Computational Methods for Underdetermined Convolutive Speech Localization and Separation via Model-based Sparse Component Analysis

Afsaneh Asaei[1], Hervé Bourlard[1,2], Mohammad J. Taghizadeh[3], Volkan Cevher[2]

[1] *Idiap Research Institute, Martigny, Switzerland*
[2] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
[3] *Huawei European Research Center, Munich, Germany*
{afsaneh.asaei, herve.bourlard}@idiap.ch, mohammad.taghizadeh@huawei.com, volkan.cevher@epfl.ch

Abstract

In this paper, the problem of speech source localization and separation from recordings of convolutive underdetermined mixtures is studied. The problem is cast as recovering the spatio-spectral speech information embedded in a microphone array compressed measurements of the acoustic field. A model-based sparse component analysis framework is formulated for sparse reconstruction of the speech spectra in a reverberant acoustic resulting in joint localization and separation of the individual sources. We compare and contrast the computational approaches to model-based sparse recovery exploiting spatial sparsity as well as spectral structures underlying spectrographic representation of speech signals. In this context, we explore identification of the sparsity structures at the auditory and acoustic representation spaces. The auditory structures are formulated upon the principles of structural grouping based on proximity, autoregressive correlation and harmonicity of the spectral coefficients and they are incorporated for sparse reconstruction. The acoustic structures are formulated upon the image model of multipath propagation and they are exploited to characterize the compressive measurement matrix associated with microphone array recordings.

Three approaches to sparse recovery relying on combinatorial optimization, convex relaxation and Bayesian methods are studied and evaluated based on thorough experiments. The sparse Bayesian learning method is shown to yield better perceptual quality while the interference suppression is also achieved using the combinatorial approach with the advantage of offering the most efficient computational cost. Furthermore, it is demonstrated that an average autoregressive model can be learned for speech localization and exploiting the proximity structure in the form of block sparse coefficients enables accurate localization. Throughout the extensive empirical evaluation, we confirm that a large and random placement of the microphones enables significant improvement in source localization and separation performance.

Keywords: Structured sparse representation, Model-based sparse recovery, Reverberation, Source localization and separation, Sparse component analysis, Computational auditory scene analysis

1. Introduction

Source localization and separation are central problems in various microphone array applications. This work takes place at the intersection of sparse component analysis and computational auditory scene analysis (CASA). Motivated by the commonalities between these two approaches to source separation, we consider structural dependencies influencing our auditory system for structured sparse recovery to develop a model-based sparse component analysis framework.

We consider the microphone array acquisition model as a linear convolutive mixing process stated as

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

where the signal of each microphone x_m is characterized as a superposition of the signal of individual sources s_n, $\forall n \in \{1, \cdots, N\}$ convolved with the acoustic channel, h_{mn}, between the position of source n and microphone m; N and M denote the number of sources and microphones, respectively.

This formulation is stated in time domain. To exploit the sparsity structures of sound as we shall see in this paper, the frequency domain representation is considered. Hence, the short time Fourier transform is applied on the microphone array signals. The convolution-multiplication property of the Fourier transform leads to the following mixing model

$$X_m = \sum_{n=1}^{N} H_{mn} S_n, \quad \forall m \in \{1, ..., M\} \tag{2}$$

where X_m, H_{mn} and S_n are the frequency domain representations of x_m, h_{mn} and s_n respectively.

The goal is to recover the individual source signals from M recorded mixtures. There is no prior knowledge about N, M and the acoustic channels H_{mn} and the estimation of the signals can only be achieved under assumptions about the signal or channel characteristics. Furthermore, the linear system expressed in (2) is underdetermined if $N \geqslant M$. Hence, additional assumptions are required to circumvent the ill-posedness of the separation problem. In the next section, we overview some of the prior works on multichannel techniques for speech separation.

1.1. Prior Work

The signal of individual sources can be recovered through multichannel linear filtering. These techniques can be grouped in two categories: *independent component analysis* and *beamforming*. The alternative non-linear strategies to demixing rely on extraction of the descriptions of individual sounds within the framework of *computational auditory scene analysis* and *sparse component analysis*. In the following, we study the assumptions underlying each approach and the scope of their application.

Independent component analysis (ICA) relies on the assumption that the signals are statistically independent. Hence, the objective is formulated to estimate an inverse/demixing filter such that the recovered

source signals are statistically independent (Comon and Jutten, 2010). This approach typically requires the number of source and microphones to be known in advance. In addition, the system has to be (over-) determined, i.e. $\mathsf{M} \geqslant \mathsf{N}$ so that an inverse filter exists. Furthermore, the mixing matrix must remain the same (stationary acoustic assumption) for a relatively long period of time to provide a reasonable estimate of a large number of model parameters. This assumption is difficult to fulfill in the realistic scenarios in which speakers turn their heads or move around.

Buchner et al. proposed to incorporate characterization of the room acoustics in the separation process (Buchner et al., 2007). Their approach exploits the statistical independence assumption of the sources to perform joint deconvolution and separation of the signals in overdetermined scenarios. Nesta et al. proposed an extension for underdetermined scenario where multiple complex valued 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 (Nesta and Omologo, 2012). The method 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 (Nesta and Omologo, 2012; Wang et al., 2011). Other extensions of ICA for the underdetermined scenarios consist in integration with sparse masking techniques within a hierarchical separation framework (Araki et al., 2004; Davies and Mitianoudis, 2004).

Beamforming is a geometric method to speech recovery that relies on steering/forming the beam pattern of the microphone array towards the desired source. This process can spatially filter out interferences from other directions regardless of the signal nature. Due to the spatial directivity, it can also mitigate the effect of reverberation which causes a field of dispersed signals. The limitation of beamforming is that separation is not possible when multiple sounds come from directions that are the same or near to each other (Wolfel and McDonough, 2009; Parra and Alvino, 2002).

Unlike the ICA approach, the beamforming requires information about the microphone array configuration and the sources (such as the direction of the desired source). However, there is no need to determine the number of spatially spread and reverberant interferences. It has been shown that beamforming can attain excellent separation performance in determined or overdetermined time-invariant demixing problems (Kumatani et al., 2011; Taghizadeh et al., 2012). However, only partial interference suppression is possible in underdetermined cases. Recent work considers non-linear combination of beamformers which incorporate sparsity of the spectro-temporal coefficients to address the underdetermined demising (Dmour and Davies, 2011). The application of this method is however limited to the anechoic scenarios and the performance is degraded in reverberant condition.

Huang proposed to exploit the acoustic channel to achieve speech separation and dereverberation (Huang et al., 2005). Their method applies a blind channel identification where the mixing procedure is delineated with a multiple-input multiple-output 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 the temporal echoes. The drawback of this approach is that it can only perform channel identification from single talk periods and it requires a high input signal-to-noise ratio.

Computational auditory scene analysis (CASA) aims to recover the source signals from one or two recordings of an acoustic scene by exploiting the principles of human auditory scene analysis (Wang and Brown, 2006). The capability of individual listening in a clutter of sounds is attributed to the ability of listeners to "glimpse" the target voice during intervals in the competing interferences (Brungart, 2001). To state it more specifically, the sparse distribution of speech energy in spectro-temporal plane results in gaps in the spectrum of the overlapping sounds during which listeners can obtain an uncorrupted estimate of the target speech signal. The contribution of CASA is thus identifying the spectro-temporal regions that are dominated by a single sound source (Kollmeier et al., 2008).

The CASA approach may be regarded as a two-stage process. The first stage is decomposition of the acoustic input into a collection of local spectro-temporal regions. The second stage is grouping the segments of the spectro-temporal scene that are likely to have arisen from the same environmental source into a perceptual structure. Bregman recognizes the major primitive grouping principles relying on (1) proximity in frequency and time (2) harmonicity (3) continuous or smooth transition (forming a continuous trajectory) (4) common onset and offset (5) amplitude and frequency modulation (Bregman, 1990; Parsons, 1976). These modeling mechanisms has been the source of various algorithmic approaches to deal with the complex listening situation (Faller and Merimaa, 2004). Many CASA systems achieve source separation by computing a mask to weight a spectro-temporal representation of the acoustic input and this has been regarded as the computational goal of CASA (Hu and Wang, 2001; Roman et al., 1991).

Sparse Component Analysis (SCA) relies upon the assumption that the signals have sparse representation (Zibulevsky and Pearlmutter, 2001). The sparsity implies that the representation of signal occupies only a small part of a larger space so the mixtures of sparse components are likely to be disjoint. In many cases that the canonical representation of the signals do not exhibit the sparsity properties, a linear transformation of the signal such as Fourier or wavelet transform yields sparse representation. Yilmaz et al. exploited the sparsity and disjoint characteristics of spectro-temporal representation of speech mixtures to perform separation (Yilmaz and Rickard, 2004). Their method relies on delay and attenuation differences between the signals acquired by two microphones to construct a binary mask and extract the individual signals. The extensions of this approach have been proposed for M-channel microphone array and convolutive mixtures recorded in a reverberant room by Melia et al in (Melia and Rickard, 2007) and Abrard et al. in (Abrard and Deville, 2005).

The key differences between different SCA techniques boil down to the method of clustering the components for mixing matrix estimation as well as mask construction or sparse recovery to achieve source separation (Jafari et al., 2006; Mourad and Reilly, 2010). Generalized to CASA, in many SCA approaches, a soft mask is applied thus the assumption that each spectro-temporal coefficient belongs to the same

source is relaxed and the recovered speech does not suffer from the musical noise and missing components in their spectrographic representation (Araki et al., 2005; Kearns et al., 1997). Recent advances consider the structures underlying the sparse coefficients to improve the separation quality in a model-based sparse component analysis framework while reducing the number of required measurements (Asaei, 2013; Baraniuk et al., 2010).

1.2. Contributions and Outline

This study takes place in the context of sparse component analysis. Motivated by the success of structured sparse recovery, our goal is to address two important questions with regard to application of SCA on speech recordings. The first question is:

1. *What are the sparsity structures underlying spatio-spectral representation of multiparty speech data?*

To that end, we elucidate the type of structures pertained to propagation and hearing of sound that can be incorporated to group the sparse coefficients of multichannel signal in the spectral domain. A key source of inspiration emerges from the psychoacoustic models studied in the field of auditory scene analysis. Building upon the theories of CASA and SCA, some auditory and acoustic structures are formulated in such a way that they can be integrated in the framework of sparse recovery. Then, the next question that we address is:

2. *What are the computational methods to sparse recovery that exploit these structures and how do they perform?*

To that end, different model-based sparse optimization procedures are investigated. This study focuses on the combinatorial, convex geometric and sparse Bayesian learning approaches to sparse recovery. The procedure underlying each method is briefly elaborated and thorough experiments are devised to evaluate the performance of each method in terms of source localization and speech separation quality under clean and noisy conditions.

