Structured Sparsity Models for Reverberant Speech Separation

Afsaneh Asaei, *Student Member, IEEE,* Mohammad Golbabaee, Hervé Bourlard, *Fellow, IEEE* and Volkan Cevher, *Senior Member, IEEE*

Abstract

We tackle the speech separation problem through modeling the acoustics of the reverberant chambers. Our approach exploits structured sparsity models to perform speech recovery and room acoustic modeling from recordings of concurrent unknown sources. The speakers are assumed to lie on a two-dimensional plane and the multipath channel is characterized using the image model. We propose an algorithm for room geometry estimation relying on localization of the early images of the speakers by sparse approximation of the spatial spectrum of the virtual sources in a free-space model. The images are then clustered exploiting the low-rank structure of the spectro-temporal components belonging to each source. This enables us to identify the early support of the room impulse response function and its unique map to the room geometry. To further tackle the ambiguity of the reflection ratios, we propose a novel formulation of the reverberation model and estimate the absorption coefficients through a convex optimization exploiting joint sparsity model formulated upon spatio-spectral sparsity of concurrent speech representation. The acoustic parameters are then incorporated for separating individual speech signals through either structured sparse recovery or inverse filtering the acoustic channels. The experiments conducted on real data recordings of spatially stationary sources demonstrate the effectiveness of the proposed approach for multi-party speech recovery and recognition.

Index Terms

Afsaneh Asaei and Hervé Bourlard are with Idiap Research Institute and also affiliated with École Polytechnique Fédérale de Lausanne, Switzerland. Mohammad Golbabaee is with CEntre de REcherches en MAthématiques de la DEcision (CEREMADE), Université Paris IX, Paris Dauphine, France. Volkan Cevher is with École Polytechnique Fédérale de Lausanne, Switzerland.

E-mails: {afsaneh.asaei, herve.bourlard}@idiap.ch, volkan.cevher@epfl.ch
golbabaee@ceremade.dauphine.fr

Multi-party reverberant recordings, Structured sparse recovery, Room acoustic modeling, Image model, Distant speech recognition

I. INTRODUCTION

RECOVERY of speech signals from an acoustic clutter of unknown competing sound sources plays a key role in many applications involving distant-speech recognition, scene analysis, video-conferencing, hearing aids, surveillance, sound-field equalization and sound reproduction. Despite the vast efforts devoted to the issues arising in real-world conditions, development of systems to operate in the presence of overlapping sound sources yet remains a demanding challenge [1].

In this paper, we consider distant-talking speech recognition in a multi-party environment where multiple sound sources talk simultaneously. The common existence of overlapped speech segments has been shown to increase the speech recognition word error rate up to 30% for a large vocabulary task [2] hence, it is required to incorporate an effective source separation technique to segregate the desired speech from the competing signals prior to recognition. We assume that the signals are acquired by an array of calibrated microphones.

Previous approaches to multi-channel speech separation can be broadly group into three classes. The first category incorporates a prior knowledge about mutual independence and statistical characteristics of the source signals to identify the mixing model and to recover the individual sources [3]. These techniques are usually confined to the scenarios where the number of microphones is greater than or equal to the number of sources referred to as *overdetermined* or *determined* mixtures respectively and their performance degrades due to reverberation [4].

The second category relies on spatial filtering based on beamforming or steering a microphone array beam-pattern towards the target speaker to enable suppression of the undesired sources [5, 6]. The underlying assumption of this approach is that there is no reverberation so the beamforming techniques are formulated upon direct path acquisition of the signals. These geometric approaches to speech separation can work with any number of microphones including the scenarios in which the number of sources exceeds the number of sensors referred to as *underdetermined* mixtures. A limitation is that the standard beamforming techniques overlook the model of acoustic multipath and they are less effective in reverberant condition [7, 8].

The third category is based on sparse representation of the source signal, also known as *sparse component analysis* [9, 10]. These techniques exploit a prior assumption that the sources have a sparse representation in a known basis or frame. The notion of sparsity opens a new road to address the

underdetermined unmixing problem to estimate the unknown variables from fewer known data. Since the underdetermined linear system admits infinitely many solutions, the answer ought to be the sparest solution measured in terms of the sparsity inducing norms [10, 11]. The prior art on multichannel speech recovery exploiting sparsity models are largely confined to the recovery of the signals at individual frequency level and ignore the higher-level structures exhibited in data representation.

The approach that we propose in this paper relies on structured sparsity models underlying multiparty multi-channel recordings in reverberant environments. We discretize the planar area of the room into a grid of uniform cells where each of the speakers is located at one of the cells. If there are N speakers in the room and given a fine grid of G cells such that the cell's occupancy is exclusive, the distribution of the sources in the room is sparse; i.e., out of G cells only $N \ll G$ contain the sound sources. This implies the spatial sparsity model as depicted in Fig. 1.

Denoting the signal attributed to the source located at cell g as S_g and concatenating the signals corresponding to each cell, the signal vector coming from all over the room can be formed as $\mathcal{S} = [S_1^T, ..., S_G^T]^T$ where T stands for the transpose operator. If we consider one instance of recordings from N speakers, $\mathcal{S}$ is a sparse vector with only N non-zero elements. The support of $\mathcal{S}$ corresponds to the N cells where the sources are located. If we consider F instances (e.g. frequency bins) of recordings and assume that sources are immobile, each instance of the signal of a particular source implies sparsity in exactly the same manner as every other instances as they all correspond to the one particular cell where the source is located. This extra restriction imposes a constraint on the structure of the elements in $\mathcal{S}$ which goes beyond simple sparsity. We characterize sparsity with such constraints as structured sparsity. Fig. 1 illustrates the particular block sparsity model exhibited in representation of the signals coming from all over the grid as described above.

This paper exploits structured sparsity models to recover the unknown individual speech signals: $S_g, g \in \{1, ..., G\}$ from a few known mixed recordings when the speakers are talking simultaneously. In addition to the spatial sparsity and block dependency, we exploit harmonic sparsity of spectral components. The spectral structure of voiced speech typically comprises a small number of spectral peaks at harmonics of a fundamental frequency; at other frequencies the energy is typically low or negligible. We can therefore model the distribution of energy over frequencies as being sparse. Furthermore, we exploit the structured sparsity underlying the acoustic channel of the room characterized by the image model of the multipath effect. The contribution of this paper is ultimately to introduce a unified theory of spatio-spectral speech separation formulated as a problem of sparse recovery of information embedded in multichannel recordings exploiting structured sparsity models.

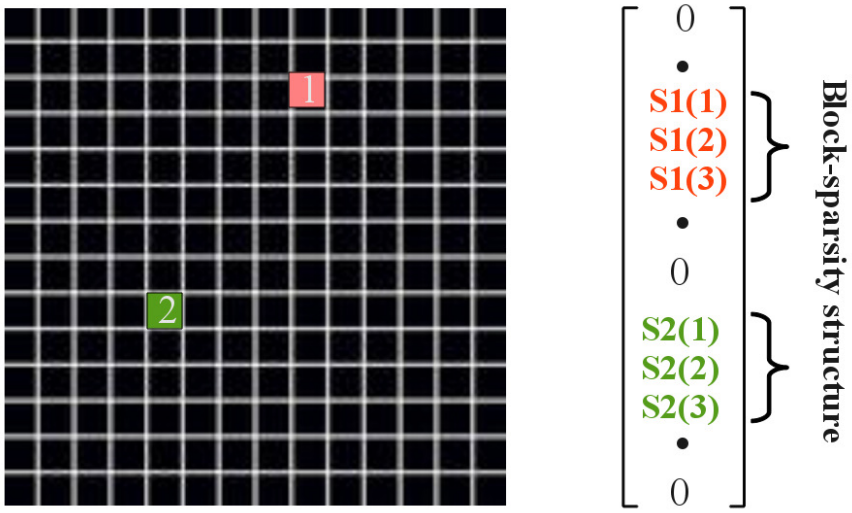

Fig. 1: The spatial sparsity of the speakers inside the room is illustrated through discretization of the planar area of the room into a grid of G cells. The sources occupy only two cells marked as 1 and 2. Hence, the spatial representation of the source signals generated inside the room is sparse.
Assuming that the sources are immobile, if we denote an arbitrary F (e.g. $F = 3$) instances of the signal attributed to the speaker at cell g as $S_g(f), f \in \{1, \cdots, F\}$ and concatenate the signals corresponding to each cell, the signal vector of the room can be formed as $\mathcal{S} = [S_1^T, ..., S_G^T]^T \in \mathbb{C}^{GF\times 1}$. We can see that support of $\mathcal{S}$ exhibits the block-sparsity structure as there are only two blocks of non-zero elements corresponding to the two speakers. The size of each block is the number of recording instances.

II. State-of-the-Art

This paper tackles the multi-party speech recovery problem through modeling the acoustic of the enclosure and exploiting sparsity models. The room acoustic characterization was earlier incorporated in the method proposed in [12]. Their approach relies on statistical independence assumption of the sources to estimate the acoustic channel of the enclosure and perform joint deconvolution and separation of speech signals; its applicability is limited to overdetermined scenarios. This assumption has been relaxed in the method proposed in [13] where multiple complex valued independent component analysis adaptations jointly estimate the mixing matrix and the temporal activities of multiple sources in each frequency band to exploit the spectral sparsity of speech signals. However, it does not explicitly rely on identification of the acoustic channel and recovery of the desired source imposes a permutation problem due to mis-alignment of the individual source components [13].

