Learning Dictionaries of Stable Autoregressive Models for Audio Scene Analysis

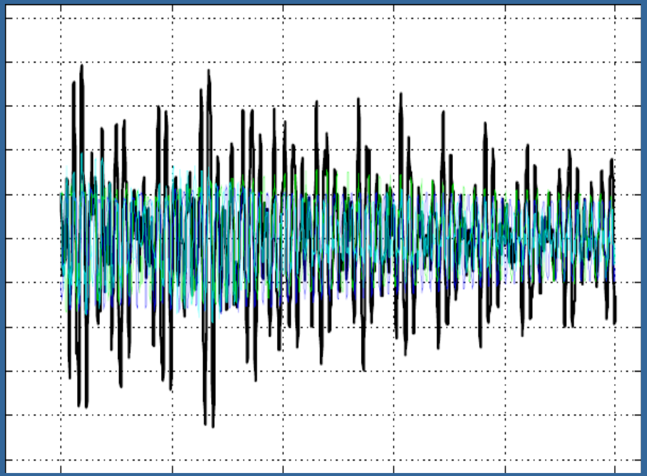
Youngmin Cho & Lawrence Saul

{yoc002,saul}@cs.ucsd.edu

University of California, San Diego

Audio Scene Analysis

- What do you hear?
  1. Toilet flushing 
  2. Doorbell ringing 
  3. Dog barking 
  4. Glass breaking 
  5. Shotgun firing 



- What didn't you hear?




- Long-term goal : annotating audio libraries
- This work : preliminary exploration

Outline

- Problem description
  - Audio scene analysis
- Our approach
  - Inference - Basis pursuit w/ autoregressive models
  - Learning – Regularized least squares
- Experimental results
- Summary and future work

Audio Scene Analysis

- How to detect when certain sounds are present in mixed signals?

- Assumptions

  Single microphone recordings

  Large number K of possible sources

  Sparse coding : out of many possible sources, only a few $k \ll K$ appear.

Main Issues

- Scaling with dictionary size K
  How to avoid $K!/(k!(K-k)!)$ combinatorial search?
- Modeling acoustic variability of sources
  How to represent it efficiently?
- Learning dictionaries from examples
  How to estimate stable models?

Outline

- Problem description
  Audio scene analysis
- Our approach
  Inference - Basis pursuit w/ autoregressive models
  Learning – Regularized least squares
- Experimental results
- Summary and future work

Basis Pursuit (BP)

- Analyzes signal as optimal superposition of overcomplete dictionary elements.

- Given: observed signal $x \in \mathbb{R}^T$,
  dictionary elements $\{s_i \in \mathbb{R}^T\}_{i=1}^K$,

$$\min \sum_{i=1}^{K} |\beta_i| \text{ subject to } x = \sum_{i=1}^{K} \beta_i s_i$$

: L^1-norm penalty favors sparse solutions.

BP for Audio Scene Analysis?

- Not suited for dictionaries whose entries store waveforms of natural sounds.

- Such sounds are likely to exhibit many variations.

- Representing these variations by different entries would explode the dictionary size.

- Big idea : store models, not waveforms as dictionary entries.

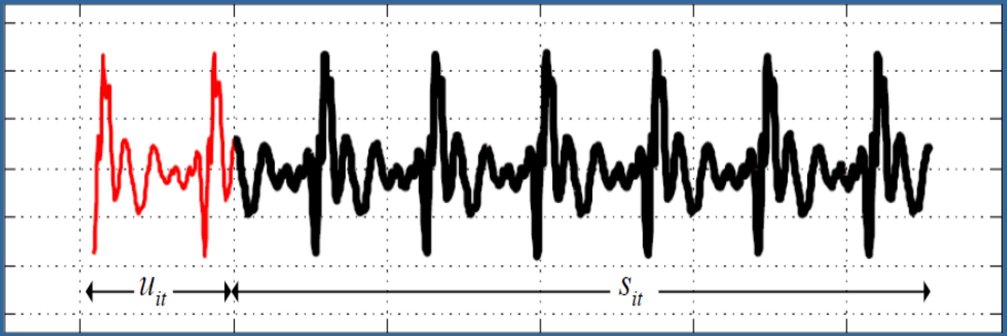
Autoregressive Models

- Linear predictive modeling

  Predicts a target value by a linear combination of previous samples
- Assumption: i^{th} source $\{s_{it}\}_{t=1}^{T}$ can be approximated by a linear predictive model.

