An Efficient Posterior Regularized Latent Variable Model for Interactive Sound Source Separation

Nicholas J. Bryan, Stanford University

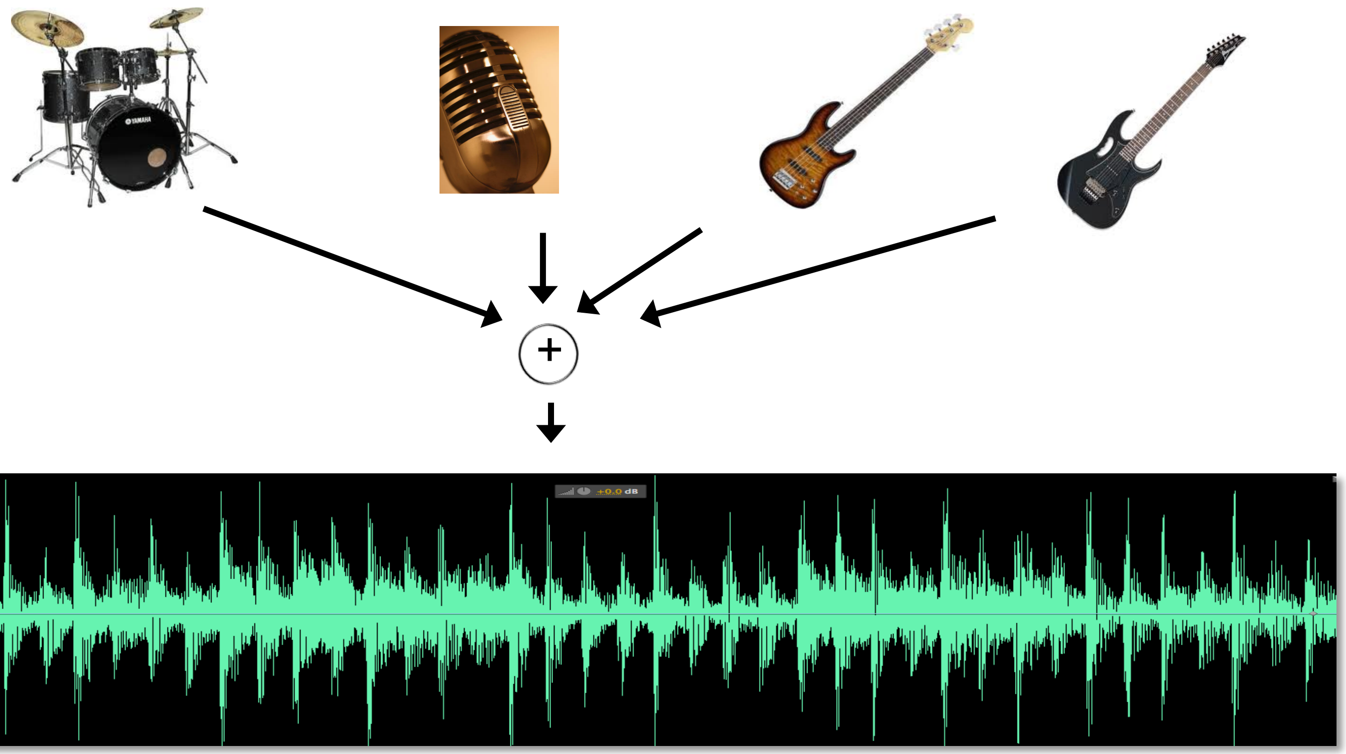
Gautham J. Mysore, Adobe Research

ICML 2013

Sound Check

Motivation I

- Real world sounds are mixtures of many individual sounds



Current State-of-the-Art

