MODEL-BASED SPARSE COMPONENT ANALYSIS FOR REVERBERANT SPEECH LOCALIZATION

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

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{afsaneh.asaei, herve.bourlard, mohammad.taghizadeh}@idiap.ch, volkan.cevher@epfl.ch

ABSTRACT

In this paper, the problem of multiple speaker localization via speech separation based on model-based sparse recovery is studies. 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 obtained 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

Speaker 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, 3]. 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 [4]. 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 [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 side-lobe 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, 14].

In this paper, we adopt our speech separation framework using sparse component analysis [15] and conduct the evaluations in terms of speech localization [16]. 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 acoustic measurements associated to the microphone array recordings are characterized using the image source 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.
◇ $\{S, \mathcal{S}\}$: spectral representation of single/all source signals.
◇ $\{X, \mathcal{X}\}$: spectral representation of single/all micro. signals.
◇ Φ: microphone array manifold matrix.

AA and MJT are supported by the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2). VC is supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof and SNF 200021-132548, SNF 200021-146750 and SNF CRSII2-147633.

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 $N \ll G$. The spectrographic signals corresponding to each source, $S_g \in \mathbb{C}^{F \times 1}, \forall g \in \{1, \dots, G\}$ are concatenated to form a spatio-spectral representation of sources as $\mathcal{S} = [S_1^T ... S_G^T]^T \in \mathbb{C}^{GF \times 1}$ where $.^T$ denotes the transpose operator and F is the number of frequency components. The spatio-spectral representation has a sparse support corresponding to the location of the sources. We express the signal ensemble at microphone array as a single vector $\mathcal{X} = [X_1^T ... X_M^T]^T$ where each $X_m \in \mathbb{C}^{F \times 1}$ designates the spectral representation of recorded signal at microphone m. The sparse vector $\mathcal{S}$ generates the underdetermined ($M < G$) microphone mixture observations as $\mathcal{X} = \Phi \mathcal{S}$ where Φ is the microphone array measurement matrix consisted of the acoustic projections associated to the distant signal acquisition.

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 the *image model* [17] 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 frequency-dependent projection associated with the source located at ν_g and captured by microphone located at μ_m is characterized by the media Green's function and denoted by $\xi^f_{\nu_g \to \mu_m}$. More details are explained in [18]. We construct matrix $\Xi_{\nu_g \to \mu_m}$ for the measurement of the F consecutive frequencies as $\Xi_{\nu_g \to \mu_m} = \text{diag}\,(\xi^1_{\nu_g \to \mu_m} \;\dots\; \xi^F_{\nu_g \to \mu_m})$. Hence, the projections associated with the acquisition of the source signals located on the grid by microphone m is $\phi_m = [\Xi_{\nu_1 \to \mu_m} ... \Xi_{\nu_g \to \mu_m} ... \Xi_{\nu_G \to \mu_m}]$ and the measurement matrix of M-channel microphone array is obtained as $\Phi = [\phi_1 \dots \phi_M]^T$. 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 $\mathcal{S}$ [18, 16]. 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 [15, 16]. The source locations are determined from the support of the high energy components of $\mathcal{S}$ corresponding to the cells on the grid.

2.3. Computational Approaches to Sparse Recovery

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

$$\hat{\mathcal{S}} = \underset{\mathcal{S} \in \mathbb{M}}{\text{argmin}} \|\mathcal{S}\|_0 \quad \text{s.t.} \quad \mathcal{X} = \Phi \mathcal{S} + \nu \tag{1}$$

where the counting function $\|.\|_0 : \mathbb{R}^G \to \mathbb{N}$ returns the number of non-zero components in its argument and ν is a noise vector.

The major classes of computational techniques for solving sparse approximation problem are *Greedy pursuit*, *Convex optimization* and *Sparse Bayesian learning* [19]. In a greedy approach, the nonzero components of $\mathcal{S}$ are estimated in an iterative procedure by modifying one or several coefficients chosen to yield a substantial improvement in quality of the estimated signal. Alternatively, the counting function in (1) is replaced with a sparsity inducing convex norm that exploits the structure underlying $\mathcal{S}$. Therefore, a convex objective is obtained which can be solved using convex optimization. In a Bayesian approach, a prior distribution of $\mathcal{S}$ is considered with sparsity inducing hyperparameters and a maximum a posteriori estimation is derived.

