ISSE – An Interactive Source Separation Editor, Part II

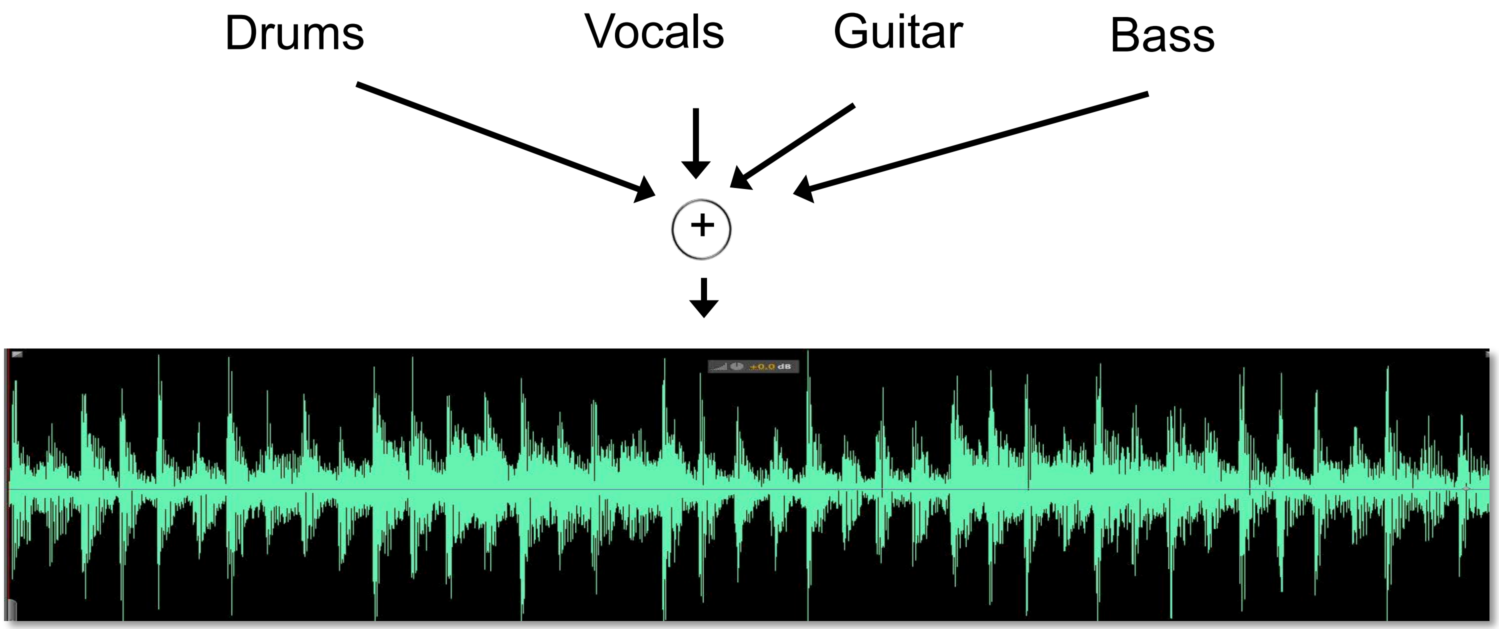
Nicholas J. Bryan
Stanford University

Overview

- Introduction
- Background
- Approach
- Algorithm
- Evaluation
- Conclusion

Motivation

- Real world sounds are mixtures of many individual sounds.



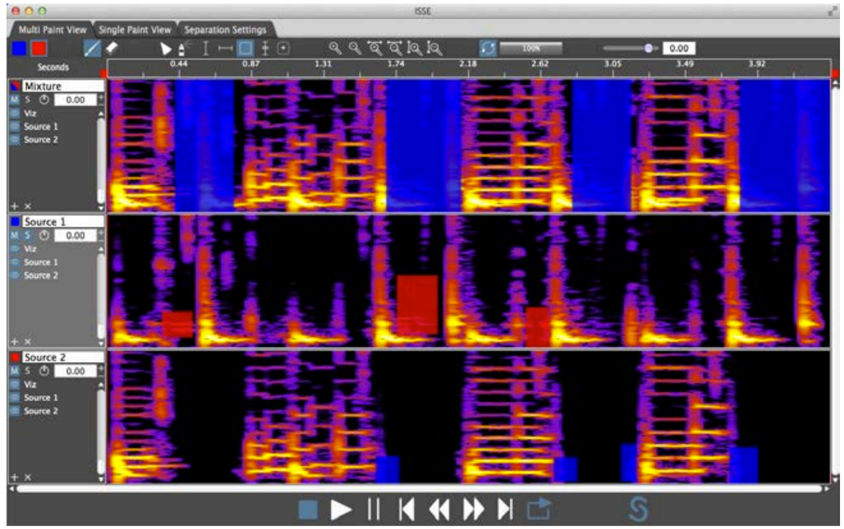
Applications I

- Denoising
- Audio post-production and remastering
- Spatial audio and upmixing
- Music Information Retrieval

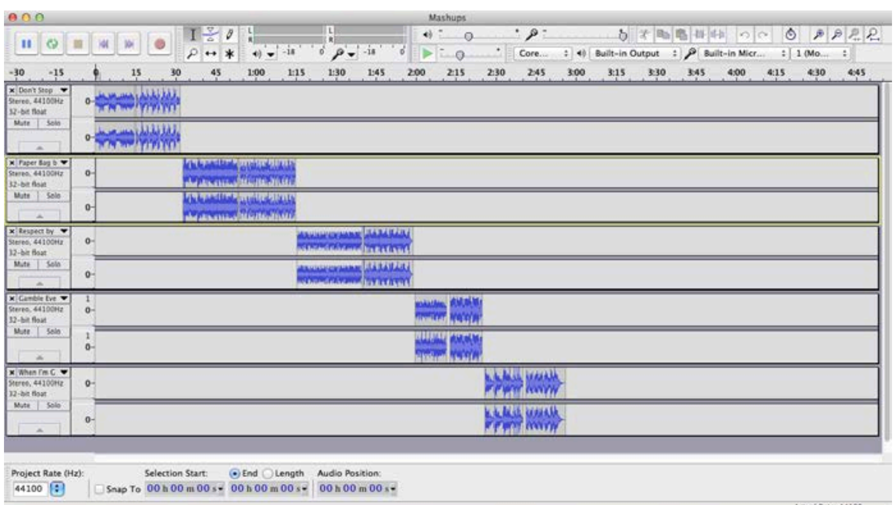
Applications

- Music remixing and content creation.
- Human-computer interaction perspective.
- How does a end-user perform source separation?

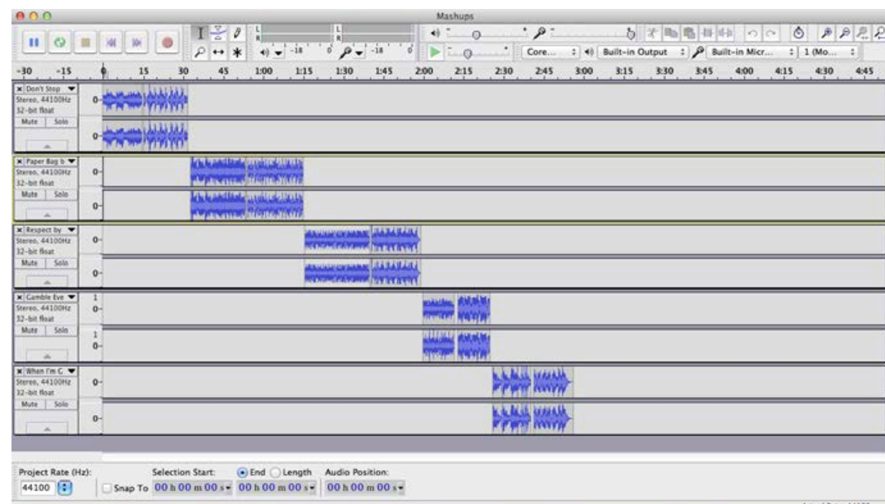
Live Demonstration + Sound Examples



Live demo



Vocal extraction remixes



Piano, coughing, denoising

Note

- Machine learning algorithm that **adapts** to user annotations.
- **Not copying** the pixel data underneath the annotations.
- A **local** annotation can have a **global** effect.

Overview

- Motivation
- Background
- Approach
- Algorithm
- Evaluation
- Conclusion

Overview of Techniques

Independent component analysis

Microphone arrays

Adaptive signal processing

Computational auditory scene analysis

Spectral processing

Sinusoidal modeling

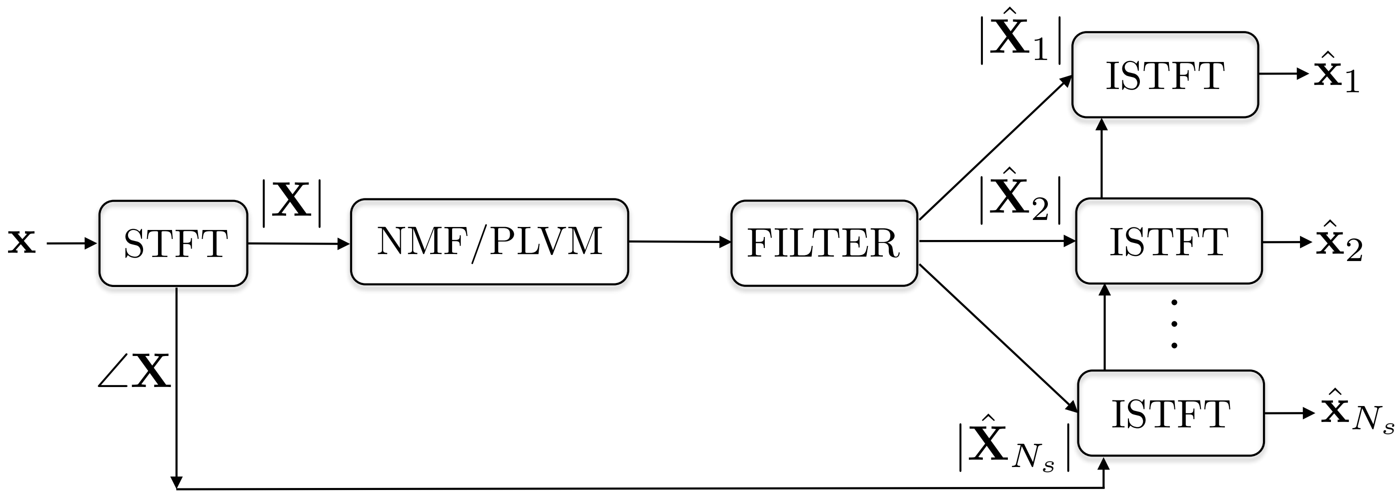
Time-frequency selection

Classical denoising and speech enhancement

Non-Negative Matrix Factorization (NMF) and Related Probabilistic Latent Variable Models (PLVM)

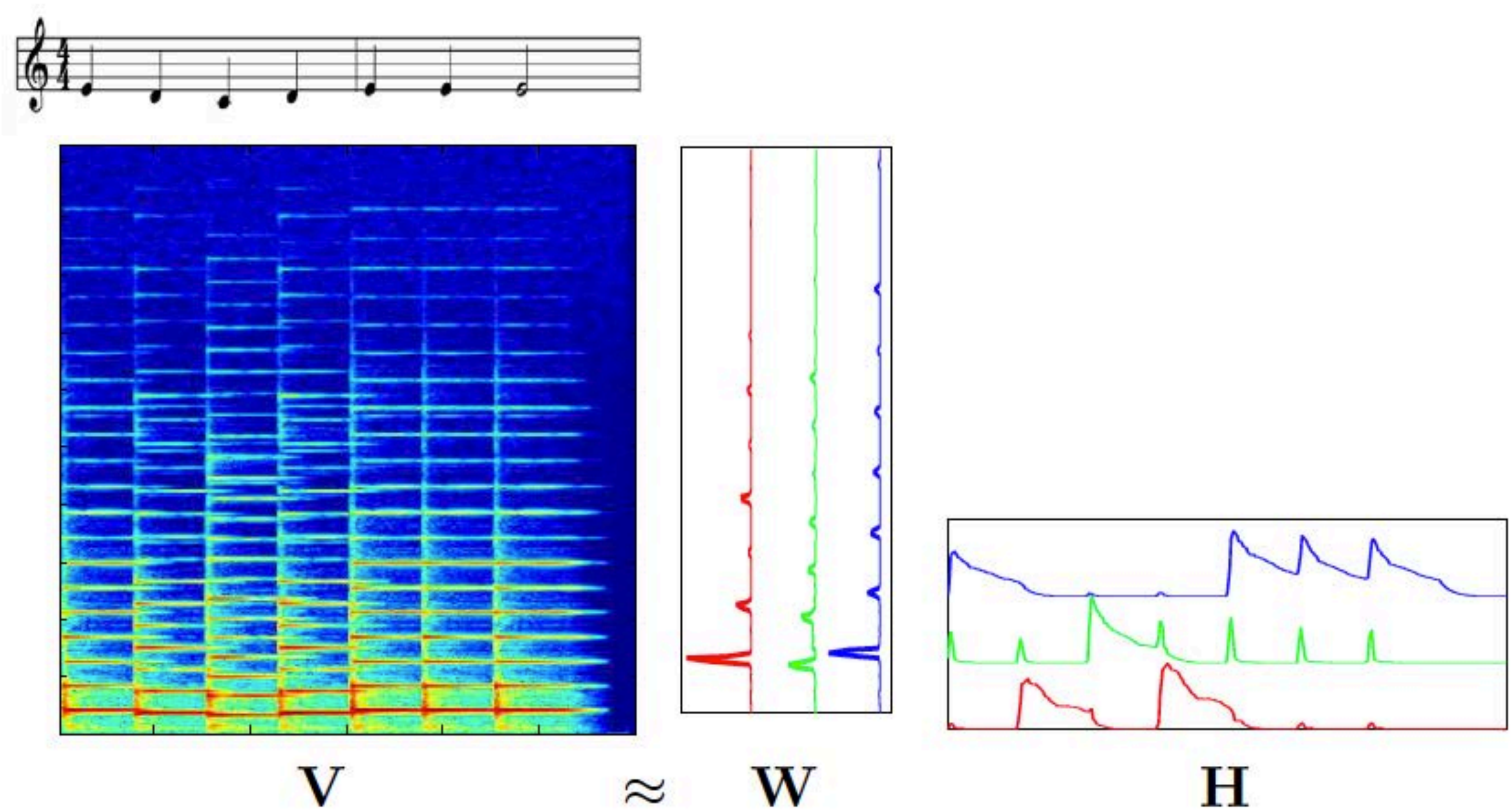
- **Machine learning**, data-driven, basis decomposition, dictionary.
- **Model** each sound **source** within a mixture.
- **Linear combination** of prototypical frequency **spectra**.
- Well suited to our **motivation**.
- **Monophonic** and/or stereophonic recordings.
- One of the most **promising** separation methods of the past decade.
  - NMF [Lee & Seung, 1999, 2001; Smaragdis & Brown 2003]
  - PLVM [Raj & Smaragdis 2005, Smaragdis et al., 2006]

