PHASE-OPTIMIZED K-SVD FOR SIGNAL EXTRACTION FROM UNDERDETERMINED MULTICHANNEL SPARSE MIXTURES

Antoine Deleforge and Walter Kellermann
University of Erlangen-Nuremberg, Germany

ABSTRACT

We propose a novel sparse representation for heavily underdetermined multichannel sound mixtures, *i.e.*, with much more sources than microphones. The proposed approach operates in the complex Fourier domain, thus preserving spatial characteristics carried by phase differences. We derive a generalization of K-SVD which jointly estimates a dictionary capturing both spectral and spatial features, a sparse activation matrix, and all instantaneous source phases from a set of signal examples. The dictionary can then be used to extract the learned signal from a new input mixture. The method is applied to the challenging problem of ego-noise reduction for robot audition. We demonstrate its superiority relative to conventional dictionary-based techniques using recordings made in a real room.

1. INTRODUCTION

Most interesting signals are *structured*. This is what distinguishes them from mere random noise. This structure can often be expressed in terms of *sparsity* in a particular basis. More precisely, if $\mathbf{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_T]$ represents T signal examples, there must exist a set of K atoms or a *dictionary* $\mathbf{D} = [\boldsymbol{d}_1, \dots, \boldsymbol{d}_K]$ such that each signal is a linear combination of only a few atoms, *i.e.*, $\mathbf{Y} \approx \mathbf{D}\mathbf{X}$ where $\mathbf{X}$ is sparse. Estimating $\mathbf{D}$ and $\mathbf{X}$ from $\mathbf{Y}$ is a sparse instance of *matrix factorization*. In audio signal processing, it is natural to seek such a factorization in the non-negative *power spectral density* (PSD) domain, since the magnitude spectra of natural sounds such as speech often feature redundancy and sparsity. This approach gave rise to a large number of methods for audio signal representation and extraction within the framework of *non-negative matrix factorization* [1–4]. In contrast, complex spectra are usually considered uninformative and therefore not investigated.

While single-channel signals are well represented by their PSD only, disregarding the phase comes with a substantial loss of information in multichannel signals. Indeed, phase differences carry important spatial cues. For this reason, nearly all existing NMF-based methods are limited to monaural sound processing, although recent multichannel extensions of NMF have been proposed for music signal separation [3, 4]. These method do not rely on sparsity and must be tuned to a known, relatively small number of target sources.

In this paper, we propose a new sparse representation for multichannel signals in the complex Fourier domain. The key novelty is to *estimate* the instantaneous phases of all involved signal spectra instead of ignoring them. The proposed decomposition may be viewed as blindly unmixing a mixture of $K \gg M$ sources where M is the number of microphones. The fundamental assumption is here that each source contributes a specific complex-valued spectral component to the observation, and that the gain and the activation of each source is sparse along the time axis. We first derive a phase-optimized generalization of the well-known *orthogonal matching pursuit* (OMP) method [5, 6]. This generalization allows sparse coding of a multichannel signal $\boldsymbol{y}_t$ given a dictionary $\mathbf{D}$, independently of instantaneous source phases. Moreover, we show that an optimal dictionary $\mathbf{D}$ as well as all instantaneous source phases and sparse activations can be blindly estimated from a set of signal examples $\mathbf{Y}$ only. This is achieved by deriving a phase-optimized generalization of the popular K-SVD algorithm [7].

The proposed representation is applied to noise reduction in the context of a humanoid robot producing self-noise ('*ego-noise*') when performing motor actions [8]. This problem is extremely challenging for two reasons: First, the noise signal is highly non-stationary due to fast and irregular motions and collisions. This rules out the use of conventional spectral-subtraction methods such as [9]. Second, ego-noise signals exhibit nonzero spatial coherence between microphones [8]. However, they cannot be modeled by a single point source, nor even by a small set of point sources. In the case of a walking robot, clicks generated by collisions with the floor as well as full body and microphone movements are producing sounds arriving at the microphones from many directions with unknown transfer functions. This seriously limits the usefulness of spatial filtering methods such as beamforming [10] or blind source separation [11, 12].

On the positive side, motor noise signals are strongly structured. This has been exploited using noise template databases [13, 14]. These approaches are based on vector quantization, which can be seen as a particular instance of K-SVD [7]. Some approaches estimate the instantaneous noise PSD using Gaussian process models [15] or neural networks [13]. These methods rely on synchronized motor state information, which may not be reliably available in practice.

