Title: CAREER: Scaling Source Separation to Big Audio Data

Project Summary: The world we live in is composed out of mixed signals. It is practically impossible to effortlessly obtain a clean recording of speech, music, environmental, mechanical, underwater, or biomedical sounds. Because of that, we often resort to source separation and mixture analysis methods that allow us to enhance a signal in order to facilitate further processing. A very successful approach to this type of processing makes use of training data that can assist in the extraction of a source or its parameters. However, for these methods we do not have algorithms that can scale to "big data" levels and are constrained to use small data sets that do not allow us to take advantage of extensive training. The goal of this proposal is to develop the foundations for efficient single-channel mixture-analysis, and to apply them on large-data problems. More specifically, what is proposed is a fundamental rethinking of popular non-negative factorization methods, taking advantage of recent developments in manifold structure, deflation, hashing and quantization approaches, which will for the first time enable the application of these techniques on both large-data and computation-constrained settings. We will focus on two fundamental algorithmic improvements: efficient model building from prohibitively large data sets, and tractable algorithms for deploying very large source models. Our contributions are also applicable to a large family of related models in wide use in acoustical signal processing (e.g. non-negative dynamical models, N-HMMs, convolutive decompositions, etc.), and will also enable them to operate in "big data" regimes, something that is currently impractical. We will validate our work with applications operating on large collections of recordings and performing popular mixture signal tasks such as source extraction, recognition of concurrent sources, and parameter estimation from noise-contaminated data.

Intellectual Merit: Single-channel source separation has long been a research area driven by accuracy metrics. However, with an abundance of available data, we see the additional need for efficiency. This work will undertake the first systematic study of algorithmic efficiency in this critical new "big audio data" context. It will draw from recent developments in efficient data processing and combine them with traditional signal processing and linear algebra primitives that are in wide use today. It will in effect answer two important questions. First, what are the algorithmic principles that will enable mixture-signal processing on a "big data" scale, and make it accessible to resource-constrained devices? And second, what new regime of results can we achieve if we use methods that can train on large-scale data and make use very large source models? These are both unexplored domains in acoustic signal processing, and hold the promise of validating the effectiveness of large-scale data usage in yet another research area.

Broader Impact: Being able to process mixed signals is a central process in our day-to-day life. It means we can design speech recognizers that are truly noise-invariant; medical sensors that require vastly less prep time; hearing aids that reliably allow the wearer to navigate in acoustically difficult environments; multimedia databases that can index huge, noisy audio assets quickly and correctly; surveillance systems that can track simultaneous targets, etc. Being able to do so with efficient processing models will translate to increased applicability of such technology to real-life problems. Additionally, and due to the fundamental nature of this work, these new models will also apply to other time-series such as sonar, communications, and bio-signals and are expected to be of broad use to these communities as well.

Keywords: *Audio signal processing, source separation, audio recognition*

Project description

Introduction

The ability to work with mixtures of signals is a necessity in today's acoustical systems. For many real-life applications we are presented with a target signal alongside interfering sources. This is the norm for a wide-range of fields such as speech processing, underwater acoustics, music information retrieval, bioacoustics, mechanical machinery monitoring, etc., where desired information is buried inside a mixture of signals. When presented with single-channel mixtures, a popular way to analyze them involves the use of a training set which represents the statistical characteristics of the target sources in the mixture. Recent advances have resulted in powerful mixture analysis tools, but they also command significant computational complexity. This complexity necessitates more processing power and time, and has hindered our exploration of big-data scales for such methods. Motivated by this problem, the main issues that we will address in this project are two: a) the ***development of efficient computational approaches that enable the deployment of single-channel mixture models on large amounts of data***, and b) the ***first study that charts the effects of large-scale training and models on such methods***.

Background

Single-channel acoustical source separation is the problem of extracting a target signal from a single-channel recording that contains the superposition of multiple signals. This is a very common problem that one finds in a wide breadth of applications, in which we might want to extract a signal (e.g. speech from noise, submarines from ocean clutter), or maybe just some of its parameters (e.g. a phoneme sequence, or notes played in a music piece). Both of these problems have been historically addressed using algorithms that perform denoising, source separation, or otherwise include some level of mixture-invariance. Although there is a considerable amount of such research in general (e.g. multichannel-systems using array methods, or domain-specific methods based on parameterizations), here we will focus on the general single-channel case for acoustical signals due to its wide applicability and the demanding computational complexity that it commands.

Extracting a target source from a single-channel recording is an ill-defined problem. More specifically, for a given mixture signal $m(t) = x(t) + n(t)$ that is comprised out of a target signal $x(t)$ and a set of interferences $n(t)$ we wish to extract the only target signal by itself. The main obstacle in this problem is to specify the target signal so that we can isolate it. Unlike multi-channel methods, there is no spatial domain or cross-sensor statistics to help us guide an algorithm to the target source. Thus, the majority of work on this problem has been centered on finding a way to effectively define the target signal.

During the last decade a particularly effective approach for analyzing single-channel mixture acoustical signals has been through the use of Non-Negative Factorizations, and their probabilistic formulations (Virtanen et al.). These methods construct an additive set of dictionaries for describing sounds and use these as models that specify what to extract from a mixture. The most basic of formulation starts with constructing a model of the target and interference signals using training data. For a specific type of signal (e.g. speech) one can learn a speech-specific dictionary using multiple clean speech signals, and likewise do the same for, say, the interfering street ambience. Once this training data is collected it is transformed to an energy frequency domain representation (such as a magnitude or power short-time Fourier transform) and represented as matrices $\mathbf{F}_i$ whose elements $f_{\omega,t}$ contain an energy measure at frequency ω and time t, for each source i. For each source, a *spectral dictionary* $\mathbf{W}_i$ is learned via the decomposition:

$$\begin{gathered}\mathbf{F}_i \approx \mathbf{W}_i \cdot \mathbf{H}_i \\ \mathbf{F}_i \in \mathbb{R}_+^{M\times N}, \mathbf{W}_i \in \mathbb{R}_+^{M\times K}, \mathbf{H}_i \in \mathbb{R}_+^{K\times N}\end{gathered} \tag{1}$$

where $\mathbb{R}_+^{M\times N}$ is the space of non-negative M by N matrices (i.e. matrices whose elements are all greater or equal to zero). This is known as the Non-Negative Matrix Factorization model (Lee and Seung 1999). The approximation in equation (1) can be defined in various ways, but in practice it is often expressed in terms of a modified Kullback-Leibler or the Itakura-Saito divergence (Lee and Seung 2001, Févotte and Idier 2011). Alternatively one can use a parameterized form employing Bregman divergences (Cichocki and Amari 2010), which includes the previous measures as special cases. Another popular option is the Euclidean distance, however in practice this results in significantly lower quality performance for acoustical source separation problems and will be explicitly avoided in our work. Depending on the selected divergence, the estimates for **W** and **H** will be obtained using an iterative process that we will discuss later on.