Block Diagram



- Transform signal via the short-time Fourier transform (**STFT**).
- Compute a **NMF/PLVM.**
- **Filter** mixture sound.
- Inverse **STFT.**

The STFT and NMF



- The basis vectors capture prototypical frequency content.
- The weights capture the gains of the basis vectors.

Non-Negative Matrix Factorization

$$\overset{\text{Data}}{\left[\quad \mathbf{V} \quad\right]} \approx \overset{\text{Basis Vectors}}{\left[\quad \mathbf{W} \quad\right]} \overset{\text{Weights}}{\left[\quad \mathbf{H} \quad\right]}$$

- A matrix factorization where everything is non-negative.
- $\mathbf{V} \in \mathrm{R}_+^{F \times T}$ - original non-negative data
- $\mathbf{W} \in \mathrm{R}_+^{F \times K}$ - matrix of basis vectors, dictionary elements
- $\mathbf{H} \in \mathrm{R}_+^{K \times T}$ - matrix activations, weights, or gains
- $K < F < T$ (typically)

Optimization Formulation

$$\underset{\mathbf{W},\mathbf{H}}{\arg\min} \;\; D(\mathbf{V} \,|\, \mathbf{W}\,\mathbf{H})$$

$$\text{subject to}\, \mathbf{W} \geq 0, \mathbf{H} \geq 0$$

- Minimize the *divergence* between V and WH.

$$D_{EUC}(\mathbf{V} \,|\, \mathbf{W}\,\mathbf{H}) = \sum_f \sum_t (V_{ft} - [\mathbf{W}\,\mathbf{H}]_{ft})^2$$

$$D_{KL}(\mathbf{V} \,|\, \mathbf{W}\,\mathbf{H}) = \sum_f \sum_t (V_{ft} \log \frac{V_{ft}}{[\mathbf{W}\,\mathbf{H}]_{ft}} - V_{ft} + [\mathbf{W}\,\mathbf{H}]_{ft})$$

$$D_{IS}(\mathbf{V} \,|\, \mathbf{W}\,\mathbf{H}) = \sum_f \sum_t \left(\frac{V_{ft}}{[\mathbf{W}\,\mathbf{H}]_{ft}} - \log \frac{V_{ft}}{[\mathbf{W}\,\mathbf{H}]_{ft}} - 1 \right)$$

- At best, find a local optima (not convex).

Iterative Numerical Optimization

- How do we solve for W and H?
- Use block coordinate descent.
  - Solve for W
  - Solve for H
  - Repeat
- Use Majorization-Minimization.
  - Lower bounding algorithm
  - Use rules of convexity
  - Converges to local optima
- Alternative optimization methods.
  - Projected gradient descent
  - Projected Newton's methods
  - Interior point methods (overkill)

$$\arg\min_{\mathbf{W}} \; D(\mathbf{V} \,|\, \mathbf{W}\,\mathbf{H}) \quad \text{subject to} \quad \mathbf{W} \geq 0$$

$$\arg\min_{\mathbf{H}} \; D(\mathbf{V} \,|\, \mathbf{W}\,\mathbf{H}) \quad \text{subject to} \quad \mathbf{H} \geq 0$$

NMF Parameter Estimation via MM

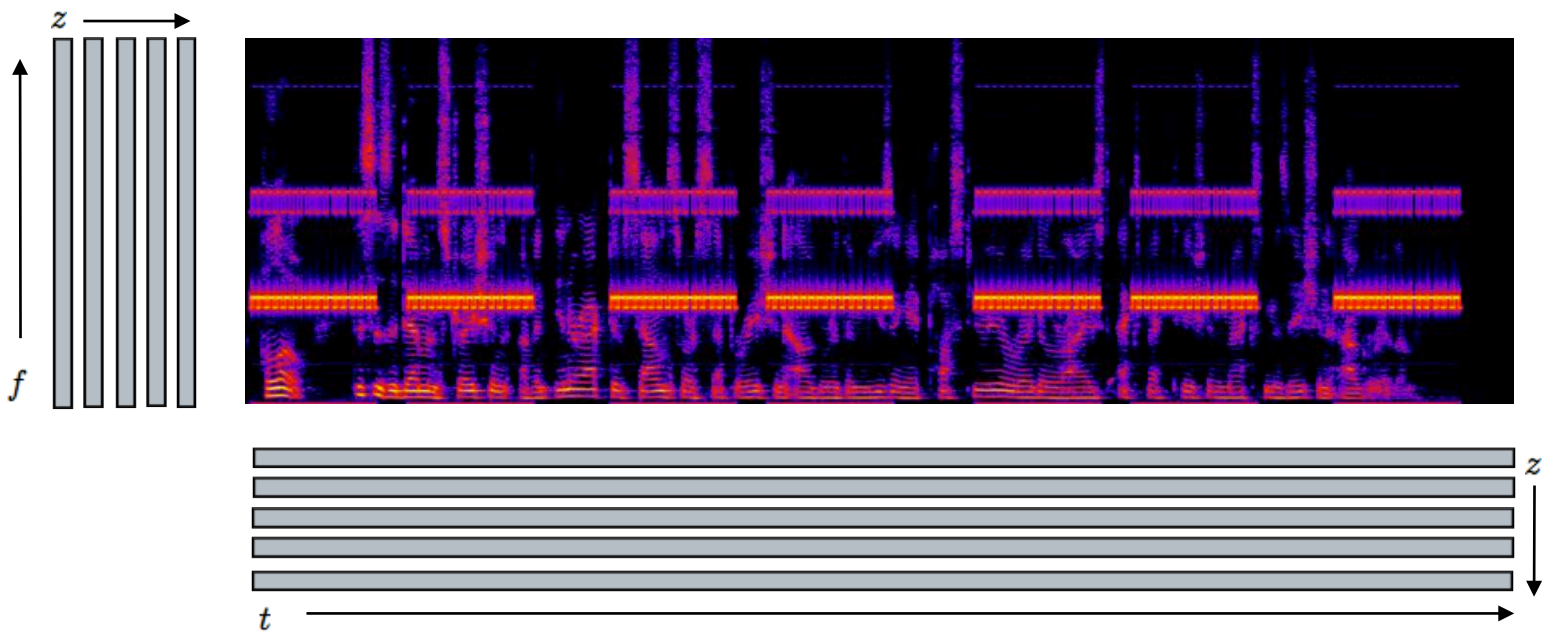
- Initialize to positive random.
- Repeat until convergence.

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\left(\frac{\mathbf{V}}{\mathbf{W}\,\mathbf{H}}\right)\mathbf{H}^{\mathrm{T}}}{\mathbf{1}\,\mathbf{H}^{\mathrm{T}}}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^{\mathrm{T}}\left(\frac{\mathbf{V}}{\mathbf{W}\,\mathbf{H}}\right)}{\mathbf{W}^{\mathrm{T}}\,\mathbf{1}}$$

Non-Negative Matrix Factorization

$$\mathbf{V} \approx \mathbf{WH}$$



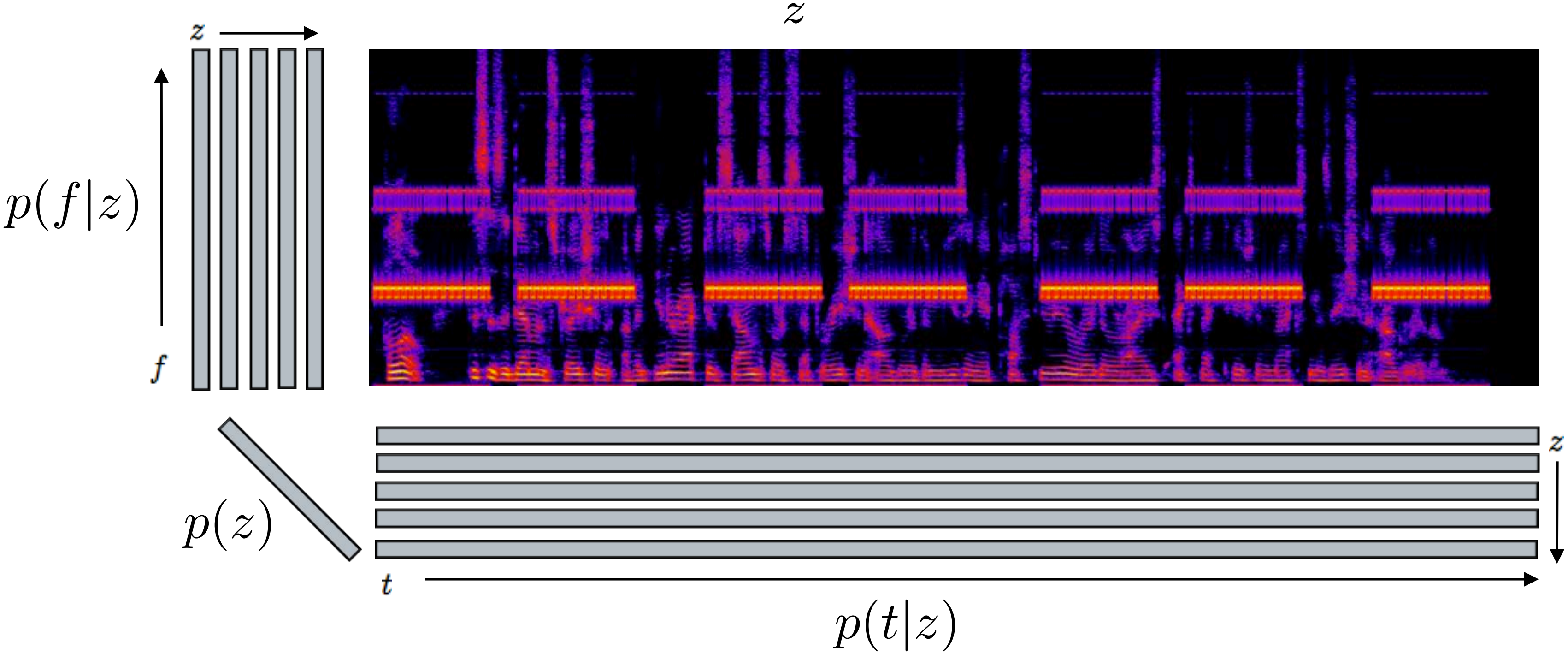
$\mathbf{W}$ Basis vectors, frequency components, dictionary

$\mathbf{H}$ Time activations or gains

Probabilistic Latent Variable Model (PLVM)

- Probabilistic latent component analysis (PLCA).

$$\mathbf{V} \approx p(f,t) = \sum_z p(z)p(f|z)p(t|z)$$

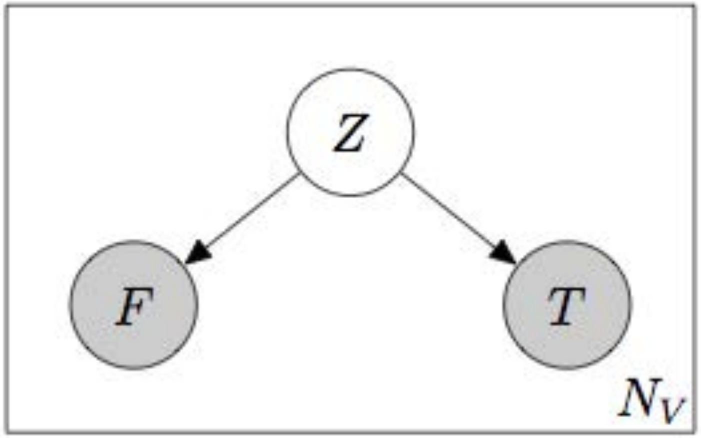


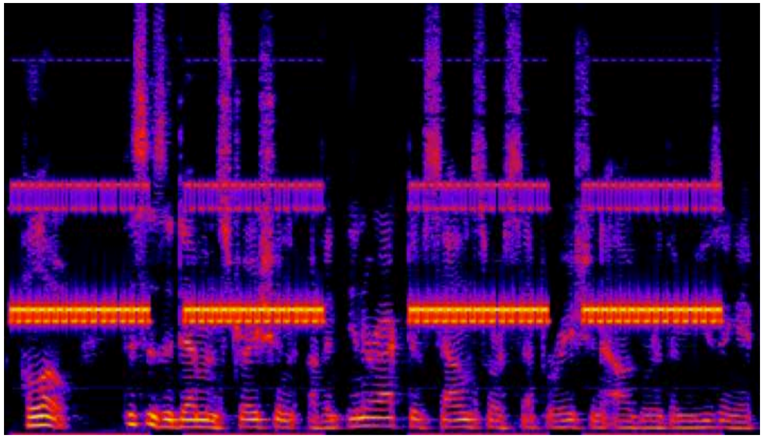
$p(f|z)$ Basis vectors, frequency components, dictionary

$p(z)$ Latent component weights

$p(t|z)$ Time activations or gains

Generative Model





1. For $n = 1, \ldots, N_V$ times, where $N_V = \sum_f \sum_t V_{ft}$,

   (a) Generate a latent variable $z^{(n)} \sim p_Z(z) := Multinomial(N_V, \boldsymbol{\pi}^{(z)})$.