The present work considers the iterative hard thresholding [20], an extension of basis pursuit algorithm [21] as well as the sparse Bayesian learning framework proposed in [22] for model-based sparse recovery incorporating the sparsity structures underlying spectrographic speech.

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 [23, 16].

Harmonic structure captures the dependency among the frequencies which are harmonics of a fundamental frequency as particularly exhibited in the voiced parts of speech. Imposing a harmonic structure in recovering vector $\mathcal{S}$ 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

$$\mathcal{F}_H \triangleq \{k f_0 | 1 < k < K\}, \tag{2}$$

Block structure indicates that the neighboring discrete frequencies collaborate on a common (spatial) sparsity profile. Imposing this structure in recovering vector $\mathcal{S}$ requires that adjacent 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

$$\mathcal{F}_B \triangleq \{[f_1, ..., f_B], \; \dots, \; [f_{F-B+1}, ..., f_F]\}. \tag{3}$$

AR dependency is exhibited due to the correlation among the block coefficients corresponding to each source through an autoregressive process

$$S_g(b) = \sum_{t=1}^{\mathcal{R}} \beta_g(t) S_g(b - t) + u(b), \tag{4}$$

which indicates that $S_g(b)$ can be regressed on $\mathcal{R}$ most recent, consecutive values of the process; $u(b)$ denotes an input sequence. We define the auto regressive (AR) model of order $\mathcal{R}$ as

$$\mathcal{F}_{AR} \triangleq [1, \beta_g(1), \beta_g(2), \dots, \beta_g(\mathcal{R})] \tag{5}$$

where $\beta_g \in (-1, 1)$ denotes the AR coefficients. Thereby,

$$S_g(b) = \mathcal{F}_{AR}\, [u(b), S_g(b-1), \dots, \; S_g(b - \mathcal{R})]^T \tag{6}$$

and the covariance matrix $\mathcal{B}_g$ of the corresponding block to each source is a Toeplitz matrix identified by the AR coefficients.

4. MODEL-BASED SPARSE RECOVERY

We consider different model-based sparse recovery algorithms to recover the sparse vector incorporating the structures defined above. In particular, we exploit Iterative hard thresholding *IHT* [24], L_1L_2 convex optimization [21] as well as Block Sparse Bayesian Learning framework, *BSBL* [22].

IHT: Iterative hard thresholding (IHT) offers a simple yet effective approach to estimate the sparse vectors. It seeks an $\mathcal{N}$-sparse approximation $\hat{\mathcal{S}}$ matching the observation $\mathcal{X}$ by minimizing the residual error through an iterative procedure. We use the algorithm proposed in [20] which is an accelerated scheme for hard thresholding methods with the following recursion

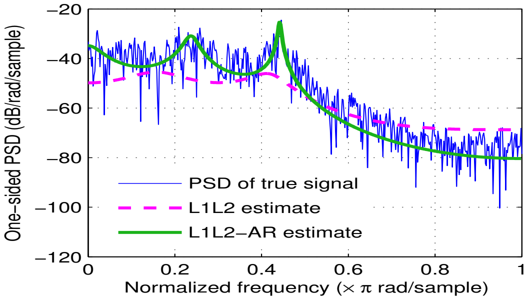

Fig. 1: Incorporating AR dependencies in basis pursuit sparse recovery.

$$\begin{aligned} \hat{\mathcal{S}}^0 &= 0, \quad R^i = \mathcal{X} - \Phi\hat{\mathcal{S}}^i \\ \hat{\mathcal{S}}^{i+1} &= \mathcal{M}^{\mathcal{F}.}\left(\hat{\mathcal{S}}^i + \kappa\Phi^T R^i\right) \end{aligned} \tag{7}$$

where the step-size κ 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 $\mathcal{M}^{\mathcal{F}.}$ is obtained by reweighting and thresholding the energy of the components of $\hat{\mathcal{S}}$ with either $\mathcal{F}_B$ or $\mathcal{F}_H$ structures.

