Multi-party Speech Recovery Exploiting Structured Sparsity Models

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

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

Abstract

We study the sparsity of spectro-temporal representation of speech in reverberant acoustic conditions. This study motivates the use of structured sparsity models for efficient speech recovery. We formulate the underdetermined convolutive speech separation in spectro-temporal domain as the sparse signal recovery where we leverage model-based recovery algorithms. To tackle the ambiguity of the real acoustics, we exploit the Image Model of the enclosures to estimate the room impulse response function through a structured sparsity constraint optimization. The experiments conducted on real data recordings demonstrate the effectiveness of the proposed approach for multi-party speech applications.

Index Terms: speech sparsity, structured sparsity models, underdetermined convolutive speech separation, Image Model

1. INTRODUCTION

Hands-free speech communication using microphone arrays has been an active research area, playing a key role in many applications involving distant-speech recognition, scene analysis and videoconferencing. Despite the vast efforts devoted to the issues arise in real-world conditions, development of systems to operate in acoustic clutter of unknown competing sound sources yet remains a demanding challenge.

Previous approaches to speech separation can be loosely grouped into three categories. The first category relies on spatial filtering techniques based on beamforming to capture a specific target by steering the beam pattern of a microphone array [1]. The second category incorporates the statistical characteristics of the sources to identify the mixing model. The sources are usually recovered from the mixtures by least square optimization or matrix pseudo-inversion [2]. The third category is based on sparse representation of the signal, also known as sparse component analysis (SCA) [3]. These techniques basically exploit a prior assumption that the sources have a sparse representation in a known basis or frame. The assumption 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.

In [4], we show that sparse component analysis is in fact a highly potential approach to deal with overlapping problem in speech recognition. We observe compelling evidence that sparse recovery formulations could preserve the salient information to recognize speech with conventional speech recognition systems. In this paper, we investigate the theoretical guarantees of sparse recovery algorithms for spectrographic representation of speech. Our analysis motivates the use of structured sparsity models in sparse component analysis of speech signal. We then propose a framework of joint localization-separation of convolutive speech mixtures where we leverage model-based sparse recovery algorithms. To tackle the ambiguity of the real acoustics, the room impulse response has been modeled exploiting the structured sparsity obtained by the Image Model of the enclosures. We incorporated this model in a structured sparsity constrained formulation of the cross-relation optimization [5] to estimate the impulse response of the room for a particular source position as well as the corresponding reflection ratios of the large surfaces. This model enables us to identify the room impulse response for all locations in the room and we subsume it into our formulation for separation of the desired speech using model-based sparse recovery.

The rest of the paper is organized as follows: Sparsity of spectro-temporal representation of speech is investigated in Section 2. We adopt our formulation of the convolutive source separation presented in [4] and propose a structured sparse acoustic modeling approach to tackle the ambiguity of the real acoustics in Section 3. The experimental results are covered in Section 4. Conclusions are drawn in Section 5.

2. Analysis of Speech Sparsity

A signal $Z \in \mathbb{R}^G$ is N-sparse if only $N \ll G$ entries of Z are nonzero. We call the set of indices corresponding to the nonzero entries as the support of Z. Many natural and manmade signals are not strictly sparse, but they can be closely approximated as such if the absolute value of their support when sorted decay according to the power law [6]

$$|Z_{I(i)}| \leq Pi^{-1/r}, \quad i = 1, ..., N, \quad \text{and} \quad r < 1, \qquad (1)$$

where I indexes the sorted coefficients in magnitude. Defining the Z_K as the best K-term approximation of Z, which is obtained by keeping just the first K terms in $Z_{I(i)}$, the K-sparse approximation error when measured in the ℓ_p norm would have a power law decay exponent ρ as K increases:

$$\|Z - Z_K\|_p \leq (r\rho)^{-1/p} PK^{-\rho}, \qquad (2)$$

with $\rho = \frac{1}{r} - \frac{1}{p}$.

