Structured Sparsity Models for Multiparty Speech Recovery from Reverberant Recordings

Structured Sparsity Models for Multiparty Speech Recovery from Reverberant Recordings

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

Abstract

We tackle the multi-party speech recovery problem through modeling the acoustic of the reverberant chambers. Our approach exploits structured sparsity models to perform room modeling and speech recovery. We propose a scheme for characterizing the room acoustic from the unknown competing speech sources relying on localization of the early images of the speakers by sparse approximation of the spatial spectra 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 demonstrate the effectiveness of the proposed approach for multi-party speech recovery and recognition.

Index Terms

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

Afsaneh Asaei and Hervé Bourlard are with Idiap Research Institute and also affiliated with École Polytechnique Fédérale de Lausanne, Switzerland. Mohammad Golbabaee and Volkan Cevher are with École Polytechnique Fédérale de Lausanne, Switzerland.

E-mails: afsaneh.asaei@idiap.ch, mohammad.golbabaei@epfl.ch, herve.bourlard@idiap.ch, volkan.cevher@epfl.ch

I. INTRODUCTION

RECOVERY of speech signal 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 competing sound sources yet remains a demanding challenge [1].

This paper considers distant-talking speech recognition in 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 dichotomized 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, 4]. The method proposed in [5] exploit the statistical characteristics to estimate the acoustic channel of the enclosure and performs joint deconvolution and separation of speech signals. The underlying assumption of the approaches belonging to the first category is the statistical independence of the sources. Moreover, these techniques are confined to the scenarios where the number of microphones is greater than or equal to the number of sources also known as overdetermined or determined mixtures respectively [6].

The second category relies on spatial filtering techniques based on beamforming or steering a microphone array beam-pattern towards the target speaker thus resulting in the suppression of the undesired sources [7, 8]. The underlying assumption of this approach is that there is no reverberation so the beamforming techniques are formulated upon upon direct path acquisition of the signals. These geometric techniques can work with any number of microphone including the scenarios in which the number of sources exceeds the number of sensors thereby, we have underdetermined mixtures [9].

The third category is based on sparse representation of the source signal, also known as Sparse Component Analysis (SCA) [10, 11]. 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 a fewer number of known data. As there are many solutions to such systems, the answer ought to be the sparest solution measured

in terms of the sparsity inducing norms [11, 12, 13]. The prior art on multichannel speech recovery through 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 gird 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 i as S_i 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 transpose. If we consider 1 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 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 here.

This paper exploits structured sparsity models to recover the unknown individual speech signals: $S_i, i \in 1, ..., G$ from a few known multi-party recordings when the speakers are talking simultaneously. In addition to the spatial sparsity, we will exploit sparsity in spectral domain. 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 model the sparsity underlying the acoustic model of the room characterized by the *Image Model* of multipath effect. The contribution of this paper is ultimately to introduce a unified theory of multiparty speech recovery formulated as a problem of signal recovery by exploiting structured sparsity models underlying the representation of information embedded in multichannel recordings.

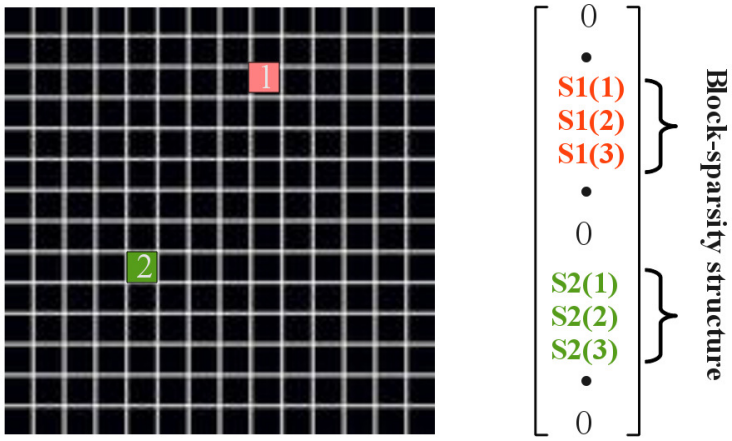

Fig. 1: The spatial sparsity of the speakers inside the room is illustrated through discretization of the planar area of the room into a grid of G cells. The sources occupy only two cells marked as 1 and 2. Hence, the spatial representation of the source signals generated inside the room is sparse.
Assuming that the sources are immobile, if we denote the 3 instances of the signal attributed to the speaker at cell i as $S_i(n) \in \mathbb{C}^{3\times 1}, n \in \{1,2,3\}$ and concatenate the signals corresponding to each cell, the signal vector of the room can be formed as $\mathcal{S} = [S_1^T, ..., S_G^T]^T \in \mathbb{C}^{3G\times 1}$. 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 [5]. Their approach relies on statistical independence assumption of the sources to perform joint deconvolution and separation of speech signals and it is limited to overdetermined scenarios. This assumption has been relaxed in the method proposed in [14] where multiple complex valued Independent Component Analysis (ICA) 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 [14].

A blind channel identification approach for speech separation and dereverberation is proposed in [15]. 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 (UNMCFLMS) algorithm [16] and de-reverberation 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) [17]). A real-time implementation of this approach has been presented in [18], where the optimum inverse filtering is substituted by an iterative technique, which is computationally more efficient and allows the inversion of long RIRs in real-time applications [18]. 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 (WPE) for joint de-reverberation and separation of individual speech sources from determined and overdetermined mixtures [19]. This method does not perform channel estimation and it does not perform well in estimation of the acoustic channel 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 [20].

This paper takes a new perspective to analysis of multi-channel recordings. We cast the microphone array acquisition as compressive sensing the information embedded in acoustic field and we leverage the theory of model-based sparse recovery for characterization of the acoustic measurements and recovering the speech components. Our approach features 4 contribution:

- We separate the individual speech sources from the underdetermined convolutive mixtures exploiting various algorithmic approaches to structured sparse recovery while incorporating different types of spectral, spatial as well as acoustic multi-path structures.
- We estimate the geometry of the reflective surfaces from recording of multiple unknown sources located at unknown positions exploiting sparse recovery and low-rank clustering techniques.
- We propose a new formulation of the reverberation model and estimate the absorption factors of the surfaces using structured sparse recovery.
- We analyze how the performance of speech recovery is entangled with the design of microphone array layout.

In this paper, we first overview the problem statement and characterization of multi-party multi-channel recordings in Section III along with the main assumption under consideration. The structured sparse speech recovery algorithms are described in Section IV. We set up the formulation of the structured sparse acoustic modeling in Section V. We elaborate on the theory of room geometry estimation in Section V-A and propose the approaches to absorption coefficient estimation in Sections V-B and V-C. The experimental analysis of the proposed techniques are discussed in Section VII. The conclusions are drawn in Section VIII.