This decomposition approximates the time-frequency distribution of a certain type of signal using spectra constructed from a spectral *dictionary* $\mathbf{W}_i$, and a corresponding set of dictionary element *activations* $\mathbf{H}_i$. A key element of this decomposition is the parameter non-negativity. Due to that, the approximation of any input will have to be based on a superposition of spectra and will not make use of cross-cancellations. This turns out to be a crucial element as compared to other types of decompositions, one that forces the discovered spectra and activations to correlate better with our cognitive representations of audio mixtures (we think of sounds as an additive-only combination of elements, with no concept of removing parts to explain what we hear). Given adequate training data (usually in the order of a minute or so), one can learn a spectral dictionary that can adequately describe other recordings of the same type of sound. That means that a well-trained dictionary for a specific speaker, or a musical instrument, or an environmental sound will often be good enough to describe other instances of that type of sound.

When confronted with a matrix **M** that contains an energy time/frequency representation of a mixture containing previously known sources (i.e. sources for which we have a set of $\mathbf{W}_i$'s), we can use the same model to explain this mixture in terms of the contribution of each known source, using:

$$\mathbf{M} \approx \mathbf{W}_1 \cdot \mathbf{H}_1 + \mathbf{W}_2 \cdot \mathbf{H}_2 \tag{2}$$

Where now, $\mathbf{W}_1$ and $\mathbf{W}_2$ are fixed and known from a preceding training stage and $\mathbf{H}_1$ and $\mathbf{H}_2$ can be estimated using the same process as above (note that this equation is the same model as in equation (1), but this time one of the factors – the **W**'s – is known). Upon estimation of the two activation matrices **H**, we can approximate the observed mixture as a sum of $\mathbf{W}_1 \cdot \mathbf{H}_1$ and $\mathbf{W}_2 \cdot \mathbf{H}_2$, each of which will describe the time/frequency energy of each of the two sources as defined by their dictionaries. These two representations can be inverted back to a time-domain signal as described in (Virtanen et al.), and will result in separating the two sources in the original mixture. A necessary condition in order to have a successful separation is that the two dictionary subspaces do not overlap significantly. In practice, even when separating speech-only mixtures (with both dictionaries representing speech, but trained on different speakers), we see a good enough performance so this isn't a significant limitation. In real-life, we might not know all the sources that make up a mixture, in which case we can use a semi-supervised formulation (Smaragdis, Raj, and Shashanka 2007) that permits us to have only one dictionary (which can describe either the target source, or the interferences). An example of this process is shown in Figure 1, wherein a hydrophone recording of a whale is extracted from a noisy ocean recording.

This model also allows us to perform parameter estimation on a source that is part of a mixture. For example, the sum of the activation values in $\mathbf{H}_i$ corresponding to the i-th source can be used to infer that source's amplitude in a mixture. Additionally, if we have training data that is annotated, we can associate each learned dictionary element (each column of $\mathbf{W}$) with a set of parameters, e.g. pitch, phoneme, class type, etc. Using the principles described in (Smaragdis 2011), one can estimate these parameters for a source that is observed inside a mixture, something that has been put to use for a wide range of problems, such as pitch estimation on multiple instruments in music, speech recognition of simultaneous speakers and sound recognition from dense mixtures (Nam, Mysore, and Smaragdis 2012, Smaragdis 2011, Smaragdis and Raj 2012). What is very important in this case, is that one does not need to perform source separation and then run pitch tracking on its output. This is known to be a suboptimal approach in which source separation artifacts can interfere with any subsequent estimates. By using the aforementioned estimation methods, one can truly perform analysis on mixture signals directly in a single step, something that is highly desirable in many problems that involve acoustic analysis.

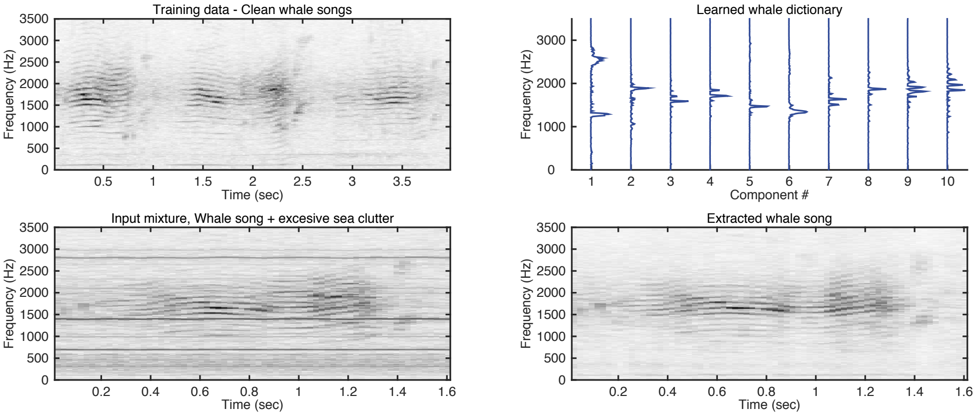


Figure 1. An example of source separation in an underwater sensing context. Recordings of whale songs (top left) are used to learn a dictionary of whale sounds (top right). Given an unseen mixture containing a new whale song contaminated with excessive sea clutter interference (bottom left) we can isolate the whale components (bottom right) by fitting the learned whale components and simultaneously estimating the interference source (see (Smaragdis et al. 2014) for details).

This approach of operating on mixtures using non-negative modeling was originally introduced by the PI and has since been put to considerable use in the audio processing community. There are by now thousands of papers on this representation, and it is being used commercially with critical acclaim[†]. Further developments in this area have resulted in more elaborate models that can better model more signal attributes. These include the probabilistic and Bayesian formulations (Hoffman 2012, Shashanka, Raj, and Smaragdis 2008), convolutive and shift-invariant models (Smaragdis 2007), Hidden Markov Model formulations (Mysore, Smaragdis, and Raj 2010), dynamical models (Mohammadiha, Taghia, and Leijon 2012), and tensor formulations (Cichocki, Zdunek, and Amari 2007), amongst many others.