Relying on the key characteristic of sparse representation, the G-dimensional data can be stably recovered from $M = O(N \log(G/N))$ dimensionality reducing while information preserving measurements through efficient optimization algorithms which search for the sparsest solution [4, 6]. The sparse signal recovery algorithms offer provable guarantees for the recovery error of the signals characterized in (1) as

$$\|Z - \hat{Z}_K\|_2 \leq C_1 \|Z - Z_K\|_2 + C_2 \frac{1}{\sqrt{K}} \|Z - Z_K\|_1; \quad (3)$$

with the constants C_1 and C_2 depending on the sparse recovery algorithm.

To investigate the sparsity of spectro-temporal representation of speech signal, 25 speech utterances are taken from

TIMIT database and analyzed by short Time Fourier Transform (STFT) with different window sizes. Table 1 summarizes the percentage of the coefficients required for 10 dB reconstruction of the spectro-temporal components where $C_1 = 0$ and $C_2 = 2$ are the theoretical lower-bounds. In addition to the clean speech, we have studied the reverberant speech in two acoustic conditions with moderate reverberation and high reverberation with the corresponding 200 ms and 500 ms reverberation time. The concept of room Reverberation Time basically defines the time required for the multi-path energy to decay 60 dB from the direct path; hence denoted by RT_{60}.

As the results indicate, the sparsity is preserved in reverberant speech. The maximum sparsity is obtained for 64 ms analysis window for clean as well as moderately reverberant speech while in highly reverberant conditions, larger windows seem to give sparser coefficients. We have already seen in [4] that with less than 30% of the time-frequency coefficients, it is possible to perform word recognition with more than 90% accuracy. This observation verified the hypothesis that the information bearing components for Automatic Speech Recognition (ASR) are sparse; hence could be applied in the framework of SCA for multi-party ASR. The discrepancy between speech sparsity in terms of speech reconstruction and recognition, motivates ASR-specific sparse representation. These results further motivate exploiting the underlying structure of the sparse coefficients to reduce the number of required measurements and improve the recovery performance from very few observations [6, 7].

Table 1: Percentage of coefficients required for 10 dB reconstruction of speech spectro-temporal representation

STFT Window (ms)	32	64	128	256
Clean	32.79	31.72	32.63	34.79
RT_{60} = 200ms	31.92	30.34	30.46	32.09
RT_{60} = 500ms	38	36.38	35.6	35.57

3. Multi-party Speech Recovery

3.1. Problem Statement

We consider an approximate model of the real environment as a linear convolutive mixing process, stated concisely as:

$$x_j(n) = \sum_{i=1}^{N} h_{ji}(n) * s_i(n), \quad j = 1, ..., M; \tag{4}$$

where s_i refers to the source signal i passing through the room acoustic channel and recorded at sensor j (x_j). The notation $*$ stands for convolution. The number of sources is N and the number of microphones is M. The room impulse response from source i to sensor j is approximated by the filter h_{ji}. This formulation is stated in time domain. To represent it in a sparse domain, we consider the spectro-temporal representation of speech signals.

Our objective is to separate the N sources from M convolutive mixtures while $M < N$. Neither the number of sources nor the source signals are assumed known so the scenario is blind. We cast the underdetermined source separation problem in spectro-temporal domain as a sparse signal recovery where we exploit the underlying structure of the sparse coefficients to obtain more efficient signal reconstruction. The structured sparsity models are discussed in Sections 3.2 and 3.3.

3.2. Structured Sparse 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 $N \ll G$. We then define a G-dimensional vector with components of the signal coming from each grid. We now 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 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. Hence, the spatio-spectral representation is a vector with $F \times G$ components obtained as $Z = [Z_1 ... Z_G]^T$.

Note that due to the spatial sparsity, there is a block-structure underlying the sparse coefficients which could be exploited in sparse recovery algorithms to improve the efficiency of the sparse recovery by limiting the degrees of freedom of the sparse signal within a block configuration [6, 7].

