MODEL-BASED COMPRESSIVE SENSING FOR MULTI-PARTY DISTANT SPEECH RECOGNITION

Afsaneh Asaei[1,2]*, Hervé Bourlard*[1,2] *and Volkan Cevher*[1,2]

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

ABSTRACT

We leverage the recent algorithmic advances in compressive sensing, and propose a novel source separation algorithm for efficient recovery of convolutive speech mixtures in spectro-temporal domain. Compared to the common sparse component analysis techniques, our approach fully exploits structured sparsity models to obtain substantial improvement over the existing state-of-the-art. We evaluate our method for separation and recognition of a target speaker in a multi-party scenario. Our results provide compelling evidence of the effectiveness of sparse recovery formulations in speech recognition.

Index Terms— Model-Based Compressive Sensing, Sparse Component Analysis, Sparse Recovery, Overlapping Speech, Speech Recognition

1. INTRODUCTION

One of the major challenges of speech recognition systems in realistic environments and distant-talking applications is the common existence of overlapped speech segments which has been shown to increase the speech recognition word error rate up to 30% for a large vocabulary task [1]. Therefore, to design a speech recognition system for multi-speaker environments, it is required to incorporate an effective source separation technique in the front-end processing to separate the desired speech from the competing signals prior to recognition.

Sparse Component Analysis (SCA) is a Blind Source Separation (BSS) technique exploiting a priori 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 BSS) [2, 3]. The common SCA practice for degenerate unmixing is a two-step procedure: (1) Estimation of the mixing process and (2) Separation of the sources. While both of these steps take advantage of the sparsity of representation, they are usually performed independent of each other and the use of sparse recovery algorithms has been confined to the source separation at individual frequency level [2, 4].

In [5], we showed that sparse component analysis is in fact a highly potential approach to deal with overlapping problem in speech recognition systems. We achieved excellent word recognition rate with conventional speech recognition under the assumptions that there in no reverberation in the room hence the spatial cues are reliable to estimate the mixing process and recover the sources. It has been shown in [6] that despite the degradation of the spatial cues in reverberant conditions, the components of the overlapping speech in spectro-temporal domain remain disjoint and sparse. This observation motivated us in this research to formulate the under-determined source separation as a sparse recovery problem from dimensionality reducing measurements where we could leverage the compressive sensing theory. Contrary to the common SCA practice, our formulation merges the two steps of mixing process estimation and source separation as a joint localization-separation framework.

In this paper, we consider the echoic mixture of competing signals in spectro-temporal domain. We adopt a localization framework proposed by [7] in the context of sensor networks and discretize the planar area of the room into dense grids where the representation of sources exhibits spatial sparsity. We exploit spatial sparsity in tandem with spectral sparsity to obtain a sparse representation of signal where the sparse coefficients hold a block inter-dependency structure. We further exploit this structure in our sparse recovery algorithm for an efficient source extraction.

The rest of the paper is organized as follows: In Section 2 we explain the fundamental premises underlying compressive sensing. We set up the formulation of the convolutive source separation using compressive measurements in Section 3. Section 4 covers the details of the experiments. Conclusions are drawn in Section 5.

2. COMPRESSIVE SENSING BACKGROUND

Compressive Sensing (CS) exploits sparsity to acquire high-dimensional signals using very few linear non-adaptive measurements. A signal Z in a G-dimensional space is N-sparse if only $N \ll G$ entries of Z are nonzero. We call the set of indices corresponding to the non-zero entries as the support of Z. The CS theory indicates that such a signal can be sampled and reconstructed with only $M = O(N \log(G/N))$ linear measurements:

$$X = \Phi Z \tag{1}$$

where X are the measurements and Φ is an $M \times G$ measurement matrix. A sufficient but not necessary condition on Φ to recover the signal is a Restricted Isometry Property (RIP). Defining ℓ_p-norm of Z as $\|Z\|_{\ell_p} := (\sum_i |Z_i|^p)^{1/p}$ and an isometry constant δ_N of a matrix Φ as the smallest number such that

$$(1 - \delta_N)\|Z\|_{\ell_2}^2 \leq \|\Phi Z\|_{\ell_2}^2 \leq (1 + \delta_N)\|Z\|_{\ell_2}^2, \tag{2}$$

the matrix Φ holds RIP property if δ_N is not too close to one. This property implies that all pairwise distances between N-sparse signals must be well preserved in the measurement space or equivalently all subsets of N columns taken from the measurements are in fact nearly orthogonal.

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.

Relying on the two ingredients (1) sparse representation and (2) incoherent measurement, CS guarantees to circumvent the ill-posedness of the problem and recover the N-sparse signal stably from the compressive measurements by efficient optimization algorithms which search for the sparsest signal that agrees with those measurements.