This family of acoustical signal processing methods is also closely related to many similar models from other fields. Most notably, the factorization in equation (1) can be related to dimensionality reduction methods, sparse learning and compressive sensing. Models like Principal Component Analysis (PCA) and Independent Component Analysis (ICA) (Oja and Hyvarinen 2000) are very closely related, but have

[†] E.g. the Audionamix tools, used in the production of Hollywood movies, and Adobe Audition, which is used by million of audio engineers worldwide.

eventually been deemed inappropriate for mixture analysis since they make use of cross-cancellation to reconstruct the inputs. Additionally, their results are not as easily decipherable as with the non-negative models. Along the same lines is research on sparse coding and compressive sensing (Baraniuk 2007, Donoho and Tanner 2005), whose models are also using a form similar to equation (1). Just as with PCA and ICA, these models excel at approximating the input, but result in non-intuitive cross-cancelling decompositions that have not been found to be as useful for acoustical processing applications (made worse by the lack of early research on dictionary learning, and the inappropriateness of using the Euclidean distance cost function). Along the same lines, random projection methods have not found much success in acoustic signal processing. More similar is work has taken place with tensor models (Sidiropoulos and Kyrillidis 2012, Cichocki, Zdunek, and Amari 2007), which in many cases embraces non-negativity and is also flexible enough to accommodate the divergence functions that seem to work better in such problems. Finally, there is also a strong connection to the topic modeling literature (Hoffman 2012, Blei, Ng, and Jordan 2003). Especially though the probabilistic formulations of non-negative models, e.g. (Shashanka, Raj, and Smaragdis 2008, Hoffman 2012), one can see a very similar structure in which the dictionary and the activations are interpreted as probabilities. In some cases, models from these two areas can be theoretically identical, although there is significantly more work in the acoustical processing domain to address signal-specific attributes (e.g. Markov properties, smoothness constraints, etc.). Finally, non-negative decompositions have also been examined in the algorithms literature (Arora et al. 2012), where many interesting (and very efficient) strategies have been developed. Unfortunately, these approaches use a Euclidean distance cost function that is known to be suboptimal for acoustics applications and does not help in resolving many of the issues presented herein.

The problem

Despite the efficacy and success of non-negative representations, these methods do not scale well computationally. They are still constrained to workstation-level computing and are applied on relatively small data sizes. This not only precludes their use on low-power devices, but also hinders our exploration of their behavior when trained and tested on large data collections. As part of this proposal we will develop the necessary computational framework to allow the application of such methods on large data sets, and enable the exploration and better understanding of the effects of such large-scale analyses.

To appreciate the involved complexities, consider the computational requirements of these models. For the factorization model in equation (1) and using the modified KL divergence to measure fit, we need to employ the following iterative computations to progressively refine our current estimates of **W** and **H**:

$$\begin{aligned}\mathbf{V} &= \mathbf{F} \oslash \mathbf{W} \cdot \mathbf{H} \\ \mathbf{W}^{new} &= \mathbf{W} \odot \left(\mathbf{V} \cdot \mathbf{H}^{\top} \right) \\ \mathbf{H}^{new} &= \mathbf{H} \odot \left(\mathbf{W}^{\top} \cdot \mathbf{V} \right)\end{aligned} \tag{3}$$

Here, the symbols $\oslash$ and $\odot$ denote element-wise division and multiplication respectively. In practice one needs to repeat the above computations a few hundred times to obtain stable dictionary and activation estimates. For a large input matrix **F** this implies a considerable number of floating point operations. For example, a 1023 by 10,000,000 **F** matrix corresponds to about 16 hours of high-resolution speech, which means that for an order $K = 100$ model we would be employing more than six TFLOPs for each iteration. If one were to consider large corpora that contain content in the order of days or months (e.g. telephone conversations, oceanographic recordings, music archives, battlefield communications), clearly these methods become impractical or even intractable. Alternative approaches include other optimization schemes (Kim, He, and Park 2014), but they still exhibit prohibitive levels of computational complexity.

One possible approach to address large inputs is based on online methods (Lefevre, Bach, and Févotte 2011, Duan, Mysore, and Smaragdis 2012). These approaches operate on short sliding windows over the input matrix, and accumulate results in order to approximate a factorization on the entire matrix. This makes online processing feasible, and removes some of the bandwidth constraints when dealing with large inputs. However, the computational complexity of these methods is on par with batch approaches, while exhibiting a noticeable loss in estimation accuracy.

Keeping the above problems in mind, we will focus on a set of novel algorithmic developments that will allow us to address the efficiency issue. The goal is not only to gain efficiency for processing of large data, but also to make such methods available on embedded and low-power devices.

Research plan

Our research plan is comprised out of three main components. One of these addresses the issue of ***rapid dictionary learning when confronted with large data sets***, the other addresses the issue of ***efficiently deploying a system that uses a large dictionary***, and the third one considers the ***application of these findings on the much wider array of related algorithms*** that have been developed in this field. In the process of investigating some important theoretical underpinnings of these models ***we will perform the first investigations of the effect of large-scale training on the performance of mixture analysis models***.

Component I: Enabling Rapid Dictionary Learning

Problem: A considerable pain point with non-negative models relates to the rank of the decomposition. Practical models could use $K = 50$ to 500 dictionary elements that, due to the nature of all current algorithms, have to be estimated simultaneously. This creates a problem, since in many situations the optimal value of K is not known advance and one needs to perform multiple factorizations for different values of K until a satisfactory model is found.

Proposal: For this part of our work we will address the issue of rapidly building non-negative libraries. In doing so, we will advance the state of the art in efficient dictionary building, and we will examine the benefits of learning dictionaries from large acoustic data sets – something that has not been feasible so far.

Approach: A novel approach to simplifying dictionary learning is using a deflationary approach, in which only one component is extracted at a time and one can iterate until a desired level of performance is reached. Such models have been used in other forms of linear decompositions, e.g. with the PCA/SVD and ICA (Brand 2002, Oja and Hyvarinen 2000) and also considered for non-negative factorizations in (Biggs, Ghodsi, and Vavasis 2008, Gillis 2011). Unfortunately, the non-negative model approaches do not necessarily offer a concrete computational advantage, and are developed using a Euclidean distance cost function that is known to be suboptimal for acoustical signal applications.