3.3. Structured Sparse Acoustic Modeling

We consider the room acoustics as a rectangular enclosure consisting of finite-impedance walls. Taking into account the physics of the signal propagation and multi-path effects modeled by the Image Method [8], the Room Impulse Response (RIR) is the time domain Green's function with the particular form of

$$h(n, \mu, \nu) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu - \nu_r\|^\kappa} \delta\left(n - \frac{\|\mu - \nu_r\|}{c}\right), \tag{5}$$

where ι corresponds to the reflection ratio of the walls when the signal is reflected r times. In practice, ι has a different value for each wall, but for the sake of brevity we keep it constant here. The ν_r refers to the source distances to the microphone: ν_0 corresponds to the direct-path, and $\nu_{1,...,R}$ refer to the multi-path effect due to the contributing images within a radius given by the speed of sound (c) times the reverberation time. The attenuation constant κ depends on the nature of the propagation and is considered in our model equal to 1 which corresponds to the spherical propagation. Note that this model implies the sparsity of the high energy components of the acoustic channel. Given the geometry of the reflection surfaces, the support of the sparse coefficients is known.

Our objective is to estimate the RIR function for all grid locations. We assume that the geometry of the room is known. Hence, given the impulse response for a particular point, we can estimate the reflection coefficients of our model stated in (5) by least squares fitting. We use the cross relation technique presented in [5] for the blind estimation of the impulse response. Assuming that there is only one source active, the recorded signal at a pair of microphones can be expressed as:

$$x_i(n) = h_i(n) * s(n), \quad x_j(n) = h_j(n) * s(n). \tag{6}$$

It is straightforward to see that

$$x_i(t) * h_j(t) = x_j(t) * h_i(t); \tag{7}$$

considering an L-tap acoustic filter, for $n = L, ..., T$, where T is the length of the recorded signal, (7) becomes:

$$[\chi_i(L) - \chi_j(L)] \begin{bmatrix} h_j \\ h_i \end{bmatrix} = 0, \tag{8}$$

where $h := [h(L), ..., h(0)]^T$ and

$$\chi(L) = \begin{bmatrix} x(L) & x(L+1) & \dots & x(2L) \\ x(L+1) & x(L+2) & \dots & x(2L+1) \\ \vdots & \vdots & \ddots & \vdots \\ x(N-L) & x(N-L+1) & \dots & x(N) \end{bmatrix}. \tag{9}$$

This equation forms the basic idea for blind channel identification by least square optimization [5]. We propose a structure sparsity constraint to capture the main reflections characterized by the Image Model. The structured sparsity is theoretically motivated due to the fact that the multi-path signal energy is a function of the reflective areas. Hence, for the general environment of the meeting rooms, many objects are acoustically transparent [9]. In addition to the theoretical evidence, we empirically verified the effectiveness of the structure sparsity constraint for identification of the real acoustic impulse responses from noisy reverberant data generated by the impulse responses available at AIR database [10].

Given the room geometry and the source location, the support of the highest energy components of the RIR is determined by the Image Model and denoted by Ω_d which refers to the direct path component calculated precisely as α and Ω_r which refers to the support of the reflections. We define $\Pi := [\chi_i(L) - \chi_j(L)]$ and $H := [h_j h_i]^T$. The structure sparse acoustic filter will be obtained through the following optimization

$$\begin{aligned} &\hat{H} = \arg\min \|H\|_1 \\ &s.t. \quad \|\Pi H\|_2 \leq \epsilon, \quad H(\Omega_d) = \alpha, \quad H(\Omega_r) > 0 \end{aligned} \tag{10}$$

The estimated RIR is then used for estimating the reflection ratios by minimizing the mean squared error between the estimated RIR and all possible filters that our Image Model could generate. Hence, the RIR function is identified for all grid positions.

3.4. Model-based Sparse Signal Recovery

Given the source-sensor transfer function characterized in (5) and estimated through (10), we define $\Xi_{\nu_i \to \mu_j}$, a diagonal matrix consisting of the Fourier coefficients of the RIR function between sensor position ν_i and source position μ_j. Following from the convolution-multiplication property of the Fourier transform, the observation mixture recorded at sensor i can be expressed as $X_i = \phi_i Z$ where

$$\phi_i = [\Xi_{\nu_1 \to \mu_i} ... \Xi_{\nu_g \to \mu_i} ... \Xi_{\nu_G \to \mu_i}], \tag{11}$$

and ϕ_i is the i^{th} sensor's measurement matrix. We express the signal ensemble as a single vector $X = [X_1^T ... X_M^T]^T$, where each X_m is an $F \times 1$ vector consisting of the spectral coefficients of the signal at microphone m. Similarly, we concatenate each sensor measurement into a single measurement matrix $\Phi = [\phi_1 ... \phi_M]^T$. The sparse vector Z generates the signal ensemble as $X = \Phi Z$.

