

[8] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," in *IEEE proceedings*, vol. 98, no. 6, June 2010, pp. 1045–1057.

[9] M. Elad, *Sparse and redundant representation, from theory to application in signal and image processing*. Springer, 2010.

[10] J. Màlek, Z. Koldovskỳ, and P. Tichavskỳ, "Semi-blind source separation based on ica and overlapped speech detection," in *Latent Variable Analysis and Signal Separation*, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds. Springer Berlin / Heidelberg, 2012, vol. 7191, pp. 462–469.

[11] A. Asaie, H. Bourlard, and V. Cevher, "Model-based compressive sensing for multi-party distant speech recognition," in *Proceeding of ICASSP*, 2011, pp. 4600–4603.

[12] M. Fakhry and F. Nesta, "Underdetermined source detection and separation using a normalized multichannel spatial dictionary," *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, pp. 1 –4, sept. 2012.

[13] J. Allen and D. Berkeley, "Image method for efficiently simulating small-room acoustics," *Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.

[14] Y. Pati, R. Rezaifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *27th Asilomar Conference Signals, Systems and Computer*, November 1993.

[15] F. Nesta, M. Matassoni, and R. F. Astudillo, "A flexible spatial blind source extraction framework for robust speech recognition in noisy environments," in *In Proc. of CHIME*, Vancouver, 2013.

[16] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America*, vol. 124(1), pp. 269–277, Jul. 2008.

[17] H. Sawada, S. Araki, and R. Makino, "Blind extraction of dominant target sources using ica and time-frequency masking," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2165–2173, November 2006.

[18] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of International Symposium on ICA and Blind Signal Separation*, San Diego, CA, USA, Dec. 2001.

[19] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 236–243, 1984.

[20] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duang, "The 2010 signal separation evaluation campaign (SiSEC2010) –part II–: Audio source separation challenges," in *Proc. LVA/ICA*, 2010.