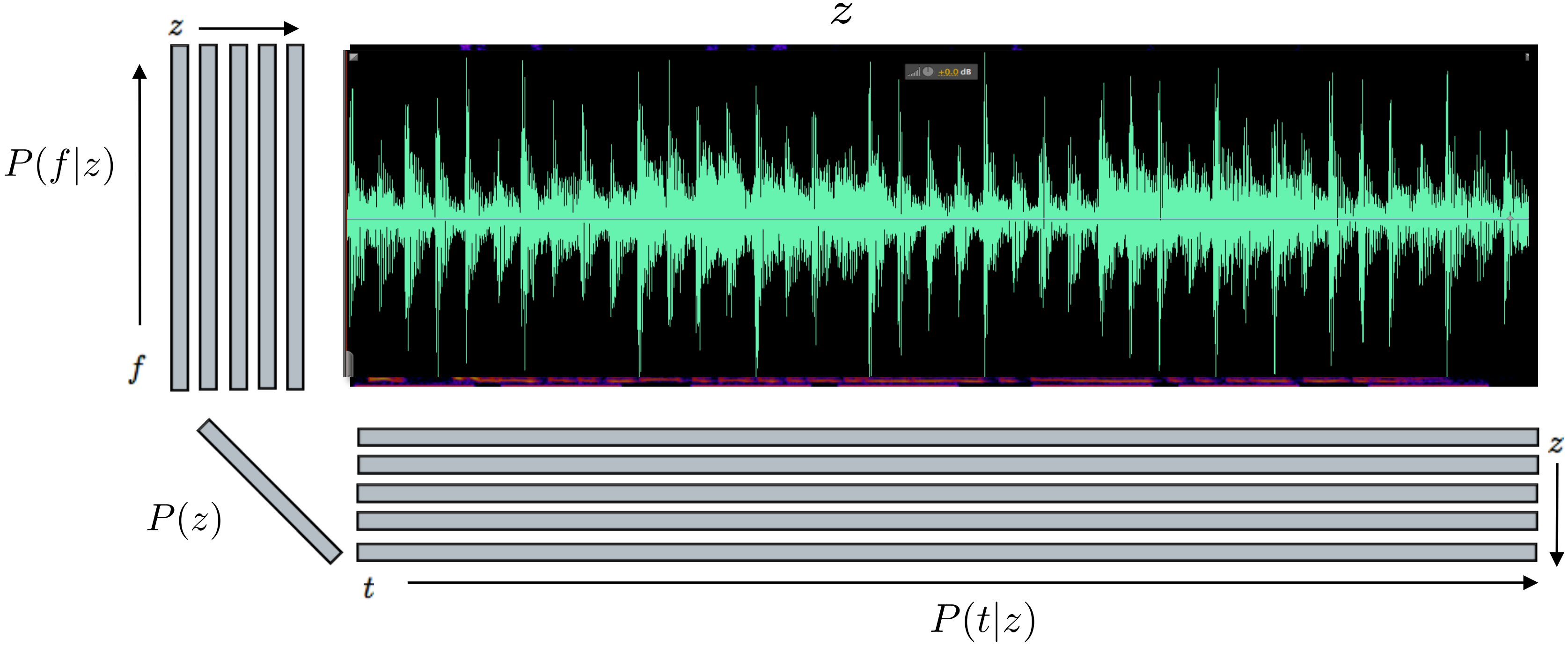
- Non-negative matrix factorization (NMF)
  [Lee & Seung, 2001; Smaragdis & Brown 2003]

- Related latent variable models (LVM)
  [Raj & Smaragdis 2005, Smaragdis et al., 2006]

Latent Variable Model

- Probabilistic latent component analysis (PLCA) [Smaragdis et al., 2006]

