Deep Unfolding for Multichannel Source Separation

Scott Wisdom, John Hershey, Jonathan Le Roux, Shinji Watanabe

Department of Electrical Engineering
University of Washington
Seattle, WA, USA

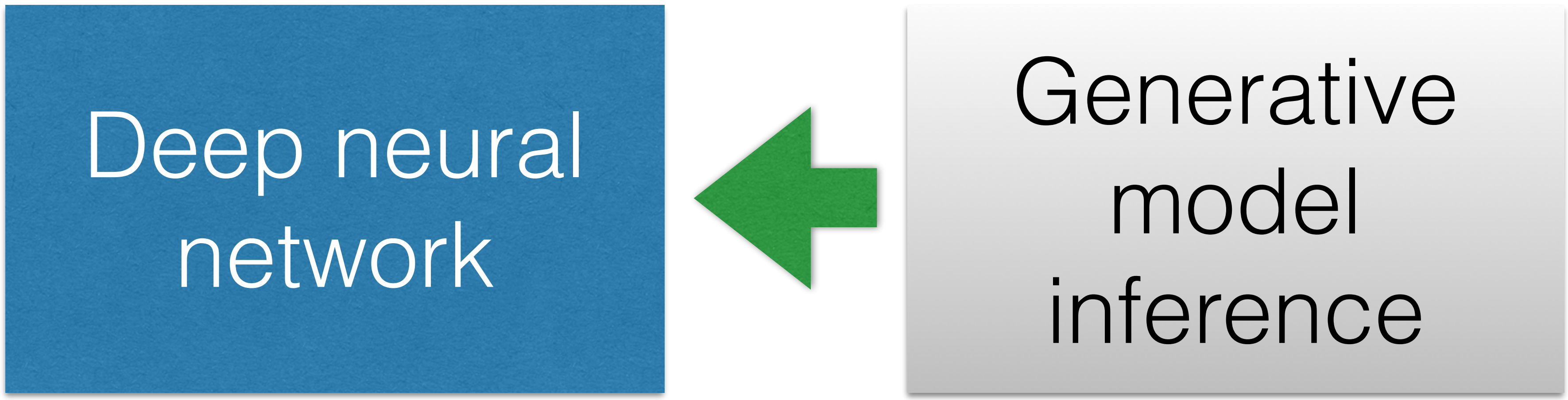
Mitsubishi Electric Research Labs (MERL)
Cambridge, MA, USA

This work was done while S. Wisdom was an intern at MERL and at JSALT 2015 in UW, Seattle, which was supported by JHU via grants from NSF (IIS), Google, Microsoft, Amazon, Mitsubishi Electric, and MERL.

Motivation

- Deep neural networks work well for a wide variety of tasks
  - Computer vision, speech recognition, speech enhancement
- But not many for microphone arrays (only [1]-[4]). How to incorporate domain knowledge?
- How can we use deep neural networks for multichannel source separation?

[1] Nugraha et al. 2015
[2] Hoshen et al. 2015
[3] Sainath et al. 2015, 2016
[4] Xiao et al. 2016



Approach:

- We use a new method called “deep-unfolding” to create deep neural networks from generative models
  - Recently used for NMF [1] and LDA topic modeling [2]
- We can improve the generative model, which tells us how to change the architecture of the neural network

[1] Le Roux et al. 2015
[2] Chen et al. 2015

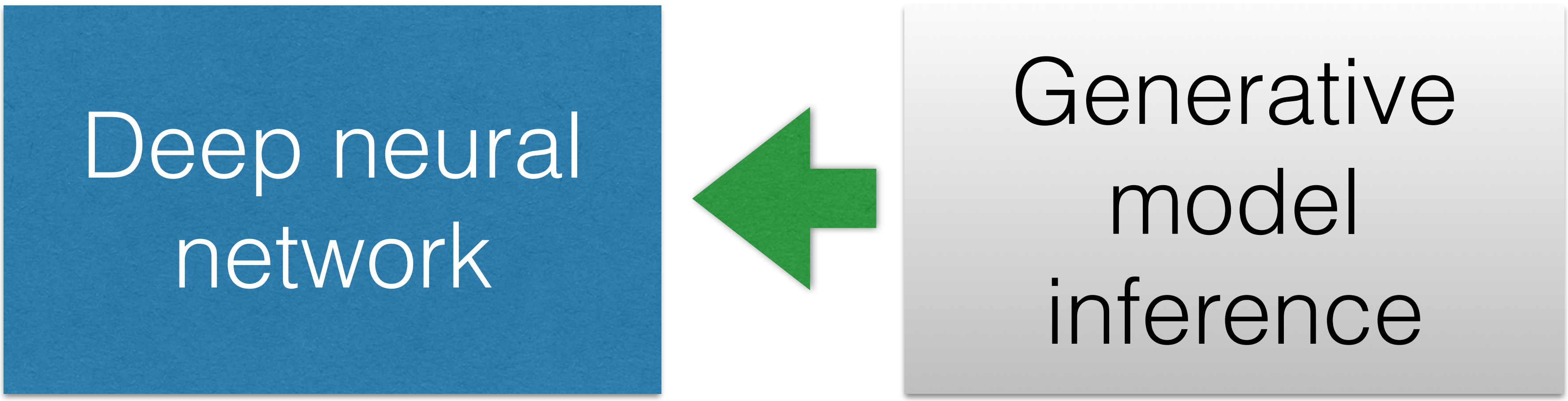
Results:

- A meaningful and interpretable deep network that can separate sources in complex-valued multichannel frequency-domain.
- Discriminative training improves performance of the original inference algorithm.