The notation used in this paper will be as follow:

- x_m: signal of the m^{th} microphone in time domain
- X_m: signal of the m^{th} microphone in frequency domain
- s_n: signal of the n^{th} source in time domain
- S_n: signal of the n^{th} source in frequency domain
- h_{mn}: acoustic channel between the m^{th} microphone and n^{th} source in time domain
- H_{mn}: acoustic channel between the m^{th} microphone and n^{th} source in frequency domain
- Φ: microphone array manifold matrix; it characterizes the acoustic projections associated to the acquisition of source signals inside the enclosure
- l: each time sample
- f: each frequency bin
- F: number of Fourier coefficients
- τ: each frame of speech
- $\mathcal{T}$: number of speech frames
- $\circledast$: convolution operation
- T: transpose operation
- $*$: conjugate transpose operation
- $\dagger$: pseudo-inverse operation
- N: number of sources
- M: number of microphones
- G: number of cells
- c: speed of sound assumed to be constant
- R: order of reflections in a reverberant room

- D: number of reflective surfaces within the enclosure

III. Multiparty Reverberant Recordings

A. Problem Statement

In the present paper, we deal with 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

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

This formulation is stated in time domain; to represent it in a sparse domain, we apply the Gabor expansion, i.e., the discrete Short-Time Fourier Transform (STFT) of speech signals. Following from the convolution-multiplication property of the Fourier transform, the mixtures in frequency domain can be written as

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

Our objective is to recover the individual source signals $S_{\cdot}$ from the distant microphone recordings. There is no prior information about the number of sources and the acoustic mixing channels.

B. Multi-party Speech Representation

We consider a scenario in which N speakers are distributed in a planar area 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 spectra of the sources is defined as a vector with a sparse support indicating the components of the signal corresponding to each cell of the gird.

We consider 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 $S_n \in \mathbb{C}^{F \times 1}$ denotes the spectral representation or signal of the n^{th} source (located at cell number n) in Fourier domain. We express the signal ensemble at 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 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.

C. Acoustic Measurement Characterization

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 [21]. Taking into account the physics of the signal propagation and multi-path effects, the projections associated with the source located at the cell g where ν_g represents the position of the center of the cell and captured by microphone i located at position μ_i are characterized by the media Green's function and denoted as $\xi_{\nu_g \to \mu_i}$ defined by

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

where $j = \sqrt{-1}$ and ι is the reflection coefficient; ι^r is the reflection coefficient after r reflections of the walls. The attenuation constant α depends on the nature of the propagation and is considered in our model to equal 1 which corresponds to the spherical propagation. This formulation assumes that if $s_1(l) = s(l)$ and $s_2(l) = s(l-\rho)$, then $S_2(f,\tau) \approx \exp(-jf\rho) S_1(f,\tau)$.

Given the source-sensor projection defined in Equation (3), we construct matrix $\Xi_{\nu_g \to \mu_i}$ for the measurement of the F consecutive frequencies as

$$\Xi_{\nu_g \to \mu_i} = \begin{bmatrix} \xi^1_{\nu_g \to \mu_i} & 0 & \dots & 0 \\ 0 & \xi^2_{\nu_g \to \mu_i} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \xi^F_{\nu_g \to \mu_i} \end{bmatrix} \tag{4}$$

Hence, the projections associated to the acquisition of the source signals located on the grid by microphone i is $\phi_i = [\Xi_{\nu_1 \to \mu_i} ... \Xi_{\nu_g \to \mu_i} ... \Xi_{\nu_G \to \mu_i}]$ and the measurement matrix of M-channel microphone array would be defined as

$$\Phi = \begin{bmatrix} \phi_1 \\ \\ \vdots \\ \\ \phi_M \end{bmatrix} \tag{5}$$

As indicated by Equation 3, characterizing the acoustic projections amounts to identifying the location of the *source images* as well as the absorption factors of the reflective surfaces. We exploit this parametric model to address the speech recovery problem in this paper.

In Section III-B we mentioned that the sparse vector $\mathcal{S}$ generates the microphone measurements as $\mathcal{X} = \Phi\mathcal{S}$. Our goal is to recover $\mathcal{S}$ from a small number of measurements (i.e., $M < G$). There are many solutions to this problem, we thus exploit the prior information on sparse properties of $\mathcal{S}$ to circumvent the ill-posedness of the problem.

We cast the underdetermined speech recovery problem as sparse approximation where we exploit the underlying structure of the sparse coefficients to recover the signal components more efficiently from fewer number of measurements [22]. This is the topic of the following Section IV.

IV. Structured Sparse Speech Recovery

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

$$\hat{\mathcal{S}} = \underset{\mathcal{S}\in\mathbb{M}}{\operatorname{argmin}} \|\mathcal{S}\|_0 \quad \text{s.t.} \quad \mathcal{X} = \Phi\mathcal{S} \tag{6}$$

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

The major classes of computational techniques for solving sparse approximation problem stated in Equation (6) include greedy pursuit, convex relaxation, non-convex optimization, and Bayesian algorithms [23]. This paper considers greedy algorithms and convex optimization, which offer provable correct solutions under well-defined conditions. 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 Equation (6) by replacing the counting function with a sparsity inducing norm.

A. Structured Sparsity models

We focus on two types of structures underlying the sparse coefficients:

- The first structure is the block-dependency model which is exhibited if some interconnections between the adjacent frequencies exist. In case of the vector $\mathcal{S}$, the *block sparsity structure* 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

$$\mathcal{F}_B = \{[f_1, ..., f_b], [f_{b+1}, ..., f_{2b}], [f_{F-b+1}, ..., f_F]\}. \tag{7}$$

- The second structure is the harmonic-dependency model exhibited if there are some interconnections between frequencies which are the harmonics of a fundamental frequency. In voiced speech, most of the energy in the speech signal 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 $\mathcal{F}_H$ structure defined as

$$\mathcal{F}_H = \{kf_0 | 1 < k < K\}, \tag{8}$$

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

B. Model-based Sparse Recovery

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

- *IHT*: Iterative hard thresholding (IHT) offers a simple yet effective approach to estimate the sparse vectors [24]. 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 [25] which is an accelerated scheme for hard thresholding methods with the following recursion

$$\begin{cases} \hat{\mathcal{S}}_0 = 0 \\ \mathrm{r}_\mathrm{i} = \mathcal{X} - \Phi\hat{\mathcal{S}}_\mathrm{i} \\ \hat{\mathcal{S}}_{\mathrm{i}+1} = \mathcal{M}^{\mathcal{F}_\cdot}\left(\hat{\mathcal{S}}_\mathrm{i} + \kappa\Phi^\mathsf{T}\mathrm{r}_\mathrm{i}\right) \end{cases} \tag{9}$$

where the step-size κ is the Lipschitz gradient constant to guarantee the fastest convergence speed [26]. 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 $\mathcal{F}_\mathrm{B}$ or $\mathcal{F}_\mathrm{H}$ structures [25].

