# Learning Dictionaries of Stable Autoregressive Models for Audio Scene Analysis

**Youngmin Cho** yoc002@cs.ucsd.edu
**Lawrence K. Saul** saul@cs.ucsd.edu
CSE Department, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0404 USA

## Abstract

In this paper, we explore an application of basis pursuit to audio scene analysis. The goal of our work is 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 sources, a small but unknown number combine and overlap to yield the observed signal. To infer which sounds are present, we decompose the observed signal as a linear combination of a small number of active sources. We cast the inference as a regularized form of linear regression whose sparse solutions yield decompositions with few active sources. We characterize the acoustic variability of individual sources by autoregressive models of their time domain waveforms. When we do not have prior knowledge of the individual sources, the coefficients of these autoregressive models must be learned from audio examples. We analyze the dynamical stability of these models and show how to estimate stable models by substituting a simple convex optimization for a difficult eigenvalue problem. We demonstrate our approach by learning dictionaries of musical notes and using these dictionaries to analyze polyphonic recordings of piano, cello, and violin.

## 1. Introduction

In this paper we study the problem of detecting when certain sounds are present in a mixed audio signal. We imagine that out of a large number of possible sounds, a small but unknown number are present and overlapping in the observed signal. We seek to detect the constituent sources, choosing from among entries that have been catalogued and represented in a large audio dictionary. We explore a novel framework in which the dictionary entries are themselves autoregressive models of time domain waveforms; these models characterize the acoustic variability of individual sources.

The problem we study arises in a number of settings and at many different levels of analysis. For example, at a low level of analysis, we might try to recover the score of a polyphonic recording using a dictionary whose entries represent individual musical notes (Cont, 2006). At a higher level of analysis, we might try to tag and index the frames of a movie soundtrack using a dictionary whose entries represent particular voices and sound effects (Chechik et al., 2008). We believe that the problem has many potential applications of commercial interest.

A large literature of related work addresses 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. The problem of source separation has been studied using methods from independent component analysis (Hyvarinen et al., 2001), computational auditory scene analysis (Wang & Brown, 2006), and probabilistic inference (Roweis, 2000). Compared to these other approaches, we generally start from different assumptions and work toward different goals. However, as shown in section 5, our approach can potentially be used for source separation.

Our paper builds on recently proposed frameworks for sparse decomposition of mixed audio signals (Nakashizuka, 2008; Cho & Saul, 2009). Previous studies addressed the problem of *inference* in this framework: given autoregressive models of time domain waveforms for a large number $K$ of possible sources, how to identify which $k \ll K$ of these sources occur in single microphone recordings? Using ideas from basis pursuit (Chen et al., 1998), this problem can be studied without performing an exponential

Appearing in *Proceedings of the 26^th International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

search through all possible $K!/(k!(K-k)!)$ combinations of sources. In this paper, we also address the problem of *learning* in this framework: how to estimate stable autoregressive models of possible sources that comprise the dictionary for basis pursuit?

The organization of this paper is as follows. In section 2, we review the generalized formulation of basis pursuit (Cho & Saul, 2009) in which each dictionary entry stores the coefficients of an autoregressive model. Given a mixed audio signal, a sparse decomposition into constituent sources is computed by balancing the modeling errors from individual sources against a regularization term on their initial conditions. The required optimizations are convex and can be solved efficiently by iterative descent algorithms.

In section 3, we discuss the problem of learning in this framework. We can estimate the coefficients of autoregressive models using simple least squares methods, but these methods do not guarantee that the resulting models are *stable*: i.e., that their predictions will not diverge as the models are evolved through time. Imposing stability constraints on the estimation procedure leads to a non-convex optimization. We review previous approaches to this problem in the more general setting of linear dynamical systems (Lacy & Bernstein, 2002; Siddiqi et al., 2008), then describe a simple approach that works well in our setting. In particular, by appealing to Gershgorin circle theorem for eigenvalue bounds (Golub & Loan, 1996), we show that stability is guaranteed by imposing a simple $\ell_1$-norm constraint on the model coefficients.

