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

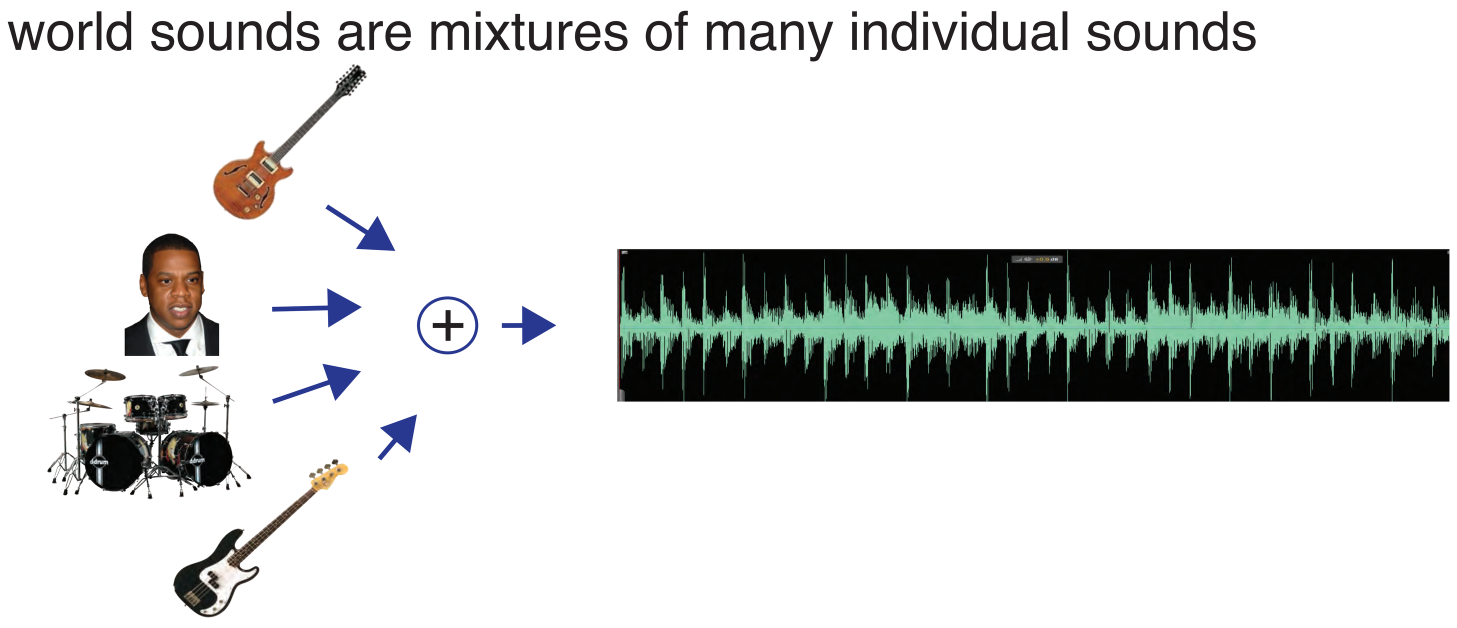
Nicholas J. Bryan[1] & Gautham J. Mysore[2]

[1]Center for Computer Research in Music and Acoustics, Stanford University

[2]Adobe Research

Introduction

- Real world sounds are mixtures of many individual sounds



- It's useful to separate a mixture into its respective sources

music transcription **audio denoising**

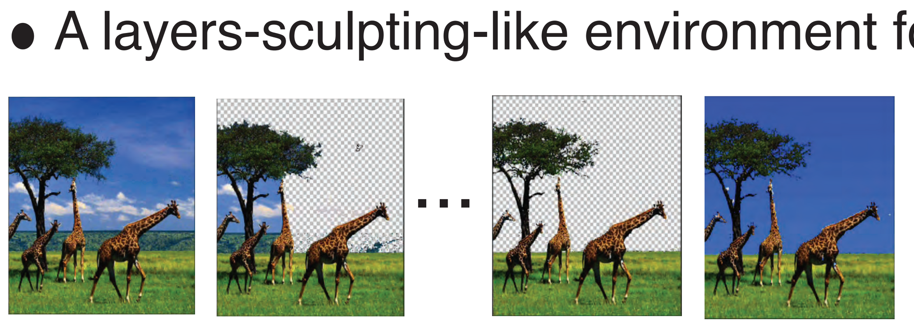
audio-based forensics **music remixing**