We use a model-based sparse recovery approach proposed in [7] to recover Z. To incorporate the underlying structure of the sparse coefficients, a model approximation is performed along with a gradient calculation at each iteration. Since the sparse coefficients in our model live in at most N blocks, an N-block-sparse signal is approximated by reweighting and thresholding the energy of the blocks [7]. The recovered signal Z contains the contribution of each speaker to the actual sensor observations in the block corresponding to the speaker position. We refer to our method as Blind Source Separation via Model-based Sparse Recovery (BSS-MSR).

4. Experiments

4.1. Overlapping Speech Database

The experiments are performed in the framework of Multichannel Overlapping Numbers Corpus (MONC) [11]. This database is acquired by playback of utterances from the original Numbers corpus. The recordings were made in a $8.2m \times 3.6m \times 2.4m$ rectangular room containing a centrally located $4.8m \times 1.2m$ rectangular table. The positioning of loudspeakers was designed to simulate the presence of 3 competing 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 recorded the mixtures. The speech signals are recorded at 8 kHz sampling frequency.

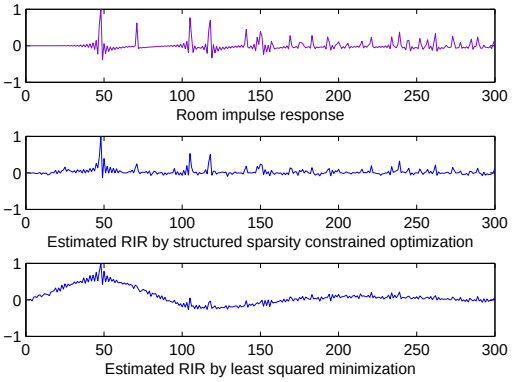

Figure 1: Room Impulse Response (RIR) estimation from noisy measurements

4.2. Room Acoustic Modeling

We used the CVX software package [12] for optimization formulated in (10) while sigma is chosen 0.1. The data was provided by concatenating 20 single speaker speech utterances.

The super-resolution source localization is performed based on the energy recovered from each grid location with our sparse recovery algorithm while the forward model consists in the direct path [4]. The support of the sparse coefficients was determined considering a 6 sided model of an enclosure with a known geometry. We assumed that the reflections of the carpet floor are trapped under the table; hence, the meeting table was considered as the floor in our Image Model. The room reverberation time is measured about 100 ms from the energy decay curve of the estimated RIR and the reflection coefficients are estimated as 0.1 for the walls as well as the ceiling and 0.6 for the meeting table. Our estimation matches the empirical Sabin-Franklin's formula [13]:

$$RT_{60} = \frac{24 \ln(10) V}{c \sum_{i=1}^{6} S_i (1 - \iota_i^2)}, \tag{12}$$

where V denotes the volume of the room, and ι_i and S_i denote the reflection coefficient and the surface of the i^{th} wall, respectively. Though our method is blind, we verified the estimated impulse response and the corresponding reflection coefficients through adaptive filtering technique using the original clean speech provided at MONC from the original Numbers corpus. Figure 1 shows the effectiveness of the room impulse response estimation with the structured sparsity constraints and the alternative least squared optimization from noisy data.

4.3. Speech Separation Performance

In [4], we provide speech recognition results of the separated speech using model-based sparse recovery algorithm. The study presented here provides a complementary insight to analyze the effectiveness of our proposed BSS-MSR approach with real data recordings while the structured sparsity models are incorporated to tackle the ambiguity of the room acoustics. We evaluate the quality of the recovered speech using Weighted Spectral Slope (WSS) distortion measure as well as Perceptual Evaluation of Speech Quality (PESQ) [14]. These objective measures are motivated due to the highest correlation (above 84%) with word recognition rate in low to moderately reverberant conditions [14].

