Structured Sparse Coding for Microphone Array Location Calibration

Afsaneh Asaei[1,2,3], *Bhiksha Raj*[3], *Hervé Bourlard*[1,2], *Volkan Cevher*[2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland
[3]Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
afsaneh.asaei@idiap.ch,bhikshap@cs.cmu.edu,herve.bourlard@idiap.ch,volkan.cevher@epfl.ch

Abstract

We address the problem of microphone location calibration where the sensor positions have a sparse spatial approximation on a discretized grid. We characterize the microphone signals as a sparse vector represented over a codebook of multi-channel signals where the support of the representation encodes the microphone locations. The codebook is constructed of multi-channel signals obtained by inverse filtering the acoustic channel and projecting the signals onto a array manifold matrix of the hypothesized geometries. This framework requires that the position of a speaker or the track of its movement to be known without any further assumption about the source signal. The sparse position encoding vector is approximated by model-based sparse recovery algorithm exploiting the block-dependency structure underlying the broadband speech spectrum. The experiments conducted on real data recordings demonstrate the effectiveness of the proposed approach and the importance of the joint sparsity models in multi-channel speech processing tasks.

Index Terms: Microphone array calibration, Structured Sparse coding, Model-based sparse recovery, Multi-party speech signals

1. Introduction

Microphone array calibration is the fundamental initial step in multi-channel speech processing systems. In this paper we focus on the problem of calibrating the location of microphones, i.e. the determination of array topology, in order to enable spatial filtering for high-quality speech acquisition. Before describing our work, we first overview previous approaches for calibration of the microphone arrays to identify some of the practical challenges faced by them.

The research leading to these results has received funding from the European Union under the Marie-Curie Training project SCALE (Speech Communication with Adaptive LEarning), FP7 grant agreement number 213850. We also thank Mohammad J. Taghizadeh for his inputs on calibration techniques and implementation of the MDS-based method proposed in [1].

Previous Work. McCowan and Lincoln proposed a calibration method based on a diffuse noise field model [1]. A diffuse noise field is characterized by noise signals that propagate with equal probability from all locations. The coherence in any frequency band between the noise arriving at any two microphones can hence be shown to be a Sinc function of the distance between the microphones. McCowan and Lincoln propose to compute the inter-microphone distances by fitting the measured coherence of the noise with a Sinc function in least-squared error sense. To increase the robustness, the noise frames are extracted and classified by k-means clustering. Although their method does not require any specific set-up or signals to be transmitted, the performance is limited to very compact microphone arrays in an enclosure where the diffuse noise model holds.

Alternative approaches incorporate transmission of a known signal for microphone calibration. For instance, Flanagan and Bell proposed a method which integrates self-calibration and source localization. Their method estimates the source directions of arrival (DOAs) along with the sensors locations using the Weiss-Friedlander technique [2]. The estimation of sensor location and DOAs are performed alternatively until the algorithm converges. Sachar et. al. presented an experimental setup for calibration of the microphones by pulsed acoustic excitation using an array of five domed tweeters as sources [3]. A test pulse is used to record and measure the transmission times between speakers and microphones. Although the microphone array calibration techniques based on transmission of known signal are usually capable of performing some level of gain and phase calibration, their applicability is limited due the restricted scenario of microphone recordings they employ [4, 5].

On the other hand, some approaches have been proposed to calibrate the full network of ad-hoc microphones given only partial information about the pairwise distances. In a practical method known as Multi-Dimensional Scaling (MDS)-MAP, the shortest paths between all pairs of nodes is approximated using information of a partly known network topology. To refine the approximation, it applies singular value decomposition and

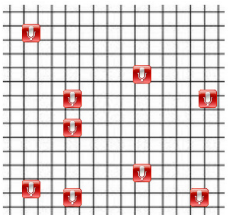

Figure 1: Microphone (colored boxes) positions on a dicretized grid. The occupancy of grid locations by microphones is sparse.

reconstructs a low-rank matrix by truncating the singular values [6]. The classical MDS is then applied to estimate the microphone coordinates.