   (b) Generate a frequency $f^{(n)}|z^{(n)} \sim p_{F|Z}(f|z) := Multinomial(N_V, \boldsymbol{\pi}^{(f|z)})$.

   (c) Generate a time $t^{(n)}|z^{(n)} \sim p_{T|Z}(t|z) := Multinomial(N_V, \boldsymbol{\pi}^{(t|z)})$.

2. Set V_{ft} equal to the count of the occurrence of each outcomes value pair (f, t). Discard all samples of the latent variable z.

Maximum Likelihood Parameter Estimation

- Formulate the log-likelihood of our model.

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\Theta}|\,\mathbf{V}) &= \ln p(\mathbf{V}\,|\boldsymbol{\Theta}) \\
&= \ln \frac{(\sum_f \sum_t V_{ft})!}{V_{11}!V_{12}!\dots V_{ft}!} \prod_{f=1}^{N_F} \prod_{t=1}^{N_T} p(f,t)^{V_{ft}} \\
&= \ln \frac{(\sum_f \sum_t V_{ft})!}{V_{11}!V_{12}!\dots V_{ft}!} \prod_{f=1}^{N_F} \prod_{t=1}^{N_T} \Big[\sum_z p(z)p(f|z)p(t|z)\Big]^{V_{ft}} \\
&= \sum_{f=1}^{N_F} \sum_{t=1}^{N_T} V_{ft} \ln \Big[\sum_z p(z)p(f|z)p(t|z)\Big] + const.
\end{aligned}
$$

- Maximize w.r.t. the parameters (take derivative, set to zero, etc.).

Expectation Maximization Parameter Estimation I

- Formulate the log-likelihood of our model $\mathcal{L}(\boldsymbol{\Theta}|\,\mathbf{V})$.
- Form an auxiliary function that lower bounds the log-likelihood.

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\Theta}|\,\mathbf{X}) &= \ln p(\mathbf{X}\,|\boldsymbol{\Theta}) \\
&= \mathcal{F}(q, \boldsymbol{\Theta}) + \mathrm{KL}(q||p) \\
&\geq \mathcal{F}(q, \boldsymbol{\Theta})
\end{aligned}
$$

$$
\mathcal{F}(q, \boldsymbol{\Theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}\,|\boldsymbol{\Theta})}{q(\mathbf{Z})} \right\}
$$

$$
\begin{aligned}
\mathrm{KL}(q||p) &= \mathrm{KL}(q(\mathbf{Z})\ ||\ p(\mathbf{Z}\,|\,\mathbf{X}, \boldsymbol{\Theta})) \\
&= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}\,|\,\mathbf{X}, \boldsymbol{\Theta})}{q(\mathbf{Z})} \right\}
\end{aligned}
$$

Expectation Maximization Parameter Estimation II

- Iteratively maximize lower bound in two steps (coordinate ascent).
- E Step:

Compute the posterior $p(\mathbf{Z} \mid \mathbf{X}, \mathbf{\Theta})$

$$\begin{aligned} q^{n+1} &= \arg\max_q \mathcal{F}(q, \mathbf{\Theta}^n) \\ &= \arg\min_q \ \mathrm{KL}(q||p) \end{aligned}$$

Compute posterior

$$P(z|f,t)$$

- M Step:

$$\mathbf{\Theta}^{n+1} = \arg\max_{\mathbf{\Theta}} \mathcal{F}(q^{n+1}, \mathbf{\Theta})$$

Update model params

$$P(f|z)$$

$$P(t|z)$$

$$P(z)$$

- Converges to local optima.

PLCA Parameter Estimation via EM

- Initialize to random probabilities.
- Repeat until convergence.
  - E step

$$P(z|f,t) = \frac{P(z)P(f|z)P(t|z)}{\sum_z P(z)P(f|z)P(t|z)}$$

  - M step

$$P(z) = \frac{\sum_f \sum_t V_{ft} P(z|f,t)}{\sum_z \sum_f \sum_t V_{ft} P(z|f,t)}$$

$$P(f|z) = \frac{\sum_t V_{ft} P(z|f,t)}{\sum_f \sum_t V_{ft} P(z|f,t)}$$

$$P(t|z) = \frac{\sum_f V_{ft} P(z|f,t)}{\sum_f \sum_t V_{ft} P(z|f,t)}$$

Relationship between NMF and PLCA

- Equivalent up until init., normalization, reordering of updates.
- PLCA update equations in matrix notation vs. KL-NMF.

$$\begin{aligned}
\mathbf{Z} &\leftarrow \frac{\mathbf{V}}{\mathbf{W}\,\mathbf{H}} \\
\mathbf{W} &\leftarrow \mathbf{W} \odot \frac{\mathbf{Z}\,\mathbf{H}^{\mathrm{T}}}{\mathbf{1}\,\mathbf{H}^{\mathrm{T}}} \\
\mathbf{H} &\leftarrow \mathbf{H} \odot (\mathbf{W}^{\mathrm{T}}\,\mathbf{Z})
\end{aligned}$$

PLCA update equations

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{(\frac{\mathbf{V}}{\mathbf{W}\,\mathbf{H}})\mathbf{H}^{\mathrm{T}}}{\mathbf{1}\,\mathbf{H}^{\mathrm{T}}}$$

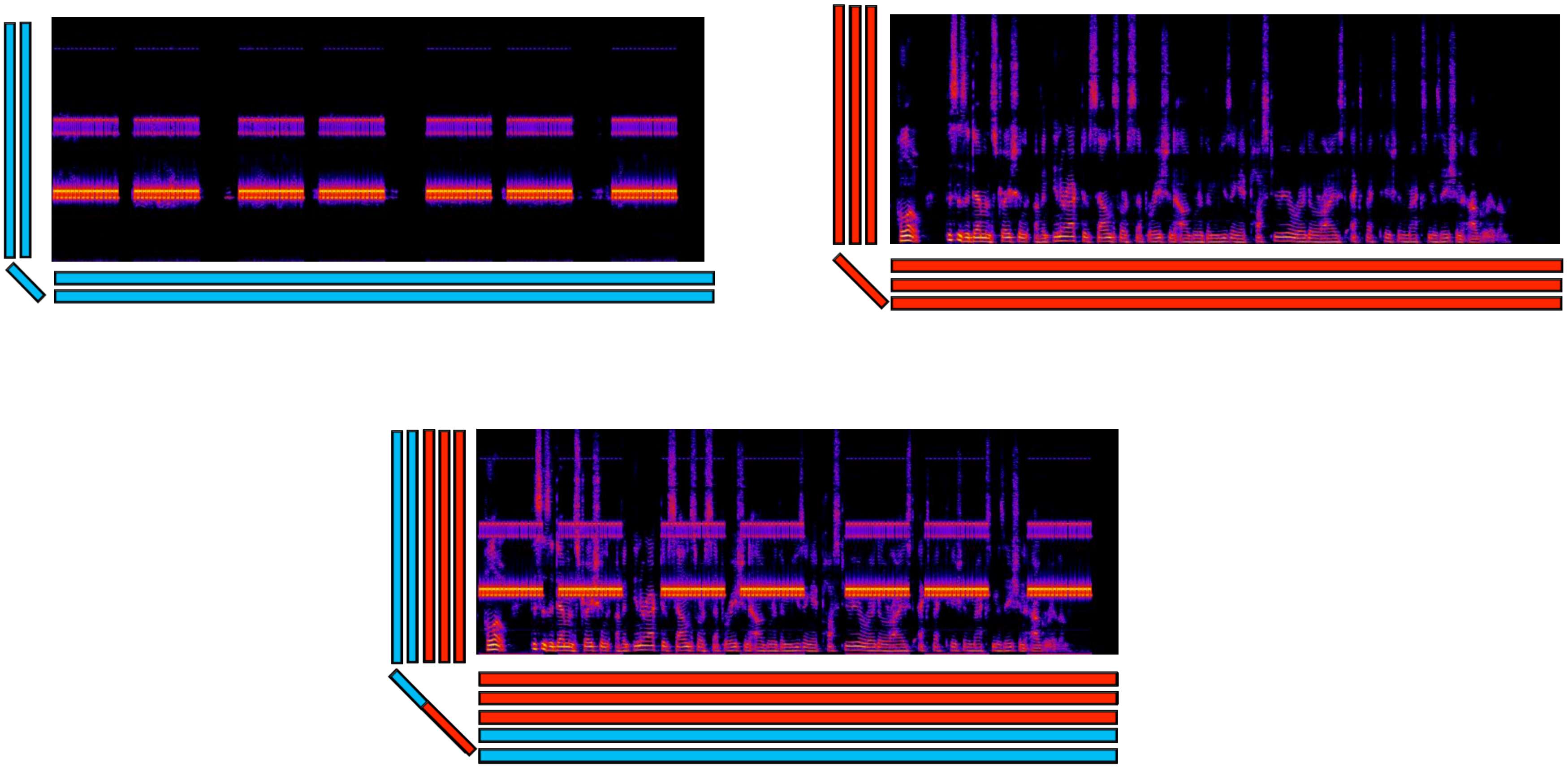
$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^{\mathrm{T}}(\frac{\mathbf{V}}{\mathbf{W}\,\mathbf{H}})}{\mathbf{W}^{\mathrm{T}}\,\mathbf{1}}$$

KL-NMF update equations

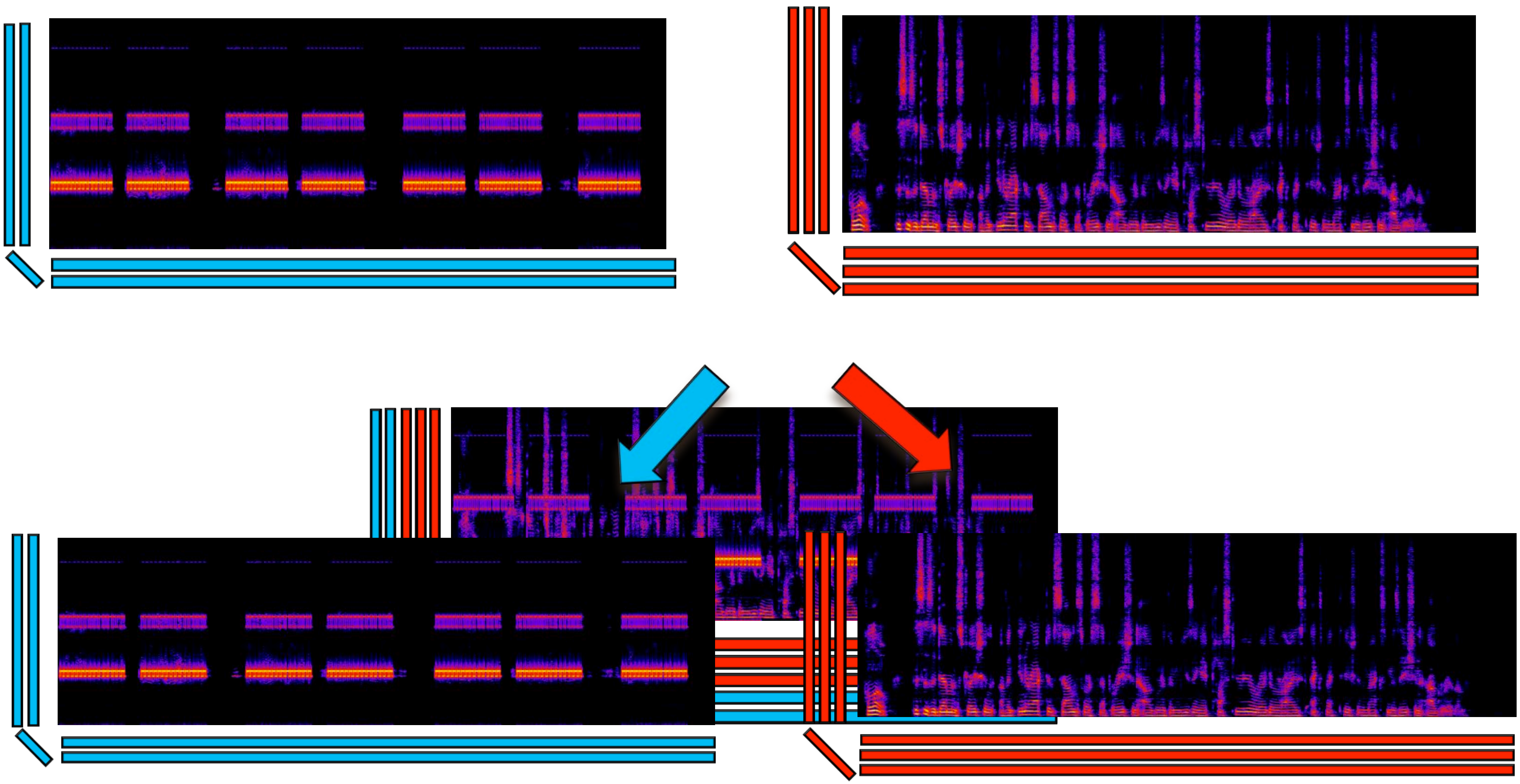
Modeling and Separating Mixtures

- Model each source within a mixture independently.
- Given a mixture, fix frequency distributions and estimate weights.
- Three general classes of techniques [Smaragdis 2007]:
  - Supervised separation
  - Semi-supervised separation
  - Unsupervised separation
- Use NMF/PLVM output to filter mixture.

