

Binary Sparse Coding of Convolutional Mixtures for Sound Localization and Separation via Spatialization

Afsaneh Asaei, *Member, IEEE*, Mohammad J. Taghizadeh, *Student Member, IEEE*

Saeid Haghighatshoar, *Member, IEEE*, Bhiksha Raj, *Senior Member, IEEE* Hervé Bourlard, *Fellow, IEEE*

Volkan Cevher, *Senior Member, IEEE*

Abstract—We propose a sparse coding approach to address the problem of source-sensor localization and speech reconstruction. This approach relies on designing a dictionary of spatialized signals by projecting the microphone array recordings into the array manifolds characterized for different locations in a reverberant enclosure using the image model. Sparse representation over this dictionary enables identifying the subspace of the actual recordings and its correspondence to the source and sensor locations. The speech signal is reconstructed by inverse filtering the acoustic channels associated to the array manifolds. We provide rigorous analysis on the optimality of speech reconstruction by elucidating the links between inverse filtering and source separation followed by deconvolution. This procedure is evaluated for localization, reconstruction and recognition of simultaneous speech sources using real data recordings. The results demonstrate the effectiveness of the proposed approach and compare favorably against beamforming and independent component analysis techniques.

Index Terms—Sparse coding, Reverberation, Source localization and separation, Multiparty (Overlapping) speech recognition, Microphone array, Sound spatialization.

I. INTRODUCTION

Many of the sound technologies rely on distant signal acquisition using an array of microphones deployed in the acoustic scene. This flexible acquisition setup requires effective methodologies to recover the desired information from an acoustic clutter of unknown source signals. The forward model of distant sound recordings is often approximated as a linear operation through $\mathcal{X} = \mathcal{H}\mathcal{S}$ where $\mathcal{S} \in \mathbb{C}^{N \times 1}$ denotes N original source signals; $\mathcal{H} \in \mathbb{C}^{M \times N}$ consists of the microphone array manifold vectors associated to N sources with respect to M microphones and $\mathcal{X} \in \mathbb{C}^{M \times 1}$ represents the microphone array signals. We work with the frequency domain spectral representation of the signals. Our goal is to recover the

original source signals from the mixture recordings provided by the microphone array.

The distant microphones enable spatial sampling through projection of the spectral information embedded in the acoustic scene into the spatial subspaces of microphone array manifold vectors. In this context, source localization and separation (demixing) amounts to recovering the spatio-spectral representation of the individual sources. This problem is known to be ill-posed, thus prior knowledge/assumptions on either the characteristics/structure of the sources or the acoustic mixing channels are often exploited.

A plethora of demixing methodologies rely on statistical independence of the individual sources referred to as independent component analysis (ICA). The separation is thus formulated as estimating a demixing matrix such that the recovered source signals are statistically independent [1]. To account for the multipath propagation, characterization of the room acoustics is integrated in the separation process [2]. However, the ICA-based separation schemes require large amount of data recorded in a stationary acoustic condition to provide a reasonable estimate of the model parameters. Also, ICA performs poorly in underdetermined case. In addition, it imposes a permutation problem due to misalignment of the individual source components [3, 4, 5]. To further extend this framework to non-invertible matrices, additional prior on sparse representation of the sources can be incorporated [6, 7].

An alternative geometric demixing strategy is devised based on the capability of a microphone array for directional acquisition or beamforming. This procedure achieves source separation by steering the beam pattern of the microphone array towards the desired source thus filtering out the interferences regardless of their signal nature [8, 9]. The underlying hypothesis is that the sources are uncorrelated; this assumption is vulnerable to reverberation so the beamformer can mitigate or cancel the desired signal in a reverberant acoustic. An additional limiting factor is the spatial resolution for resolving closely located sources. Furthermore, unlike the ICA approach, beamforming requires precise information about the microphone array configuration and the desired source location. Recent work considers non-linear mixture of beamformers which incorporates sparsity of the spectro-temporal coefficients to address the underdetermined demixing [10]. The application of this method is limited to the anechoic mixtures and the performance is degraded due to reverberation.

Afsaneh Asaei is with Idiap Research Institute, Martigny, Switzerland. Mohammad J. Taghizadeh is with Huawei European Research Center, Munich, Germany. Saeid Haghighatshoar is with Technische Universität Berlin, Germany. Bhiksha Raj is with Language Technology Institute, Carnegie Mellon University, USA. Hervé Bourlard is with Idiap Research Institute and also affiliated with École Polytechnique Fédérale de Lausanne, Switzerland. Volkan Cevher is with École Polytechnique Fédérale de Lausanne, Switzerland. E-mails: {afsaneh.asaei, herve.bourlard}@idiap.ch, mohammad.taghizadeh@huawei.com, saeid.haghighatshoar@tu-berlin.de, bhiksha.raj@cs.cmu.edu, volkan.cevher@epfl.ch.

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Furthermore, the multi-channel non-negative matrix factorization (MC-NMF) has been studied for source separation in convolutive mixtures [11, 12, 13]. The MC-NMF considers the source model as a summation of elementary components each representative of a characteristic spectral signature. This method is particularly suited to polyphony and applied for stereo source separation of music and speech mixtures [11]. One limitation is that decomposition is applied on squared or magnitude STFT coefficients, thus the angle information is almost lost. To address this issue, MC-NMF is applied in a beamspace domain where the input data are represented by direction of arrivals (DOAs) obtained from phase information [12]. More recent developments integrate beamforming-based DOA and NMF through modeling the spatial covariance of individual sources as a weighted combination of DOA kernels where the magnitude of the sources is obtained from NMF [13]. Although this approach improves the spatial coherence of the NMF separated sources, it only exploits the direct path for DOA kernel formulation that limits its theoretical accuracy in convolutive mixture modeling.

A blind channel identification approach for speech separation and dereverberation is proposed in [14]. The mixing procedure is delineated with a multiple-input multiple-output mathematical model decomposed into sequential procedures to remove spatial interference followed by deconvolution of temporal echoes [15]. The major drawback of such implementation is that it can only perform channel identification from single talk recordings in high input signal-to-noise ratio. Another approach extends the maximum likelihood criteria applied in weighted prediction error modeling for joint dereverberation and separation of individual speech sources from determined and overdetermined mixtures [16]. This method does not require channel estimation and assumes that source spectral components are uncorrelated across time frames. It also relies on a single source assumption and thus can not achieve dereverberation when there are multiple sound sources [17].

More recently, the sparse component analysis (SCA) techniques are developed upon the assumption that the sources admit a sparse representation [18]. Sparsity indicates that the representation of signal occupies only a small part of a larger space and the mixtures of sparse components are likely to be disjoint. In practice, if the canonical representation of the sources do not hold the sparsity premise, a linear transformation of the signal, through either a generic transformation such as STFT, DCT or a data driven approach such as sparse coding dictionary [19] yields sparse representation. The initial techniques incorporate the geometric distinction to separate the sources by exploiting the delay and attenuation differences among the signals captured by microphones; this information is used to construct a binary mask and extract the individual signals [20]. The extensions of this algorithm have been proposed for convolutive mixtures in [21, 22]. New advances on SCA incorporates sparsity structures identified through the principles of auditory segregation and multipath acoustic with a model-based sparse component analysis framework [23, 24].

Earlier works in the context of sparse component analysis for source localization and separation require prior knowledge

on the microphone array geometry. In [25], we investigated the duality between source and sensor localization problems and showed that if the source location is given, the microphone array geometry can be calibrated using the sparse coding approach. In this paper, we extend our earlier framework for joint source and sensor localization. The sparsity-inducing dictionary is devised through spatialization of the speech signals. Sparse representation over this dictionary leads to identification of the array manifold corresponding to the distant signal recordings and its associated source and sensor locations in the acoustic scene. Thus, it enables source-sensor localization. To account for the multipath propagation in a reverberant environment, the image model of the acoustic is exploited through the technique elaborated in [24]. The proposed scheme yields speech reconstruction through inverse filtering the acoustic channel. The optimality of this approach is investigated by showing its equivalence to speech separation and channel deconvolution. In contrast to the prior works, designing/identifying a linear filter for source separation and dereverberation in time domain is not the objective of this paper. Instead, we process each frequency of the signal independently thus demixing is achieved in frequency domain.

The rest of the paper is organized as follows. The formulation of the source-sensor localization via speech reconstruction is stated in Section II. The sparse coding approach to tackle this problem is elaborated in Section III. The sparse coding approach relies on adaptive construction of a sparsifying dictionary via spatialization of the signals and it can be efficiently implemented through a greedy sparse recovery algorithm as explained in Section III. The equivalence of reconstruction via inverse filtering to speech separation and deconvolution is worked out in Section IV. This framework is evaluated on real data recordings for speech source localization and separation and overlapping speech recognition; the results are provided in Section V. Finally, the conclusions are drawn in Section VI.

II. PROBLEM STATEMENT

A. Signal Model

We consider a scenario in which an unknown signal $S_n(f)$ at frequency f emanates from an unknown location ν_n in an enclosure and impinges on a microphone located at μ_m . The response of the acoustic channel from the source location ν_n to the microphone location μ_m is denoted by $H(f, \nu_n, \mu_m)$ and the signal acquired by the distant microphone is

$$X_m(f) = H(f, \nu_n, \mu_m)S_n(f) \quad (1)$$

Let $\Upsilon = \{\nu_1, \dots, \nu_N\}$ denotes the position of N sources and $\Pi = \{\mu_1, \mu_2, \dots, \mu_M\}$ denotes the position of M microphones. We define

$$S(f) = [S_1(f) \cdots S_N(f)]^\top, \quad \mathcal{X}(f) = [X_1(f) \cdots X_M(f)]^\top \quad (2)$$

where \cdot^\top stands for the transpose operator; we can characterize the forward model of distant recordings as

$$\mathcal{X}(f) = \mathcal{H}_{\Upsilon, \Pi}(f)S(f), \quad (3)$$

where

$$\mathcal{H}_{\Upsilon, \Pi}(f) = \begin{bmatrix} H(f, \nu_1, \mu_1) & \cdots & H(f, \nu_N, \mu_1) \\ \vdots & \ddots & \vdots \\ H(f, \nu_1, \mu_M) & \cdots & H(f, \nu_N, \mu_M) \end{bmatrix}$$

is known as the array *narrow-band* manifold matrix which is specific to the source and microphone locations.