Our Contribution. The present study provides a new perspective on microphone array position calibration problem. We assume that the microphones already have their gain and phase information calibrated and their mutual coupling effects are small. In addition, we assume we know the location of one source. We propose a sparse recovery framework for microphone location estimation where the unknown sensor locations are approximated over a discrete grid. The key idea is that the microphone locations are sparse over the discretized area. This idea is illustrated in Figure 1. We hypothesize a set of locations corresponding to the unknown sensor and calculate their weights using the Iterative Hard Thresholding sparse recovery algorithm [7]. Compared to the relevant microphone array calibration work in the literature, our approach is fundamentally different, as we provide a sparse approximation for the sensor locations as opposed to the continuous solutions. The mathematical formulation used in our calibration approach is a dual of the solution of joint localization-separation via sparse approximation, which recovers multiple speech sources using known sensor positions [8, 9, 10, 11].

The paper follows with the statement of the theoretical framework of microphone calibration problem in Section 2. The theory of the our structured sparse coding solution is elaborated in Section 2.3. The experimental evaluations on various practical scenarios are presented in Section 3 along with the analysis of the empirical and theoretical performance bounds. The conclusions are drawn in Section 4.

2. Sparse Coding for Microphone Calibration

2.1. Problem Statement

We consider a scenario that an unknown sound signal $S(f)$ at frequency f emanates from a known location in an enclosure and impinges on an array of M microphones located at $\mathcal{L} = \{l_1, l_2, \cdots, l_M\}$ on a 2-D plane. The room response H_{l_m} from the source location to the location l_m is known for each of the M microphone locations. The signal captured by a microphone located at l_m would therefore be

$$X_{l_m}(f) = H_{l_m}(f)S(f)$$

and representing $X_{\mathcal{L}}(f) = [X_{l_1}(f) \cdots X_{l_M}(f)]^\top$ and $H_{\mathcal{L}}(f) = [H_{l_1}(f) \cdots H_{l_M}(f)]^\top$, we can write

$$X_{\mathcal{L}}(f) = H_{\mathcal{L}}(f)S(f). \quad (1)$$

$H_{\mathcal{L}}$ is also known as the *array manifold* vector and is specific to the source location and the locations of the microphones at $\mathcal{L} = \{l_1, l_2, \cdots, l_M\}$. The microphone-calibration problem is that the *location* of the microphone is not known and must be estimated.

We can obtain an estimate of the source as $\hat{S}(f) = H_{\mathcal{L}}(f)^\dagger X_{\mathcal{L}}(f)$, where $H_{\mathcal{L}}(f)^\dagger$ represents the pseudo-inverse of $H_{\mathcal{L}}(f)$. Given that the estimate $\hat{S}(f)$ obtained using any $H_{\mathcal{L}}(f)$ is correct, then

$$X_{\mathcal{L}}(f) = H_{\mathcal{L}}(f)\hat{S}(f) = H_{\mathcal{L}}(f)H_{\mathcal{L}}(f)^\dagger X(f); \quad (2)$$

this now gives us an effective handle to estimate $\mathcal{L}$ as

$$\mathcal{L} = \underset{l_1, l_2, \cdots, l_M}{\arg\min} \|X_{\mathcal{L}}(f) - \hat{X}_{\mathcal{L}}(f)\|_2^2, \quad (3)$$

where

$$\hat{X}_{\mathcal{L}}(f) = H_{\mathcal{L}}(f)H_{\mathcal{L}}(f)^\dagger X_{\mathcal{L}}(f) \quad (4)$$

is the projection of $X_{\mathcal{L}}(f)$ onto the array manifold vector $H_{\mathcal{L}}(f)$.

The discerning reader may note that the objective function of Equation (3) is merely $\|(I - H_{\mathcal{L}}H_{\mathcal{L}}^\dagger)X(f)\|_2^2$, which is minimized if $H_{\mathcal{L}}$ is chosen such that the solitary non-unity singular value of $I - H_{\mathcal{L}}H_{\mathcal{L}}^\dagger$ goes to zero. This may appear to be independent of $X_{\mathcal{L}}(f)$; however this is not so – the corresponding eigenvector must also be maximally aligned to $X_{\mathcal{L}}(f)$ for the objective to be minimized. Nevertheless, the formulation expressed in Equation (3) introduces greater dependence on data.