Our preliminary investigations have produced an efficient deflation method that is efficient and appropriately tailored for acoustic processing by optimizing a cost function using the β-divergence (a parameterized divergence form (Févotte and Idier 2011) that can take the form of both the Kullback-Leibler and Itakura-Saito divergences according to the value of β). This model is based the following optimization:

$$\begin{aligned}&\underset{w^{(k)},h^{(k)},R^{(k)}}{\arg\min}\,\mathcal{D}\left(\mathbf{F}^{(k)}\middle|\mathbf{w}^{(k)}\cdot\mathbf{h}^{(k)}+\mathbf{R}^{(k)}\right)+\lambda\left|\mathbf{R}^{(k)}\right|_F\\&\text{s.t.}\quad \mathbf{w}^{(k)}\geq 0,\ \ \mathbf{h}^{(k)}\geq 0,\ \ \mathbf{R}^{(k)}\geq 0\end{aligned}\tag{4}$$

Here the (k) superscript denotes the dictionary element index and $\mathcal{D}$ is the cost function we choose to employ (in this case any form of a β-divergence). For $k = 1 \ldots K$ we solve the above optimization problem and in each step we obtain a new dictionary element $\mathbf{w}^{(k)}$ and its corresponding activations $\mathbf{h}^{(k)}$. These will correspond to columns of **W** and rows of **H** respectively. There is also a non-negative residual component $\mathbf{R}^{(k)}$ which is the result of removing the one-rank approximation $\mathbf{w}^{(k)}\,\mathbf{h}^{(k)}$ from the input $\mathbf{F}^{(k)}$. For each successive iteration we use $\mathbf{F}^{(k)} = \mathbf{R}^{(k\text{-}1)}$, except for the first iteration in which $\mathbf{F}^{(1)}$ will be the original time/frequency input to analyze. The regularization term on $\mathbf{R}^{(k)}$ is there to discourage the trivial solution where $\mathbf{R}^{(k)} = \mathbf{F}^{(k)}$. This approach effectively approximates the input as best as possible with a rank one decomposition while leaving a non-negative residual, then obtains a new rank-one decomposition on that residual and repeats until the desired number of dictionary elements has been extracted. The numerical solution to the above problem has a multiplicative update that is not significantly complex (shown here for $\beta = 1$):

$$\begin{aligned}
\mathbf{w}^{(k)} &\leftarrow \mathbf{w}^{(k)} \odot \left[\left(\mathbf{F}^{(k)} \oslash \left(\mathbf{w}^{(k)}\mathbf{h}^{(k)} + \mathbf{R}^{(k)}\right)\right)\cdot \mathbf{h}^{(k)^\top}\right] \oslash \left(\mathbf{1}\cdot \mathbf{h}^{(k)^\top}\right)\\
\mathbf{h}^{(k)} &\leftarrow \mathbf{h}^{(k)} \odot \left[\mathbf{w}^{(k)^\top}\cdot\left(\mathbf{F}^{(k)} \oslash \left(\mathbf{w}^{(k)}\mathbf{h}^{(k)} + \mathbf{R}^{(k)}\right)\right)\right] \oslash \left(\mathbf{w}^{(k)^\top}\cdot \mathbf{1}\right)\\
\mathbf{R}^{(k)} &\leftarrow \mathbf{R}^{(k)} \odot \left[\mathbf{F}^{(k)} \oslash \left(\mathbf{w}^{(k)}\mathbf{h}^{(k)} + \mathbf{R}^{(k)}\right)\right] \oslash \left(1 + \lambda \mathbf{R}^{(k)}\right)
\end{aligned} \tag{5}$$

In terms of FLOPs these updates are more efficient than the regular factorization updates (Lee and Seung 2001), since they do not include any matrix/matrix products. This results in a dramatic speedup since we only need to compute a rank-1 factorization at a time, and as we show below, we also obtain faster convergence behavior this way.

In Figure 2 we show the convergence behavior of this approach as compared to a regular factorization. In that figure's simulation we seek to find an adequately good low-rank approximation to an acoustic recording. Using a regular full-rank approach we need to successively compute a rank-1, rank-2 and rank-3 decomposition until we see obtain a reconstruction error below 10% of the input's energy. Using the deflation method we see that we obtain faster convergence and a lower error, while additionally requiring roughly 2/3rds of the iterations to reach comparable performance. In addition to that, the update equations for the deflation method are numerically significantly faster to compute. Since in real-life problems we routinely compute dictionaries with hundreds of elements, the savings of this approach can be significant. Although the results obtained through deflation are not quantitatively the same as with full-rank methods‡, they result in qualitatively the same performance for dictionary-based tasks such as speech denoising.

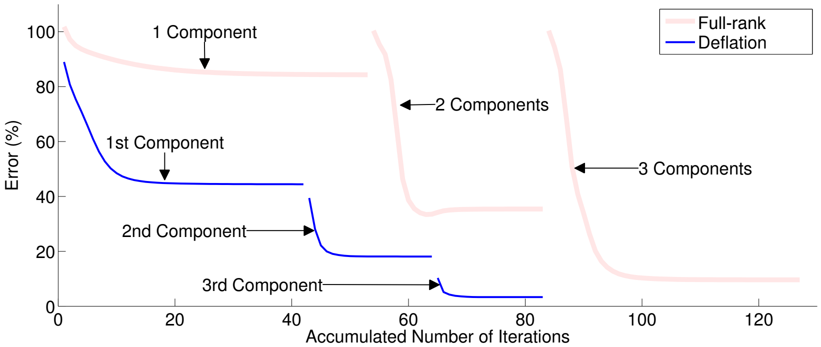


Figure 2. Comparing the deflation approach to regular factorizations. The pink lines show the reconstruction error as we perform successive factorizations with an increasing rank in order to discover an optimal *K*. The blue lines show the reconstruction error using the deflation approach, which converges faster and uses much faster computations.

‡ It should be noted here that for most practical non-negative models in acoustical signal processing, there is no uniqueness. Most methods converge to local optima that in practice result in uniform performance.

Validation: As part of our plan we will investigate this approach further to better understand how it behaves and how additional computational savings can be gained. We plan to launch a formal investigation in the computational complexity and convergence behavior of this process, and to run thorough simulations that examine how these estimates relate to full-rank models. We will focus on problems of dictionary learning, which is the process that stands to gain the most. In order to validate our findings in practice, we will apply this method on a natural sound data set that consists of multiple days of content. Our goal is to learn a dictionary for this data set and to then use it to perform source separation and sound classification on mixtures of sounds created from that database. Since existing methods cannot operate on large scales of data we will do a direct performance comparison using a smaller subset. We will measure differences between respective runtimes and also of literature-standard metrics for separation and classification problems (Févotte, Gribonval, and Vincent 2005, Giannoulis et al. 2013). Finally, we will perform the first ever analysis of a corpus consisting of days of content, and will note how a dictionary trained on a substantially larger set of data performs with respect to the current state of the art.

Deliverables: By the end of this investigation we will have a lightweight dictionary learning algorithm, new theoretical results on the design of deflation-based non-negative models, and a better understanding of the quality of dictionaries trained on large data sets.