In this paper, we build on our earlier work on room acoustic modeling using the image method [24]. The image method indicates that the multipath channel of an enclosure is identified if the enclosure boundary, absorption coefficients of the reflective surfaces and the source and sensor locations are known. Given the room geometry and the absorption coefficients, our objective is to estimate the location of the sources (Υ) and sensors (Π) and reconstruct the original signals (\mathcal{S}) from the distant microphone recordings.

B. Spatialization Approach to Localization

In this section, we formulate a new localization approach based on sound spatialization or reconstruction of the distant speech signal. Given an estimate of the array manifold $\hat{\mathcal{H}}_{\Upsilon, \Pi}(f)$ and assuming an over-determined system, i.e., $N \leq M$, the original source signal can be reconstructed as [26]

$$\hat{\mathcal{S}}_s(f) = (\hat{\mathcal{H}}_{\Upsilon, \Pi}^* \hat{\mathcal{H}}_{\Upsilon, \Pi}(f))^{-1} \hat{\mathcal{H}}_{\Upsilon, \Pi}^* \mathcal{X}(f) = \hat{\mathcal{H}}_{\Upsilon, \Pi}^\dagger(f) \mathcal{X}(f),$$

where $*$ denotes the conjugate transpose operator and $\hat{\mathcal{H}}_{\Upsilon, \Pi}^\dagger(f)$ represents the pseudo-inverse of $\hat{\mathcal{H}}_{\Upsilon, \Pi}(f)$. Thereby, the projection of the distant recording into the array manifold is obtained as

$$\hat{\mathcal{X}}(f) = \hat{\mathcal{H}}_{\Upsilon, \Pi}(f) \hat{\mathcal{S}}_s(f) = \hat{\mathcal{H}}_{\Upsilon, \Pi}(f) \hat{\mathcal{H}}_{\Upsilon, \Pi}^\dagger(f) \mathcal{X}(f); \quad (4)$$

this now gives us an effective handle to estimate Υ, Π :

$$\Upsilon, \Pi = \arg \min_{\nu_1, \dots, \nu_N, \mu_1, \dots, \mu_M} \|\mathcal{X}(f) - \hat{\mathcal{X}}(f)\|_2^2, \quad (5)$$

For processing a broadband signal such as speech, the above formulation can be succinctly stated in matrix form as follows. Let $\mathcal{F} = \{f_1, f_2, \dots, f_B\}$ represent a set of B adjacent frequencies. The *broad-band* array manifold matrix $\mathcal{H}_{\Upsilon, \Pi}(\mathcal{F})$ is obtained by stacking a set of $B \times B$ diagonal matrices obtained from

$$\begin{aligned} & H^{\text{diag}}(\mathcal{F}, \nu_n, \mu_m) = \\ & \text{diag}([H(f_1, \nu_n, \mu_m) \ H(f_2, \nu_n, \mu_m) \ \cdots \ H(f_B, \nu_n, \mu_m)]), \\ \mathcal{H}_{\Upsilon, \Pi}(\mathcal{F}) &= \begin{bmatrix} H^{\text{diag}}(\mathcal{F}, \nu_1, \mu_1) & \cdots & H^{\text{diag}}(\mathcal{F}, \nu_N, \mu_1) \\ \vdots & \ddots & \vdots \\ H^{\text{diag}}(\mathcal{F}, \nu_1, \mu_M) & \cdots & H^{\text{diag}}(\mathcal{F}, \nu_N, \mu_M) \end{bmatrix}. \end{aligned}$$

We define

$$\begin{aligned} X_m(\mathcal{F}) &= [X_m(f_1) \ X_m(f_2) \ \cdots \ X_m(f_B)]^\top, \\ \mathcal{X}(\mathcal{F}) &= [X_1^\top(\mathcal{F}) \ \cdots \ X_m^\top(\mathcal{F}) \ \cdots \ X_M^\top(\mathcal{F})]^\top, \\ S_n(\mathcal{F}) &= [S_n(f_1) \ S_n(f_2) \ \cdots \ S_n(f_B)]^\top, \\ \mathcal{S}(\mathcal{F}) &= [S_1^\top(\mathcal{F}) \ \cdots \ S_n^\top(\mathcal{F}) \ \cdots \ S_N^\top(\mathcal{F})]^\top. \end{aligned}$$

The broadband extension equivalent of (3) is given by

$$\mathcal{X}(\mathcal{F}) = \mathcal{H}_{\Upsilon, \Pi}(\mathcal{F}) \mathcal{S}(\mathcal{F}). \quad (6)$$

The source-sensor locations and the signal can be estimated

as

$$\begin{aligned} \Upsilon, \Pi &= \arg \min_{\nu_1, \dots, \nu_N, \mu_1, \dots, \mu_M} \|\mathcal{X}(\mathcal{F}) - \hat{\mathcal{H}}_{\Upsilon, \Pi}(\mathcal{F}) \hat{\mathcal{H}}_{\Upsilon, \Pi}^\dagger(\mathcal{F}) \mathcal{X}(\mathcal{F})\|_2^2 \\ \mathcal{S}(\mathcal{F}) &= \hat{\mathcal{H}}_{\Upsilon, \Pi}^\dagger(\mathcal{F}) \mathcal{X}(\mathcal{F}). \end{aligned} \quad (7)$$

This formulation indicates a parametric approach to source-sensor localization and signal reconstruction by minimizing the objective function stated in (7). It defines the locations as continuous random vectors in space and results in a non-linear objective which is difficult to optimize. In this paper, we resort to a non-parametric method and cast the joint localization and signal reconstruction problem as sparse coding over a dictionary of spatialized signals.

III. SPATIALIZED SPARSE CODING

The enclosure area of sensor deployment and source positions is discretized into a grid of G cells such that $M, N \ll G$. The dictionary, $\mathcal{C}(\mathcal{F})$ is constructed of the spatialized signals for all possible source-sensor locations (manifold matrices) as expressed in

$$\mathcal{C}(\mathcal{F}) = \left[\mathcal{H}_{\Upsilon, \Pi}(\mathcal{F}) \hat{\mathcal{H}}_{\Upsilon, \Pi}^\dagger(\mathcal{F}) \mathcal{X}(\mathcal{F}) \right]_{(\nu_n, \mu_m)}, \quad \mathcal{C}(\mathcal{F}) \in \mathbb{C}^{M \times B \times \mathcal{D}} \quad (8)$$

where \mathcal{D} denotes the number of atoms and it is equal to the combinatorial selection of M microphones and N sources out of possible G grid cells. A binary (selector) vector is defined as

$$P = [p_g], \quad p_g \in \{0, 1\}^{\mathcal{D}} \quad (9)$$

The source-sensor localization problem stated in (7) is cast as the following sparse coding problem:

$$\hat{P} = \arg \min_P \|\mathcal{X}(\mathcal{F}) - \mathcal{C}(\mathcal{F})P\|_2^2 \quad \text{s.t.} \quad \|P\|_0 = \mathcal{K}, \quad (10)$$

where the counting function $\|\cdot\|_0$ returns the number of non-zero components in its argument and $\mathcal{K} = B$ is the number of active frequency components. If a component of $p_g = 1$, it indicates that a sensor/source exists at the particular location associated to that atom of the dictionary. Considering the broad range of B frequencies, the components of P has a block structure underlying its support. In the following Section III-B an efficient sparse recovery algorithm to find the solution of (10) is proposed.

A. Greedy Pursuit

The dictionary $\mathcal{C}(\mathcal{F})$ consists of the projections of the microphone signals into the manifold matrices corresponding to the source-sensor locations on the grid of G cells. Without loss of generality and for the clarity of presentation, we assume that $N = 1$ ¹. The possible number of sensor positions is $\binom{G}{M}$. Considering that a source could be co-located with the sensor, the total combinations of source-sensor positions adds up to $G \binom{G}{M}^2$. Corresponding to each of these $G \binom{G}{M}$ arrangements,

¹By the end of this section, we explain why this is actually a practical assumption for processing the recordings of simultaneous speech sources; larger values are evaluated for the experimental results presented in Section V.

²In case of having the source being co-located with the microphone, a closed form exact solution to microphone localization problem exists which is explained in Appendix I.

there is an array manifold matrix which characterizes the subspace of the microphone array recordings. The complexity of the combinatorial search among all the possible locations is very high so we take an iterative greedy sparse recovery approach.

The basic idea of the greedy pursuit algorithm to solve the problem stated in (10) is to find the location of a small subset of K -microphones at each iteration through a combinatorial search; In the first iteration, all possible source-sensor locations are considered for spatialization and the one selected by sparse coding is indicated as the source location. In the rest of the iterations, the source position is fixed and only the sensor positions are altered. If the location of $M - K$ of the sensors is known *a priori* and only K sensor locations are unknown, then the choice of possible manifold matrices reduces to $\binom{G}{K}$. In the discussion below, we assume $K = 1$ for simplicity, but the extension is straightforward for higher values of K .