A blind channel identification approach for speech separation and dereverberation is proposed in [14]. In this paper, the mixing procedure is delineated with a multiple-input multiple-output (MIMO) mathematical model. The authors propose to decompose the convolutive source separation problem into sequential

procedures to remove spatial interference at the first step followed by deconvolution of temporal echoes. To separate the speech interferences, the MIMO system of recorded overlapping speech in reverberant environment is converted into the single-input-multi-output (SIMO) systems corresponding to the channel associated with each speaker. The SIMO channel responses are then estimated using the blind channel identification through the unconstrained normalized multi-channel frequency-domain least mean square algorithm [15] and dereverberation can be performed based on the Bezout theorem also known in the context of room acoustics as the multiple-input/output inverse-filtering theorem (MINT) [16]. A real-time implementation of this approach has been presented in [17], where the optimum inverse filtering is substituted by an iterative technique, which is computationally more efficient and allows the inversion of long room impulse responses in real-time applications [17]. The major drawback of such implementation is that it can only perform channel identification from single talk periods and it requires a high input signal-to-noise ratio. Another approach to perform joint dereverberation and speech separation extends the maximum likelihood criteria applied in weighted prediction error method using determined and overdetermined mixtures [18]. This method assumes that the source spectral components are uncorrelated across time frames and it relies on a single source assumption for estimation of the acoustic channel, thus it can not achieve dereverberation when there are multiple sound sources [19].

This paper takes a new perspective to the objective of multi-channel processing as recovery of the information embedded in the acoustic field from compressive acquisition provided by microphone array. We derive a spatio-spectral representation of concurrent sound sources and characterize the acoustic reverberation model to formulate a model-based sparse component analysis framework for identification of the source locations and separation of the individual spectral components. The proposed framework incorporates the model underlying spectrographic speech representation as well as the acoustic channel for extraction of the information bearing components. More specifically, our approach features the following contributions:

- ◇ Model-based sparse component analysis framework for speech separation and localization incorporating spectral, spatial and acoustic multipath structures.
- ◇ Room geometry estimation algorithm from recordings of multiple unknown sources relying on sparse recovery and low-rank clustering.
- ◇ Formulation of the reverberation model factorized into free-space propagation and source permutation to model the multipath effect.
- ◇ Room absorption coefficient estimation algorithm from recordings of multiple unknown sources

using model-based sparse recovery.

- ◇ Analysis of the performance of computational approaches to model-based sparse recovery considering speech-specific structures.
- ◇ Analysis of the performance of speech recovery considering the design of microphone array layout.

The rest of the paper is organized as follows: The problem statement and sparse representation of multiparty reverberant recordings is explained in Section III. We set up the formulation of the structured sparse acoustic modeling in Section IV; the room geometry estimation algorithm is elaborated in Section IV-B and a formulation of reverberation model for absorption coefficient estimation is derived in Section IV-C. The structured sparse speech recovery algorithms are described in Section V and the theoretical analysis of the performance bound is explained in Section VI. The experimental analysis are presented in Section VII and finally, the conclusions are drawn in Section VIII.

III. Reverberant Speech Recordings

A. Problem Statement

We address the problem of separating the signals of an unknown number of speakers from multi-channel recordings in a reverberant room. We consider an approximate model of the acoustic observation as a linear convolutive mixing process, stated concisely as

$$x_m = \sum_{n=1}^{N} h_{mn} \circledast s_n, \quad m = 1, ..., M \tag{1}$$

where x_m and s_n denote the time domain signal of the m^{th} microphone and n^{th} source respectively; h_{mn} denotes the acoustic channel between signal and microphone and $\circledast$ is the convolution operator. M and N indicate the total number of microphones and sources respectively. This formulation is stated in time domain; to represent it in a sparse domain, we apply the discrete Short-Time Fourier Transform (STFT) on speech signals. Following from the convolution-multiplication property of the Fourier transform, the mixtures in frequency domain can be written as

$$X_m(f, \tau) = \sum_{n=1}^{N} H_{mn} S_n(f, \tau), \quad m = 1, ..., M \tag{2}$$

where X_m, S_n and H_{mn} denote the microphone and source signals and their corresponding acoustic channel in Fourier domain. f and τ indicate the frequency and frame index respectively. Our objective is to recover the individual source signals S_n from the distant microphone recordings. There is no prior information about the number of sources and the acoustic mixing channels.

B. Spatio-Spectral Sparse Representation

To obtain the *sparse representation* of multiparty speech sources, we consider a scenario in which N speakers are distributed in a planar area (at the same height in three-dimensional space) spatially discretized into a grid of G cells. We assume to have a sufficiently dense grid so that each speaker is located at one of the cells thus $N \ll G$. The spatial spectrum of the sources is defined as a vector with a sparse support indicating the components of the signal corresponding to each cell of the grid.

We consider the spectro-temporal representation of multi-party speech and entangle the spatial representation of the sources with the spectral representation of the speech signal to form vector $\mathcal{S} = [S_1^T ... S_G^T]^T \in \mathbb{C}^{GF \times 1}$. Each $S_g \in \mathbb{C}^{F \times 1}$ denotes the spectral representation of the signal of the g^{th} source (located at cell number g) in Fourier domain. We express the signal ensemble at the microphone array as a single vector $\mathcal{X} = [X_1^T ... X_M^T]^T$ where each $X_m \in \mathbb{C}^{F \times 1}$ denotes the spectral representation of the recorded signal at microphone number m. The sparse vector $\mathcal{S}$ generates the microphone observations as $\mathcal{X} = \Phi\mathcal{S}$. Φ is the microphone array measurement matrix consisted of the acoustic projections associated to the acquisition of source signals located on the grid. In the following Section IV, we propose a method to characterize the acoustic measurements.

IV. Structured Sparse Acoustic Modeling

A. Characterizing the Acoustic Measurements

We assume the room to be a rectangular enclosure consisting of finite impedance walls. The point source-to-microphone impulse responses of the room are calculated using the *image model* technique [20]. Taking into account the physics of the signal propagation and multipath effects, the projections associated with the source located at cell g where ν_g represents the position of the center of the cell and captured by microphone m located at position μ_m are characterized by the media Green's function and denoted as $\xi^f_{\nu_g \to \mu_m}$ defined by

$$\xi^f_{\nu_g \to \mu_m} : X(f,\tau) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu_m - \nu_g^r\|^\alpha} \exp(-jf\frac{\|\mu_m - \nu_g^r\|}{c}) S(f,\tau), \tag{3}$$

where $j = \sqrt{-1}$ and ν_g^r designates the location of the r^{th} virtual source corresponding to the actual source located at cell g with the corresponding reflective energy ratio of ι^r. R denotes the number of source images and c is the speed of sound. The attenuation constant α depends on the nature of the propagation and is considered in our model to equal 1 which corresponds to spherical propagation. This

formulation assumes that if $s_1(l) = s(l)$ and $s_2(l) = s(l-\rho)$, then $S_2(f,\tau) \approx \exp(-jf\rho)S_1(f,\tau)$; hence, the frame size should be greater than the length of the impulse response for this assumption to hold.

Given the source-sensor projection defined in (3), we construct the matrix of the F consecutive frequencies as $\Xi_{\nu_g \to \mu_m} = \text{diag}(\xi^1_{\nu_g \to \mu_m}, \ldots, \xi^F_{\nu_g \to \mu_m})$. Hence, the projections associated to the acquisition of the source signals located on the grid by microphone m is $\phi_m = [\Xi_{\nu_1 \to \mu_m} \ldots \Xi_{\nu_g \to \mu_m} \ldots \Xi_{\nu_G \to \mu_m}]$ and the M-channel microphone array manifold matrix is obtained as $\Phi = [\phi_1^T \ldots \phi_M^T]^T$. Thereby, characterizing the acoustic projections amounts to identifying the *location of the source images* as well as the *absorption coefficients* of the reflective surfaces. We exploit this parametric model to address the speech recovery problem. In the following Section IV-B, we estimate the geometry of the room to identify the location of the source images. In Section IV-C, we estimate the absorption coefficients of the reflective surfaces.

B. Estimation of the Room Geometry

The projection expressed in (3) corresponds to characterization of the forward model of the room acoustic channel as

$$H(f, \mu_m, \nu_g) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu_m - \nu_g^r\|} \exp(jf \frac{\|\mu_m - \nu_g^r\|}{c}) \tag{4}$$

$H(f, \mu_m, \nu_g)$ indicates the room impulse response function between the microphone located at μ_m and a source located at ν_g. Hence, identifying the locations of the R images of the source enables identifying the temporal support of the room impulse response function. According to the image model, if the geometry of the enclosure is known, it is possible to identify the source images up to any arbitrary order [20].

Recent studies have shown that the impulse response function is a unique signature of the room and the geometry can be reconstructed given that up to second order of reflections are known [21, 22]. Relying on this observation, we propose to localize the source images by sparse recovery with a free-space measurement model, i.e., $R = 0$, while the deployment of the grid captures the location of early reflections. The time support of the acoustic channel, $\{\nu^r \,|\, 1 < r < R\}$ corresponds to the cells where the recovered energy of the signal is maximized. We consider the localized sources in a *close proximity* to the microphone array within a pre-specified distance range as the actual sources generating the signals $S_n, n = \{1, ..., N\}$. The localized images are sorted up to the order of $D(D+1)/2$ where D indicates the number of reflective surfaces according to the cosine angle between the estimated signals and the source signal (S_g) and considered as the images associated to the g^{th} source. The cosine angle is the appropriate distance measure to cluster the components which are geometrically aligned, i.e., images of

the same source. The bound of $\mathrm{D}(\mathrm{D}+1)/2$ guarantees a unique map to the geometry of the enclosure as proved in [21]. Given the location of the source images, we estimate the room geometry by brute-force search to identify the dimensions which generate the least-squares approximation of the location of virtual sources [22]. Algorithm 1 summarizes the steps to implement room geometry estimation.

Algorithm 1 Room geometry estimation

(i) Run sparse source localization algorithm with a free-space measurement model.
(ii) Run k-means clustering using cosine angle as the distance metric.
  ▷ Select the centroid of the clusters as the nearest (actual) sources to the array center.
  ▷ Measure the cosine angle between components of virtual and actual sources.
  ▷ Keep the closest $\mathrm{D}(\mathrm{D}+1)/2$ sources as the cluster members.
(iii) Find the room geometry by identifying the dimensions which yield the best approximation of the location of source images in least-squares sense.

The approach that we presented in this section can estimate the room geometry if a single source or multiple unknown sources exist in the room. Applying the image model to a rectangular room, a lattice of virtual sources is obtained. As the temporal support of room impulse response is attributed to the source images, the image model of multipath propagation insinuates temporal sparsity of the early part of the room impulse response function with a particular structure. We refer to this property as the acoustic structured sparsity and exploit it to address the problem of estimating the absorption coefficients.

