COMPUTATIONAL METHODS FOR STRUCTURED SPARSE COMPONENT ANALYSIS OF CONVOLUTIVE SPEECH MIXTURES

Afsaneh Asaei^{1,2}, *Michael E. Davies*^{3}, *Hervé Bourlard*^{1,2}, *Volkan Cevher*^{2}

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland
[3]Institute for Digital Communications, The University of Edinburgh, United Kingdom

afsaneh.asaei@idiap.ch,mike.davies@ed.ac.uk,herve.bourlard@idiap.ch,volkan.cevher@epfl.ch

ABSTRACT

We cast the under-determined convolutive speech separation as sparse approximation of the spatial spectra of the mixing sources. In this framework we compare and contrast the major practical algorithms for structured sparse recovery of speech signal. Specific attention is paid to characterization of the measurement matrix. We first propose how it can be identified using the Image model of multipath effect where the acoustic parameters are estimated by localizing a speaker and its images in a free space model. We further study the circumstances in which the coherence of the projections induced by microphone array design tend to affect the recovery performance.

***Index Terms*—** Structured sparse signal recovery, convolutive source separation, Image model, sparse microphone array

1. INTRODUCTION

Blind separation of the speech signal from an acoustic clutter of unknown competing sound sources plays a key role in many applications involving distant-speech recognition, scene analysis, videoconferencing, hearing aids and surveillance.

Previous approaches to tackle this problem can be loosely grouped into three categories. The first category relies on spatial filtering techniques based on beamforming or steering the microphone array beam pattern towards the target speaker and suppression of the undesired sources [1]. The second category incorporates the statistical characteristics of the sources to identify the mixing model [2]. The sources are usually recovered from the mixtures by least square optimization or matrix pseudo-inversion. The third category is based on sparse representation of the signal, also known as sparse component analysis. These techniques exploit a prior assumption that the sources have a sparse representation in a known basis or frame. The notion of sparsity opens a new road to address the degenerate unmixing problem when the number of sensors is less than the number of speakers, also known as under-determined source separation [3].

Recent studies of Lawrence Carin recognized a physical manifestation of the compressive sensing (CS) measurements through the projections associated with the media Green's function [4, 5]. In [6] we leveraged the CS theory in a novel formulation of underdetermined convolutive speech separation from dimensionality reducing measurements. Our framework exploits a prior knowledge of the room geometry for recovery of the reverberant signal [6, 7]. In practice however, such information is barely available. So in this follow-up, we first fix that limitation by recovering a speech signal and its images using a free space model. The images are then incorporated to estimate the geometry of the field. We further study various algorithmic approaches to sparse recovery and analyse how the performance guarantees are entangled with the design of microphone array layout.

The paper follows with the problem statement and characterization of the microphone array measurement matrix in Section 2.3. We introduce the greedy vs. convexified alternatives of our structured sparse component analysis framework in Section 2.4 and study the theoretical relationship between the microphone array measurements and the sparse recovery performance in Section 2.5. The experiments are presented in Section 3 and the conclusions are drawn in Section 4.

2. STRUCTURED SPARSE COMPONENT ANALYSIS

2.1. Problem Statement

We consider an approximate model of the acoustic observation as a linear convolutive mixing process, stated concisely as

$$x_m(t) = \sum_{n=1}^{N} h_{mn}(t) * s_n(t), \quad m = 1, ..., M; \qquad (1)$$

where s_n refers to the source signal n convolved through the acoustic channel, h_{mn} and recorded at microphone m (x_m). N and M denote the number of sources and microphones respectively. This formulation is stated in time domain; to represent it in a sparse domain, we consider the spectro-temporal representation of speech signal.

Our objective is to separate the N sources from M convolutive mixtures while $M < N$. We cast the underdetermined speech separation problem in the spectro-temporal domain as a sparse approximation where we exploit the underlying structure of the sparse coefficients in the recovery algorithm [8].