Given the multi-channel signal recording $\mathcal{X}(\mathcal{F}) \in \mathbb{C}^{MB}$ and assuming that the position of source and $M - 1$ microphones are known, we construct a dictionary denoted by $\mathcal{C}(\mathcal{F}) \in \mathbb{C}^{MB \times GB}$, composed of projections of $\mathcal{X}(\mathcal{F})$ into G array manifold matrices as given in (4). The g^{th} manifold matrix corresponds to a microphone array with $M - 1$ microphones at known positions and the M^{th} microphone at cell g . More specifically, the following set of array manifold matrices is composed:

$$\{\mathcal{H}_{\gamma, \Pi_{M-1, \mu_1}}(\mathcal{F}), \mathcal{H}_{\gamma, \Pi_{M-1, \mu_2}}(\mathcal{F}), \dots, \mathcal{H}_{\gamma, \Pi_{M-1, \mu_G}}(\mathcal{F})\}$$

where $\mathcal{H}_{\gamma, \Pi_{M-1, \mu_g}}(\mathcal{F})$ represents the manifold matrix for the array configuration where the first $M - 1$ microphones are in their known locations, and the g^{th} microphone is at $\mu_g, \forall g \in \{1, \dots, G\}$. We now write the dictionary as

$$\begin{aligned} \mathcal{C}(\mathcal{F}) = & [\mathcal{H}_{\gamma, \Pi_{M-1, \mu_1}} \mathcal{H}_{\gamma, \Pi_{M-1, \mu_1}}^\dagger(\mathcal{F}) \mathcal{X}(\mathcal{F}), \dots \\ & \dots \mathcal{H}_{\gamma, \Pi_{M-1, \mu_G}} \mathcal{H}_{\gamma, \Pi_{M-1, \mu_G}}^\dagger(\mathcal{F}) \mathcal{X}(\mathcal{F})]; \end{aligned} \quad (11)$$

P is now a $GB \times 1$ vector, with the property that it is *B-sparse* with a block structure: at most B consecutive entries beginning at index B_g can be non-zero. In other words, the support of P is a *1-block-sparse* vector that corresponds to the location of the microphones.

B. Model-based Sparse Recovery

The sparse coding approach expressed in (10) indicates that once the dictionary is constructed of all the spatial projections of the multi-channel signals, finding the unknown locations amounts to sparse approximation of the encoding vector P which selects the projections corresponding to the right array manifold matrix. Since the dictionary is constructed of B adjacent frequencies, the non-zero components of P has a block structure corresponding to the common support/cell where the unknown microphone is located. To incorporate the underlying structure of the sparse coefficients, we use the block sparse recovery algorithm proposed in [27] which is an accelerated scheme for hard thresholding methods with the following recursion

$$P_{i+1} = \mathcal{M}_\Theta (P_i + \kappa \mathcal{C}^\top (\mathcal{X} - \mathcal{C}P_i)), \quad (12)$$

where the step-size κ is the Lipschitz gradient constant to guarantee the fastest convergence speed. To incorporate for the underlying block structure, the model projection operator \mathcal{M}_Θ thresholds and retains only the one (or more generally K) B -block with the highest energy, with subsequent renormalization [27]. The support of estimated \hat{P} determines the microphone location. To estimate the *1-block-sparse* solution, it is also possible to find the combinatorial solution of (10) through a linear search. We performed the combinatorial optimization during the experimental analysis and the results were similar to what we obtained by hard thresholding expressed in (12). The summary of the spatialized sparse coding procedure is stated in Algorithm 1.

Algorithm 1 The greedy sparse coding algorithm for source-sensor localization

1. Initialize K as the number of microphones used for dictionary construction.
 2. Construct the dictionary of spatial signal projections for K microphones and one source.
 3. Find the *1-block-sparse* solution to (10) for $M = K$ along with the source location.
 4. Choose another microphone at an unknown position and construct the dictionary and set $K = K + 1$.
 5. Find the *1-block-sparse* solution to (10) for $M = K$.
 6. Repeat 3 - 5 (multi-source) or 4 - 5 (single-source) until all microphones/sources are localized.
-

This algorithm finds the location of the microphones one by one by taking into account the already localized microphones; however, it generalizes trivially to the case of K unknown microphone locations. Assuming that the sources are stationary, the source locations are fixed at the initial iterations leading to a substantial decrease in computational cost.

The estimation of distant signal based on single source projection at each frequency is accurate if the sources have fast decaying auto-correlation and are sufficiently separated. Due to the sparsity of speech signal in Fourier domain, this condition holds for processing segments of overlapping speech recordings and it enables accurate multi-source localization [23, 28].

Given the location of the sources and microphones, the signals are reconstructed by inverse filtering the acoustic channel. In Section IV, we provide rigorous analysis of the optimality of speech reconstruction in multi-source scenario by showing the equivalence of inverse filtering to speech separation followed by channel deconvolution.

IV. EQUIVALENCE TO SOURCE SEPARATION AND DEREVERBERATION

In this section, we show the equivalence of inverse filtering to a two stage procedure: *speech separation* followed by *dereverberation*. For the sake of brevity, the proof is explained for two sources recorded by three sensors in Section IV-A. The generalized theory for larger values of M and N is elaborated in Section IV-B.

$$\mathcal{J}\mathcal{C}^\dagger = \frac{1}{|\mathcal{J}\mathcal{C}^\top\mathcal{J}\mathcal{C}|} \overbrace{\begin{bmatrix} \sum_{i=1}^3 H_{i2}^2 & \sum_{i=1}^3 H_{i1}H_{i2} \\ \sum_{i=1}^3 H_{i1}H_{i2} & \sum_{i=1}^3 H_{i1}^2 \end{bmatrix}}^{\mathcal{A}} \begin{bmatrix} H_{11} & H_{21} & H_{31} \\ H_{12} & H_{22} & H_{32} \end{bmatrix} \quad (14)$$

$$\mathcal{A} = \begin{bmatrix} H_{11} \sum_{i=1}^3 H_{i2}^2 - H_{12} \sum_{i=1}^3 H_{i1}H_{i2} & H_{21} \sum_{i=1}^3 H_{i2}^2 - H_{22} \sum_{i=1}^3 H_{i1}H_{i2} & H_{31} \sum_{i=1}^3 H_{i2}^2 - H_{32} \sum_{i=1}^3 H_{i1}H_{i2} \\ -H_{11} \sum_{i=1}^3 H_{i1}H_{i2} + H_{12} \sum_{i=1}^3 H_{i1}^2 & -H_{21} \sum_{i=1}^3 H_{i1}H_{i2} + H_{22} \sum_{i=1}^3 H_{i1}^2 & -H_{31} \sum_{i=1}^3 H_{i1}H_{i2} + H_{32} \sum_{i=1}^3 H_{i1}^2 \end{bmatrix}$$

A. Simple Example

Consider the multiple input multiple output (MIMO) finite impulse response acoustic system which is represented in matrix notation as

$$\underbrace{\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}}_{\mathcal{X}} = \underbrace{\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \\ H_{31} & H_{32} \end{bmatrix}}_{\mathcal{J}\mathcal{C}} \underbrace{\begin{bmatrix} S_1 \\ S_2 \end{bmatrix}}_{\mathcal{S}} \quad (13)$$

Hence, $\mathcal{S} = \mathcal{J}\mathcal{C}^\dagger\mathcal{X}$ where $\mathcal{J}\mathcal{C}^\dagger$ is defined in (14).

Thereby, S_1 is obtained by

$$S_1 = \frac{1}{|\mathcal{J}\mathcal{C}^\top\mathcal{J}\mathcal{C}|} \sum_{j=1}^3 \left(H_{j1} \sum_{i=1}^3 H_{i2}^2 - H_{j2} \sum_{i=1}^3 H_{i1}H_{i2} \right) X_j$$

$$= \frac{1}{|\mathcal{J}\mathcal{C}^\top\mathcal{J}\mathcal{C}|} [H_{22}(H_{11}H_{22} - H_{12}H_{21}) + H_{32}(H_{11}H_{32} - H_{12}H_{31})]X_1 +$$

$$[H_{12}(H_{21}H_{12} - H_{22}H_{11}) + H_{32}(H_{21}H_{32} - H_{22}H_{31})]X_2 +$$

$$[H_{12}(H_{31}H_{12} - H_{32}H_{11}) + H_{22}(H_{31}H_{22} - H_{32}H_{21})]X_3 \quad (15)$$

We show that the equation on the right-hand side can be factorized such that

$$S_1 = F^{-1}G\mathcal{X}$$

$$F = \begin{bmatrix} H_{22}H_{11} - H_{12}H_{21} \\ H_{32}H_{11} - H_{12}H_{31} \\ H_{32}H_{21} - H_{22}H_{31} \end{bmatrix}, \quad G = \begin{bmatrix} H_{22} & -H_{12} & 0 \\ H_{32} & 0 & -H_{12} \\ 0 & H_{32} & -H_{22} \end{bmatrix} \quad (16)$$

The factorized system is shown in Figure 1. Defining $\mathcal{Y} = G\mathcal{X}$, we obtain

$$Y_{S_1,1} = H_{22}X_1 - H_{12}X_2,$$

$$Y_{S_1,2} = H_{32}X_1 - H_{12}X_3,$$

$$Y_{S_1,3} = H_{32}X_2 - H_{22}X_3$$

Thereby, S_2 is separated from the three channel outputs Y_1, Y_2, Y_3 by the first linear operation $G\mathcal{X}$. The second operation performs channel deconvolution hence,

$$S_1 = \frac{1}{|F|} [(H_{22}H_{11} - H_{12}H_{21})Y_{S_1,1} +$$

$$(H_{32}H_{11} - H_{12}H_{31})Y_{S_1,2} +$$

$$(H_{32}H_{21} - H_{22}H_{31})Y_{S_1,3}]$$

Substituting Y_1, Y_2, Y_3 with their values, we obtain

$$S_1 = \frac{1}{|F|} [H_{22}(H_{11}H_{22} - H_{12}H_{21}) + H_{32}(H_{11}H_{32} - H_{12}H_{31})]X_1 +$$

$$[H_{12}(H_{21}H_{12} - H_{22}H_{11}) + H_{32}(H_{32}H_{21} - H_{22}H_{31})]X_2 +$$

$$[H_{12}(H_{31}H_{12} - H_{32}H_{11}) + H_{22}(H_{31}H_{22} - H_{32}H_{21})]X_3 \quad (17)$$

We see that the nominator is equivalent to the nominator in

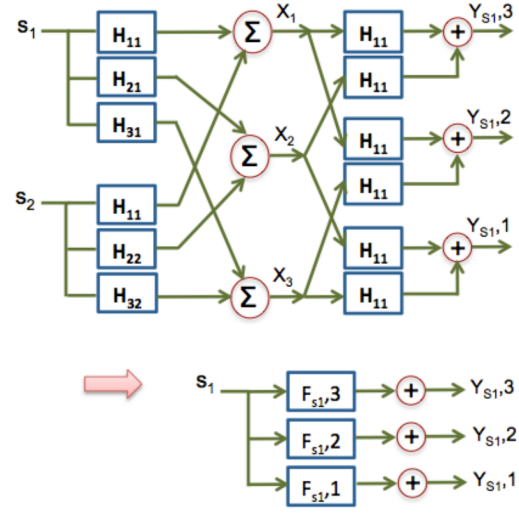


Fig. 1: Source separation filters for a MIMO system: $F_{S_1,1}$, $F_{S_1,2}$ and $F_{S_1,3}$ are calculated from the acoustic channel corresponding to multiple sources as stated in (16).