Supervised Separation



Supervised Separation



Filtering I

- Convert source reconstruction into time-varying linear filter.

$$\mathbf{F}_s = \frac{\mathbf{W}_s\,\mathbf{H}_s}{\mathbf{W}\,\mathbf{H}} = \frac{\sum_{z \in Z_s} p(z)p(f|z)p(t|z)}{\sum_{z \in Z} p(z)p(f|z)p(t|z)}$$

- Filter mixture in time-frequency domain.

$$|\hat{\mathbf{X}}_s| = \mathbf{F}_s \odot |\mathbf{X}|$$

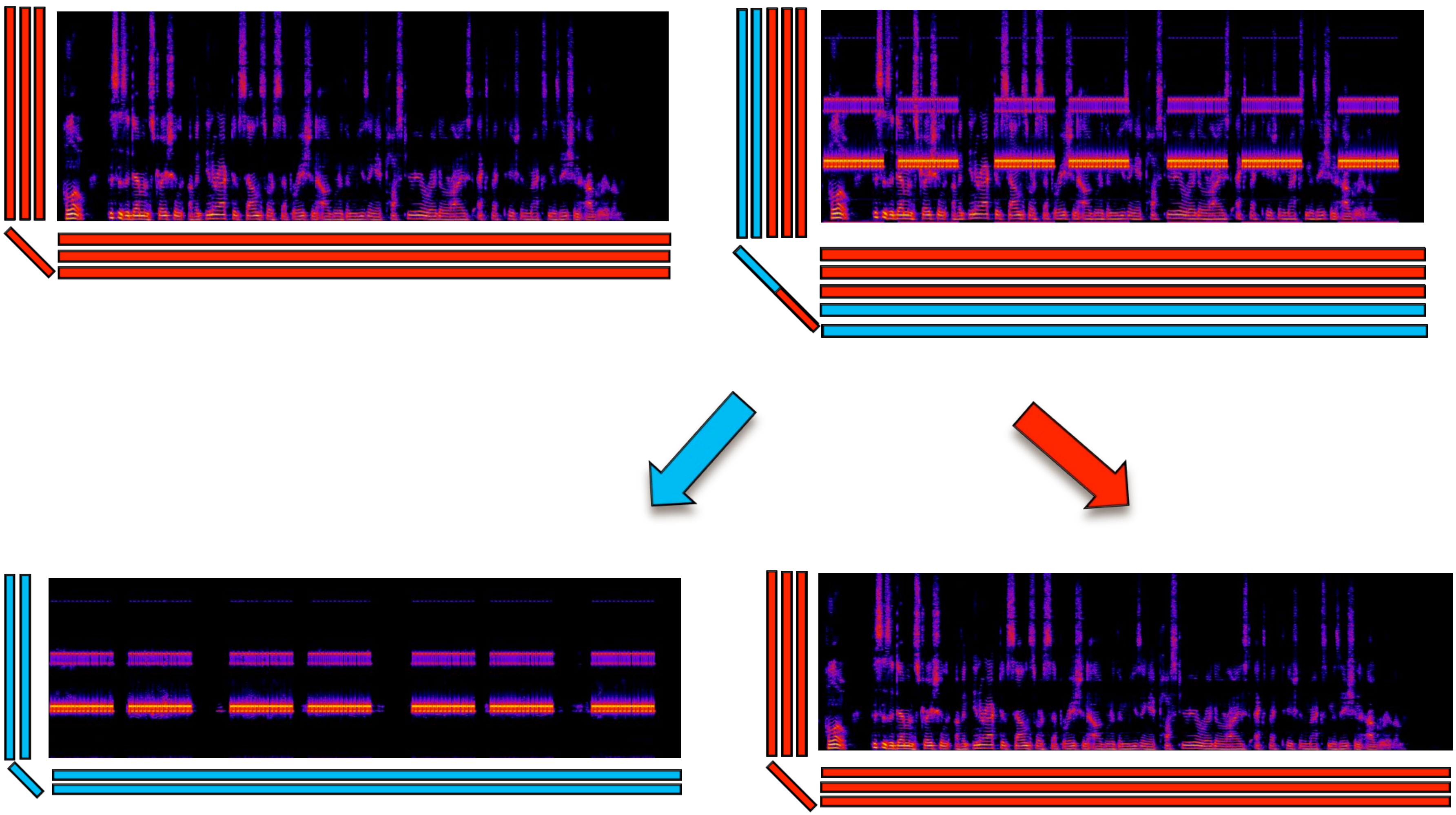
- Inverse STFT with mixture phase $\angle \mathbf{X}$.

- Overlap-add (OLA) processing to filter mixture [Smith 2011].

Filtering II

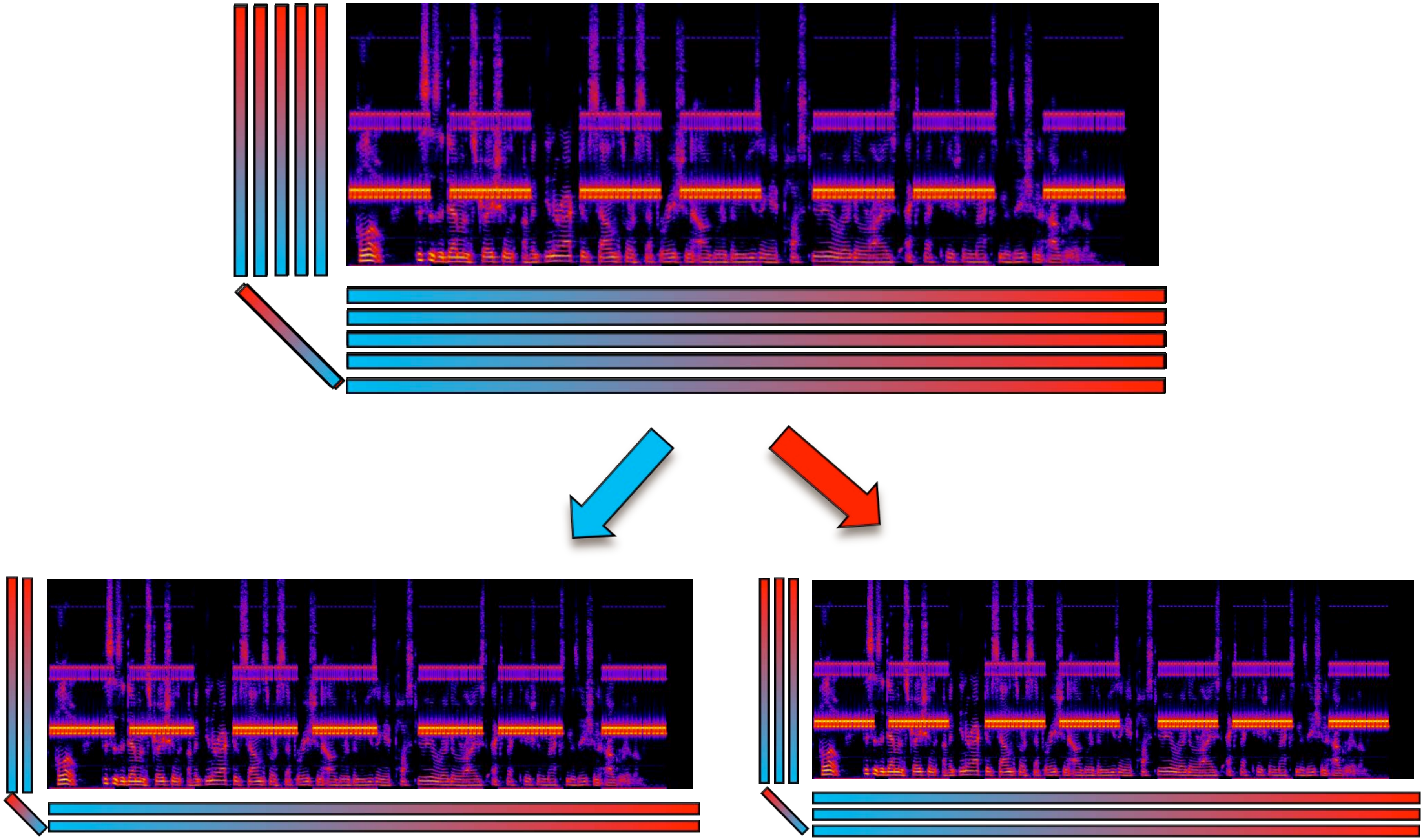
- Sharp discontinuities in the filter frequency response.
- Time-aliasing and other unwanted audible artifacts.
- Convert filters to a alias-free form via optimal filter design [Smith 2011].
- Incorporate STFT consistency constraints [Le Roux 2013].

Semi-Supervised Separation



Unsupervised Separation

- Without training data....difficult!



General Problems

- Overall a **very difficult**, ill-posed problem.
- **Requires** isolated **training** data.
- **No** auditory or **perceptual models** of hearing.
- **Cannot correct** for poor results (even if obvious).

Overview

- Motivation
- Background
- Approach
- Algorithm
- Evaluation
- Conclusion

Approach

- Improve upon NMF/PLVM separation.
- Informed source separation.
  - Spatial information [Ozerov & Fevotte 2009]
  - Score information [Woodruff et al. ‘06, Ganesman et al. ‘10, Duan & Pardo ’11]
  - Temporal dynamics [Mysore et al. 2010]
  - User-guidance

User-Guided Source Separation

- Examples:
  - Singing/humming [Smaragdis 2009, Smaragdis and Mysore 2009]
  - Binary time region annotations [Ozerov et al. 2011, 2012]
  - Fundamental frequency annotations [Durrieu and Thiran 2012]
  - Binary time-frequency region annotations [Lefèvre et al. 2012]

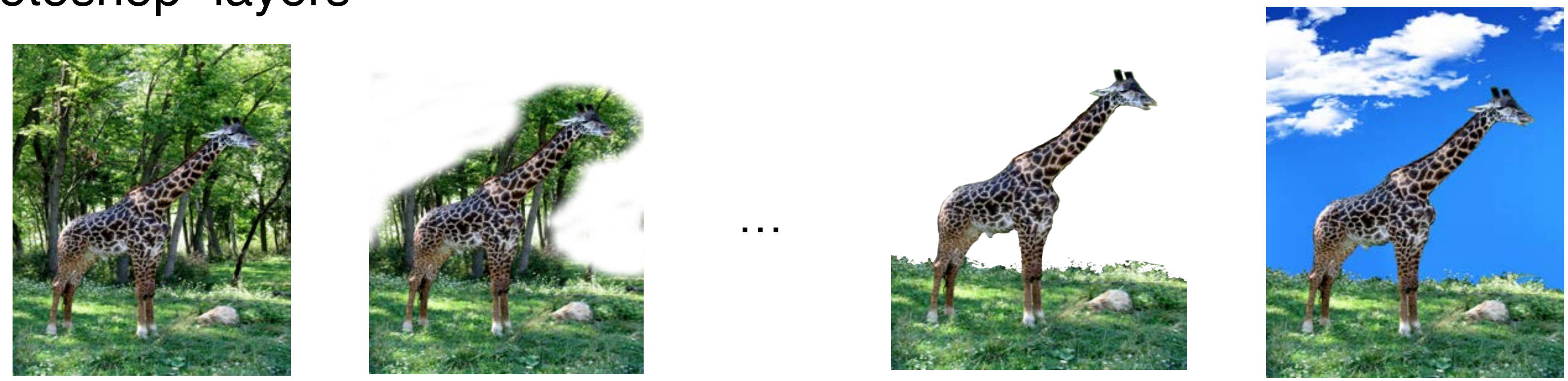
- Typically no user-feedback, refinement, and/or iteration.