C. Estimation of Absorption Coefficients

This section elaborates on a novel formulation of the reverberant recordings which entangles the structured sparsity indicated by the image model and the spatio-spectral sparsity of multiparty recordings for joint estimation of the absorption coefficients and recovery of the sources. This approach enables us to estimate the frequency-dependent absorption factors in a multi-source environment.

1) Factorized Formulation of the Reverberant Recordings: We formulate the reverberation model factorized into permutation (corresponding to the source images) and attenuation (corresponding to the absorption factors) of the sources in an unbounded space.

We assume that the G-cells grid of the room containing N sources is expanded into $\mathcal{G}$-cells free-space discretization where the actual-virtual sources are active. If each of the sources have R images, $\mathsf{N}(\mathsf{R}+1)$ actual-virtual sources are active. Given the geometry of the room, the image model maps the position index $\mathfrak{i} \in \{1,\ldots,\mathsf{G}\}$ of each source to a group $\Omega_{\mathfrak{i}} \in \{1,\ldots,\mathcal{G}\}$ containing the location indices of this source and its images (the corresponding virtual sources) in $\mathcal{G}$-points. Consequently, a *free-space*

propagation model can be considered between $\mathcal{G}$ actual-virtual source locations and the positions of M microphones. Hence, the forward model between sources and the microphone recordings can be concisely stated as follows:

$$\mathcal{X} = OP\mathcal{S}. \tag{5}$$

This model holds for each particular independent frequency f of the speech spectrum so we discard the frequency dependency in our mathematical formulation for the sake of brevity. Given $\mathcal{X} \in \mathbb{C}^{M \times \mathcal{T}}$, the *observation* matrix of $\mathcal{T}$ frames consisting of the spectro-temporal representation of M microphones at a particular frequency band, we decompose the microphone recordings into the following terms:

- $\mathcal{S} \in \mathbb{C}^{G \times \mathcal{T}}$ is the *source* matrix whose rows contain $\mathcal{T}$ frames of the spectro-temporal representation of the *actual* sources located in G positions inside the room. Given a fine discretization of the room such that each source occupy an exclusive cell, only $N \ll G$ cells are occupied with active sources and contain nonzero elements and the *support* of $\mathcal{S}$ represents the position of those N active sources is sparse. In other words, the *spatial sparsity* indicates $\mathcal{S}$ to be a row-sparse matrix with a support corresponding to the position of the actual sources.
- $P \in \mathbb{R}_{+}^{\mathcal{G} \times G}$ is the *permutation* matrix such that its i^{th} column contains the absorption factors of $\mathcal{G}$ points on the grid of actual-virtual sources with respect to the reflection of the i^{th} actual source. Since the image model characterizes the source groups, each column $P_{.,i}$ is consequently supported only on the corresponding group Ω_i i.e., $\forall i \in \{1 \ldots, G\}$, $\forall j \notin \Omega_i, P_{j,i} = 0$.
- $O \in \mathbb{C}^{M \times \mathcal{G}}$ is the *free-space Green's function* matrix such that each $O_{j,i}$ component indicates the sound propagation coefficients, i.e. the attenuation factors and the phase shift due to the direct path propagation of the sound source located at cell i (on a $\mathcal{G}$-point grid of actual-virtual sources) and recorded at the j^{th} microphone. Given the $\mathcal{G}$-cell discretization, O is computed from the propagation formula stated in (3) and it is equal to Φ when $R = 0$.

2) Source Localization and Absorption Coefficient Estimation: Relying on the spatio-spectral sparsity of multiple competing sources, the covariance matrix of the reverberant recordings exhibits structured sparsity determined by the image model. We exploit this structured sparsity to identify the location of the active sources and their corresponding absorption coefficients consisting of the columns of P. Given the model of the microphone recordings stated in (5), the covariance matrix of the observations is

$$\begin{aligned} C = \mathcal{X}\mathcal{X}^* &= O\Sigma O^* \\ &= \sum_{i=1}^{G} O_{.,\Omega_i} \Sigma_{\Omega_i,\Omega_i} O^*_{.,\Omega_i}, \end{aligned} \tag{6}$$

where $.^*$ denotes conjugate transpose and $\Sigma = P\mathcal{S}\mathcal{S}^*P^*$. Note that the spatio-spectral sparsity of concurrent speech sources implies that $\mathcal{S}\mathcal{S}^*$ is a diagonal matrix whose diagonal elements specify the energy of the individual sources - Section VII-A provides some empirical insights on the properties of the covariance matrix. The second equation follows because of the structure of the permutation-attenuation matrix P which indicates that Σ is supported only on the set $\bigcup_i \Omega_i \times \Omega_i$ i.e.,

$$\begin{aligned} \Sigma_{j,i} &= 0 \qquad \forall (j,i) \notin \bigcup_{i=1}^{G} \Omega_i \times \Omega_i, \\ \Sigma_{\Omega_i,\Omega_i} &= \|\mathcal{S}_{i,.}\|_2^2 P_{\Omega_i,.} P^*_{\Omega_i,.}, \end{aligned} \tag{7}$$

where $\|\mathcal{S}_{i,.}\|_2 = \sqrt[2]{\mathcal{S}_{i,.}\mathcal{S}^*_{i,.}}$. As we can see, recovering the diagonal elements of $\Sigma_{\Omega_i,\Omega_i}$ is sufficient to determine the energy of the corresponding source i and the absorption coefficients $P_{\Omega_i,.}$. We thus focus on recovering these sub-matrices for all $i \in \{1,\ldots,G\}$ from the observation covariance matrix C. Using the property of the Kronecker product, we can rewrite (6) as

$$\begin{aligned} C_{vec} &= \underbrace{\begin{bmatrix} B(1) & B(2) & \ldots & B(G) \end{bmatrix}}_{\mathcal{B}} \underbrace{\begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(G) \end{bmatrix}}_{\mathcal{V}} \\ &\forall i \in \{1\ldots,G\} : v(i) \triangleq (\Sigma_{\Omega_i,\Omega_i})_{vec}, B(i) \triangleq \overline{O}_{.,\Omega_i} \otimes O_{.,\Omega_i}. \end{aligned} \tag{8}$$

where $\otimes$ denotes the Kronecker product between two matrices and $\overline{O}_{.,\Omega_i}$ is the *element-wise* conjugate of $O_{.,\Omega_i}$. In a typical problem setup, very few microphones are used for recording, i.e. $M \ll \mathcal{G} < \sum_{i=1}^{G} |\Omega_i|$ where operator $|.|$ indicates the cardinality of the set; thus recovering $\Sigma_{\Omega_i,\Omega_i}$ requires solving an underdetermined system of linear equations and therefore, in general (6) admits infinitely many solutions and recovery is not feasible.

To circumvent the ill-posedness of the inverse problem, we exploit yet another kind of *block-sparsity structure* that is exhibited in our formulation of the reverberant multi-party recordings. The block sparsity of the actual-virtual sources implies that only $N \ll G$ groups of $v(i)$s (or correspondingly $\Sigma_{\Omega_i,\Omega_i}$) contain nonzero elements, and thus, identifying those groups equivalently determines the positions of the active sources $\mathcal{S}$.

In addition, by recovering the corresponding elements of $\mathcal{V}$ and then normalizing them by the sources energies, we can identify the absorption coefficients (i.e., the columns of P) which correspond to the attenuation for each source due to the multipath reflections.

We simplify the notation by using $\Sigma^i \triangleq \Sigma_{\Omega_i,\Omega_i} \in \mathbb{R}^{|\Omega_i|\times|\Omega_i|}$. Our block-sparse recovery approach can then be formulated by the following convex minimization problem:

$$\begin{aligned} \underset{\Sigma^1,\ldots,\Sigma^G}{\arg\min} \quad & \sum_{i=1}^{G} \left\| \Sigma^i_{vec} \right\|_{L_2} \\ \text{subject to} \quad & \|C_{vec} - \mathcal{B}\mathcal{V}\|_{L_2} \leqslant \varepsilon \\ & \left(\mathcal{V} = \left[(\Sigma^i_{vec})^T, , \ldots, (\Sigma^i_{vec})^T \right]^T \right) \\ & \Sigma^i = (\Sigma^i)^* \quad \forall i \in \{1,\ldots,G\} \\ & \Sigma^i_{l,j} \geqslant 0 \qquad \forall l,j,i \end{aligned} \tag{9}$$

We recall that minimizing the sum of the L_2 norms of a group of vectors induces the block-sparsity structure in the solution so that, only few subsets of vectors in the group (i.e. few Σ^is) contain nonzero elements. Indeed, if Σ^is have the same size (i.e. $|\Omega_1| = |\Omega_2| = \ldots = |\Omega_G|$) the objective function of (9) becomes equivalent to the L_1L_2 norm[1] of a matrix whose rows are populated by $(\Sigma^i_{vec})^T$, which as mentioned earlier is a popular convex approach for block (group) sparse approximation. We solve (9) by using the iterative proximal splitting algorithm [23].

To summarize, we obtain the location of the sources and their images. The components of $\Sigma_{\Omega_i,\Omega_i}$ normalized by the energy of the sources corresponds to the attenuation factors. We entangle the room geometry with the absorption coefficients to characterize the acoustic projections *for any order of desired reflections* R, and construct the microphone array measurement matrix Φ as described in Section IV-A. In a scenario where $N < M$, we apply inverse filtering to perform joint speech separation and deconvolution as explained in the following Section IV-C3.