- *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 [27, 24, 28]. The signal estimation algorithm would thus have the following recursion

$$\begin{cases} \Lambda_0^{\mathcal{F}_\cdot} = 0 \\ \lambda_\mathrm{i} = \underset{\lambda\in\Phi_{\setminus\Lambda_{\mathrm{i}-1}^{\mathcal{F}_\cdot}}}{\operatorname{argmin}} \|\mathcal{X} - \Phi_{\Lambda_{\mathrm{i}-1}^{\mathcal{F}_\cdot}\cup\lambda}\Phi^\dagger_{\Lambda_{\mathrm{i}-1}^{\mathcal{F}_\cdot}\cup\lambda}\mathcal{X}\|_2 \\ \Lambda_\mathrm{i}^{\mathcal{F}_\cdot} = \Lambda_{\mathrm{i}-1}^{\mathcal{F}_\cdot} \cup \lambda_\mathrm{i} \\ \hat{\mathcal{S}}_\mathrm{i} = \Phi^\dagger_{\Lambda_\mathrm{i}}\mathcal{X} \end{cases} \tag{10}$$

- $\mathrm{L}_1\mathrm{L}_2$: Another fundamental approach to sparse approximation replaces the combinatorial counting function in the mathematical formulation stated in Equation (6) with the L_1 norm, yielding convex optimization problems that admit a tractable algorithm referred to as basis pursuit [29]. We use a multiple-measurement version of basis pursuit algorithm by re-arranging the components of $\mathcal{S}$ as a row-sparse matrix with the $\mathfrak{n}^{\mathcal{F}_\cdot}$ columns corresponding to the common sparsity structure $\mathcal{F}_\cdot$ referring to either $\mathcal{F}_\mathrm{B}$ or $\mathcal{F}_\mathrm{H}$. Hence, the optimization problem to recover the structured sparse coefficients would be the following

$$\hat{\mathcal{S}} = \arg\min \|\mathcal{S}\|_{L_1,L_2} \quad \text{s.t.} \quad \mathcal{X} = \Phi\mathcal{S},$$
$$\|\mathcal{S}\|_{L_1,L_2} = \left(\sum_{i=1}^{G} \left[\sum_{j=1}^{n^{\mathcal{F}_\cdot}} \mathcal{S}^2(i,j) \right] \right)^{1/2}. \tag{11}$$

The speech recovery approach as described in this section, requires identification of the acoustic measurements. To tackle this problem, we incorporate the *Image Model* of multipath effect, as stated in Equation (3). We elaborate on characterization of the room acoustic in the next section.

V. Structured Sparse Acoustic Modeling

Recall from Section III-C that characterizing the acoustic projections amounts to identifying the location of the *source images* as well as the *absorption factors* of the reflective surfaces. In Section V-A, we estimate the geometry of the room to identify the location of the *source images*. In Sections V-B and V-C, we address the problem of *absorption coefficient* estimation.

A. Estimation of the Room Geometry

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

$$H(f,\mu_i,\nu_g) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu_i - \nu_g^r\|} \exp(jf\frac{\|\mu_i - \nu_g^r\|}{c}). \tag{12}$$

$H(f,\mu_i,\nu_g)$ indicates the room impulse response function between the microphone located at μ_i and a source located at ν_g. Hence, identifying the locations of the R Images of the source corresponds to identifying the 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 [21].

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 [30]. Relying on this observation, we propose to localize the *source images* using the sparse recovery algorithm with a free space measurement model, i.e., $R = 0$, while the deployment of the grid captures the location of early reflections. The 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 source signals in a *close proximity*

TABLE I: Room geometry estimation procedure

- Run sparse source localization algorithm with a free-space measurement model.
- Run k-means clustering using Cosine angle as the distance metric.
  - The centroid of the clusters are selected as the location of the nearest sources to the center of the array.
  - The Cosine angle is measured between components of the signal attributed to the actual sources w.r.t. the components corresponding to the source images.
  - The number of each cluster members is limited to $D(D+1)/2$.
- Find the room geometry by identifying the dimensions which yield the best approximation of the location of source images in least-squares sense.

to the microphone array within a distance d as the actual sources generating the signals $S_n, n = 1, ..., I$. The localized images are sorted up to the order of $D(D+1)/2$ according to the Cosine angle between the estimated signals and the source signal (S_n) and considered as the images associated to the n^{th} source. 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 source images from the location of the actual sources. Table I summarizes the steps to implement room geometry estimation.

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. The *Image Model* indicates the sparsity of the room impulse response function with a particular structure imposed by the vertical reflections. We refer to this property as the acoustic structured sparsity and exploit it to address the problem of estimating the absorption coefficients. In Section V-B, we propose an approach for estimating the absorption factors if there is only a single talker. In Section V-C, we elaborate on a novel model of the room reverberation which enables accurate estimation of the absorption factors from a plurality of speech sources.

B. Single-Source Absorption Coefficient Estimation

We consider the linear convolutive model of the reverberant enclosure and denote the time-domain acoustic channel between the source and microphone i as $h_i(l)$. Hence, the signal of the microphone i, $x_i(l)$ is a filtered version of source signal $s(l)$ as $h_i(l) \circledast s(l)$. It is straightforward to see that

$$x_i(l) \circledast h_j(l) = x_j(l) \circledast h_i(l); \tag{13}$$

Considering an L-tap acoustic filter, for $l = L, ..., \mathcal{L}$, where $\mathcal{L}$ is the length of the recorded signal, (13) becomes:

$$[\chi_i(L) - \chi_j(L)] \begin{bmatrix} h_j \\ h_i \end{bmatrix} = 0, \tag{14}$$

where $h := [h(L), ..., h(0)]^T$ and

$$\chi(L) = \begin{bmatrix} x(L) & x(L+1) & \dots & x(2L) \\ x(L+1) & x(L+2) & \dots & x(2L+1) \\ \vdots & \vdots & \ddots & \vdots \\ x(\mathcal{L}-L) & x(\mathcal{L}-L+1) & \dots & x(\mathcal{L}) \end{bmatrix}. \tag{15}$$

This equation forms the basic idea for blind channel identification by least squares optimization [31]. Relying on the structured sparsity model as indicated by the *Image Model* of multipath effect, we propose the optimization algorithm constrained on the structured sparsity to capture the main reflections characterized by the *Image Model*.

Despite the existence of various reflective objects inside the room, the structured sparsity model obtained through the room geometry is theoretically sound due to the fact that the multi-path signal energy is a function of the reflective areas. Hence, for the general environment of the meeting rooms, many objects are acoustically transparent [32]. In addition to the theoretical evidence, we empirically verified the effectiveness of the structured sparsity constraint for identification of the real acoustic impulse responses from noisy reverberant data generated by the impulse responses available in Aachen Impulse Response (AIR) database [33].

Given the room geometry and the source location, the support of the highest energy components of RIR is determined by the *Image Model* and denoted by Ω_d which refers to the direct path component calculated precisely as ϑ and $\Omega_\mathcal{I}$ which refers to the support of the reflections. We define $\Pi := [\chi_i(L) \quad -\chi_j(L)]$ and $\mathcal{H} := [h_j^T \quad h_i^T]^T$. The structured sparse acoustic filter will be obtained by the following optimization

$$\begin{aligned} &\hat{\mathcal{H}} = \arg\min \|\mathcal{H}\|_1 \\ &\text{s.t.} \quad \|\Pi\mathcal{H}\|_2 \leqslant \epsilon, \quad \mathcal{H}(\Omega_d) = \vartheta, \quad \mathcal{H}(\Omega_r) > 0 \end{aligned} \tag{16}$$

The estimated RIR is then used to estimate the absorption coefficients of our model stated in Equations (3-5) by least squares fitting and to characterize the acoustic channels of all cell positions in order to identify the microphone array measurement matrix. The speech recovery is then achieved by structured sparse recovery algorithms as explained in Section IV.

C. Multi-Source Absorption Coefficient Estimation

This section elaborates on a novel formulation of the reverberant recordings which entangles 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. Generalized to the algorithm proposed in Section V-B, we can estimate the frequency-dependent absorption factors in a multi-source environment.