In practice, signals may not be exactly sparse but they could be applied in CS framework if the support of the coefficients have a rapid power-law decay when sorted hence, called compressible signals. A new paradigm in CS exploits the inter-dependency structure underlying the support of the sparse coefficients in recovery algorithms to reduce the number of required measurements and to better differentiate true signal information from recovery artifacts which leads to a more robust and efficient recovery [8, 9].

3. BLIND SOURCE SEPARATION FROM COMPRESSIVE MEASUREMENTS

3.1. Problem Definition

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

$$x_j(n) = \sum_{i=1}^{N} \sum_{l=1}^{L} h_{ji}(l) s_i(n-l+1), \quad j = 1, ..., M; \quad (3)$$

where s_i refers to the source signal i passing through the room acoustic channel and recorded at sensor j (x_j). 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 an L-tap filter h_{ji}. This formulation is stated in time domain.

To represent it in a sparse domain, we consider the Gabor expansion, i.e., the discrete Short-Time Fourier Transform (STFT) of speech signals. Following from the convolution-multiplication property of the Fourier transform, the mixtures in frequency domain can be written as

$$X_j(\omega, \tau) = \sum_{i=1}^{N} H_{ji}(\omega) S_i(\omega, \tau), \quad j = 1, ..., M; \quad (4)$$

where $S_i(\omega, \tau)$ and $X_j(\omega, \tau)$ are the STFT of the original source i recorded at distant microphone j where the analysis window is centered at time τ and ω indicates the frequency. $H_{ji}(\omega)$ is the frequency domain source-sensor transfer function.

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 formulate the underdetermined source separation problem in spectro-temporal domain as a sparse approximation using compressive measurements. We already mentioned the two fundamental ingredients underlying CS: (1) sparse representation and (2) incoherent sensing. We briefly review these topics in sections 3.2 and 3.3.

3.2. Spatio-Spectral Sparse Representation

I. Spatial Sparsity: 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 grid selector vector θ with components θ_g that are 1 or 0 depending on whether or not a source is present at grid point g. With this notation, note that the number of sources N is equal to the ℓ_0-norm of θ, which is a pseudo-norm defined as the number of non-zero elements in the vector.

II. Spectral Sparsity: We consider the time-frequency (t-f) representation of speech signal. The analysis coefficients are compressible and exhibit a power-law decay [5]. Due to the spectral sparsity we make an assumption that at each t-f point, the number of active speakers is less than the number of mixtures (M). From the fact that speech representation in spectro-temporal domain is approximately disjoint [6, 5], this assumption is a reasonable hypothesis. In fact the assumption used in [6] that only one source is active at each t-f point is relaxed here to $K \leq M$ sources contribute to the same t-f component.

III. Spatio-Spectral Sparsity: We now entangle the spatial representation of the sources with the spectral representation and define a vector Z whose support is the t-f contribution of each source signal located at grid g. Suppose that the number of analysis coefficients is F. Considering the whole t-f components of each source signal, each element of z_g is an $F \times 1$ vector which carry the spectral coefficients coming from grid g. Hence the spatio-spectral representation is a vector with $F \times G$ components obtained as

$$Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_G \end{bmatrix}. \quad (5)$$

Note that due to the spatial sparsity, there is a block-structure underlying the sparse coefficients which could be exploited in CS recovery algorithms to improve the efficiency of the sparse recovery by limitting the degrees of freedom of the sparse signal within a block configuration [9, 8].

3.3. Incoherent Measurements

We consider the room acoustic as a rectangular enclosure consisted of finite-impedance walls. The point source-to-microphone impulse responses are calculated using Image model technique [10]. To model the measurements with distant microphones, we define a linear convolution operator for signal propagation denoted by $\xi_{\nu \to \mu}$, which takes a source signal s and its location ν and calculates the observed signal x at the location μ via

$$x = \xi_{\nu \to \mu}[s]. \quad (6)$$

Taking into account the physics of the signal propagation and multi-path effects, ξ represents the Green's function in frequency domain with the particular form of

$$\xi^{\omega}_{\nu \to \mu} : X(\omega, \tau) = \sum_{r=1}^{R} \frac{\iota^r}{\|\mu - \nu_r\|^{\alpha}} \exp(-j\omega \frac{\|\mu - \nu_r\|}{c}) S(\omega, \tau), \quad (7)$$

where $j = \sqrt{-1}$ and ι corresponds to the reflection ratio of the walls when the signal is reflected r times. Hence, for the direct-path signal, r is equal to 0. 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 times the reverberation time. The attenuation constant α depends on the nature of the propagation and considered in our model equal to 1 which corresponds to the spherical propagation. Given the source-sensor transfer function defined in (7), the observation mixture captured at sensor i can be expressed as $X_i = \phi_i Z$,