Component II: Accelerating Large Dictionary Deployment

Problem: In many cases we are confronted with having to simultaneously use multiple dictionaries, each with a few hundred or thousands of elements. One such situation is when performing source classification from mixtures. In this case, each source class will come with its own dictionary, and all of these dictionaries will be used at the same time. Another case is that of the Universal Speech Model (Sun and Mysore 2013), which is a speech model trained on a lot of different speakers and concatenates their dictionaries to form a universal speech model. These cases are equivalent to a factorization that is done with a very wide matrix $\mathbf{W}$, which can result in a significant computational cost and to date has limited the deployment of such systems with large dictionaries.

Proposal: Our recent work (Kim and Smaragdis 2013), has shown an efficient way to maintain the benefits of a large dictionary while shedding the extra computational burden. We will extend this idea and develop a framework that will allow us to compress large dictionaries to significantly smaller sizes, while still maintaining their high performance characteristics. This will allow us to deploy systems that are using significantly more complex models without requiring the necessary computational load. This idea will also let us perform simulations that have been impossible in the past, and will provide us with a better understanding of how very large dictionaries can improve acoustic mixture analysis.

Approach: The key observation here is that dictionary elements of acoustical sources tend to lie on highly structured manifolds. With no loss of generality, we will consider the model where all the columns of $\mathbf{F}$, $\mathbf{W}$ and $\mathbf{H}$ are constrained to sum to 1. This is a common convention in the literature since we are usually invariant to absolute scale in the acoustic domain. This is often achieved by normalizing the columns of the input matrix, performing an analysis, and subsequently applying the same column gains to modulate the output back to the original input scale. This process also transforms the factorization model to the Probabilistic Latent Component Analysis (PLCA) model, which is a probabilistic version of a non-negative factorization (Shashanka, Raj, and Smaragdis 2008).

An illustrative case of how such models work is shown in Figure 3.a. Because of the normalizations, all dictionary elements (as well as data) will lie inside a simplex. Shown in the left plot is an illustrative low-dimensional dictionary that lies on a manifold. Each point corresponds to a dictionary element, which in this case would be a 3-dimensional spectrum. Any data points in that source's space will be explained by a convex combination of the dictionary elements. There are two problems in this setting: a) for a large dictionary we would need considerable memory and processing resources to apply it on a mixture, and b) this dictionary can reconstruct parts of the space that are outside of the source's manifold (e.g. any part between two dictionary elements). The latter point can be an issue since one of our primary concerns with such models is to have minimal overlap between multiple dictionaries. If a dictionary can reconstruct points that are outside of a source's manifold that maximizes the probability that it can overlap with another source's dictionary. A solution to this problem would be to use sparse activations on the dictionary elements, thereby approximating areas close to each dictionary element and ensuring that all potential reconstructions are lying on the dictionary manifold (Smaragdis, Shashanka, and Raj 2009). This approach results in highly improved results for speech separation tasks, but unfortunately it introduces additional computations that prohibit large-data processing.

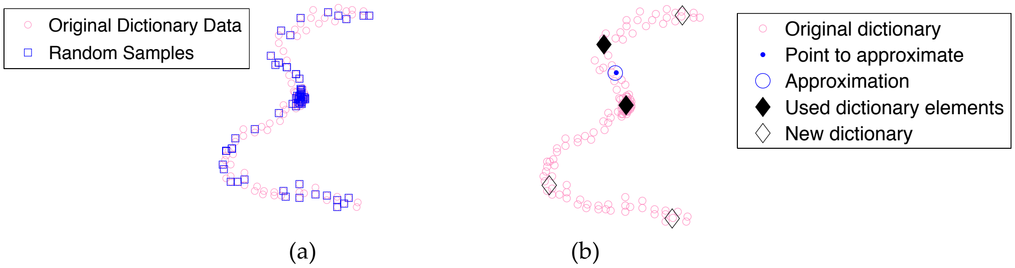


Figure 3. Experiments in subsampling large dictionaries. In plot (a) the light pink points are the original dictionary elements, and the blue points are the subsampled points. We see that there is a fair amount of redundancy due to a high concentration of original dictionary elements near the center of the manifold. When we use the proposed method, plot (b), we use a very small amount of dictionary elements whose locally linear combinations can trace the original manifold structure. When approximating a data point as shown, we only use a local neighborhood of dictionary elements and guarantee manifold preservation.

Our more recent investigations (Kim and Smaragdis 2013) focus on resolving such problems while maintaining these desirable performance characteristics. One obvious approach is to employ random sampling on the source dictionary. With a substantially large dictionary, one can sub-sample it and use the resulting samples as a new smaller dictionary. In the Figure 3.a, we show what this looks like. It can result in a dictionary with fewer elements that is amenable to faster processing. However, there is no guarantee that we will sample points from the crucial parts of the large dictionary and run the risk of suboptimal performance. To address that, we propose a different method that maintains performance while drastically reducing the necessary dictionary size. This involves the following decomposition:

$$\begin{aligned} &\mathbf{W} \approx \mathbf{W} \cdot \mathbf{S} \cdot \mathbf{H} \\ &\mathbf{W} \in \mathbb{R}_+^{M \times K},\ \mathbf{S} \in \mathbb{R}_+^{K \times L},\ \mathbf{H} \in \mathbb{R}_+^{L \times N} \end{aligned} \tag{6}$$

Where $L << K$ and the matrices **S** and **H** are sparse along their columns. This decomposition effectively approximates **W** using a much smaller dictionary $\mathbf{W}_2 = \mathbf{W} \cdot \mathbf{S}$ of size $M \cdot L$, as opposed to **W**'s $M \cdot K$. Each dictionary element of $\mathbf{W}_2$ will be a linear combination of only a few elements of **W**, and will ideally capture the most salient points of the original large dictionary. The estimation equations of this process are very similar to the basic factorization model and of similar complexity. Of course this isn't a complete solution, we still have two major problems to address.

First, we are still faced with a large number of computations for completing this decomposition. This can be overcome using a prioritization step based on hashing. We note that by design **S** and **H** have sparse columns, meaning that the elements in $\mathbf{W} \cdot \mathbf{S}$ will each include only a small number of the columns of **W**.

As a result, most of the dot products that are necessary to estimate **S** and **H** will be performed between vectors that are mostly uncorrelated. Therefore, we can prioritize computations so that we only perform useful dot products. Therefore, in every iteration of the estimation process we use a hashing-based search to find which parts of **W**, **S** and **H** will produce approximately non-zero dot products and only compute these. This allows us to significantly reduce the size of necessary computations and still obtain a good estimate of **S** and **H**. Using preliminary experiments, we found that using this approach we can reduce necessary computations up to two orders of magnitude without an appreciable loss of performance.

