Structured Sparse Acoustic Modeling for Speech Separation

Afsaneh Asaei*, Mohammad Golbabaee†, Hervé Bourlard*, and Volkan Cevher‡
*Idiap Research Institute, Martigny and École Polytechnique Fédérale de Lausanne, Switzerland
†Centre de Recherches en Mathmatiques de la Decision, Paris, France
‡Laboratory for Information and Inference Systems, École Polytechnique Fédérale de Lausanne, Switzerland
Emails: afsaneh.asaei@idiap.ch, golbabaee@ceremade.dauphine.fr, herve.bourlard@idiap.ch, volkan.cevher@epfl.ch

***Abstract*—A novel formulation of acoustic multipath is proposed for estimation of the room acoustic using recordings of unknown concurrent speech sources at unknown locations. The framework exploits sparsity and low-rank structures characterized by the Image method for estimation of the geometry and the absorption factors of the reflective surfaces. The experiments conducted on real data recordings demonstrate the effectiveness of the method for modeling the room acoustic and its application for speech separation and dereverberation.**

I. Introduction

We assume the room to be a rectangular enclosure consisting of finite impedance walls. The planar area of the room is discretized into grid of G-cells such that the sources occupy exclusive cells. A signal spectrum S_g is attributed to each cell thus $\mathcal{S} = [S_1^T \dots S_G^T]^T$ forms the signal of the acoustic field. The signal recorded by the microphone i is denoted by X_i and $\mathcal{X} = [X_1^T \dots X_M^T]^T$ denotes the microphone array recordings. A sparse vector $\mathcal{S}$ (i.e., sources) generates the recorded signals by the linear model $\mathcal{X} = \Phi\mathcal{S}$ where Φ characterizes the compressive acoustic projections consorted to acquisition of the acoustic scene data. The objective is to identify Φ from recordings of few unknown source signals at unknown locations.

II. Factorized Multipath Acquisition Model

The point source-to-microphone impulse responses of the room are calculated using the *Image Method* and the acoustic projections are characterized by the media Green's function [**?**]. We assume that the G-cells grid of the room containing N sources is expanded into $\mathcal{G}$-cells free-space discretization where the *actual-virtual* sources are active. The room geometry can be estimated by sparse approximation and low-rank clustering of the individual source images [**?**]. Given the geometry of the room, a *free-space* Image Model maps the position index $i \in \{1, \dots, G\}$ of each source to a group $\Omega_i \subset \{1, \dots, \mathcal{G}\}$ containing the location indices of this source and its images. Hence, we can factorize the acoustic projections as $\Phi = OP$, where $O \in \mathbb{C}^{M \times \mathcal{G}}$ is the free-space Green's function matrix and $P \in \mathbb{R}_+^{\mathcal{G} \times G}$ is the permutation matrix such that its i^{th} column contains the absorption factors of $\mathcal{G}$ points on the grid of actual-virtual sources with respect to the reflection of the i^{th} actual source. Since the Image Model characterizes the source groups, each column $P_{.,i}$ is consequently supported only on the corresponding (non-overlapping) group Ω_i i.e., $\forall i \in \{1 \dots, G\}$, $\forall j \notin \Omega_i, P_{j,i} = 0$.

III. Source Localization and Reflection Estimation

Relying on spatio-spectral sparsity of concurrent speech sources, the covariance matrix of the reverberant recordings exhibits structured sparsity such that $C = \mathcal{X}\mathcal{X}^* = O\Sigma O^* = \sum_{i=1}^{G} O_{.,\Omega_i} \Sigma_{\Omega_i,\Omega_i} O^*_{.,\Omega_i}$, where $.^*$ denotes conjugate transpose and $O_{.,\Omega_i}$ corresponds to the row elements of Ω_i^{th} column of matrix O. $\Sigma = P\mathcal{S}\mathcal{S}^*P^*$. The third equality follows because of the structure of the permutation-attenuation matrix P which indicates that Σ is supported only on the set $\bigcup_i \Omega_i \times \Omega_i$ i.e.,

$$\Sigma^i \triangleq \Sigma_{\Omega_i,\Omega_i} = \|\mathcal{S}_{i,.}\|_2^2 P_{\Omega_i.,} P^*_{\Omega_i,.}, \text{ and } \forall i \neq j \ \Sigma_{\Omega_i,\Omega_j} = 0 \quad (1)$$