It may be noted that the sparse recovery framework presented in this paper entangles the *synthesis sparsity* of the spatial domain with the *analysis sparsity* of the spectral domain. The canonical statement of the spatial representation of the signals exhibits sparsity in the space of source positions, a property referred to as *synthesis sparsity*. Furthermore, the measurements and the unknown acoustic signals are processed in frequency domain hence, the short time Fourier transformation (STFT) is applied that yields *analysis sparsity*. Although the theoretical implications of the synthesis and analysis sparsity models are not within the scope of this work, we would like to bear in mind that when we talk about analysis sparsity, the transformation always plays a key role and requires further investigations (Tan and Fevotte, 2005).

The rest of the paper is organized as follows. The model-based sparse component analysis is stated in Section 2.1. The structured sparsity models underlying the perception and propagation of sound are studied

in Sections 2.2 and 2.3. We will see that the acoustic sparsity enables characterization of the compressive measurement mechanism associated with microphone array recordings and the auditory sparsity leads to incorporation of auditory-inspired models for speech signal reconstruction. In Section 3, we explain the premises of different computational approaches to model-based sparse recovery that exploit these structures for joint localization and separation of speech sources. Section 4 covers the details of the experimental analysis. Finally, the conclusions are drawn in Section 5.

2. Model-based Sparse Component Analysis

2.1. Problem Definition

We assume that the planar area of the room is discretized as a 2D grid with G uniform cells and N speakers are distributed such that each one of them occupies an exclusive cell and $\mathsf{N} \ll \mathsf{G}$[1]. The *spatial spectrum* of an acoustic scene is defined as a vector formed by concatenation of the spectrum from each cell. If no active source is positioned on a cell, the corresponding spectrum is a vector of all zeros. Hence, in a typical scenario that the number of sources is smaller than the number of cells, this vector is very sparse, i.e. the support of non-zero components is a small subset of its actual dimension.

We consider the frequency domain representation of the speech signals. The spatial spectrum of the acoustic scene is formed from concatenation of the spectral representations of the individual source signals on all grid cells defined as

$$\mathcal{S} = [\mathsf{S}_1^\top, \cdots, \mathsf{S}_\mathsf{g}^\top, \cdots, \mathsf{S}_\mathsf{G}^\top]^\top \in \mathbb{C}^{\mathsf{GF}\times 1} \tag{3}$$

where $\mathsf{S}_\mathsf{g} \in \mathbb{C}^{\mathsf{F}\times 1}$ denotes the spectral coefficients of the source located at cell g and F is the number of discrete frequencies; $.^\top$ stands for the transpose operator. We express the signal ensemble at the microphone array as a single vector

$$\mathcal{X} = [\mathsf{X}_1^\top, \cdots, \mathsf{X}_\mathsf{m}^\top, \cdots, \mathsf{X}_\mathsf{M}^\top]^\top \in \mathbb{C}^{\mathsf{MF}\times 1} \tag{4}$$

where each $\mathsf{X}_\mathsf{m} \in \mathbb{C}^{\mathsf{F}\times 1}$ denotes the spectral representation of recorded signal at microphone m.

The sparse vector $\mathcal{S}$ generates the microphone observations as $\mathcal{X} = \Phi\mathcal{S}$; $\Phi \in \mathbb{C}^{\mathsf{MF}\times\mathsf{GF}}$ is the acoustic measurement matrix consisting of the microphone array manifold vectors associated with the source located at each cell on the grid; as the matrix is typically very wide, i.e. $\mathsf{M} \ll \mathsf{G}$, we also refer to it as the *compressive* measurement matrix. The rigorous definition of Φ is expressed in Section 2.2 (Asaei et al., 2014).

In this paper, we assume that the acoustic measurement matrix is characterized based on the image model of multipath propagation. Hence, the spatio-spectral speech recovery problem amounts to sparse reconstruction of $\mathcal{S}$ given the compressed measurements in $\mathcal{X}$. Once $\mathcal{S}$ is estimated, the support of the high

[1]There is no algorithmic impediment to consider a 3D grid, but the number of required measurements (microphones) and the computational cost will be higher.

energy components indicates the source locations corresponding to the cells on the grid and the coefficients constitutes the individual source signals.

Contrary to the common SCA practice, this framework merges the two steps of source localization and separation through a joint objective of sparse reconstruction of the spatio-spectral representation of the acoustic scene. In addition, the underlying structures can be incorporated to achieve a more efficient recovery scheme. In this context, the objective of this paper is to identify and formulate the structures pertained to the propagation and perception of sound. Furthermore, the computational approaches to model-based sparse recovery are studied in order to compare and contrast their performance in terms of speech localization and separation using underdetermined recordings in a reverberant environment.

2.2. Sparsity Structures of Sound Propagation

To study the acoustic structures associated with multipath propagation of sound, the *image model* is considered (Allen and Berkley, 1979). The image model enables characterizing the acoustic field of a reverberant enclosure. It asserts that the sound field generated by a point source in a "shoe-box" room can be represented as the superposition of the original sound field and the ones generated by mirror images of the source with respect to the surrounding walls. Therefore, we can represent a path involving reflections by multiple straight line (free-space) paths connecting the microphone to the actual and virtual sources/images. This model is applicable to general polyhedra assuming that the reflective surfaces are piecewise planar (Borish, 1984).

Now lets assume that the grid is expanded over the boundaries of the enclosure; applying the image model to a source at a particular cell leads to a sparse lattice of virtual sources. Given the room geometry and the source location, a spatial structure underlies the multipath recordings which is known for each position inside the room. This structure is referred to as the *acoustic multipath sparsity structure* exploited in the spatial characterization of the microphone array recordings (Asaei et al., 2014).

To state it more precisely, we define $H_{mg}(f, \mu_m, \nu_g)$ as the multipath channel response at frequency f between a source g located at ν_g and the microphone m located at μ_m. Based on the image model, we can characterize the channel as a superposition of free-space Green's function stated as

$$H(f, \nu_g, \mu_m) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu_m - \nu_g^r\|^\alpha} \exp(-j2\pi f \frac{\|\mu_m - \nu_g^r\|}{c}) \tag{5}$$

where ν_g^r indicates the position of the r^{th} virtual sources with the reflective energy ratio of ι^r. Let $F' = \{f_1, \cdots, f_\pi\}$ represents a set of π frequencies, the compressive measurement matrix associated with the

microphone array for a broadband source positioned on the grid is obtained through

$$
\begin{aligned}
&H^{\mathrm{diag}}(F', \nu_n, \mu_m) = \\
&\mathrm{diag}\left([H(f_1, \nu_n, \mu_m)\ H(f_2, \nu_n, \mu_m) \cdots H(f_\pi, \nu_n, \mu_m)]\right), \\
&\Phi = \begin{bmatrix} H^{\mathrm{diag}}(F', \nu_1, \mu_1) & \cdots & H^{\mathrm{diag}}(F', \nu_N, \mu_1) \\ \vdots & \ddots & \vdots \\ H^{\mathrm{diag}}(F', \nu_1, \mu_M) & \cdots & H^{\mathrm{diag}}(F', \nu_N, \mu_M) \end{bmatrix}
\end{aligned} \tag{6}
$$

The spatial sparsity of the sources enables recovery of the signals $\mathcal{S}$ from the compressed measurements $\mathcal{X}$ using the compressive measurement model $\mathcal{X} = \Phi\mathcal{S}$. The unique map of the source location to the acoustic multipath structure guarantees the exact localization/separation (Dokmanić et al., 2013).

2.3. Sparsity Structures of Auditory Perception

The perceptual mechanism that governs the auditory system guide us to identify the structures underlying spectrographic speech representation. We rely on Bregman's framework for auditory scene analysis to characterize the auditory sparsity structures (Bregman, 1990; Wang and Brown, 2006); the theory asserts that the acoustic signal in the form of spectro-temporal scene is decomposed into a collection of segments, which are subsequently grouped to form coherent streams. The mechanism underlying grouping can be either a bottom-up process relying on the intrinsic structure of environmental sound or a top-down process based on prior knowledge on schema patterns (e.g. syllabic or linguistic). We investigate and formulate some of the bottom-up structures that can be exploited for sparse recovery. These structures are exhibited in the form of harmonicity, proximity and autoregressive dependency. The top-down structures often require language dependent processing modules and integration of the top-down structures in sparse recovery is an interesting subject of future studies.

2.3.1. Harmonicity

The set of acoustic components that are harmonically related (i.e. have frequencies which are integer multiples of the same fundamental frequency) tend to be grouped together. The harmonic structure is manifested in voiced speech as it 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 having harmonic sparsity structure. To state it more precisely, the support of $\mathcal{S}$ in sparse recovery must conform to the following model

$$
\mathcal{F}_H \triangleq \{k f_0 | 1 < k < K\} \tag{7}
$$

where the group of $\mathcal{F}_H$ frequencies correspond to a single cell; f_0 denotes the fundamental frequency and K is the number of harmonics to be grouped together.

2.3.2. Proximity

The set of adjacent frequencies tend to be grouped together while we are listening to the sound in a complex environment (Wang and Brown, 2006). Based on the principle of proximity, the closer acoustic components are in frequency/time, the greater is the tendency to group them into a single stream/source. To formulate this model for sparse reconstruction of $\mathcal{S}$, a block of $\mathtt{B}$ neighboring discrete frequencies is defined as

$$\mathcal{F}_{\mathtt{B}} \triangleq \{[\mathtt{f}_1, ..., \mathtt{f}_{\mathtt{B}}],\ \ldots,\ [\mathtt{f}_{\mathtt{F}-\mathtt{B}+1}, ..., \mathtt{f}_{\mathtt{F}}]\} \tag{8}$$

where all these frequencies are recovered as the spectral components of a single source corresponding to one cell of the grid. Hence, this block structure indicates that the spatial sparsity is the same at all adjacent frequencies.

Considering a single frequency component of multiple frames, the block structure boils down to *simultaneous sparsity* which is a special case of block structure when the manifold vectors corresponding to a group of sparse components are all the same. This structure enables capturing the sequential structure of the coefficients while processing multiple frames together. It is also referred to as multiple measurement model (Wipf and Rao, 2007).

2.3.3. Autoregressive Correlation

Another principle of structural grouping relies on amplitude modulation. Frequency components that exhibit the same modulation tend to be grouped together by our auditory system while analyzing the acoustic scene. This dependency can be captured through linear prediction by fitting an autoregressive (AR) model to find the optimal linear combination of a fixed-length history to predict the next component.

