SPARSE DECOMPOSITION OF MIXED AUDIO SIGNALS BY BASIS PURSUIT WITH AUTOREGRESSIVE MODELS

Youngmin Cho and Lawrence K. Saul

Department of Computer Science and Engineering
University of California, San Diego
{yoc002,saul}@cs.ucsd.edu

ABSTRACT

We develop a framework to detect when certain sounds are present in a mixed audio signal. We focus on the regime where out of a large number of possible sounds, a small but unknown number are combined and overlapped to yield the observed signal. To infer which sounds are present, we attempt to decompose the observed signal as a linear combination of a small number of sources. To encourage sparse solutions with this property, we balance the modeling errors from individual sources against an ℓ_1-norm penalty of the type used in basis pursuit and regularized linear regression with grouped variables. Our approach can be viewed as a novel generalization of basis pursuit in two ways: first, with a dictionary of fixed size, we attempt to model acoustic waveforms of potentially variable duration; second, for dictionary entries, we do not store basis vectors representing static templates, but the coefficients of autoregressive models that characterize the acoustic variability of individual sources. We derive the required optimizations in this framework and present experimental results on combinations of periodic and aperiodic sources.

Index Terms— Signal detection, Machine learning, Sparsity

1. INTRODUCTION

In this paper we consider the problem of detecting when certain sounds are present in a mixed audio signal. The problem arises in many settings, but here we focus on the particular regime that seems most relevant to the indexing and annotation of large digital audio libraries. For each recording in this library, we imagine that out of a large number of possible sounds, a small but unknown number are present and possibly overlapping in the observed signal. The goal in this application is to tag and index each recording by the sounds it contains. With this goal in mind, we develop a theoretical framework for the sparse decomposition of mixed audio signals and present preliminary but positive results demonstrating its feasibility.

The problem we consider in this paper is related to the problem of source separation, or the "cocktail party" problem, in which the goal is to recover the individual sources from a mixed audio signal. Models of source separation have been studied in many different communities. Researchers in *blind* source separation do not assume prior knowledge of individual sources, but merely exploit their statistical independence. However, popular methods such as independent components analysis typically assume the availability of multiple microphone recordings [1]. Researchers in machine learning attempt to estimate the statistics of individual sources from training data, then decompose mixed signals by performing probabilistic inference in a generative model. With sufficient training data, such methods have yielded successful results in source separation from only single microphone recordings [2]; however, the complexity of exact inference scales exponentially with the number of presumed sources [3]. Finally, researchers in computational auditory scene analysis (CASA) attempt to build models that mimic the workings of human audition, exploiting the same psychoacoustic cues (e.g., harmonicity, onsets/offsets, binaural differences) as human listeners [4]. Ultimately, these models could be expected to achieve human levels of performance, but they are currently limited by our incomplete understanding of human audition.

Compared to previous studies, our approach starts from different assumptions and works toward different goals. We assume prior knowledge of a large number K of possible sources and seek to identify which $k \ll K$ of these sources occur in single microphone recordings. Though the model we propose can be used for source separation, our main goal is not separation, but detection. For this problem, our main contribution is to propose a framework whose required optimizations scale efficiently with the number of possibly active sources. In particular, our approach avoids an exponential search through all possible $K!/(k!(K-k)!)$ combinations of sources.

Our approach can be viewed as an extension of previous work on basis pursuit [5]. The problem in this work is to reconstruct a vector of fixed length from a small number of basis vectors in an overcomplete dictionary. In basis pursuit, a sparse decomposition is achieved by balancing the reconstruction error against an ℓ_1-norm penalty on the linear reconstruction coefficients. The penalty serves as a regularizer, favoring sparse solutions. The required optimization is convex, with no local minima, and it scales efficiently with the dictionary size.

The problem we study differs from the problem of basis pursuit in two crucial respects. First, with a dictionary of fixed size, we are attempting to reconstruct acoustic waveforms of potentially variable duration. Second, we cannot model individual sources by simple static templates because they themselves exhibit many degrees of variability besides amplitude (e.g., duration, phase, timbre). To handle these differences, we have adapted the basic ideas behind basis pursuit to our setting, building also on recent work for mixtures of periodic sources [6]. Like basis pursuit, our approach computes a sparse decomposition of mixed signals in terms of their constituent sources, the possibilities of which are catalogued by a large dictionary. In our approach, however, the entries in this dictionary are not static templates, but autoregressive models whose parameters characterize the acoustic properties of different sources, as well as their degrees of variability.