Note that if $\|\mathcal{S}_{i,.}\|_2^2 \neq 0$ then the corresponding matrix Σ^i is rank one. As we can see, recovering the diagonal elements of $\Sigma_{\Omega_i,\Omega_i}$ is sufficient to identify the energy of the corresponding source i and the absorption coefficients $P_{\Omega_i,.}$. We thus focus on recovering these sub-matrices for all $i \in \{1, \dots, G\}$ from the observation covariance matrix C. Using the property of the Kronecker product, we can write

$$C_{vec} = \underbrace{\Big[B(1)B(2)\dots B(G)\Big]}_{\mathcal{B}} \underbrace{\Big[\nu(1)^T \nu(2)^T \dots \nu(G)^T\Big]^T}_{\mathcal{V}} \quad (2)$$
$$\forall i \in \{1 \dots, G\} : \nu(i) \triangleq \left(\Sigma_{\Omega_i,\Omega_i}\right)_{vec}, \; B(i) \triangleq \overline{O}_{.,\Omega_i} \otimes O_{.,\Omega_i}.$$

where $\otimes$ denotes the Kronecker product between two matrices and $\overline{O}_{.,\Omega_i}$ is the *element-wise* conjugate of $O_{.,\Omega_i}$. The simultaneous low-rank and joint sparse recovery can then be formulated by the following convex problem with the tuning parameter λ:

$$\begin{aligned} \underset{\Sigma^1,\dots,\Sigma^G}{\arg\min} \ \sum_{i=1}^{G} \left\|\Sigma^i_{vec}\right\|_{L_2} & \text{ subject to } \quad \|C_{vec} - \mathcal{B}\mathcal{V}\|_{L_2} \leqslant \varepsilon \\ \Sigma^i = (\Sigma^i)^* \quad \forall i \in \{1, \dots, G\} & \qquad \Sigma^i_{l,j} \geqslant 0 \quad \forall l, j, i \end{aligned} \quad (3)$$

We solve (3) using the iterative proximal splitting algorithm [**?**].

IV. Experiments and Conclusion

Given the location of the sources and the characterized room acoustic channel, the desired signal can be recovered by inverse filtering. The Zelinsky post-filtering is applied to the recovered speech to reduce the effect of late reverberation. We evaluate the performance of the method on separation of three concurrent speech sources in terms of Perceptual Evaluation of Speech Quality (PESQ), Source to Interference Ratio (SIR) to measure the amount of interference cancellation and Word Recognition Rate (WRR). The proposed method relying on room acoustic modeling via joint sparse recovery (RAM-SR) is compared with superdirective (SD) beamforming. The results are summarized in Table I.

TABLE I: Performance evaluation of speech separation

Meas.	Baseline	Lapel Mic.	SD Beamf.	RAM-SR
PESQ	1.6	2.27	2.48	2.62
SIR	-0.7	18.35	10	14.2
WRR%	39.92	68.13	61.45	79.21

We can see from (1) that $\Sigma^i = \|\mathcal{S}_{i,.}\|_2^2 P_{\Omega_i.,} P^*_{\Omega_i,.}$ is a rank one matrix so we can replace the objective of (3) by $\sum_{i=1}^{G} \left\|\Sigma^i\right\|_*$ to perform low-rank recovery. We further compare the results with joint sparse recovery and explore the advantages of incorporating the low-rank structures for room acoustic modeling and dereverberation.

References

[1] 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.

[2] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for multiparty speech recovery from reverberant recordings," *arXiv:1210.6766*, 2012.

[3] M. Golbabaee and P. Vandergheynst, "Compressed sensing of simultaneous low-rank and joint-sparse matrices," *arXive:1211.5058*, 2012.