1) Factorized Formulation of the Reverberant Recordings: We propose a novel formulation of 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[1]. 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 $\Omega_{\mathfrak{i}} \subset \{1, \ldots, \mathcal{G}\}$ 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 could be concisely stated as follows:

$$\mathcal{X} = \mathsf{OP}\mathcal{S}. \tag{17}$$

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 $\mathcal{X} \in \mathbb{C}^{\mathsf{M} \times \mathcal{T}}$, the *observation* matrix of $\mathcal{T}$ frames consisted of spectro-temporal representation of M microphones at a particular frequency band, we decompose the microphone recordings into the following terms:

- $\mathcal{S} \in \mathbb{C}^{\mathsf{G} \times \mathcal{T}}$ 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 $\mathsf{N} \ll \mathsf{G}$ cells are occupied with active sources and contain nonzero elements and the *support* set $\mathcal{S} \subset \{1, \ldots, \mathsf{G}\}$ representing the position of those N active sources is sparse. In other words, the *spatial sparsity* indicates S to be a row-sparse matrix with a support corresponding to the position of the actual sources.
- $\mathsf{P} \in \mathbb{R}_{+}^{\mathcal{G} \times \mathsf{G}}$ is the *permutation* matrix such that its $\mathfrak{i}^{\mathfrak{th}}$ column contains the absorption factors of $\mathcal{G}$

[1]If each of the sources have R images, $\mathsf{N}(\mathsf{R}+1)$ actual-virtual sources are active.

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 Equation (3) and it is equal to Φ when $R = 0$.

2) Source Localization and Absorption Coefficient Estimation: Relying on 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 the columns of P. Given the model of the microphone recordings stated in (17), the covariance matrix of the observations is

$$\begin{aligned} C = \mathcal{X}\mathcal{X}^* &= O\Sigma O^* \\ &= \sum_{i=1}^{G} O_{.,\Omega_i} \Sigma_{\Omega_i,\Omega_i} O^*_{.,\Omega_i}, \end{aligned} \tag{18}$$

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 specifies the energy of the individual sources - Section VII-B 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.,

$$\begin{aligned} \Sigma_{j,i} &= 0 \qquad \forall (j,i) \notin \bigcup_{i=1}^{G} \Omega_i \times \Omega_i, \\ \Sigma_{\Omega_i,\Omega_i} &= \|\mathcal{S}_{i,.}\|_2^2 P_{\Omega_i,.} P^*_{\Omega_i,.}, \end{aligned} \tag{19}$$

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 identify 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 (18) as

$$C_{vec} = \underbrace{\begin{bmatrix} B(1) & B(2) & \dots & B(G) \end{bmatrix}}_{\mathcal{B}} \underbrace{\begin{bmatrix} \nu(1) \\ \nu(2) \\ \vdots \\ \nu(G) \end{bmatrix}}_{\mathcal{V}} \tag{20}$$

$$\forall i \in \{1 \dots, G\}:$$

$$\nu(i) \triangleq (\Sigma_{\Omega_i,\Omega_i})_{vec},$$

$$B(i) \triangleq \overline{O}_{\cdot,\Omega_i} \otimes O_{\cdot,\Omega_i}.$$

where $\otimes$ denotes the Kronecker product between two matrices and $\overline{O}_{\cdot,\Omega_i}$ is the *element-wise* conjugate of $O_{\cdot,\Omega_i}$. In a typical problem setup, very few microphones are used for recording, i.e. $M \ll \mathcal{G} < \sum_{l=1}^{G} |\Omega_i|$; thus recovering $\Sigma_{\Omega_i,\Omega_i}$ requires solving an underdetermined system of linear equations and therefore, in general (18) 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 $\nu(i)$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:

$$\begin{aligned} \underset{\Sigma^1,\ldots,\Sigma^G}{\arg\min} \quad & \sum_{i=1}^{G} \left\| \Sigma^i_{vec} \right\|_{L_2} \\ \text{subject to} \quad & \|C_{vec} - \mathcal{B}\mathcal{V}\|_{L_2} \leqslant \varepsilon \\ & \left(\mathcal{V} = \left[(\Sigma^i_{vec})^T, , \ldots, (\Sigma^i_{vec})^T \right]^T \right) \\ & \Sigma^i = (\Sigma^i)^* \quad \forall i \in \{1, \ldots, G\} \\ & \Sigma^i_{l,j} \geqslant 0 \qquad \forall l, j, i \end{aligned} \tag{21}$$

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 (21) becomes equivalent to the L_1L_2 norm[2] 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 (21) by using the iterative proximal splitting algorithm [34].

To summarize, we obtain the location of the sources and their images which also corresponds to the support of the room impulse response function for multiple sources. 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* R, as stated through Equations (3)-(5). In a scenario where $N < M$, we apply inverse filtering to perform joint speech separation and deconvolution as explained in the following Section V-C3.

3) Speech Recovery by Inverse Filtering the Acoustic Channel: The approach presented in Sections V-C1 and V-C2 enables us to localize the sources and model the mixing channels. Thereby, 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 characterized by matrix H, we recover the desired signal by inverse filtering stated as

$$\hat{S} = (H^T H)^\dagger H^T X \tag{22}$$

This operation performs exact deconvolution of the signal from the early room impulse response

[2]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 (11)

function [15, 17]. The late reverberation can be statistically modeled as an exponentially decaying white Gaussian noise [35] which also possess the diffuse characteristics [36].

To reduce the effect of late reverberation and enhance the signal, we apply the post-processing proposed in [37]. Among several post-filtering methods proposed in the literature [38, 39], the Zelinski post-filtering [37] 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-F. The dereverberation of the early impulse response achieved by inverse filtering the acoustic channels enables a more efficient post-filtering as formulated in [37].

VI. Compressive Acoustic Measurements

The approach that we have taken in this paper to address the multi-party speech recovery as studied throughout Sections III-V, 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 theory for characterization of the acoustic measurements and recovering the speech components. In this framework, the theoretical analysis of the performance bounds of our approach is entangled with the performance of the sparse recovery algorithms. A fundamental property to guarantee the theoretical performance bounds is the coherence of the measurement matrix [23] defined as

$$\mu(\Phi) = \max_{1 \leqslant j,k \leqslant G, j \neq k} \frac{|\langle \phi_j, \phi_k \rangle|}{\|\phi_j\| \|\phi_k\|}. \tag{23}$$

The coherence quantifies the smallest angle between any pairs of the columns of Φ. The number of recoverable non-zero coefficients (K) using either convexified or greedy sparse recovery is inversely proportional to μ [23] as

$$K < \frac{1}{2}(\mu^{-1} + 1)$$

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 [40] that the free-

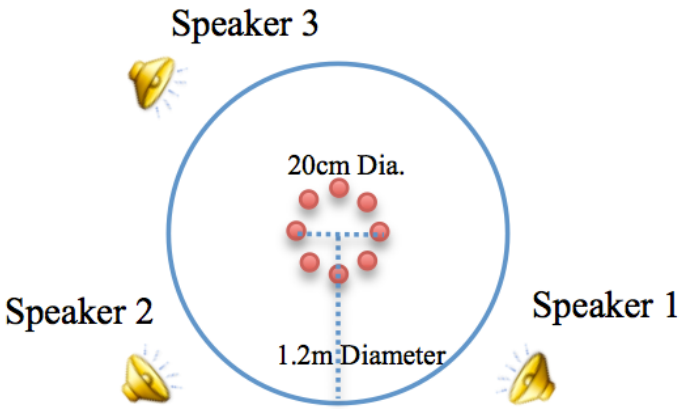

Fig. 2: Loudspeaker and microphone placement used for recording MONC corpus [42].

space Green's function constituted projections given that the inter-element spacing is large enough exhibit 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 [40, 41]. 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 addressed in Section VII.