3) Speech Recovery by Inverse Filtering the Acoustic Channel: The approach presented in Sections IV-C1 and IV-C2 enables us to localize the sources and model the mixing channels. In a scenario where the number of sources is less than the number of microphones (i.e., $M > N$), we can use the frequency domain deconvolution to reverse the attenuation and phase shift induced by the acoustic propagation. Given the frequency domain impulse response function $H(f, \mu_m, \nu_g)$ between microphone

[1]The $\|.\|_{L_1L_2}$ mixed-norm of a matrix is defined as the sum of the L_2 norms of its rows as defined in (16)

located at μ_m and source located at ν_g as expressed in (4), we construct

$$\mathcal{H} = \begin{bmatrix} H(f, \mu_1, \nu_1) & \dots & H(f, \mu_1, \nu_N) \\ \vdots & & \vdots \\ H(f, \mu_M, \nu_1) & \dots & H(f, \mu_M, \nu_N) \end{bmatrix}$$

The desired source is recovered by inverse filtering stated as

$$\hat{\mathcal{S}} = (\mathcal{H}^T \mathcal{H})^\dagger \mathcal{H}^T \mathcal{X} \tag{10}$$

This operation performs exact deconvolution of the signal from the early room impulse response function [14, 16]. The late reverberation can be statistically modeled as an exponentially decaying white Gaussian noise which possess the diffuse characteristics [24].

To reduce the effect of late reverberation and enhance the signal in terms of speech quality and recognition rate, we apply the Zelinski post-processing proposed in [25]. Among several post-filtering methods proposed in the literature, 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. We implemented the Zelinski post-filter for the experiments described in Section VII-D. The dereverberation of the early impulse response achieved by inverse filtering the acoustic channels enables a more efficient post-filtering as formulated in [25]. The experimental analysis are presented in Section VII-D1.

In the alternative *underdetermined scenario* where the number of sources exceeds the number of available recordings (i.e., $M < N$), solving the system of $\mathcal{X} = \Phi\mathcal{S}$, requires solving an ill-posed and degenerate system of linear equations which can take infinitely many answers, we thus exploit prior information on the sparse properties of $\mathcal{S}$ to circumvent the ill-posedness of the problem. We cast the underdetermined speech recovery problem as sparse signal reconstruction where we exploit the underlying structure of the sparse coefficients to recover the signal components more efficiently from a few number of measurements [26]. The details are elaborated in the following Section V.

V. Structured Sparse Speech Recovery

A. Computational Approaches

The objective is to estimate the structured sparse coefficient vector $\mathcal{S}$ such that $\mathcal{X} = \Phi\mathcal{S}$. This problem can be stated precisely as

$$\hat{\mathcal{S}} = \underset{\mathcal{S}\in\mathbb{M}}{\operatorname{argmin}} \|\mathcal{S}\|_0 \quad \text{s.t.} \quad \mathcal{X} = \Phi\mathcal{S} \tag{11}$$

where $\mathbb{M}$ specifies the union of all vectors with a particular support structure. The counting function $\|.\|_0 : \mathbb{C}^{GF} \rightarrow \mathbb{N}$ returns the number of non-zero components in its argument.

The major classes of computational techniques for solving the sparse approximation problem stated in (11) include greedy pursuit, convex relaxation, non-convex optimization, and Bayesian algorithms. This paper considers greedy algorithms and convex optimization, which offer provable correct solutions under well-defined conditions [27]. The greedy pursuit method iteratively refines the current estimate for the coefficient vector $\mathcal{S}$ by modifying one or several coefficients chosen to yield a substantial improvement in quality of the estimated signal. The Convex optimization approach solves a convex relaxation of (11) by replacing the counting function with a sparsity inducing norm.

B. Structured Sparsity models

We focus on two types of structures underlying the sparse coefficients: *block-dependency* and *harmonicity*.

- The block-dependency model is exhibited if some interconnections between the adjacent frequencies exist. In case of the vector $\mathcal{S}$, it indicates that the spatial sparsity structure is the same at all neighboring discrete frequencies. In other words, a block of b consecutive frequencies corresponds to the same cell so the signal of the individual sources is recovered with a structure of independent blocks defined as

$$\mathcal{F}_B = \{[f_1, ..., f_b], [f_{b+1}, ..., f_{2b}], [f_{F-b+1}, ..., f_F]\} \tag{12}$$

- The harmonic-dependency model is exhibited if there are some interconnections between frequencies which are the harmonics of a fundamental frequency. In voiced speech, most of the signal energy occurs at harmonics of a fundamental frequency. The *harmonic sparsity structure* captures this model: it indicates that at any cell of the grid, energy is present in all frequencies that can be expressed as harmonics of a fundamental frequency. To state it more precisely, the support of vector $\mathcal{S}$ has the

following $\mathcal{F}_{\mathrm{H}}$ structure defined as

$$\mathcal{F}_{\mathrm{H}} = \{\mathsf{k}\mathsf{f}_0 | 1 < \mathsf{k} < \mathsf{K}\}, \tag{13}$$

where f_0 is the fundamental frequency and K is the number of harmonics.

C. Model-based Sparse Recovery

Sparse recovery methods have been proposed to incorporate the underlying structure of the sparse coefficients in recovering the unknown sparse vector. We use model-based sparse recovery algorithms explained as follows:

- *IHT*: Iterative hard thresholding (IHT) offers a simple yet effective approach to estimate the sparse vectors. It seeks an N-sparse representation $\hat{\mathcal{S}}$ of the observation $\mathcal{X}$ iteratively to minimize the residual error. We use the algorithm proposed in [28] which is an accelerated scheme for hard thresholding methods with the following recursion

$$\begin{cases} \hat{\mathcal{S}}_0 = 0 \\ \mathsf{r}_\mathsf{i} = \mathcal{X} - \Phi\hat{\mathcal{S}}_\mathsf{i} \\ \hat{\mathcal{S}}_{\mathsf{i}+1} = \mathcal{M}^{\mathcal{F}_\cdot}\left(\hat{\mathcal{S}}_\mathsf{i} + \kappa\Phi^\mathsf{T}\mathsf{r}_\mathsf{i}\right) \end{cases} \tag{14}$$

  The step-size κ is the Lipschitz gradient constant to guarantee the fastest convergence speed [28]. To incorporate for the underlying structure of the sparse coefficients, the model approximation $\mathcal{M}^{\mathcal{F}_\cdot}$ is defined as reweighting and thresholding the energy of the components of $\hat{\mathcal{S}}$ with either $\mathcal{F}_{\mathrm{B}}$ or $\mathcal{F}_{\mathrm{H}}$ structures.

- *OMP*: The Orthogonal Matching Pursuit (OMP) is a greedy pursuit algorithm which iteratively refines a sparse solution by successively identifying one or more components that yield the greatest improvement in quality. To describe our model-based OMP in mathematical formulation, we consider an index set Λ which selects a subset of columns from Φ. Denoting the set difference operator as $\setminus$, the columns of $\Phi_{\setminus\Lambda}$ corresponding to either $\mathcal{F}_{\mathrm{B}}$ or $\mathcal{F}_{\mathrm{H}}$ structures are searched per iteration and Λ is expanded so as the mean-squared error of the signal approximation is minimized through the left pseudo-inverse operation denoted by $\Phi_\cdot^\dagger$ [27, 29]. The signal estimation algorithm would thus

have the following recursion

$$\begin{cases} \Lambda_0^{\mathcal{F}_\cdot} = 0 \\ \lambda_i = \underset{\lambda \in \Phi_{\setminus \Lambda_{i-1}^{\mathcal{F}_\cdot}}}{\operatorname{argmin}} \ \|\mathcal{X} - \Phi_{\Lambda_{i-1}^{\mathcal{F}_\cdot} \cup \lambda} \Phi^{\dagger}_{\Lambda_{i-1}^{\mathcal{F}_\cdot} \cup \lambda} \mathcal{X}\|_2 \\ \Lambda_i^{\mathcal{F}_\cdot} = \Lambda_{i-1}^{\mathcal{F}_\cdot} \cup \lambda_i \\ \hat{\mathcal{S}}_i = \Phi^{\dagger}_{\Lambda_i} \mathcal{X} \end{cases} \tag{15}$$

- L_1L_2: Another fundamental approach to sparse approximation replaces the combinatorial counting function in the mathematical formulation stated in (11) with the L_1 norm, yielding convex optimization problems that admit a tractable algorithm referred to as basis pursuit [30]. We use a group version of basis pursuit algorithm with the number of group components $n^{\mathcal{F}_\cdot}$ determined by each structure $\mathcal{F}_\cdot$ referring to either $\mathcal{F}_B$ or $\mathcal{F}_H$. The optimization problem to recover the structured sparse coefficients $\hat{\mathcal{S}}$ is formulated as follows

$$\begin{aligned} \hat{\mathcal{S}} &= \arg\min \|\mathcal{S}\|_{L_1,L_2} \quad \text{s.t.} \quad \mathcal{X} = \Phi\mathcal{S}, \\ \|\mathcal{S}\|_{L_1,L_2} &= \left(\sum_{g=1}^{G} \left[\sum_{b=1}^{n^{\mathcal{F}_\cdot}} \mathcal{S}_g^2(b) \right]^{1/2} \right) \end{aligned} \tag{16}$$

The speech recovery approach as described in this section, requires characterization of the acoustic measurements and the performance bound is entangled with the properties of the microphone array manifold matrix.

VI. Performance Bound

The approach that we have taken in this paper to address the reverberant speech separation as studied throughout Sections III-IV, relies on casting the problem as reconstructing the high-dimensional spatio-spectral information embedded in the acoustic scene from a compressive acquisition provided by the array of microphones. We leveraged model-based sparse recovery framework for characterization of the compressive acoustic measurements and recovering the speech components. In this framework, the theoretical analysis of the performance bounds is entangled with the performance of the sparse recovery algorithms [27]. We adopt the notion that ϕ^j represents the j^{th} column of Φ. A key property to guarantee the theoretical performance bound is the coherence of the measurement matrix defined as

$$\gamma(\Phi) = \max_{1 \leqslant j,k \leqslant G, j \neq k} \frac{|\langle \phi^j, \phi^k \rangle|}{\|\phi^j\| \|\phi^k\|} \tag{17}$$

The coherence quantifies the smallest angle between any pairs of the columns of Φ. The number of recoverable non-zero coefficients (N) using either convexified or greedy sparse recovery is inversely proportional to γ as $\mathrm{N} < \frac{1}{2}(\gamma^{-1} + 1)$ [27]. Hence, to guarantee the performance of sparse recovery algorithms, it is desired that the coherence is minimized. As the measurement matrix is constructed of the location-dependent projections, this property implies that the contribution of the source to the array's response is small outside the corresponding sensor location or equivalently the resolution of the array is maximized. It has been shown in [31] that the free-space Green's function constituted projections given that the inter-element spacing is large enough, exhibits an optimal design, and the columns of the measurement matrix corresponds to a sampled Fourier basis function. It has been further pointed out that a large-aperture random design of sensor array yields the projections to be mutually incoherent [31]. Thereby, the projections are spread across all the acoustic scene and each sensor captures the information about all components of $\mathcal{S}$. These studies elucidate that the performance of our sparse approximation framework is entangled with the microphone array construction design. This issue is investigated in Section VII.