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

$$\Xi_{\nu_g \to \mu_i} = \begin{bmatrix} \xi^1_{\nu_g \to \mu_i} & 0 & \dots & 0 \\ 0 & \xi^2_{\nu_g \to \mu_i} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \xi^{\Omega}_{\nu_g \to \mu_i} \end{bmatrix}, \quad (8)$$

where 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 consisted of the whole t-f components of the signal at microphone m. Similarly, we concatenate each sensor measurement into a single measurement matrix

$$\Phi = \begin{bmatrix} \phi_1 \\ ... \\ \phi_M \end{bmatrix}. \quad (9)$$

The sparse vector Z generates the signal ensemble as $X = \Phi Z$. We consider a linear mapping of the observation to ensure incoherency [11]. Let Λ denote this mapping, i.e. $X' = \Lambda X$ and

$$\Lambda = P\Phi^{\dagger} \quad (10)$$

where $P = \text{orth}(\Phi^T)^T$. Note that orth$(A)$ is an orthogonal basis for the column space of matrix A. Then Z can be well recovered from X' since

$$X' = P\Phi^{\dagger}X = P\Phi^{\dagger}\Phi Z = PZ, \quad (11)$$

and P is an orthogonal matrix.

3.4. Block-based Sparse Recovery

We use a model-based CS recovery approach proposed in [8]. This algorithm is inspired by the development of the first order methods in optimization, most notably based on the algebra used in Nesterov's optimal gradient and smoothing techniques. The sparse recovery is performed by a gradient type of method where the Lipschitz gradient constant is used as the step size to guarantee the fastest convergence speed. To incorporate for 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 [8]. 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. Speech Database

The experiments are all performed in the framework of Aurora 2 [12]. This database is designed to evaluate the performance of speech recognition algorithms in noisy conditions. A fixed HTK back-end was trained on multi-condition data with different noise types including those of Subway, Babble, Car and Exhibition at 5 SNR levels as well as clean data. Overlapping speech was synthesized by mixing clean Aurora 2 test utterances with interfering sentences from HTIMIT database. The broad phonetic space of HTIMIT allows the results of our Aurora 2 framework to be generalizable for the task of digit recognition in overlapping conditions. The interference utterances are chosen randomly from a pool of 40 sentences balanced among male and female. For each test sample, interferences are randomly chosen out of this subset to construct two mixtures. The interference files are scaled prior to mixing to achieve the particular baseline and looped to compensate for the difference between the file lengths.

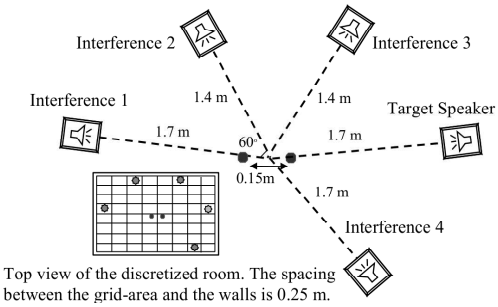

Fig. 1. Overhead view of the room set-up

4.2. Acoustic Parameters

The planar area of a room with dimension 3×4 is divided into grids with 50 cm spacing (hence 48 grids in total). The sources are assumed to have the same elevations as the sensors (located in the middle with 1.5 m height) and distributed as depicted in Fig. 1. The stereo mixtures are recorded from the room center. Room impulse responses are generated with the Image model technique [10] using intra-sample interpolation, up to 15th order reflections and omni-directional microphones. The corresponding reflection ratio, β used by the Image model was calculated via 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, c is the speed of sound in the air ($\approx 342m/s$) and T is the room reverberation time. In our experiments T = 200ms.

4.3. Analysis Parameters

The speech signals are recorded at 8 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. The separated speech is then reconstructed back into time domain. We found this reconstruction a convenient mean of changing the FFT size and period. The separated speech is then presented to the standard Aurora 2 speech recognition system. The speech signal is processed in blocks of 25 ms with a shift of 10 ms to extract 13 MFCC cepstral coefficients per frame. These coefficients after cepstral mean/variance normalization are appended to their delta and delta-delta derivatives to obtain a 39 dimensional feature vector for every 10 ms of speech.

4.4. Speech Recognition Performance

We conducted some experiments to evaluate the performance of the proposed method for speech recognition systems. Although any number of measurements could be accommodated in our framework, we carried out the experiments with stereo mixtures to compare the results with our previous work on SCA for speech recognition [5].

