Applying Compressive Sensing to the Cocktail Party Problem

Eli Saracino
Department of Computer Science
Boston University
Boston, MA 02135
esaracin@bu.edu

Andy Huynh
Department of Computer Science
Boston University
Boston, MA 02135
ndhuynh@bu.edu

Abstract

A Compressed Sensing framework to approach blind source separation is investigated. We compare previous attempts, such as Nonnegative Matrix Factorization and Independent Component Analysis, and convey the potential advantages of Compressive Sensing: namely, it providing a more efficient reconstruction that relies far less on the necessity of learning.

1 Introduction

The applications of the Cocktail Party Problem are far-reaching: from surveying and separating radio signals, to imagining and examining neural signals in a medical setting, source separation plays an important role in an ever-expanding number of fields. An efficient approach to solving the Problem, then, has implications that extend far beyond the realm of Computer Science.

Past techniques applied to solve this problem have, at a high level, relied largely on the learning of a specific dictionary, and the utilization of this dictionary to reconstruct distinct signals from potentially mixed sources. One such approach is Nonnegative Matrix Factorization, hereby NMF. As suggested, this method implies that prior information about speech sources is known in order to work properly, and, perhaps even more to its detriment, does not provide a well-defined solution in the case of over-complete dictionaries, which are so often utilized in the Compressive Sensing framework to minimize the number of measurements needed to reconstruct a given signal. **[1]** The NMF problem statement effectively tries to optimize the following cost function:

$$E = ||\mathbf{Y} - \bar{\mathbf{D}}\mathbf{H}||_F^2 + \lambda \sum_{ij} \mathbf{H}_{ij} \text{ s.t } \mathbf{D{,}H} \geq 0$$

where $\bar{D}$ is a column-wise normalized dictionary matrix and **H** is the sparse solution upon which an L_1 norm penalty is induced. In keeping with the theme of NMF, both $\bar{D}$ and **H** must be nonnegative. In the above statement, λ is a parameter that controls the degree of sparsity.

Notably, the problem statement is not totally dissimilar to that of the standard Compressive Sensing model, albeit with some minor differences. For example, the dictionary $\bar{D}$ must be learned, usually by training on a large dataset of a single speaker. This, of course, sacrifices efficiency and casts framework of the reconstruction as a learning problem.

With the advent of Compressive Sensing, however, we can model this problem in a new light. To do this, we first have make an assumption of sparsity; that is, in analyzing a mixed audio signal, at any given instant in time, we must assume that noise is coming from only a single source. In this way, the *frequency* domain of a mixed signal will be sparse, even if it should be the case that the the signal is

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

quite robust in the time domain. With this assumption, we can frame the problem as a case of sparse signal recovery, defined as follows:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

where **S** is our set of source signals and **X** is a mixture of these signals as determined by our mixing matrix, **A**. As we will go on to show, determining **A** becomes a key step in the Compressive Sensing approach to solving the problem.

While solving this problem subject to the minimization the L_0 norm of **s** is notably NP-hard, we can obtain an approximate solution using any one of several greedy algorithms. Specifically, we apply Orthogonal Matching Pursuit to reconstruct the separate sparse signals, and transform them back into the time domain to complete the recovery.

In the coming sections, we will describe in more detail the compressive sensing approach used and it's computational benefits over such techniques as NMF, as well as some of its own shortcomings.

2 Method

The following is an overview of the algorithmic steps we take in applying Compressive Sensing to the Cocktail Party problem to reconstruct distinct signals from mixed input sources.

In the following steps, we'll assume we have a mixed signal of size *T*. That is, there are *T* discrete time instants over which our signal is heard.

In keeping with our assumption of a frequency-sparse input signal, the first step we take is to compute the Short-time Fourier transform of our source signal in order to cast it into its time-frequency domain. Shown below is a figure that demonstrates the utility of this transformation: an input signal that was originally robust becomes relatively sparse, with distinct structure forming along specific axes.

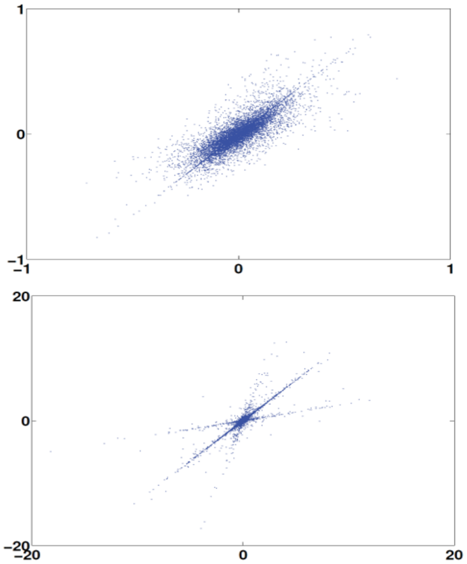

Figure 1: Above, a mixed signal in the Time domain. Below, in the Transform domain **[1]**

It now becomes our job to generate a mixing matrix, defined as **A** in the Introduction's problem statement, in order to create a mixture of signals, **X**, that is able to be used in the reconstruction. The first step in this process it to apply k-means clustering to the normalized STFT computed in the previous step. By setting *k* to the number of original sources we hope to separate, we are able to effectively partition the frequency domain of the transformed signal along the axes defined in the above figure.