2.2. Multi-party Speech Representation

We consider a scenario in which N speakers are distributed in a planar area discretized into G grids. We assume to have a sufficiently dense grid so that each speaker is located at one of the grid points and $N \ll G$. Hence, the spatial spectra of the sources is

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. AA also thanks the Institute for Digital Communications (IDCoM) and the Centre for Speech Technology Research (CSTR) at the university of Edinburgh for hosting her as the SCALE secondment.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

a G-dimensional vector consisted of components of the signal corresponding to each grid. We consider time-frequency representation of multi-party speech and entangle the spatial representation of the sources with the spectral representation of the speech signal and define a vector Z whose support is the time-frequency contribution of each source signal located at grid point g. Suppose that the number of analysis coefficients is F, each element of z_g is an $F \times 1$ vector which carries the spectral coefficients coming from grid g. Thereby, a spatio-spectral representation of the original sources would be obtained as a vector with $F \times G$ components denoted by $Z = [Z_1^T ... Z_G^T]^T$. We express the signal ensemble at microphone array as a single vector $X = [X_1^T ... X_M^T]^T$ where each X_m is an $F \times 1$ vector consisted of the time-frequency components of the signal recorded at microphone m. The sparse vector Z generates the sensor observations as $X = \Phi Z$.

2.3. Measurement Matrix Identification

We consider the room acoustic as a rectangular enclosure consisted of finite-impedance walls. The point source-to-microphone impulse responses are calculated using the Image model technique [9]. Taking into account the physics of the signal propagation and multi-path effects, the projections associated with the source located at ν and captured by the microphone at μ are characterized by the media Green's function and denoted as $\xi_{\nu \to \mu}$ defined by

$$\xi^f_{\nu \to \mu} : X(f, \tau) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu - \nu_r\|^\alpha} \exp(-jf \frac{\|\mu - \nu_r\|}{c}) S(f, \tau), \tag{2}$$

where $j = \sqrt{-1}$ and ι corresponds to the reflection ratio of the walls when the signal is reflected r times. f denotes the frequency and c is 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 corresponds to the spherical propagation. This formulation assumes that if $s_1(t) = s(t)$ and $s_2(t) = s(t - l)$ then for all $l < L_{max}$, $S_2(f, \tau) \approx \exp(-jfl) S_1(f, \tau)$. Given the source-sensor projection defined in (2), the measurement matrix associated to M-channel microphones array would be defined as $\Phi = [\phi_1 ... \phi_M]^T$ where

$$\begin{aligned} \phi_i &= [\Xi_{\nu_1 \to \mu_i} ... \Xi_{\nu_g \to \mu_i} ... \Xi_{\nu_G \to \mu_i}], \\ \Xi_{\nu_g \to \mu_i} &= \begin{bmatrix} \xi^1_{\nu_g \to \mu_i} & 0 & \cdots & 0 \\ 0 & \xi^2_{\nu_g \to \mu_i} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \xi^F_{\nu_g \to \mu_i} \end{bmatrix}. \end{aligned} \tag{3}$$

The projection expressed in (2) corresponds to characterization of the forward model of the acoustic channel as

$$h(t, \mu, \nu) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu - \nu_r\|} \delta(t - \frac{\|\mu - \nu_r\|}{c}). \tag{4}$$

Hence, identification of the measurement matrix boils down to recovery of the support of the forward model which is entangled with localization of the R Images of the source and estimation of the reflection coefficients.

It has been shown in [10] that the impulse response function is a unique signature of the room and the geometry can be reconstructed provided that up to second order of reflections are known. They propose a technique to estimate the room geometry in a very high signal-to-noise ratio. Relying on this observation and assuming that there are segments of non-overlapping speech, we localize the source images using the sparse recovery algorithm in a free space measurement model, i.e., $R = 0$, while the deployment of the grids captures the location of early reflections. The support of the acoustic channel, $\{\nu_r | 1 < r < R\}$ corresponds to the grids where the recovered energy of the signal is maximized. Given the support of the channel, we estimate the best fit geometry of the room entangled with an approximation of the attenuation constant in mean-square sense. This regression assumes that the reflection coefficients are fixed and independent of frequency.

2.4. Structured Sparse Signal Recovery

The major classes of computational techniques for solving sparse approximation problems are greedy pursuit, convex relaxation, non-convex optimization and Bayesian algorithms [11]. This paper focuses on greedy algorithms, in particular Iterative Hard Thresholding (IHT) and Orthogonal Matching Pursuit (OMP) as well as convex optimization, which offer provable correct solutions under well-defined conditions [11].