The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2 013) under grant agreement n° 609465 (EARS project).

2. PHASE-OPTIMIZED DICTIONARY LEARNING

2.1. Modeling Large and Sparse Multichannel Mixtures

Let $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T] \in \mathbb{C}^{MF \times T}$ be an observed $M-$channel spectrogram with F frequency bins and T time frames. We use the decomposition $\boldsymbol{y}_t = [\boldsymbol{y}_{1t}^\top, \ldots, \boldsymbol{y}_{Ft}^\top]^\top$ where $\boldsymbol{y}_{ft} \in \mathbb{C}^M$ is the captured M-channel signal at (f, t). We assume that $\mathbf{Y}$ is the recording of a finite but potentially large mixture of K sound sources, each emitting a specific spectral shape. One intuitive interpretation of this model in the target application of a robot recording its own noise is that the sources correspond to all possible sounds that can be emitted from the various mechanical parts of the moving robot, effectively forming a signal with an intricate spatial distribution. We denote by $\boldsymbol{a}_{fk} \in \mathbb{C}^M$ the transfer function from source k to the M microphones at frequency f. We denote by $\phi_{ft,k} \in \mathbb{C}$ with $|\phi_{ft,k}| = 1$ the instantaneous phase of source k at (f, t). At time t, we assume that each source k emits a fixed magnitude spectrum $\boldsymbol{p}_k = [p_{1k}^\top, \ldots, p_{Fk}^\top]^\top \in \mathbb{R}^{+F}$ multiplied by an *activation factor* (gain) $x_{kt} \geq 0$. A central assumption of our model is that the activation vector $\boldsymbol{x}_t = [x_{1t}^\top, \ldots, x_{Kt}^\top]^\top$ is *sparse, i.e.,* only a small number $S_{\text{max}} \ll K$ of sources is active at time t, and $\boldsymbol{x}_t$ has at most S_{max} nonzero elements ($\|\boldsymbol{x}_t\|_0 \leq S_{\text{max}}$). For all f and t, the mixing model reads:

$$\boldsymbol{y}_{ft} = \sum_{k=1}^{K} \phi_{ft,k} p_{fk} \boldsymbol{a}_{fk} x_{kt} + \boldsymbol{e}_{ft}, \tag{1}$$

where $\boldsymbol{e}_{ft} \in \mathbb{C}^M$ represents some residual noise at (f, t). To simplify this expression, the transfer function and the magnitude spectrum of source k at frequency f are combined into a single vector $\boldsymbol{d}_{fk} = p_{fk} \boldsymbol{a}_{fk} \in \mathbb{C}^M$. We denote by $\mathbf{D} \in \mathbb{C}^{MF \times K}$ the signal's *dictionary* whose columns or *atoms* are the vectors $\boldsymbol{d}_k = [\boldsymbol{d}_{1k}^\top, \ldots, \boldsymbol{d}_{Fk}^\top]^\top \in \mathbb{C}^{MF}$. In practice, each atom k can be normalized so that $\|\boldsymbol{d}_k\|_2 = 1$ and that each entry of $\boldsymbol{d}_k$ associated to the first channel is real-valued and positive. This comes without loss of generality because the source activations and instantaneous phases compensate this normalization in equation (1).

Let $\mathbf{\Phi}_t = [\boldsymbol{\phi}_{t,1}, \ldots, \boldsymbol{\phi}_{t,K}] \in \mathbb{C}^{F \times K}$ denote the matrix of all source phases at frame t. We define the *phase-corrected dictionary* at frame t by:

$$\mathbf{D}\{\mathbf{\Phi}_t\} = \begin{bmatrix} \phi_{1t,1}\boldsymbol{d}_{11} & \ldots & \phi_{1t,K}\boldsymbol{d}_{1K} \\ \vdots & & \vdots \\ \phi_{Ft,1}\boldsymbol{d}_{F1} & \ldots & \phi_{Ft,K}\boldsymbol{d}_{FK} \end{bmatrix} \in \mathbb{C}^{MF \times K}. \tag{2}$$