VII. Experimental Analysis

A. Orthogonality of Spectrographic Speech

We carried out experiments to investigate the orthogonality of multiple speech sources in the frequency domain. In this experiment, five speech signals are obtained by random concatenation of 100 utterances from Wall Street Journal speech corpus and they are normalized prior to analysis. The length of each speech signal is 2 min and the signals are analyzed in frames of size 256ms (fft-size = 2048) with 50% overlap; thus we obtain five 1024×900 matrices corresponding to the STFT of each source. The orthogonality is measured for each frequency band independently. We construct the matrix $\mathcal{S}_{5\times 900}$ where each row corresponds to each source and has the frequency components of a particular band along 900 frames. In case of perfectly orthogonal sources, $\mathcal{C} = \mathcal{S}\mathcal{S}^*$ is diagonal and the energy of the diagonal elements of the matrix is equal to the matrix Frobenius norm. Fig. 2-right-hand-side illustrates the diagonal-L_2-norm/matrix-Frobenius-norm.

In addition, we compute a pointwise multiplication of the STFTs of two utterances and plot the histograms of the resulted values. Fig. 2-left-hand-side illustrates the obtained histogram. We can see the distribution mass of the energy of the point-wise multiplication values is localized around 0. This phenomenon indicates that the majority of the high energy components in the spectro-temporal domain

are non-overlapping or disjoint. The orthogonality of spectrographic speech is exploited in our acoustic modeling approach explained in Section IV-C.

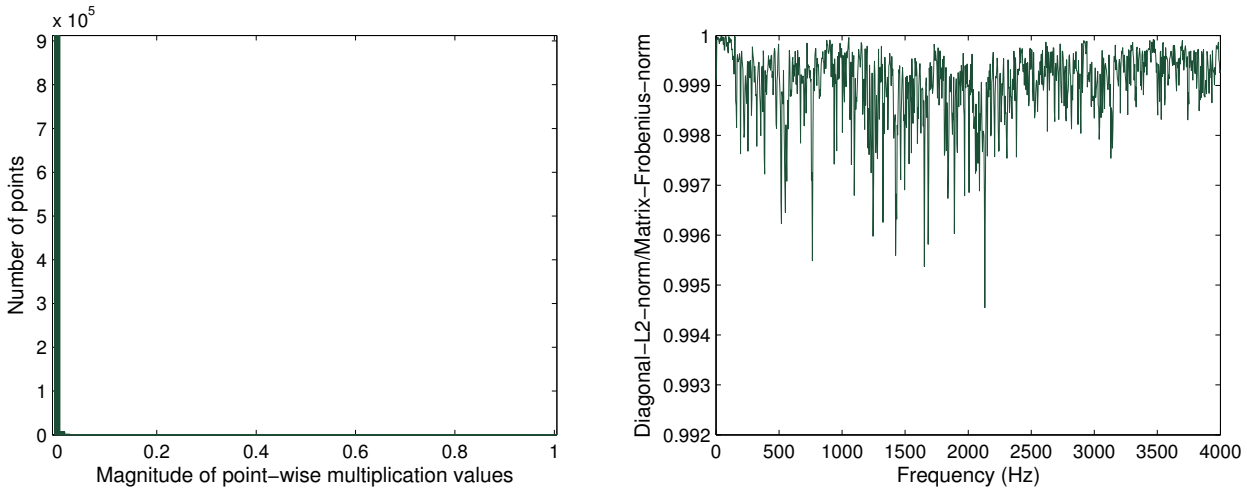

Fig. 2: Orthogonality of multiple speech utterances in spectro-temporal domain: Left-hand-side illustrates the energy histogram of the component-wise multiplication of speech utterances and Right-hand-side illustrates the diagonal-L_2-norm/matrix-Frobenius-norm of the covariance matrix constructed per frequency.

B. Data Recordings Set-up

Experiments were performed in the framework of the Multichannel Overlapping Numbers Corpus (MONC). This database was collected by outputting 10136 utterances from Numbers Corpus release 1.0 (telephone quality speech, 30-word vocabulary), prepared by the Center for Spoken Language Understanding at the Oregon Graduate Institute on one or more loudspeakers, and recording the resulting sound field using a microphone array and various lapel microphones [32]. The recordings were made in a 8.2 m $\times$ 3.6 m $\times$ 2.4 m rectangular room containing a centrally located 4.8 m $\times$ 1.2 m rectangular table. The positioning of loudspeakers was designed to simulate the presence of 3 competing speakers seated around a circular meeting room table of diameter 1.2 m. The loudspeakers were placed at 90° spacings at an elevation of 35 cm (distance from table surface to center of main speaker element). An eight-element, 20 cm diameter, circular microphone array placed in the center of the table recorded the mixtures. The recording scenario is illustrated in Fig. 3. The average signal to noise ratio (SNR) of the recordings is 10 dB.

This database is collected to evaluate distant speech recognition performance in overlapping condition. 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 3050 training utterances,

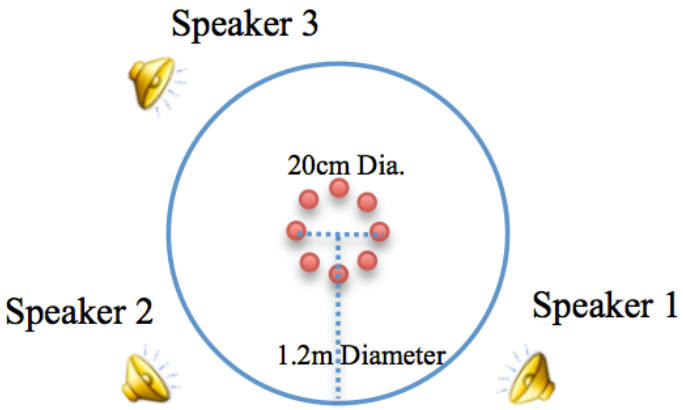

Fig. 3: Microphones and speakers placement

1018 development set and 1044 test files. Separation of utterances into train, devtest and test sets was done according to the Numbers release 1.0 documentation (i.e. based on utterance ID number modulo 5). Generated *competing speaker* utterance ID lists was performed by 500-utterance circular shift of the ordering. The word loop grammar is used and the task is speaker independent.

C. Reverberant Acoustic Modeling

1) Room Geometry Estimation: The first step to characterize the room acoustics is to estimate the room geometry. We accomplish this step through localization of the images of multiple sources in a large extended area using the sparse recovery framework with a free space model. The location of the source images corresponds to the temporal support of the room impulse response function. The energies of the recovered signals are sorted and truncated to the order of $\mathrm{D}(\mathrm{D}+1)/2$ to include the early reflections of the walls and guarantee the uniqueness of the solution. The estimated support of the room impulse response function is then used for estimation of the room rectangular geometry by generating the room impulse responses for various room dimensions and identify the best fit to the estimated support in least-squares sense. The brute-force finding of the room geometry has a computational cost depending on the number of dimensions of the spatial search space. We can employ some heuristic approaches and start from an initial guess about the boundaries as the half-way wall between the source and its earliest images. The estimates are then refined around the initial state through least square regression of all virtual sources. The method is implemented to find the location of the surrounding walls excluding the floor and the ceiling. There is no algorithmic impediment to perform room modeling using a volumetric grid, although more number of microphones is required. For the purpose of the experiments presented in this paper,

we assume the heights to be known (to enable three-dimensional acoustic modeling) for reducing the dimensionality and computational cost and enabling an exhaustive experimental analysis.

The planar area of the room is divided into square cells with 25 cm spacing. The maximum distance from the center of the array to identify the actual sources is 1 m; therefore, if a source is localized at a distance greater than 1 m, it is considered as a virtual source or source image. To achieve a better estimation, we restrict our discretized grid to the orthogonal subspaces corresponding to the orthogonal walls. We could estimate the geometry of the room up to 50 cm error per dimension (i.e., 25 cm per wall) from the recordings of three sources in a close proximity to the microphone array as depicted in Fig. 3. The experiments are run on MATLAB 7.14 on 4 Core(TM) i7 CPU @ 2.8-GHz, 11.8-GiB RAM PC; the required absolute elapsed time to perform geometry estimation by searching 1.5 m around the initial guess for estimating the two- and three-dimensional geometry were 0.91 and 11 seconds respectively; it shows a linear growth proportional to the number of augmented search levels. Once the geometry is estimated, the whole session is recorded in one place.