The AR inter-dependency implies a correlation among the block coefficients of each source, that we model using an AR process of order $\mathcal{R}$ characterized as the following

$$\mathcal{F}_{\mathtt{AR}} \triangleq [1, \beta_{\mathtt{g}}(1), \beta_{\mathtt{g}}(2), \ldots, \beta_{\mathtt{g}}(\mathcal{R})] \tag{9}$$

where $\beta_{\mathtt{g}} \in (-1, 1)$ denotes the AR coefficients. The sources $\mathcal{S}_{\mathtt{g}}$ are mutually independent, but each source satisfies an AR model as

$$\mathcal{S}_{\mathtt{g}}(\mathfrak{b}) = \mathcal{F}_{\mathtt{AR}}\,[\mathfrak{u}(\mathfrak{b}), \mathcal{S}_{\mathtt{g}}(\mathfrak{b}-1),\ \ldots,\ \mathcal{S}_{\mathtt{g}}(\mathfrak{b}-\mathcal{R})]^{\top} \tag{10}$$

where $\mathfrak{u}(\mathfrak{b})$ denotes the input sequence and $\mathfrak{b}$ corresponds to the indices of adjacent frequencies or a single frequency at multiple frames. We will see how the AR correlation is taken into account in sparse reconstruction in Section 3.2.4 and study the empirical evaluation results in Section 4.3.2.

In this section, we explained about the sparsity structures underlying the spatio-spectral representation of convolutive speech mixtures recorded by an array of microphone. In the following section, we overview some of the computational methods to sparse recovery and explain how the spectrographic sparsity structures can be incorporated for sparse reconstruction of the individual speech sources.

3. Model-based Sparse Recovery

Defining a set $\mathbb{M}$ as the union of all vectors with a particular support structure, estimation of the structured sparse coefficient of vector $\mathcal{S}$ from the microphone recordings $\mathcal{X}$ can be expressed 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 the counting function $\|.\|_0 : \mathbb{R}^{\mathsf{GF}} \rightarrow \mathbb{N}$ returns the number of non-zero components in its argument.

3.1. Computational Methods

The major classes of computational techniques for solving sparse reconstruction problems rely on combinatorial, geometric or probabilistic frameworks (Tropp and Wright, 2010).

Combinatorial optimization: The nonzero components of $\mathcal{S}$ are estimated through an iterative procedure by modifying one or several coefficients chosen to yield a substantial improvement in quality of the estimated signal. In this paper, we consider the algorithms based on iterative hard thresholding (IHT) (Blumensath and Davies, 2009; Kyrillidis and Cevher, 2011) as well the agglomorative greedy approach of orthogonal matching pursuit (OMP) to evaluate the combinatorial method to speech localization and separation incorporating the sparsity structures underlying spectrographic coefficients (Gribonval and Bacry, 2003).

Convex relaxation: The counting function in (11) is replaced with a sparsity inducing convex norm defined for the convex hull of the union of sparse vectors $\mathcal{S}$. Therefore, a convex objective is obtained which can be solved using convex optimization (McCoy et al., 2014). We consider a particular extension of basis pursuit algorithm which relies on L_1-norm (defined as the sum of absolute values of the vector elements) relaxation of the counting objective and $\mathsf{L}_1\mathsf{L}_2$ formulation of group sparse recovery (Berg and Friedlander, 2008).

Bayesian methods: A prior distribution is associated with $\mathcal{S}$ with sparsity inducing hyperparameters and a maximum a posteriori estimation is obtained given the distant microphone measurements, $\mathcal{X}$. We consider in particular the Multiple measurement FOCal Underdetermined System Solver,(MFOCUSS) (Cotter et al., 2005) and the sparse Bayesian learning, (SBL) framework (Wipf and Rao, 2007; Zhang and Rao, 2011, 2012).

3.2. Sparse Recovery Algorithms

In this section, the model-based sparse recovery algorithms that are used for our empirical study in Section 4 are briefly explained.

3.2.1. IHT

The iterative hard thresholding (IHT) offers a simple yet effective approach to estimate the sparse vectors. We use the algorithm proposed in (Kyrillidis and Cevher, 2011) which is an accelerated scheme for hard thresholding methods with the following recursion

$$\begin{aligned} &\hat{\mathcal{S}}^0 = 0, \quad \mathcal{R}^{\mathfrak{i}} = \mathcal{X} - \Phi\hat{\mathcal{S}}^{\mathfrak{i}} \\ &\hat{\mathcal{S}}^{\mathfrak{i}+1} = \mathcal{M}^{\mathcal{F}_\cdot}\left(\hat{\mathcal{S}}^{\mathfrak{i}} + \kappa\Phi^{\mathsf{T}}\mathcal{R}^{\mathfrak{i}}\right) \end{aligned} \tag{12}$$

where the step-size κ is the Lipschitz gradient constant to guarantee the fastest convergence speed; $\mathfrak{i}$ denotes the iteration variable. To incorporate for the underlying structure of the sparse coefficients, the model approximation operator $\mathcal{M}^{\mathcal{F}_\cdot}$ is defined as reweighting and thresholding the energy of the components of $\hat{\mathcal{S}}$ with either $\mathcal{F}_{\mathrm{H}}$ or $\mathcal{F}_{\mathrm{B}}$ structures defined in (7) and (8).

In Sections 4.3 and 4.4, this method is evaluated for sparse recovery considering the proximity/block structure underlying the spectral and temporal components as well as the harmonicity of the spectral components of $\mathcal{S}$.

3.2.2. $\mathrm{L}_1\mathrm{L}_2$

An extension of the basis pursuit algorithm for group sparse recovery can be used for model-based sparse recovery. The optimization problem to recover the group sparse coefficients $\hat{\mathcal{S}}$ is formulated as follows:

$$\begin{aligned} \hat{\mathcal{S}} =& \underset{\mathcal{S}}{\operatorname{argmin}} \{\|\mathcal{S}\|_{\mathrm{L}_1\mathrm{L}_2} \text{ s.t. } \mathcal{X} = \Phi\mathcal{S}\}, \\ \|\mathcal{S}\|_{\mathrm{L}_1\mathrm{L}_2} =& \sum_{g=1}^{G} \sqrt{\sum_{b=1}^{\mathfrak{n}^{\mathcal{F}_\cdot}} (\mathcal{S}_g(b))^2} \end{aligned} \tag{13}$$

where the number of group members $\mathfrak{n}^{\mathcal{F}_\cdot}$ is determined by the underlying structure of the grouped components. From the formulation of (13), we can see that the inner sum calculates the L_2 norm of the group components whereas the outer sum calculates the L_1 norm of the groups. Hence, the $\mathrm{L}_1\mathrm{L}_2$ objective imposes sparsity at a group level while the coefficients for each individual group can be dense.

Like IHT, this method is also evaluated for sparse recovery considering the proximity/block structure underlying the spectral and temporal components as well as the harmonicity of the spectral components of $\mathcal{S}$ in Sections 4.3 and 4.4.

3.2.3. MFOCUSS

The Multiple measurement FOCal Underdetermined System Solver (MFOCUSS) algorithm employs an L_ρ-norm like diversity measure; it is formulated to recover the sparse matrices with a simultaneous sparsity structure (Section 2.3.2) via

$$\hat{\mathsf{S}} = \arg\min_{\mathsf{S}} \|\mathsf{X} - \Phi\mathsf{S}\|_{\mathcal{F}}^2 + \lambda \sum_{r=1}^{G} (\|\mathsf{S}_{r\cdot}\|)^\rho \tag{14}$$

where S and X are formed from the spectral vectors $\mathcal{S}$ and $\mathcal{X}$ of multiple frames: $\mathsf{S} = [\mathcal{S}_1, \cdots \mathcal{S}_L] \in \mathbb{C}^{G \times L}$ and $\mathsf{X} = [\mathcal{X}_1, \cdots \mathcal{X}_L] \in \mathbb{C}^{M \times L}$ considering a single frequency. The $\rho \in [0, 1]$ is a user-defined parameter[2], to prevent models with many nonzero rows.

The factored-gradient iterative algorithm to minimize the objective stated in (14) is summarized as

$$\begin{aligned} & c^i_{.l} = \|X^i_{.l}\| = \left(\sum_{l=1}^{L} (X^i_{.l})^2\right)^{1/2}, \forall l \in \{1, \cdots, L\} \\ & \mathcal{W}^{i+1} = \text{diag}\,((c^i_{.l})^{1-\rho/2}), \\ & Q^{i+1} = (\Phi \mathcal{W}^{i+1})^{\dagger} X \\ & \hat{S}^{i+1} = \mathcal{W}^{i+1} Q^{i+1} \end{aligned} \tag{15}$$

where the notion $X_{r.}$ refers to the r^{th} row of matrix X and $(.)^{\dagger}$ stands for the pseudo-inverse operation. The update rules stated in (15) are guaranteed to converge monotonically to a local minimum of the objective (Cotter et al., 2005).

The MFOCUSS objective function (14) can be derived using a generalized Gaussian prior on the row norms of spectro-temporal components. Thereby, it admits the Bayesian analysis on maximum a posteriori (MAP) estimation of the sparse coefficients given the observation $\mathcal{X}$ and linear measurement matrix Φ (Saab et al., 2007). The difficulty with this procedure is twofold: either the prior is not sufficiently sparsity-inducing (Cevher, 2009) and the MAP estimates sometimes fail to be sparse enough, or a combinatorial number of suboptimal local solutions must be dealt with if a highly sparse prior is chosen. Hence, alternative Bayesian strategy based on sparse Bayesian learning framework are developed (Wipf and Rao, 2007) that is considered in this paper.

In Sections 4.3 and 4.4, the MFOCUSS method is evaluated for sparse recovery considering the proximity/block structure underlying the temporal components of $\mathcal{S}$.