We focus on two classes of structures underlying the sparse coefficients, namely block inter-dependency, i.e., connection between adjacent frequencies, as well as harmonicity exhibited in spectrographic representation of speech signal. To state it more precisely, the vector Z consisted of $F \times G$ components is recovered as $\mathcal{F} \times G$ independent components where $\mathcal{F}$ corresponds to adjacent frequencies in a block-sparse model as defined in

$$\mathcal{F}_B = \{[f_1, ..., f_b], [f_{b+1}, ..., f_{2b}], [f_{F-b+1}, ..., f_F]\} \tag{5}$$

and b denotes the size of the blocks. Similarly, if the harmonicity of the spectral coefficients is incorporated in the structured sparse recovery algorithm, then the $\mathcal{F}$ contains a harmonic subset of frequencies and would be expressed as

$$\mathcal{F}_H = \{k f_0 | 1 < k < K\}, \tag{6}$$

where f_0 is the fundamental frequency and K is the number of harmonics. Relying on the two proposed structures for recovery of the speech frames, the model-based sparse recovery approaches to exploit these structures would be as follow

IHT: We use the algorithm proposed in [12] which is an accelerated scheme for hard thresholding methods with the following recursion:

$$Z_{i+1} = \mathcal{M}\left(Z_i + \kappa \Phi^t (X - \Phi Z_i)\right), \tag{7}$$

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 $\mathcal{M}$ is defined as reweighting and thresholding the energy of the $\mathcal{F}_\mathcal{X}$ structures [12] where $\mathcal{X}$ corresponds to B or H.

OMP: Suppose that Λ indexes the subset of components from Φ, with the matrix Φ_Λ composed of the corresponding subset of components. The signal estimation at each step would be

$$\begin{aligned} \hat{Z} &= \hat{\Phi}^\dagger_\Lambda X \\ \hat{\Phi}_\Lambda &= \arg\min \|X - \Phi_\Lambda \Phi^\dagger_\Lambda X\|_2. \end{aligned} \tag{8}$$

Denoting the set difference operator as $\backslash$, the corresponding $\mathcal{F}_\mathcal{X}$ structures of $\Phi_{\backslash \Lambda}$ are searched per iteration and Λ is expanded so as the mean-squared error of the residual is minimized [13, 14].

L_1L_2: Sparsity inducing convex norms are a major alternative to the greedy approaches; we use a multiple-measurement version of basis pursuit algorithm with the following objective [15]

$$\hat{Z} = \arg\min \|Z\|_{1,2} \quad s.t. \quad X = \Phi Z,$$
$$\|Z\|_{1,2} = \left(\sum_{i=1}^{G} \left[\sum_{j=1}^{\mathcal{F}} Z^2(i,j) \right]^{1/2} \right). \tag{9}$$

2.5. Compressive Sensing Point of View

The inspiring analogy between the natural projections manifested by the media Green's function and the type of measurements associated with CS enables us to exploit the generic theory of CS for quantitative assessment of the sparse recovery performance in microphone array recordings [4, 5]. From the CS point of view, it is desired that the coherence between the columns of the measurement matrix is minimized. Since the microphone array measurements are constructed of the location-dependent projections, this property implies that the contribution of the source to the array's response is small outside its corresponding location or equivalently the resolution of the array has to be maximized. Carin shows that the Green's function constituted projections given that the inter-element spacing is large enough exhibit an optimal design and the columns of the measurement matrix corresponds to a sampled Fourier basis function [4]. It has been further pointed out that a large-aperture random design of sensor array yields the projections to be mutually incoherent. So the projections are spread across all the acoustic scene and each sensor captures the information about all components of Z. The CS theory further implies that minimizing the mutual coherence is beneficial to reduce the number of microphones required for signal recovery [4, 5]. Motivated by these insights, the performance of our sparse approximation framework is entangled with the microphone array construction design. This issue is addressed in Section 3.

3. EXPERIMENTS

3.1. Evaluation Set-up

The synthesized overlapping speech scenario is depicted in Figure 1. The sampling frequency is 8 kHz. The spectro-temporal representation is obtained by windowing the signal in 256 ms frames using a Hann function with 50% overlapp. The length of the speech signal is 15s. The planar area of the room with dimension $3m \times 3m \times 3m$ is divided into grids with 60cm spacing.