The above modification can be succinctly stated in matrix form as follows. Let $F = \{f_1, f_2, \cdots, f_B\}$ represent a set of B adjacent frequencies within a band. We define an array manifold *matrix* for M sensors in locations $\mathcal{L} = \{l_1, l_2, \cdots, l_M\}$ as the $MB \times B$ matrix $H_{\mathcal{L}}(F)$ obtained by stacking a set of diagonal matrices obtained from $H_{l_1}(f)$ to $H_{l_M}(f)$. Let $H_{l_m}^{\text{diag}}(F) = \text{diag}([H_{l_m}(f_1)) \; H_{l_m}(f_2) \cdots H_{l_m}(f_B)])$. $H_{\mathcal{L}}(F) = [H_{l_1}^{\text{diag}}(F) \cdots H_{l_M}^{\text{diag}}(F)]^\top$. We define $X_{l_m}(F) = [X_{l_m}(f_1) \; X_{l_m}(f_2) \cdots X_{l_m}(f_B)]^\top$; $X_{\mathcal{L}}(f) = [X_{l_1}(f) \cdots X_{l_M}(f)]^\top$. We define $S(F) = [S(f_1) \; S(f_2) \cdots S(f_B)]^\top$. The extended equivalent of Equation (1) is given by

$$X_{\mathcal{L}}(F) = H_{\mathcal{L}}(F)S(F). \quad (5)$$

The location of the two microphones can be estimated as

$$\mathcal{L} = \underset{l_1, l_2, \cdots, l_M}{\arg\min} \|X_{\mathcal{L}}(F) - H_{\mathcal{L}}(F)H_{\mathcal{L}}(F)^\dagger X_{\mathcal{L}}(F)\|_2^2. \quad (6)$$

This formulation indicates a parametric approach to microphone calibration problem where $\mathcal{L}$ is estimated directly by minimizing the objective function stated in (6). It defines the source locations as continuous random vectors in a 2-D plane and results in a non-linear objective which is difficult to optimize. In the this paper, we resort to a non-parametric method and we formulate the microphone calibration problem as structured sparse coding where we leverage the sparse recovery algorithms to find the optimal solution. This idea is described in the following Sections.

2.2. Sparse Calibration Model

We consider a scenario in which M microphones are distributed on a discrete grid of G points sufficiently dense so that each microphone can be assumed to lie at one of the grid points and $M \ll G$. We then define a G-dimensional grid selector vector P with components P_i that are 1 or 0 depending on whether or not a microphone is present at grid point i. With this notation, note that the number of microphones M is equal to the ℓ_0 norm of P, which is defined as the number of non-zero elements in the vector. Thereby, the microphone calibration problem can be converted into a linear regression and the solution could be formulated as follows [11, 12]

$$\hat{P} = \arg\min\{\|X - \mathcal{C}P\|_2^2 : \quad P \in \{0,1\}, ||P||_0 = M\} \tag{7}$$

where $\mathcal{C} = H(F)H(F)^\dagger X(F)$ and we drop the parenthesized (F) here for brevity. The possible number of combinations of microphone positions is therefore $\binom{G}{M}$, since each of the microphones can lie at each of the G positions. Corresponding to each of these $\binom{G}{M}$ arrangements is an array manifold vector. Any one of these could represent the *true* array manifold vector for the array. The complexity of this problem is very high so we take a greedy sparse recovery approach.

If the location of $M - K$ of the sensors is known *a priori* and only K sensor locations are unknown, then the choice of possible manifold vectors reduces to $\binom{G}{K}$. In the discussion below we assume $K = 1$ for simplicity, but the argument is easily extended to higher values of K. Given the multi-channel signal recording $X \in \mathbb{C}^{M\times 1}$ and assuming that the position of $M - 1$ of the microphones are known, we construct a codebook denoted by $\mathcal{C} \in \mathbb{C}^{M\times G}$, composed from projections of X onto G array manifold vectors as given by Equation 4. The i^{th} manifold vector corresponds to a microphone array with $M - 1$ microphones at known positions and the M^{th} microphone at the i^{th} grid locations. Since the support of P corresponds to the location of the microphone on the grid it is a 1-*sparse* vector.