3.2.4. BSBL

This algorithm is an extension of Sparse Bayesian Learning (SBL) incorporating block sparsity thus denoted by BSBL. The empirical Bayesian sparse learning method (Zhang and Rao, 2012) assumes that the component of $\mathcal{S}$ are mutually independent and each has a Gaussian distribution. Hence, the prior distribution for $\mathcal{S}$ has a multivariate normal distribution. If we assume that the components of $\mathcal{S}$ consist in blocks of B adjacent coefficients (e.g. multiple frequencies or frames of a single source), the joint distribution of each block is $p(S_g; \gamma_g, \beta_g) \sim \mathcal{N}(0, \gamma_g \mathcal{B}_g)$ where γ_g is a non-negative hyper-parameter controlling the block-sparsity of $\mathcal{S}$; $\gamma_g = 0$ indicates that the associated block in $\hat{\mathcal{S}}$ is zeros or no source is located on the corresponding cell. The $\mathcal{B}_g \in \mathbb{R}^{B \times B}$ is a positive definite matrix that captures the correlation structure of $\mathcal{S}_g$. For example to account for the AR dependency model stated in (10), the covariance matrix $\mathcal{B}_g$ of each

[2]A typical value for speech-specific task can be selected as 0.6 (Saab et al., 2007) or 0.8 as suggested by the authors of the MFOCUSS algorithm (Cotter et al., 2005).

source has the following Toeplitz structure

$$\mathcal{B}_g = \begin{bmatrix} 1 & \beta_g & \dots & \beta_g^{B-1} \\ \beta_g & 1 & \dots & \beta_g^{B-2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_g^{B-1} & \beta_g^{B-2} & \dots & 1 \end{bmatrix} \tag{16}$$

Under the assumption that the blocks are mutually uncorrelated, the prior for $\mathcal{S}$ is $p(\mathcal{S}; \gamma_g, \mathcal{B}_g, \forall g)$ given by

$$\mathcal{N}(0, \Sigma_0),\ \Sigma_0 = \text{diag}\left([\gamma_1 \mathcal{B}_1 \dots \gamma_G \mathcal{B}_G]\right)$$

Assuming the Gaussian likelihood for the block sparse model as $p(\mathcal{X}|\mathcal{S}; \sigma^2) \sim \mathcal{N}(\Phi\mathcal{S}, \sigma^2 I)$ and applying the Bayes rule, we obtain the posterior density of $\mathcal{S}$, which is also Gaussian, thus $p(\mathcal{S}|\mathcal{X}; \sigma^2, \{\gamma_g, \mathcal{B}_g\}_{g=1}^G) \sim \mathcal{N}(\mu_s, \Sigma_s)$ where

$$\Sigma_s = (\Sigma_0^{-1} + \frac{1}{\sigma^2}\Phi^T S)^{-1}\mathcal{X}$$

Having all the hyperparameters σ^2, γ_g, $\mathcal{B}_g$, the MAP estimate of $\mathcal{S}$ is given by the mean defined as

$$\hat{\mathcal{S}} \triangleq \mu_s = \Sigma_0\Phi^T(\sigma^2 I + \Phi\Sigma_0\Phi^T)^{-1}\mathcal{X} \tag{17}$$

The hyperparameters can be efficiently estimated through the EM procedure (Zhang and Rao, 2012). We will see in Section 4.3 that the AR model can be estimated offline for the specific task of speech localization.

3.2.5. TMSBL

This algorithm exploits the temporal structure in multiple measurements sparse Bayesian learning thus coined as TMSBL. The multiple measurements share simultaneous sparsity as the support of sparse coefficients are the same for all measurements. To exploit the temporal structure in a form of simultaneous sparsity, the matrix of measurements, X obtained by concatenation of the coefficients along multiple consecutive frames is transformed to a vector through

$$\begin{aligned} \mathcal{X} &= \text{vec}(X) \in \mathbb{C}^{ML\times 1}, \\ \Psi &= \Phi \otimes I_L, \\ \mathcal{S} &= \text{vec}(S) \in \mathbb{C}^{GL\times 1} \end{aligned}$$

The transformation leads to $\mathcal{X} = \Psi\mathcal{S}$ with a block structure obtained as

$$\mathcal{X} = [\phi_1 \otimes I_L, ..., \phi_G \otimes I_L][S_1^T, ..., S_G^T]^T = \sum_{g=1}^{G} (\phi_g \otimes I_L)S_g$$

where $S_g \in \mathbb{C}^{L\times 1}$ is the g^{th} block in $\mathcal{S}$ and $S_g = S_{g.}^T$. Having N nonzero rows in S means N nonzero blocks in $\mathcal{S}$. Thus, $\mathcal{S}$ is block-sparse. Given the vectorized formulation, the rest of the procedure is similar to BSBL. This algorithm is referred to as temporal multiple measurement SBL or TMSBL (Wipf and

Rao, 2007) that is used for evaluation of speech localization and separation incorporating the temporal proximity/block sparsity structure taking into account the AR correlation of the source-specific coefficients. If the autoregressive correlation of the coefficients is completely ignored, the basic multiple measurement SBL (MSBL) algorithm (Zhang and Rao, 2011) is used for evaluation in sections 4.3 and 4.4.

4. Experimental Analysis

The experiments are conducted to evaluate the performance of the model-based sparse recovery algorithms in terms of source localization, speech separation as well as computational cost. More specifically, the goal of our experimental analysis is to study the performance of model-based sparse recovery of multiparty speech signals considering the following issues:

- ◇ Sparsity structures underlying spectrographic speech representations.
- ◇ Different computational methods exploiting these structures for sparse reconstruction.
- ◇ Robustness to noise and mismatch in characterization of a reverberant acoustic field.
- ◇ Sensitivity to coherence of the measurements entangled with the geometry of the microphone array.

4.1. Acoustic and Analysis Setup

The speech utterances are taken from the Wall Street Journal (WSJ) corpus (Lincoln et al., 2005). The WSJ corpus consists of 20000 words from read Wall Street Journal sentences. The sentences are read by a range of speakers (34 in total) with varying accents (including a number of non-native English speakers). This database provides a broad phonetic space for speech separation evaluation. The utterances used for the experiments presented here are selected arbitrarily. The length of the signal used for speech separation evaluation is about 2.5 min.

The overlapping speech was synthesized by mixing sentences from the test set with interfering sentences from the development set. The interference files are normalized in speaker localization evaluation and scaled to yield $-20\,\mathrm{dB}$ source to interference ratio (SIR) for speech separation experiments.

The planar area of a room with dimension $3\mathrm{m}\times3\mathrm{m}\times3\mathrm{m}$ is divided into cells with $50\,\mathrm{cm}$ spacing. The acoustic recordings are synthesized for two four-channel microphone arrays. The scenarios include *random* and *compact* placement of the microphones at different noise levels. The random placement has a large aperture and it is motivated by the theoretical insights on the performance of sparse recovery algorithms and the importance of random and incoherent measurements. The data collection setup is depicted in Figure 1. The center of the compact microphone array is located at the room center. The number of sources is $\mathsf{N} = 5$ for speech separation experiments. The source-microphone configuration has the topology depicted in Figure 1. For the source localization experiments, the number of sources varies as $\mathsf{N} \in \{5, \cdots, 10\}$ and all the combinatorial unique topologies are considered while the microphone locations are fixed.

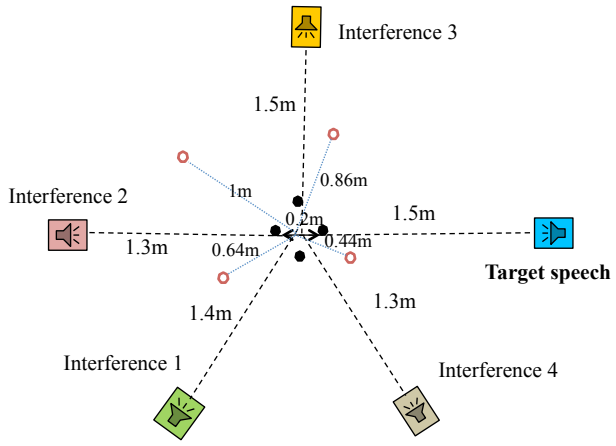

Figure 1: Overhead view of the room set-up for uniform (black dots) and random microphone array (white dots). The angle of the random microphones (white dots) located in the first to fourth quadrants are 80, 140, 200 and 340 degrees with respect to the horizontal axis.

The room impulse responses are generated with the image model (Allen and Berkley, 1979) using intra-sample interpolation, up to 15^{th} order reflections and omni-directional microphones for a room reverberation time equal to 200 ms; the speed of sound is assumed to be $\mathbf{c} = 343$ m/s. The speech signals are recorded at 16 kHz sampling frequency and the spectro-temporal representation for source separation is obtained by windowing the signal in 250 ms frames using Hann function with 50% overlapping for applying the short time Fourier transform. The number of FFT points is 4096. This relatively long window is selected based on the reverberation time and its importance in modeling the acoustic field using image method (Asaei et al., 2014).

4.2. Coherence of the Measurements

The compressive acoustic measurement matrix is forced by the laws of physics governing the propagation of sound and it is a function of source and microphone placements, enclosure properties as well as the emanating signal frequency. Exploiting the broadband frequencies of speech provides us a handle to select a microphone array measurement matrix meeting the desired spectral properties.

The theory of sparse signal reconstruction requires the measurements to be incoherent. This condition can partly be fulfilled by designing a microphones array geometry to minimize the coherence in spatial sampling which is out of the scope of this paper. Hence, we rely on an intuitive design based on randomness and distant positioning of the microphones.

To analyze the measurement matrix for the broadband speech spectrum, we compute the condition number of Φ for the random array setup at different frequency bands; the condition number is defined as the ratio of the maximum to the minimum singular values. The results are illustrated in Figure 2. The results suggest subband processing in order to increase the efficiency of sparse recovery algorithms and in particular it demonstrates the importance of high frequencies to obtain a less coherent measurement matrix.

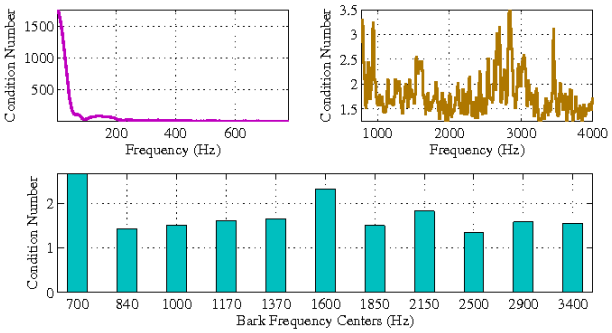

Figure 2: Condition number computed per frequency band for the ad hoc microphone array measurement matrix. The upper plots correspond to the linear frequencies whereas the lower bars are centered at Bark frequencies. It is evident that the high frequencies results in less correlation and the measurement matrix exhibits better spectral properties.

4.3. Source Localization Performance

Recall from 2.1 the sparse vector $\mathcal{S}$ consisting of the source-specific coefficients positioned on the grid cells (cf. (3)). This vector is reconstructed using sparse recovery methods elaborated in 3.2. The support of the high energy coefficients indicates the source locations. In this section, we study the localization performance using different sparse recovery algorithms.

4.3.1. Evaluation Criterion

The localization accuracy is quantified as the ratio of the number of sources found at the right locations divided by the total number of sources. For example, if there are 5 simultaneous sources, the coefficients of $\mathcal{S}$ are sorted based on their energy and the top five are considered for localization; if only 4 of the actual locations are present in the top five high energy cells, the localization accuracy is 80%. The localization accuracy is measured for each frame independently. The length of the speech signal for frame-based localization is about two seconds and the average of the frame-based accuracies is reported.