(15). We can further see that the determinant of F is

$$|F| = (H_{22}H_{11} - H_{12}H_{21})^2 + (H_{32}H_{11} - H_{12}H_{31})^2 + (H_{32}H_{21} - H_{22}H_{31})^2$$

which is equivalent to determinant of $H^\top H$ as

$$|H^\top H| = (H_{11}^2 + H_{21}^2 + H_{31}^2)(H_{12}^2 + H_{22}^2 + H_{32}^2)$$

$$- (H_{11}H_{12} + H_{21}H_{22} + H_{31}H_{32})^2$$

$$= \underbrace{H_{11}^2 H_{22}^2 + H_{21}^2 H_{12}^2 - 2H_{11}H_{12}H_{21}H_{22}}_{(H_{22}H_{11} - H_{12}H_{21})^2}$$

$$+ \underbrace{H_{11}^2 H_{32}^2 + H_{31}^2 H_{12}^2 - 2H_{11}H_{12}H_{31}H_{32}}_{(H_{32}H_{11} - H_{12}H_{31})^2}$$

$$+ \underbrace{H_{21}^2 H_{32}^2 + H_{31}^2 H_{22}^2 - 2H_{11}H_{22}H_{31}H_{32}}_{(H_{32}H_{21} - H_{22}H_{31})^2} = |F|^2$$

Similar to what was shown for S_1 , the same equality holds for S_2 . The equivalence of inverse filtering to the two stage procedure (1) separation of interference signals and (2) deconvolution of the signal from the early room impulse response function enables analysis of the stability and performance of the system [14]. Having the interferences separated as $\mathcal{Y} = G\mathcal{X}$, multiple interference-free speech signals are obtained $Y_{S_1,1}, Y_{S_1,2}, Y_{S_1,3}$ which are corresponding to the following acoustic paths:

$$F_{S_1,1} = H_{22}H_{11} - H_{12}H_{21}$$

$$F_{S_1,2} = H_{32}H_{11} - H_{12}H_{31} \quad (18)$$

$$F_{S_1,3} = H_{32}H_{21} - H_{22}H_{31}$$

Hence, the separated signals can sound more reverberant due to the prolonged impulse response of the equivalent channels. To perform exact deconvolution of the signals, Bezout theorem for dereverberation of an acoustic system is exploited in [26] referred to as multiple-input/output inverse-filtering (MINT) theory of acoustic channel deconvolution. The Bezout theorem is mathematically expressed as follows:

$$\begin{aligned} \gcd [F_{S_{m,1}}, F_{S_{m,2}}, \dots, F_{S_{m,P}}] &= 1 \\ \Leftrightarrow \exists U_{S_{m,1}} U_{S_{m,2}} \dots, U_{S_{m,P}} : \sum_{p=1}^P F_{S_{m,p}} U_{S_{m,p}} &= 1. \end{aligned} \quad (19)$$

where $\gcd[\cdot]$ denotes the greatest common divisor of the involved polynomials. According to the Bezout theorem, if the polynomials $F_{S_{m,p}}$ ($p = 1, 2, \dots, P$) have no common divisor or equivalently acoustic channel responses $H_{n,m}$, $n = 1, 2, \dots, N$ do not share any common zeros, it is possible to equalize each of the M channels. Therefore, having multiple channel responses eliminates the requirement on minimum-phase impulse response for perfect dereverberation. The considerations raise when channel responses have common zeros. This can happen if $C_{S_m} = \gcd [H_{12}, H_{22}, H_{32}] = \gcd [F_1, F_2, F_3] \neq 1$; having common zeros, the theorem can only partially dereverberate the speech signal up to the polynomial C_{S_m} [14].

B. Generalized Theory

In Section IV-A, we presented the proof of equivalence of inverse filtering the room acoustic channel with speech separation and dereverberation considering 3 microphones and 2 sources. In this section, we show that inverse filtering of any $M \times N, M > N$ MIMO system is equivalent to a two step procedure: *source separation* followed by *channel deconvolution*.

Denoting the transfer function between microphone m and source n as $H_{m,n}$, a MIMO finite impulse response (FIR) acoustic system can be represented in matrix notation as

$$\underbrace{\begin{bmatrix} X_1 \\ \vdots \\ X_M \end{bmatrix}}_x = \underbrace{\begin{bmatrix} H_{11} & \dots & H_{1N} \\ \vdots & & \vdots \\ H_{M1} & \dots & H_{MN} \end{bmatrix}}_{\mathcal{H}} \underbrace{\begin{bmatrix} S_1 \\ \vdots \\ S_N \end{bmatrix}}_s \quad (20)$$

Hence, the minimum mean square estimation (MMSE) of the source is given by

$$\hat{s} = \arg \min_s (\mathcal{X} - \mathcal{H}s)^2 = \arg \min_s \frac{\sum_{i=1}^{(M)} (\mathcal{X}_i - \mathcal{H}_i s)^2}{\binom{M-1}{N-1}} \quad (21)$$

where submatrix \mathcal{H}_i is an $N \times N$ matrix consisting of N selected rows from \mathcal{H} so we have $\binom{M}{N}$ possible construction for \mathcal{H}_i . Corresponding to each \mathcal{H}_i , we construct \mathcal{X}_i hence, we obtain

$$\mathcal{X}_i = \mathcal{H}_i \hat{s}_i \quad \forall i = 1, \dots, \binom{M}{N} \quad (22)$$

where \hat{s}_i is an estimation of S using \mathcal{X}_i and \mathcal{H}_i . To estimate

S , we compute the derivative of (21) with respect to S

$$\begin{aligned} \frac{1}{\binom{M-1}{N-1}} \frac{\partial}{\partial S} \sum_{i=1}^{(M)} (\mathcal{X}_i - \mathcal{H}_i S)^2 &= 0 \\ \Rightarrow \sum_{i=1}^{(M)} \mathcal{H}_i^T \mathcal{X}_i &= \sum_{i=1}^{(M)} \mathcal{H}_i^T \mathcal{H}_i \hat{s} \\ \Rightarrow \hat{s} &= \frac{\sum_{i=1}^{(M)} \mathcal{H}_i^T \mathcal{H}_i \hat{s}_i}{\sum_{i=1}^{(M)} \mathcal{H}_i^T \mathcal{H}_i} \end{aligned} \quad (23)$$

We now move onto show the equivalence to a two step procedure of *source separation* and *dereverberation*. Without the loss of generality, we first separate S_1 from $S_j, j = 2, \dots, N$. From *Cramer's rule* in linear algebra we have

$$\left[\mathcal{X}_i \mathcal{H}_i^{/1} \right] = |\mathcal{H}_i| \tilde{S}_1^i = y_{S_1, i} \quad (24)$$

where $\mathcal{H}_i^{/1}$ is \mathcal{H}_i matrix with removing the first column of it. Hence $y_{S_1, i}$ is only dependent on S_1 while the components of S_2, \dots, S_N are separated. By repeating (24) for all possible \mathcal{H}_i , we get the set of all estimates of S_1 as $y_{S_1, 1}, \dots, y_{S_1, \binom{M}{N}}$.

In the second step, we perform deconvolution to extract \hat{S}_1 . We define the matrices

$$\begin{aligned} Y &\triangleq [y_{S_1, 1}, \dots, y_{S_1, \binom{M}{N}}] \\ F &\triangleq [|\mathcal{H}_{(1)}| \dots |\mathcal{H}_{\binom{M}{N}}|] \end{aligned} \quad (25)$$

Hence, deconvolution of the channel from the separated source is achieved via

$$\begin{aligned} \hat{S}_1 &= \frac{Y F^T}{F^T F} = \frac{Y [|\mathcal{H}_{(1)}| \dots |\mathcal{H}_{\binom{M}{N}}|]^T}{\sum_{i=1}^{(M)} |\mathcal{H}_i|^2} \\ &= \frac{[|\mathcal{H}_{(1)}| \tilde{S}_1^1 \dots |\mathcal{H}_{\binom{M}{N}}| \tilde{S}_1^{\binom{M}{N}}] [|\mathcal{H}_{(1)}| \dots |\mathcal{H}_{\binom{M}{N}}|]^T}{\sum_{i=1}^{(M)} |\mathcal{H}_i|^2} \\ &\Rightarrow \hat{S}_1 = \frac{|\mathcal{H}_{(1)}|^2 \tilde{S}_1^1 + \dots + |\mathcal{H}_{\binom{M}{N}}|^2 \tilde{S}_1^{\binom{M}{N}}}{\sum_{i=1}^{(M)} |\mathcal{H}_i|^2} \end{aligned} \quad (26)$$

If we calculate $\hat{S}_2, \dots, \hat{S}_N$ in the same manner, we have

$$\begin{aligned} \hat{S}_2 &= \frac{|\mathcal{H}_{(1)}|^2 \tilde{S}_2^1 + \dots + |\mathcal{H}_{\binom{M}{N}}|^2 \tilde{S}_2^{\binom{M}{N}}}{\sum_{i=1}^{(M)} |\mathcal{H}_i|^2} \\ &\vdots \\ \hat{S}_N &= \frac{|\mathcal{H}_{(1)}|^2 \tilde{S}_N^1 + \dots + |\mathcal{H}_{\binom{M}{N}}|^2 \tilde{S}_N^{\binom{M}{N}}}{\sum_{i=1}^{(M)} |\mathcal{H}_i|^2} \end{aligned} \quad (27)$$

By concatenation we stated in (20), we obtain (28); the equality to (23) is proved in Appendix VI.

$$\hat{s} = \frac{|\mathcal{H}_{(1)}|^2 \hat{S}_1 + \dots + |\mathcal{H}_{\binom{M}{N}}|^2 \hat{S}_{\binom{M}{N}}}{\sum_{i=1}^{(M)} |\mathcal{H}_i|^2} = \frac{\sum_{i=1}^{(M)} |\mathcal{H}_i|^2 \hat{s}_i}{\sum_{i=1}^{(M)} |\mathcal{H}_i|^2} \quad (28)$$

It may be noted that decomposition of inverse filtering in two steps leading to separation and dereverberation is particularly important in the context of source separation since implementation of the filters requires different practical considerations [14]. Nevertheless, the theoretical results established in this section are general results applicable for all methods based on inverse filtering.