VII. Experimental Analysis

A. Data Recordings Set-up

Experiments were performed in the framework of the Multichannel Overlapping Numbers Corpus (MONC). This database is acquired by playback of utterances from the Numbers Corpus release 1.0, prepared by the Center for Spoken Language Understanding at the Oregon Graduate Institute [42].

The recordings were made in a 8.2m $\times$ 3.6m $\times$ 2.4m rectangular room containing a centrally located 4.8m $\times$ 1.2m 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 1.2m. The loudspeakers were placed at 90° spacings at an elevation of 35cm (distance from table surface to center of main speaker element). An eight-element, 20cm diameter, circular microphone array placed in the center of the table recorded the mixtures. The recording scenario is illustrated in Fig. 2. One hour of speech signals are recorded at 8 kHz sampling frequency. The average signal to noise ration (SNR) of the recordings is 9dB.

B. Orthogonality of Spectrographic Speech

We carried out experiments to investigate the orthogonality of multiple speech sources in Fourier domain. In this experiment, 3 speech signals, 9s each, is analyzed in frames of size 128ms (fft-size = 1024) with 50% overlap; thus we obtain 3 matrices of 512 by 140 corresponding to the STFT of each source. The orthogonality is measured for each frequency band independently. We construct the matrix $\mathcal{X}_{3\times 140}$ where each row corresponds to each source and has the frequency components of a particular band along 140 frames. In case of perfectly orthogonal sources, $\mathrm{C} = \mathcal{X}\mathcal{X}^*$ is Identity and the energy of the diagonal of the matrix is equal to the matrix Frobenius norm. Fig. 3-right-hand-side illustrates the diagonal-L_2-norm/matrix-Frobenius-norm.

In addition, we performed some experiments by pointwise multiplication of the STFTs of two utterances and plot the histograms of the resulted values. Fig. 3-left-hand-side illustrates the obtained histogram. As 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 spectro-temporal domain are non-overlapping or disjoint.

C. Room Geometry Estimation

The first step to characterize the room acoustic is to estimate the room geometry. We accomplish this step through localization of the images of multiple speakers in a large extended area using the sparse recovery framework with a free space model as followed by the least-squares regression of the room geometry as explained in Section V-A.

The location of the source images corresponds to the support of the room impulse response function. The energies of the recovered signals are sorted and truncated to the order of $\mathrm{D}(\mathrm{D}+1)/2$ where D denotes the number of reflective surfaces and it is equal to 6 in our study to cover the support of the early reflections of the walls to guarantee the uniqueness of the solution [30]. 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 geometries and identify the best fit to the estimated support in least-squares sense.

The planar area of the room is divided into cells with 25cm spacing. The distance threshold to identify the actual sources is selected as 1m. To achieve a higher estimation, we restricted our discretized gird to the orthogonal subspaces corresponding to the orthogonal walls. We could estimate the geometry of the room up to 50cm error from the recordings of 3 sources in a close proximity to the microphone array as depicted in Fig. 2.

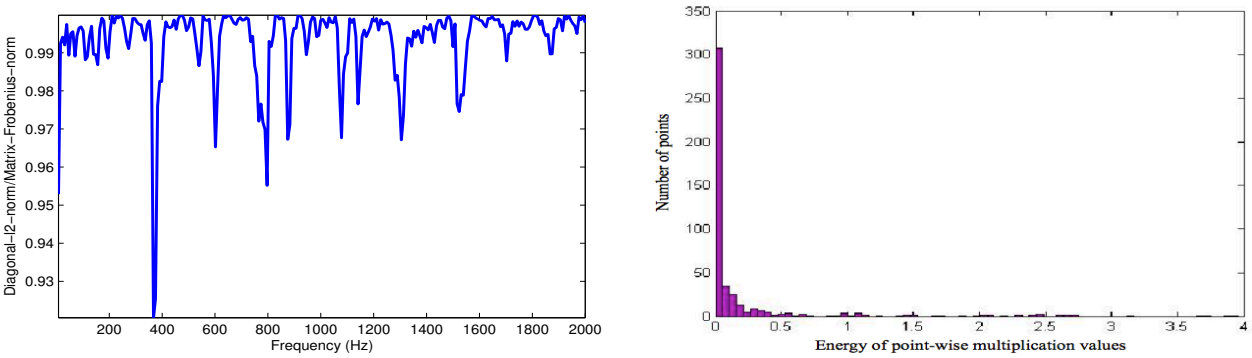

Fig. 3: Speech orthogonality

D. RIR Estimation

The second step to characterize the room acoustic is to estimate the absorption coefficients of the reflective surfaces. We accomplish this step through estimation of the room impulse response (RIR) function by implementing the technique explained in Section V-B. We used the CVX software package [43] for optimization formulated in (16) while sigma is chosen 0.1. The data was provided by concatenating 20 single speaker speech utterances.

The super-resolution source localization is performed based on the energy recovered from each cell using sparse recovery framework while the forward model corresponds to the direct path propagation and the support of the RIR function was determined considering a 6-sided model of an enclosure with the known geometry. We assumed that the reflections of the carpet floor are trapped under the table; hence, the meeting table was considered as the floor in our Image Model. The room reverberation time is measured about 100 ms from the energy decay curve of the estimated RIR and the reflection coefficients are estimated as 0.1 for the walls as well as the ceiling and 0.6 for the meeting table. Our estimation matches the empirical Sabin-Franklin's formula [44]:

$$\mathrm{RT}_{60} = \frac{24\ln(10)V}{\mathrm{c}\sum_{\mathrm{i}=1}^{6} W_{\mathrm{i}}(1-\iota_{\mathrm{i}}^{2})}, \tag{24}$$

where V denotes the volume of the room, ι_{i} the reflection coefficient, and W_{i} the surface of the i^{th} wall.