Interactive Source Separation

- Extension of **user-guided** separation.
- **Subtle**, but significant difference.
- **Two-way** communication between user and algorithm.
- Emphasize on **user-feedback**, refinement, and iteration.
- **Re-compute** each interaction.
- Requires **speed.**

Interaction Analogy

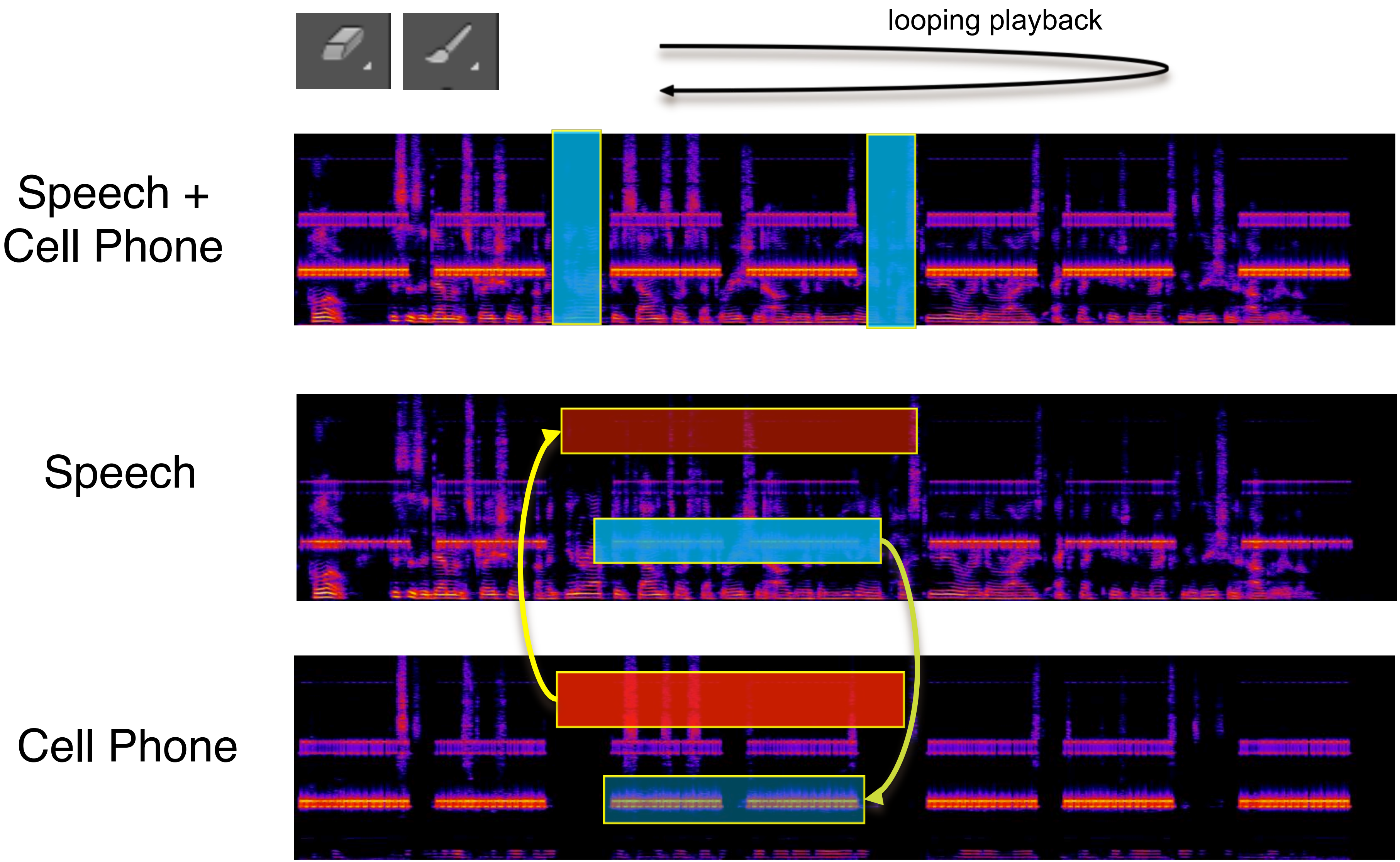
- Photoshop "layers"



- 3D Sculpting

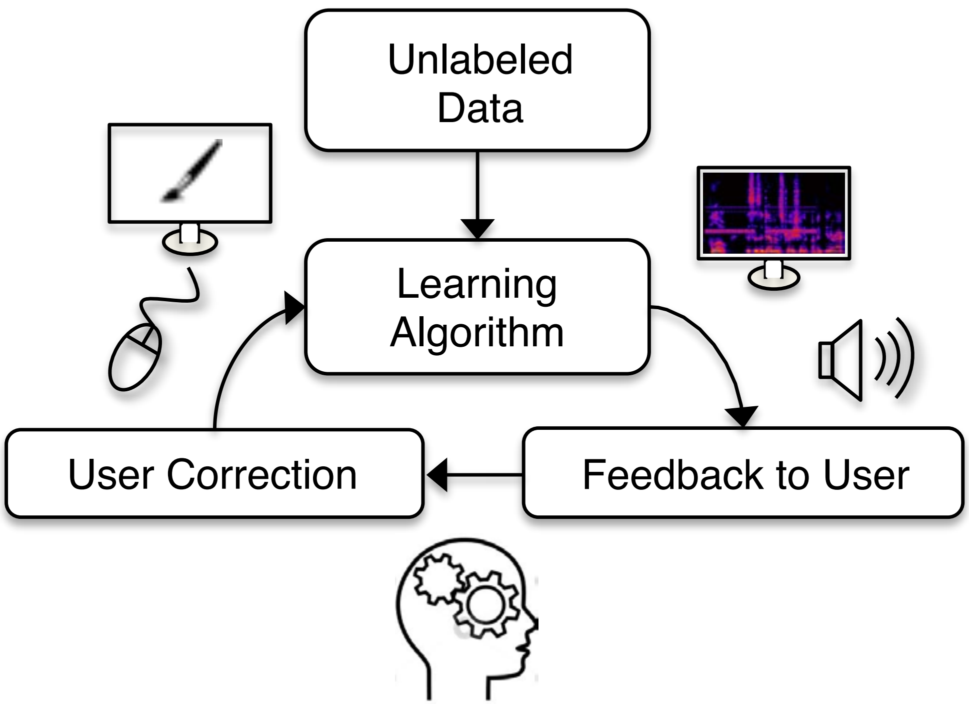
- User-feedback is key!

A Layers-Sculpting-Like Interaction for Audio



Interactive Machine Learning

- Machine learning (**ML**) and human-computer interaction (**HCI**).
- **User-perspective** of ML (train and test).
- We can elicit **more** information **than** a **class label!**
- Found great success across **several domains** including:
  - [Fails & Olsen 2003]
  - [Fogarty et al. 2008]
  - [Cohn et al. 2008]
  - [Settles 2011]
  - [Fiebrink 2011]

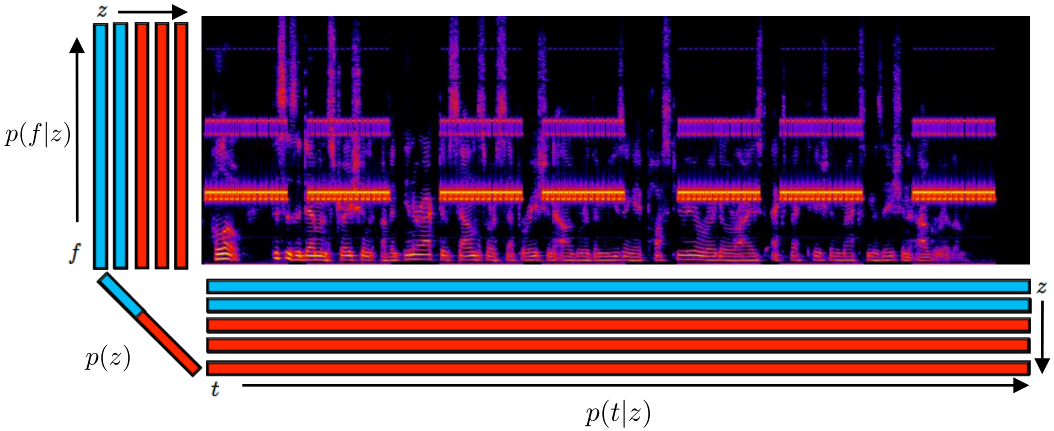


Overview

- Motivation
- Background
- Approach
- Algorithm
- Evaluation
- Conclusion

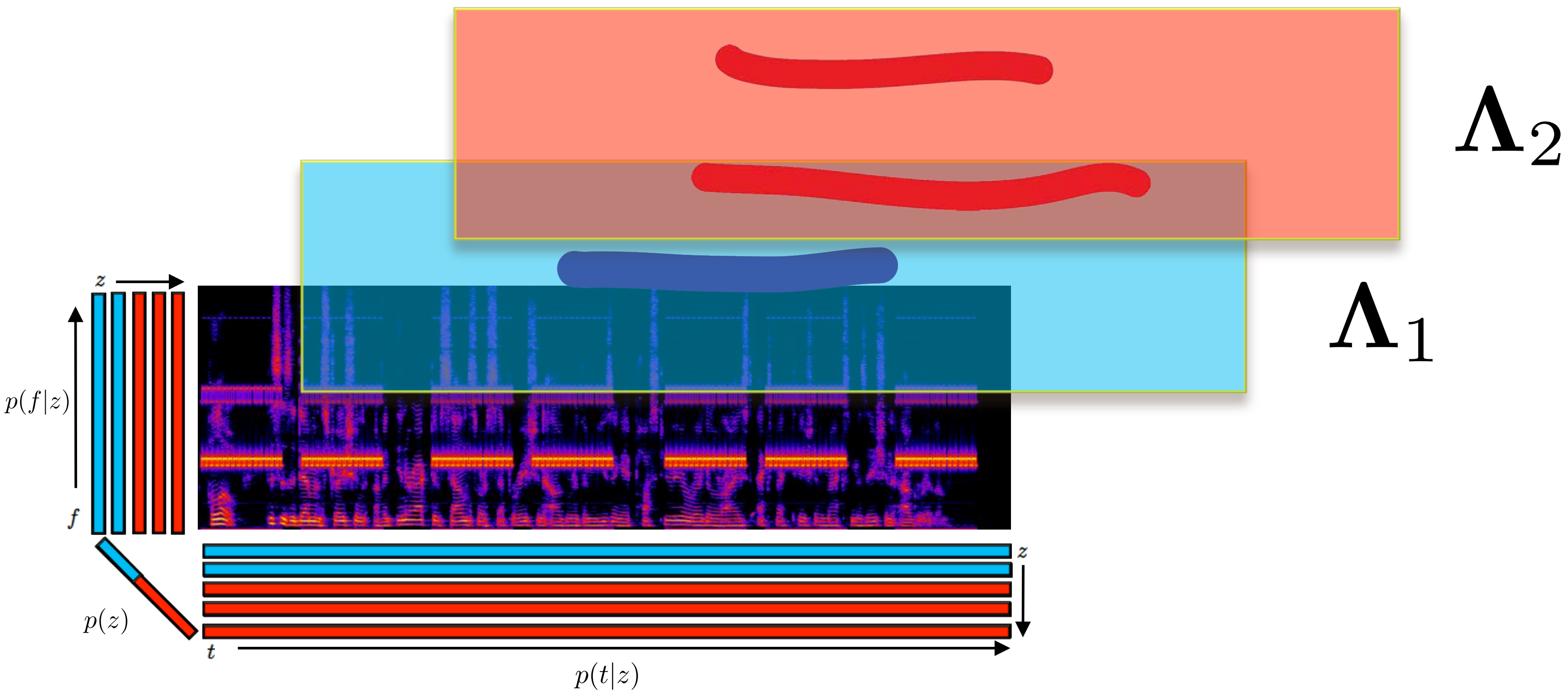
Probabilistic Model

$$\mathbf{V} \approx P(f,t) = \sum_z P(z)P(f|z)P(t|z)$$



Probabilistic Model w/Painting Constraints

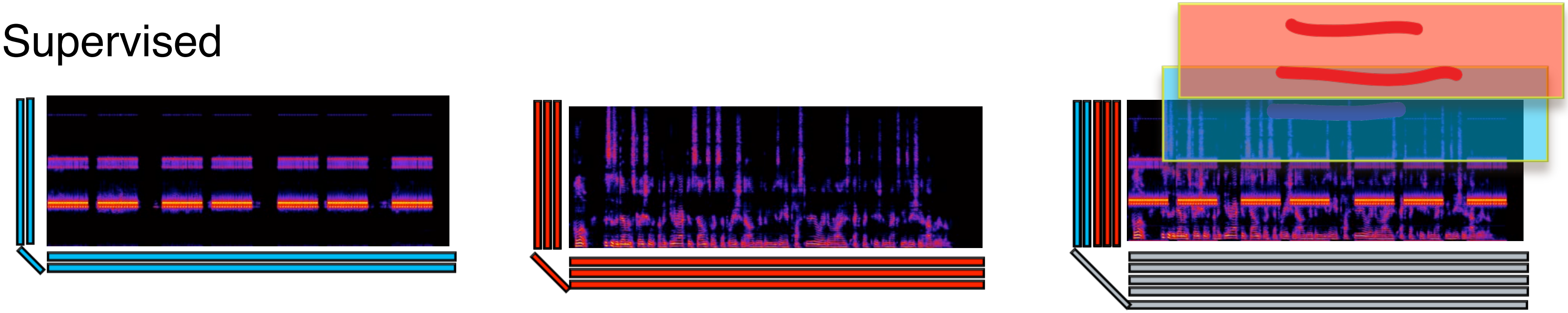
$$\mathbf{V} \approx P(f,t) = \sum_z \tilde{P}(z)\tilde{P}(f|z)\tilde{P}(t|z)$$



