MODEL-BASED SPARSE COMPONENT ANALYSIS FOR REVERBERANT SPEECH LOCALIZATION

Afsaneh Asaei¹, Hervé Bourlard¹,², Mohammad J. Taghizadeh¹,² and Volkan Cevher²

¹Idiap Research Institute, Martigny, Switzerland
²Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

{afsaneh.asaei, herve.bourlard, mohammad.taghizadeh}@idiap.ch, volkan.cevher@epfl.ch

ABSTRACT

This paper studies the problem of multiple speaker localization via speech separation based on model-based sparse recovery. We compare and contrast computational sparse optimization methods incorporating harmonicity and block structures as well as autoregressive dependencies underlying spectrographic representation of speech signals. The results demonstrate the effectiveness of block sparse Bayesian learning framework incorporating autoregressive correlations to achieve a highly accurate localization performance. Furthermore, significant improvement is achieved using ad-hoc microphones for data acquisition set-up compared to the compact microphone array.

Index Terms— Structured sparsity, Reverberant speech localization, Autoregressive modeling, Ad-hoc microphone array

1. INTRODUCTION

Speech localization in the clutter of voice and acoustic multipath is an active area of research on microphone arrays for hands-free speech communication. The accurate knowledge of the speaker location is essential for an effective beampattern steering and interference suppression [1, 2]. We briefly review the main approaches to address this problem.

High Resolution Spectral Estimation: These approaches are based on analysis of the received signals’ covariance matrix and impose a stationarity assumption for accurate estimation [3]. Important techniques applied for speech localization include minimum variance spectral estimation as well as eigen-analysis methods such as multiple signal classification (MUSIC). The underlying hypotheses are not quite realistic in reverberant speech localization and alternative strategies have been usually considered [4, 5].

Time Difference Of Arrival (TDOA) Estimation: Another approach is based on TDOA estimation of the sources with respect to a pair of sensors. The generalized cross correlation (GCC) is the most common technique for TDOA estimation where the idea is basically to map the peak location of the cross-correlation function of the signal of two microphones to an angular spectrum. A weighting scheme is usually employed to increase the robustness of this approach to noise and multi-path effects. Maximum likelihood estimation of the weights has been considered as an optimal approach in the presence of uncorrelated noise, while the phase transform (PHAT) has been shown to be effective to overcome reverberation ambiguities [6, 7]. In addition to the GCC-PHAT, identification of the speaker-microphone acoustic channel has been incorporated for TDOA estimation and reverberant speech localization [8, 9]. However, despite of being practical and robust, TDOA-based techniques do not offer a high update rate. Alternative strategies have thus been sought for multiple-target tracking and adaptive beam-steering [10, 11].

Beamformer Steered Response Power (SRP): In this approach, the space is scanned by steering a microphone array beam-pattern and finding the direction associated to the maximum power. Delay-and-sum, minimum variance beamformers, and generalized sidelobe canceler have been the most effective methods for speaker localization [12]. The SRP-based approaches have a higher effective update rate compared to TDOA-based methods, and are applicable in multi-party scenarios using phase-transform weighting scheme [13].

In this paper, we adopt our speech separation framework using sparse component analysis [14] and conduct the evaluations in terms of speech localization [15]. We analyze the reverberant mixtures of speech signals in spectro-temporal domain. The planar area of the room is discretized into a dense grid such that the speakers are located at particular cells exclusively. A spatio-spectral sparse representation is obtained by concatenating the spectral components attributed to the sources located on the grid. The compressive acoustics measurements associated to the microphone array recordings are characterized using Image model of multipath propagation. The spatio-spectral sparse representation is estimated from the compressive array measurements using sparse optimization methods where the supports of high energy components indicate the source locations. The computational approaches to model-based sparse recovery of spectrographic speech are compared and contrasted considering block, harmonic as well as autoregressive dependencies.