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

$$s_{it} = u_{i|t|} \quad \textbf{for } t \leq 0.$$



- $\{\alpha_i\}_{i=1}^{K}$ are stored as dictionary entries.

Extension of BP with AR Models

- Given: observed signal $x \in \mathbb{R}^T$,
  dictionary elements $\{\alpha_i\}_{i=1}^K$,

$$\min_{s,u} \left\{ \frac{1}{2} \sum_{i=1}^{K} \sum_{t=1}^{T} \left(s_{it} - \sum_{\tau=1}^{m} \alpha_{i\tau}\, s_{it-\tau} \right)^2 + \gamma \sum_{i=1}^{K} \sqrt{\sum_{\tau=0}^{m-1} u_{i\tau}^2} \right\}$$

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

- Objectives
  Fit individual sources to autoregressive models.
  Favor sparse solutions.
  Balance objectives by regularization parameter γ.
- Constraints
  Sources must reconstruct signal.
  Sources must match initial conditions.

Outline

- Problem description

  Audio scene analysis

- Our approach

  Inference - Basis pursuit w/ autoregressive models

  Learning – Regularized least squares

- Experimental results

- Summary and future work

Dictionary Learning

- How to learn AR model α for particular acoustic source $s \in \mathbb{R}^T$?

- Unconstrained least squares

$$\min_{\alpha} \sum_{t=m+1}^{T} \left(s_t - \sum_{\tau=1}^{m} \alpha_\tau \, s_{t-\tau} \right)^2.$$

Learning Stable Models

- Option #1 : Preprocessing the waveform (e.g., windowing)

- Option #2 : Postprocessing the model (e.g., scaling)

- Option #3 : Integrating stability into estimation (e.g., our approach)

Stability Constraint

- Least squares with stability constraint
- Representing AR model as linear dynamical system A

$$\max|\lambda(A)| \leq 1 \ \textbf{where} \ A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \vdots & & & \vdots & & \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \alpha_m & \alpha_{m-1} & \alpha_{m-2} & \cdots & \alpha_2 & \alpha_1 \end{bmatrix}.$$

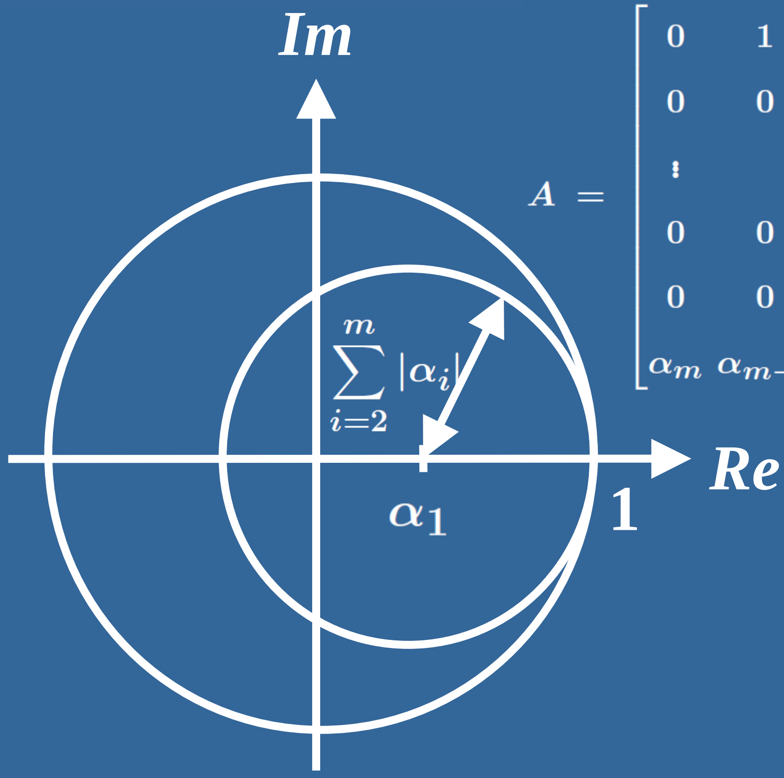
Not convex!