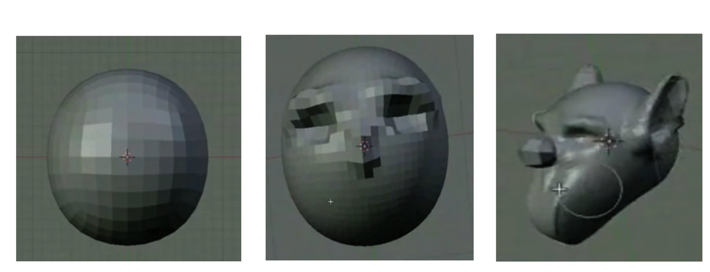
- Current non-negative matrix factorization and related probabilistic models methods can perform well, but:

  -require training data
  -may also yield poor results
  -are typically a one-shot process w/no user-feedback

Analogy

- A layers-sculpting-like environment for audio



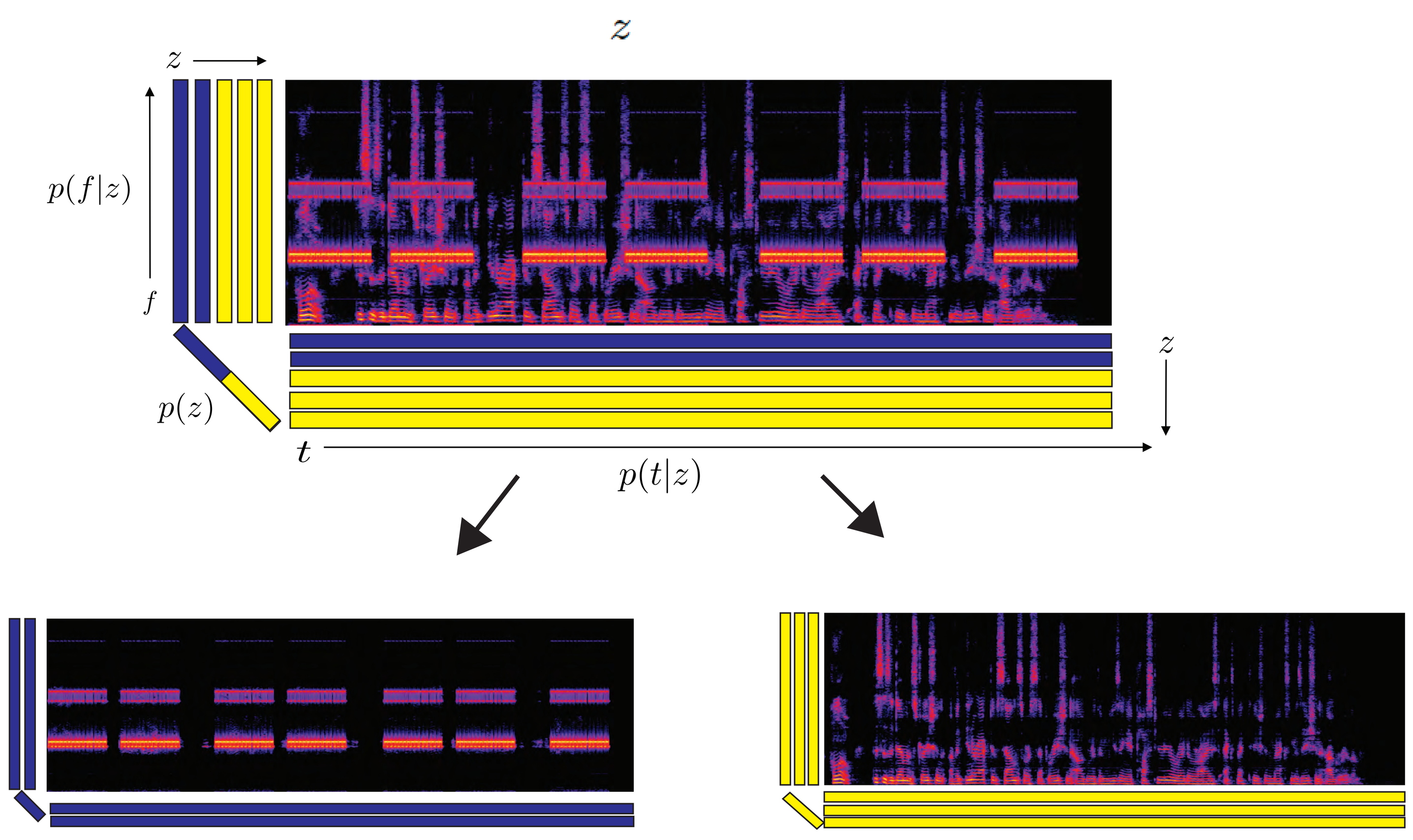


- Remove burden of being perfect the first time and interact w/algorithm

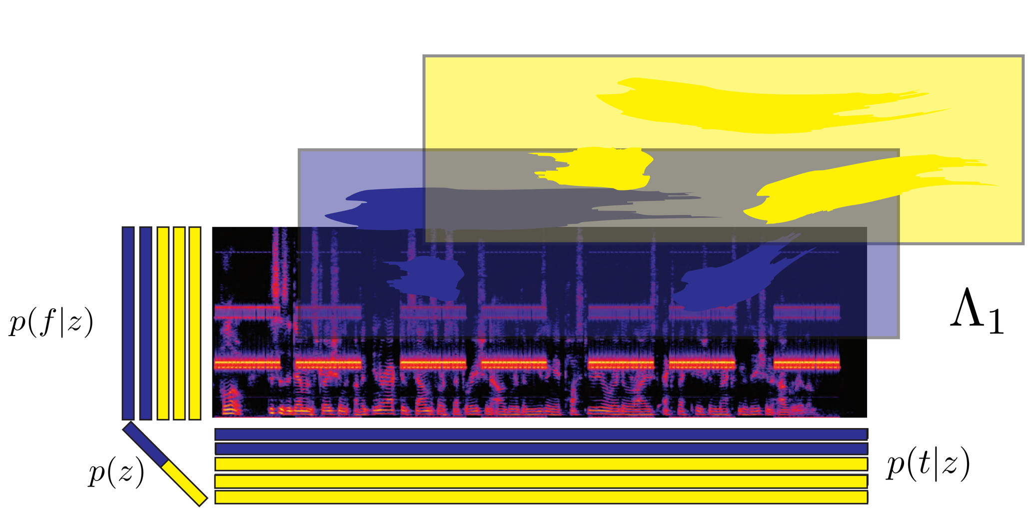
Proposed Method

- Probabilistic model of audio spectrogram data

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



- Interactively constrain/regularize the model via painting annotations



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

- Parameter estimation via expectation-maximization
- No explicit training data needed

Posterior Regularization

- Incorporate painting annotations as penalty constraints
- Difficult to encode time-frequency-source constraints via priors
- Use framework of posterior regularization for EM algorithms
- Contraints on the posterior (E step) as oppose to standard priors (M step)

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

- Map painting annotations to linear grouping expectation constraints
- Within a single E step, solve for each time-frequency point:

$$\begin{aligned} &\arg\min_{\mathbf{q}} && -\mathbf{q}^{\mathrm{T}}\ln\mathbf{p} + \mathbf{q}^{\mathrm{T}}\ln\mathbf{q} + \mathbf{q}^{\mathrm{T}}\boldsymbol{\lambda} \\ &\text{subject to} && \mathbf{q}^{\mathrm{T}}\mathbf{1} = 1,\ \mathbf{q} \succeq 0 \end{aligned} \qquad \boldsymbol{\lambda} = [\alpha,\alpha,\beta,\beta,\beta]$$