$$\mathbf{X} \approx P(f,t) = \sum_{z} P(z)P(f|z)P(t|z)$$



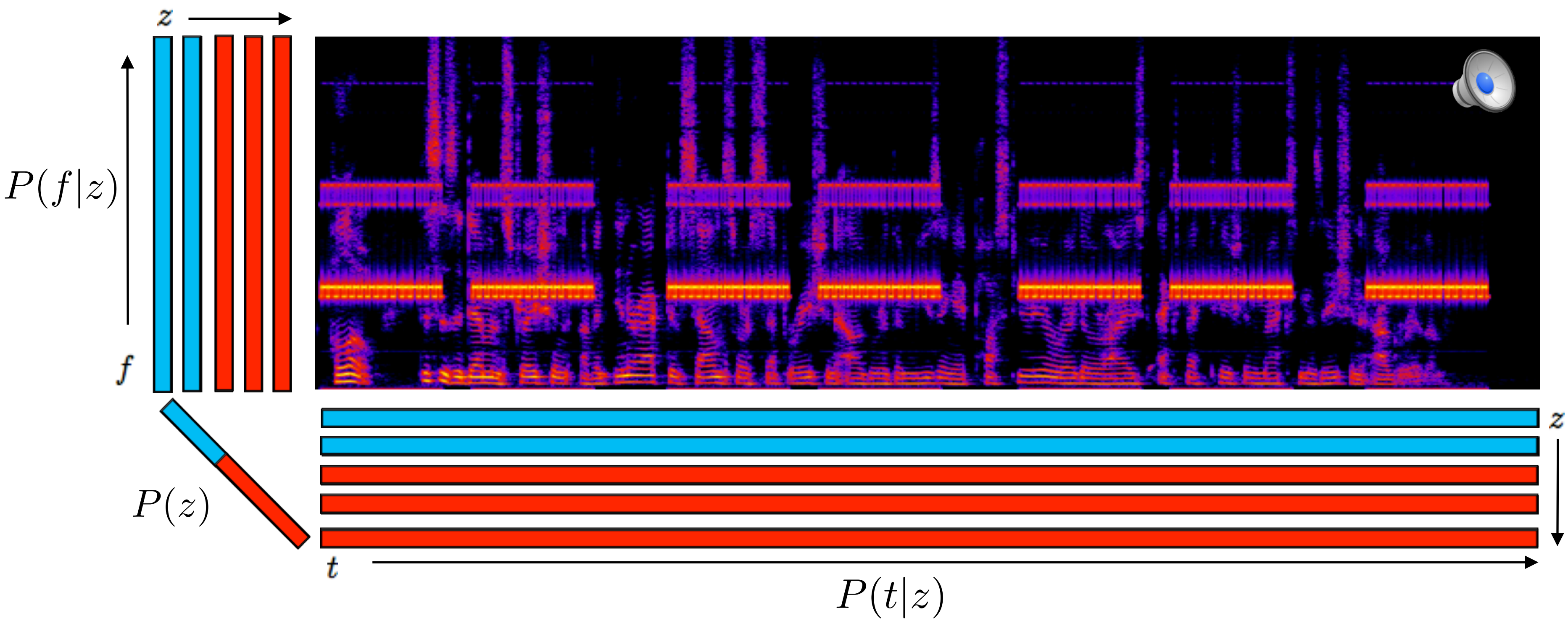
$P(f|z)$ Basis vectors, frequency components, dictionary

