A Compressive Sensing Framework for Solving 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

This paper aims to propose a solution to the cocktail party problem via framing the problem in a compressive sensing setting. The solution proposed is based on the assumption voices are generally non-overlapping in a mixed measurement for a small enough time window and that voices in general only take on a set amount of frequencies. As a result the mixing matrix can be learned in this sparse domain in hopes of reversing the mixing of the sources. Additionally a dictionary for the sparse representation of the signal must be learned, thus we define a way to retrieve this mixing and sampling matrix in this paper.

1 Introduction

Human beings are excellent at being able to separate sound sources. Given our binaural hearing, we are able to easily tune out and focus on one voice within a group of many. One of the biggest open problems in signal processing is asking if it's possible to mimic this phenomenon on paper. This problem, of separating a single signal source from many is otherwise known as the Cocktail Party problem. Solutions to the Cocktail Party Problem could provide innovate a multitude of disciplines: 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.

Past techniques applied to solve this problem have, at a high level, relied 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 (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:

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

where $\bar{\mathbf{D}}$ is a column-wise normalized dictionary matrix and $\mathbf{H}$ is the sparse solution upon which an L_1 norm penalty is induced. In keeping with the theme of NMF, both $\bar{D}$ and $\mathbf{H}$ 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 the framework of the reconstruction as a learning problem.

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

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. For example, if the mixed signal is a bunch of voices, at any infintesimal time window only one voice is active. 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 quite robust in the time domain. With this assumption, we can frame the problem as a case of sparse signal recovery, defined as follows:

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

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.

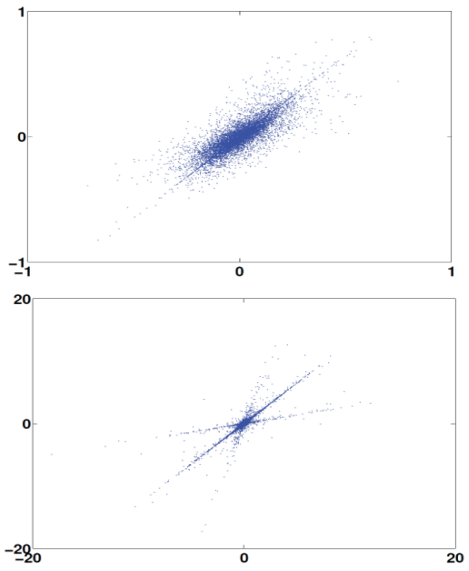


Figure 1: Above, a mixed signal in the Time domain. Below, in the Transform domain **[1]**

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 $C = \begin{pmatrix} c_1 & c_2 & \dots & c_k \end{pmatrix}$ 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: $M = \begin{pmatrix} L_1 & L_2 & \dots & L_k \end{pmatrix}$

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. The main .wav file used for the tests was a mixture of Clint Eastwood and Graham speaking over each other. Other samples included data from the Interspeech 2006 source separation challenge, and are comprised of concise sentences following the same general structure. The source .wav files were combined using the sox command line program.

In practice, the results of using Compressive Sensing for blind source separation were underwhelming. The figures below depict both the original, mixed, signal, and the resulting signals following the Compressed Sensing framework. In general, Independent Component Analysis (ICA), an industry standard of addressing the problem of blind source separation, is known to significantly outperform these results.

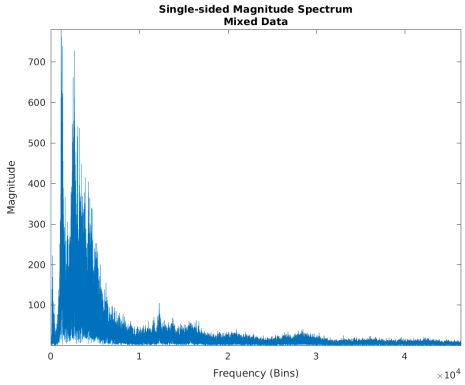
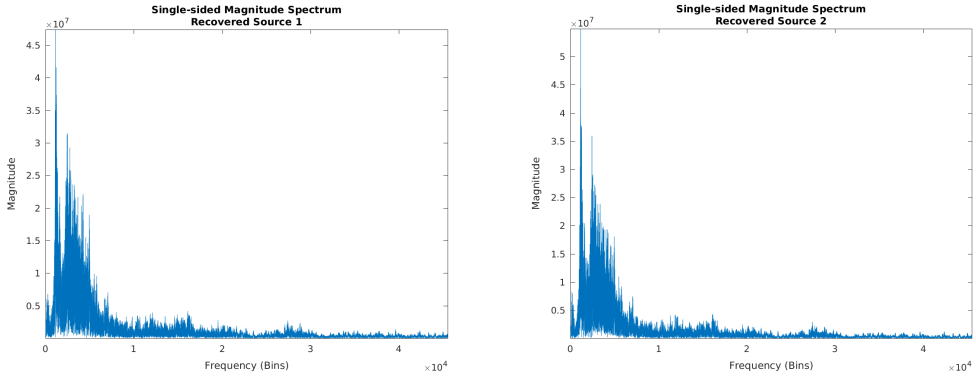


Figure 2: Original Signal



(a) SNR: -17.926 (b) SNR: -19.863

Figure 3: Separated Sources

Unfortunately these power spectrums shows the lower frequencies of the separated sourcs absolutely dominating. This is an indication that the noise tends to overpower any separation found in the proposed method. Comparing he magnitude of the recovered sources to that of the original signal shows this noise is much more prominent in the recovered. Thus our algorithm tends to be highly sensitive to any noise of the original signal. SNR being negative is also another indicative factor of the separation not going smoothly.

While these results are less than ideal, its worth noting that, though ICA provides more accurate source separation, it requires more a priori information to perform its separation than the compressed sensing approach. Namely, the number of measurements of the mixed signal must be at least equal to the number of distinct sources *in* that signal. In our framework, only a single measurement is necessary to attempt a separation.

The Cocktail Party Problem continues to be a fascinating and challenging problem today. It is clear from these experiments that, even with innovations such as compressed sensing, there is much work left to do to achieve perfect blind source separation.

Conclusion

This study provides an indepth analysiso on one possible technique for solving the Cocktail Party problem. Unfortunately results show this method is not as robust and provides poor results. In general standard optimization techniques or componenet analysis techniques may provide a better alternative. The problem stems from a model that relies heavily on too much inference off of a small amounts of data. Althought the compressive sensing framework gives us theoretical bounds on the information needed to recover a signal, there exist still complications in implementation as speech may not follow all the necessary assumptions.

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/