The rest of the paper is organized as follows: Section 2 explains the premises underlying model-based sparse component analysis of reverberant recordings, and sets up the formulation of reverberant speech source localization. The structured sparsity models underlying speech components are elaborated in Section 3 followed by the computational approaches to model-based sparse recovery in Sections 4. Section 5 presents the details of the experiments. Conclusions are drawn in Section 6. The notations used in this paper are as follows:

-○ $g \in \{1, \ldots, G\}$: number of a cell on a grids.
-○ $n \in \{1, \ldots, N\}$: number of source; $N \ll G$.
-○ $m \in \{1, \ldots, M\}$: number of microphones; $M < N$.
-○ $f \in \{1, \ldots, F\}$: number of spectral coefficients.
-○ $\{S, \tilde{S}\}$: spectral representation of single/all source signals.
-○ $\{X, \tilde{X}\}$: spectral representation of single/all micro. signals.
-○ $\Phi$: microphone array manifold matrix.
2. SPARSE COMPONENT ANALYSIS OF REVERBERANT SPEECH MIXTURES

2.1. Spatio-Spectral Sparse Representation

The scenario that we consider is consisted of N speakers 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 supposed to be located at the center of a cell, and N ≪ G. The signals corresponding to each cell are concatenated to form a spatial representation of sources. Hence, the energy of the signals on the grid define a spatial spectrum with a sparse support denoting the location of the sources. We consider the spectro-temporal representation of signal and entangle the spatial representation of the sources with the spectral representation of the speech signal to form the complex vector $\mathbf{s} = [\mathbf{s}_1^T, ..., \mathbf{s}_G^T]^T \in \mathbb{C}^{G \times 1}$ where $\mathbf{s}_g$ corresponds to the transpose operator. Each $\mathbf{s}_g \in \mathbb{C}^{F \times 1}$ denotes the spectral representation of the $g^{th}$ source located at cell number $g$ in Fourier domain. We express the signal ensemble at microphone array as a single vector $\mathbf{x} = [\mathbf{x}_1^T, ..., \mathbf{x}_G^T]^T$ where each $\mathbf{x}_m \in \mathbb{C}^{F \times 1}$ denotes the spectral representation of recorded signal at microphone number $m$. The sparse vector $\hat{\mathbf{s}}$ generates the underdetermined ($M < G$) microphone mixture observations as $\mathbf{x} = \Phi \hat{\mathbf{s}}$ where $\Phi$ is the microphone array measurement matrix consisted of the acoustic projections associated to the acquisition of the spatio-temporal sources.

2.2. Acoustic Measurement Characterization

To characterize the acoustic measurements, the room is modeled as a rectangular enclosure consisting of finite impedance walls. The point source-to-microphone impulse responses are calculated using Image Model technique [16] where a reverberant signal is represented as superposition of the signals attributed to the source images with respect to the reflective surfaces. Taking into account the physics of multipath propagation, the projections associated with the source located on the cell $g$ where $\mathbf{v}_g$ represents the position of the center of the cell and captured by microphone $m$ located at position $\mu_m$ are characterized by the media Green’s function through

$$\mathbf{L}_{\mathbf{v}_g \rightarrow \mu_m} \cdot \mathbf{x}(f) = \sum_{r=1}^{T} \frac{\tau_r \mu_m - \mathbf{v}_g}{\mu_m - \mathbf{v}_g} \exp(-\sqrt{\frac{\mu_m - \mathbf{v}_g}{c}}) S(f),$$

where $\tau_r$ is the reflection ratio associated to the $r^{th}$ image source located at $\mathbf{v}_g$. The attenuation constant $\alpha$ 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 $s_2(l) = s(l-p)$, then $s_2(f) \approx \exp(-j\rho)S_1(f)$.