The last three sections of the paper present our empirical findings and conclusions. In section 4, we present experimental results on the learning of dictionaries of musical notes. Here we contrast the results obtained by autoregressive modeling in the time domain to those obtained by nonnegative matrix factorization (NMF) (Lee & Seung, 2001) in the magnitude frequency domain (Smaragdis & Brown, 2003; Cont, 2006; Cheng et al., 2008). In section 5, we use the dictionaries learned by autoregressive modeling and NMF to analyze polyphonic musical recordings. We compare the performance of these competing approaches in terms of the precision and recall of musical notes detected in sliding audio windows of each recording. Finally, in section 6, we conclude by suggesting several directions for future work.

## 2. Model

In this section we review how ideas from basis pursuit (Chen et al., 1998) have been extended to audio scene analysis with autoregressive models as dictionary elements (Nakashizuka, 2008; Cho & Saul, 2009).

### 2.1. Review of basis pursuit

Basis pursuit (BP) is a popular method for decomposing a signal $\vec{x}$ into an optimal superposition of dictionary elements $\{\vec{s}_i\}_{i=1}^{K}$. In particular, BP computes a sparse set of linear coefficients $\{\beta_i\}_{i=1}^{K}$ such that:

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

Typically, the dictionary is overcomplete, with the number of dictionary elements $K$ exceeding 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 $\ell_1$-norm of the linear coefficients subject to the constraint in eq. (1). It is convex with no local minima.

BP models the observed signal by an additive combination of dictionary elements (i.e., basis vectors) with varying amplitudes. This model is very well suited to data compression using Fourier or wavelet dictionaries. However, it is not ideally suited to audio scene analysis with dictionaries whose entries store waveforms of 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. While variations in phase can be efficiently encoded using shift-invariant schemes (Grosse et al., 2007), variations in duration and timbre do not admit similar solutions. Representing these variations explicitly by different dictionary elements would explode the dictionary size.

### 2.2. Extension to autoregressive modeling

For audio scene analysis, we wish to decompose the waveform $\{x_t\}_{t=1}^{T}$ of a mixed audio signal into the waveforms of its constituent sources. The waveforms of individual acoustic sources may exhibit considerable variability from one realization to the next. We can characterize this variability in a simple way using autoregressive models with linear predictive coefficients (Makhoul, 1975). Specifically, suppose that waveforms $\{s_{it}\}_{t=1}^{T}$ from the $i^{\text{th}}$ source approximately satisfy the $m^{\text{th}}$ order linear recursion relation

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

for samples at times $t > m$. The particular realization of the $i^{\text{th}}$ 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:

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

Note that each autoregressive model describes an $m$-dimensional family of signals in which the differing 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 $T$ can be described by simply evolving the recursion relation in eq. (3) for different numbers of time steps.

Basis pursuit can be extended by viewing these autoregressive models as dictionary entries for individual acoustic sources (Cho & Saul, 2009). Specifically, we can compute sparse decompositions of mixed audio signals $\{x_t\}_{t=1}^{T}$ in terms of a few ($k \ll K$) active sources by the following optimization:

$$\min_{s,u} \left\{ \frac{1}{2}\sum_{i=1}^{K}\sum_{t=1}^{T}\Big(s_{it} - \sum_{\tau=1}^{m}\alpha_{i\tau}\; s_{it-\tau}\Big)^2 + \gamma\sum_{i=1}^{K}\|u_i\|_2 \right\}$$

$$\text{subject to } x_t = \sum_{i=1}^{K} s_{it} \text{ and } s_{it} = u_{i|t|} \text{ for } t \leq 0. \tag{5}$$

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 final term (Nakashizuka, 2008) in the cost function is an $\ell_2$-norm penalty on 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 favors sparse solutions in which many sources have zero excitation (i.e., all zero initial conditions, with $u_{i\tau} = 0$ for all $\tau$); we interpret such sources as inactive. The two terms measuring modeling error and sparsity are balanced by the regularization parameter $\gamma > 0$ that also appears in eq. (5).