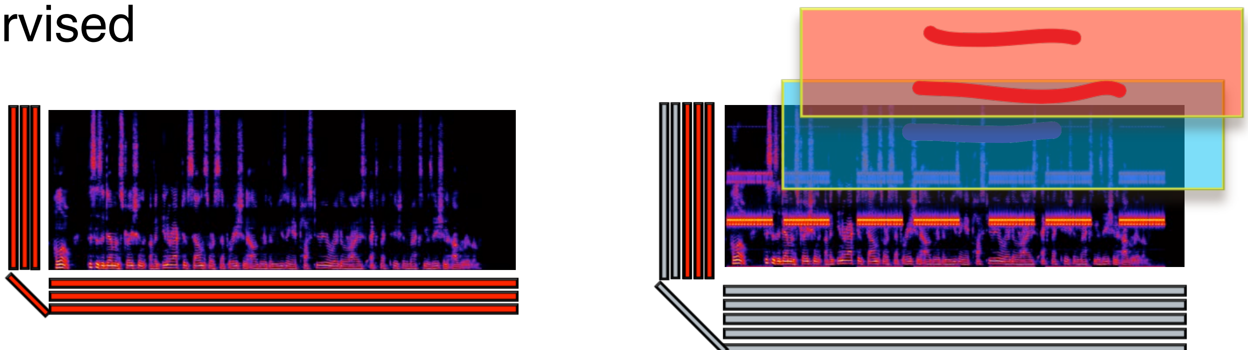
- Color → source
- Opacity → strength

Supervised, Semi-Supervised, & Unsupervised Learning

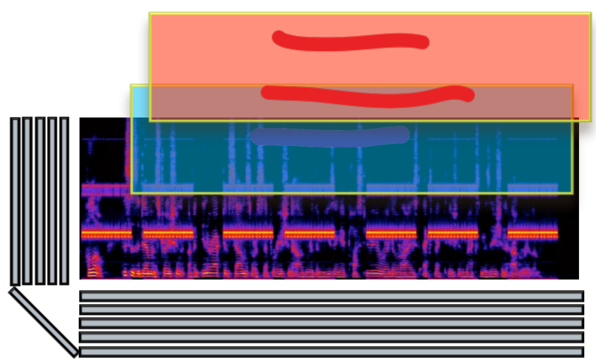
- Supervised



- Semi-Supervised



- Unsupervised



Constraints

- Constraints typical encoded as: $P(f|z)$ $P(t|z)$ $P(z)$
  - Prior probabilities on model parameters (e.g. Dirichlet priors)
  - Direct observations
- Does not (reasonably) allow time-frequency constraints
- Posterior regularization [Graça et al., 2007, Ganchev et al., 2010]
  - Complementary method that allows time-frequency constraints $P(z|f,t)$
  - Iterative optimization procedure for each E step
  - Well suited for our problem

Expectation Maximization

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}|\,\mathbf{X}) &= \ln p(\mathbf{X}\,|\boldsymbol{\Theta}) \\ &= \mathcal{F}(q,\boldsymbol{\Theta}) + \mathrm{KL}(q||p) \\ &\geq \mathcal{F}(q,\boldsymbol{\Theta})\end{aligned}$$

E Step:

$$\begin{aligned}q^{n+1} &= \arg\max_{q} \mathcal{F}(q,\boldsymbol{\Theta}^{n}) \\ &= \arg\min_{q}\ \mathrm{KL}(q||p)\end{aligned}$$

M Step:

$$\boldsymbol{\Theta}^{n+1} = \arg\max_{\boldsymbol{\Theta}} \mathcal{F}(q^{n+1},\boldsymbol{\Theta})$$

Expectation Maximization w/Posterior Constraints I

$$
\begin{aligned}
\mathcal{L}(\mathbf{\Theta}|\,\mathbf{X}) &= \ln p(\mathbf{X}\,|\mathbf{\Theta}) \\
&= \mathcal{F}(q, \mathbf{\Theta}) + \mathrm{KL}(q||p) \\
&\geq \mathcal{F}(q, \mathbf{\Theta})
\end{aligned}
$$

E Step:

$$
\begin{aligned}
q^{n+1} &= \arg\max_{q \in \mathcal{Q}} \mathcal{F}(q, \mathbf{\Theta}^n) \\
&= \arg\min_{q \in \mathcal{Q}} \ \mathrm{KL}(q||p)
\end{aligned}
$$

M Step:

$$
\mathbf{\Theta}^{n+1} = \arg\max_{\mathbf{\Theta}} \mathcal{F}(q^{n+1}, \mathbf{\Theta})
$$

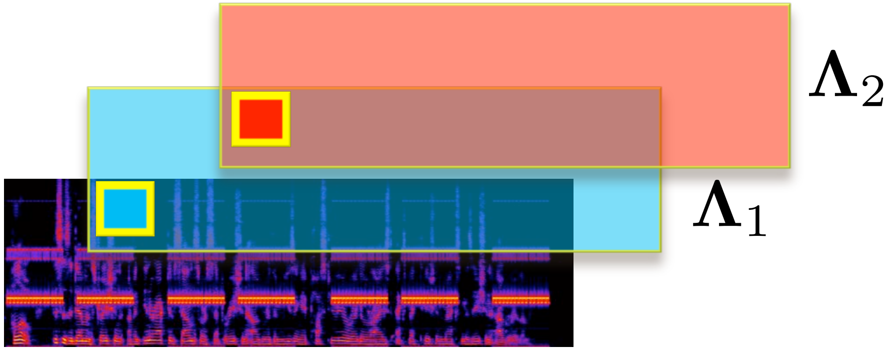
Linear Grouping Expectation Constraints

$$\arg\min_{q \in \mathcal{Q}} \ \mathrm{KL}(\ q(z|f,t) \ || \ p(z|f,t)\ )$$

- For each time-frequency point, solve

$$\arg\min_{\mathbf{q}} \quad -\mathbf{q}^{\mathrm{T}} \ln \mathbf{p} + \mathbf{q}^{\mathrm{T}} \ln \mathbf{q}$$

$$\text{subject to} \quad \mathbf{q}^{\mathrm{T}} \mathbf{1} = 1, \ \mathbf{q} \geq 0$$



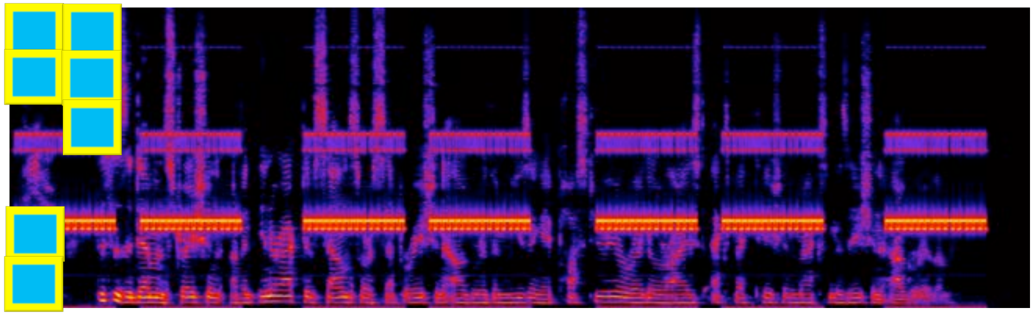
$$\lambda^{\mathrm{T}} = [\mathbf{\Lambda}_{1_{ft}} \ \mathbf{\Lambda}_{1_{ft}} \ \mathbf{\Lambda}_{1_{ft}} \ \ldots \ \mathbf{\Lambda}_{2_{ft}} \ \mathbf{\Lambda}_{2_{ft}} \ \mathbf{\Lambda}_{2_{ft}}]$$

Big Picture

E Step: Compute posterior $p(z|f,t)$

$$\forall f,t \quad \arg\min_{\mathbf{q}} \; -\mathbf{q}^{\mathrm{T}}\ln\mathbf{p} + \mathbf{q}^{\mathrm{T}}\ln\mathbf{q} + \mathbf{q}^{\mathrm{T}}\boldsymbol{\lambda}$$

$$\text{subject to} \quad \mathbf{q}^{\mathrm{T}}\mathbf{1} = 1, \; \mathbf{q} \geq 0$$



M Step:

$$\boldsymbol{\Theta}^{n+1} = \arg\max_{\boldsymbol{\Theta}} \mathcal{F}(q^{n+1}, \boldsymbol{\Theta})$$

Fast, Closed-Form Updates

- With simple penalty, both E and M steps are in closed form.
- Reduces to simple, fast multiplicative updates vs. NMF.
- Roughly the same computational cost as without constraints.

expectation step
for all z, f, t **do**

$$Q(z|f,t) \leftarrow \frac{P(z)P(f|z)P(t|z)}{\sum_{z'} P(z')P(f|z')P(t|z')}$$

end for

expectation step
for all z, f, t **do**

$$Q(z|f,t) \leftarrow \frac{P(z)P(f|z)P(t|z)\tilde{\Lambda}_{(f,t,z)}}{\sum_{z'} P(z')P(f|z')P(t|z')\tilde{\Lambda}_{(f,t,z')}}$$

end for

- In general, constrained inference would require numerical opt.

Overview

- Motivation
- Background
- Approach
- Algorithm
- Geometric Interpretation
- Evaluation
- Conclusion

Evaluation

- Initial results
- Signal Separation Evaluation Campaign (SiSEC) 2013
- User tests

Evaluation Metrics

- BSS-EVAL metrics [Vincent et al., 2006]
  - (SDR) Signal-to-Distortion Ratio → Overall separation quality
  - (SIR) Signal-to-Interference Ratio → Amount of reduction from unwanted source
  - (SAR) Signal-to-Artifact Ratio → Amount of artifacts introduced by algorithm
- Baselines
  - Ideal, oracle algorithm (soft mask)
  - No user-annotation
  - Past high-performing algorithms

Initial Results

- Supervised, semi-supervised, & unsupervised separation comparison

Example	Ideal	Supervised	Semi-Supervised	Unsupervised
Cell	30.7	29.2 / 27.6	28.4 / 06.5	28.8 / -0.6
Drum	14.8	09.7 / 08.5	07.7 / 03.9	10.0 / 00.2
Cough	15.8	14.0 / 12.5	12.0 / 10.5	13.8 / -2.1
Piano	26.1	26.0 / 21.6	14.9 / 08.4	23.1 / 01.1
Siren	27.8	23.8 / 18.9	21.0 / 19.9	24.2 / -4.2

Table 1: SDR (dB) with and without interaction vs. ideal results.

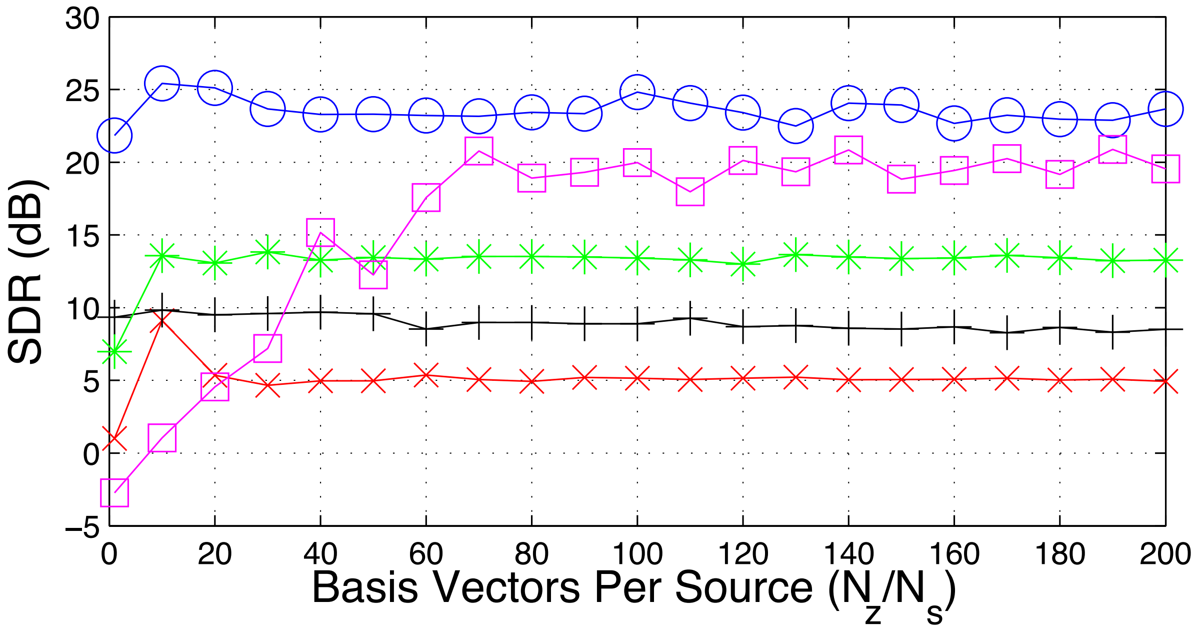
- Outperformed prior SiSEC 2011 vocals state-of-the-art [Durrieu 2012]

Example	Ideal	Baseline	Lefévre	Durrieu	Proposed
S1	13.2	-0.8	7.0	9.0	9.2
S2	13.4	0.2	5.0	7.8	11.1
S3	11.5	-0.2	3.8	6.4	7.8
S4	12.5	1.4	5.0	5.9	7.9

Table 2: SDR (dB) results for the four SiSEC rock/pop songs.