Given the observations and the codebook of the signal projections onto the manifold vectors corresponding to G grid locations, calibration of the unknown microphone position amounts to sparse approximation of P. The solution to Equation 7 finds the location of one microphone, given the locations of the remaining; however it generalizes trivially to the case of K unknown microphone locations. In the following Section 2.3, we elaborate on construction of the codebook from the observations.

2.3. Codebook of Spatial Signals

The design of the code book $\mathcal{C}$ is based on the reconstruction of the acoustic field from multi-channel recordings. Consider a source signal S from a known location, which is recorded by each of M microphones. Let the location of i^{th} microphone be $l_{p(i)}$. $p(i)$ is unknown. The signal X_i captured by the i^{th} microphone is obtained by passing S through the acoustic channel of the room from the source location to $l_{p(i)}$, $H_{p(i)}$. Hence, we have a linear model of the M microphone observations in spectral domain stated as

$$\begin{bmatrix} X_1(f) \\ \vdots \\ X_M(f) \end{bmatrix} = \begin{bmatrix} H_{p(1)}(f) \\ \vdots \\ H_{p(M)}(f) \end{bmatrix} S(f), \tag{8}$$

or, more succinctly, representing $X(f) = [X_1(f) \cdots X_M(f)]^\top$ and $H(f) = [H_{p(1)}(f) \cdots H_{p(M)}(f)]^\top$ as earlier, $X(f) = H(f)S(f) + E$, where E represents measurement error. We will generally assume that the error is isotropic. We refer to this equation as the *forward* model.

The spectral components are obtained by Short Time Fourier Transform (STFT). As each frame is processed independently, the frame indices are omitted in this notation for brevity. This formulation relies on the narrowband assumption that if the source is delayed in time domain, i.e. if $s_2(t) = s_1(t - l)$ then for all $l < L_{max}$, $S_2(f, \tau) \approx \exp(-jfl)S_1(f, n)$ where $S_i(f, n)$ is the STFT of the time domain signal $s_i(t)$, n indicates the current frame index.

Given the formulation of Equation 8, the least-squares approximation to the source signal $S(f)$ is given by $\hat{S}(f) = H^\dagger(f)X(f)$ as given in Equation 2.

In order to characterize the forward model, we consider the recording environment to be a rectangular enclosure consisting of finite-impedance walls. The point source-to-microphone impulse responses $H_i(f)$ to each of the grid locations are calculated using the Image Model technique [13]. Taking into account the properties signal propagation and multi-path effects, the frequency response of the acoustic channel between a source located at ν and a microphone located at l_i is identified as

$$H_i(f) = \sum_{r=1}^{R} \frac{\iota^r}{\|l_i - \nu_r\|^\alpha} \exp(-jf\frac{\|l_i - \nu_r\|}{\tau}), \tag{9}$$

where $j = \sqrt{-1}$, ι represents the reflection ratio of the walls, ι^r is the cumulative reflection ratio when the signal is reflected r times and τ denotes the speed of sound. The attenuation constant α depends on the nature of the propagation and is considered in our model to equal 1, which represents spherical propagation. Hence, characterization of the forward model amounts to localization of the R Images of the source along with the absorption ratios ι associated to the reflective surfaces. Although the point-source assumption does not hold in practice, evaluations on real data verify that estimation of the early support of the room impulse response function enables estimation of ι and R, and that these can be applied to determine the parameters of the forward model with sufficient accuracy to enable efficient recovery of speech by sparse approximation. Details of the procedure can be found in [14, 15] and are not repeated here; for now we will assume that ι and R have been well estimated and are known.

Assuming we know the locations of $M-1$ microphones and only the M^{th} microphone must be located, there are only G possible valid array configurations to consider in the construction of the codebook $\mathcal{C}$ in Equation 7. We compose the corresponding set of array manifold matrices $H_1(F), H_2(F), \cdots, H_G(F)$, where $H_i(F)$ represents the manifold matrix for the array configuration where the first $M-1$ microphones are in their known locations, and the M^{th} microphone is at l_i. We now write the codebook as

$$\mathcal{C} = [H_1(F)H_1^{\dagger}(F)X(F), \ldots H_G(F)H_G^{\dagger}(F)X(F)]. \tag{10}$$

The X in the calibration model of Equation 7 must correspondingly be taken to actually represent $X(F)$. P is now a $GB\times 1$ matrix, with the property that it is B-*sparse* with a block structure: at most B consecutive entries beginning at index Bm can be non-zero, where m is an integer.