The optimization in eq. (5) is convex. It can be most efficiently solved by first eliminating the variables $\{s_{it}\}_{t>0}$ representing source waveforms. These variables appear quadratically with a simple linear constraint on their sum at each time $t$. Thus they can be eliminated analytically, leaving an unconstrained optimization to be performed over the source initial conditions $\{u_{i\tau}\}_{\tau=0}^{m-1}$. The final optimization over these variables has the same form as regularized linear regression in group Lasso problems (Yuan & Lin, 2006). More details on block coordinate relaxation methods (Sardy et al., 2000) for these problems are discussed elsewhere (Nakashizuka, 2008; Cho & Saul, 2009).

## 3. Learning

When prior knowledge is not available about individual acoustic sources, their autoregressive models must be learned from audio examples. In this section, we describe a stable alternative to simple least squares methods for this problem.

### 3.1. Least squares

To learn an autoregressive model for a particular acoustic source, we must estimate the linear predictive coefficients $\{\alpha_{i\tau}\}_{\tau=1}^{m}$ in eq. (3) from audio examples of that source's time domain waveforms. In this paper, we consider the simplest learning scenario in which the autoregressive models for different sources are estimated independently of one another. Then, after these different models are estimated, we simply collect and combine them to constitute the dictionary entries for basis pursuit. Because we focus on learning one autoregressive model at a time, in the rest of this section, we simplify notation by dropping the index $i$ indicating the particular source being modeled.

The most straightforward way to estimate each autoregressive model is to minimize the fitting error from the approximation in eq. (3). For simplicity, we assume that each model is being estimated from a single time domain waveform. In this case, we can estimate the linear predictive coefficient by solving the unconstrained optimization:

$$\min_{\alpha}\; \frac{1}{2}\sum_{t=m+1}^{T}\Big(s_t - \sum_{\tau=1}^{m}\alpha_\tau\; s_{t-\tau}\Big)^2. \tag{7}$$

Eq. (7) can be solved by simple least squares and pseudoinverse methods.

### 3.2. Stable least squares

An autoregressive model is stable if its predictions over time do not diverge. There are two reasons why we wish to estimate autoregressive models with this property. First, in practice we have observed that the numerical optimizations for inference in section 2.2 can be ill-conditioned or unstable when the autoregressive models are not constrained in this way. Second, we can view stability as a form of prior knowledge about the sources we are modeling; namely, we expect their waveforms to remain bounded over time. Though simple least squares methods suffice to minimize the approximation error in eq. (3), they do not necessarily yield stable models. Thus, for stability, we must impose other constraints on the estimation procedure.

The stability of an autoregressive model depends on the values of its linear predictive coefficients. We can see this dependence by formulating these models in a slightly different way. In particular, let $S_t = [s_t\ s_{t+1} \ldots s_{t+m-1}]^\top$ denote the vector of $m$ consecutive samples starting at time $t$. Then we can rewrite the recursion relation in eq. (3) as an $m^{\text{th}}$ order linear dynamical system that acts on these vectors. To specify this system, we define the update matrix:

$$A \;=\; \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \vdots & & & \vdots & & \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \alpha_m & \alpha_{m-1} & \alpha_{m-2} & \cdots & \alpha_2 & \alpha_1 \end{bmatrix}. \tag{8}$$

The nonzero elements of the update matrix $A$ consist of ones just above the diagonal, as well as the linear predictive coefficients $\alpha_\tau$, which appear in the bottom row. In terms of this matrix, we can rewrite the recursion relation in eq (3) as simply:

$$S_t \approx A^{t-1} S_1. \tag{9}$$

Eq. (9) shows that in general the model predictions will diverge unless $A$ has no eigenvalues of magnitude greater than one. Thus for stability we require:

$$\max |\lambda(A)| \leq 1. \tag{10}$$

The stability constraint in eq. (10) is easy to state but not to impose. In particular, the constraint is not convex, and the least squares problem in eq. (7) cannot be solved in closed form with eq. (10) imposed as an additional constraint.