L_1L_2: Another fundamental approach to sparse approximation is based on replacing the combinatorial counting function in the mathematical formulation stated in (1) with the L_1 norm, resulting in a convex optimization problem that admits a tractable algorithm referred to as basis pursuit [21]. We use the group basis pursuit algorithm where the number of group components $n^{\mathcal{F}.}$ is determined by each structure. The optimization problem to recover the block sparse coefficients $\hat{\mathcal{S}}$ is formulated as follows:

$$\hat{\mathcal{S}} = \underset{\mathcal{S}}{\operatorname{argmin}}\{\|\mathcal{S}\|_{L_1,L_2} \text{ s.t. } \mathcal{X} = \Phi\mathcal{S}\},\ \|\mathcal{S}\|_{L_1,L_2} = \sum_{g=1}^{G}\sqrt{\sum_{b=1}^{n^{\mathcal{F}.}} S_g^2(b)} \tag{8}$$

To incorporate the AR dependencies of the block coefficients of $\mathcal{S}$, Φ in (8) is revised to $\tilde{\Phi}$ consisted of $\tilde{\Xi}_{\nu_g\to\mu_m}$ where the diagonal elements are multiplied by $\mathcal{F}_{\mathcal{AR}}$; thus $\tilde{\Xi}_{\nu_g\to\mu_m} = \Xi_{\nu_g\to\mu_m}\mathcal{F}_{\mathcal{AR}}$ and the vector $[u(b), S_g(b-1), \ldots, S_g(b-\mathcal{R})]$ is recovered based on the observation at $\mathcal{X}(b)$. The signal can be reconstructed by filtering the recovered u [25] while the AR model parameters can be estimated from the recovered signal components through an iterative procedure. Fig. 1 demonstrates an example of an AR signal of order four recovered using the proposed procedure. More details are discussed in Section 5.2.

BSBL: The correlation among the coefficients modeled through an AR process is incorporated in the framework of block sparse Bayesian learning [26, 22]. Due to the AR dependency model, each block of S_g is assumed to satisfy a multivariate Gaussian distribution as $p(S_g; \gamma_g, \mathcal{B}_g) \sim \mathcal{N}(0, \gamma_g\mathcal{B}_g)$ where γ_g is a non-negative hyper-parameter controlling the block-sparsity of $\mathcal{S}$ and $\mathcal{B}_g \in \mathbb{R}^{B\times B}$ is a positive definite matrix that captures the correlation structure of the g^{th} block.

Under the assumption that the sources S_g are mutually uncorrelated, the prior of $\mathcal{S}$ is given by $p(\mathcal{S}; \gamma_g, \mathcal{B}_g, \forall g) \sim \mathcal{N}(0, \Sigma_0)$, where Σ_0 is $\operatorname{diag}([\gamma_1\mathcal{B}_1 \ldots \gamma_G\mathcal{B}_G])$. Assume the noise vector satisfies $p(\nu) \sim \mathcal{N}(0, \sigma^2 I)$, we have $p(\mathcal{X}|\mathcal{S}; \sigma^2) \sim \mathcal{N}(\Phi\mathcal{S}, \sigma^2 I)$. By applying the Bayes rule, we obtain the posterior density of $\mathcal{S}$, which is also Gaussian, $p(\mathcal{S}|\mathcal{X}; \sigma^2, \{\gamma_g, \mathcal{B}_g\}_{g=1}^{G}) = \mathcal{N}(\mu_s, \Sigma_s)$ with the covariance matrix $\Sigma_s = (\Sigma_0^{-1} + \frac{1}{\sigma^2}\Phi^T\Phi)^{-1}$. Having all the hyper-parameters σ^2, γ_g, $\mathcal{B}_g$, the MAP estimate of $\mathcal{S}$ is given by the mean defined as [22]

$$\hat{\mathcal{S}} \triangleq \mu_s = \Sigma_0\Phi^T(\sigma^2 I + \Phi\Sigma_0\Phi^T)^{-1}\mathcal{X}, \tag{9}$$

The block sparsity of $\hat{\mathcal{S}}$ is controlled by γ_g in Σ_0; during the estimation procedure, $\gamma_g = 0$ indicates that the associated block in $\hat{\mathcal{S}}$ 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

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×3m×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 [17] using intra-sample interpolation, up to 15^{th} 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 thus the number DFT points is 2048.