The expected localization performance is quantified by averaging the results over an exhaustive set of unique configurations consisting of $\mathsf{N} \in \{5-10\}$ sources. The exhaustive evaluation is necessary to quantify the expected performance as the deterministic bounds of sparse recovery algorithms are too pessimistic and they are particularly derived for the worst case scenario which is not likely to occur (Asaei, 2013)[3].

[3]More specifically, the deterministic performance bounds are guaranteed when the number of measurements are in order of $\mathsf{C}\,\mathsf{N}\log(\mathsf{G}/\mathsf{N})$ where C is a constant more than 2. Considering $\mathsf{N} = 5$ sources on a grid of $\mathsf{G} = 36$ cells, it indicates that the number of required measurements is at least eight. On the other hand, the experimental evaluation shows that only four microphones is enough for localization and separation. This observation can be justified through two reasons: (1) the deterministic bounds are pessimistic and derived according to the worst case scenarios and (2) the number of non-zero coefficients in each time-frequency point is often associated with only one source due to the disjointness property of overlapping speech sources (Asaei et al., 2010, 2014) hence, $\mathsf{N} = 1$ and four measurements are essential based on the theory.

4.3.2. Average Autoregressive Model

In this section, we study an average *temporal* AR model for speech signal and exploit it for source localization. To obtain the average AR model, the spectro-temporal representation of speech signal of length 10 min is first obtained using short time Fourier transform. Next, starting from the first frame and shifting by one frame, similar frequency coefficients of all consecutive frames (segments) of length 13 are used to learn an AR model of order 13. Finally, the AR coefficients of all segments at all discrete frequencies are averaged. Figure 3 illustrates the average temporal AR coefficients. The AR model is estimated from the voiced speech segments using an energy-based voice activity detector. The speech signal is recorded in clean condition.

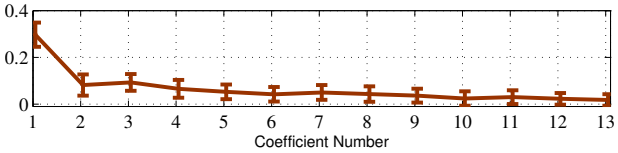

Figure 3: The 13-order average *temporal* AR coefficients estimated for 10 min speech. The cross lines show the variance of the estimates.

The first-order AR coefficient is estimated as 0.3. We can observe that a first-order average AR model is a good approximation to capture the temporal correlation across multiple frames. This correlation structure is applicable for general speech sources and it can be attributed to the shape of human's vocal tract.

The autoregressive correlation of the sparse coefficients is considered in the framework of sparse Bayesian learning. The TMSBL method is developed to learn the AR parameters during the optimization procedure (Zhang and Rao, 2011). However, the algorithm is expensive in terms of computational cost and we empirically observe that it does not yield better performance than MSBL (Wipf and Rao, 2007) where the AR correlation is completely ignored. Hence, we use the average first-order AR coefficient estimated above as an input to the TMSBL algorithm alleviating the requirement for frame-wise AR estimation. By doing so, the algorithm performs both faster and more accurate. The gain in accuracy is more noticeable in compact microphone array scenario. Figure 4 illustrates the improvement obtained using an average AR model. We further estimated an average AR model per frequency, but it did not yield a higher localization accuracy.

In addition to the temporal correlation, we investigate the spectral correlation of the adjacent spectral coefficients. To estimate the AR coefficients, the frequency band is split into blocks of size 16 and processed independently. Figure 5 demonstrates the average *spectral* AR model for 10 min speech signal. The first-order coefficient is estimated as 0.45. Similar to the temporal AR model, we verify that modeling the blocks as a first-order AR process can be sufficient to incorporate the intra-block correlation.

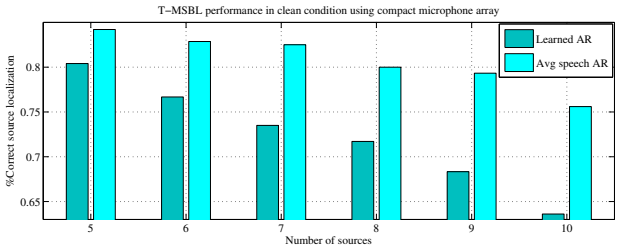

Figure 4: TMSBL localization accuracy using $\mathsf{L} = 13$ frames. The light (blue color) bars correspond to the performance when the average AR model parameter is fixed, i.e. first-order model with a coefficient equal to 0.3; the algorithm does not update the correlation coefficient. The dark (blue color) bars shows the performance when the AR correlation coefficients is learned by the algorithm. We can see that it is beneficial to estimate the average AR coefficients offline and use the first-order coefficient.

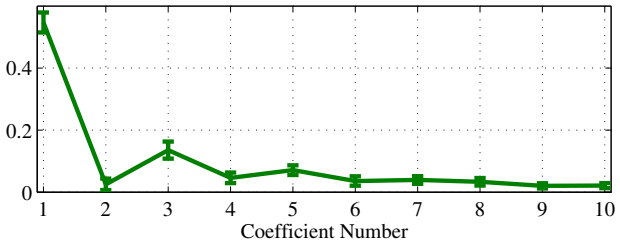

Figure 5: The 10-order average *spectral* AR coefficients estimated for 10 min speech signal. The cross lines illustrate the variance of the estimates.

4.3.3. Localization Accuracy

We first study the localization accuracy exploiting the proximity structure associated with simultaneous recovery of multiple frames. For this experiment, L frames of microphone array recordings are considered. The localization performance of all sparse recovery algorithms improves with $\mathsf{L} > 1$. We can empirically observe that $\mathsf{L} \in \{10, ..., 15\}$ is a good choice to maximize the performance of different algorithms; the localization results for $\mathsf{L} = 13$ in clean and noisy conditions are illustrated in Figure 6. The noisy condition includes both the effect of additive noise as well as mismatch in acoustic modeling. The SNR of noisy scenario is 10 dB after adding white Gaussian noise and the absorption coefficients of the forward model are 25% deviated from the actual parameters.

Although the number of microphones is limited to 4, we can localize up to 9 sources in noisy condition and uncertainties in acoustic channel modeling. Given that each frame is 250 ms and the frames are 50% overlapping, $\mathsf{L} = 13$ frames indicates that 1.6 seconds of speech recordings yields perfect localization from highly incomplete spatial information of only 4 microphones. There is no noticeable difference between IHT, $\mathsf{L}_1\mathsf{L}_2$ and MFOCUSS algorithms. However, it is clear that the large random setup leads to significant improvement in source localization. We can not explain why sparse Bayesian learning framework does not perform comparable to the other methods.

The frame-based localization accuracy exploiting spectral structures are illustrated in Figure 7. In this

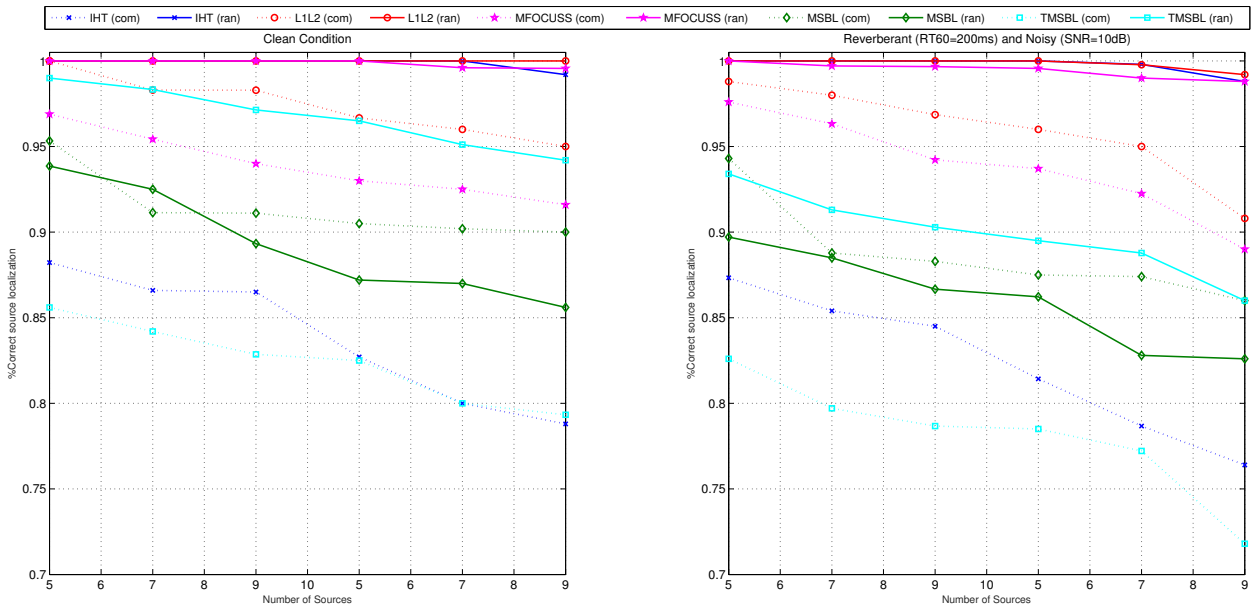

Figure 6: Expected localization accuracy evaluated for 5-10 simultaneous sources exploiting *temporal* structured sparse recovery. The expected performance is quantified by averaging the results over an exhaustive set of unique configurations consisting of $\mathrm{N} \in \{5-10\}$ sources.

experiment, we choose the size of each block as $\mathrm{B} = 16$. The algorithms are run for the stopping threshold fixed to 1e-2 and the maximum iteration of 150. The value of p is selected 0.8 as suggested by the authors of MFOCUSS (Cotter et al., 2005). We evaluated other values as suggested in (Saab et al., 2007), but no difference was obtained.

We can see that exploiting the frequency structure yields high accuracy in localization of multiple simultaneous sources using only one frame. The $\mathrm{L_1L_2}$, MFOCUSS and IHT are the best performing algorithms. Note that the large and random placement of the microphones is essential to achieve highly accurate localization results. The number of microphones is only 4 whereas we can localize up to 9 sources with 95% accuracy. These results are beyond the deterministic performance bounds. The orthogonality or disjointness of spectrographic speech signals is a key property to achieve this performance (Asaei et al., 2014).

Exploiting the temporal structures require the sources to be stationary. Comparing the results illustrated in Figures 7 and 6 shows that we can relax this requirement and rely on spectral sparsity models to achieve highly accurate localization. The results of harmonic sparse recovery were comparable to the block-sparse recovery demonstrated in Figure 7 thus, they are not further elaborated here.