V. EXPERIMENTS

The experiments are conducted to evaluate the performance of the proposed approach for multi-source localization and signal reconstruction. The location of the microphones is unknown and estimated through the proposed procedure. The performance is evaluated in terms of source localization and separation and speech recognition.

A. Microphone Array Corpus

The evaluation is carried out on two databases collected on multi-channel overlapping numbers corpus (MONC) [29] as well as multi-channel wall street journal (MC-WSJ) corpus [30]. A brief overview of both databases is stated in the following subsections. The readers are referred to the specific documentations published in [29] and [30] for more details.

1) *MONC*: This database was collected by outputting utterances from the Numbers corpus (telephone quality speech, 30-word vocabulary) on one or more loudspeakers, and recording the resulting sound field using a microphone array and various lapel and table-top microphones. The energy levels of all utterances in the Numbers corpus were normalized to ensure a relatively constant desired speech level across all recordings. The corpus was then divided into 6049-utterances as the training set, 2026-utterances as the cross validation set, and 2061-utterances as the test set. Competing-sources of the cross-validation and test sets were also produced by rearranging the order of their respective utterances. The word loop grammar is used and the task is speaker independent. The acoustic of the enclosure is mildly reverberant and the average SNR is 10 dB for single source recordings. The speech signals are recorded at 8 kHz sampling frequency. The geometrical set-up of the recordings are exactly measured and provided in [29].

2) *MC-WSJ*: This database was used for PASCAL Speech Separation Challenge II [31, 32]. The subset consists of two concurrent sources who are simultaneously reading sentences from the Wall Street Journal and being recorded with 8-channel circular microphone array. The sources are either seating or standing at 1.5×1.5 table and a circular microphone array with diameter 20 cm is located at the center of the table. The total number of utterances is 356 (or 178, respectively given that two sentences are read at a time). The ground-truth of the geometrical set-up is not provided hence, source-sensor localization results can not be evaluated in the framework of MC-WSJ. We use this corpus for overlapping speech recognition due to its phonetically rich data.

B. Source-Sensor Localization Performance

The single source utterances are used for sensor localization. The spectro-temporal representation is obtained by windowing the microphone signals in 256 ms frames using a Hann function with 25% overlap. To perform localization of 8-channel microphones, the greedy pursuit algorithm summarized in Table 1 is used. First, the position of the source along with the two broadside microphones are estimated. Thereafter, the location of the other microphones are estimated one per iteration. To increase the resolution of the estimates while keeping the dimensionality of the sparse vector bounded, a coarse-to-fine strategy is taken [33]. More specifically, the area is discretized into 5 cm cells. The localized microphones are then re-located in 1 cm accuracy using a finer discretization. The average norm of source-sensor localization error is 8.9 mm. Given that the complexity of the combinatorial optimization increases as $\mathcal{O}(G^M)$ whereas the greedy sparse recovery has a complexity of $\mathcal{O}(GM)$, it is crucial to employ the sparse recovery algorithms to enable accurate localization in our set-up.

In addition, the method proposed in [34, 35] is implemented and used to find the location of the circular array. This method relies on diffuse noise model to find the topology of the array. The results obtained for localization of circular microphones is about 1.2 cm using about 10 s recording of diffuse noise field. In practice however, the level and length of the available diffuse noise might be challenging to employ the technique proposed in [34]. Hence, our proposed approach which requires only a few speech frames (less than 1 s) provides a higher applicability and accuracy.

C. Speech Reconstruction and Recognition Performance

Given the location of the sources and the characterized room acoustic channel, the desired signal is recovered by inverse filtering to perform speech recognition. Assuming that the channel responses do not have common zeros, we can deconvolve the signal from the early room impulse response function through least-squared (psuedo) inverse filtering. The optimality of this approach by showing its relation to weighted inverse filtering and equivalence to the two-step operation of source separation and channel deconvolution is established in Section IV.

Based on the proposed theories, the effect of early reverberation is mitigated. The late reverberation can be statistically modeled as an exponentially decaying white Gaussian noise [36] which also possess the diffuse characteristics [37, 38]. To reduce the effect of late reverberation and enhance the signal, we can apply the post-processing techniques. Among several post-filtering methods proposed in the literature [39, 38], the Zelinski post-filtering is a practical implementation of the optimal Wiener filter; while a precise realization of the later requires knowledge about the spectrum of the desired signal, the Zelinski post-filtering method uses the auto- and cross-power spectra of the multi-channel input signals to estimate the target signal and noise power spectra under the assumption of *zero cross-correlation* between noise on different sensors [40]. We implemented the

TABLE I: Quality evaluation of the recovered speech in terms of signal to interference ratio (SIR), perceptual evaluation of speech quality (PESQ) and word recognition rate (WRR) using beamforming, vs. inverse filtering of the room acoustic channel. N denotes the number of concurrent sources. The numbers in parentheses show the amount of improvement due to denoising achieved by post-filtering.

N	Meas.	Directpath filtering	Multipath deconvolution
1	SIR	18.52 (0.02)	16.5 (-0.4)
	PESQ	3.3 (0)	2.97 (0.06)
	WRR%	95 (0)	93.3 (-0.37)
2	SIR	11.33 (-0.47)	17.5 (5)
	PESQ	2.69 (0.09)	2.8 (0.2)
	WRR%	68.16 (-2.03)	87.93 (4.56)
3	SIR	10 (-0.2)	14.2 (4.1)
	PESQ	2.48 (0.08)	2.62 (0.22)
	WRR%	61.45 (-1.55)	79.21 (8.33)

Zelinski post-filter for the experiments conducted in Section V. The dereverberation of the early impulse response achieved by inverse filtering the acoustic channels enables a more efficient post-filtering as we will see shortly in this section.

The automatic speech recognition scenario was designed to broadly mirror that of Moore and McCowan [41]. A typical front-end was constructed using the HTK toolkit [42] with 25 ms frames at a rate of 10 ms. This produced 12 mel-cepstra plus the zeroth coefficient and the first and second time derivatives; 39 features in total. Cepstral Mean Normalization is applied to the feature vectors, resulting in speech recognition performance improvement of about 15% relative. The back-end consists of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. The ASR accuracy on the clean speech data is about 95%. We perform MAP adaptation by training directly on recovered data. The Zelinski post-filtering is applied on the recovered speech prior to recognition.

In addition to the speech recognition, we evaluate the quality of the recovered speech using signal to interference ratio (SIR) [43] as well as perceptual evaluation of speech quality (PESQ) [44]. As the speech recovery approach relies on the principles of spatial diversity, we compare it with beamforming which possesses similar trait. We use the super-resolution source localization based on sparse recovery to perform near-field beamforming. Room acoustic modeling is achieved through our method explained in [24]. As the results indicate, inverse filtering the acoustic multipath projections followed by post-filtering yields the maximum interference suppression and highest perceptual quality of the recovered speech in multi-party scenarios as quantified in terms of SIR, PESQ and word recognition rate (WRR). The Zelinski post-processing is derived to reduce the effect of *uncorrelated noise*. The change of the objective evaluation metric attributed to post-filtering is indicated in parentheses in Table I. We observe that the improvement in performance obtained after deconvolution of the early part of the room acoustic channel is higher than what we can achieve after standard beamforming based on direct-path filtering.

Furthermore, we conduct some experiments based on a subset of MC-WSJ corpus [30] used for PASCAL Speech Separation Challenge II [31, 32]. The speech recognition system

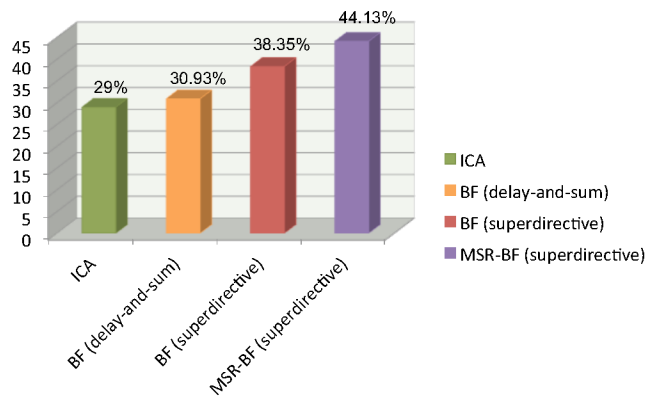


Fig. 2: Word recognition rate (WRR) using independent component analysis (ICA), delay-and-sum and superdirective beamforming and model-based sparse recovery incorporated for superdirective beamforming (MSR-BF). The benchmark WRR for the headset microphone is 76.55% and for the distant microphone is 0%.

used in the experiments is identical to the one in [45, 32], except that three passes are applied (instead of four). The first pass is unadapted speech recognition. In the second pass, unsupervised maximum likelihood linear regression (MLLR) feature space adaptation is applied and the third pass employs full MLLR adaptation. The estimated source positions are the same ones used in [45, 32] for superdirective and delay-and-sum beamforming. The room reverberation time is 0.7 s and the corpus is highly noisy.

In addition to the conventional beamforming, we employed independent component analysis. The model-based sparse recovery (MSR) provides the exact location of the sources on a discretized grid [28]. The speech recognition results are presented in Fig 2. The results show that providing the accurate positioning achieved by sparse coding improves the performance of the conventional beamforming methods. Exact deconvolution via inverse filtering was not useful in this setup due to missing information on geometrical setup; in particular, the height information is missing which incurs inaccuracies in calculation of the room acoustic channel to perform inverse-filtering. This information is hard to infer using the 2D microphone array setup used for recording this database.