Model Selection

- How many basis vectors?
- Set it to a large number (50)



Signal Separation Evaluation Campaign 2013

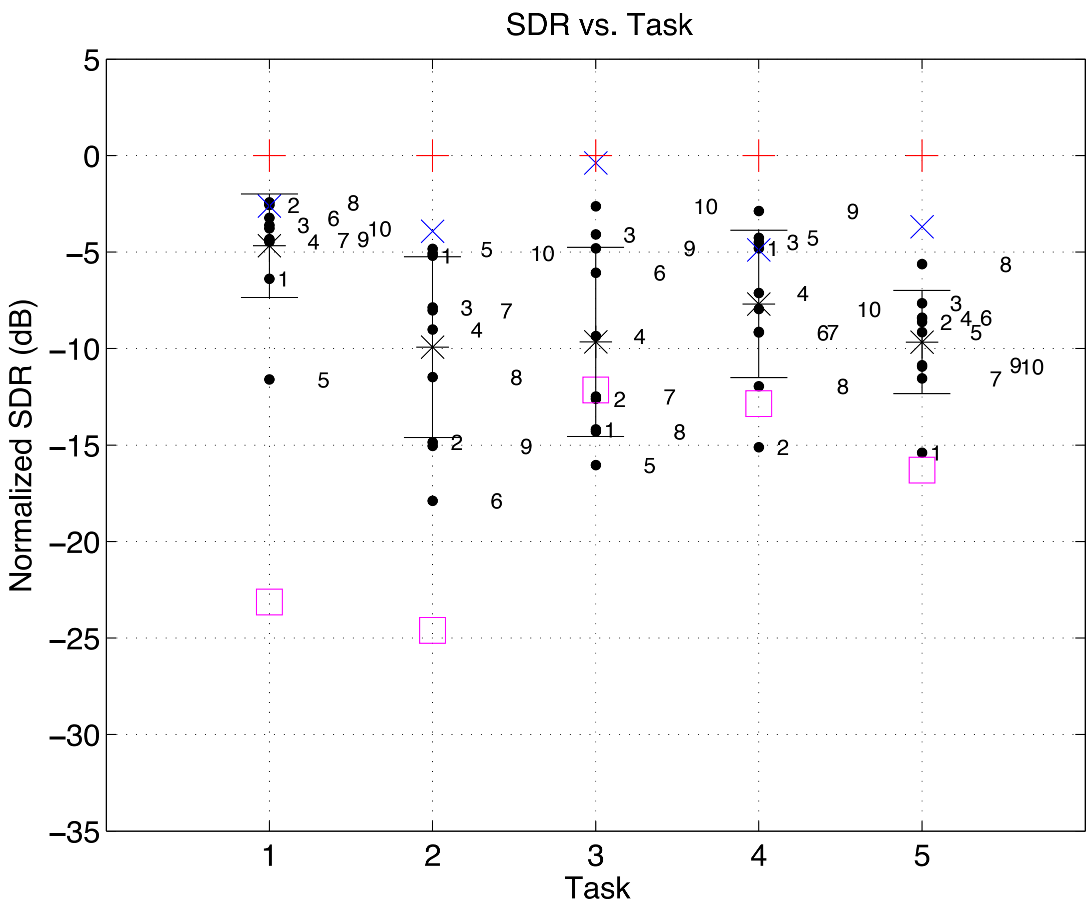
- Task 1: Professionally produced music recordings
  - 15 submissions
  - Variety of stereo music recordings
  - Vocals, drums, bass, guitar, piano, other
- State-of-the-art performance
  - Best overall SDR 16/24 times. Next closest 4/24 times.
  - Best vocal SDR 6/7 times. Outperformed algorithms specifically design for vocals.
  - Best drum SDR 2/5 times.
  - Best bass SDR 2/5 times.
  - Best piano SDR 1/2 times.
  - Best guitar SDR 1/1 times.
  - Best other SDR 4/4 times.
  - Recordings are stereo-channel. Our algorithm is monophonic applied to stereo.

Novice User Evaluation

- How well a novice can perform separation?
- 10 inexperienced users
- 1 hour long study
  - Introduction and explanation
  - 5 separation tasks, 10 minutes each, increasing difficulty
  - Exit survey
- Measure separation quality per example per user
- Compare against expert user
- Tasks:
  - Cell phone + speech
  - Siren + speech
  - Drums + bass
  - Orchestra + cough
  - Vocals + guitar

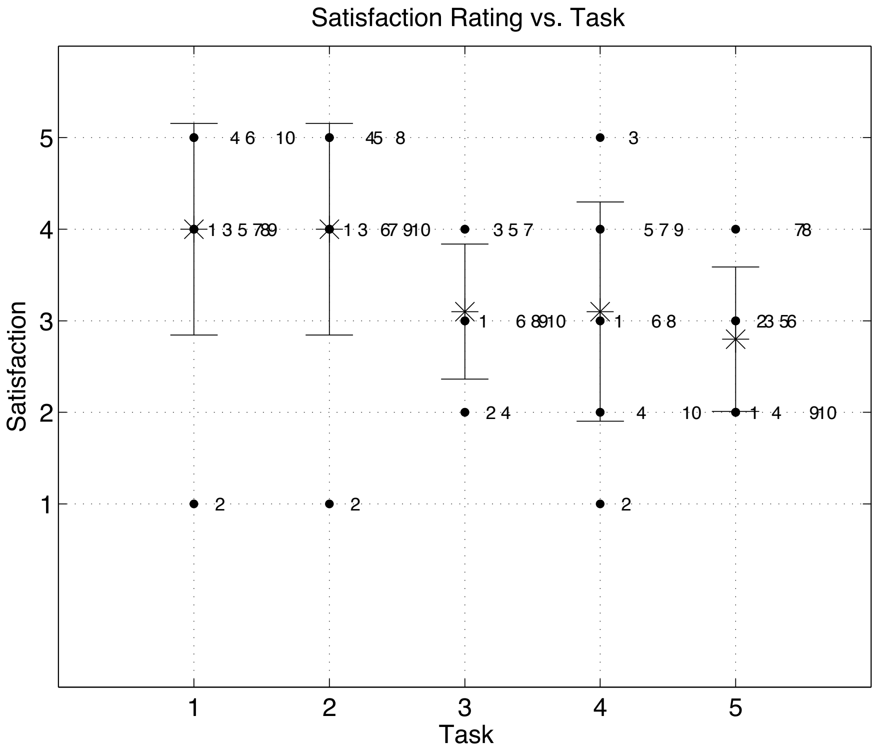
Novice User Results I

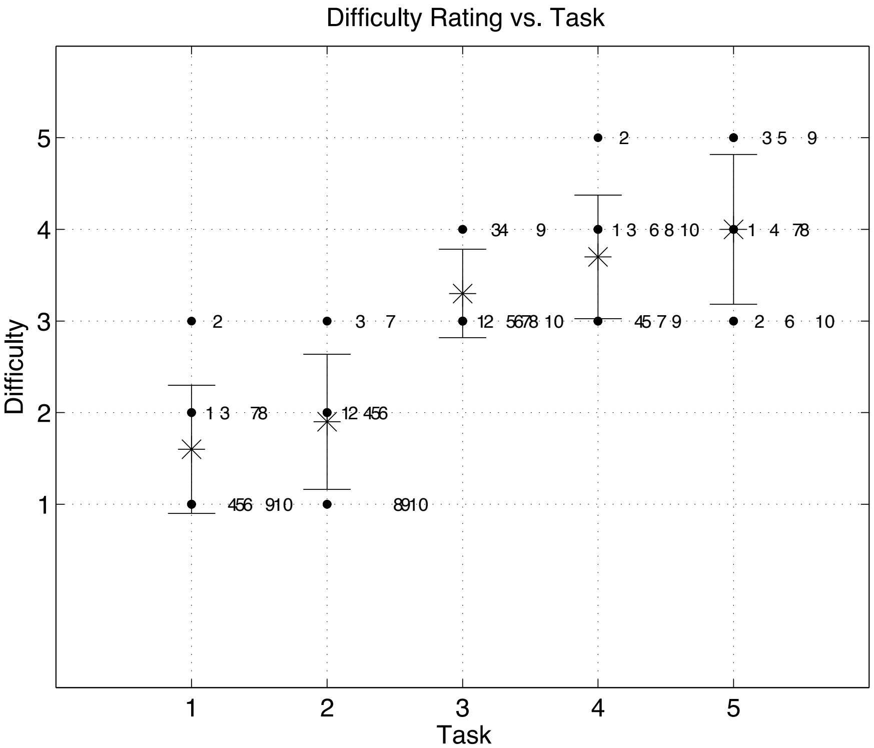
- In some cases, novices outperformed the expert!
- Most cases, the expert was best.



Novice User Results II

- The more difficult the task, the more unsatisfying





Overview

- Motivation
- Background
- Approach
- Algorithm
- Evaluation
- Conclusion

Interactive Approach: Benefits

- **Reduces** manual **effort.**
- **Improves automatic** approaches (correct for poor results).
- **No training** data needed!
- Indirectly incorporate a **perceptual model.**

Interactive Approach: Problems

- Requires a **user + learning** curve!
- **No guarantee** of high-quality results.
- Overall computation time can be **slow**.
- **ALL** machine learning algorithms **require a user.**
  - **Who**: engineer, scientist, end-user, audio engineer
  - **What**: class labeled data, feature labels, other
  - **Where**: research laboratory, recording studio, other
  - **When**: train and testing occur separately or simultaneously
  - **Why**: applications can be different or the same

Overall Contributions

- **Interactive** source separation approach.
- NMF/PLVM + painting via **posterior regularization**.
- With or **without training** data (unsup., semi-sup., or sup.).
- Relatively **insensitive** to model selection.
- Open-source, freely available, cross-platform **software**.
- **State-of-the-art** separation and user studies.

General and high performing separation method.

Publications

1. N. J. Bryan, G. J. Mysore. **"Interactive User-Feedback for Sound Source Separation."** *ACM Int. Conf. on Intelligent User-Interfaces, Workshop on Interactive Machine Learning*, 2013.
2. N. J. Bryan, G. J. Mysore. "**An Efficient Posterior Regularized Latent Variable Model for Interactive Sound Source Separation.**" *Int. Conf. on Machine Learning*, 2013.
3. N. J. Bryan, G. J. Mysore. "**Interactive Refinement of Supervised and Semi-Supervised Sound Source.**" *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2013.
4. N. J. Bryan, G. J. Mysore. "**Signal Separation Evaluation Campaign (SiSEC) Submission.**" http://sisec.wiki.irisa.fr, 2013.
5. N. J. Bryan, G. J. Mysore, G. Wang. **"Source Separation of Polyphonic Music With Interactive User-feedback on a Piano Roll Display."** *Int. Society of Music Inf. Retrieval*, 2013.
6. (submitted) N. J. Bryan, G. J. Mysore, G. Wang. **"ISSE: An Interactive Source Separation Editor."** *Conf. on Human Factors in Computing Systems*, 2014.

Software + Code

- http://isse.sourceforge.net
- Application + Code
  - OSX, Windows, Linux
  - C++ and Matlab code
  - User forum, wiki, user manual, audio and video demonstrations
- Application Web Statistics
  - 2000+ downloads (60+ countries, 36% Japan, 28% USA)
  - 3600+ Soundcloud listens (13+ hours of audio listened)
  - 4000+ Youtube views (10+ days of video watched)
  - 8000+ webpage visits (14.5+ days of viewing)

Thank you!

Work advised by:
Gautham J. Mysore &
Prof. Ge Wang

References I

- [Lee & Seung, 1999] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization.” Nature, 1999.
- [Lee & Seung, 2001] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization.” NIPS, 2001.
- [Smaragdis & Brown 2003] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription.” WASPAA, 2003.
- [Fails & Olsen 2003] J. A. Fails and D. R. Olsen, “Interactive machine learning.” IUI, 2003.
- [Raj & Smaragdis 2005] B. Raj and P. Smaragdis, “Latent variable decomposition of spectrograms for single channel speaker separation.” WASPAA, 2005.
- [Smaragdis et al., 2006] P. Smaragdis, B. Raj, and M. Shashanka, “A probabilistic latent variable model for acoustic modeling.” NIPS Workshop on Acoustic Processing, 2006.
- [Woodruff et al. 2006] J. F. Woodruff, B. Pardo, and R. B. Dannenberg, “Remixing stereo music with score-informed source separation.” ISMIR, 2006.