$P(z)$ Latent component weights

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

Latent Variable Model

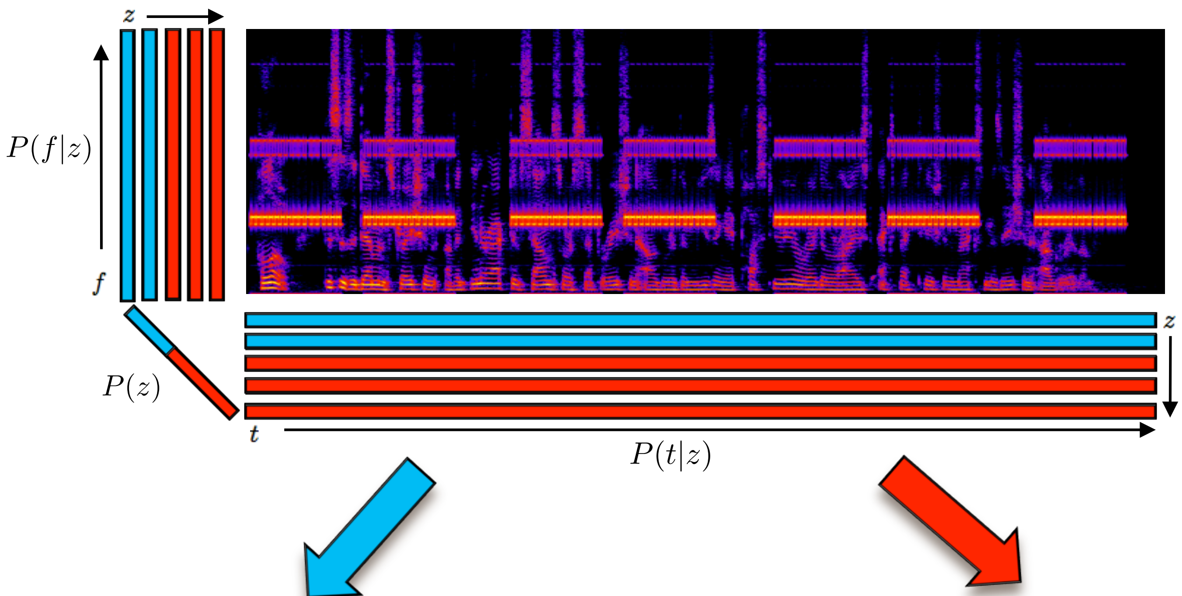
$$\mathbf{X} \approx P(f,t) = \sum_z P(z)P(f|z)P(t|z)$$