This paper is organized as follows. In section 2, we begin by reviewing basis pursuit then show how its basic ideas can be extended to our setting. Our formulation of the problem balances the modeling errors from individual sources against an ℓ_1-norm penalty of the

type used in regularized linear regression with grouped variables [7]. We derive the required optimizations in our framework and describe their efficient solution by group Lasso and block coordinate descent algorithms [8]. In section 3, we present the results of experiments on mixtures of both periodic and aperiodic sources. Finally, in section 4, we conclude by listing several directions for future work.

2. MODEL

2.1. Review of basis pursuit

Basis pursuit (BP) is a popular method for decomposing a signal into an optimal superposition of dictionary elements [5]. In particular, denoting the signal by $\vec{x}$ and the dictionary elements by $\{\vec{s}_i\}_{i=1}^K$, BP attempts to compute a sparse set of (scalar) linear coefficients $\{\beta_i\}_{i=1}^K$ such that:

$$\vec{x} \;=\; \sum_{i=1}^{K} \beta_i \vec{s}_i. \tag{1}$$

In the usual regime of interest, the dictionary is overcomplete: that is, the number of dictionary elements K exceeds the dimensionality of the signal $\vec{x}$. Of the many decompositions satisfying eq. (1), BP favors the sparse decomposition that minimizes:

$$\min \sum_{i=1}^{K} |\beta_i| \;\text{ subject to }\; \vec{x} = \sum_{i=1}^{K} \beta_i \vec{s}_i. \tag{2}$$

The optimization in eq. (2) minimizes the ℓ_1-norm of the linear coefficients subject to the constraint in eq. (1). Because the optimization is convex, it has no local minima.

BP models the observed signal by an additive combination of dictionary elements with varying amplitudes. This model is very well suited to data compression using Fourier and/or wavelet dictionaries. However, it is not well suited to analyzing mixed audio signals in terms of sources from naturally occurring sounds. These sounds are likely to vary not only in terms of amplitude from one realization to the next, but also in terms of duration, phase, and timbre. To represent these variations explicitly by different dictionary elements would explode the dictionary size. We discuss a way around this problem in the next section.

2.2. Extension to autoregressive models

We extend basis pursuit by using autoregressive models to parameterize the variability of individual acoustic sources. In particular, we imagine that waveforms $\{s_{it}\}_{t=1}^T$ from the i^{th} source approximately satisfy the m^{th} order linear recursion relation:

$$s_{it} \;\approx\; \sum_{\tau=1}^{m} \alpha_{i\tau}\, s_{it-\tau}. \tag{3}$$

Strictly speaking, the above equation is only defined for $t > m$ since the waveform $\{s_{it}\}_{t=1}^T$ is only defined for $t > 0$. The particular realization of the i^{th} source's waveform is determined by the m *initial conditions* to the recursion relation which we denote by $\{u_{it}\}_{t=0}^{m-1}$. Eq. (3) can be extended to all times t by making the identification:

$$s_{it} \;=\; u_{i|t|} \quad \text{for } t \le 0. \tag{4}$$

Note that each dictionary entry models an m-dimensional family of signals in which the initial conditions can not only parameterize variations in amplitude (by scaling u_{it}), but also variations in phase (by shifting u_{it}) and timbre (by reweighting u_{it}). Finally, signals of variable duration T are accommodated simply by evolving the recursion relation in eq. (3) for different numbers of time steps.

In this paper, we assume that the m linear coefficients $\{\alpha_{i\tau}\}_{\tau=1}^m$ of each source's autoregressive model are known a priori and stored as one of the K entries in our dictionary. With this dictionary in hand, our goal is to compute sparse decompositions of mixed audio signals $\{x_t\}_{t=1}^T$ in terms of a few ($k \ll K$) active sources. For this problem, we propose the following optimization:

$$\begin{aligned} &\min_{s,u} \left\{ \frac{1}{2} \sum_{i=1}^{K} \sum_{t=1}^{T} \Big(s_{it} - \sum_{\tau=1}^{m} \alpha_{i\tau}\, s_{it-\tau} \Big)^2 \;+\; \gamma \sum_{i=1}^{K} \|u_i\|_2 \right\} \\ &\text{subject to } \; x_t = \sum_{i=1}^{K} s_{it} \quad \text{and} \quad s_{it} = u_{i|t|} \text{ for } t \le 0. \end{aligned} \tag{5}$$