2) Absorption Coefficients Estimation: The initial evaluations are conducted on synthesized recordings to enable quantification of the performance bound in a controlled set-up. We consider the following scenarios: (1) 8-channel circular microphone array positioned in the middle of the room, (2) 12-channel microphone array: two sets of 6-channel circular arrays, each located 1 m far apart with respect to the center of the room, (3) 16-channel microphone array: two sets of 8-channel circular arrays, each located 1 m far apart with respect to the center of the room. We considered about 3 cm displacement of the microphones with respect to the Euclidean coordinates used for computing the microphone array manifold matrix. The reverberant channel is simulated using the code available in [33] for a four-sided $3 \times 4\,\mathrm{m}^2$ enclosure. The area of the room is discretized into a grid of uniform cells of size $0.5 \times 0.5\,\mathrm{m}^2$ adding up to 40 cells inside the room. The reflection coefficients of the walls are selected as 0.4, 0.6, 0.8 and 0.9. Evaluations are carried out using $\mathrm{N} = \{1, 2, 3\}$ omni-directional sources distributed arbitrarily in the room with the following characteristics (a) Spectrum of orthogonal random broad-band sources at 52 auditory-centered frequencies and (b) Spectrum of independent speech sources at the frequency-bands which contain 80% of the total energy. Fig. 4 demonstrates the estimated room acoustic impulse response from recordings of two concurrent speech sources recorded by 8-channel microphone array using our structured sparse acoustic modeling technique explained in Section IV. Alternatively, the blind channel impulse response estimation referred to as the Cross-Relation technique [34] is used to recover the channel from recording of a single source; the normalized distance quantified as $\|\mathrm{H} - \hat{\mathrm{H}}\|_2 / \|\mathrm{H}\|_2$

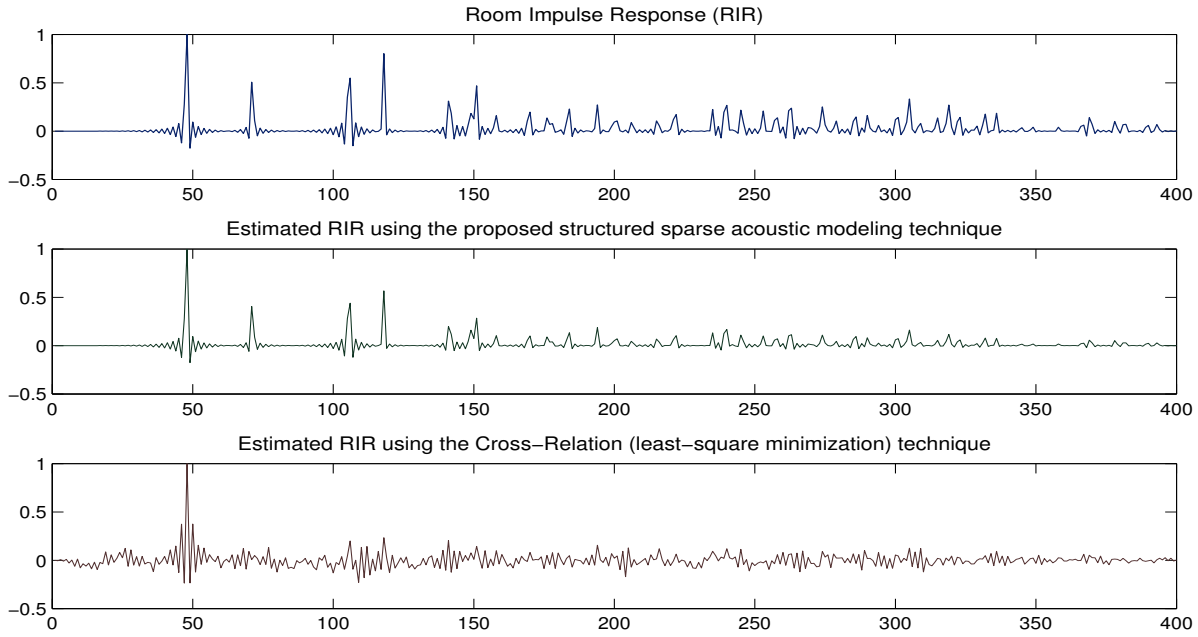

Fig. 4: (top) Example of a simulated room impulse response (RIR) [33], (middle) Estimated RIR using the proposed structured sparse acoustic modeling technique, and (bottom) Estimated RIR using the Cross-Relation (least-squared optimization) technique [34]. The normalized distances between the actual RIR and estimated RIR using structured sparse recovery and least-squared optimization are 0.33 and 0.92 respectively.

is calculated as 0.33 and 0.92 respectively. To our knowledge, the state-of-the-art techniques can not recover the acoustic channel from recordings of multiple unknown speech sources.

The results of source localization (SL), absorption coefficients estimation (AC) and signal recovery (SR) are illustrated in Figs. 5 (orthogonal sources) and Fig. 6 (speech sources). The results of Fig. 5 demonstrates the performance bound of the algorithm presented in Section IV-C. We can see that in noiseless condition, SL is achieved almost 100% correct per frequency band for any number of (one to three) sources. However, estimates of the absorption coefficients are not exact; the root mean square error (RMSE) is proportional to the number of microphones used to collect the data. The best estimate is achieved when 16 microphones are used; increasing the number of concurrent sources results in about 5% error increase in estimation of AC. Similarly, estimations of the source coefficients (SR) is obtained up to 4% error if there is only one source active. Increasing the number of sources reduces the accuracy about 5% per added source. Contrasting these results with the bar charts obtained for speech sources does not show any degradation in 16-microphones scenario. In more under-sampled regimes, the degradation is less than 5% in SL and upper bounded by 10% in AC and SR.

If we consider adding white Gaussian noise to the recorded signals, the errors in AC estimation and

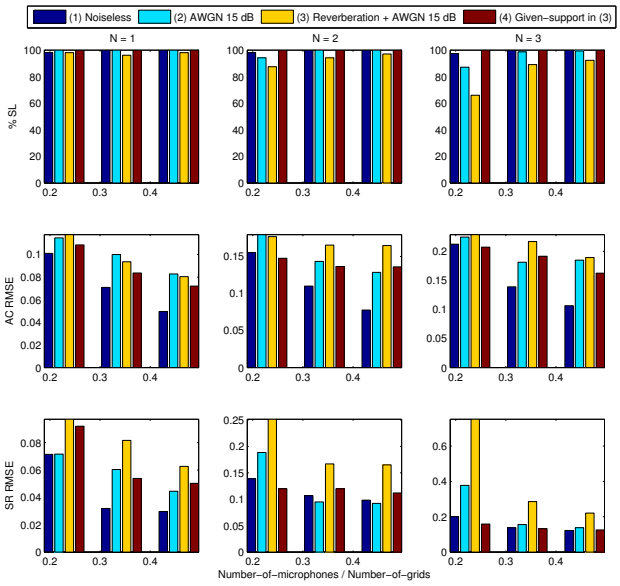

Fig. 5: Performance of the algorithm in terms of Source Localization (SL), Root Mean Squared Error (RMSE) of Absorption Coefficients (AC) estimation as well as Signal Recovery (SR). The test data are random *orthogonal sources* and the measurement matrix is consisted of free-space Green's function. The SNR of noisy condition is 15 dB.

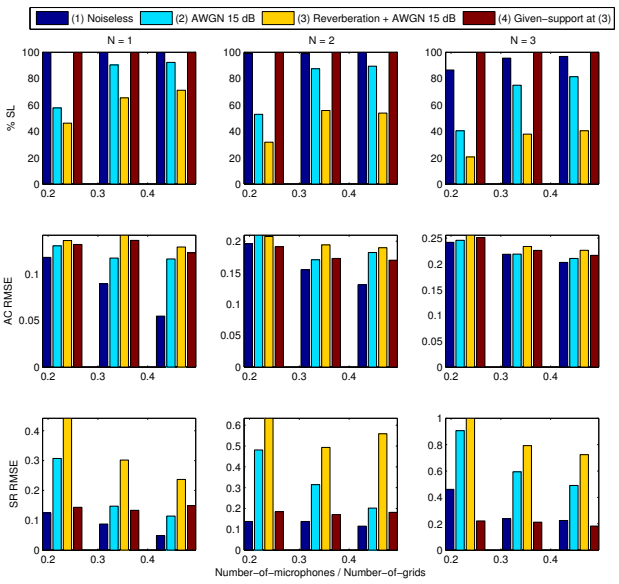

Fig. 6: Performance of the algorithm in terms of Source Localization (SL), Root Mean Squared Error (RMSE) of Absorption Coefficients (AC) estimation as well as Signal Recovery (SR). The test data are random *speech sources* and the measurement matrix is consisted of the free-space Green's function. The SNR of noisy condition is 15 dB.

SR are increased up to 8% and 50% respectively. In a similar way, considering the effect of additive noise and reverberation mismatch (obtained by adding noise to the AC coefficients), the distortion of AC estimates is bounded by the noise level whereas the recovered source coefficients (SR) are highly degraded. Contrasting these results with the speech bar charts demonstrates up to 40% SR distortion using only 8 microphones whereas AC estimation is achieved more accurately and degraded less than 5% using the approximately orthogonal speech sources; the average error of AC estimation is expected around 10-20% in noisy and reverberant condition. These results show a good robustness with increasing the number of concurrent sources. In addition, we observe a noticeable reduction in the performance of support recovery or localization (SL) of speech sources per frequencies; this effect could be justified as the spectrographic speech is approximately sparse and many of the components have a small energy which are drawn in noise. Hence, exploiting model-based sparse recovery or considering the broadband speech spectrum is crucial to achieve a reasonable localization performance [35].

Given that support recovery (i.e., SL) is obtained 100% correct by considering the broadband spectrum of speech signal and assuming that the sources are immobile, we can use the identified support for AC

estimation and speech recovery. The resulted accuracy is upper bounded by noise level and in particular it enables a great improvement in SR. Hence, we carried out the AC estimation experiments on real data recordings where the support of the sparse coefficients (i.e. location of the active sources) is estimated from the first initial (< 5) frames and absorption coefficients are recovered given the support. If the number of microphones is more than the number of sources, then support estimation (source localization) enables very accurate results for estimation of the absorption coefficients [35]. Similarly for speech recovery, we can perform inverse filtering to separate the individual sources. This scenario is investigated in Section VII-D1. We computed the average time per frequency for the absorption coefficient estimation of six-sided walls using 8-channel microphone array as 17.16 seconds. This computational cost grows linearly with the dimension of the sparse vector and the number of microphones. Estimating the support from the first initial frames, enables estimation of the coefficients by pseudo-inversion which decreases the computational cost to a fraction of a second.