- Results in closed-form, efficient E and M steps ⟶ interactive speeds

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

$$Q(z|f,t) \leftarrow \frac{P(z)P(f|z)P(t|z)}{\sum_{z'} P(z')P(f|z')P(t|z')} \quad (8)$$

  end for
  maximization step
  for all z, f, t **do**

$$P(f|z) \leftarrow \frac{\sum_t \mathbf{X}_{(f,t)} Q(z|f,t)}{\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')} Q(z|f',t')} \quad (9)$$

$$P(t|z) \leftarrow \frac{\sum_f \mathbf{X}_{(f,t)} Q(z|f,t)}{\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')} Q(z|f',t')} \quad (10)$$

$$P(z) \leftarrow \frac{\sum_f\sum_t \mathbf{X}_{(f,t)} Q(z|f,t)}{\sum_{z'}\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')} Q(z'|f',t')} \quad (11)$$

  end for
until convergence

⟶

precompute: $\tilde{\mathbf{\Lambda}} \leftarrow \exp\{-\mathbf{\Lambda}\}$
repeat
  expectation step
  for all z, f, t **do**

$$Q(z|f,t) \leftarrow \frac{P(z)P(f|z)P(t|z)\tilde{\mathbf{\Lambda}}_{(f,t,z)}}{\sum_{z'} P(z')P(f|z')P(t|z')\tilde{\mathbf{\Lambda}}_{(f,t,z')}} \quad (28)$$

  end for
  maximization step
  for all z, f, t **do**

$$P(f|z) \leftarrow \frac{\sum_t \mathbf{X}_{(f,t)} Q(z|f,t)}{\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')} Q(z|f',t')} \quad (29)$$

$$P(t|z) \leftarrow \frac{\sum_f \mathbf{X}_{(f,t)} Q(z|f,t)}{\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')} Q(z|f',t')} \quad (30)$$

$$P(z) \leftarrow \frac{\sum_f\sum_t \mathbf{X}_{(f,t)} Q(z|f,t)}{\sum_{z'}\sum_{f'}\sum_{t'} \mathbf{X}_{(f',t')} Q(z'|f',t')} \quad (31)$$