Next, we use the one-dimensional cluster centers computed from the run of k-means to construct the mixing matrix. We'll define the following notation to express this process:

Let $C = \left(c_1 \; c_2 \; ... \; c_k \right)$ be the cluster centers previously computed.

Let L_i be a T x T diagonal matrix, of which there are k, and whose diagonal entries are equivalent to the corresponding i^{th} entry of C as defined above.

We can finally define the mixing matrix M, then, by concatenating these matrices horizontally, as follows: $M = \left(L_1 \; L_2 \; ... \; L_k \right)$

As a final preparatory step, we multiply this matrix, M, by a T x T DCT matrix to further sparsify the high-frequency components of human speech **[2]**. With this step, the preparation of our problem is complete, and the reconstruction of our separated signals can begin.

In practice, the reconstruction is often performed using a windowing method,as the dimensionality of the input signal, T, can often be quite high, and the matrices described, then, can be quite large. Choosing some l such that you compute the mixing matrix in l x l sections, and use it to reconstruct partial signals in a sectionalized way is a viable strategy, and it is what we employed in our reconstructions.

At each windowing phase, we perform Orthogonal Matching Pursuit to reconstruct the portion of the signal responsible for the window being processed. The concatenation of these windows together, then, yields the full, separated, original signal.

3 Results

In this section, we review the results of our reconstruction and compare it to other common methods of blind source separation, analyzing the optimal window.

calculate psnr for each window size, maybe chart for best reconstruction? results (comparison to other techniques (ICA) and NMF)

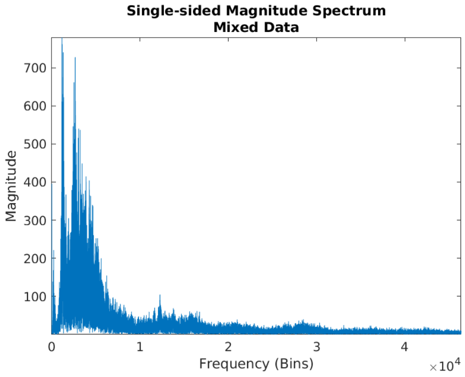

Figure 2: Sample figure caption.

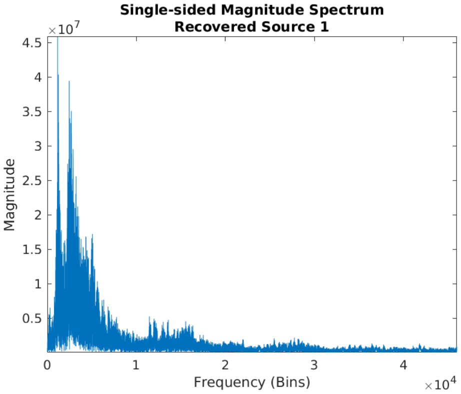

Figure 3: Sample figure caption.

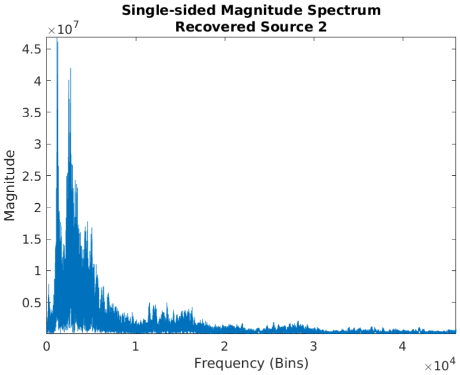

Figure 4: Sample figure caption.

References

[1] Schmidt, Mikkel N. and Rasmus Kongsgaard Olsson. "Single-channel speech separation using sparse non-negative matrix factorization." *INTERSPEECH* (2006).

[2] BAO, G., YE, Z., XU, X. AND ZHOU, Y. *A Compressed Sensing Approach to Blind Separation of Speech Mixture Based on a Two-Layer Sparsity Model* - IEEE Journals & Magazine Bao, G., Ye, Z., Xu, X. and Zhou, Y. (2017). A Compressed Sensing Approach to Blind Separation of Speech Mixture Based on a Two-Layer Sparsity Model - IEEE Journals & Magazine. [online] Ieeexplore.ieee.org. Available at: http://ieeexplore.ieee.org/document/6384713/