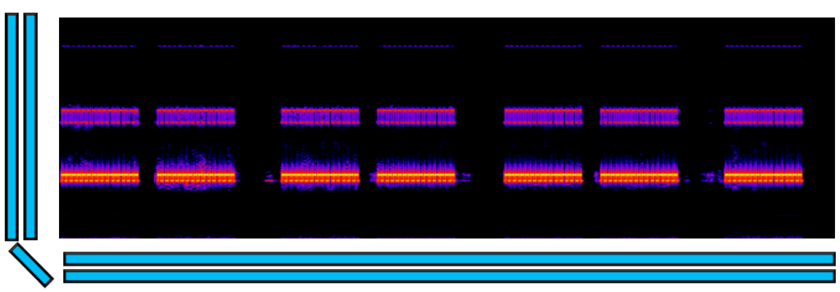


- Solve via an expectation-maximization (EM) algorithm

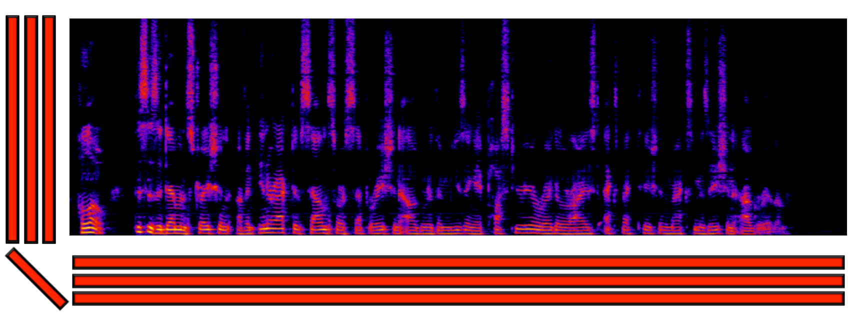
Latent Variable Model

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





$P(s = s_1|f,t)$



$P(s = s_2|f,t)$

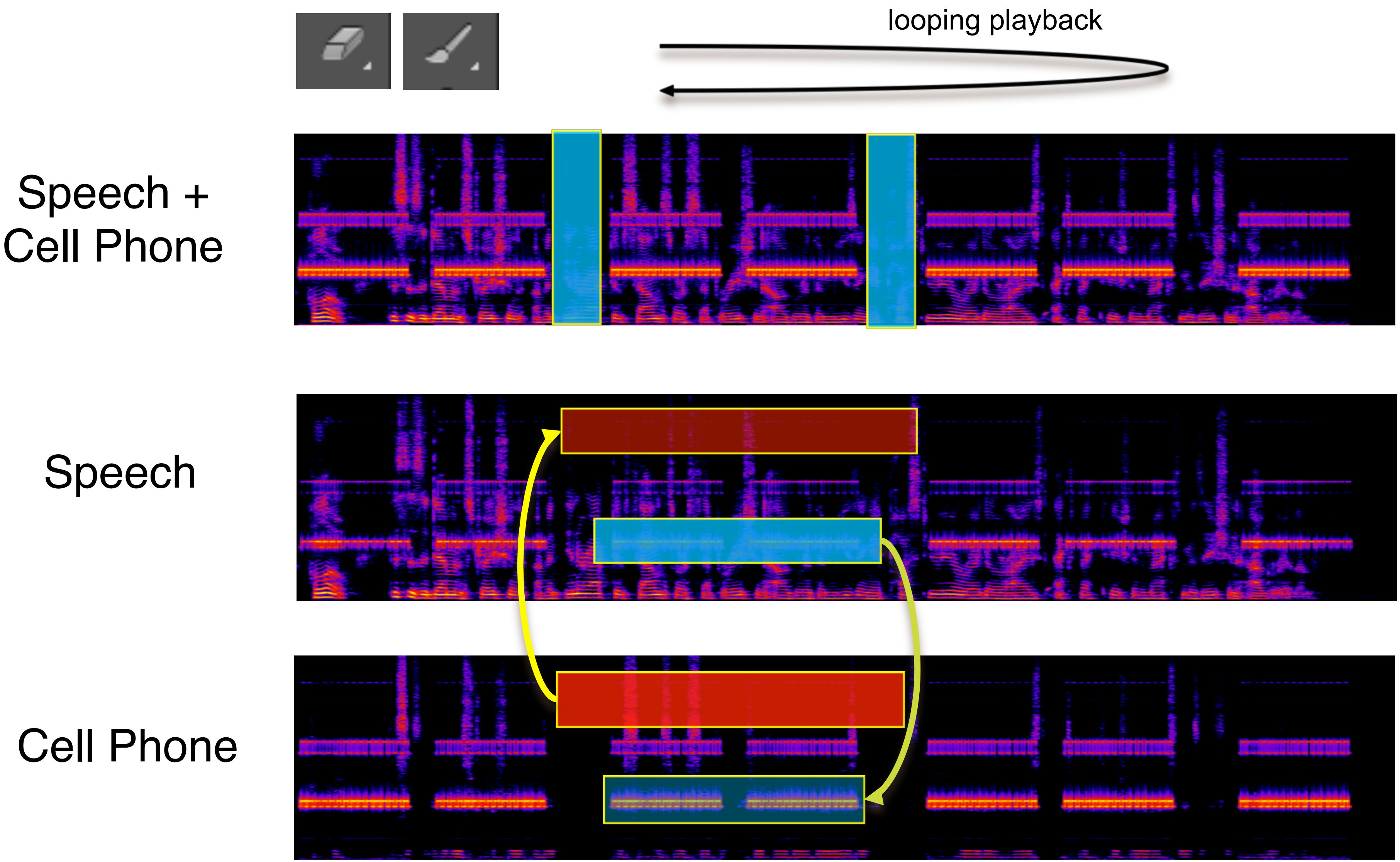
Problems

- Requires isolated training data (supervised/semi-supervised)
- Don’t incorporate auditory/perceptual models of hearing
- One-shot process, cannot correct for poor results
- Very difficult, underdetermined problem