The model (1) can now be rewritten as $\boldsymbol{y}_t = \mathbf{D}\{\mathbf{\Phi}_t\}\boldsymbol{x}_t + \boldsymbol{e}_t$ where $\boldsymbol{x}_t$ is a sparse vector. If $\mathbf{D}\{\mathbf{\Phi}_t\}$ is known, estimating $\boldsymbol{x}_t$ in order to minimize $\boldsymbol{e}_t$ given $\boldsymbol{y}_t$ is known as a *sparse coding* problem [6, 7]. However in the considered case of a multichannel sound mixture, this would require the prior knowledge of not only the K sources' transfer functions and magnitude spectra contained in $\mathbf{D}$, but also their instantaneous phases contained in $\mathbf{\Phi}_t$. The latter randomly varies over time, is hard to predict, and should therefore be estimated from the signal. This yields the following novel optimization problem, which will be referred to as *phase-optimized sparse coding*:

$$\begin{aligned} &\underset{\mathbf{\Phi}_t, \boldsymbol{x}_t}{\operatorname{argmin}} \|\boldsymbol{y}_t - \mathbf{D}\{\mathbf{\Phi}_t\}\boldsymbol{x}_t\|_2 \quad \text{subject to:} \\ &\|\boldsymbol{x}_t\|_0 \leq S_{\text{max}} \text{ and } \forall\, f, k, x_{kt} \geq 0 \;\; , |\phi_{ft,k}| = 1. \end{aligned} \tag{3}$$

Note that due to the sparsity of $\boldsymbol{x}_t$, only those $\phi_{ft,k}$ for which $x_{kt} > 0$ intervene in the target function. The others can be ignored, leading to a sparse matrix $\mathbf{\Phi}_t$. Moreover, the non-negativity constraint on x_{kt} is not necessary, since for any complex pair $(x_{kt}, \boldsymbol{\phi}_{t,k})$, the pair $(|x_{kt}|, \frac{x_{kt}}{|x_{kt}|}\boldsymbol{\phi}_{t,k})$ leaves the cost function unchanged. This constraint is thus relaxed in the remainder of the paper.

2.2. Phase-Optimized Orthogonal Matching-Pursuit

Although finding an exact solution to sparse coding was proven to be NP-hard [16], a number of efficient approximate methods have been proposed [5, 6, 17, 18], among which *orthogonal matching pursuit* (OMP) [5, 6] is one of the most widely used due to its simplicity and high practical efficiency. In this section, we propose an algorithm inspired by OMP that addresses the phase-optimized sparse coding problem (3). This is referred to as *phase-optimized orthogonal matching pursuit* (PO-OMP) and summarized in Alg. 1.

Similarly to OMP, PO-OMP is a greedy algorithm that selects the best matching dictionary atom, indexed by $k(i)$, at each iteration i. This is repeated either until i reaches a maximum desired sparsity number S_{max} or when the cost function (3) falls below a desired reconstruction threshold τ. To avoid carrying large sparse matrices, we use the variables $\widetilde{\boldsymbol{x}}_t^{(i)} \in \mathbb{C}^i$, $\widetilde{\mathbf{\Phi}}_t^{(i)} \in \mathbb{C}^{F \times i}$ and $\widetilde{\mathbf{D}}^{(i)} \in \mathbb{C}^{MF \times i}$. They respectively correspond to $\boldsymbol{x}_t$, $\mathbf{\Phi}_t$ and $\mathbf{D}$ in which only rows or columns indexed by $k(1) \ldots k(i)$ are kept. Let $\boldsymbol{r}_t^{(i)} = [\boldsymbol{r}_{1t}^{(i)\top}, \ldots, \boldsymbol{r}_{Ft}^{(i)\top}]^\top \in \mathbb{C}^{MF}$ be the residual vector at iteration i, *i.e.,* $\boldsymbol{r}_t^{(i)} = \boldsymbol{y}_t - \widetilde{\mathbf{D}}^{(i)}\{\widetilde{\mathbf{\Phi}}_t^{(i)}\}\widetilde{\boldsymbol{x}}_t^{(i)}$ and $\boldsymbol{r}_t^{(0)} = \boldsymbol{y}_t$. As in OMP, each iteration of PO-OMP consists of two steps. In the first step, the dictionary atom that best approximates the current residual is found. This requires to solve:

$$\underset{k,\, \boldsymbol{\phi}_{t,k},\, x_{kt}}{\operatorname{argmin}} \|\boldsymbol{r}_t^{(i-1)} - \boldsymbol{d}_k\{\boldsymbol{\phi}_{t,k}\}x_{kt}\|_2 \text{ s.t. } \forall f,\ |\phi_{ft,k}| = 1 \tag{4}$$

where $\boldsymbol{d}_k\{\boldsymbol{\phi}_{t,k}\}$ denotes the k-th column of $\mathbf{D}\{\mathbf{\Phi}_t\}$. Using the Lagrange multiplier method to enforce the constraints on $\boldsymbol{\phi}_{t,k}$, the solution of (4) is obtained through lines 4-8 of Alg. 1. In the second step, all values in $\widetilde{\boldsymbol{x}}_t^{(i)}$ and $\widetilde{\mathbf{\Phi}}_t^{(i)}$, including values found in previous iterations, are optimized according to the i atoms selected so far. This requires to solve:

Algorithm 1 PO-OMP

Input: Signal $\boldsymbol{y}_t \in \mathbb{C}^{MF}$, dictionary $\mathbf{D} \in \mathbb{C}^{MF \times K}$, sparsity number S_{max} and reconstruction threshold τ.
Output: Sparse activation vector $\boldsymbol{x}_t \in \mathbb{R}^{+K}$ and sparse phase corrections $\boldsymbol{\Phi}_t \in \mathbb{C}^{F \times K}$ so that $\boldsymbol{y}_t \approx \mathbf{D}\{\boldsymbol{\Phi}_t\}\boldsymbol{x}_t$.

1: $\widetilde{\boldsymbol{x}}_t^{(0)} := [\,]$; $\widetilde{\boldsymbol{\Phi}}_t^{(0)} := [\,]$; $\widetilde{\mathbf{D}}^{(0)} := [\,]$; $\boldsymbol{r}_t^{(0)} := \boldsymbol{y}_t$; $i := 0$;
2: **while** $i \leq S_{\text{max}}$ and $\|\boldsymbol{r}_t^{(i)}\|_2 > \tau$ **do**
3: $\quad i := i + 1$;
4: $\quad \forall f, k,\ b_{fk} := \langle \boldsymbol{r}_{ft}^{(i-1)} | \boldsymbol{d}_{fk} \rangle$;
5: $\quad \forall k,\ c_k := |\sum_{f=1}^{F} b_{fk} |b_{fk}|^{-1}|$;
6: $\quad k(i) := \text{argmax}_k(c_k)$;
7: $\quad \widetilde{\boldsymbol{x}}_t^{(i)} := [\widetilde{\boldsymbol{x}}_t^{(i-1)\top}, c_{k(i)}]^\top$;
8: $\quad \forall f,\ \widetilde{\boldsymbol{\phi}}_{ft}^{(i)} := [\widetilde{\boldsymbol{\phi}}_{ft}^{(i-1)}, b_{fk(i)} |b_{fk(i)}|^{-1}]$;
9: $\quad \widetilde{\mathbf{D}}^{(i)} := [\widetilde{\mathbf{D}}^{(i-1)}, \boldsymbol{d}_{k(i)}]$;
10: $\quad$ **repeat**
11: $\quad\quad \widetilde{\boldsymbol{x}}_t^{(i)} := (\widetilde{\mathbf{D}}^{(i)}\{\widetilde{\boldsymbol{\Phi}}_t^{(i)}\})^\dagger \boldsymbol{y}_t$; *// ($\dagger$ = pseudo-inverse)*
12: $\quad\quad \boldsymbol{r}_t^{(i)} := \boldsymbol{y}_t - \widetilde{\mathbf{D}}^{(i)}\{\widetilde{\boldsymbol{\Phi}}_t^{(i)}\}\widetilde{\boldsymbol{x}}_t^{(i)}$;
13: $\quad\quad \forall j, f,\ \widetilde{\phi}_{ft,j}^{(i)} := \dfrac{\langle \boldsymbol{r}_{ft}^{(i)} | \boldsymbol{d}_{fk(j)} \rangle + \widetilde{\phi}_{ft,j}^{(i)} \widetilde{x}_{jt}^{(i)}}{|\langle \boldsymbol{r}_{ft}^{(i)} | \boldsymbol{d}_{fk(j)} \rangle + \widetilde{\phi}_{ft,j}^{(i)} \widetilde{x}_{jt}^{(i)}|}$;
14: $\quad$ **until** $\Delta(\|\boldsymbol{r}_t^{(i)}\|_2) < \epsilon$
15: **end while**
16: **return** sparse $\boldsymbol{x}_t$ and $\boldsymbol{\Phi}_t$ obtained from $\widetilde{\boldsymbol{x}}_t^{(i)}$ and $\widetilde{\boldsymbol{\Phi}}_t^{(i)}$.