2.4. Calibration by Structured Sparse Recovery

The calibration model expressed in (7) indicates that once the codebook is constructed of all the spatial projections of the multi-channel signals, calibration of the unknown microphone position amounts to sparse approximation of the encoding vector P which selects the projections corresponding to the right location. Since the codebook is constructed of F adjacent frequencies, the non-zero components of P has a block structure corresponding to the common support/grid where the unknown microphone is located. To incorporate the underlying structure of the sparse coefficients, we use the model-based sparse recovery algorithm proposed in [12] which is an accelerated scheme for hard thresholding methods with the following recursion:

$$P_{i+1} = \mathcal{M}\left(P_i + \kappa \mathcal{C}^{\top}(X - \mathcal{C}P_i)\right), \tag{11}$$

where the step-size κ is the Lipschitz gradient constant to guarantee the fastest convergence speed. To incorporate for the underlying block structure, the model projection operator $\mathcal{M}$ thresholds and retains only the one (or more generally K) B-block with the highest energy, with subsequent renormalization [12]. The support of the finally estimated P determines the microphone location.

3. Experimental Analysis

3.1. Real Recordings Set-up

We perform some evaluations on the Multichannel Overlapping Numbers Corpus (MONC) [16]. This database is acquired by playback of utterances from the original Numbers corpus. The recordings were made in a 8.2m×3.6m×2.4m rectangular room containing a centrally located 4.8m ×1.2m rectangular table. The positioning of loudspeakers was designed to simulate the presence of the speakers seated around a circular meeting room table of diameter 1.2m. The loudspeakers were placed at 90° spacings at an elevation of 35cm (distance from table surface to center of main speaker element). An eight-element, 20cm diameter, circular microphone array placed in the center of the table was used to record the mixtures.

3.2. Calibration Results

The speech signals are recorded at 8kHz sampling frequency. The spectro-temporal representation is obtained by windowing the signal in 256ms frames using a Hann function with 50% overlap. We used the algorithm published in [15] to characterize the forward model. As mentioned earlier, to jointly localize all M microphones, we will require a codebook with $\binom{G}{M}$ entries. As this is computationally infeasible, we take an incremental approach. We first locate two microphones, which only requires a codebook of size $\binom{G}{2}$, representing the array manifold vectors for all possible pairs of locations. Thereafter, we incrementally locate additional microphones until all microphones are calibrated. To increase the resolution of the estimates while keeping the dimensionally of the sparse vector bounded, we take a coarse-to-fine strategy [17]. We discretize the area into 5cm grids. The localized microphones are then re-located in 1cm accuracy using a finer discretization. We calibrate the first two channels at the array broad-side. We then move onto the next channel and continue until the full network is calibrated. The average norm of calibration error for the relative geometry is 8.9mm.

To calibrate a two-channel microphone, it is possible to find the combinatorial solution of Equation (7). We performed the microphone calibration by combinatorial optimization and the results were similar to what we obtained by hard thresholding expressed in Equation (11). Given that the complexity of the combinatorial optimiza-

tion increases as $\mathcal{O}(G^M)$ whereas the greedy sparse recovery has a complexity of $\mathcal{O}(GM)$, it is crucial to employ the structured sparse recovery algorithms to enable microphone array calibration in our set-up.

In addition, we use the method proposed in [1] for calibration of the circular array used for MONC recordings. This method relies on diffuse noise model to find the topology of the array and it can not perform calibration with the diffuse noise recorded in MONC database. We conducted some data recordings with the similar microphone array set-up. The results obtained for calibration of circular microphone is about 1.2cm using about 10s recording of diffuse noise field. In practice however, the level and length of the available diffuse noise might be challenging to employ the technique proposed in [1]. Hence, our approach which requires only a few speech frames (less than 1s) provides a higher applicability and accuracy.