Stable Least Squares

- Least squares with L^1-norm regularization

$$\min_{\alpha} \sum_{t=m+1}^{T} \left(s_t - \sum_{\tau=1}^{m} \alpha_\tau \, s_{t-\tau} \right)^2 \quad \textbf{subject to} \ \| \alpha \|_1 \leq 1.$$

- Stability implied by Gershgorin circle theorem, which locates eigenvalues in complex plane.



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

Outline

- Problem description

  Audio scene analysis

- Our approach

  Inference - Basis pursuit w/ autoregressive models

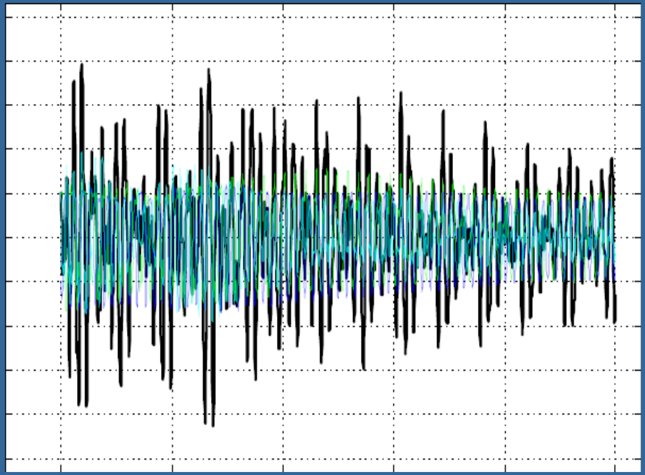
  Learning – Regularized least squares

- Experimental results

- Summary and future work

Audio Scene Analysis

- What do you hear?
  1. Toilet flushing 
  2. Doorbell ringing 
  3. Dog barking 
  4. Glass breaking 
  5. Shotgun firing 



- What didn't you hear?




- Long-term goal : annotating audio libraries
- This work : preliminary exploration

Musical Analysis as Simple Benchmark

- Entries : K=73 notes (C2-C8) on the piano

- Training data : 90 msec clips of 22050 Hz
  piano recordings

- Testing data
  Matched : Chopin & Joplin (fast solo piano)
  
  Mismatched : Verdi (slower violin-cello duet)

- Precision = # correct notes / # detected notes
- Recall = # correct notes / # true notes