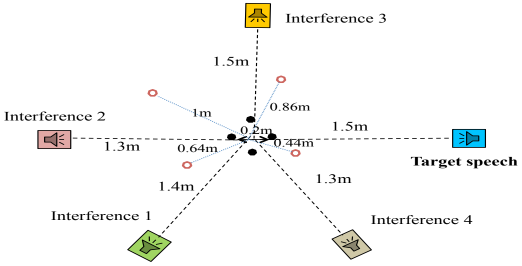

Fig. 2: Overhead view of the room set-up for uniform (black dots) and random (red circles) microphone array. The center of the uniform array is positioned at the room center.

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]. The localization accuracy is measured as the number of times that 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; however, 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 is split into blocks of size 16 processed

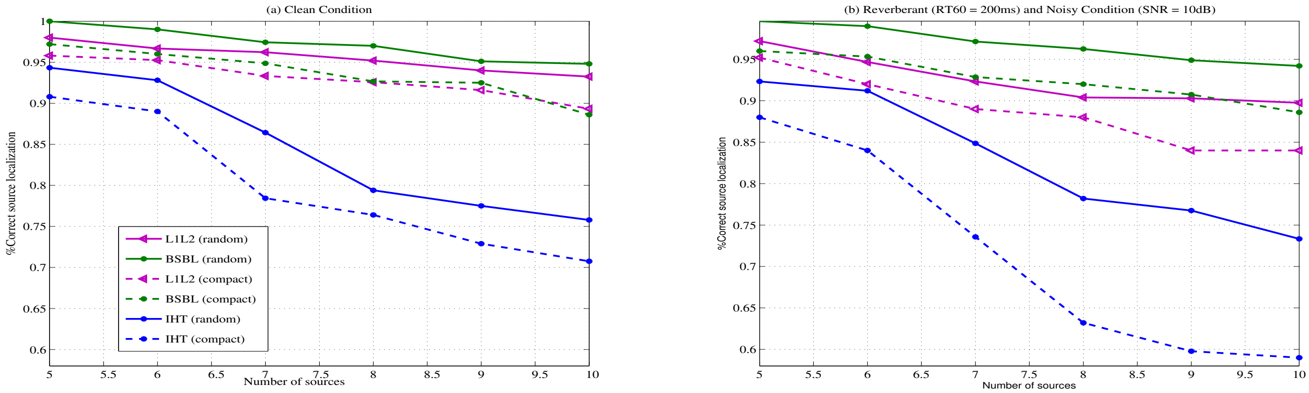

Fig. 3: Speaker localization performance evaluated for 5-10 sources in (a) clean and (b) reverberant and noisy condition

independently. Fig. 4 illustrates the frequency domain average AR model for 10 min 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. In addition, we can assume that all sources have similar correlation structure. The experimental analysis on speech-specific average AR model as depicted in Fig. 4 shows very small variance around the AR coefficients and supports this approximation. Hence, the corresponding covariance matrix of any block $\mathcal{B}_g$ in (9) is a Toeplitz matrix with the form of

$$\text{Toeplitz}([1, \beta, \beta, \ldots, \beta^{\mathcal{R}}]) = \begin{bmatrix} 1 & \beta & \ldots & \beta^{\mathcal{R}} \\ \vdots & & & \vdots \\ \beta^{\mathcal{R}} & \beta^{\mathcal{R}-1} & \ldots & 1 \end{bmatrix}$$

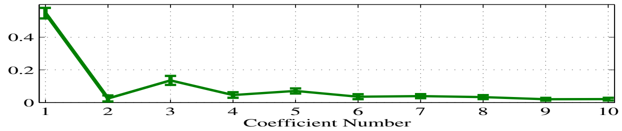

Fig. 4: 10-order AR coefficients estimated for 10min speech signal. The cross lines illustrate the variance of estimates

The results of multi-speaker localization exploiting block structure are illustrated in Fig. 3 for B = 16. All the algorithms are run for the stopping threshold fixed to 1e-2 and the maximum iteration of 150. We can see that exploiting the frequency structures yield very strong results. The number of microphones is only 4 whereas we can localize up to 9 sources with 95% accuracy. The performance compares favorably with the recent results reported on multi-source TDOA estimation and localization in reverberant acoustic [30, 31, 32, 33]. The orthogonality or disjointness of spectrographic speech signals is a key property to achieve this bound of performance [15, 16, 34]. The typical computing time for IHT [20] is real time on a modern desktop computer. However, the BSBL [22] is less than 2 times and L_1L_2 [21] is about 30 times slower than real time.