The general problem of estimating stable dynamical systems (for arbitrary matrices $A$) has been widely studied. One tractable approach replaces the eigenvalue constraint in eq. (7) by the stricter constraint that singular values of $A$ are bounded by one (Lacy & Bernstein, 2002). This constraint is convex and equivalent to bounding the eigenvalues of $AA^\top$:

$$\max |\lambda(AA^\top)| \leq 1. \tag{11}$$

An even more successful approach relaxes the constraint in eq. (11), incrementally generating and imposing only as many linear constraints on the update matrix $A$ as are required to obtain a stable solution (Siddiqi et al., 2008).

Our problem requires a different starting point than the singular value constraint in eq. (11); in particular, this constraint is actually degenerate for matrices of the special form in eq. (8). For our problem, we instead appeal to the Gershgorin circle theorem, which accounts for how well the diagonal elements of a matrix approximate its eigenvalues (Golub & Loan, 1996). The theorem states that every eigenvalue of a matrix $A$ lies in the region:

$$\begin{aligned} &\lambda(A) \;\subseteq\; \bigcup_{i=1}^{m} D_i \\ &\text{where}\;\; D_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}. \end{aligned} \tag{12}$$

For matrices of the form in eq. (8), the Gershgorin circle theorem implies that the stability constraint in eq. (10) is guaranteed by the simple $\ell_1$-norm constraint:

$$\|\alpha\|_1 \leq 1. \tag{13}$$

We can therefore estimate stable autoregressive models by replacing the unconstrained optimization in eq. (7) by the constrained optimization:

$$\begin{aligned} &\min_{\alpha}\; \frac{1}{2} \sum_{t=m+1}^{T} \Big(s_t - \sum_{\tau=1}^{m} \alpha_\tau\ s_{t-\tau}\Big)^2 \\ &\text{subject to}\;\; \|\,\alpha\,\|_1 \;\leq\; 1. \end{aligned} \tag{14}$$

The $\ell_1$-norm regularization in eq. (14) not only yields stable autoregressive models, but also favors sparse solutions in which only a few of the $\alpha_\tau$ are non-zero. Such solutions have been observed to avoid overfitting in many other contexts (Tibshirani, 1996). The required optimization is convex and can be solved by standard packages.

## 4. Musical note dictionaries

We experimented with the methods in the previous section to learn autoregressive models of musical notes. The models were trained on publicly available recordings (Fritts, 1997) of 73 individual notes (C2–C8) on the piano. Specifically, each note's model was trained on a 90 msec window clipped from the loudest part of the note's recording and sampled at 22050 Hz. Musical notes provide a useful testbed for these methods because their waveforms can be idealized as periodic

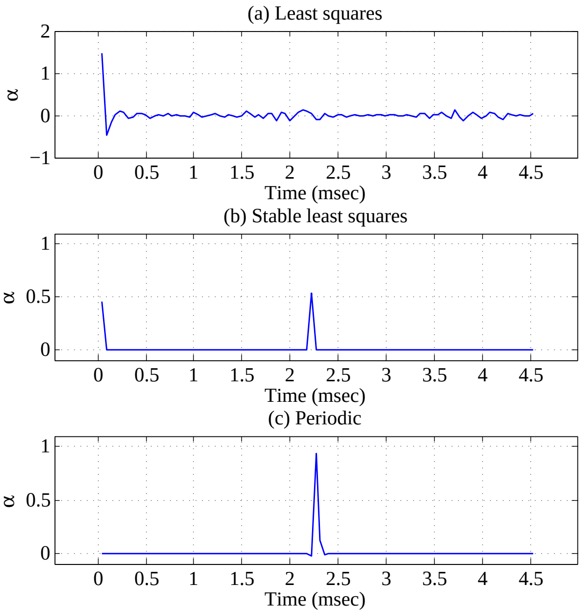

*Figure 1.* Coefficients of different autoregressive models for the piano note A4 with period 2.27 msec. Only the first 100 of $m = 350$ total coefficients are shown.

signals with known frequencies. This prior knowledge makes it easier to visualize, interpret, and compare the results of different training procedures. We also compared the results from autoregressive modeling to the spectral templates of musical notes estimated in the magnitude frequency domain.