Although our method is blind, we verified the estimated impulse response and the corresponding reflection coefficients through adaptive filtering technique using the original clean speech provided at MONC from the original Numbers corpus. Figure 4 shows the effectiveness of the room impulse response

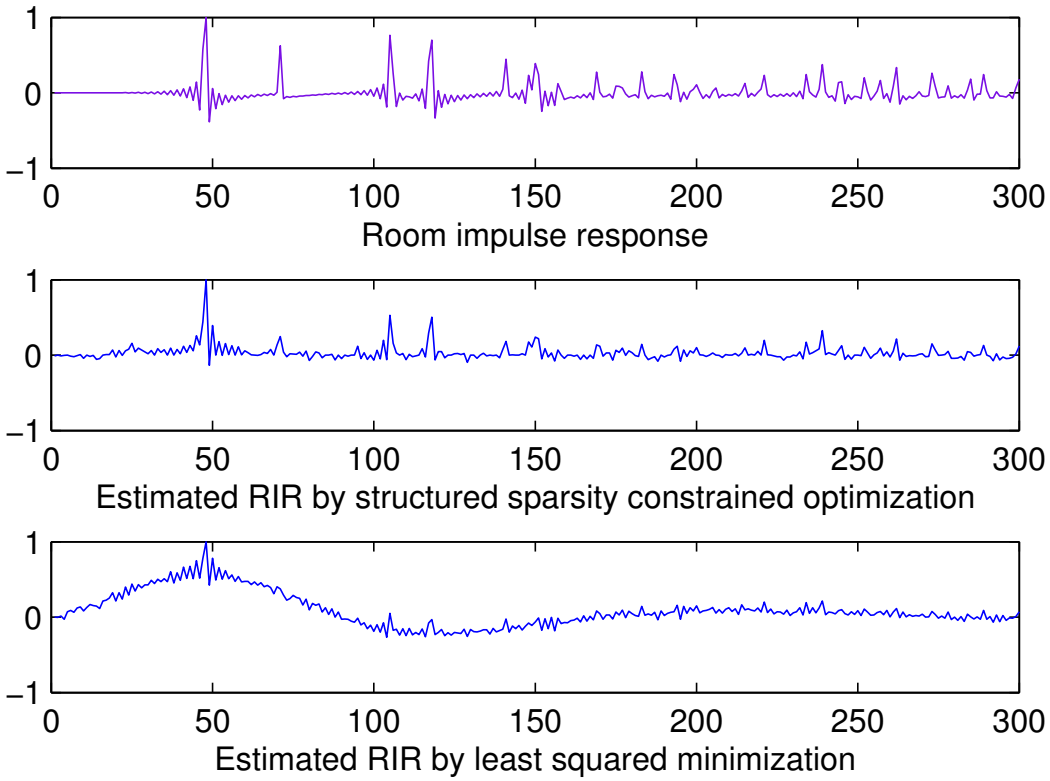

Fig. 4: Room Impulse Response (RIR) estimation from noisy measurements

estimation with the structured sparsity constraints and the alternative least-squared optimization from noisy data.

E. Structured Sparse Speech Recovery Performance

The speech recovery experiments are performed using different sparse recovery approaches to incorporate the block inter-connection as well as harmonicity of the spectro-temporal coefficients of speech signal. The spectro-temporal representation required for speech recovery is obtained by windowing the signal in 256ms frames using a Hann function with 25% overlap. The quality evaluation results in terms of Signal to Interference Ratio (SIR) [45] and Perceptual Evaluation of Speech Quality (PESQ) [46] are summarized in Figure 5. The block-size b was set to 4 as it was shown yielding the best results, especially for B-OMP and B-L_1L_2.

In the harmonic model, we consider that $f_0 \in [150 - 400]$ Hz. Those frequencies that are not the harmonics of f_0 are recovered independently in H-IHT and H-L_1L_2. We also considered that the harmonic

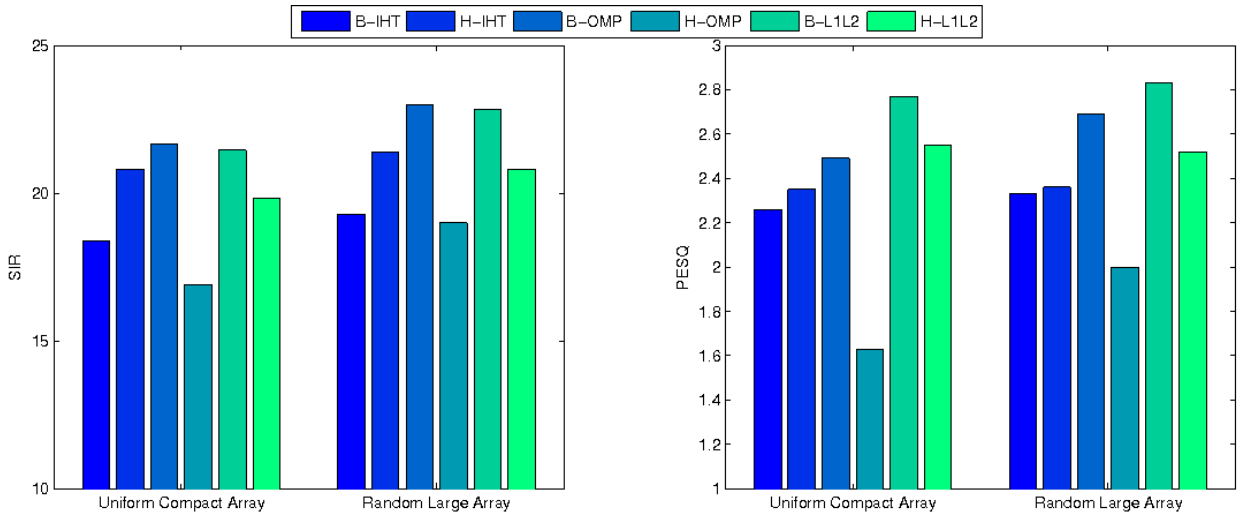

Fig. 5: Quality evaluation of the separated speech using different sparse recovery approaches in terms of SIR and PESQ. The baseline measures are -3.68 and 1.44 respectively

structures are non-overlapping and k spans the full frequency band. For H-OMP, the harmonic subspaces are used to select the bases while projection is performed for the full frequency band.

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 [47]. 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 [25, 24] and a reasonable performance [48].

Considering the speech signal model consisted of voiced and unvoiced segments, the block-interdependency mostly corresponds to the unvoiced speech while the harmonicity is exhibited in the voiced segments; hence we expect that a combination of both of the structures is beneficial for efficient speech recovery.

Comparing the results with the conventional uniform-array, we observe that the random setting of microphone array can significantly improve the quality of the separated speech. Hence, the compact uniform microphone array set-up is not an optimal design from the sparse reconstruction standpoint and the present study motivates more investigation on sparse and ad-hoc microphone array layouts.

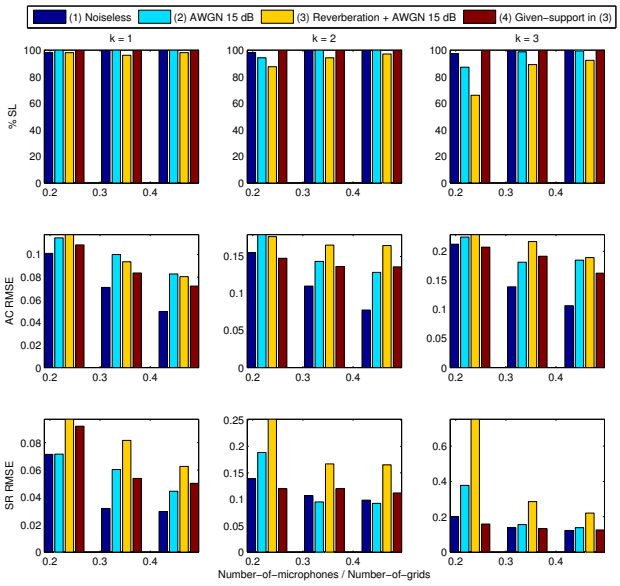

Fig. 6: Performance of the algorithm in terms of Source Localization (SL), Root Mean Squared Error (RMSE) of Absorption Coefficients (AC) estimation as well as Signal Recovery (SR). The test data are random orthogonal sources and the measurement matrix is the free-space Green function