VI. CONCLUSIONS

This paper presented a sparse coding approach to source-sensor localization via spatialization and sound reconstruction. An adaptive strategy for designing a dictionary of distant signals is devised based on projection of the array recordings into the array manifolds characterized for the hypothesized locations on a grid. The projections amount to reconstruction of the distant signals by inverse filtering the acoustic channel. Source and sensor localization is achieved in a greedy fashion through sparse coding over this dictionary of spatialized signals. Furthermore, a closed-form solution is derived for the case of known source signal such as co-location of the source with one of the sensors. The equivalence of inverse filtering to speech separation and deconvolution was shown rigorously leading to an optimal solution if the room impulse responses

do not have common zeros. The experimental analysis verified the applicability of the proposed framework using real data recordings of multiparty speech reconstruction and recognition.

APPENDIX I. CLOSED-FORM SOLUTION TO SENSOR LOCALIZATION

Consider a scenario where a single source signal and its position ν_1 is known; for instance the source is co-located with one of the microphones. We can formulate an alternative solution to the sensor localization problem which can be admits a closed-form solution. Defining

$$\begin{aligned} \mathbf{S} &= \text{diag}[S(f_1) \cdots S(f_B)], \mathbf{S} \in \mathbb{C}^{B \times B} \\ \mathbf{X} &= [X_1(\mathcal{F}) \cdots X_M(\mathcal{F})]^T, \mathbf{X} \in \mathbb{C}^{M \times B} \\ \mathbf{H} &= \begin{bmatrix} H(f_1, \nu_1, \mu_1) & \cdots & H(f_B, \nu_1, \mu_1) \\ \vdots & \ddots & \vdots \\ H(f_1, \nu_1, \mu_M) & \cdots & H(f_B, \nu_1, \mu_M) \end{bmatrix}_{G \times B}, \end{aligned}$$

we can write

$$\mathbf{X} = \mathbf{C}\mathbf{H}\mathbf{S} \quad (29)$$

where $\mathbf{C} \in \{0, 1\}^{M \times G}$ is a matrix consisting of binary values with a single 1 at each row and all the other components equal to 0 such that the column indices are exclusive and $\|\mathbf{C}\|_0 = M$; by this definition, $C_{i,j} = 1$ selects one of the G cells where a microphone is located. For example, if we have two microphones ($M = 2$) located at cells 10 and 20, only $C_{1,10}$ and $C_{2,20}$ are 1. Thereby, it reveals the microphone locations on the grid.

The formulation stated in (29) considers a rearrangement of the spectral coefficients in $\mathcal{X}(\mathcal{F})$ to constitute the rows of \mathbf{X} (as opposed to vector concatenation in (6)). \mathbf{S} is the diagonal matrix consisting of B frequency components of the source. Matrix \mathbf{H} consists of the acoustic channel associated to the source located at ν_1 and the microphone located at either of the G cells. Given this formulation of the microphone localization problem, we can find the exact solution for \mathbf{C} through least-square:

$$\mathbf{C}^T = (\mathbf{S}\mathbf{H}^T)^{-1} \mathbf{X}^T \quad (30)$$

The matrix \mathbf{C} encodes the location of the microphones. This solution requires that $B \geq G$ which hold for the broadband signal and a reasonable resolution of the discretization.

APPENDIX II. THEORY OF LEAST SQUARES

A. Notation and auxiliary results

We need to define some notation before proceeding to the main problem. For $n \in \mathbb{N}$, we define $[n] = \{1, 2, \dots, n\}$ to be the set of all integers from 1 to n . Let $p \subset [n]$ be a nonempty subset of $[n]$ of size k and let A be an $n \times m$ matrix. We denote by A_p a $k \times n$ matrix obtained by selecting the rows of matrix A belonging to p , e.g. for $p = \{1, 2\}$ A_p is a $2 \times m$ matrix consisting of the first and the second row of A . Let B be an arbitrary $m \times m$ matrix and let $p \subset [n]$ of size $|p| = m$. We denote by $\text{emb}(B, p, n)$ the embedding of columns of B inside an $m \times n$ matrix. To explain more precisely, let us assume that the components of p are sorted

with $p_1 < p_2 < \dots < p_m$. Then $\text{emb}(B, p, n)$ is an all-zero $m \times n$ matrix except that its p_i -th column, $i = 1, 2, \dots, m$, is equal to the i -th column of B .

Let $r, c \in \mathbb{N}$ be arbitrary numbers. We define the linear space of all $r \times c$ real-valued matrices by $M_R(r, c)$ with the traditional matrix addition and scalar-matrix multiplication. For arbitrary matrices $M, N \in M_R(r, c)$, we define the following bilinear form $\langle M, N \rangle = \text{tr}(MN^T) = \sum_{i,j} M_{ij}N_{ij}$. One can simply check that \langle, \rangle defines an inner product on $M_R(r, c)$. We need the following lemma from linear algebra that we prove for sake of completeness.

Lemma 1. *Let $r, s \in \mathbb{N}$ and let $M \in M_R(r, c)$. If $\langle M, N \rangle = 0$ for every $N \in M_R(r, c)$ then $M = \mathbf{0}$.*

Proof: Let $i \in [r], j \in [c]$ be arbitrary numbers and let N be an all-zero matrix except for its ij -th component which is $N_{ij} = 1$. It results that

$$0 = \langle M, N \rangle = \sum_{k,\ell} M_{k\ell}N_{k\ell} = M_{ij} = 0,$$

which immediately implies that $M = \mathbf{0}$. \blacksquare

Lemma 2. *Let M be an square invertible matrix whose components depend on a parameter t . Then $\frac{\partial}{\partial t}M^{-1} = -M^{-1}(\frac{\partial}{\partial t}M)M^{-1}$.*

Proof: Let I be the identity matrix of the same order as M . Taking derivative from both sides of the identity $I = MM^{-1}$, one obtains $0 = \frac{\partial}{\partial t}MM^{-1} + M\frac{\partial}{\partial t}M^{-1}$ which implies that $\frac{\partial}{\partial t}M^{-1} = -M^{-1}(\frac{\partial}{\partial t}M)M^{-1}$. \blacksquare

Lemma 3. *Let A be an square matrix of order d whose components depend on a parameter t . Then $\frac{\partial}{\partial t}\det(A) = \text{tr}(A^{-1}\frac{\partial}{\partial t}A)$*

Proof: Let A_{ij} the component of A in row i and column j . We first find $\frac{\partial}{\partial A_{ij}}\det(A)$ and use the chain rule to obtain

$$\frac{\partial}{\partial t}\det(A) = \sum_{i,j \in [d]} \frac{\partial}{\partial A_{ij}}\det(A) \frac{\partial}{\partial t}A_{ij}.$$

Notice that in order to compute $\det(A)$, we can expand it with respect to the i -th row where we obtain

$$\det(A) = \sum_{k \in [d]} (-1)^{i+k} A_{ik} \det(A_{\tilde{i}\tilde{k}}),$$

where $A_{\tilde{i}\tilde{k}}$ is a $(d-1) \times (d-1)$ matrix obtained after removing the i -th row and the k -th column of the matrix A . In particular, it can be immediately checked that the only term that depends on A_{ij} in the summation is $(-1)^{i+j} \det(A_{\tilde{i}\tilde{j}})$, thus one obtains

$$\frac{\partial}{\partial A_{ij}}\det(A) = (-1)^{i+j} \det(A_{\tilde{i}\tilde{j}}) = \text{adj}(A)_{ji},$$

where $\text{adj}(A)$ denotes the *adjoint* of matrix A . Moreover, from the formula $A^{-1} = \frac{\text{adj}(A)}{\det(A)}$ for the inverse of matrix A , one immediately obtains that

$$\frac{\partial}{\partial A_{ij}}\det(A) = \det(A)(A^{-1})_{ji}.$$

Applying the chain-rule, it results that

$$\frac{\partial}{\partial t} \det(A) = \det(A) \sum_{i,j \in [d]} (A^{-1})_{ji} \frac{\partial}{\partial t} A_{ij} = \text{tr}(A^{-1} \frac{\partial}{\partial t} A),$$

where tr denotes the trace operator and where $\frac{\partial}{\partial t} A$ denotes the component-wise partial derivative of A with respect to t . ■

Lemma 4. *Let M and S be $n \times n$ matrices where S is symmetric. Then $\text{tr}(AM) = \text{tr}(AM^\top)$.*

Proof: The proof simply follows from the properties of the trace operator:

$$\text{tr}(SM) = \text{tr}((SM)^\top) = \text{tr}(M^\top S^\top) = \text{tr}(M^\top S) = \text{tr}(SM^\top),$$

where we used the symmetry of S and the fact that for arbitrary square matrices K, L of the same dimension, $\text{tr}(KL) = \text{tr}(LK)$. ■

Theorem 1 (Cauchy-Binet). *Let A and B be $n \times m$ matrices with $n \geq m$. Then*

$$\det(A^\top B) = \sum_{p \subset [n], |p|=m} \det(A_p) \det(B_p), \quad (31)$$

where $|p|$ denotes the cardinality (number of elements) of $p \subset [n]$.

B. The main theorem and connections with overdetermined set of linear equations

Let H be an $n \times m$ matrix where $n \geq m$. We assume that H has full column rank which in particular implies that there is a subset of rows of H of size m , namely $p \subset [n], |p| = m$, such that $\det(H_p) \neq 0$. Recall that H_p is a sub-matrix of H consisting of all those rows with index in p . Let $x \in \mathbb{R}^n$ be an arbitrary column vector and consider an overdetermined set of linear equations given by $x = Hs$ to be solved for the unknown variables $s \in \mathbb{R}^m$. If $m = n$ the solution of this equation is simply given by $s = H^{-1}x$. For $n > m$, the equations might be inconsistent, thus the set of linear equations $x = Hs$ might not have a solution. However, a solution exists if and only if $x \in \text{span}\{H_1, H_2, \dots, H_m\}$, where span denotes the subspace spanned by a set of vectors and where H_i , $i = 1, 2, \dots, m$, denote the i -th column of H .