4.4. Speech Separation Performance

In this section, the quality of speech separation using convolutive mixtures provided by four microphones is evaluated. The number of simultaneous sources is five. The simulated scenario is depicted in Figure 1. The criteria for evaluation are source-to-interference ratio (SIR) (Vincent et al., 2006), source-to-noise ratio

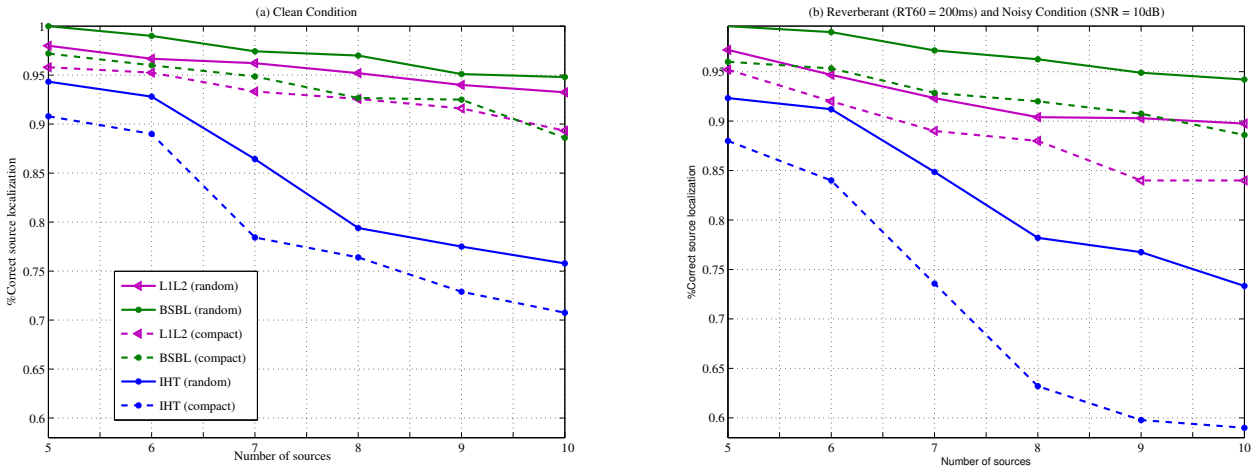

Figure 7: Localization accuracy evaluated for 5–10 simultaneous sources exploiting *spectral* structured sparse recovery. The expected performance is quantified by averaging the results over an exhaustive set of unique configurations consisting of $\mathsf{N} \in \{5-10\}$ sources.

(SNR), perceptual speech quality (PESQ) (ITU-T, 2001) and weighted spectral slope (WSS) (Persia et al., 2008). This wide range of objective measures is considered to enable thorough evaluation and analysis of the six different computational methods stated in Section 3 in different scenarios of noise and microphone array geometries.

4.4.1. Evaluation Criteria

The quality of the separated speech is measured according to the performance metrics proposed in (Vincent et al., 2006). They rely on decomposition of the estimated signal into distinct components of target, interference and noise. The decomposition is performed by orthogonal projection of the estimated source signal onto the subspace spanned by specific components of the true source, interference and noise vectors accordingly. For example, if the sources are mutually orthogonal and a time-invariant filtered version of estimated signal of the true source is denoted by $\tilde{\mathsf{S}}_{\text{target}}$, the error due to interferences is defined as $\mathsf{e}_{\text{interf}} = \sum_{\mathsf{j}=1}^{\mathsf{N}} \langle \tilde{\mathsf{S}}_{\text{target}}, \mathsf{S}_{\mathsf{j}} \rangle \mathsf{S}_{\mathsf{j}} / \|\mathsf{S}_{\mathsf{j}}\|$ where $\langle ., . \rangle$ stands for the inner product. The subspace of a filtered version of the true signal is characterized by the true signal vector and all of its delayed versions upto the filter length; the filter length is chosen to be 512 for our evaluations. Note that to quantify these objective measures we need to have access to the original source signal and interference signals. The recipe of (Vincent et al., 2006) also requires to input the noise signal. We did not use any input noise and the spatial (or filtering) distortion component is used instead; this spatial error is obtained as the difference between true source signal and the filtered version of the estimated source signal.

The relative amount of these four terms is computed numerically as the energy ratios expressed in decibels. The objective measures that we evaluate here include source-to-interference ratio (SIR) and source-

to-noise ratio (SNR), defined as

$$\mathrm{SIR} := 10\log_{10}\frac{\|\mathrm{S}_{\mathrm{target}}\|_2^2}{\|\mathrm{e}_{\mathrm{interf}}\|_2^2}, \quad \mathrm{SNR} := 10\log_{10}\frac{\|\mathrm{S}_{\mathrm{target}}\|_2^2}{\|\mathrm{e}_{\mathrm{noise}}\|_2^2} \tag{18}$$

In addition, we evaluate the perceptual speech quality in terms of PESQ (ITU-T, 2001). The PESQ measure was found to have high correlations with subjective rating of overall quality of recovered speech signals. The databases used during training included various kinds of noises, including babble, so it is a valid measure to evaluate the results of speech separation.

The PESQ quantification procedure is specifically designed to predict the perceived quality that would be given to a processed speech signal in a subjective listening test. The original and the recovered speech signals are pre-processed to compensate for the time-delay, frequency response and gain that are irrelevant for a subjective score. Then, the signals are transformed to a short-time spectrum with perceptual frequency (Bark) and intensity (loudness) scales. In particular, the calculation of the short-time loudness spectrum considers the effect of simultaneous masking in the human auditory system (ITU-T, 2001).

The differences in the internal short-time representations between both signals are computed; the (signed) difference over time is called disturbance density. Two disturbance vectors, one for positive and one for negative differences are computed and the values of both disturbance vectors are aggregated over various time spans with different norms and combined to a raw PESQ score in the range [-0.5, 4.5]. Following the standardization of PESQ as Recommendation P.862, the standardization sector of the International Telecommunications Union (ITU-T) created a simple mapping function to allow comparisons between the raw PESQ score and the traditional MOS (Mean Opinion Score) (ITU-T, 2003).

Finally, the spectral distortion can be quantified in terms of weighted spectral slope distance measure (WSS) (Persia et al., 2008). This distance measure is based on comparison of smoothed spectra from the clean and distorted speech samples. The smoothed spectra can be obtained from either linear prediction analysis, or filter-bank analysis. One implementation of WSS can be defined as follows,

$$\mathrm{d}_{WSS} = \frac{1}{\mathrm{L}+1}\sum_{\mathrm{l}=0}^{\mathrm{L}}\frac{\sum_{\mathrm{f}=1}^{\mathrm{B}} W(\mathrm{f},\mathrm{l})(\mathrm{S}(\mathrm{f},\mathrm{l})-\hat{\mathrm{S}}(\mathrm{f},\mathrm{l}))^2}{\sum_{\mathrm{f}=1}^{\mathrm{B}} W(\mathrm{f},\mathrm{l})} \tag{19}$$

where B is the number of bands, L is the total number of frames, and $\mathrm{S}(\mathrm{f},\mathrm{l})$ and $\hat{\mathrm{S}}(\mathrm{f},\mathrm{l})$ are the spectral slopes (typically the spectral differences between neighboring bands) of the f^{th} band in the l^{th} frame for clean and recovered speech, respectively. $W(\mathrm{f},\mathrm{l})$ are weights, which can be calculated described in (Klatt, 1982). The WSS of clean speech signal is 0.

4.4.2. Speech Separation Quality

The speech separation performance exploiting temporal structures in clean and noisy conditions are summarized in Table 1. The noisy condition includes both the effect of additive noise as well as mismatch in

acoustic modeling. The SNR of noisy scenario is 10 dB after adding white Gaussian noise and the absorption coefficients of the forward model are 25% deviated from the actual parameters. The number of frames, L is set to be 2 as it yields the best performance. The simulations are done in MATLAB 7.14 on 4 Core(TM) i7 CPU @ 2.8-GHz, 11.8-GiB RAM PC and the absolute elapsed times (in seconds) are measured for each algorithm.

The evaluation of the speech separation quality exploiting temporal proximity and autoregressive correlation for simultaneous sparse recovery shows that the MFOCUSS algorithm performs the best in speech reconstruction. The amount of interference suppression is higher while the signal is less distorted and it preserves better perceptual quality. We can also see that if the coherence of the measurements is small (i.e. large and random microphone array), the algorithm is not very sensitive to noise and mismatch in acoustic model characterization.

If the measurements are highly correlated which is the case in a compact uniform microphone array setup, the greedy methods such IHT achieves higher interference suppression whereas the sparse Bayesian learning approach performs the best in terms of quality and distortion. Although, in clean condition, the interference suppression performance drops by 2 dB, this degradation is more pronounced in noisy condition; for instance, the SIR of MFOCUSS is decreased by 10 dB. The discrepancy between the performance using compact microphone recording and a large random array is more noticeable for sparse Bayesian learning whereas it is less exhibited using greedy methods in particular OMP; for example, the SIR using MSBL is reduced by 10 dB in clean condition while it drops by less than 1 dB using OMP.

In addition, we can see that MSBL outperforms the BSBL framework simply by ignoring the autoregressive correlation. In terms of computational cost, OMP is the most efficient algorithm and it offers computational speed up to 25 times faster than MFOCUSS and 176 times faster than the group sparse basis pursuit using L_1L_2 objective.

Furthermore, we evaluate the quality of speech separation exploiting spectral structures. The formulation of MFOCUSS and MSBL are developed for simultaneous sparsity model and it can not exploit the spectral block sparse structure. Hence, our experiments are confined to L_1L_2, IHT, OMP and BSBL methods. The block-size, B is equal to 2 as it yields the best results. The results are summarized in Tables 2. It is clear that greedy sparse recovery is the best method to achieve high interference suppression whereas in terms of quality and distortion, L_1L_2 and BSBL perform better. In the clean scenario where there is no noise and mismatch in acoustic model, L_1L_2 tends to perform better. However, it demonstrates higher sensitivity to noise and BSBL achieves the best results under mismatch and acoustic model uncertainties. It may be noted that the BSBL recovery performance in spectral grouping is far better than exploiting the temporal proximity and autoregressive correlation, particularly in noisy condition.

Comparing the results with simultaneous sparse reconstruction exploiting temporal structures, demonstrates the effectiveness of the spectral sparse recovery approach as the perceived quality in terms of PESQ

Table 1: Speech separation quality in clean and noisy conditions exploiting temporal proximity and AR correlation models. The SNR of noisy data is 10 dB by adding white Gaussian noise and the absorption coefficients required for characterizing the acoustic by the image model are 25% deviated from the actual parameters. "LR" corresponds to the large random placement of the microphones and "UC" indicates the uniform compact microphone array.