Focus

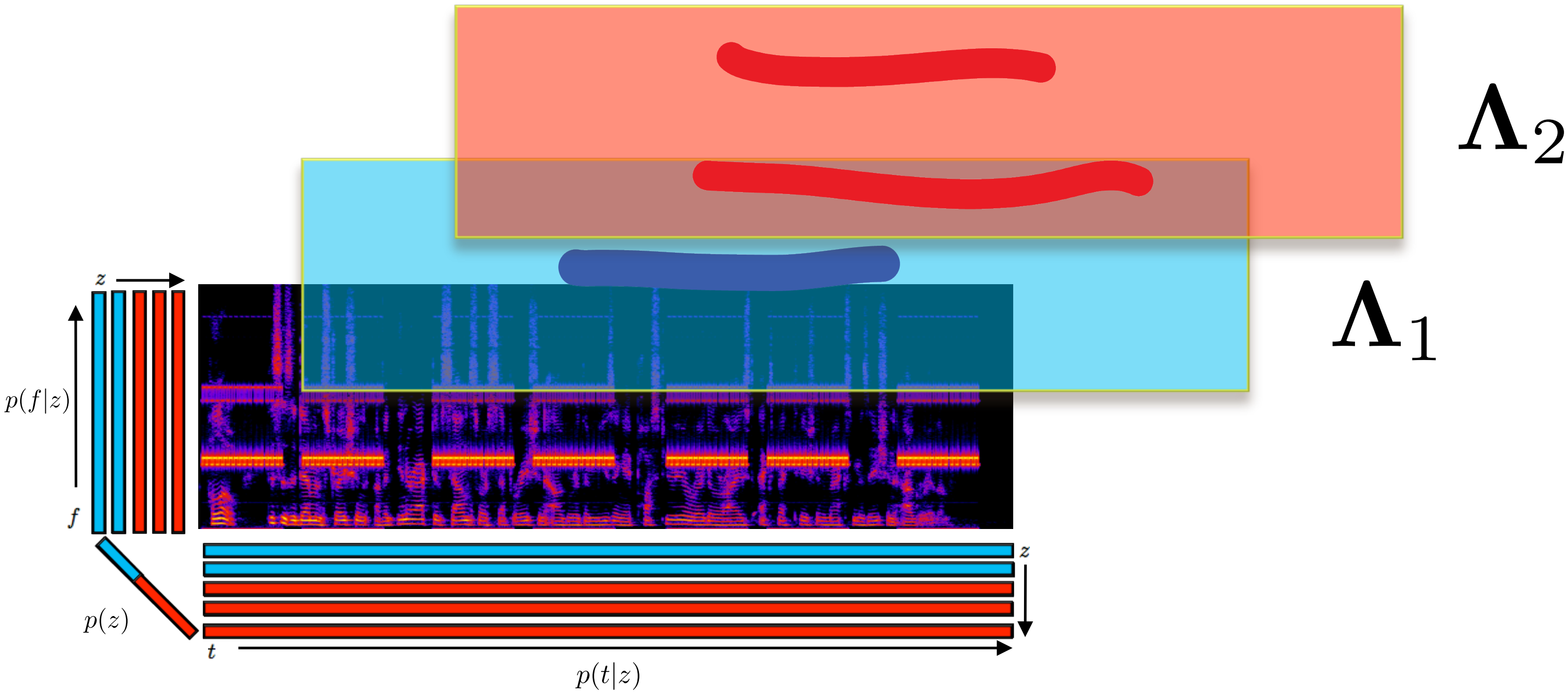
- Eliminate the need to explicit training data
- Method of user feedback to guide separation
- Algorithm to incorporate the user feedback

Paradigm: Listen, Paint, Remove



Latent Variable Model w/Painting Constraints

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



- Incorporate painting annotations into the model

Constraints

- Constraints typical encoded as: $P(f|z) \quad P(t|z) \quad P(z)$
  - Prior probabilities on model parameters
  - Direct observations
- Does not (reasonably) allow time-frequency constraints
- Posterior regularization [Graça et al., 2007, 2009]
  - 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

$$\ln P(\mathbf{X}|\mathbf{\Theta}) = \mathcal{F}(Q, \mathbf{\Theta}) + \mathrm{KL}(Q||P)$$

$$\ln P(\mathbf{X}|\mathbf{\Theta}) \geq \mathcal{F}(Q, \mathbf{\Theta})$$

E Step:

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

M Step:

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

Expectation Maximization w/Posterior Constraints I

$$\ln P(\mathbf{X}|\boldsymbol{\Theta}) = \mathcal{F}(Q, \boldsymbol{\Theta}) + \mathrm{KL}(Q||P)$$

$$\ln P(\mathbf{X}|\boldsymbol{\Theta}) \geq \mathcal{F}(Q, \boldsymbol{\Theta})$$

E Step:

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

M Step:

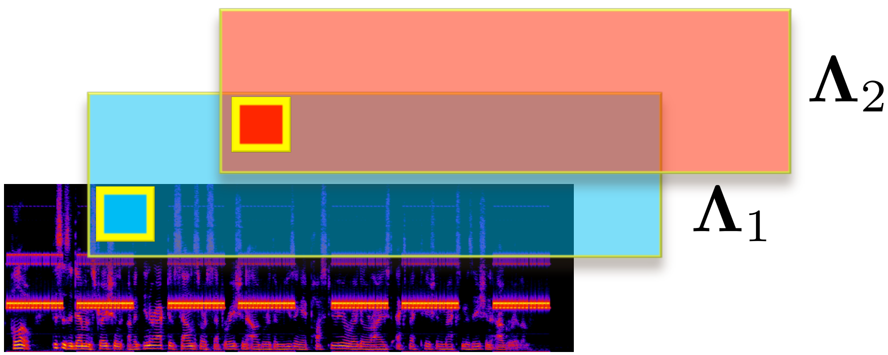
$$\boxed{\boldsymbol{\Theta}^{n+1} = \arg\max_{\boldsymbol{\Theta}} \mathcal{F}(Q^{n+1}, \boldsymbol{\Theta})}$$

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

$$\begin{aligned} &\arg\min_{\mathbf{q}} && -\mathbf{q}^{\mathrm{T}} \ln \mathbf{p} + \mathbf{q}^{\mathrm{T}} \ln \mathbf{q} \\ &\text{subject to} && \mathbf{q}^{\mathrm{T}} \mathbf{1} = 1, \;\; \mathbf{q} \succeq 0 \end{aligned}$$