### 4.1. Autoregressive models in the time domain

Fig. 1 shows the linear predictive coefficients of three different autoregressive models for the piano note A4. The model in the bottom plot was constructed by prior knowledge, so we explain it first as a useful reference. The musical note A4 has a frequency of 440 Hz. This periodicity is not perfectly realized by audio samples from any actual instrument recording, but we can construct a fairly accurate autoregressive model with this idealization in mind. Note that for a periodic signal whose period $\rho = k\Delta$ is exactly equal to an integer multiple $k$ of the sampling resolution $\Delta$, a simple but perfectly accurate autoregressive model (Nakashizuka, 2008) has zero coefficients everywhere except at $\alpha_k = 1$. More generally, for any periodic signal, we can construct a highly accurate autoregressive model by setting all its coefficients to be zero except at lags very close to the signal's period. The values of these non-zero coefficients can be determined by a simple polynomial (e.g., cubic) interpolation (Cho & Saul, 2009). The bottom plot in Fig. 1 shows these idealized coefficients for the musical note A4.

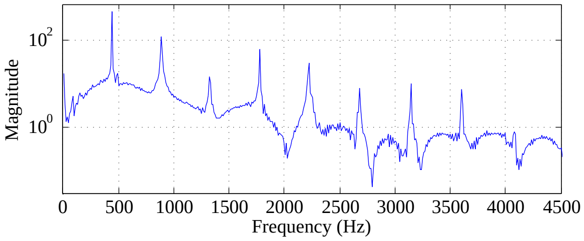

*Figure 2.* A spectral template for the piano note A4 with frequency 440 Hz.

The autoregressive model in the top plot of Fig. 1 was estimated by minimizing the approximation error in eq. (7). Though this least squares solution yields the smallest fitting error on the training samples, the coefficients do not reflect any underlying periodic structure. In fact, despite having been trained on multiple cycles of the note's period, the largest coefficient in this model is $\alpha_1$, indicating that the model attempts to predict each sample mainly from the immediately preceding one. This property was also observed for the models of other musical notes estimated by least squares.

The autoregressive model in the middle plot of Fig. 1 was estimated by the constrained optimization in eq. (14). The $\ell_1$-norm regularization in this optimization alleviates the susceptibility to noise of the least squares estimator. The coefficients of this model are not only stable, but also succeed in recovering the underlying periodic structure of the training data. Most of the coefficients are zero, except at very small lags and at lags near the actual period of A4. This behavior was also observed for the models of other musical notes trained by $\ell_1$-regularized least squares.

### 4.2. Spectral templates in the magnitude frequency domain

Fig. 2 shows the magnitude power spectrum of the piano note A4 for the same audio samples used to estimate the models in the previous section. The spectrum has peaks at integer multiples of 440 Hz, corresponding to the note's fundamental frequency. In addition, the shape of the spectrum conveys information about the timbre of this note on the piano.

To establish another reference for the autoregressive models in the previous section, we also constructed a dictionary of spectral templates, one for each note, in the magnitude frequency domain. These templates can be used to analyze mixtures of musical notes using ideas from nonnegative matrix factorization (NMF) (Smaragdis & Brown, 2003; Cont, 2006;

Cheng et al., 2008). Specifically, let $y_\mu$ denote the magnitude spectrum of an observed signal, and let $h_{i\mu}$ denote the spectral template for the $i^{\text{th}}$ musical note. We can compute the nonnegative weights $w_i \geq 0$ that best reconstruct $y_\mu \approx \sum_i w_i h_{i\mu}$ by minimizing one of two error measures:

$$\varepsilon_{\text{LS}} = \sum_\mu \left( y_\mu - \sum_i w_i h_{i\mu} \right)^2, \tag{15}$$

$$\varepsilon_{\text{KL}} = \sum_\mu \left( y_\mu \log \frac{y_\mu}{\sum_i w_i h_{i\mu}} - y_\mu + \sum_i w_i h_{i\mu} \right) \tag{16}$$