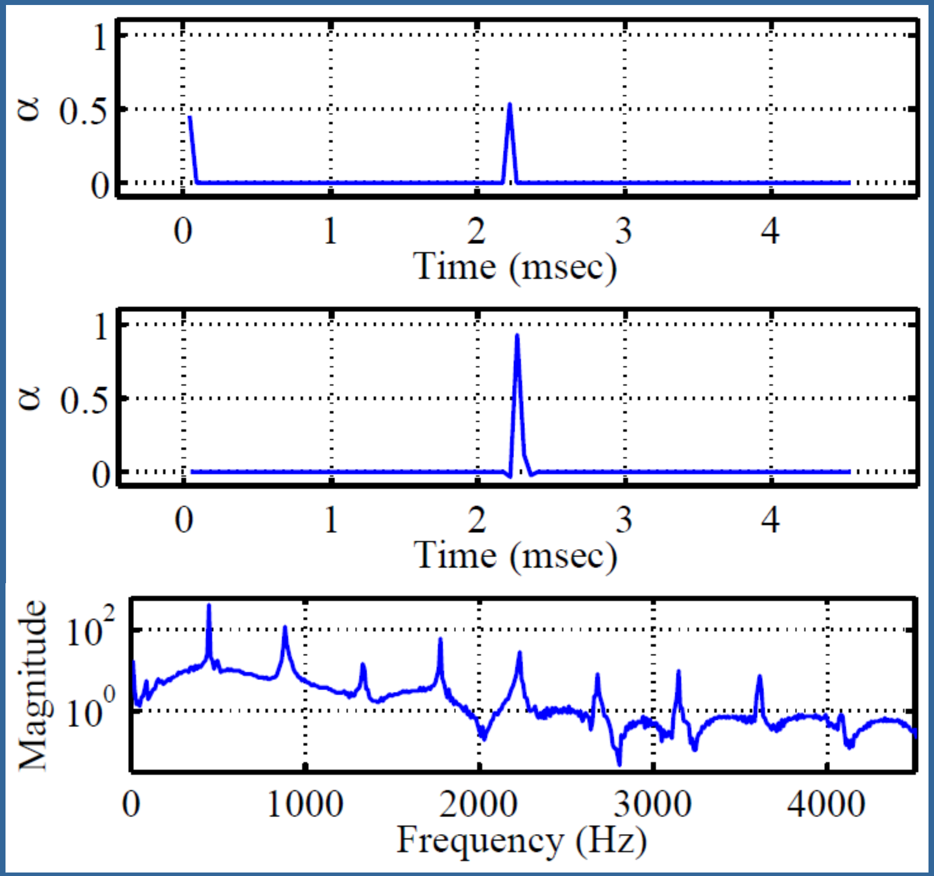
Musical Note Dictionaries

- Example : A4 with period 2.27 msec (440 Hz)
- Models

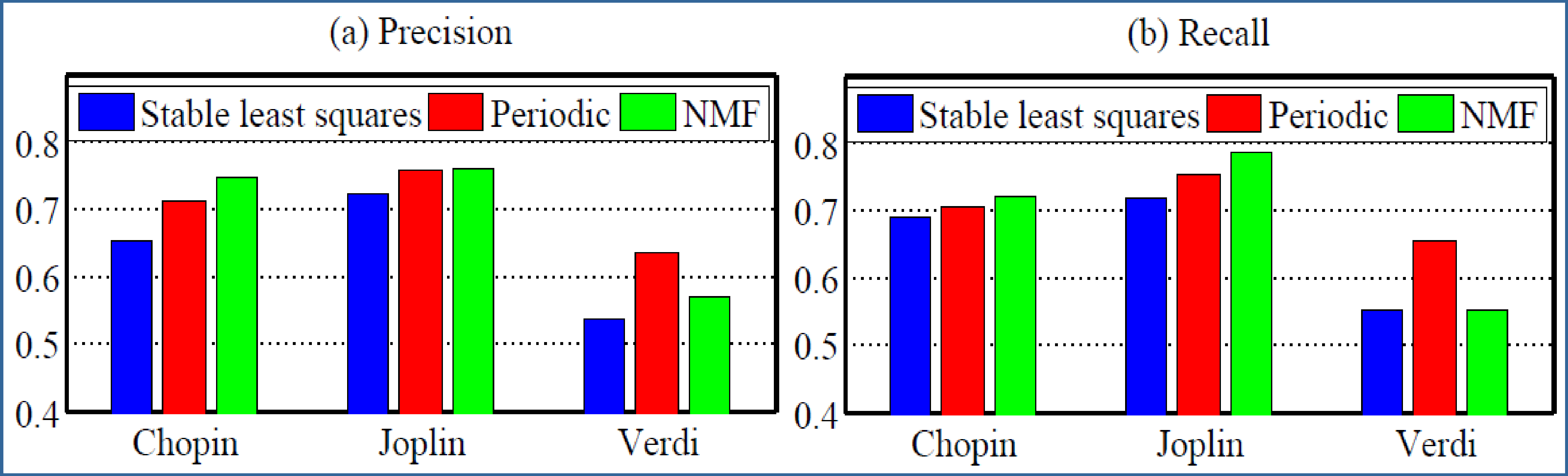
  1) Stable least squares
     (L^1-norm regularization)

  2) Periodic
     (prior knowledge)

  3) Non-negative matrix factorization
     (magnitude frequency domain)



Song Transcription



- What works
  - Inferring constituent sources on the whole
- What needs improvement
  - Learning timbre of musical notes
- Typical errors
  - Octave confusions and note boundaries

Summary & Future Work

- What we have done
  - Extending BP using autoregressive models
  - Learning stable autoregressive models
- What next
  - Large dictionaries of diverse sounds (non-musical, non-periodic)
  - Statistical modeling of initial conditions
  - Unsupervised learning