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

Fast 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**

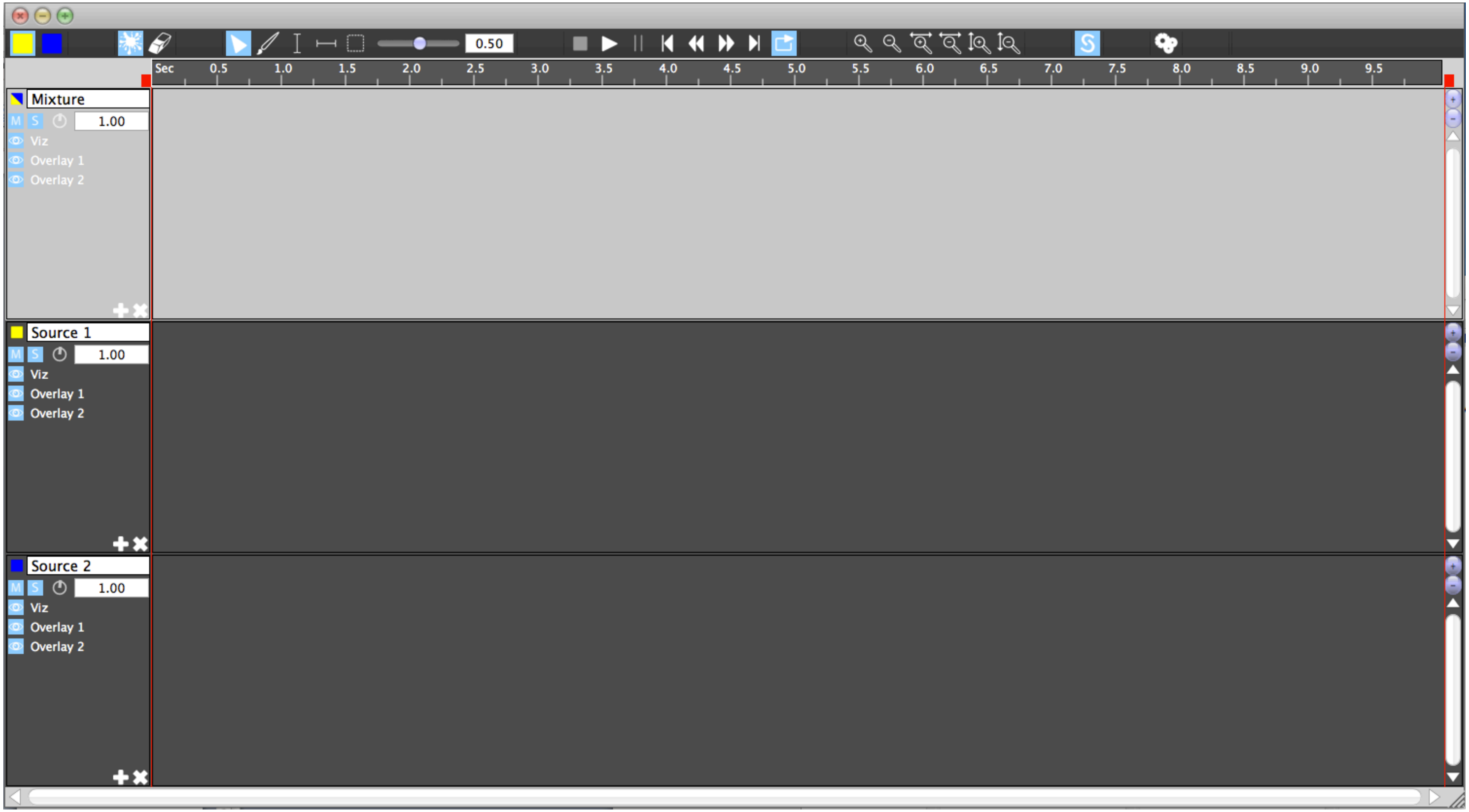
$$Q(z|f,t) \leftarrow \frac{P(z)P(f|z)P(t|z)\boxed{\tilde{\Lambda}_{(f,t,z)}}}{\sum_{z'} P(z')P(f|z')P(t|z')\boxed{\tilde{\Lambda}_{(f,t,z')}}}$$

end for

Evaluation

- BSS-EVAL metrics [Vincent et al., 2006]
  - Signal-to-Distortion Ratio (SDR)
  - Signal-to-Interference Ratio (SIR)
  - Signal-to-Artifact Ratio (SAR)
- Test material
  - Cell phone + speech (C), drums + bass (D), orchestra + cough (O), piano + wrong note (P), siren + speech (S)
  - Vocals + background music (S1, S2, S3, S4)
- Results
  - Outperformed prior state-of-the-art on tested material
  - Outperformed SiSEC 2011 vocals + background music winner

Live Demonstration



Jackson 5 Remix

Jackson 5's "I want You Back"

Cher Llyod's "Want U Back"

Remix

A Look Back

- Perceptual domain, objective evaluation is difficult
- Human evaluation within the learning process
- Processing training data only

Conclusion

- Sound source separation algorithm
  - Time-frequency constraints via posterior regularization
  - No explicit training data
  - Efficient, interactive algorithm w/closed-form update equations
  - Improved separation quality over prior work
  - Open source software
- Poster ID: 348
- Demos at ccrma.stanford.edu/~njb/research/iss

An Efficient Posterior Regularized Latent Variable Model for Interactive Sound Source Separation

Nicholas J. Bryan, Stanford University

Gautham J. Mysore, Adobe Research

ICML 2013