$$\underset{\widetilde{\boldsymbol{x}}_t^{(i)}, \widetilde{\boldsymbol{\Phi}}_t^{(i)}}{\text{argmin}}\ \|\boldsymbol{y}_t - \widetilde{\mathbf{D}}^{(i)}\{\widetilde{\boldsymbol{\Phi}}_t^{(i)}\}\widetilde{\boldsymbol{x}}_t^{(i)}\|_2 \text{ s.t. } \forall f, j,\ |\widetilde{\phi}_{ft,j}^{(i)}| = 1.$$

We could not find a general closed-form solution to this problem. However, it can be solved iteratively by sequentially minimizing the objective function with respect to $\widetilde{\boldsymbol{x}}_t^{(i)}$ and each column of $\widetilde{\boldsymbol{\Phi}}_t^{(i)}$ separately. Convergence is considered reached when the relative variation of the residual error $\Delta(\|\boldsymbol{r}_t^{(i)}\|_2)$ falls below a preset threshold ϵ, *e.g.*, less than 0.1%. The values of $\widetilde{\boldsymbol{x}}_t^{(i)}$ and $\widetilde{\boldsymbol{\Phi}}_t^{(i)}$ found in previous iterations provide a good initialization for this procedure. Closed form solutions for this sequential minimization are given in lines 11-13 of Alg. 1. In practice, the residual vector $\boldsymbol{r}_t^{(i)}$ is reupdated after each new estimation of $\widetilde{\boldsymbol{\phi}}_{t,j}^{(i)}$ in order to improve convergence. The overall algorithm is guaranteed to decrease the residual error $\|\boldsymbol{r}_t^{(i)}\|_2$ at each step. As in OMP, a local minimum may be reached due to the non-convexity of the problem. However, OMP is known to perform well if $S_{\text{max}} \ll K$, and the same was observed with PO-OMP.

2.3. Phase-Optimized K-SVD

PO-OMP requires a known dictionary $\mathbf{D}$, capturing the spectral shapes and transfer functions of the K sources in the mixture. This may not be available in practice. This section addresses the challenging problem of *training* such a dictionary, based on a set of examples $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T] \in \mathbb{C}^{MF \times T}$. More formally, we seek a solution to:

$$\underset{\mathbf{X}, \mathbf{D}, \boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_T}{\text{argmin}} \sum_{t=1}^{T} \|\boldsymbol{y}_t - \mathbf{D}\{\boldsymbol{\Phi}_t\}\boldsymbol{x}_t\|_2^2 \quad \text{subject to:}$$
$$\|\boldsymbol{x}_t\|_0 \leq S_{\text{max}} \text{ and } \forall\ f, k, |\phi_{ft,k}| = 1. \tag{5}$$

This is reminiscent of a sparse dictionary learning problem [19], except that $\mathbf{D}$ is corrected by $\boldsymbol{\Phi}_t$ at each t. Dictionary learning has been widely investigated, and the most popular method is probably K-SVD [7], due to its simplicity and high efficiency. Following these lines, we propose a method that solves for (5), referred to as *phase-optimized K-SVD* (PO-KSVD). The corresponding algorithm is summarized in Alg. 2.

Algorithm 2 PO-KSVD

Input: Signal examples $\mathbf{Y} \in \mathbb{C}^{MF \times T}$, sparsity number S_{max} and reconstruction threshold τ.
Output: Matrices $\mathbf{D} \in \mathbb{C}^{MF \times K}$, $\mathbf{X} \in \mathbb{R}^{+K \times T}$ (sparse) and $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_t \in \mathbb{C}^{F \times K}$ (sparse) so that $\boldsymbol{y}_t \approx \mathbf{D}\{\boldsymbol{\Phi}_t\}\boldsymbol{x}_t\ \forall t$.