	Temporal proximity & AR - Clean						Temporal proximity & AR - Noisy					
LR	L_1L_2	IHT	OMP	BSBL	MSBL	MFOC	L_1L_2	IHT	OMP	BSBL	MSBL	MFOC
SIR	14.52	14.24	12.47	12.15	24.16	**26.4**	10.39	12.45	10.63	11.13	**15.37**	14.71
SNR	4.99	4.66	4.54	3.53	26.35	**28.3**	4.61	4.32	4.32	3.14	15.86	**18.35**
PESQ	1.80	1.69	1.80	2.80	1.72	**3.09**	1.73	1.73	1.74	1.8	2.09	**2.19**
WSS	62.59	93.88	86.50	69.23	40.39	**34.25**	70.62	74.94	74.13	68.67	69.89	**63.70**
UC	L_1L_2	IHT	OMP	BSBL	MSBL	MFOC	L_1L_2	IHT	OMP	BSBL	MSBL	MFOC
SIR	14.09	14.16	12.72	8.24	14.57	**24.12**	8.26	**9.78**	9.07	6.47	4.2	4.08
SNR	4.75	4.23	3.62	2.67	15.7	**23.41**	**4.29**	4.00	3.28	2.22	3.99	3.22
PESQ	1.86	1.79	1.85	1.79	2.59	**2.96**	1.66	1.72	1.79	**1.82**	1.05	1.34
WSS	66.46	93.15	87.92	67.55	49.32	**35.3**	75.02	82.73	86.8	**69.20**	87.78	71.46
Time	132.46	4.62	**0.75**	17.6	2.96	2.55	136.3	7.17	**0.75**	18.3	3.6	2.88

is higher, the distortion in terms of SNR and WSS is smaller and the interference suppression is achieved better as quantified by SIR.

To evaluate the effect of harmonic sparse recovery, we consider the fundamental frequency $f_0 \in [150-400]$ Hz. The frequencies that are not harmonics of f_0 are recovered independently in IHT and L_1L_2 methods. We also consider that the harmonic structures are non-overlapping and k spans the full frequency band. The harmonic sparse recovery approach does not require estimation of f_0. We start from $f_0 = 50$ and consider all of its harmonics within the frequency band (i.e., $f \leqslant 8000$); hence, a block of size $K = 80$ of harmonics of $f_0 = 50$ are recovered jointly. Then we move to $f_0 = 51$ for the next group sparse reconstruction and proceed up to $f_0 = 400$. Therefore, the size of the blocks are variable. To prevent overlapping, the priority is given to the first seen frequency components; for example, if a particular frequency is first included in the harmonics of $f_0 = 50$, it is excluded from the harmonics of $f_0 = 100$. The remaining frequency components are recovered independently. For the OMP method, 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. The results are summarized in Tables 3.

We can see that harmonic sparse recovery can lead to better interference suppression compared to block sparse recovery whereas the overall distortion is more and the perceived quality is less indicating some artifacts. The best results are often achieved by L_1L_2 sparse recovery algorithm. Similar experiments on Numbers corpus (Mccowan, 2003) yields better separation quality using the harmonic model (Asaei et al., 2014). The difference can be justified as the harmonicity of numbers (pronunciation of $0-20$) are generally higher than the average phonetically rich speech utterances provided in MC-WSJ corpus. Moreover, like

Table 2: Speech separation quality in clean and noisy conditions exploiting the spectral proximity and AR correlation models. The SNR of noisy data is 10 dB by adding white Gaussian noise and the absorption coefficients required for characterizing the acoustic by the image model are 25% deviated from the actual parameters.

	Spectral proximity & AR - Clean			
Large-Random	L_1L_2	IHT	OMP	BSBL
SIR	20.00	**22.69**	16.59	15.71
SNR	**25.52**	13.54	11.22	11.33
PESQ	**3.11**	2.45	2.46	2.58
WSS	**31.01**	104.5	59.71	50.46
Uniform-Compact	L_1L_2	IHT	OMP	BSBL
SIR	**15.85**	15.25	15.19	9.30
SNR	**20.58**	14.21	8.61	7.02
PESQ	**2.98**	2.38	2.39	2.49
WSS	**33.57**	87.91	75.74	53.62
Time	148.2	4.25	**0.97**	5.85

	Spectral proximity & AR - Noisy			
Large-Random	L_1L_2	IHT	OMP	BSBL
SIR	16.45	**19.32**	14.31	12.63
SNR	**16.14**	13.36	8.98	8.03
PESQ	2.31	2.32	2.27	**2.45**
WSS	52.38	66.74	59.49	**50.75**
Uniform-Compact	L_1L_2	IHT	OMP	BSBL
SIR	11.72	**14.67**	9.92	7.25
SNR	**13.65**	10.32	9.27	4.95
PESQ	2.04	2.26	2.07	**2.36**
WSS	62.85	73.74	67.53	**57.55**
Time	139.2	4.79	**0.846**	7.2

Table 3: Speech separation quality in clean and noisy conditions exploiting spectral harmonic structure. The SNR of noisy data is 10 dB by adding white Gaussian noise and the absorption coefficients required for characterizing the acoustic by the image model are 25% deviated from the actual parameters.

	Spectral Harmonicity - Clean			
Large-Random	L_1L_2	IHT	OMP	BSBL
SIR	19.75	14.40	9.99	**20.85**
SNR	5.68	4.68	**6.02**	5.38
PESQ	**2.44**	1.99	1.94	2.16
WSS	**32.83**	50.91	43.55	46.49
Uniform-Compact	L_1L_2	IHT	OMP	BSBL
SIR	15.55	10.32	6.49	**21.43**
SNR	5.37	4.40	**5.97**	5.16
PESQ	**2.36**	1.97	1.86	2.11
WSS	**35.09**	54.00	50.38	42.16

	Spectral Harmonicity - Noisy			
Large-Random	L_1L_2	IHT	OMP	BSBL
SIR	**16.69**	15.46	9.35	15.60
SNR	5.46	4.57	**6.05**	5.35
PESQ	**1.45**	1.42	1.27	1.37
WSS	52.29	**51.86**	58.04	62.33
Uniform-Compact	L_1L_2	IHT	OMP	BSBL
SIR	**14.08**	11.13	5.89	3.84
SNR	5.14	4.31	**5.95**	3.89
PESQ	**1.45**	1.45	1.34	0.73
WSS	**56.72**	56.77	70.24	81.12

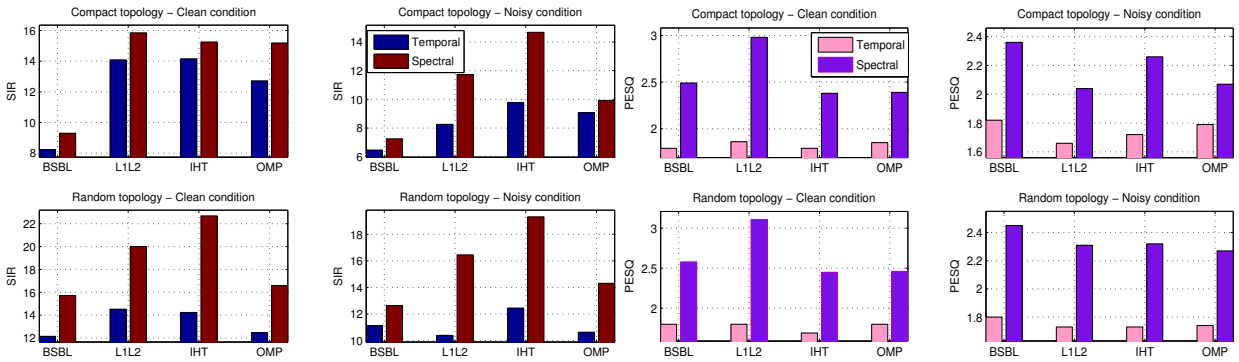

Figure 8: Speech recovery performance in terms of source to interference ratio (SIR) and perceptual evaluation of speech quality (PESQ). SIR measures the amount of interference suppression. PESQ is a perceptually motivated metric which shows high correlation with speech recognition performance (Persia et al., 2008).

the former experiments, comparing the results obtained from the recording of a large random microphone array with that of a uniform compact microphone array, demonstrates that the large random array setup enables higher quality of the separated speech.

From the evaluation results listed in Tables 1– 3, we can say that although the sparse Bayesian learning framework shows higher sensitivity to the measurement coherence for temporal sparse recovery, the performance degrades almost equally for all the algorithms in spectral sparse recovery results.

To get the essence of this thorough empirical evaluation and better contrast the performance of different computational sparse recovery methods exploiting either the temporal or spectral proximity structures along with the random and compact microphone placement, some graphs are extracted from the tabulated results and demonstrated in Figure 8. The bar charts illustrate the amount of interference suppression (SIR) (Vincent et al., 2006), perceptual quality (PESQ) (ITU-T, 2001) as well as weighted spectral slope distance measure (WSS). It is evident that incorporating the spectral structures yields superior performance to the temporal structures in terms of SIR and PESQ; hence, the spectrographic speech representation holds stronger spectral dependencies among the coefficients that affect the reconstruction performance. In addition, we observe that the highest perceptual quality is achieved using the sparse Bayesian learning framework (BSBL) and convex optimization (L_1L_2). It can be due to the zero-forcing limitation of the IHT and OMP methods. This deficiency of forcing the small coefficients to zero is particularly exhibited for speech-like signals which do not possess high compressibility. However, in some applications such as speech recognition, where the reconstruction of the signal is not desired, we can exploit the sparsity of the information bearing components in greedy sparse recovery approaches which offer a noticeable computational speed in efficient implementation and a reasonable performance (Asaei et al., 2011).

Further experiments show that increasing the size of the blocks leads to degradation in the quality of the separated speech although in some cases, it increases the signal to interference ratio (SIR); this observation is in accordance with the perceptual grouping principles relying on proximity in spectro-temporal

space (Wang and Brown, 2006; Asaei, 2013). The block structure underlies the phonetic information (Yang and Hermansky, 2000) which motivates further consideration for speech recognition applications.

5. Concluding Remarks

In this paper, the underdetermined convolutive speech localization and separation is formulated as structured sparse recovery from compressive acoustic measurements. In this context, our studies aim to address two important issues: (1) Identification of the sparsity structures underlying spatio-spectral representation of multiparty speech recordings and (2) Evaluation of the computational methods that incorporate these structures for model-based sparse recovery.