The scenario of the real data evaluations is explained in Section VII-B which is similar to the first set-up described above. The location of the desired source is fixed through out the whole session (i.e. stationary condition). The estimated absorption coefficients are plotted using the data in the following conditions: (I) single speech utterances, (II) Two simultaneous speech utterances, (III) Three simultaneous speech utterances. The estimates are run over 9000 speech files of MONC corpus [32]; the absorption coefficients are computed and averaged for each frequency-band independently. The estimated frequency-dependent absorption coefficients (computed at a resolution of 4 Hz) are illustrated in Fig. 7. To enable estimation of the three-dimensional acoustic parameters, we considered two parallel grids at given heights corresponding to the first order reflection of the table and ceiling; the reflections of the carpet floor are trapped under the table hence, the meeting table was considered as the floor in our image model [36]. Thereby, the algorithm explained in Section IV-C is run for a six-sided enclosure. To be more illustrative, the absorption coefficients are depicted for four surrounding walls, although we performed three-dimensional acoustic modeling. The absorption coefficients are estimated independently per frame hence, our method is applicable to the dynamic scenarios where the speaker changes the position at a rate slower than the frame-size.

There is no ground truth of the actual acoustic parameters available. The plots show a consistent estimation using recordings of one, two and three concurrent sources. The database is noisy (SNR$\approx$10 dB); the synthetic data evaluations reported in Fig. 6 show an expected 10-20% error in absorption coefficient estimation in noisy condition. Similar uncertainty of the coefficients is observed on real data recordings. Nevertheless, we use an average estimate of acoustic parameters for speech recovery tests conducted in

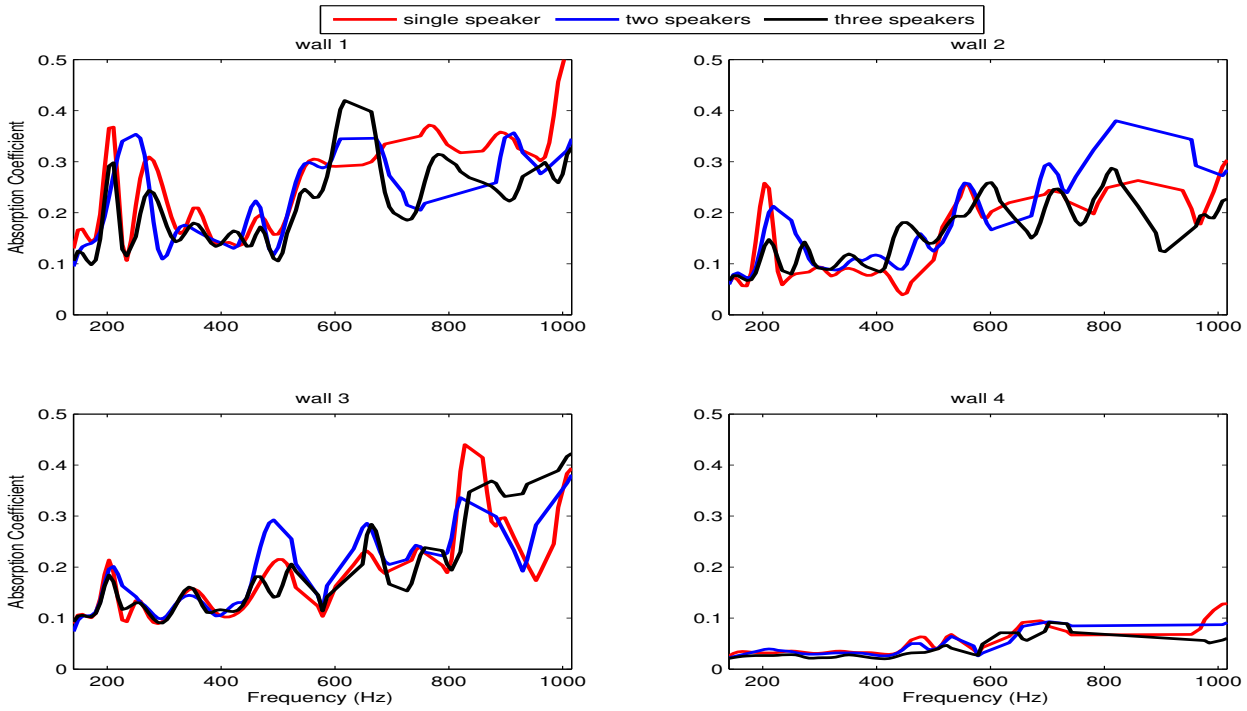

Fig. 7: Frequency-dependent absorption coefficients computed for each wall from the utterances of 3 competing speakers for the third speaker.

Section VII-D1.

D. Reverberant Speech Separation

1) Overdetermined Scenario: Given the location of the sources and the characterized room acoustic channel obtained from the formulation stated in Section IV, we recovered the desired signal by inverse filtering and perform speech recognition. We used overlap-add (OLA) to reconstruct a time domain signal after speech separation. The signal is again transformed to Fourier domain using a short window size appropriate for speech feature extraction (e.g. MFCC). The OLA can be considered as a convenient mean of changing the DFT size and period.

The automatic speech recognition (ASR) scenario was designed to broadly mirror that of Moore and McCowan [37]. A typical front-end was constructed using the HTK toolkit [38] 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 HMMs with 3 emitting states per triphone and 12 mixtures per state. The ASR accuracy on

the clean speech data is about 95%. We performed MAP adaptation by applying each technique on the training data for the corresponding experiments. The Zelinski post-filtering is applied on the separated speech prior to the recognition [25].

In addition to the speech recognition, we evaluated the quality of the recovered speech using signal to interference ration (SIR) [39] as well as perceptual evaluation of speech quality (PESQ) [40]. As our approach relies on the principles of spatial diversity, we compare it with beamforming which possess similar essence. We used the super-resolution speaker localization based on sparse recovery to perform near-field beamforming. The resulting speech recovery performance is summarized in Table I.

TABLE I: Quality evaluation of the recovered speech in terms of Source to Interference Ratio (SIR), Perceptual Evaluation of Speech Quality (PESQ) and Word Recognition Rate (WRR) using near-field Super Directive (SD) beamforming before and after applying post-filtering (PF), vs. inverse filtering of Room Acoustic Model (RAM)

N	Meas.	Baseline	Lapel	SD	SD-PF	RAM	RAM-PF
1	SIR	12.3	19.19	18.5	18.52	16.7	16.1
	PESQ	2.7	3	3.3	3.3	2.92	2.97
	WRR%	89.61	93.21	95	95	93.9	93.3
2	SIR	2.6	18.29	11.8	11.33	13	17.5
	PESQ	2	2.35	2.7	2.69	2.65	2.8
	WRR%	55.19	74.53	70.19	68.16	83.8	87.93
3	SIR	-0.7	18.35	10.2	10	10.1	14.2
	PESQ	1.6	2.27	2.48	2.48	2.4	2.62
	WRR%	39.92	68.13	63	61.45	70.88	79.21

As the results indicate, speech separation and deconvolution obtained by inverse filtering of the room acoustic channels followed by post-filtering (RAM-PF) yields the maximum interference suppression and highest perceptual quality of the recovered speech in multi-party scenarios as quantified in terms of SIR and PESQ. It also outperforms other techniques in terms of word recognition rate. The Zelinski post-processing is derived to reduce the effect of uncorrelated noise. We can observe that the improvement in performance obtained after deconvolution of the room acoustic channel is higher than what we can achieve after standard beamforming.

2) Underdetermined Scenario: To consider the generalized scenario of underdetermined mixtures, we incorporate the room acoustic model for structured sparse speech recovery explained in Section V. The scenario similar to Fig. 3 is synthesized using five uniformly situated sources and a circular array with 4 elements is used for recordings. Alternative to the uniform compact array, a *random large array* is simulated where the distances of the four microphones to the array center is multiplied by $2, 3, 4, 5$ respectively. The sampling frequency is 8 kHz. The recording condition is clean; the goal of

this experiment is to evaluate the underdetermined speech separation performance comparing various computational strategies and speech-specific models in different microphone array topologies.

The spectro-temporal representation is obtained by windowing the signal in 256 ms frames using a Hann function with 50% overlapping. The length of the speech signal is 15 s. The speech separation experiments are performed using different sparse recovery approaches to incorporate the block dependency as well as harmonicity of the spectro-temporal coefficients of speech signal. The quality evaluation results in terms of SIR [39] and PESQ [40] are summarized in Fig. 8. The block-size $\mathtt{b}$ was set to 4 as it was shown to yield the best results, especially for B-OMP and B-$\mathrm{L_1L_2}$. The average time for recovery of a 256 ms speech frame using $\mathrm{L_1L_2}$ [30], IHT [28] and OMP [41] were 148.29, 4.25 and 0.973 seconds respectively.

In the harmonic model, we consider that $\mathrm{f_0} \in [150-400]$ Hz. Those frequencies that are not harmonics of $\mathrm{f_0}$ are recovered independently in H-IHT and H-$\mathrm{L_1L_2}$. We also considered that the harmonic structures are non-overlapping and $\mathtt{k}$ spans the full frequency band. The harmonic sparse recovery approach does not require estimation of $\mathrm{f_0}$. We start from $\mathrm{f_0} = 50$ and consider all of its harmonics within the frequency band (i.e., $\mathrm{f} \leqslant 4000$); hence, a block of size $\mathtt{K} = 80$ of harmonics of $\mathrm{f_0} = 50$ are recovered jointly. Then we move to $\mathrm{f_0} = 51$ and proceed up to $\mathrm{f_0} = 400$. Therefore, the size of the blocks are variable. To prevent overlapping, the priority is given to the first seen frequency components. In other words, if a particular frequency is first included in the harmonics of $\mathrm{f_0} = 50$, it is excluded from the harmonics of $\mathrm{f_0} = 100$. The remaining frequency components are recovered independently. For H-OMP, the harmonic subspaces are used to select the bases while projection is performed for the full frequency band. This procedure is applied on all of the frames regardless of the voiced/unvoiced characteristics. Therefore, we expect the model to be more effective if the ratio of the voiced segments is greater than the unvoiced segments; a combination of *block* and *harmonic* model could be considered for effective model-based speech recovery.

We observe that the highest quality in terms of SIR and PESQ are obtained by convex optimization. This could be due to the zero-forcing spirit of greedy approaches. This deficiency is particularly exhibited for speech-like signals, which do not possess high compressibility [42, 43]. However, in some applications such as speech recognition, where the reconstruction of the signal is not required, we can exploit the sparsity of the information bearing components in greedy sparse recovery approaches, which offer a noticeable computational speed in efficient implementations and a reasonable performance [26]. Comparing the results of ad-hoc microphones with the conventional compact topology suggests that uniform compact microphone array is not an optimal design from sparse recovery perspective and using the recordings of an ad-hoc large microphone array yields better performance.