1: Initialize $\mathbf{D}$ with K normalized, random columns of $\mathbf{Y}$;
2: **repeat**
3: $\quad \forall t,\ [\boldsymbol{x}_t, \boldsymbol{\Phi}_t] = \text{po\_omp}(\boldsymbol{y}_t, \mathbf{D}, S_{\text{max}}, \tau)$;
4: $\quad \forall k,\ \boldsymbol{s}_*^k = \text{sparse}(\boldsymbol{x}_*^k)$; *// Non-zero indicator of $\boldsymbol{x}_*^k$*
5: $\quad$ **for** $k = 1 \rightarrow K$ **do**
6: $\quad\quad$ **repeat**
7: $\quad\quad\quad$ Compute $\mathbf{E}_k$; *// Large-bracketed term in (6)*
8: $\quad\quad\quad$ Obtain $\boldsymbol{d}_k$ and $\boldsymbol{x}_*^k$ from $\text{svd}(\mathbf{E}_k\{\bar{\boldsymbol{\Phi}}_*^k\}_{/\{\boldsymbol{s}_*^k\}})$;
9: $\quad\quad\quad \forall f, t,\ \phi_{ft,k} = \dfrac{\langle \boldsymbol{e}_{ft,k} | \boldsymbol{d}_{fk} x_{kt} \rangle}{|\langle \boldsymbol{e}_{ft,k} | \boldsymbol{d}_{fk} x_{kt} \rangle|}$;
10: $\quad\quad$ **until** $\Delta(\|\mathbf{E}_k\|_F) < \epsilon$
11: $\quad$ **end for**
12: **until** $\Delta(\Sigma_{t=1}^{T} \|\boldsymbol{y}_t - \mathbf{D}\{\boldsymbol{\Phi}_t\}\boldsymbol{x}_t\|_2^2) < \epsilon$
13: **return** matrices $\mathbf{D}$, $\mathbf{X}$ and $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_t$.

Similarly to K-SVD, PO-KSVD alternates between a sparse-coding step, *i.e.*, (3) and a *dictionary update* step. Since the former is solved by PO-OMP, we now focus on the latter. The key idea responsible for the efficiency of K-SVD is to sequentially update each atom and associated activations, while preserving the non-zero support of $\mathbf{X}$ found during the sparse-coding step. Let $\boldsymbol{x}_*^k$ denote the k-th row vector of $\mathbf{X}$ and $\boldsymbol{s}_*^k = \text{sparse}(\boldsymbol{x}_*^k) \in \{0;1\}^{1 \times T}$ denote the binary row vector indicating the non-zero elements of $\boldsymbol{x}_*^k$ after PO-OMP. Let $\boldsymbol{\Phi}_*^k = \{\phi_{ft,k}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F \times T}$ denote the sparse matrix of source k's instantaneous phases. For each atom k, the associated optimization problem can be written:

$$\underset{\boldsymbol{d}_k, \boldsymbol{x}_*^k, \boldsymbol{\Phi}_*^k}{\text{argmin}} \left\| \left(\mathbf{Y} - \sum_{j \neq k} (\boldsymbol{d}_j \boldsymbol{x}_*^j)\{\boldsymbol{\Phi}_*^j\} \right) - (\boldsymbol{d}_k \boldsymbol{x}_*^k)\{\boldsymbol{\Phi}_*^k\} \right\|_F$$
$$\text{s.t.:}\quad \text{sparse}(\boldsymbol{x}_*^k) = \boldsymbol{s}_*^k \ \text{ and } \ \forall\ f, t,\ |\phi_{ft,k}| = 1. \tag{6}$$

	Waving noise				Walking noise				CTS	
Method used	SDR (dB)	SIR (dB)	PESQ	CKR	SDR (dB)	SIR (dB)	PESQ	CKR	train	test
PO-KSVD+	**2.31±4.2**	**22.5±2.8**	**2.09±0.3**	**82.9**	**1.83±4.5**	**22.3±3.2**	**2.00±0.2**	**88.4**	9.99	0.59
PO-KSVD	1.38±3.9	14.4±3.5	2.06±0.4	81.1	1.45±4.3	19.8±3.6	1.80±0.2	87.8	9.99	0.59
NMF	0.07±2.6	7.01±4.8	1.38±0.2	50.6	1.62±3.2	17.9±3.1	1.51±0.2	65.2	4.13	0.01
K-SVD	-3.91±3.6	-1.31±4.2	1.46±0.3	45.7	1.10±4.4	6.08±2.4	1.38±0.1	70.1	0.22	0.04
mixture	-5.37±4.0	-3.87±4.8	1.42±0.3	43.9	0.76±4.2	4.78±2.4	1.33±0.1	67.1	-	-