We exploit the spatial sparsity of multipath propagation for characterization of the acoustic measurements based on its unique relation to the source locations inside the enclosure. The spectrographic sparsity structures are formulated upon the auditory principles of structural grouping. It is demonstrated that localization of simultaneous speech sources is more accurate, and quality of the separated speech is better preserved, if the proximity, autoregressive correlation and harmonicity of the spectral or temporal coefficients are incorporated in the sparse reconstruction procedure.

Through the source localization evaluation, we found that grouping a few consecutive frames enables exact localization of up to 9 simultaneous sources from only 4 microphone recordings using either of the convex, combinatorial or Bayesian methods. We could also achieve highly accurate localization using a single frame of overlapping speech by taking into account the spectral structures. To exploit the autoregressive correlation model, it was shown that an average model exists at both temporal and spectral levels and it is more efficient than learning the AR coefficients online. The BSBL method yields the best localization results per frame.

Furthermore, the source separation evaluation under clean and noisy scenarios with acoustic model mismatch confirmed the effectiveness of the sparse Bayesian learning via BSBL algorithm to achieve the best results in terms of perceived speech quality and distortion whereas a high inference suppression is also obtained by IHT and OMP.

Although the performance of both localization and separation improve by incorporating the sparsity structures, learning the hyper-parameters for these structures is application-specific. For example, it is desirable to exploit a large block (e.g. size = 16) of frequencies or frames to enable accurate localization whereas the quality of the reconstructed speech is degraded as the block size goes beyond 2. In addition, the AR correlation model can be computed offline for the source localization task, but it needs to be learned online through the BSBL procedure for high quality speech separation.

The computational cost of the greedy combinatorial methods developed for IHT and OMP demonstrates the efficiency required for real time implementation. Given the accurate localization and high interference

suppression achieved by this approach, in some applications that the perceptual quality of the separated speech is not a desired objective, the greedy combinatorial methods may be considered for sparse recovery.

Throughout the extensive experimental evaluation, we observe that the layout of microphone placement has a great impact on localization and separation performance. Indeed, a random placement of the microphones at a large pairwise distance enables better results compared to a uniform compact setup. The optimal design of the microphones for sparse localization and separation framework remains an open question for future research.

Acknowledgment

The research leading to these results has received funding from by SNSF project on "Parsimonious Hierarchical Automatic Speech Recognition (PHASER)" grant agreement number 200021-153507. The authors would like to acknowledge Raphael Ullmann at Idiap Research Institute for the helpful discussions and inputs on PESQ.

We also would like to acknowledge the anonymous reviewers for the insightful comments and remarks to improve the quality and clarity of the manuscript.

References

Abrard, F., Deville, Y., 2005. A time-frequency blind signal separation method applicable to under-determined mixtures of dependent sources. Elsevier, Signal Processing .

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. Journal of Acoustic Society of America 65.

Araki, S., Makino, S., Blin, A., Mukai, R., Sawada, H., 2004. Underdetermined blind separation for speech in real environments with sparseness and ICA, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Araki, S., Sawada, H., Mukai, R., Makino, S., 2005. A novel blind source separation method with observation vector clustering, in: Proceedings of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC).

Asaei, A., 2013. Model-based Sparse Component Analysis for Multiparty Distant Speech Recognition. Ph.D. thesis. Ecole Polytechnique Federal de Lausanne (EPFL).

Asaei, A., Bourlard, H., Cevher, V., 2011. Model-based compressive sensing for distant multi-party speech recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Asaei, A., Bourlard, H., Garner, P.N., 2010. Sparse component analysis for speech recognition in mullti-speaker environment, in: Proceeding of INTERSPEECH.

Asaei, A., Golbabaee, M., Bourlard, H., Cevher, V., 2014. Structured sparsity models for reverberant speech separation. IEEE/ACM Transactions on Audio, Speech and Language Processing 22.

Baraniuk, R.G., Cevher, V., Duarte, M.F., Hegde, C., 2010. Model-based compressive sensing. IEEE Transactions in Information Theory .

Berg, E.V.D., Friedlander, M.P., 2008. Probing the pareto frontier for basis pursuit solutions,. SIAM Journal on Scientific Computing .

Blumensath, T., Davies, M.E., 2009. Iterative hard thresholding for compressed sensing. Applied and Computational Harmonic Analysis 27(3).

Borish, J., 1984. Extension of the image model to arbitrary polyhedra. Journal of the Acoustic Society of America 75(6).

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambride, MA.

Brungart, D.S., 2001. Information and energetic masking effects in the perception of two simultaneous talkers. Journal of Acoustical Society of America, 2001.

Buchner, H., Aichner, R., Kellermann, W., 2007. TRINICON-based blind system identification with application to multiple-source localization and separation. volume 13. In Blind Speech Separation, S. Makino, T.-W. Lee, and H. Sawada, Eds. New York: Springer.

Cevher, V., 2009. Learning with compressible priors, in: Neural Information Processing Systems (NIPS).

Comon, P., Jutten, C., 2010. Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press.

Cotter, S.F., Rao, B.D., Kjersti, E., Kreutz-Delgado, K., 2005. Sparse solutions to linear inverse problems with multiple measurement vectors. IEEE Transactions on Signal Processing .

Davies, M., Mitianoudis, N., 2004. Simple mixture model for sparse overcomplete ica. IEE Proceedings on Vision, Image and Signal Processing 151 (1).

Dmour, M.A., Davies, M.E., 2011. A new framework for underdetermined speech extraction using mixture of beamformers. IEEE Transactions on Audio, Speech, and Language Processing 19, 2011.

Dokmanić, I., Parhizkar, R., Walther, A., Lu, Y.M., Vetterli, M., 2013. Acoustic echoes reveal room shape. Proceedings of the National Academy of Sciences 110.

Faller, C., Merimaa, J., 2004. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. Journal of the Acoustical Society of America 116(5).

Gribonval, R., Bacry, E., 2003. Harmonic decomposition of audio signals with matching pursuit. IEEE Transactions on Signal Processing 51.

Hu, G., Wang, D., 2001. Speech segregation based on pitch tracking and amplitude modulation, in: IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics.

Huang, Y., Benesty, J., Chen, J., 2005. 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 13(5).

ITU-T, 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. International Telecommunications Union, Geneva, Switzerland .

ITU-T, 2003. Itu-t rec. p.862.1, mapping function for transforming p.862 raw result scores to mos-lqo. International Telecommunications Union, Geneva, Switzerland .

Jafari, M.G., Abdallah, S.A., Plumbley, M.D., Davies, M.E., 2006. Sparse coding for convolutive blind audio source separation, in: The 6th International Conference on Independent Component Analysis and Blind Source Separation.

Kearns, M., Mansour, Y., Ng, A.Y., 1997. An information-theoretic analysis of hard and soft assignment methods for clustering, in: The13th conference on Uncertainty in Artificial Intelligence (UAI-97), Morgan Kaufmann.

Klatt, D.H., 1982. Prediction of perceived phonetic distance from critical-band spectra: a first step, in: IEEE International Conference on Acoustics, Speech, and Signal Processing.

Kollmeier, B., Brand, T., Meyer, B., 2008. Perception of speech and sound. Springer Handbook of Speech Processing.

Kumatani, K., McDonough, J.W., Raj, B., 2011. Maximum kurtosis beamforming with a subspace filter for distant speech recognition, in: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU).

Kyrillidis, A., Cevher, V., 2011. Recipes on hard thresholding methods, in: Proceedings of the fourth international workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP).

Lincoln, M., McCowan, I., Vepa, I., Maganti, H.K., 2005. The multi-channel wall street journal audio visual corpus (mc-wsj-av):

Specification and initial experiments, in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

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

McCoy, M., Cevher, V., Dinh, Q., Asaei, A., Baldassarre, L., 2014. Convexity in source separation: Models, geometry, and algorithms. Signal Processing Magazine, IEEE 31.

Melia, T., Rickard, S., 2007. Underdetermined blind source separation in echoic environment using desprit. EURASIP Journal on Advances in Signal Processing .

Mourad, N., Reilly, J.P., 2010. Modified hierarchical clustering for sparse component analysis, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Nesta, F., Omologo, M., 2012. Convolutive Underdetermined Source Separation through Weighted Interleaved ICA and Spatio-temporal Source Correlation. volume 7191. In: Yeredor, A. et al. (eds.) LVA/ICA 2012. LNCS, Springer, Heidelberg.

Parra, L.C., Alvino, C.V., 2002. Geometric source separation: Merging convolutive source separation with geometric beamforming. IEEE Transactions on Acoustics, Speech and Signal Processing 10(6).

Parsons, T.W., 1976. Separation of speech from interfering speech by means of harmonic selection. Journal of the Acoustical Society of America 60(4).

Persia, L.D., Milone, D., Rufiner, H.L., Yanagida, M., 2008. Perceptual evaluation of blind source separation for robust speech recognition. Signal Processing .

Roman, N., Wang, D., Brown, G.J., 1991. Speech segregation based on sound localization. Journal of the Acoustical Society of America 114(4).

Saab, R., Yilmaz, O., Mckeown, M.J., Abugharbieh, R., 2007. Underdetermined anechoic blind source separation via ℓ_q-basis-pursuit with $q < 1$. IEEE Transactions on Signal Processing .

Taghizadeh, M.J., Garner, P.N., Bourlard, H., 2012. Broadband beampattern for multi-channel speech acquisition and distant speech recognition, in: IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM).

Tan, V.Y.F., Fevotte, C., 2005. A study of the effect of source sparsity for various transforms on blind audio source separation performance, in: Proceeding of Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS).

Tropp, J.A., Wright, S.J., 2010. Computational methods for sparse solution of linear inverse problems,. Proceedings of the IEEE 98.

Vincent, E., Gribonval, R., Fevotte, C., 2006. Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing 14.

Wang, D., Brown, G.J., 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press.

Wang, L., Ding, H., Yin, F., 2011. A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures. IEEE Transactions on Acoustics, Speech and Signal Processing 19(3), 2011.

Wipf, D.P., Rao, B.D., 2007. An empirical bayesian strategy for solving the simultaneous sparse approximation problem. IEEE Transactions on Signal Processing 55(7).

Wolfel, M., McDonough, J., 2009. Distant speech recognition. New York: John Wiley & Sons, .

Yang, H.H., Hermansky, H., 2000. Search for information bearing components in speech, in: Advances in Neural Information Processing Systems 9.

Yilmaz, O., Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing 52.

Zhang, Z., Rao, B.D., 2011. Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning. IEEE Journal of Selected Topics in Signal Processing 5(5).

Zhang, Z., Rao, B.D., 2012. Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation.

IEEE Trans. on Signal Processing .

Zibulevsky, M., Pearlmutter, B.A., 2001. Blind source separation by sparse decomposition in a signal dictionary. Neural Computation 13(4).