We evaluate the quality of the recovered speech using the Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), Source-to-Artifact Ratio (SAR) and source Image to Spatial distortion Ratio (ISR) as described in [16]. In addition, we measure the Perceptual Evaluation of Speech Quality (PESQ) [17]. The PESQ of the clean speech is 4.5.

3.2. Room Geometry Estimation

The first step requires an initialization of the system to infer the room geometry. We estimate the room geometry by localizing the images of a single speaker in an extended area of dimension $9m \times 9m$. An utterance of a single source co-located with the microphone array is recorded in a reverberant room so we can ignore the direct-path. The source images are then localized by L_1 minimization using our sparse recovery framework in a free space model. The location of the source images corresponds to the support of the room impulse response function. The energy of the recovered signal is sorted and truncated to the order of $D(D+1)/2$, D denotes the number of reflective surfaces and it is equal to 6 in our case; hence it corresponds to the support of the early reflections of the walls [10]. The estimated support of the room impulse response function is then used for estimation of the room rectangular geometry in least-squares sense. However, the estimated room acoustic is not always exact. We did the experiments for 10 random utterances of the length 1-2s; the maximum error obtained in estimating the geometry was 50cm. Taking into account this uncertainty, our experiments are conducted for the worst case where the room geometry is considered as $3.5m \times 3.5m \times 3.5m$ in the Image model for measurement characterization. The reflected coefficients are assumed to be known and equal to 0.8 which corresponds to 180ms reverberation time according to the Eyring's formula:

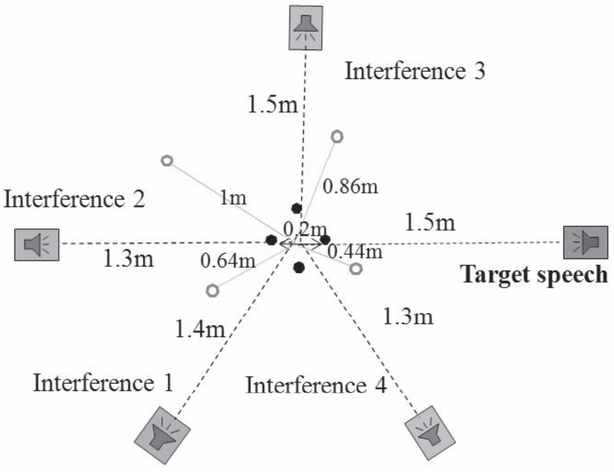

Fig. 1: Overhead view of the room set-up for uniform (black dots) and random microphone array (white dots)

$$\beta = \exp(-13.82/[c(L_x^{-1} + L_y^{-1} + L_z^{-1})T]), \tag{10}$$

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

3.3. Speech Separation Performance

The speech separation experiments are performed using different sparse recovery approaches to incorporate the block inter-connection as well as harmonicity of the spectro-temporal coefficients of speech signal. The quality evaluation results are summarized in Table 1. The block-size (B) is equal to 4 as it yields the best results specially for B-OMP and B-L_1L_2. In the harmonic model, we consider that $f_0 \in [150 - 400]$ Hz. The frequencies which are not the harmonics of f_0 are recovered independently in H-IHT and H-L_1L_2; we also considered that the harmonic structures are non-overlapping and k spans the full frequency band. For H-OMP, the harmonic subspaces are used to select the bases while projection is performed for the full frequency band.

As the results indicate, we observe that the least distortion (SDR) and the highest perceptual quality (PESQ) are obtained by convex optimization. This could be due to the zero-forcing spirit of greedy approaches. This deficiency is particularly exhibited for speech-like signals which do not possess high compressibility [7].