Table 1.Average and standard deviations (Avg±Std) of the signal-to-distortion-ratios (SDR), signal-to-interfer-ratios (SIR), PESQ measures and correct keyword recognition rates in % (CKR) over 82 target speech signals, for waving and walking noises. The last columns show the average computation times (in secs) per second of signal (CTS) for training and testing the methods using MATLAB on a conventional PC.

Here, $\|.\|_F$ denotes the Frobenius norm. If $\mathbf{E}_k$ denotes the matrix between large brackets, the above cost function is equal to $\|\mathbf{E}_k\{\bar{\mathbf{\Phi}}_*^k\} - \boldsymbol{d}_k\boldsymbol{x}_*^k\|_F$ where $\bar{\mathbf{\Phi}}_*^k$ denotes the complex conjugate of $\mathbf{\Phi}_*^k$, and $\boldsymbol{d}_k\boldsymbol{x}_*^k$ is an $MF \times T$ rank-1 matrix. For fixed phases $\mathbf{\Phi}_*^k$, the solution of $\boldsymbol{d}_k$ and $\boldsymbol{x}_*^k$ is obtained via singular value decomposition (SVD) of $\mathbf{E}_k\{\bar{\mathbf{\Phi}}_*^k\}_{/\{\boldsymbol{s}_*^k\}}$ where $/\{\boldsymbol{s}_*^k\}$ means that columns corresponding to $s_{kt} = 0$ have been removed (more details in [7]). For fixed $\boldsymbol{d}_k$ and $\boldsymbol{x}_*^k$, the update of $\mathbf{\Phi}_*^k$ is closed-form, and provided at line 9 of Alg. 2. This sequential minimization is iterated until convergence of $\|\mathbf{E}_k\|_F$, similarly to Alg. 1. As in K-SVD, the convergence of PO-KSVD relies on the ability of PO-OMP to decrease the residual error with respect to the dictionary update solution. While this is not guaranteed, it can be solved by *external inference*, *i.e.*, for each t, the output of PO-OMP is only kept if it improves the reconstruction of $\boldsymbol{y}_t$. Convergence is then guaranteed.

3. EXPERIMENTAL RESULTS

The commercial robot NAO of Aldebaran Robotics [20] was used to gather 4-channel recordings downsampled to 16 kHz in a real room with moderate reverberation level (T60≈200ms). Two one-minute recordings of NAO walking on place or repeatedly waving the right arm were used for training. The fan of the robot was on, resulting in additional stationary ego-noise which was reduced using multichannel Wiener filtering, as described in [8]. For testing, 82 recordings lasting approximately 1s each of a loudspeaker emitting speech utterances from the GRID corpus [21] were made with the fan turned off. The loudspeaker was placed 1 meter away in front of the robot, at null elevation. These speech recordings were summed with out-of-training waving or walking sequences to generate test mixtures. Spectrograms were computed using 64ms Hamming windows with 50% overlap.

PO-KSVD was used to learn a dictionary for each of the two training signals, using several values of $K \in [5, 100]$ and $S_{\text{max}} \in [1, 5]$. Best performances were obtained with $K = 40$, $S_{\text{max}} = 3$ for waving, $K = 10$, $S_{\text{max}} = 2$ for walking. The reconstruction threshold was fixed to a low value $\tau = 10^{-4}$. Once ego-noise dictionaries were trained, the ego-noise signals were estimated from test mixtures using PO-OMP. The residuals were used as desired speech output. To further improve output signals, time-frequency points for which the residual PSD was less than the estimated ego-noise PSD were set to the average background noise magnitude, while preserving the phase. This masking technique is referred to as PO-KSVD+.