The optimization is to be performed over all K source waveforms $\{s_{it}\}_{t=1}^T$ and initial conditions $\{u_{i\tau}\}_{\tau=0}^{m-1}$. Most of the terms in the optimization are already familiar. The first term measures the fidelity of each source to its autoregressive model. The constraints enforce the identification in eq. (4), as well as the fact that the sum of the sources must reproduce the observed signal. The novel term in the cost function is the ℓ_1-norm penalty on the ℓ_2-norm of each source's initial conditions:

$$\sum_{i=1}^{K} \|u_i\|_2 \;=\; \sum_{i=1}^{K} \sqrt{\sum_{\tau=0}^{m-1} u_{i\tau}^2}. \tag{6}$$

This term originates in previous work on sparse decompositions of mixtures of periodic sources [6]. It favors sparse solutions in which many sources have zero excitation (i.e., all zero initial conditions, with $u_{i\tau} = 0$ for all τ); we interpret such sources as inactive. The two terms measuring modeling error and sparsity are balanced by the regularization parameter $\gamma > 0$.

2.3. Efficient optimization

The optimization in eq. (5) appears considerably more complicated than the optimization for BP in eq. (2). In particular, whereas BP merely computes a scalar amplitude β_i for each source, our approach computes a vector of initial conditions $\{u_{i\tau}\}_{\tau=0}^{m-1}$ and an extended waveform $\{s_{it}\}_{t=1}^T$. The extra complexity arises from the expressiveness of our approach, which explicitly models the potential variabilities of each source as opposed to representing them by fixed basis vectors. Though more complex than BP, the required optimization can be massaged into a simple tractable form which we describe in this section.

The first step to optimize eq. (5) is to eliminate the variables $\{s_{it}\}_{t>0}$ representing source waveforms. To do so, we introduce a Lagrange multiplier λ enforcing the sum constraint, thereby obtaining an unconstrained, continuously differentiable quadratic minimization over the variables $\{s_{it}\}_{t>0}$ and λ. We then eliminate these variables by expressing their optimal values in terms of the source initial conditions $\{u_{i\tau}\}_{\tau=0}^{m-1}$. The final result of this derivation leaves an unconstrained optimization to be performed over the source initial conditions. After some algebra, the remaining cost function (up to an additive constant) can be written in the form:

$$\mathcal{L}(u) = \; \frac{1}{2} \|Y - Zu\|_2^2 \;+\; \gamma \sum_{i=1}^{K} \|u_i\|_2, \tag{7}$$

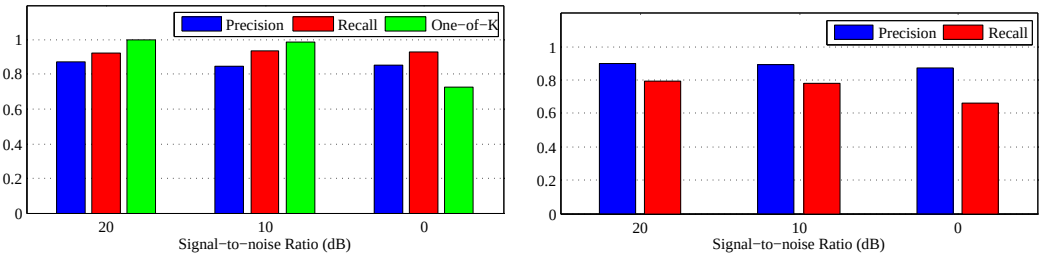

Fig. 1. Precision and recall results from analyzing signals containing a single aperiodic source (*left*) and a mixture of five aperiodic sources (*right*). See text for details.

where the variable u is the concatenation of the variables $u_{i\tau}$ across all sources and lags, and Y and Z are matrices expressed in terms of the observed signal x_t and dictionary coefficients $\alpha_{i\tau}$.

We briefly discuss how to minimize $L(u)$ using block coordinate relaxation methods [8] and ideas developed for group LASSO problems [7]. Basically, we select one set of initial conditions u_j from $\{u_i\}_{i=1}^K$ and minimize $L(u)$ with respect to these variables while holding all others fixed. We then iterate this procedure over all sets of initial conditions and repeat this loop until reaching the minimum of $L(u)$. In the key innermost step of this loop, we "shrink" u_j and set it to zero if doing so satisfies:

$$\|{Z_j}^\top (Y - Zu)\|_2 \quad \leq \gamma, \tag{8}$$

where the other components of u are held fixed to their current values while u_j is set to zero, and where Z_j is a matrix of size $T \times m$ derived from the j^{th} sub-block of the matrix Z. On the other hand, if the above condition is not satisfied, then we update u_j to a new non-zero value that minimizes the stationarity condition:

$${Z_j}^\top (Y - Zu) = \frac{\gamma u_j}{\|u_j\|_2}. \tag{9}$$

Though nonlinear, this equation is straightforward to solve. In particular, simple algebra yields a one-dimensional nonlinear equation for the magnitude $\|u_j\|_2$ (which can be solved by Newton's method); finally, given this magnitude, eq. (9) reduces to a linear set of equations for u_j.