Due to the limited number of measurements, we first performed a basis reduction using the 20% high-energy coefficients of the observed data. In this procedure, each t-f coefficient is assigned to one of the grids by ℓ_1-minimization over the whole grid points. The

number of coefficients assigned to each grid denotes the activity of that region. Based on the activity obtained as such some of the grids are hypothesized as active and the rest are discarded in BSS-MSR. The number of active grids is upper-bounded with 15 (we assumed that the number of simultaneous speakers is always less than 15). This activity detection results in reducing the dimensionality of the sparse recovery problem and increases its efficiency. The source separation is then performed by block-sparse recovery of the spatio-spectral coefficients defined in 5. The target speech is selected based on the proximity to the position of interest. The forward model of the room impulse response is not given in the tests and it is approximated by taking an arbitrary value for the reflection coefficients. We repeated the experiments for β = 0.6, 0.7, 0.9 and 1 (considering $\beta \in [0.6, 1]$ for the typical room acoustic) and averaged the results.

We compared our method with the improved version of Degenerate Unmixing Estimation Technique (DUET) for speech recognition [6, 5]. We also ran the experiments on Line Orientation Separation Technique (LOST) [13] since the separation of sparse components in this method is based on ℓ_1-minimization at each t-f point independently. Tables 1-2 give the recognition results of the demixed speech for the clean and multi-condition training of Aurora 2 database. The baseline is word recognition rate of overlapping speech.

Table 1. Word accuracy of the separated speech for stereo echoic mixtures of 3 sources (interferences 1-2 and target speech).

Train Cond. (TC)	Baseline	DUET	LOST	BSS-MSR
Clean (C)	59.3	56.34	61.2	89.3
Multi-Cond. (MC)	61.78	77.25	64.3	92.7

Table 2. Word accuracy of the separated speech for echoic mixtures of 5 sources (interferences 1-4 and target speech). BSS-MSR^1 refers to the stereo recording and BSS-MSR^2 refers to 4-channel circular microphone array recording with the radius = 0.15m.

TC	Baseline	DUET	LOST	BSS-MSR^1	BSS-MSR^2
C	47.3	31.72	49.2	81.7	88.7
MC	58.19	53.32	50.6	91	94

As the results indicate, the proposed method based on model-based sparse recovery can effectively recover the desired speech from the overlapping mixtures. The estimation of the mixing process in reverberant enclosures is highly erroneous within the scheme of DUET and LOST and it results in a poor speech separation performance. However, BSS-MSR incorporates the reverberant mixing model in separation of the speech components.

The underlying block-sparse structure exists in the signal ensemble recorded by microphone array and could be applied in any sparse component analysis technique using multi-channel recordings. When a frequency sparse signal is recorded by an array of microphones, all of the recorded signals contain the same Fourier frequencies but with different amplitudes and delays. Such a signal ensemble can be vectorized by concatenation, and the coefficients can be rearranged so that the concatenated vector exhibits block sparsity [9]. Similarly, the virtual source images used in Image model technique to study the multi-path effect of the room acoustic exhibit a block-sparsity which could be exploited in further analysis of the room acoustic. The proposed method could also be used for accurate localization of simultaneous speakers.

5. CONCLUSIONS

We proposed a novel source separation technique for efficient recovery of convolutive speech mixtures in spectro-temporal domain using model-based compressive sensing theory. Contrary to the common practice in sparse component analysis, our formulation merges the two steps of mixing model estimation and source separation as a single step joint localization-separation based on spatio-spectral sparsity of overlapping speech signals. The method has been used for separation and recognition of a target speaker in a multi-party scenario. The results show that the model-based sparse recovery formulation is very effective for distant speech recognition systems in the presence of competing signals. The underlying block-sparse structure that we exploit exists in any signal ensemble recorded by microphone array. It is also exhibited in virtual source images due to the acoustic multi-path effect. The success of our proposed method motivates incorporation of this structure in sparse component analysis techniques using multi-channel recordings in reverberant conditions.

6. REFERENCES

[1] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proceedings of EUROSPEECH*, 2001.

[2] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, 2001.

[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] R. Saab, O. Yilmaz, M. J. Mckeown, and R. Abugharbieh, "Underdetermined anechoic blind source separation via ℓ_q-basis-pursuit with $q < 1$," *IEEE Transactions on Signal Processing*, 2007.

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

[6] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[7] 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.

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

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

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

[11] C. Feng, S. Valaee, and Z. Tan, "Multiple target localization using compressive sensing," in *Proceedings of IEEE Global Telecommunications Conference*, 2009.

[12] D. Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ICSLP*, 2000.

[13] P. O'Grady and B. A. Pearlmutter, "Soft-lost: EM on a mixture of oriented lines," *Proceedings of ICA, Lecture Notes in Computer Science, Springer-Verlag*, 2004.