3.3. Empirical Performance Bounds

To establish the empirical performance bounds, we carry out the experiments on synthetic data recordings using ad-hoc microphones distributed in a 0.4m×0.4m area as illustrated in Figure 1. The reference point speaker is located at either 0.5m or 1.5m distance to the center of the grid corresponding to a near-field or far-field speaker respectively. We considered a 3m×3m×3m room and synthesized the room impulse responses with the Image Model [13] with reflective factors of 0.8 for the six walls, which corresponds to 180ms reverberation time according to Eyring's formula:

$$\beta = \exp(-13.82/[c(L_x^{-1} + L_y^{-1} + L_z^{-1})T]), \quad (12)$$

where L_x, L_y and Lz are the room dimensions, τ is the speed of sound in the air ($\approx 342m/s$) and T is the room reverberation time.

We consider the scenario in which the recording condition is perfectly known. To evaluate the sensitivity of our approach to the uncertainties in the estimation of forward model parameters (i.e. uncertainties or errors in estimates of speaker location, room geometry and absorption coefficients) we consider two mismatched test conditions. In the first scenario (Mismatched1), the observations were generated with a forward model where the codebook is constructed of the spatial projections using a models with up to 25% error in the absorption coefficients corresponding to each of the walls. In the second scenario (Mismatched2), we assume that the room geometry is estimated with an error of 10cm and the absorption coefficients are estimated with 25% error on each of the six reflective walls. The performance of the microphone calibration in terms of Root Mean Squared Error (RMSE) is listed in the Table 1. The parameters δ indicates the resolution of the grid which is in our case equal to 5cm. We considered all pairs of combinations to quantify an average expected error to calibrate the first two-channels. We then select a third channel for calibration. We observe that the scenario of the speaker positioned at a far-field distance with respect to the microphone array is less sensitive to the mismatched conditions.

Table 1: RMSE (cm) of microphone array calibration. Two combinations (Co.) are considered: Pairs (P) and Triples (T). δ indicates the resolution of the grid and is equal to 5cm in our experiments

Acoustic condition	Co.	Near-field	Far-field
Match. & Mismatch.1	P	δ	δ
	T	δ	δ
Mismatched2	P	14.7	12.3
	T	14	6

3.4. Theoretical Performance Bounds

Relying on the formulation of the microphone calibration as sparse coding expressed in (7), the theoretical analysis of the performance of our approach amount to the sparse recovery guarantees and it is tied to the properties of the codebook matrix $\mathcal{C}$. A fundamental property of $\mathcal{C}$ is the *coherence* between the columns defined as [18]

$$\mu(\mathcal{C}) = \max_{1 \leq j,k \leq G, j \neq k} \frac{|\langle c_j, c_k \rangle|}{\|c_j\| \|c_k\|}. \quad (13)$$

The coherence quantifies the smallest angle between any pairs of the columns of $\mathcal{C}$ and the number of recoverable non-zero coefficients (K) using either convexified or greedy sparse recovery is inversely proportional to μ as $K < \frac{1}{2}(\mu^{-1} + 1)$ [18]. Hence, to guarantee sparse recovery performance, it is desired that the coherence is minimized. Since the codebook is constructed of locations and frequency dependent Green's function projections, this property implies that the contribution of the source to the array's response is small outside the corresponding sensor location or equivalently the resolution of the array is maximized. Recent studies have shown that the ad-hoc microphone arrays distributed randomly yield significant improvements in the sparse signal reconstruction performance [14, 10]. Thus the performance of our sparse approximation framework is entangled with the design of the grid points for codebook construction as well as the frequency of the signal. To analyze the codebook for the broadband speech spectrum, we compute μ for different frequency bands. The results are illustrated in Figure 2.

This study shows that the coherence of the codebook is smaller for the higher frequencies and suggests sub-band processing of the speech signal. Alternatively, joint sparsity models enable us to reduce the ambiguity while exploiting the synergy of the broadband components. This issue is investigated in Figure 3.