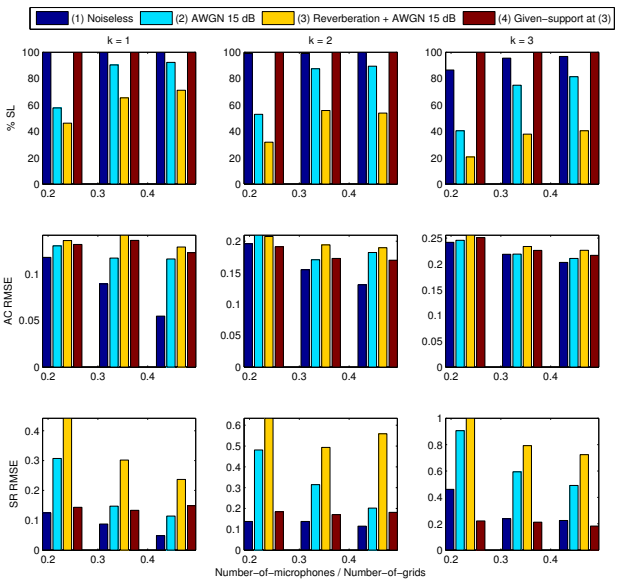

Fig. 7: Performance of the algorithm in terms of Source Localization (SL), Root Mean Squared Error (RMSE) of Absorption Coefficients (AC) estimation as well as Signal Recovery (SR). The test data are random speech utterances and the measurement matrix is the free-space Green function

F. Multi-party Acoustic Modeling and Speech Recovery

1) Synthetic data evaluations: We perform some initial evaluations on synthetic data in various noisy and reverberant conditions to validate our approach explained in Section V-C. The results of these experiments elucidate the empirical performance bounds for absorption coefficient estimation and signal recovery using block sparse recovery algorithm.

We consider the following recording set-ups: (1) 8-channel circular microphone array positioned in the middle of the room, (2) 12-channel microphone array: two sets of 6-channel circular array located far apart, (3) 16-channel microphone array: two sets of 8-channel circular array located far apart. We considered about 3cm displacement of the microphones. Evaluations are carried out using 1-3 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. The results of source localization (SL), absorption coefficients estimation (AC) and signal recovery (SR) are illustrated in Figs. (6) and (7).

2) Speech Recovery Performance: Given the location of the sources and the characterized room acoustic channel, we recover the desired signal by inverse filtering and perform speech recognition.

The automatic speech recognition (ASR) scenario was designed to broadly mirror that of Moore and McCowan [49]. A typical front-end was constructed using the HTK toolkit [50] with 25ms frames at a rate of 10ms. This produced 12 mel-cepstra plus the zero$^{\mathrm{th}}$ coefficient and the first and second time derivatives; 39 features in total. Cepstral Mean Normalization (CMN) is applied to the feature vectors, resulting in speech recognition performance improvement of about 15% relative. The ASR accuracy on the clean speech data is about 95%. We perform MAP adaptation by training directly on recovered data. The Zelinsky post-filtering is applied on the recovered speech prior to the recognition [37].

In addition to the speech recognition, we evaluate the quality of the recovered speech using SIR [45] as well as PESQ [46]. As our methods rely on the principles of spatial diversity, we compare them with beamforming techniques which possess similar essence. We used the super-resolution speaker localization based on sparse recovery to perform near-field beamforming. In addition, we compared our method with the sparse RIR-estimation approach described in Section V-B which relies on least-squares fitting of the Image Model (RIR-LS) [47] for estimation of the absorption factors. The resulting speech recovery performance is summarized in Table II.

As the results indicate, the proposed RAM-SR method 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 results support importance of the structured sparsity models to recover the spatio-spectral information from the multi-channel recordings. The spectral dependencies of the speech components could be further parametrized through auto-regressive (AR) models where we could characterize the dependencies along the temporal sequences or oblique structures in spectro-temporal domain [51]. We will further incorporate these structures into the framework of sparse recovery to devise the speech-specific algorithms which enable more efficient recovery performance. We could further exploit the statistical dependencies [52] for the specific task of speech recognition. The approach that presented to model the acoustic channel relies on joint-sparse recovery. Give the low-rank structure of the problem induced by the similar signals attributed to the source and its images, a promising extension of this work would be exploiting the low-rank and joint-sparse recovery algorithms as we studied in [53].

3) Real data evaluations: The scenario of the real data tests is explained in Section VII-A which is similar to the first set-up described above. We assume the location of the desired source to be fixed throught out the whole session. 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 [42] and

TABLE II: 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, vs. inverse filtering of the RIR estimation based on the least-squared estimation of the absorption factors (RIR-LS) and the proposed Room Acoustic Modeling via block Sparse Recovery (RAM-SR)

N	Meas.	Baseline	Lapel	SD	RIR-LS	RAM-SR
1	SIR	12.3	19.19	18.52	16.5	16.1
	PESQ	2.7	3	3.3	2.91	2.97
	WRR%	89.61	93.21	95	93.67	93.3
2	SIR	2.6	18.29	11.33	12.5	17.5
	PESQ	2	2.35	2.69	2.6	2.8
	WRR%	55.19	74.53	68.16	83.37	87.93
3	SIR	-0.7	18.35	10	10.1	14.2
	PESQ	1.6	2.27	2.48	2.4	2.62
	WRR%	39.92	68.13	61.45	70.88	79.21

computed and averaged for each frequency-band independently. The estimated absorption coefficients for each frequencies (computed at a resolution of 4 Hz) are illustrated in Fig. 8.

VIII. Conclusions

We addressed recovery of the speech information from structured sparse reconstruction perspective where we exploited spatial, spectral as well as acoustic structures underlying the representation of multiparty reverberant recordings to characterize the measurements and to recover the individual signals.

Relying on the structured sparsity of Image model of multipath effect, we identified the acoustic channel through (1) estimation of the room geometry by localization of the sources and low-rank clustering of the subspaces corresponding to each source and (2) estimation of the absorption coefficients using block-sparse recovery algorithm. Given the acoustic channel, we characterized the compressive acoustic projections and cast multiparty speech recovery as structured sparse approximation where we exploited the block dependency as well as harmonicy of the spectral coefficients to recover the speech signals. In addition, we showed that in a well-defined set-up we can pereform joint separation and deconvolution by frequency domain inverse filtering of the acoustic channels.

The proposed theory is validated by quantitative assessments performed through extensive experiments. The results on real data recordings demonstrate the applicability of our method for recovery of speech in multi-party scenarios for higher level application of distant speech recognition. The results motivate incorporating further parametrized sparsity structures to devise speech-specific recovery algorithms. They also support construction designs relying on ad-hoc and sparse microphone arrays layout for efficient capturing and extraction of the information embedded in the acoustic scene.