Given the source-sensor projection defined in (1), we construct matrix $\mathbf{Z}_{\mathbf{v}_g \rightarrow \mu_m}$ for the measurement of the $g$ consecutive frequencies as $\mathbf{Z}_{\mathbf{v}_g \rightarrow \mu_m} = \text{diag} \{ \mathbf{L}_{\mathbf{v}_g \rightarrow \mu_{m_1}} \ldots \mathbf{L}_{\mathbf{v}_g \rightarrow \mu_{m_K}} \}$. Hence, the projections associated to the acquisition of the source signals located on the grid by microphone $m$ is $\Phi_m = [\mathbf{Z}_{\mathbf{v}_1 \rightarrow \mu_m} \ldots \mathbf{Z}_{\mathbf{v}_G \rightarrow \mu_m}]$ and the measurement matrix of M-channel microphone array is characterized as $\Phi = [\Phi_1 \ldots \Phi_M]^T$.

To fully identify this model, the location of the source images as well as the associated reflected ratios have been estimated and incorporated for sparse recovery of the reverberant speech signals $\hat{\mathbf{s}}$ [17, 15]. We cast the underdetermined reverberant speech localization problem as sparse approximation where we exploit the underlying structure of the sparse coefficients for efficient recovery using fewer number of measurements [18, 14, 15]. The source locations are determined from the support of the high energy components of $\hat{\mathbf{s}}$ corresponding to the cells on the grid.

2.3. Computational Approaches to Sparse Recovery

Defining a set $\mathcal{M}$ as the union of all vectors with a particular support structure, estimation of the sparse coefficient vector $\hat{\mathbf{s}}$ from the microphone recordings $\mathbf{x}$ can be expressed as

$$\hat{\mathbf{s}} = \arg \min_{\mathbf{s} \in \mathcal{M}} \| \mathbf{s} \|_0 \quad \text{s.t.} \quad \mathbf{x} = \Phi \mathbf{s}$$

where the counting function $\| \cdot \|_0 : \mathbb{R}^G \rightarrow \mathbb{N}$ returns the number of non-zero components in its argument.

The major classes of computational techniques for solving sparse approximation problem are Greedy pursuit, Convex optimization and Sparse Bayesian learning [19].

**Greedy pursuit:** The nonzero components of $\mathbf{s}$ are estimated in an iterative procedure by modifying one or several coefficients chosen to yield a substantial improvement in quality of the estimated signal. The present work considers an extension of the iterative hard thresholding [20, 21] to incorporate sparsity structures underlying spectrographic speech.

**Convex optimization:** The counting function in (2) is replaced with a sparsity inducing convex norm that exploits the structure underlying $\mathbf{s}$. Therefore, a convex objective is obtained which can be solved using convex optimization. The present work considers extension of basis pursuit algorithm which relies on $L_1$ recovery [22].

**Sparse Bayesian learning:** A prior distribution is associated to $\mathbf{s}$ with sparsity inducing hyperparameters and a maximum a posteriori estimation is derived. The present work considers the Bayesian framework proposed in [23, 24].

3. STRUCTURED SPARSITY MODELS

We consider three types of structures underlying the spectral coefficients: **harmonicity, block structure** as well as AR dependency. These structures are supported by the evidences from the studies on computational auditory scene analysis [25, 15].

**Harmonic structure** captures the dependency among the frequencies which are harmonic of a fundamental frequency as particularly exhibited in the voiced parts of speech. Imposing a harmonic structure in recovering vector $\hat{\mathbf{s}}$ requires that at any cell of the grid, the K high energy components can be expressed as harmonics of a fundamental frequency $f_0$ defined through

$$f_H \triangleq [k f_0 | 1 < k < K],$$

**Block structure** indicates that the adjacent frequencies collaborate on a common (spatial) sparsity profile. Imposing this structure in recovering vector $\hat{\mathbf{s}}$ requires that neighboring discrete frequencies correspond to one cell on the grid. Hence, the signal of individual sources is recovered in blocks of size $B$ with the structure defined as

$$f_B \triangleq \{ [f_1, ..., f_{B}], ..., [f_{B-1+B}, ..., f_{B}] \},$$

**AR dependency** is an additional inter-dependency exhibited due to the correlation among the block entries corresponding to each source, which we model using an auto regressive (AR) process of order $\mathcal{R}$ characterized by the following model