2.4. Special case: periodic signals

Our approach builds on previous work for computing sparse decompositions of mixtures of periodic sources [6]. Specifically, this earlier work considered periodic sources whose periods were integer multiples of the sampling resolution. Such sources obey the simple recursion relation $s_{it} = s_{it-\tau_i}$, where τ_i is the integral period of the i^{th} source. Experiments showed that this approach was successful at recovering periodic sources from a mixed audio signal.

Our approach in this paper is based on autoregressive models satisfying the more general recursion relations in eq. (3). Within this framework, we can also model periodic sources with non-integer periods. In particular, for such sources, we can approximate the waveform's value at any arbitrary point in one cycle by an interpolation of its values in the preceding cycle. We develop this approximation by considering a single periodic source $\{s_t\}_{t=1}^T$ with non-integer period τ. (Here, for simplicity, we drop the index indicating its entry number in the dictionary.) Let $\{\tau_j\}_{j=0}^3$ denote the four integers closest in value to τ. Then we seek linear recursion coefficients such that:

$$s_t \;\approx\; \sum_{j=0}^{3} \alpha_{\tau_j} s_{t-\tau_j}. \tag{10}$$

Our approach here is to compute the coefficients α_{τ_j} that estimate s_t from a cubic interpolation of its "nearby" values in the preceding cycle at times $t - \tau_j$. Let $\Delta_j = t - \tau_j$ denote the lags which appear in the recursion relation, eq. (10). The coefficients which estimate s_t by cubic interpolation can be found by straightforward algebra. They are given by:

$$\begin{bmatrix} \alpha_{\tau_0} \\ \alpha_{\tau_1} \\ \alpha_{\tau_2} \\ \alpha_{\tau_3} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ \Delta_0 & \Delta_1 & \Delta_2 & \Delta_3 \\ \Delta_0^2 & \Delta_1^2 & \Delta_2^2 & \Delta_3^2 \\ \Delta_0^3 & \Delta_1^3 & \Delta_2^3 & \Delta_3^3 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{11}$$

Note that the recursion coefficients in these autoregressive models depend only on the period and the sampling rate (as opposed to any details of the periodic signals themselves). Higher-order interpolations can also be efficiently computed, though so far in our work we have not found them necessary.

3. EXPERIMENTAL RESULTS

We performed many sets of experiments on mixed signals containing aperiodic and periodic sources. All experiments were performed by analyzing 100 msec windows of signal sampled at 22050 Hz. To study the effect of varying signal-to-noise ratio (SNR), we corrupted the mixed signals by different levels of Gaussian noise. For each experiment, we optimized eqs. (5) and (7) and interpreted non-zero initial conditions $\{u_{i\tau}\}$ for the i^{th} source as a finding that the source was active. Within each set of experiments, we tuned the regularization parameter γ in eq. (5) to achieve the best average performance, which generally consisted of balancing the errors in precision and recall. For the experiments with single sources, we also report the classification performance assuming that the signal contains exactly one source and labeling it by whichever autoregressive model has the lowest fitting error—that is, without considering other sources in the decomposition. At high SNRs, this number may be viewed as an upper bound on the achievable results for precision and recall.

3.1. Mixtures of aperiodic sources

Our first experiments focused on mixtures of aperiodic sources, which we synthesized as follows. First, we constructed a dictionary of $K = 60$ sources, with each source parameterized by an autoregressive model of order $m = 32$. We randomly sampled the coefficients $\{\alpha_{i\tau}\}$ in eq. (3) from a normal distribution with zero mean and unit variance. The coefficients for each model were then rescaled so that each model was stable and its predictions would not diverge over time. We generated individual sources by randomly sampling initial conditions for their autoregressive models and then computing waveforms by evolving their recursion relations.