The other problem is that we need to guarantee that the new dictionary $\mathbf{W}_2$ is expressive enough to properly represent the space that **W** occupies. We observe that the operation that we perform so far produces new dictionary elements that are comprised out of linear combinations of the original dictionary **W**. An example of that is shown in Figure 3.b. Note that the new dictionary elements tend to occupy key positions from the original dictionary, not being spread in areas that can be easily explained by their convex combinations. To properly approximate the manifold structure of the original dictionary, we need to impose one more constraint – that the elements of the new dictionary are joint-sparsely activated. This sparsity constraint is defined so that only neighboring elements can be active at the same time. This means that each potential approximation will be defined by a small number of neighboring dictionary elements, thereby maintaining the manifold structure of the original dictionary **W**. An example of that is shown in Figure 3.b.

Using this approach, we performed experiments to measure the performance drop with respect to the size reduction of the original dictionary. In Figure 4 we see the comparison between an original full dictionary, a random sub-sampling of it, and the manifold preserving method with hashing and not. We see that as we reduce the number of sampled dictionary elements the approximation error increases. However our approach results in a much slower error increase that allows us to use significantly smaller fraction of the dictionary size while still maintaining a relatively good performance. Additionally, using the hashing-based computation prioritization we observe negligible effects in performance, while offering significant computational savings.

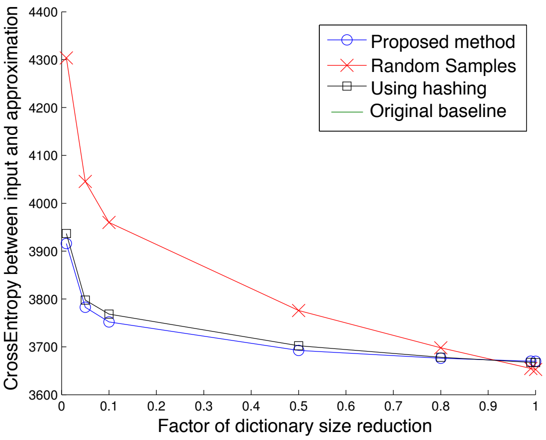


Figure 4. A comparison of the performance between randomly sampled dictionaries and manifold preserving sampling. The vertical axis measures the approximation error between the original input and the reconstruction based on these methods. Lower values mean that we are obtaining a better fit. The horizontal axis represents the number of dictionary elements as compared to the full dictionary size. One can see that although random sampling works, it isn't as effective as the proposed approach, which quickly converges close to the performance of a full dictionary using only a fraction of its size.

Even more strikingly, we used this method for extracting speech from mixtures (and we discovered that using only 5% of the size of the original dictionary we can perform just as well as using the entire dictionary, whereas random sampling started producing rapidly degraded results once we used less than 50% of the original dictionary size.

For the purposes of this proposal, we will perform a deeper theoretical analysis of the implications of manifold-preserving quantizations and we will devise optimal algorithms for this approach. So far there has been very little theoretical analysis of the manifold properties of dictionaries of acoustical signals, and how they can overlap with each other when used in factorization models. We will be the first to examine this space and its practical implications. Unlike traditional linear models (e.g. the ones used in compressive sensing and sparse coding), we cannot use existing coherence measures since they are designed to operate on a Euclidean space and they do not apply on the types of dictionaries we find in acoustical mixture processing. As we better specify what it means to have a manifold-preserving method we will be answering questions that are central to these models and have not been considered in the past.

Validation: We will validate our approach by using it for speech denoising problems trained on large corpuses. As shown in (Smaragdis, Shashanka, and Raj 2009) and (Sun and Mysore 2013) using a large dictionary trained in multiple speakers can result in a significant increase of denoising performance. However, this comes at a cost of additional computations that impose undesirable constraints. Current models are limited to being trained on a relatively small number of speakers (not more than 100) using only a few seconds of speech from each speaker. Using the aforementioned approaches we will be able to train on a much larger training set (in the order of hundreds of speakers and hours-long segments) and obtain vastly larger dictionaries. This effort will not only allow us to use such large dictionaries on weaker devices (e.g. embedded and mobile systems), but it will provide us with a glimpse of how large training data can help us improve source separation systems. We will use community-accepted measures (Févotte, Gribonval, and Vincent 2005, Taal et al. 2010) to quantify the performance of these large-dictionary systems compared to existing approaches. It is now known that a large dictionary can have a dramatically positive effect on performance; being able to scale such an approach to a much larger training set can potentially bring us to a new regime of results for mixed acoustic signal processing.

Deliverables: At the end of this investigation we anticipate to have a better way of compressing large dictionaries with no appreciable effects on performance, to have obtained a deeper understanding of the manifold structure of sound in these spaces, and to measure impact of training on large acoustic data sets for the purposes of mixture analysis.

Component III: Application of Components I/II to the Wider Family of Models

The aforementioned directions will be primarily tested on the basic factorization model in equation (1), however given the large number of extensions of this model we anticipate that these methods will have impact on a wider set of tools. As we have shown in (Smaragdis et al. 2014), there are elegant ways to generalize the basic model and describe a wide family of static and dynamic models that are based on the same principle. This includes models that are designed to decompose different types of time/frequency transforms (e.g. power spectra, magnitude spectra, learned transforms like Karhunen-Loëve models, etc.), but more importantly models that incorporate the temporal dimension. This includes Kalman filters and Hidden Markov Models, whose states are characterized by non-negative factorizations. Many of the theoretical investigations that will conduct will be directly applicable to these models as well.

More specifically, we are interested in using the aforementioned developments to speedup the parameter estimation and deployment of the N-HMM (Mysore 2010) and dynamic PLCA (Smaragdis et al. 2014) models. These are models that have been shown to be more effective than static models when it comes to analyzing temporally coherent signals (which most of acoustic data is). However, the existing algorithms are heavily slowed down by the use of large dictionaries and the difficulty of learning them rapidly. We anticipate that our analysis on fast learning and deployment of non-negative dictionaries will let us apply new ideas on these models and make their application on larger data sets feasible. For the N-HMM model, we plan to test its performance on larger dictionary models than the ones used before (Mysore 2010), and to see how its performance scales when trained on larger data sets and user with more states that contain larger dictionaries. Likewise, for the dynamic PLCA model we will consider a larger dictionary size and measure its effect on performance for tasks like mixture sound recognition and denoising as reported in (Nam, Mysore, and Smaragdis 2012) and (Mohammadiha, Taghia, and Leijon 2012).

Educational activities

We find that due to the interdisciplinary nature of this project there is potential to design educational activities that present new ways to think about audio processing (and signal processing in general).