$$f_{\mathcal{AR}} \triangleq [\beta_0, \beta_1, \ldots, \beta_{\mathcal{R}}]$$

where $\beta_g \in \{-1, 1\}$ denotes the AR coefficients. The sources $S_g$ are mutually independent, but each source satisfies an AR model as

$$S_g(b) = f_{\mathcal{AR}} [u(b), S_g(b-1), \ldots, S_g(b-\mathcal{R})]^T$$

where $u(b)$ denotes the input sequence. From (6) we can see that the covariance matrix $\boldsymbol{B}_g$ of each source is a Toeplitz matrix identified by the AR coefficients.
4. STRUCTURED SPARSE RECOVERY

We consider different model-based sparse recovery algorithms to recover the sparse vector incorporating the structures defined above. In particular, we employ Iterative hard thresholding (IHT) [26], $L_1L_2$ convex optimization [22] as well as Block Sparse Bayesian Learning (BSBL) [24].

**IHT**: Iterative hard thresholding (IHT) offers a simple yet effective approach to estimate the sparse vectors. It seeks an $N$-sparse approximation $\hat{S}$ matching the observation $X$ by minimizing the residual error. We use the algorithm proposed in [21] which is an accelerated scheme for hard thresholding methods with the following recursion

\[
\hat{S}^0 = 0, \quad \hat{X}^0 = X - \Phi \hat{S}^0
\]

\[
\hat{S}^{i+1} = \hat{X}^i + \kappa \Phi^T \hat{R}^{i+1}
\]

(7)

where the step-size $\kappa$ is the Lipschitz gradient constant to guarantee the fastest convergence speed. To incorporate for the underlying structure of the sparse coefficients, the model approximation operator $\mathcal{M}^{\hat{S}}$ is defined as reweighting and thresholding the energy of the components of $\hat{S}$ with either $\mathcal{T}_B$ or $\mathcal{T}_H$ structures.

$L_1L_2$: Another fundamental approach to sparse approximation replaces the combinatorial counting function in the mathematical formulation stated in (2) with the $L_1$ norm, resulting in a convex optimization problem that admits a tractable algorithm referred to as basis pursuit [22]. We use a group version of basis pursuit algorithm with the number of group components $n^\mathcal{G}$ determined by each structure. The optimization problem to recover the block sparse coefficients $\hat{S}$ is formulated as follows:

\[
\hat{S} = \arg\min_{\hat{S}} \|S\|_{L_1L_2} \text{ s.t. } X = \Phi \hat{S}, \quad \|S\|_{L_1L_2} = \sum_{g=1}^{G} \sum_{b=1}^{n^\mathcal{G}} (S_g(b))^2
\]

(8)

To incorporate the AR dependencies of the block coefficients of $S$, $X = \Phi S$ is solved by (8) where $\Phi$ constitutes of $\mathcal{G}_{\mathcal{V}g}^{-1} = \mu_m$ where the diagonal elements are multiplied by $\mathcal{T}_{\mathcal{A}K}$. Fig. 1 demonstrates an example of an AR signal of order 4 recovered using the proposed procedure. More details are discussed in Section 5.2.