The BSBL algorithm incorporates the AR dependency model to replace the Euclidean norms with Mahalanobis distance measure and it plays a role of whitening the sources during the learning of hyperparameters [27]. On the other hand, incorporation of AR model in the framework of L_1L_2 enables preserving these structural dependencies. As a basic example, an AR signal is generated by filtering a white Gaussian noise. The formulation of the L_1L_2-AR enables recovery of the input u (6) along with the signal components. 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 [16, 35], hence they are not further elaborated here.

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 Φ. A key property to guarantee the theoretical performance bounds is the coherence ϑ of the measurement matrix defined as the smallest angle between any pairs of the columns of Φ. The number N of recoverable non-zero coefficients using either convexified or greedy sparse recovery is inversely proportional to the coherence as $N < \frac{1}{2}(\vartheta^{-1} + 1)$ [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 $\mathcal{S}$ [36]. Furthermore, the coherence of the acoustic measurements is smaller at the high frequencies of the broadband speech spectrum, hence the bands bellow 100Hz are discarded from our localization scheme [16].

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

[1] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine, Special Issue on Fundamental Technologies in Modern Speech Recognition*, 2012.

[2] A. Asaei, M. J. Taghizadeh, M. Bahrololum, and M. Ghanbari, "Verified speaker localization utilizing voicing level in split-bands," *Signal Processing*, vol. 89(6), 2009.

[3] M. J. Taghizadeh, R. Parhizkar, P. N. Garner, and H. Bourlard, "Euclidean distance matrix completion for ad-hoc microphone array calibration," in *IEEE 18th International Conference in Digital Signal Processing*, 2013.

[4] J. Dmochowski, S. Benesty, and S. Affes, "Broadband music: opportunities and challenges for multiple source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors," *Journal of Signal Processing Systems*, vol. 63(2), 2011.

[6] M. Omologo and P. Svaizer, "Acoustic source localization in noisy and reverberant environments using CSP analysis," in *Proc. of ICASSP*, 1996.

[7] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, 2012.

[8] F. Ribeir, C. Zhang, D. Florencio, and D. Ba, "Using reverberation to improve range and elevation discrimination in sound source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18(7), 2010.

[9] S. Nam and R. Gribonval, "Physics-driven structured cosparse modeling for source localization," in *Proc. of ICASSP*, 2012.

[10] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language*, vol. 11(2), 1997.

[11] M. Omologo F. Nesta, P. Svaizer, "Cumulative state coherence transform for a robust two-channel multiple source localization," in *Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009.

[12] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, Ph.D. thesis, 1993.

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

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

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

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

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

[18] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22(3), 2014.

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

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

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

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

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

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

[25] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28(1), 1999.

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

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

[28] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, "The multichannel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2005.

[29] P. Boufounos, P. Smaragdis, and B. Raj, "Joint sparsity models for wideband array processing," in *Wavelets and Sparsity XIV, SPIE Optics and Photonics*, 2011.

[30] Charles Blandin, Alexey Ozerov, and Emmanuel Vincent, "Multisource tdoa estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[31] J. Le Roux, P. T. Boufounos, K. Kang, and J. R. Hershey, "Source localization in reverberant environments using sparse optimization," in *Proc. of ICASSP*, 2013.

[32] H. T. Do, *Robust cross-correlation-based methods for sound-source localization and separation using a large-aperture microphone array*, Ph.D. thesis, Brown University, 2011.

[33] M. B. Dehkordi, H. R. Abutalebi, and M. R. Taban, "Sound source localization using compressive sensing-based feature extraction and spatial sparsity," *Digital Signal Processing*, vol. 23(4), 2013.

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

[35] A. Asaei, M. Davies, H. Bourlard, and V. Cevher, "Computational methods for structured sparse recovery of convolutive speech mixtures," in *Proc. of ICASSP*, 2012.

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