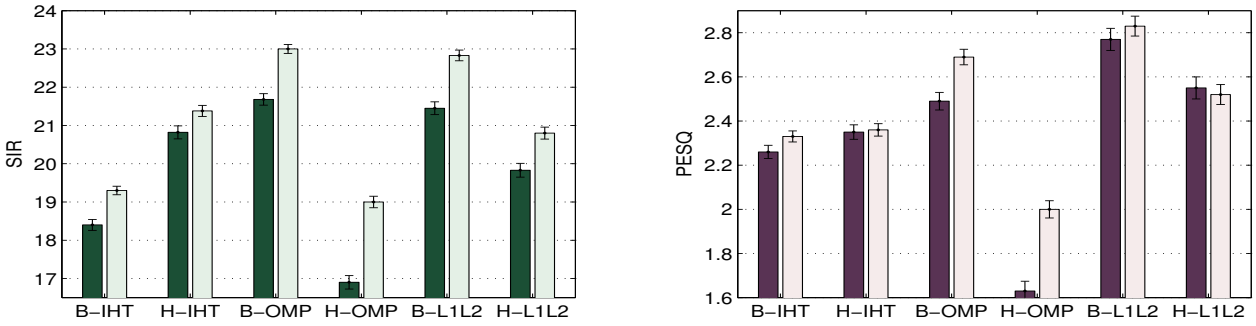

Fig. 8: Quality evaluation of the separated speech using different sparse recovery approaches in terms of SIR and PESQ. The baseline measures are -3.68 and 1.44 respectively. The errorbars depict the 90% confidence interval. The dark bars correspond to the compact uniform array whereas the light bars correspond to the ad-hoc large microphone array topology.

VIII. Conclusions

In this paper, we proposed a new convolutive speech separation framework that exploits spatio-spectral structures in reverberant recordings. This framework exploits structured sparsity models to characterize acoustic measurements obtained from an array of microphones, and to recover the individual speech sources. We estimated the acoustic response of the recording enclosure using the image model through a two-step procedure: first, estimating the room geometry and second, estimating the absorption coefficients.

For simple rectangular rooms, the room geometry was estimated by localizing virtual sources associated with discrete reflections of the original signal, followed by low-rank clustering of the subspaces resulting of each actual source. Location of the virtual sources corresponds to the temporal support of the room impulse response; these were used to estimate the geometry of the room via least square regression. The absorption coefficients associated with the reflective surfaces were then estimated via structured sparse recovery of a factorized formulation of multipath propagation model.

Given the so inferred model of the reverberant room response, we characterized microphone array recordings as compressive measurements of sound sources, and cast multiparty speech recovery as a structured sparse reconstruction problem where we exploited the block dependency, as well as harmonicity of the spectral coefficients, to recover the speech signals. Recovery may be performed through either convex optimization or greedy approaches. The results indicate that recovery through convex optimization yields the best speech quality quantified in terms of PESQ.

Interference suppression is also well achieved via greedy sparse recovery using the orthogonal matching pursuit algorithm. Hence, for applications such as speech recognition, where reconstruction error is not the objective, we can effectively employ the greedy strategies to recover the salient information bearing

components. Furthermore, we showed that in a (over)determined setup, we can achieve separation and deconvolution through inverse filtering of the acoustic channels.

We generalized our approach to large-aperture ad-hoc microphone arrays, and showed that the speech recovery performance obtained with such arrays is significantly superior to that obtained with compact arrays with uniform spacing of microphones. Hence, the compact uniform array set-up is not an optimal design for a sparse reconstruction framework and the present study motivates more investigation on sparse and ad-hoc microphone array layouts.

The success of our structured sparse recovery framework motivates incorporating other parametric models such as auto-regressive dependencies of the spectral coefficients [44, 45] or other forms of statistical dependencies [46] for speech-specific applications. Furthermore, we can extend our acoustic modeling formulation exploiting the low-rank structure of the problem induced by the similarity of signals attributed to the source and its images [47].

ACKNOWLEDGMENT

The authors would like to thank Prof. Bhiksha Raj from Machine Learning for Signal Processing (MLSP) group at Carnegie Mellon University for the valuable comments and fruitful discussions which resulted in important improvements. The research leading to these results has received funding from the European Union under the Marie-Curie Training project SCALE (Speech Communication with Adaptive LEarning), FP7 grant agreement number 213850. Volkan Cevher acknowledges Rice University for his Faculty Fellowship, MIRG-268398, ERC Future Proof, DARPA KeCoM program #11-DARPA-1055 and SNF 200021-132548. We also would like to acknowledge the anonymous reviewers for the insightful and precise comments and remarks to improve the quality and clarity of the manuscript.

REFERENCES

[1] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, and G. Nolte, "The 2011 signal separation evaluation campaign (sisec2011): Audio source separation," *Latent Variable Analysis and Signal Separation, Lecture Notes in Computer Science,*, vol. 7191, 2011.

[2] E. Shriberg, A. S. A., and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proceedings of Eurospeech*, 2001.

[3] A. Ozerov, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20(5), 2012.

[4] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

[5] 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.

[6] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation," in *Proceedings of ICASSP*, 2007.

[7] M. Wolfel and J. McDonough, "Distant speech recognition," *New York: John Wiley & Sons*, 2009.

[8] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *IEEE Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011.

[9] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13(4), 2001.

[10] R. Gribonval and S. Lesage, "A survey of sparse component analysis for blind source separation: Principles, perspectives, and new challenges," in *ESANN, 14th European Symposium on Artificial Neural Networks*, 2006.

[11] R. Saab, O. Yilmaz, M. J. Mckeown, and R. Abugharbieh, "Underdetermined anechoic blind source separation via ℓ_q-basis-pursuit with q¡1," *IEEE Transactions on Signal Processing*, 2007.

[12] 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.

[13] F. Nesta and M. Omologo, *Convolutive 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.

[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] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 36(2), 1988.

[17] R. Rotili, C. D. Simone, A. Perelli, A. Cifani, and S. Squartini, "Joint multichannel blind speech separation and dereverberation: A real-time algorithmic implementation," in *Proceedings of 6th International Conference on Intelligent Computing*, 2010.

[18] 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.

[19] 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.

[20] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustical Society of America*, vol. 60(s1), 1979.

[21] I. Dokmanic, Y. Lu, and M. Vetterli, "Can one hear the shape of a room: The 2-D polygonal case," in *Proceedings of ICASSP*, 2011.

[22] A. Asaei, M. Davies, H. Bourlard, and V. Cevher, "Computational methods for structured sparse recovery of convolutive speech mixtures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[23] P. L. Combettesand and J. C. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer-Verlag*, vol. 49, 2011.

[24] E. A. Habets, *Speech Dereverberation Using Statistical Reverberation Models*. Speech Dereverberation, Springer, 2010.

[25] I. A. McCowan and H. Bourlard, "Microphone array post-filter based n noise field coherence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 11(6), 2003.

[26] A. Asaei, H. Bourlard, and V. Cevher, "Model-based compressive sensing for multi-party distant speech recognition," in *Proceedings of ICASSP*, 2011.

[27] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems,," *Proceedings of the*

IEEE, vol. 98, 2010.

[28] A. Kyrillidis and V. Cevher, "Recipes on hard thresholding methods," in *Proceedings of CAMSAP*, 2011.

[29] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Transactions on Signal Processing*, vol. 51, pp. 101–111, 2003.

[30] E. V. D. Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions,," *SIAM Journal on Scientific Computing*, 2008, http://www.cs.ubc.ca/labs/scl/spgl1.

[31] L. Carin, D. Liu, and B. Guo, "Coherence, compressive sensing and random sensor arrays," *IEEE Antennas and Propagation Magazine*, 2011.

[32] I. A. Mccowan, "The Multichannel Overlapping Numbers Corpus," Idiap resources available online: http://www.cslu.ogi.edu/corpora/monc.pdf, 2003.

[33] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007. [Online]. Available: http://alexandria.tue.nl/extra2/200710970.pdf

[34] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, 1995.

[35] A. Asaei, "Model-based sparse component analysis for multiparty distant speech recognition," Ph.D. dissertation, Ecole Polytechnique Federal de Lausanne (EPFL), 2013.

[36] D. Ba, F. Ribeiro, C. Zhang, and D. Florencio, "L1 regularized room modeling with compact microphone arrays," in *Proceedings of ICASSP*, 2010.

[37] D. C. Moore and I. A. Mccowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings," in *Proceedings of ICASSP*, 2003.

[38] 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.

[39] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation (code available at http://www.irisa.fr/metiss/sassec07/?show=results)," *IEEE transactions on audio, speech, and language processing*, vol. 14, 2006.

[40] 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," *ITU-T Rec. p. 862*, 2001.

[41] S. Becker, "Implementation of CoSaMP and OMP for sparse recovery," in *MATLAB Central*, 2011.

[42] A. Asaei, P. N. Garner, and H. Bourlard, "Sparse component analysis for speech recognition in mullti-speaker environment," in *Proceeding of INTERSPEECH*, 2010.

[43] A. Asaei, M. J. Taghizadeh, H. Bourlard, and V. Cevher, "Multi-party speech recovery exploiting structured sparsity models," in *Proceedings of INTERPSEECH*, 2011.

[44] Z. Zhang and B. D. Rao, "Extension of sbl algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Transactions on Signal Processing*, 2012.

[45] M. Athineos and D. P. W. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Transactions on Signal Processing*, vol. 55(11), 2007.

[46] T. Peleg, Y. C. Eldar, and M. Elad, "Exploiting statistical dependencies in sparse representations for signal recovery," *IEEE Transactions on Signal Processing*, vol. 60(5), 2012.

[47] M. Golbabaee and P. Vandergheynst, "Compressed sensing of simultaneous low-rank and joint-sparse matrices," *submitted to IEEE Transactions on Information Theory, available in*, http://infoscience.epfl.ch/record/181506.