**BSBL**: The correlation among the coefficients modeled as AR dependencies is incorporated in [24] in the framework of SBL [23]. The sources $S_g$ are assumed to be Gaussian and mutually independent. The AR dependency model indicates that the linear combination of the univariate Gaussian holds a Gaussian distribution. More precisely, the joint distribution of $S_g = [S_g^1, ..., S_g^B] \sim \mathcal{N}(0, \Sigma)$ is a multivariate Gaussian, expressed by $p(S_g; \gamma_g, \beta_g) \sim \mathcal{N}(0, \gamma_g B_g)$, where $\gamma_g$ is a non-negative hyper-parameter controlling the block-sparsity of $S_g$ and $B_g \in \mathbb{R}^B \times \mathbb{B}$ is a positive definite matrix that captures the correlation structure of $S_g$ as defined in (6). Under the assumption that blocks are mutually uncorrelated, the prior for $S$ is given by $p(S; \gamma_g, B_g, \gamma_g) \sim \mathcal{N}(0, L_B)$, where $L_B$ is $\text{diag}(\gamma_1 \beta_1, ..., \gamma_g \beta_g)$. Assuming the Gaussian likelihood for the block sparse model as $p(X|S; \sigma^2) \sim \mathcal{N}(\Phi S, \sigma^2 I)$ and applying the Bayes rule, we obtain the posterior density of $S$, which is also Gaussian, $p(S|X; \sigma^2, [\gamma_g, B_g]_{g=1}) = \mathcal{N}(\mu_s, \Sigma_s)$ with the covariance matrix $\Sigma_s = (\Sigma_0^{-1} + \frac{1}{\sigma^2} \Phi^T \Sigma \Phi)^{-1}$. Having all the hyper-parameters $\sigma^2, \gamma_g, B_g$, the MAP estimate of $S$ is given by the mean defined as [24]

\[
\hat{S} \doteq \mu_s = \Sigma_0 \Phi^T (\sigma^2 I + \Phi \Sigma \Phi)^{-1} X.
\]

Clearly, the block sparsity of $S$ is controlled by $\gamma_g$ in $\Sigma_0$. During the estimation procedure, $\gamma_g = 0$ indicates that the associated block in $\hat{S}$ is zeros and no source is located on the corresponding cell. The framework proposed in [27], derives the EM-based learning rule to learn the hyperparameters. We will see in Section 5.2 that the AR-dependency matrix can be estimated offline for the specific task of speech localization.

5. EXPERIMENTAL STUDY

The experiments are conducted to quantify the performance of different structured sparse recovery algorithms on different microphone array geometric settings in terms of speech localization accuracy.

5.1. Acoustic and Analysis Setup

The overlapping speech was synthesized by mixing speech utterances taken from the Wall Street Journal (WSJ) corpus [28]. The WSJ corpus is a 20000-word corpus consisting of read Wall Street Journal sentences. The sentences are read by a range of speakers (34 in total) with varying accents. All the files are normalized prior to mixing. The microphone array recording set-up is consisted of four channels microphones. The planar area of the room with dimension 3m x 3m x 3m is divided into cells with 50 cm spacing. The data collection setup is depicted in Fig. 2. The scenarios include random and compact topologies of microphone array in clean as well as reverberant and noisy conditions. Room impulse responses are generated with the Image model technique [16] using intra-sample interpolation, up to 15th order reflections and omni-directional microphones for a room reverberation time equal to 200 ms. The number of source is known in our experiments. The speech signals of length one second 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.

![Fig. 2: Overhead view of the room set-up for uniform (black dots) and random microphone array (red dots)](image-url)

5.2. Speech Localization Performance

The probabilistic performance bounds of multi-speaker localization are obtained by averaging the results over an exhaustive and exclusive set of configurations. The results are evaluated over all configurations consisted of $N \in \{5, 10\}$ sources. The probabilistic evaluations are necessary to form a realistic expectation of our sparse recovery framework as the deterministic performance bounds are derived for the worst case scenario which is not likely to occur [29, 30]. The localization accuracy is measured as the number...
of times sources are localized correctly (the support of the recovered signal corresponds to the cell on the grid where the source is located) divided by the number of all sources.

The block sparse Bayesian learning (BSBL) algorithm can learn the AR parameters during the optimization, although, the procedure is very expensive in terms of computational cost. Hence, we carry out some studies on an average AR model for speech signal which can be exploited for source localization. To estimate the AR coefficients, the frequency band (number-of-FFT-points = 2048*4) is split into blocks of size 16 and processed independently. Fig. 4 illustrates the frequency domain average AR model for 10min speech signal. The first-order coefficient is estimated as 0.45. We can see that the higher order coefficients are small so the blocks are modeled as a first-order AR process to incorporate the intra-block correlation.