References II

- [Vincent et al., 2006] E. Vincent, R. Gribonal, C. Fevotte, “Performance measurement in blind audio source separation.” IEEE TASLP, 2006.
- [Graça et al., 2007] J. Graça, K. Ganchev, B. Taskar, “Expectation maximization and posterior constraints.” NIPS, 2007.
- [Smaragdis 2007] P. Smaragdis, B. Raj, M. Shashanka, “Supervised and semi-supervised separation of sounds from single-channel mixtures.” ICASS, 2007.
- [Fogarty 2008] J. Fogarty, D. Tan, A. Kapoor, S. Winder, “Cueflik: interactive concept learning in image search.” CHI, 2008.
- [Cohn et al. 2008] D. Cohn, R. Caruana, A. McCallum, “Semi-supervised clustering with user feedback.” Constrained Clustering: Advances in Algorithms, Theory, and Applications, 2008.
- [Smaragdis 2009] P. Smaragdis, “User guided audio selection from complex sound mixtures.” UIST, 2009.
- [Smaragdis and Mysore 2009] P. Smaragdis and G. J. Mysore, “Separation by humming: User guided sound extraction from monophonic mixtures.” WASPAA, 2009.

References III

- [Ozerov & Fevotte 2009] A. Ozerov and C. Fevotte, “Multichannel nonnegative matrix factorization in convolutive mixtures.” ICASSP, 2009.
- [Ganchev et al., 2010] K. Ganchev, J. Graça, J. Gillenwater, B. Taskar, “Posterior regularization for structured latent variable models.” JMLR, 2010.
- [Ganesman et al. 2010] J. Ganseman, G. J. Mysore, J. S. Abel, P. Scheunders, “Source separation by source synthesis.” ICMC, 2010.
- [Mysore et al. 2010] G. J. Mysore, P. Smaragdis, B. Raj, “Non-negative hidden Markov modeling of audio with application to source separation” LVA/ICA, 2010.
- [Settles 2011] “Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances.” EMNLP, 2011.
- [Fiebrink 2011] “Real-time human interaction with supervised learning algorithms for music composition and performance.” PhD Dissertation, Princeton University, 2011.
- [Smith 2011] J. Smith, Spectral Audio Signal Processing. W3K Pub., 2011.

References IV

- [Duan & Pardo 2011] Z. Duan, B. Pardo, “Soundprism: An online system for score-informed source separation of music audio.” IEEE Journal on Selected Topics in Signal Processing, 2011.
- [Ozerov et al. 2011] A. Ozerov, C. Fevotte, R. Blouet, and J.-L. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation.” ICASSP, 2011.
- [Durrieu & Thiran, 2012] J.-L. Durrieu J.-P. Thiran, “Musical audio source separation based on user-selected f0 track.” LVA/ICA, 2012.
- [Lefèvre et al. 2012] A. Lefèvre, F. Bach, and C. Fevotte, “Semi-supervised NMF with time-frequency annotations for single-channel source separation.” ISMIR, 2012.
- [Ozerov et al. 2012] A. Ozerov, N. Q. Duong, L. Chevallier, “Weighted nonnegative tensor factorization with application to user-guided audio source separation.” Tech Report, 2012.
- [Le Roux 2013] J. Le Roux, E. Vincent, “Consistent Weiner Filtering for Audio Source Separation.” IEEE Signal Processing Letters, 2013.

Extra

Alternative (Common) View of EM

- View I – expected log-likelihood, then maximize
  - E step - calculate the expected value of the log-likelihood function

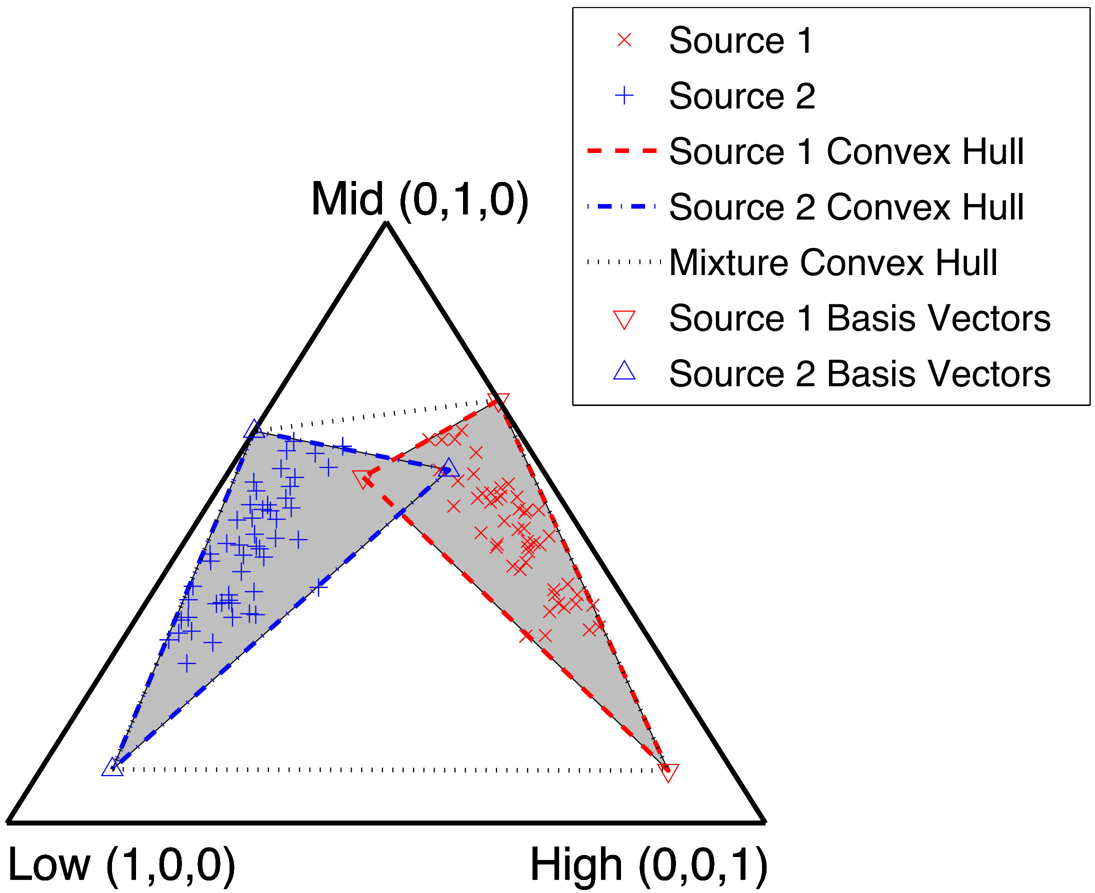
$$\mathcal{Q}(\Theta|\Theta^t) = \mathrm{E}_{\mathbf{Z}\,|\,\mathbf{X},\Theta^t}\left[\mathcal{L}(\Theta;\mathbf{X},\mathbf{Z})\right]$$

  - M step – find the parameters that maximize the expected log-likelihood

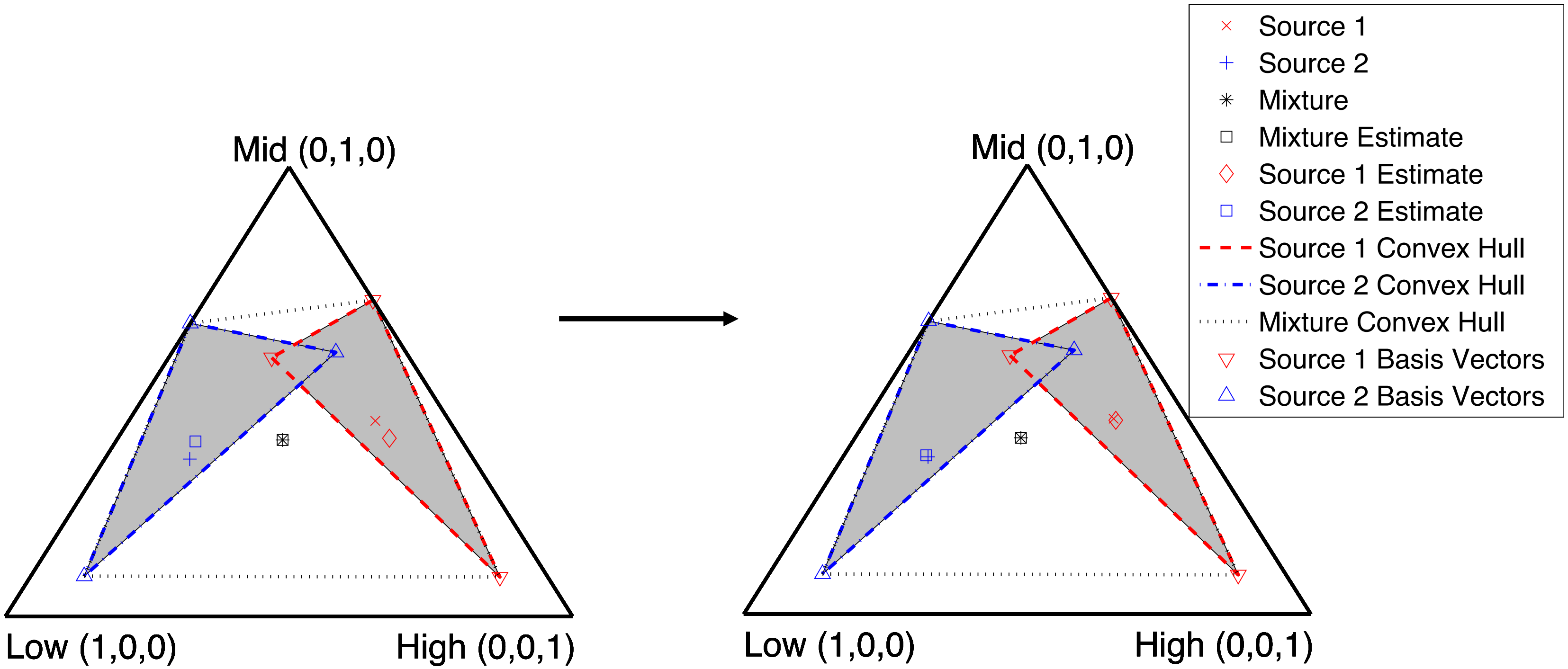
$$\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\Theta|\Theta^t)$$

  - Equivalent, but less general viewpoint

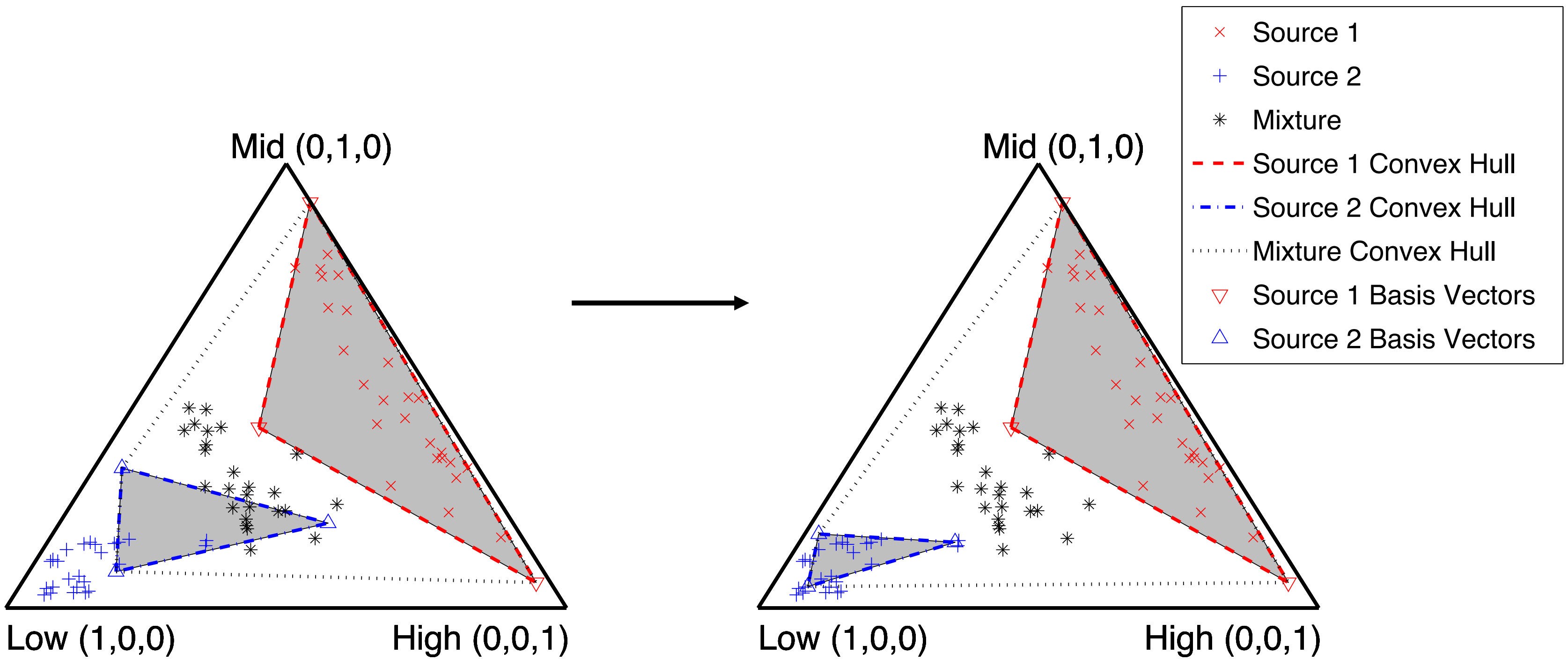
Geometric Interpretation



Simplex w/Supervised Separation



Simplex w/Semi-Supervised Separation



Simplex w/Unsupervised Separation