The results indicate that processing the frequencies

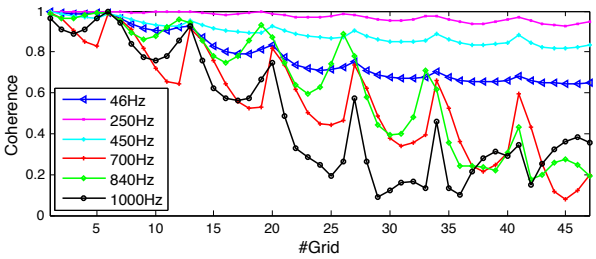

Figure 2: Coherence of the codebook for different frequencies of speech spectrum

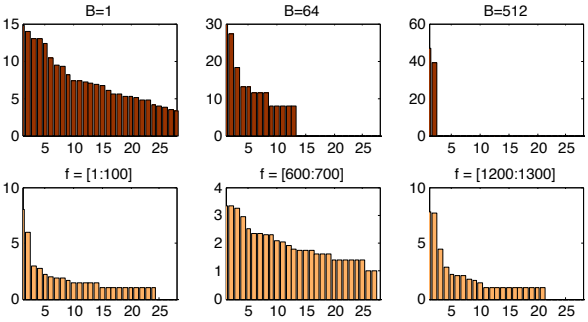

Figure 3: Partial support of P

independently is quite likely to get very ambiguous due to the high coherence of the codebook over some components. In contrast, the block-sparsity model enables very sharp estimates as a function of the block size. Hence, incorporating joint sparsity models such as the block-dependency structures improves the recovery performance in sparse modeling framework.

Our formulation of the microphone calibration enables calibration from multiple overlapping speech sources. Hence, it is possible to increase the number of sources (i.e. number of reference points) to achieve more accurate calibration results.

4. Conclusions

We cast microphone array calibration problem as sparse coding over a codebook of spatially projected multi-channel signals. The support of this representation corresponds to the arrangement of the microphone approximated on a discretized grid. This approach enables estimation of microphone array topology from recordings of an unknown speech source positioned at a known reference point in a reverberant enclosure. We demonstrated the effectiveness of our framework on real data recordings when a circular microphone array is calibrated by a line-radial grid search. We further performed an exhaustive evaluation in a general setting of ad-hoc microphones and quantified the average expected error in case of acoustic ambiguities. Our studies highlight the importance of the structured sparsity models in sparse coding framework for multi-channel speech processing tasks.

5. References

[1] I. McCowan, M. Lincoln, and I. Himawan, "Microphone array shape calibration in diffuse noise fields," *IEEE Transactions on Audio,Speech and Language Processing*, vol. 16(3), 2008.

[2] B. P. Flanagan and K. L. Bell, "Array self-calibration with large sensor position errors," *Signal Processing*, vol. 81, 2001.

[3] J. M. Sachar, H. F. Silverman, and W. R. Patterson, "Microphone position and gain calibration for a large-aperture microphone array," *IEEE Transactions on Speech and Audio Processing*, vol. 13(1), 2005.

[4] S. D. Valente, M. Tagliasacchi, F. Antonacci, P. Bestagini, A. Sarti, and S. Tubaro, "Geometric calibration of distributed microphone arrays from acoustic source correspondence," in *IEEE Workshop on Multimedia Signal Processing(MMSP)*, 2010.

[5] M. Chen, Z. Liu, L. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad-hoc microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007.

[6] Y. Shang, W. Ruml, Y. Zhang, and M. P. J. Fromherz, "Localization from connectivity in sensor networks," *IEEE Transactions on Parallel Distribute Systems*, vol. 15(11), 2004.

[7] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27:3, pp. 265–274, 2009.

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

[9] V. Cevher and R. Baraniuk, "Compressive sensing for sensor calibration," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2008.

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

[11] V. Cevher, P. Boufounos, R. G. Baraniuk, A. C. Gilbert, and M. J. Strauss, "Near-optimal bayesian localization via incoherence and sparsity," in *Proceedings of IPSN*, 2009.

[12] A. Kyrillidis and V. Cevher, "Recipes on hard thresholding methods," in *Proceedings of the fourth international workshop on Computational Advances in Multi-Sensor Adaptive Processing CAMSAP*, 2011.

[13] 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.

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

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

[16] "The multichannel overlapping numbers corpus," Idiap resources available online:, `http://www.cslu.ogi.edu/corpora/monc.pdf`.

[17] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53(2), 2005.

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