In general, even if the equations are inconsistent, one can build a candidate solution by some kind of averaging all the possible sub-solutions. More precisely, let \mathcal{P} be the set of all subsets of $[n]$ of size m , i.e. $\mathcal{P} = \{p \subset [n] : |p| = m\}$, and consider the set of all sub-matrices H_p with a nonzero determinant $\det(H_p) \neq 0$. For every H_p in this set, one can consider a sub-solution $s_p = H_p^{-1}x_p$, where x_p is the sub-vector of x consisting of the components with index in p . Taking the weighted average of all possible sub-solutions with a weighting $\omega_p \geq 0$, $p \in \mathcal{P}$, one obtains a candidate solution

$$s^\omega = \frac{\sum_{p \in \mathcal{P}} \omega_p s_p}{\sum_{p \in \mathcal{P}} \omega_p}. \quad (32)$$

Notice that by changing the associated weighting ω_p , one can obtain different solutions s^ω . We will consider the weighting

function $\omega_p = \det(H_p)^2$, which is equal to the squared determinant of the sub-matrix H_p . We define the resulting solution by

$$\hat{s}_1 = \frac{\sum_{p \in \mathcal{P}} \det(H_p)^2 H_p^{-1} x_p}{\sum_{p \in \mathcal{P}} \det(H_p)^2}. \quad (33)$$

Notice that if for a specific $p \in \mathcal{P}$, $\det(H_p) = 0$ then H_p^{-1} does not exist but, with some abuse of notation, this term does not play a role because its corresponding weight $\det(H_p)^2$ is also equal to 0.

Another approach to find a candidate solution is the least square approach where the resulting solution \hat{s}_2 is given by $\hat{s}_2 = \arg \min_{s \in \mathbb{R}^m} \|Hs - x\|_2$. Since H has full column rank, one can explicitly obtain $\hat{s}_2 = (H^\top H)^{-1} H^\top x$. In this section, we prove that exploiting the weighting $\omega_p = \det(H_p)^2$, both the estimates give exactly the same result, namely $\hat{s}_1 = \hat{s}_2$.

Theorem 2. *Suppose H is a given $n \times m$ matrix ($n \geq m$) with full column rank and assume that $x \in \mathbb{R}^n$ is an arbitrary vector. Let \hat{s}_1 be the weighted average solution given by (33) and let $\hat{s}_2 = (H^\top H)^{-1} H^\top x$ be the least square solution. Then $\hat{s}_1 = \hat{s}_2$.*

Proof: Let us label n rows of matrix H with $[n] = \{1, 2, \dots, n\}$ and let $\mathcal{P} = \{p \subset [n] : |p| = m\}$ denote the set of all subset of $[n]$ of size m . Applying the Cauchy-Binet formula as stated in Theorem 1, one obtains

$$\det(H^\top H) = \sum_{p \in \mathcal{P}} \det(H_p) \det(H_p^\top) = \sum_{p \in \mathcal{P}} \det(H_p)^2. \quad (34)$$

Using Equation (34), one can write \hat{s}_1 in the following form:

$$\begin{aligned} \hat{s}_1 &= \frac{\sum_{p \in \mathcal{P}} \det(H_p)^2 H_p^{-1} x_p}{\det(H^\top H)} \\ &= \frac{\sum_{p \in \mathcal{P}} \det(H_p)^2 \text{emb}(H_p^{-1}, p, n) x}{\det(H^\top H)}. \end{aligned}$$

Recall that for $p \in \mathcal{P}$ with elements $p_1 < p_2 < \dots < p_m$, we denote by $\text{emb}(H_p^{-1}, p, n)$ an all-zero $m \times n$ matrix except for its p_i -th column which is equal to the i -th column of H_p^{-1} , $i = 1, 2, \dots, m$. Therefore, it is essentially sufficient to prove that for every $x \in \mathbb{R}^n$, we have the following equality

$$(H^\top H)^{-1} H^\top x = \frac{\sum_{p \in \mathcal{P}} \det(H_p)^2 \text{emb}(H_p^{-1}, p, n)}{\det(H^\top H)} x. \quad (35)$$

Since this is true for every $x \in \mathbb{R}^n$, we need to prove that for every $n \times m$ matrix H with $n \geq m$ and with full column rank, the following matrix identity holds:

$$\det(H^\top H) (H^\top H)^{-1} H^\top = \sum_{p \in \mathcal{P}} \det(H_p)^2 \text{emb}(H_p^{-1}, p, n) \quad (36)$$

In order to prove the identity (36), let us define the function $F: M_{\mathbb{R}}(n, m) \rightarrow \mathbb{R}$ as follows:

$$F(H) = \det(H^\top H) - \sum_{p \in \mathcal{P}} \det(H_p)^2 \quad (37)$$

From Cauchy-Binet identity as stated in Theorem 1, it is clear that $F(H) = 0$ for every $H \in M_{\mathbb{R}}(n, m)$. Let $t = H_{ij}$ be a

parameter denoting the component of H at row i and column j . It results that $\frac{\partial}{\partial t} F(H) = 0$ and this is true for every $t = H_{ij}$, $i \in [n], j \in [m]$. One has

$$\begin{aligned}
\frac{\partial}{\partial t} \det(H^\top H) &\stackrel{(a)}{=} \det(H^\top H) \text{tr} \left\{ (H^\top H)^{-1} \frac{\partial}{\partial t} (H^\top H) \right\} \\
&\stackrel{(b)}{=} \det(H^\top H) \text{tr} \left\{ (H^\top H)^{-1} \left(\left(\frac{\partial}{\partial t} H \right)^\top H + H^\top \frac{\partial}{\partial t} H \right) \right\} \\
&\stackrel{(c)}{=} \det(H^\top H) \text{tr} \left\{ (H^\top H)^{-1} \left(H^\top \frac{\partial}{\partial t} H + H^\top \frac{\partial}{\partial t} H \right) \right\} \\
&= 2 \det(H^\top H) \text{tr} \left\{ (H^\top H)^{-1} H^\top \frac{\partial}{\partial t} H \right\} \\
&\stackrel{(d)}{=} 2 \det(H^\top H) \text{tr} \left\{ (H^\top H)^{-1} H^\top U_{ij} \right\} \\
&\stackrel{(e)}{=} 2 \det(H^\top H) \left\langle (H^\top H)^{-1} H^\top, U_{ij}^\top \right\rangle,
\end{aligned}$$

where (a) follows from Lemma (3) applied to the matrix $H^\top H$, (b) follows from the chain rule applied to $H^\top H$, (c) follows from Lemma (4) applied to the symmetric matrix $(H^\top H)^{-1}$ and matrix $\left(\frac{\partial}{\partial t} H\right)^\top H$, (d) results by taking the component-wise derivative of H with respect to $t = H_{ij}$ which we denote by U_{ij} , and where (e) results from the definition of the inner product for two matrices. One can simply check that U_{ij} is an $n \times m$ matrix with all-zero components except for ij -th component which is equal to 1. Taking the derivative of the other term in Equation (37) with respect to $t = H_{ij}$, one obtains

$$\begin{aligned}
\frac{\partial}{\partial t} \sum_{p \in \mathcal{P}} \det(H_p)^2 &= \sum_{p \in \mathcal{P}} 2 \det(H_p) \frac{\partial}{\partial t} \det(H_p) \\
&\stackrel{(a)}{=} \sum_{p \in \mathcal{P}} 2 \det(H_p) \det(H_p) \text{tr} \left(H_p^{-1} \frac{\partial}{\partial t} H_p \right) \\
&\stackrel{(b)}{=} \sum_{p \in \mathcal{P}} 2 \det(H_p)^2 \text{tr} \left(\text{embb}(H_p^{-1}, p, n) \frac{\partial}{\partial t} H \right) \\
&\stackrel{(c)}{=} 2 \text{tr} \left\{ \sum_{p \in \mathcal{P}} \det(H_p)^2 \text{embb}(H_p^{-1}, p, n) U_{ij} \right\} \\
&\stackrel{(d)}{=} 2 \left\langle \sum_{p \in \mathcal{P}} \det(H_p)^2 \text{embb}(H_p^{-1}, p, n), U_{ij}^\top \right\rangle,
\end{aligned}$$

where (a) results from Lemma (3) applied to the matrix H_p ; (b) results for the definition of embedding m columns of H_p^{-1} in an $m \times n$ matrix, specially that remaining columns of the matrix $\text{embb}(H_p^{-1}, p, n)$ are all zero which allows us to replace H_p by H , (c) results from the linearity of the trace operator tr and the definition of matrix U_{ij} , and (d) results from the definition of the matrix inner product. Therefore, one obtains that

$$0 = \frac{\partial}{\partial t} F(H) = 2 \left\langle U_{ij}^\top, \det(H^\top H) (H^\top H)^{-1} H^\top - \sum_{p \in \mathcal{P}} \det(H_p)^2 \text{embb}(H_p^{-1}, p, n) \right\rangle.$$

Notice that equality holds for all matrices U_{ij}^\top , $i \in [n], j \in [m]$. Since, U_{ij}^\top form an orthonormal basis for the linear space $M_{\mathbb{R}}(m, n)$ under the standard matrix inner product, from

Lemma 1, it immediately results that

$$\det(H^\top H) (H^\top H)^{-1} H^\top = \sum_{p \in \mathcal{P}} \det(H_p)^2 \text{embb}(H_p^{-1}, p, n).$$

From Equation 36, this is exactly what we needed to prove. \blacksquare

The content of this appendix is available online [46] for the general audiences out of the scope of the present manuscript.

ACKNOWLEDGMENT

The authors acknowledge Rahil Mahdian for the speech recognition scripts and evaluation on MC-WSJ. Afsaneh Asaei acknowledges SNSF 200021-153507 project on ‘‘Parsimonious Hierarchical Automatic Speech Recognition (PHASER)’’. Volkan Cevher acknowledges Rice University for his Faculty Fellowship, MIRG-268398, ERC Future Proof, DARPA Ke-CoM program #11-DARPA-1055 and SNF 200021-132548.