Eq. (15) computes the least squares error of the approximation, while eq. (16) computes a generalized form of the Kullback-Leibler divergence. The optimal weights that minimize both these errors can be computed by simple iterative update rules (Lee & Seung, 2001). In section 5, we also evaluate the performance of this approach.

We emphasize the basic modeling differences between this approach and the one in the last section. Minimizing the errors in eq. (15–16) requires a precise match of timbral profiles in the magnitude frequency domain, whereas the stable autoregressive models in the previous section will accurately fit any periodic signals at the modeled frequencies. In sum, the spectral templates offer increased specificity in the magnitude frequency domain, while the autoregressive models offer increased flexibility in the time domain. Our experiments in the next section evaluate the tradeoffs of these different approaches.

## 5. Evaluation

We evaluated the dictionaries of musical notes from the last section by using them to analyze polyphonic recordings. We analyzed three well-known pieces: Etude Op.25, No.2 by Chopin, “The Entertainer” by Joplin, and “La Donna e Mobile” by Verdi. The first two of these are fast solo piano pieces, while the third is a slower violin-cello duet. We worked with digital MIDI recordings[1] so that we could evaluate our inferred results against the musical score encoded by the MIDI format.

Our experiments compared the performance of five different analyses on waveforms sampled at 22050 Hz and segmented into non-overlapping 50 msec windows. For the autoregressive models in the time domain, we used the inference procedure in eq. (5) to identify constituent musical notes, interpreting notes as active if they had non-zero initial conditions. For the spectral templates in the magnitude frequency domain, we identified constituent musical notes by those having significant positive weights after minimizing the error measures in eqs. (15–16). Overall performance was measured by computing the precision and recall of identified notes averaged over all the windows in each song. For each song, we attempted to balance precision and recall by tuning the regularization parameter $\gamma$ in eq. (5) as well as the threshold value used to identify significant positive weights in the results from NMF.

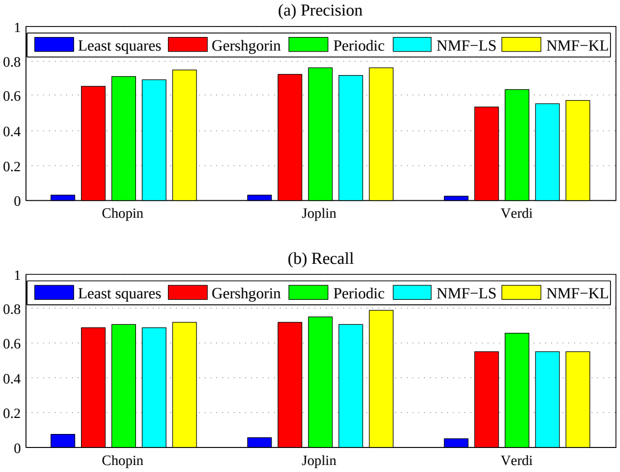

*Figure 3.* Precision and recall results from all experiments. See text for details.

Fig. 3 summarizes the precision and recall results from all of our experiments. Using the autoregressive models in the time domain, the best results are obtained by the dictionary of idealized periodic signals, followed closely by the dictionary whose entries were estimated by $\ell_1$-regularized methods. By contrast, the performance is essentially random for the dictionary of autoregressive models estimated by simple least squares methods. Using spectral templates in the magnitude frequency domain, the best results are obtained by nonnegative matrix factorization with the divergence criterion in eq. (16). The results also illustrate the tradeoffs between increased specificity in the magnitude frequency domain and increased flexibility in the time domain. On the piano pieces, the best performance is obtained from the spectral template dictionaries estimated from sampled piano notes, whereas on the violin-cello duet, the best performance is obtained by the dictionary of (instrument-independent) autoregressive models for idealized periodic signals.