The block sparse Bayesian learning (BSBL) algorithm can learn the AR parameters during the optimization, although, the procedure is very expensive in terms of computational cost. Hence, we carry out some studies on an average AR model for speech signal which can be exploited for source localization. To estimate the AR coefficients, the frequency band (number-of-FFT-points = 2048*4) is split into blocks of size 16 and processed independently. Fig. 4 illustrates the frequency domain average AR model for 10min speech signal. The first-order coefficient is estimated as 0.45. We can see that the higher order coefficients are small so the blocks are modeled as a first-order AR process to incorporate the intra-block correlation.

The block sparse Bayesian learning (BSBL) algorithm can learn the AR parameters during the optimization, although, the procedure is very expensive in terms of computational cost. Hence, we carry out some studies on an average AR model for speech signal which can be exploited for source localization. To estimate the AR coefficients, the frequency band (number-of-FFT-points = 2048*4) is split into blocks of size 16 and processed independently. Fig. 4 illustrates the frequency domain average AR model for 10min speech signal. The first-order coefficient is estimated as 0.45. We can see that the higher order coefficients are small so the blocks are modeled as a first-order AR process to incorporate the intra-block correlation.

The block sparse Bayesian learning (BSBL) algorithm can learn the AR parameters during the optimization, although, the procedure is very expensive in terms of computational cost. Hence, we carry out some studies on an average AR model for speech signal which can be exploited for source localization. To estimate the AR coefficients, the frequency band (number-of-FFT-points = 2048*4) is split into blocks of size 16 and processed independently. Fig. 4 illustrates the frequency domain average AR model for 10min speech signal. The first-order coefficient is estimated as 0.45. We can see that the higher order coefficients are small so the blocks are modeled as a first-order AR process to incorporate the intra-block correlation.

We can see that AR dependency is better preserved using the proposed procedure as illustrated in Fig. 1. However, this approach did not outperform the standard basis pursuit in terms of speech localization. Furthermore, the results of the harmonic sparse recovery were comparable to the block-sparse recovery, hence they are not further elaborated here [15].

The other important observation is that the ad-hoc layout of microphone array improves the results for all sparse recovery algorithms. It can be justified as the theoretical analysis of the performance bounds of sparse recovery algorithms is entangled with the spectral properties of \( \Phi \). A key property to guarantee the theoretical performance bounds is the coherence \( \delta \) of the measurement matrix defined as the smallest angle between any pairs of the columns of \( \Phi \). The number \( N \) of recoverable non-zero coefficients using either convexified or greedy sparse recovery is inversely proportional to the coherence as \( N \propto \frac{1}{\delta \cdot (\Theta^{-1} + 1)} \) [19]. Therefore, to guarantee the performance of sparse recovery algorithms, it is desired to minimize the coherence. As the measurement matrix is constructed of the location-dependent projections, this property implies that the performance of our localization framework is entangled with the microphone array layout. A large-aperture random design of microphone array yields the projections to be mutually incoherent, so the projections are spread across all the acoustic scene and each microphone captures the information about all components of \( S \) [33]. Furthermore, the coherence of the acoustic measurements is smaller at the high frequencies of the broadband speech spectrum, hence the bands below 100Hz are discarded from our localization scheme [15].

6. CONCLUSIONS
In this paper, we incorporated the speech-specific models for structured sparse recovery of reverberant speech sources. We outlined the fundamental computational approaches to model-based sparse recovery and evaluated their performance in terms of source localization accuracy. The numerical assessments show the block sparse Bayesian learning framework yields the best performance and an average AR model can be learned for speaker localization and specified to the algorithm to reduce the computational cost. Furthermore, we considered the impact of construction layout of the microphone array in the performance of sparse recovery framework. The theoretical insights suggest that an ad-hoc design of microphone array can better preserve the acoustic information by reducing the coherence of the acoustic measurements. The empirical evaluations confirm that considering the design specifications acknowledged by the generic theory of sparse signal recovery leads to significant improvement in speech localization performance.
References