REFERENCES

- [1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.
- [2] H. Buchner, R. Aichner, and W. Kellermann, *TRINICON-based blind system identification with application to multiple-source localization and separation*. In *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. New York: Springer, 2007, vol. 13.
- [3] F. Nesta and M. Omologo, *Convolutional Underdetermined Source Separation through Weighted Interleaved ICA and Spatio-temporal Source Correlation*. In: Yeredor, A. et al. (eds.) *LVA/ICA 2012*. LNCS, Springer, Heidelberg, 2012, vol. 7191.
- [4] L. Wang, H. Ding, and F. Yin, ‘‘A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures,’’ *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 19(3), 2011.
- [5] M. I. Mandel, D. P. Ellis, and T. Jebara, ‘‘An em algorithm for localizing multiple sound: Sources in reverberant environments,’’ *Advances in neural information processing systems 19 proceedings of the 2006 conference*, pp. 953–960, 2007.
- [6] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, ‘‘Underdetermined blind separation for speech in real environments with sparseness and ICA,’’ in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.
- [7] M. Davies and N. Mitianoudis, ‘‘Simple mixture model for sparse overcomplete ICA,’’ *IEE Proceedings on Vision, Image and Signal Processing*, vol. 151 (1), 2004.
- [8] A. Asaei, M. J. Taghizadeh, M. Bahrololum, and M. Ghanbari, ‘‘Verified speaker localization utilizing voicing level in split-bands,’’ *Signal Processing*, vol. 89(6), 2009.
- [9] M. J. Taghizadeh, P. N. Garner, H. Boursard, and H. R. Abutalebi, ‘‘An integrated framework for multi-channel multi-source speaker localization and source activity detection,’’ in *Proceedings of HSCMA*, 2011.

- [10] M. A. Dmour and M. E. Davies, "A new framework for underdetermined speech extraction using mixture of beamformers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 445–457, 2011.
- [11] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 550–563, 2010.
- [12] S. Lee, S. H. Park, and K.-M. Sung, "Beamspace-domain multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 43–46, 2012.
- [13] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.
- [14] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 13(5), 2005.
- [15] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 51(1), 2003.
- [16] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19(1), 2010.
- [17] T. Nakatani, T. Yoshioka, and K. Kinoshita, "Mathematical analysis of speech dereverberation based on time-varying gaussian source model: Its solution and convergence characteristics," in *Proceedings of IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2011.
- [18] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13(4), 2001.
- [19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *International Conference on Machine Learning (ICML)*, 2009.
- [20] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.
- [21] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environment using despritt," *EURASIP Journal on Advances in Signal Processing*, 2007.
- [22] F. Abrard and Y. Deville, "A time-frequency blind signal separation method applicable to under-determined mixtures of dependent sources," *Elsevier, Signal Processing*, 2005.
- [23] A. Asaei, "Model-based sparse component analysis for multiparty distant speech recognition," Ph.D. dissertation, Ecole Polytechnique Federal de Lausanne (EPFL), 2013.
- [24] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Speech and Audio Processing*, vol. 22 (3), pp. 620–633, 2014.
- [25] A. Asaei, B. Raj, H. Bourlard, and V. Cevher, "Structured sparse coding for microphone array position calibration," in *Proceeding of 5th ISCA workshop on Statistical and Perceptual Audition, SAPA-SCALE Conference*, 2012.
- [26] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 36(2), 1988.
- [27] A. Kyrillidis and V. Cevher, "Recipes on hard thresholding methods," in *Proceedings of CAMSAP*, 2011.
- [28] A. Asaei, H. Bourlard, M. Taghizadeh, and V. Cevher, "Model-based sparse component analysis for reverberant speech localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [29] I. A. McCowan, "The Multichannel Overlapping Numbers Corpus," Idiap resources available online: <http://www.cslu.ogi.edu/corpora/monc.pdf>, 2003.
- [30] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2005.
- [31] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Second International Workshop on Machine Learning for Multimodal Interaction*, 2005.
- [32] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wolfel, and D. Klakow, "To separate speech - a system for recognizing simultaneous speech," in *Fourth International Workshop on Machine Learning for Multimodal Interaction*, 2007.
- [33] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53(2), 2005.
- [34] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16(3), 2008.
- [35] M. J. Taghizadeh, P. N. Garner, and H. Bourlard, "Enhanced diffuse field model for ad hoc microphone array calibration," *Signal Processing*, vol. 101, 2014.
- [36] E. A. Habets, *Speech Dereverberation Using Statistical Reverberation Models*. Speech Dereverberation, Springer, 2010.
- [37] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *Journal of the Acoustical Society of America*, vol. 27(6), 1955.
- [38] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 11(6), 2003.
- [39] T. Wolff and M. Buck, "A generalized view on microphone array postfilters," in *International Workshop on Acoustic Signal Enhancement*, 2010.

- [40] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29 (6), 2012.
- [41] D. C. Moore and I. A. Mccowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," in *Proceedings of ICASSP*, 2003.
- [42] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *Journal of the Acoustical Society of America*, vol. 27(6), 1955.
- [43] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation (code available at <http://www.irisa.fr/metiss/sasse07/?show=results>)," *IEEE transactions on audio, speech, and language processing*, vol. 14, 2006.
- [44] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *International Telecommunications Union, Geneva, Switzerland*, 2001.
- [45] R. Mahdian, F. Faubel, and D. Klakow, "Multi-channel speech separation with soft time-frequency masking," in *SAPA-SCALE Conference*, 2012.
- [46] S. Haghghatshoar, M. J. Taghizadeh, and A. Asaei, "A new identity for the least-square solution of overdetermined set of linear equations," *arXiv preprint arXiv:1502.07695*, 2015.



Afsaneh Asaei Afsaneh Asaei received the B.S. degree from Amirkabir University of Technology and the M.S. (honors) degree from Sharif University of Technology, in Electrical and Computer engineering, respectively. She held a research engineer position at Iran Telecommunication Research Center (ITRC) from 2002 to 2008. She then joined Idiap Research Institute in Martigny, Switzerland, and was a Marie Curie fellow on speech communication with adaptive learning training network. She received the Ph.D. degree in 2013 from École Polytechnique Fédérale de

Lausanne. Her thesis focused on structured sparsity for multiparty reverberant speech processing, and its key idea was awarded the IEEE Spoken Language Processing Grant. Currently, she is a postdoctoral researcher at Idiap Research Institute. She has served as a guest editor of Speech Communication special issue on Advances in sparse modeling and low-rank modeling for speech processing and co-organized special issues on this topic at HSCMA'2014 and LVA/ICA'2015. Her research interests lie in the areas of signal processing, machine learning, statistics, acoustics, auditory scene analysis and cognition, and sparse signal recovery and acquisition.



Mohammad Javad Taghizadeh received his PhD from École Polytechnique Fédérale de Lausanne, Switzerland in 2015. During his doctoral studies he was a research assistant at Idiap Research Institute in Martigny, Switzerland where he conducted research on ad hoc microphone arrays signal processing techniques to enable higher level distant speech applications. Prior to doctoral studies, he was research engineer at Information Processing group at École Polytechnique Fédérale de Lausanne and Telecommunication research centres in Tehran, Iran.

He is currently a senior researcher on sound field analysis and synthesis for 3D audio technologies at Huawei European Research Centre in Munich, Germany.



Saeid Haghghatshoar (S'12) received the B.Sc. degree in electrical engineering in 2007 in Electronics and the M.Sc. degree in electrical engineering in 2009 in Communication Systems both from Sharif University of Technology, Tehran, Iran, and the Ph.D. degree in Computer and Communication Sciences from École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. His research interests lie in information theory, Communication systems, Graphical models and Compressed sensing.



speech recognition, natural language processing and machine learning

Bhiksha Raj is an associate professor in the Language Technologies Institute of Carnegie Mellon University, with additional affiliations to the Machine Learning and Electrical and Computer Engineering departments of the University. Dr Raj completed his PhD from Carnegie Mellon University in 2000. From 2001–2008 he worked at Mitsubishi Electric Research Labs in Cambridge MA, where he led the research effort on speech processing. He has been at CMU since 2008. Dr Raj's research interests include speech and audio processing, automatic



speech recognition, natural language processing and machine learning

Hervé Bourlard received the Electrical and Computer Science Engineering degree and the PhD degree in Applied Sciences both from "Faculté Polytechnique de Mons", Mons, Belgium. After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute, Full Professor at the Swiss Federal Institute of Technology Lausanne (EPFL), and Founding Director of the Swiss NSF National Centre of Competence in Research on "Interactive Multimodal Information Management (IM2)". Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI), Berkeley, CA, he is now a member of an ICSI External Fellow and a member of its Board of Trustees.

His main research interests mainly include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modeling, multimodal interaction, augmented multi-party interaction, and distant group collaborative environments.

H. Bourlard is the author/coauthor/editor of 8 books and over 330 reviewed papers (including one IEEE paper award) and book chapters. He is a Fellow of IEEE and ISCA and a Senior Member of ACM. He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of IEEE ICASSP 2002, General Chairman of Interspeech 2003) and on the Editorial Board of several journals (e.g., past co-Editor-in-Chief of "Speech Communication"). He is the recipient of several scientific and entrepreneurship awards.



Engineering Department at Rice University. His research interests include signal processing theory, machine learning, graphical models, and information theory. Dr. Cevher received a Best Paper Award at SPARS in 2009 and an ERC StG in 2011.

Volkan Cevher received a B.Sc. (valedictorian) in Electrical Engineering in 1999 from Bilkent University in Ankara, Turkey, and he received a Ph.D. in Electrical and Computer Engineering in 2005 from the Georgia Institute of Technology in Atlanta, GA. He held research scientist positions at the University of Maryland, College Park from 2006–2007 and at Rice University in Houston, TX, from 2008–2009. Currently, he is an Assistant Professor at the Swiss Federal Institute of Technology Lausanne and a Faculty Fellow in the Electrical and Computer