Curriculum Development

The project PI has been active in curriculum design, in both the CS and ECE depts. and has been charted to develop a set of new classes that fill the skills gap between them, exposing CS students to signals theory and ECE students to learning and AI. In his first year he designed a new graduate class on *Machine Learning for Signal Processing*, and a new undergraduate sophomore class on *Designing Intelligent Systems*. Both classes were hands on and cross-listed between the CS and ECE depts., but due to their breadth they were also attended by students from mechanical engineering, mathematics, music, finance, aero-astro and bioengineering. The PI received top student reviews from both of these courses during all of the semesters they were offered. We anticipate this curriculum development activity to continue throughout this project, and to use our research findings to help design more classes. More specifically, we have the following courses in mind for development:

Undergraduate CS/ECE class: *Intelligent Signal Processing*. Two admittedly weak points of CS students are their ability to manipulate signals, and their fluency with probabilistic reasoning. Both of these skills are introduced late in standard curricula (if at all), and by that point they constitute material that's difficult to grasp after the traditional exposure to deterministic and symbolic processes. Likewise, ECE students are often exposed to AI and machine learning during their upper level classes and miss the opportunity to properly internalize this way of thinking early on. Given the unprecedented prevalence of signals and learning (signals being central to rich-media manipulation and learning being one of the most dominant areas of research and commercial activity today), we believe that students should be exposed to this material early on in a practical motivating context that helps them intuitively understand the class content. This proposed class on *intelligent signal processing* will be focusing on practical applications such as image and speech recognition, biological signal processing, music information retrieval, etc., as a way to motivate students and to impart practical skills. We are taking special care not to burden such a class with traditional signals and learning theory that is often cumbersome, demotivational, and a hindrance to STEM student retention, but instead we are interested in "planting the seed" early on to help in developing an intuition which will be invaluable once more traditional signals and learning course electives are attended later on. Given the prevalence of mixed-signals and large signal collections in the real world, we anticipate that a significant part of the homework and projects will be using components that we will develop in the aforementioned research plans.

Graduate engineering class: *Big Data & Mixed-Signals Processing.* Given the wide breadth of applications that relate to the subject of this proposal we anticipate to offer a new class tailored to graduate students from all of the engineering departments. Our observation from the aforementioned *Machine Learning for*

Signal Processing class is that there is considerable interest in large-scale mixed-signals processing throughout our campus. The PI has advised (and collaborates with) students from a diverse set of departments, such as civil, biomedical, mechanical and neuro-engineering, as well as in music, physics and mathematics, on many practical and theoretical problems. We anticipate to package all the lessons learned throughout the research development of this project and to offer this graduate course in order to seek out new applications of our methods, and to address an area that we know has strong student demand. Just as before, we see a strong possibility to incorporate real-life problems with big-data and mixed-signals into the syllabus and the homework assignments. Due to the breadth of the students that can take this course we also anticipate to come across novel applications of this technology in areas outside of the PI's research domain (e.g. bioacoustics, geophysical signals, etc.).

Massive Open Online Class: *Audio Machine Learning and Signal Processing*. With the recent trend of Massive Open Online Classes (MOOCs), we have seen an unprecedented transfer of knowledge from universities to the masses. The University of Illinois has partnered with Coursera to develop online courses. The PI of this project has committed to develop a course on modern techniques for *Audio Machine Learning and Signal Processing*. This is an important and timely subject that is not covered thus far by any online offering known to us. With the increasing prevalence of audio in our lives, this is a subject that's not only interesting from an academic standpoint, but also one that is central to students seeking employment in the telecommunications, entertainment and online media industries. This course will aim to bring together common audio operations in an integrated manner, and will also incorporate findings of this project since they address one of the most central problems in audio today. Given the proliferation of large audio data sets (speech, music, environmental sounds, etc.), we find that the research problems that we will attack above will make for very motivational projects and hands-on instructional material.

Professional tutorials

Outside of the university setting we also anticipate to disseminate our findings through conference tutorials and visiting lectures. The project PI has already delivered conference tutorials on the subject of source separation and denoising (INTERSPEECH 2006, ICASSP 2011, INTERSPEECH 2012) and is a frequent guest lecturer. We anticipate continuing this dissemination effort. More specifically, we plan to deliver the following tutorials in upcoming conferences: a tutorial on *compositional methods for signal processing*, a tutorial on *mixed-signal processing for large data*, and a tutorial on *mixed-signal analysis for speech and music processing*. These tutorials will directly relate to the material we develop during the course of this project and will serve to popularize these ideas. For these tutorials we will target high-visibility venues, such as the aforementioned conferences.

Student mentoring

During the last year the PI has been mentoring a team of five graduate and three undergraduate students. He also runs an active visiting international student program that has so far hosted three more graduate students in the past few years. In addition to the above, the PI has been involved in research projects with multiple corporations (Adobe, Mitsubishi, Analog Devices, Sony), and in that capacity has mentored graduate students that perform their internships there. Due to the strong corporate ties of the PI, all of his mentored students are being placed in basic-research summer internships, where they are expected to broaden their perspective and come back with fresh ideas and new experiences.

We anticipate this specific research project to involve students from various levels. With the advent of the new aforementioned undergraduate courses, we expect to interact with greater numbers of lower-year undergraduates and to be able to offer them research opportunities since their coursework relates to the objectives of this project. We also anticipate to open new positions for graduate students and visiting students, but also to collaborate closer with existing students on campus who work on related problems. Adobe has already pledged to host summer interns to assist with their media intelligence research effort, which aligns with the goals of this proposal.

Interdisciplinary activities

Another important aspect of student mentoring will come via the incorporation of a new interdisciplinary center for audio sciences that the PI has founded with the support of the College of Engineering at the University of Illinois. This center has affiliated faculty from a wide range of departments (CS, ECE, Biomedical, Speech and Hearing, Library Sciences and Music), and is expected to act as a nexus for audio and sound research in the campus. Due to the broad interdisciplinary footprint of this center we expect to attract undergraduate and graduate students with a variety of backgrounds and research interests that will be involved in our research activities. From an educational standpoint we expect this center to broaden students' understanding of audio and to host an interdisciplinary series of seminars and invited talks that will make a well-rounded audio education more easily accessible in our campus.

In tandem with this center, we hope to take advantage of the CS + X program, a dual majors program between CS and any liberal arts program that can benefit from computational techniques. We have already developed with the School of Music a CS + Music curriculum to address the needs of undergraduate music students that have technological inclinations, and vice-versa. The strong music component of this proposal is expected to be used as a recruiting tool and motivation for students in this program, but also to demonstrate a more approachable side of otherwise stern STEM material.