PO-KSVD was compared to conventional K-SVD [7] using the same protocol and parameters. We also compared it to NMF using the versatile implementation provided by [22]. As suggested in [1], the magnitude spectra of the left microphone signal raised to the power 0.7 were used as input. The term $\lambda\|\mathbf{X}\|_1$ was added to the conventional NMF cost function to enforce sparsity. Several values of $\lambda \in [0, 4]$ and dictionary sizes $K \in [5, 40]$ were tested, and best results were obtained with $\lambda = 2, K = 20$ for waving and $\lambda = 1, K = 40$ for walking. Once a non-negative dictionary is trained, a single NMF multiplicative update can be used to estimate the ego-noise PSD from a test signal. The residuals are then used as desired magnitude spectra, while the mixture phases are preserved.

Table 1 summarizes the signal-to-distortion and signal-to-interfer ratios SDR and SIR [23], as well as the PESQ measure [24], the correct keyword recognition rate[1] (CKR) and computational times for all methods. Scores obtained from the unprocessed mixtures are given in the last row. PO-KSVD and PO-KSVD+ significantly outperforms conventional factorization methods in terms of all the metrics used. Sound excerpts and spectrograms are provided at `robot-ears.eu/po_ksvd/`.

4. CONCLUSION

To the best of the authors' knowledge, PO-KSVD is the first method that combines sparse factorization with instantaneous phase estimation in the complex Fourier domain. This paves the road to numerous applications in multichannel audio signal processing and beyond. Compared to traditional monaural approaches, this methods preserves and exploits spatial cues. In the future, we plan to further investigate this by adding spatial constraints to the dictionary in order to achieve underdetermined blind source separation and localization.

[1]The speech recognizer *pocketsphinx* [25] was used to recognize the keywords in the GRID corpus [21], as defined by the CHiME challenge [26].

5. REFERENCES

[1] M.N. Schmidt, J. Larsen, and F. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Workshop on Machine Learning for Signal Processing*. IEEE, 2007, pp. 431–436.

[2] K.W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising.," in *INTERSPEECH*, 2008, pp. 411–414.

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

[4] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[5] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *27th Asilomar Conference on Signals, Systems and Computers*. IEEE, 1993, pp. 40–44.

[6] J.A. Tropp and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[7] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[8] H. Löllmann, H. Barfuss, A. Deleforge, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *ITG Fachtagung Sprachkommunikation*, September 2014, p. 4.

[9] Rainer Martin, "Spectral subtraction based on minimum statistics," in *Proc. Eur. Signal Processing Conf.*, 1994, pp. 1182–1185.

[10] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing*, pp. 155–194. Springer, 2003.

[11] H. Buchner, R. Aichner, and W. Kellermann, "Trinicon: A versatile framework for multichannel blind signal processing," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*. IEEE, 2004, vol. 3, pp. 889–892.

[12] M.I. Mandel, R.J. Weiss, and D.P.W. Ellis, "Model-based expectation-maximization source separation and localization," *Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[13] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, "Ego noise suppression of a robot using template subtraction," in *Int. Conf. on Intelligent Robots and Systems*. IEEE/RSJ, 2009, pp. 199–204.

[14] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *Int. Conf. on Robotics and Automation*. IEEE, 2011, pp. 3517–3522.

[15] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H.G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor uav," in *Int. Conf. on Intelligent Robots and Systems*. IEEE/RSJ, 2013, pp. 3943–3948.

[16] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constructive approximation*, vol. 13, no. 1, pp. 57–98, 1997.

[17] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *Journal on scientific computing*, vol. 20, no. 1, pp. 33–61, 1998.

[18] P.O. Hoyer, "Non-negative sparse coding," in *Workshop on Neural Networks for Signal Processing*. IEEE, 2002, pp. 557–565.

[19] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*, Springer, 2010.

[20] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "The nao humanoid: a combination of performance and affordability," *CoRR abs/0807.3223*, 2008.

[21] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[22] Y. Li and A. Ngom, "Versatile sparse matrix factorization and its applications in high-dimensional biological data analysis," in *Pattern Recognition in Bioinformatics*, pp. 91–101. Springer, 2013.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*. IEEE, 2001, vol. 2, pp. 749–752.

[25] D. Huggins-Daines, M. Kumar, A. Chan, A.W. Black, M. Ravishankar, and A.I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*. IEEE, 2006, vol. 1, pp. 185–188.

[26] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*. IEEE, 2013, pp. 126–130.