  end for
until convergence

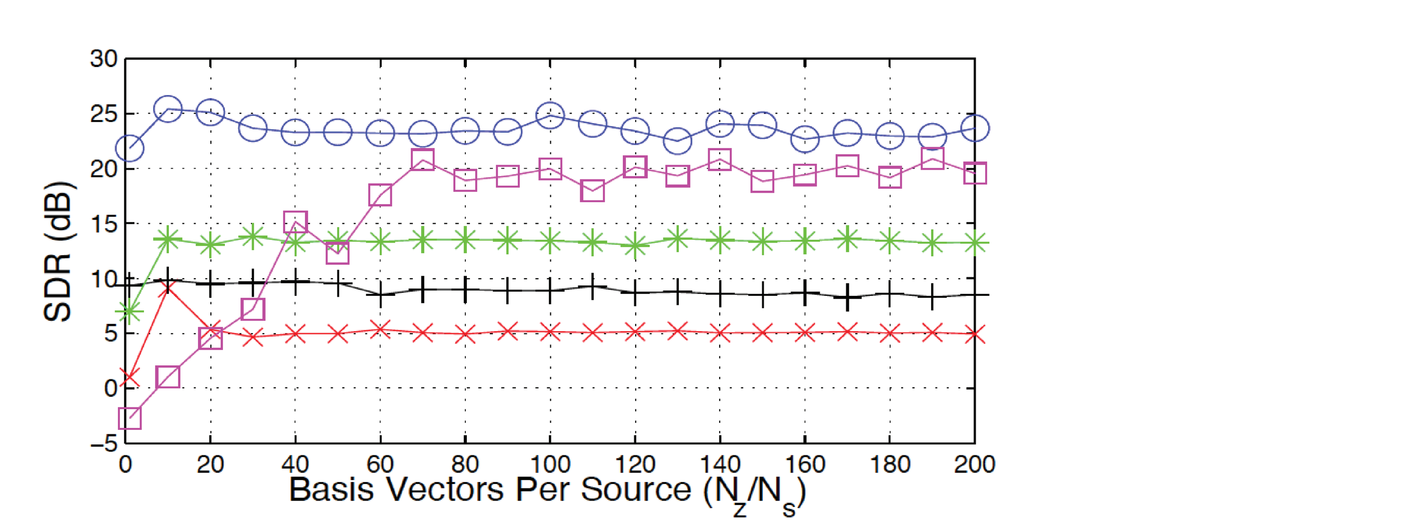
Evaluation

- Use SDR, SAR, SIR evaluation metrics for comparison
- Tested on a variety of sounds: cell phone + speech (C), drum + bass (D), orchestra + cough (O), piano + wrong note (P), siren + speech (S), vocals + background music (S1, S2, S3, S4)

Eval	Method	C	D	O	P	S
SDR	Oracle	26.9	15.1	12.2	26.1	26.7
	Baseline	-0.6	0.2	1.1	0.9	-4.1
	Proposed	24.8	11.0	9.7	22.0	21.8
SIR	Oracle	34.1	20.0	16.6	29.9	34.3
	Baseline	0.1	0.9	2.2	1.1	0.2
	Proposed	35.0	19.1	14.6	26.3	29.0
SAR	Oracle	27.9	16.8	14.6	28.8	27.6
	Baseline	14.0	12.6	10.5	17.5	7.0
	Proposed	25.8	12.6	11.7	24.3	23.2

Eval	Method	S1	S2	S3	S4
SDR	Oracle	13.2	13.4	11.5	12.5
	Baseline	-0.8	0.2	-0.2	1.4
	Lefévre	7.0	5.0	3.8	5.0
	Durrieu	9.0	7.8	6.4	5.9
	Proposed	9.2	11.1	7.8	7.9
SIR	Oracle	17.8	18.0	17.5	19.5
	Baseline	0.5	1.6	0.9	3.1
	Lefévre	13.0	14.1	8.8	11.5
	Durrieu	16.4	16.8	13.0	12.6
	Proposed	17.4	20.1	14.8	13.8
SAR	Oracle	15.4	15.4	13.1	13.6
	Baseline	8.9	8.5	8.8	10.0
	Lefévre	8.9	7.3	6.1	6.5
	Durrieu	10.5	9.0	8.0	8.3
	Proposed	10.7	12.0	9.0	9.5

- Relatively insensitive to the number of latent components (if large enough)



- On the examples tested, the proposed method outperfomed prior work

Conclusions

- Source separation algorithm that allows:

  -time-frequency constraints via posterior regularization

  -efficient, interactive algorithm

  -improved results over prior work

- For audio and video demonstrations, please see https://ccrma.stanford.edu/~njb/research/iss