Table 1: Quality evaluation of the separated speech using different sparse recovery approaches. The quality (Q.) of the baseline overlapping mixture measured at the center of the array in terms of SDR, SIR, SAR, ISR and PESQ are -3.3, -3.68, 19.55, -1.56 and 1.44 respectively. The first row corresponds to uniform compact microphone array and the second row corresponds to random microphone array set-up

Q.	B-IHT	H-IHT	B-OMP	H-OMP	B-L_1L_2	H-L_1L_2
SDR	5.93	6.9	9.96	5.8	10.52	9.85
SIR	18.4	20.82	21.68	16.9	21.45	19.83
SAR	6.12	7.35	10.56	6.05	13	12.31
ISR	13.24	13.79	18.57	13.25	14.6	14.58
PESQ	2.26	2.35	2.49	1.63	2.77	2.55
SDR	7.35	8.72	11.7	8.3	13.18	10.76
SIR	19.3	21.38	23	19	22.83	20.8
SAR	7.6	9.16	12.27	8.5	14.57	12.2
ISR	15.83	16	21.58	18.58	20.27	17.12
PESQ	2.33	2.36	2.69	2	2.83	2.52

However, in some applications such as speech recognition, where the reconstruction of the signal is not desired, we can exploit the sparsity of the information bearing components in greedy sparse recovery approaches which offer a noticeable computational speed in efficient implementations [12, 14] and a reasonable performance.

Considering the speech signal model consisted of voiced and unvoiced segments, the block-interdependency mostly corresponds to the unvoiced speech while the harmonicity is exhibited in the voiced segments; hence we expect that a combination of both of the structures is beneficial for structured sparse recovery of speech signal. Furthermore, the proposed framework can be used for multi-speaker localization. We observe that when the number of microphones is very small, considering large block sizes and harmonicity has a significant impact on localization accuracy but, in terms of signal recovery, large block sizes result in some artifacts in signal reconstruction.

Comparing the results with the conventional uniform-array, we observe that the random setting of microphone array can significantly improve the quality of the separated speech. Hence, the compact uniform microphone array set-up is not an optimal design from the CS standpoint. Since our method outperforms the highest quality achieved by optimal beamforming with real data recordings [7], the present study encourages more investigation on sparse microphone array layouts.

4. CONCLUSIONS

We compared different structured sparse recovery approaches for separation of the convolutive speech mixtures. These structures incorporate the block inter-dependency as well as harmonicity of the spectro-temporal representation of speech. In this framework, we proposed a technique to estimate the acoustic parameters by recovering a single speaker images in a free space model. We then studied the theoretical relationship between the characteristics of the measurement matrix and sparse recovery performance which motivates random microphone arrays with a large aperture size. Our experiments developed on a set-up in accordance to these insights show a substantial improvement over the conventional compact microphone arrays.

5. REFERENCES

[1] M. A. Dmour and M. E. Davies, “A new framework for underdetermined speech extraction using mixture of beamformers,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 445–457, 2011.

[2] S. Makino, T. Lee, and H. Sawada, “Blind speech separation,” *Springer*, 2007.

[3] R. Gribonval and S. Lesage, “A survey of sparse component analysis for blind source separation: Principles, perspectives, and new challenges,” in *ESANN, 14th European Symposium on Artificial Neural Networks*, 2006.

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

[5] L. Carin, D. Liu, and B. Guo, “Coherence, compressive sensing and random sensor arrays,” *IEEE Antennas and Propagation Magazine*, 2011.

[6] A. Asaei, H. Bourlard, and V. Cevher, “Model-based compressive sensing for multi-party distant speech recognition,” in *Proceedings of ICASSP*, 2011.

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

[8] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, “Model-based compressive sensing,” *IEEE Transactions in Information Theory*, 2010.

[9] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of Acoustic Society of America*, vol. 65, 1979.

[10] I. Dokmanic, Y. Lu, and M. Vetterli, “Can one hear the shape of a room: The 2-D polygonal case,” in *Proceedings of ICASSP*, 2011.

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

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

[13] R. Gribonval and E. Bacry, “Harmonic decomposition of audio signals with matching pursuit,” *IEEE Transactions on Signal Processing*, vol. 51, pp. 101–111, 2003.

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

[15] E. van den Berg and M. P. Friedlander, “Probing the pareto frontier for basis pursuit solutions,,” *SIAM Journal on Scientific Computing*, 2, 2008, code available online, `http://www.cs.ubc.ca/labs/scl/spgl1`.

[16] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation (code available at http://www.irisa.fr/metiss/sassec07/?show=results),” *IEEE transactions on audio, speech, and language processing*, vol. 14, 2006.

[17] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, “Perceptual evaluation of blind source separation for robust speech recognition,” *Signal Processing, implementation available at*, `http://www.utdallas.edu/~loizou/speech/software.htm`.