Alternative to the sparse component analysis, the geometric source separation can be performed by beamforming. We used the multi-source Steered Response Power (SRP) approach presented in [15] for speaker bearing estimation for far-field beamforming while super-resolution speaker localization based on model-based sparse recovery (MSR) enables us to perform near-field beamforming. The super-directive (SD) as well as delay-and-sum (DS) beamformers are used for separation of the desired speech signal from the competing sources. As another alternative, we have compared our method with the convolutive-BSS (C-BSS) approach proposed in [2] which has been shown to be effective for speech recognition in multi-party scenarios. The speech separation performances are summarized in Table 2. The spectro-temporal representation is obtained by windowing the signal in 128ms frames using Hann function with 50% overlapping.

Table 2: Quality evaluation of the separated speech using DS and SD beamformers while the speaker is localized either by SRP or MSR, vs. C-BSS vs. BSS-MSR; B. and L. stand for the baseline and the lapel microphone respectively.

N	M.	B.	L.	SRP-DS	SRP-SD	MSR-DS	MSR-SD	C-BSS	BSS-MSR
2	WSS	66.1	50.1	56	47.2	55.8	45.3	49.3	44
	PESQ	1.8	2.3	2.3	2.6	2.3	2.6	2.5	2.6
3	WSS	76.3	53.3	64.5	55.2	64.2	52.3	55.1	52
	PESQ	1.6	2.2	2.1	2.4	2.1	2.4	2.3	2.4

As the results indicate, the proposed method based on model-based sparse recovery yields the least distortion in terms of WSS and the highest perceptual quality. We have set up a demo page for BSS-MSR subjective evaluations [4]. The best performance of BSS-MSR is obtained through a two-step procedure: (1) speaker localization and (2) separation by projection. The number of speakers is determined by the algorithm by energy thresholding of the estimated sources. Furthermore, incorporating the MSR super-resolution localization framework into the spatial filtering improves the performance of the beamformers and they outperforms the convolutive BSS approach relying on speech characteristics. There is more to investigate on the effect of reverberation suppression of spatial filtering and the reverberation cancellation obtained by BSS-MSR.

5. Conclusions

We presented rigorous analysis of speech sparsity in spectro-temporal domain. Our observation motivates exploiting structured sparsity models for efficient speech recovery. Relying on this evidence, we propose a framework for speech separation from under-determined convolutive mixtures exploiting structured sparsity models for acoustic modeling as well as signal recovery. The results on real data recording verify the effectiveness of our proposed scheme for practical hands-free multi-party applications. We also observe that there is a discrepancy between speech sparsity in terms of speech recognition and reconstruction which motivates ASR-specific sparse representation to be integrated in sparse component analysis techniques to achieve robustness in multi-party scenarios.

6. Acknowledgements

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.

7. References

[1] L. C. Parra, C. V. Alvino, L. Tong, and T. Kailath, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, 2002.

[2] L. C. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, 2000.

[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] A. Asaei, H. Bourlard, and V. Cevher, "Model-based compressive sensing for multi-party distant speech recognition," in *Proceedings of ICASSP*, `http://www.idiap.ch/~aasaei/BSS-MSR-Demo.html`.

[5] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, 1995.

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

[7] V. Cevher, "An ALPS view of sparse recovery," in *Proceedings of ICASSP*, 2011.

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

[9] D. Ba, F. Ribeiro, C. Zhang, and D. Florencio, "L1 regularized room modeling with compact microphone arrays," in *Proceedings of ICASSP*, 2010.

[10] "Aachen Impulse Response (AIR) database - version 1.2," Institute of Communication Systems and Data Processing (IND), RWTH Aachen University, 2010, `http://www.ind.rwth-aachen.de/AIR`.

[11] "The Multichannel Overlapping Numbers Corpus," Idiap resources available online:, `http://www.cslu.ogi.edu/corpora/monc.pdf`.

[12] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," `http://cvxr.com/cvx`.

[13] E. A. P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[14] 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`.

[15] M. J. Taghizadeh, P. N. Garner, H. Bourlard, and H. R. Abutalebi, "An integrated framework for multi-channel multi-source speaker localization and source activity detection," in *Proceedings of HSCMA*, 2011.