Though basis pursuit with autoregressive modeling does not always yield the best results, we are encouraged by the overall level of performance as well as the specific breakdown of errors. Fig. 4 analyzes the er-

[1] Available at `www.piano-midi.de` and `www.8notes.com`.

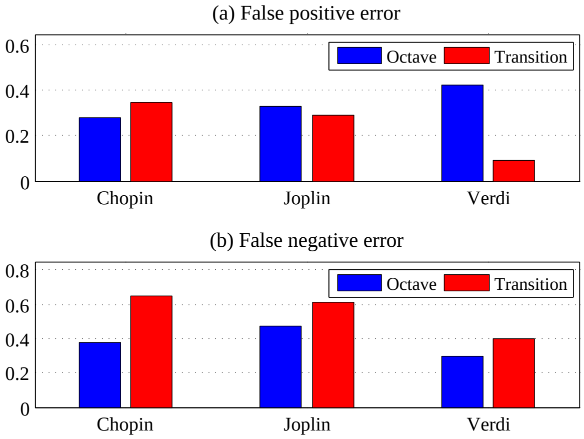

*Figure 4.* Breakdown of false positives and false negatives due to octave errors and/or note transition errors.

rors using the dictionary of autoregressive models estimated by $\ell_1$-regularized methods. In particular, the figure shows the percentages of false positives and negatives due specifically to octave and note transition errors. The former are common in musical pitch estimation; the latter are small alignment errors that arise when the model incorrectly detects notes in adjacent frames (yielding false positives), drops notes one frame before they actually disappear, or misses notes until one frame after they appear (yielding false negatives). A large percentage of the observed errors can be attributed to these causes, which do not seem especially egregious to us in the larger context of audio scene analysis. (For example, the number of transition errors appears to be directly correlated with the tempo of the analyzed pieces.) Discounting these types of errors, the precision and recall results for the $\ell_1$-regularized autoregressive models improve from roughly 70% to 90% on the two piano pieces. For improved performance, one might also consider more specialized approaches (Goto, 2006; Cont, 2006) for reducing these types of common errors in musical analysis.

Basis pursuit with autoregressive models can also be used for source separation. The framework was conceived with different goals in mind, but because it operates directly on time domain waveforms, it can be also used to reconstruct individual sources. In particular, this approach does not need to fill in missing phase information as when signals are decomposed in the magnitude frequency domain. Fig. 5 shows an example of source separation on a 50 msec window from the violin-cello duet by Verdi. At this point in the score, the violin is playing A4 and the cello is playing D3 and C4. These notes were correctly detected by the optimization in eq. (5) using the dictionary of autoregressive models estimated by $\ell_1$-regularized methods. To separate the violin from the cello, we then decomposed the observed signal in this window using only the dictionary entries for the detected notes. Finally, using prior knowledge of the frequency ranges of these instruments, the waveforms inferred for D3 and C4 were ascribed to the cello, and the waveform inferred for A4 was ascribed to the violin. The figure contrasts the predicted and actual waveforms of violin and cello in this window.

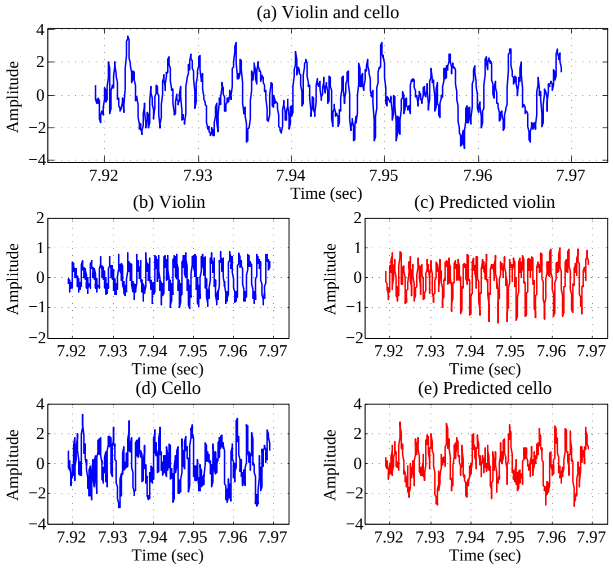

*Figure 5.* Example of source separation: cello and violin.