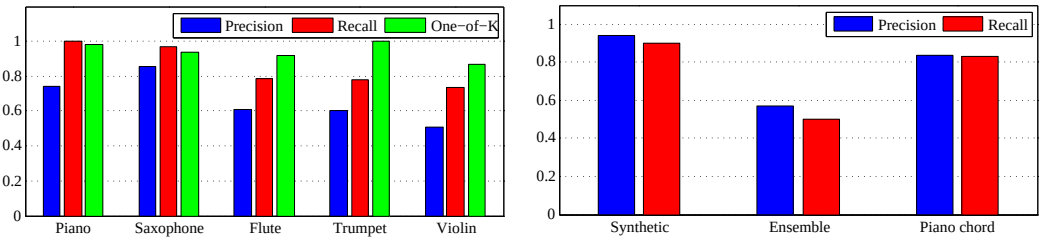

Fig. 2. Precision and recall results from analyzing signals containing a single musical note (*left*) and a mixture of five musical notes (*right*). See text for details.

We conducted two sets of experiments on aperiodic sources. In the first set, the mixed signals consisted only of a single source plus Gaussian noise. In the second set, the mixed signals consisted of a randomly chosen combination of five different sources plus Gaussian noise. We averaged our findings over 6000 experiments for single source identification (100 for each source) and 30000 experiments for mixtures of five sources. Fig. 1 shows the results in terms of precision, recall, and classification. On the whole, the results show that even from short 100 msec windows of observed signal, our approach often identifies the correct aperiodic sources and exhibits some robustness to noise as well.

3.2. Mixtures of musical notes

Our next experiments focused on mixtures of periodic signals derived from samples of musical instruments. For these experiments, we constructed a dictionary of $K = 60$ autoregressive models for periodic sources, using the cubic interpolation scheme described in section 2.4. The periods in these models corresponded to the notes C2–B6 on the musical scale. The instrument samples were taken from a public database [9]; we experimented with 100 msec clips of sampled notes from the piano, saxophone, flute, trumpet, and violin.

We conducted two sets of experiments on periodic sources. In the first set, the mixed signals consisted only of a sampled note from one instrument. The left panel of Fig. 2 shows the precision, recall, and classification results for these experiments. The bars represent results averaged over the sampled notes for each instrument: 60 for piano, 32 for saxophone, 37 for flute, 36 for trumpet, and 30 for violin. The results show that single musical notes are generally identified correctly even with instrument-independent models of idealized periodic sources. Most errors in these experiments were octave errors, as generally to be expected from results in pitch estimation.

In the second set of experiments, the mixed signals consisted of a randomly chosen combination of five different notes. We experimented with three different types of combinations: (i) random combinations of five notes, with one note sampled from each of the five available instruments; (ii) a five-finger chord on the piano, with notes that spanned at most two octaves (e.g., C4–C6); and (iii) a combination of five synthesized (perfectly periodic) sources with frequencies on the musical scale. The right panel of Fig. 1 reports averaged results over 30000 random combinations of types (i) and (iii) and all 29988 possible combinations of type (ii). The performance is best on the synthesized sources and worst (though still far better than chance) on the ensembles of notes from five different instruments.

4. CONCLUSION

In this paper we have proposed a framework for detecting sounds in mixed audio signals. Our framework extends basis pursuit by using autoregressive models to characterize the acoustic properties of different sources, as well as their degrees of variability. We demonstrated its feasibility for analyzing of mixtures of aperiodic and periodic sources. Compared to previous work on source separation, our framework was conceived to work in a somewhat different regime, where combinations of many possible sources must be considered.

Clearly, many further improvements could be imagined to the basic framework sketched here. In future work, we plan to explore algorithms for learning stable autoregressive models from sampled sounds and for integrating the inferences about active sources across multiple analysis frames. We are also interested in principled ways of setting the regularization parameter γ. Finally, we hope to assemble large dictionaries of diverse sounds and apply our approach to web-scale problems in audio information retrieval.

5. REFERENCES

[1] A. Hyvrinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems 13*. 2000, pp. 793–799, MIT Press.

[3] D. P. W. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G.J. Brown, Eds., pp. 115–146. John Wiley & Sons, Inc, 2006.

[4] D.L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, John Wiley & Sons, Inc, 2006.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[6] M. Nakashizuka, "A sparse decomposition method for periodic signal mixtures," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 91, no. 3, pp. 791–800, 2008.

[7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[8] S. Sardy, A. G. Bruce, and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 361–379, 2000.

[9] Lawrence Fritts, "The University of Iowa Musical Instrument Samples," 1997, `http://theremin.music.uiowa.edu/MIS.html`.