Outreach activities

There is no denying that audio demos can be very motivating and inspiring. Designing computers and robots that can understand sound, speech and music is an effective way to demonstrate mathematics at work in a way that everyone can appreciate. In order to help outreach, we plan to use real-world audio demos that result from our work to inspire high school and underrepresented group students and interest them in a STEM career. We plan to do so through our annual Engineering Open House, but also by employing audio-related installations in the CS building, which is frequented by students from the University High School that is across the street from our laboratory. More specifically we expect to have an autonomous demonstration audio scene analysis system which students can interact with in real-time. Due to our proximity to local high schools (with percentages of low socioeconomic students in excess of 60%) we also plan to participate in regular university visits which are designed to inspire students and attract them to the engineering and science fields.

Dissemination activities

Tool building

Since the main goal of this project is to examine scalable methods for audio mixture processing, we want to make these ideas more accessible by sharing our tools with the research community. As part of this project we expect to release MATLAB and C++ code, as well as data to replicate our experiments and facilitate further research in this area. More specifically we will provide optimized code for all the research activities described above, as well as data sets to demonstrate how it works.

Research community activities

Recently we have seen the development of a series of popular audio challenges (Christensen et al. 2010, Cooke, Hershey, and Rennie 2010), that were designed to motivate source separation researchers and to provide a standardized platform for objective comparisons. This activity is also supported by the wider signal processing community through the IEEE technical steering committees for Audio and Acoustics and Machine Learning for Signal Processing. Since during the course of this project we will be developing benchmarks and challenges to further our algorithm development, we plan to make such toolsets freely available in the form of community challenges. By doing so we hope to stimulate research in this field and to offer a set of standards that can help anchor a community.

Deliverables summary and timeline

In summary, we propose to attack three problems that will provide context and their own unique challenges to our research goals. The work involved in the above projects will be overlapping by a fair amount and does not mean that we will have to initiate and complete three distinct projects, but rather one integrated effort.

The specific **research** elements that we seek to develop are the following:

1. *Efficient dictionary learning via deflation methods.* This will involve the investigation of basic algorithms that use incremental learning to efficiently construct dictionaries from large acoustical data sets. By being the first to employ such methods on a very large natural sound corpus we also aim to obtain more insights on the effects of training set sizes on the quality of learned dictionaries, and how that affects the performance of mixture-analysis algorithms. The outcomes will be quantified in terms of processing speed relative to existing models, and by using literature-standard performance metrics to evaluate their application in problems such as source-separation and recognition of acoustic sources in mixtures.
2. *Rapid fitting of large dictionaries on new data.* This part involves taking advantage of the manifold structure in overcomplete acoustical dictionaries in order to reduce the necessary computations and minimize memory requirements. Aside from developing new algorithms, we aspire to learn more about the manifold structure of sounds, and more importantly how their geometry changes when learned through small vs. large data sets. This technology will be tested on speech denoising applications that will be trained on a large number of speakers, thus resulting on a large overcomplete dictionary. We will quantify the success of this project in terms of separation quality and speed of operation using standard metrics.
3. *Application of extensions to temporal models.* Using the results from the other two thrusts we will reformulate the H-HMM and dynamic PLCA models in order to support a larger number of states and bigger dictionaries. By applying them on various mixture problems we will be able to gauge how such larger models (also trained on larger data) can result in improved results.

On the **educational** side we plan:

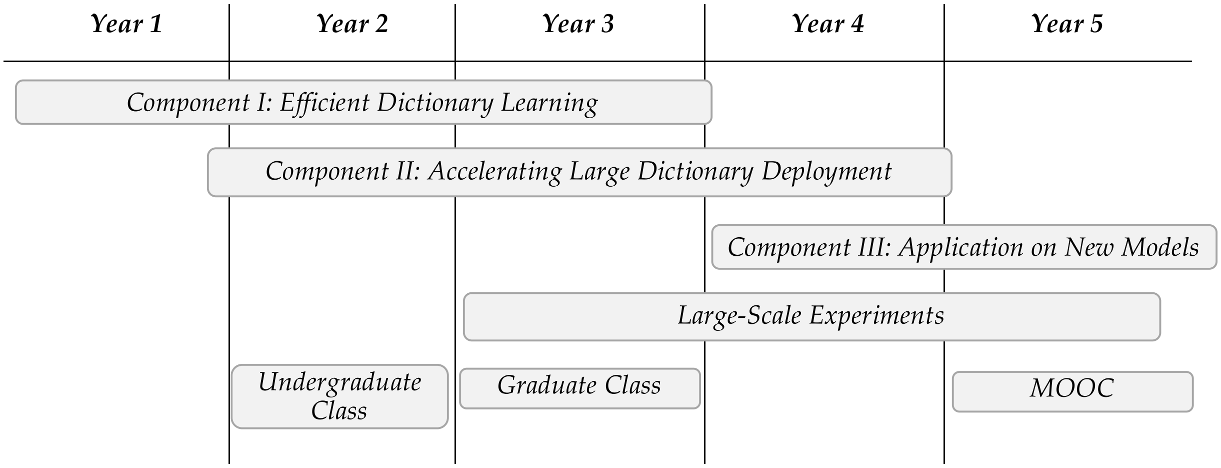
1. *A MOOC on Audio Machine Learning and Signal Processing*. An open online course served via Coursera that focuses on a unified view of common real-world audio analysis problems and their practical solutions, a big part of which will be on big-data collections and mixed-signals.
2. *A graduate course on Large-Scale Mixed-Signals Processing*. An interdisciplinary course that will involve the theory that we will develop in the context of mixed-signals processing in a variety of large-scale application domains.
3. *An undergraduate course on Intelligent Signal Processing*. An early-level hands-on course that exposes engineering students to signals and learning theory applications (the general encompassing area of this proposal).

In addition, we plan to aggressively recruit early-level undergraduates and high-school students to be involved in the construction of practical manifestations of the developed theory. Moreover, this work will take place in our newfound audio sciences center, an interdisciplinary entity, where applications of this work will be communicated to students from a wide range of backgrounds and departments.

Finally in terms of information **dissemination** we plan to:

1. *Organize conference tutorials* on the findings of this project
2. *Distribute data and code* to stimulate research on this topic.
3. *Setup a competition and benchmark processes* for mixed-signals processing.
4. *Make demos* to promote STEM involvement to high school audiences.

The anticipated timeline of this project is as follows:


Year 1
Year 2
Year 3
Year 4
Year 5
Component I: Efficient Dictionary Learning
Component II: Accelerating Large Dictionary Deployment
Component III: Application on New Models
Large-Scale Experiments
Undergraduate Class
Graduate Class
MOOC