## 6. Conclusion

In this paper, we have explored a framework to detect sounds in mixed audio signals. Our framework can be viewed as a generalization of basis pursuit in which the dictionary entries are themselves stable autoregressive models. The results in the previous section suggest several directions for further study. For example, the approach in section 4.1 has the weakness that it only learns to model the periodicity but not the timbre of musical notes. Can we remedy this deficiency by additionally incorporating a statistical model for each note that distinguishes between likely versus unlikely sets of initial conditions? More generally, for both musical and non-musical signals, many interesting questions arise naturally in this framework. Can we learn dictionaries of autoregressive models in an unsupervised fashion, not from isolated recordings of individual sources, but directly from mixed audio? Can we catalogue the types of real-world sounds that are accurately described by autoregressive models? Finally, can we model the onsets and offsets of sources across time by extending this framework with ideas from hidden Markov modeling? These questions and others are open for future work.

## Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant No. 0812576. We thank the reviewers for many helpful comments and suggestions.

## References

Chechik, G., Ie, E., Rehn, M., Bengio, S., & Lyon, D. (2008). Large-scale content-based audio retrieval from text queries. *Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval (MIR-08)* (pp. 105–112). ACM.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, *20*, 33–61.

Cheng, C., Hu, D. J., & Saul, L. K. (2008). Nonnegative matrix factorization for real time musical analysis and sight-reading evaluation. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-08)* (pp. 2017–2020).

Cho, Y., & Saul, L. K. (2009). Sparse decomposition of mixed audio signals by basis pursuit with autoregressive models. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-09)* (pp. 1705–1708).

Cont, A. (2006). Realtime multiple pitch observation using sparse non-negative constraints. *Proceedings of the International Symposium on Music Information Retrieval (ISMIR-06).*

Fritts, L. (1997). The University of Iowa Musical Instrument Samples. `http://theremin.music.uiowa.edu/MIS.html`.

Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations*. The Johns Hopkins University Press.

Goto, M. (2006). Analysis of musical audio signals. In D. Wang and G. Brown (Eds.), *Computational auditory scene analysis: Principles, algorithms, and applications*, 251–295. John Wiley & Sons, Inc.

Grosse, R., Raina, R., Kwong, H., & Ng, A. Y. (2007). Shift-invariant sparse coding for audio classification. *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)* (pp. 149–158).

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. John Wiley & Sons.

Lacy, S. L., & Bernstein, D. S. (2002). Subspace identification with guaranteed stability using constrained optimization. *Proceedings of the American Control Conference* (pp. 3307–3312).

Lee, D. D., & Seung, H. S. (2001). Algorithms for nonnegative matrix factorization. *Advances in Neural Information Processing Systems 14* (pp. 556–562). MIT Press.

Makhoul, J. J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*, 561–580.

Nakashizuka, M. (2008). A sparse decomposition method for periodic signal mixtures. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *91*, 791–800.

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

Sardy, S., Bruce, A. G., & Tseng, P. (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of Computational and Graphical Statistics*, *9*, 361–379.

Siddiqi, S., Boots, B., & Gordon, G. (2008). A constraint generation approach to learning stable linear dynamical systems. *Advances in Neural Information Processing Systems 20* (pp. 1329–1336). MIT Press.

Smaragdis, P., & Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 177–180).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, *58(1)*, 267–288.

Wang, D., & Brown, G. J. (Eds.). (2006). *Computational auditory scene analysis: Principles, algorithms, and applications.* John Wiley & Sons, Inc.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*, 49–67.