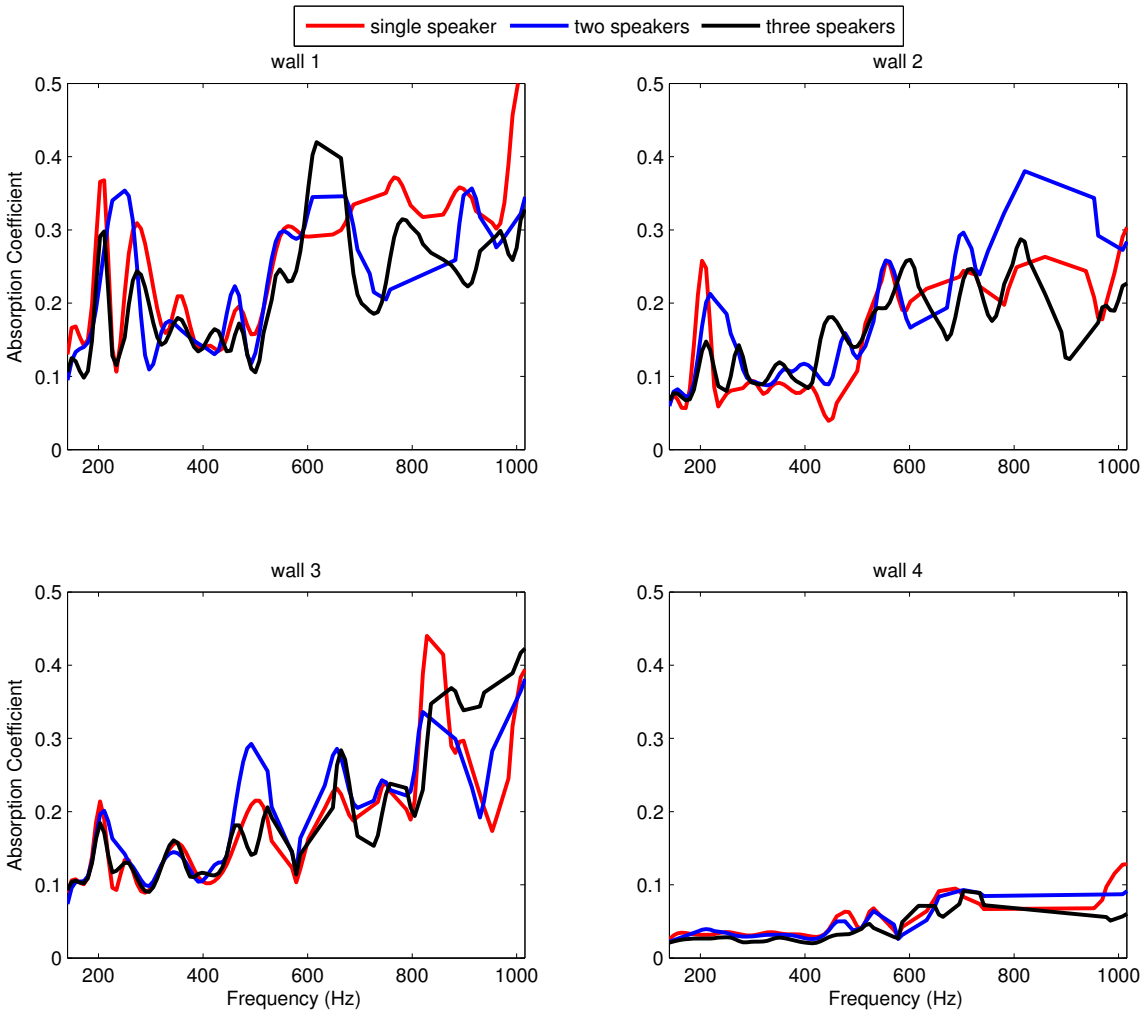

Fig. 8: Frequency-dependent absorption coefficients computed for each wall from the utterances of 3 competing speakers for the third speaker.

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.

VC acknowledges Rice University for his Faculty Fellowship, MIRG-268398, ERC Future Proof, DARPA KeCoM program #11-DARPA-1055 and SNF 200021-132548.

REFERENCES

[1] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, and G. Nolte, "The 2011 signal separation evaluation campaign (sisec2011): Audio source separation," 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 *In Proceedings of Eurospeech*, 2001.

[3] A. Ozerov, "A general flexible framework for the handling of prior information in audio source separation," vol. 20(5), 2012.

[4] S. C. Douglas, H. Sawada, and S. Makino, "Natural gradient multichannel blind deconvolution and speech separation using causal fir filters," vol. 13, 2005.

[5] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON-based blind system identification with application to multiple-source localization and separation," vol. 13, 2007.

[6] S. Makino, T. Lee, and H. Sawada, "Blind speech separation," *Springer*, 2007.

[7] 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.

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

[9] M. Wolfel and J. McDonough, "Distant speech recognition," 2009.

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

[11] 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.

[12] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, 2001.

[13] 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.

[14] F. Nesta and M. Omologo, "Convolutive underdetermined source separation through weighted interleaved ica and spatio-temporal source correlation," vol. 7191, 2012.

[15] 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," vol. 13(5), 2005.

[16] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," vol. 51(1), 2003.

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

[18] 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 *In Proceedings of 6th International Conference on Intelligent Computing*, 2010.

[19] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," vol. 19(1), 2010.

[20] 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 *In Proceedings of IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC),*, 2011.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustic Society of America*, vol. 65, 1979.

[22] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions in Information Theory*, 2010.

[23] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems,," *Proceedings of the IEEE*, 98, 2010.

[24] T. Blumensath and M. E. Davies, "Gradient pursuits," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2370–2382, 2008.

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

[26] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$, in soviet mathematics doklady," vol. 27, 1983.

[27] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," vol. 53(12), 2007.

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

[29] E. van den Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions,," *SIAM Journal on Scientific Computing*, 2, 2008, code available online, http://www.cs.ubc.ca/labs/scl/spgl1.

[30] 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.

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

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

[33] "Aachen Impulse Response (AIR) database - version 1.2," Institute of Communication Systems and Data Processing (IND), RWTH Aachen University, 2010, http://www.ind.rwth-aachen.de/AIR.

[34] P. L. Combettesand and J. C. Pesquet, "Proximal splitting methods in signal processing," vol. 49, 2011.

[35] E. A. Habets, "Speech dereverberation using statistical reverberation models," *Speech Dereverberation, Springer*, 2010.

[36] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," vol. 27(6), 1955.

[37] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *International Workshop on Acoustic Signal Enhancement*, 6, 1998.

[38] T. Wolff and M. Buck, "A generalized view on microphone array postfilters," *International Workshop on Acoustic Signal Enhancement*, 2010.

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

[40] L. Carin, "On the relationship between compressive sensing and random sensor arrays," *IEEE Antennas and Propagation Magazine*, vol. 51, pp. 72–81, 2009.

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

[42] "The Multichannel Overlapping Numbers Corpus," Idiap resources available online:, http://www.cslu.ogi.edu/corpora/monc.

pdf.

[43] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx.

[44] 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

[45] 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.

[46] L. D. Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing, implementation available at*, http://www.utdallas.edu/~loizou/speech/software.htm.

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

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

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

[50] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The htk book version 3.4," 2006.

[51] M. Athineos and D. P. W. Ellis, "Autoregressive modeling of temporal envelopes," vol. 55(11), 2007.

[52] T. Peleg, Y. C. Eldar, and M. Elad, "Exploiting statistical dependencies in sparse representations for signal recovery," vol. 60(5), 2012.

[53] M. Golbabaee and P. Vandergheynst, "Compressed sensing of simultaneous low-rank and joint-sparse matrices," http://infoscience.epfl.ch/record/181506.