Outline

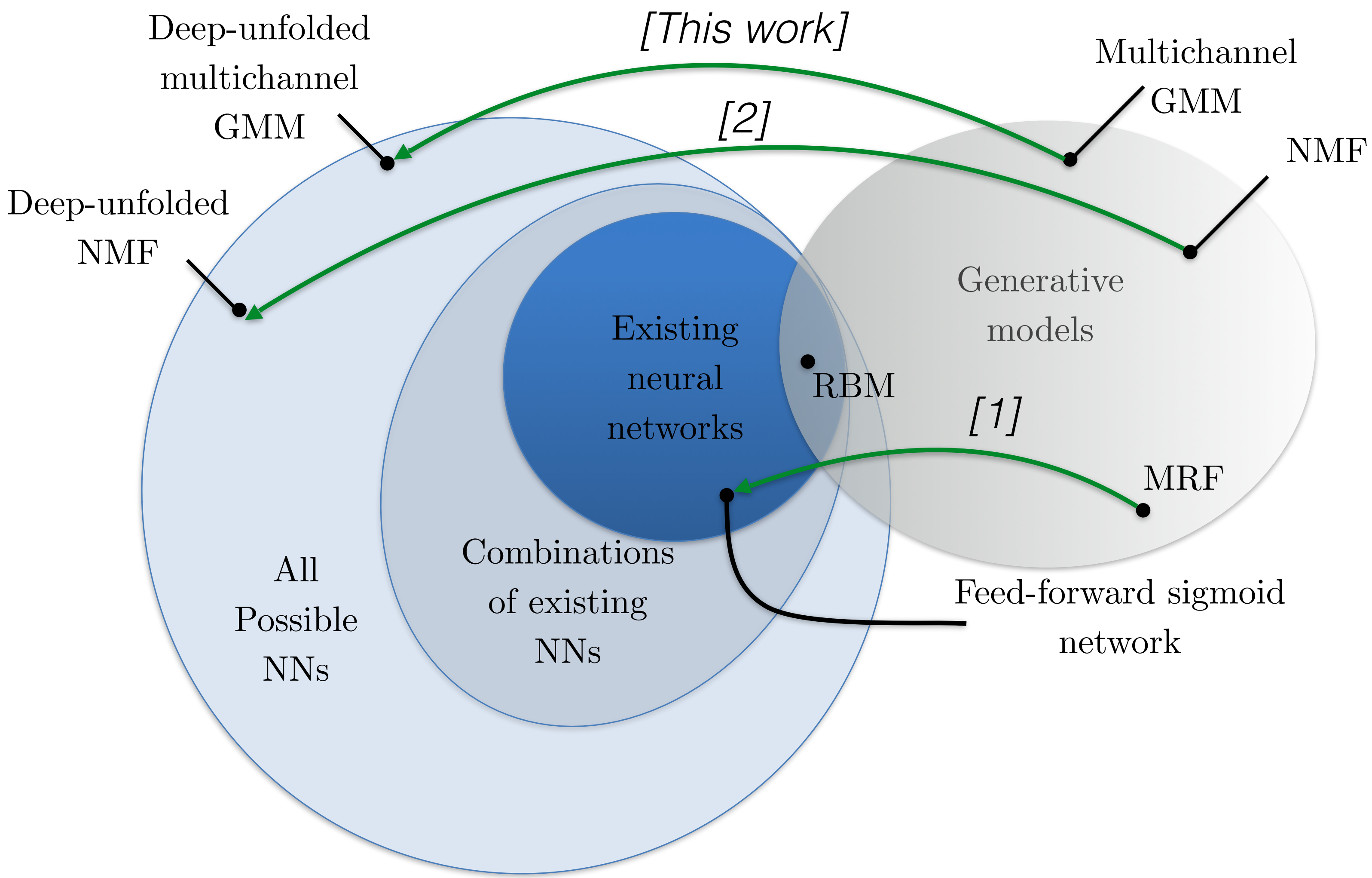
1. **Deep unfolding overview**
2. Generative model: multichannel GMM
3. Unfolding the multichannel GMM
4. Results

1. Deep unfolding overview



- Deep unfolding enables creation of principled and novel deep architectures from generative models.

1. Deep unfolding overview

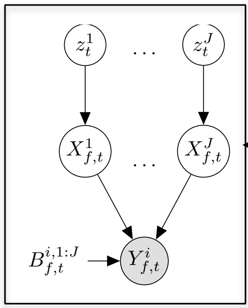


- Deep unfolding enables creation of principled and novel deep architectures from generative models.

[1] Hershey, Le Roux, Weninger 2014

[2] Le Roux, Hershey, Weninger 2015

(1) Graphical model



$\leftarrow \theta$

(1) Define model

(2) Derive iterative inference algorithm

(3) Unfold iterations into layers in a network

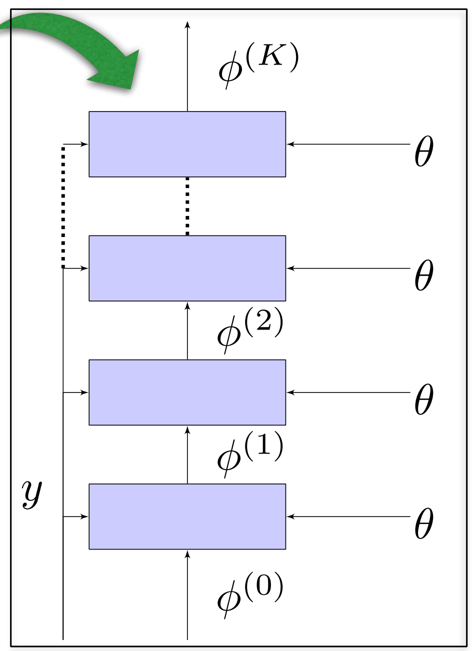
(4) Discriminatively train parameters θ, **tying** or **untying** between layers.

(2)

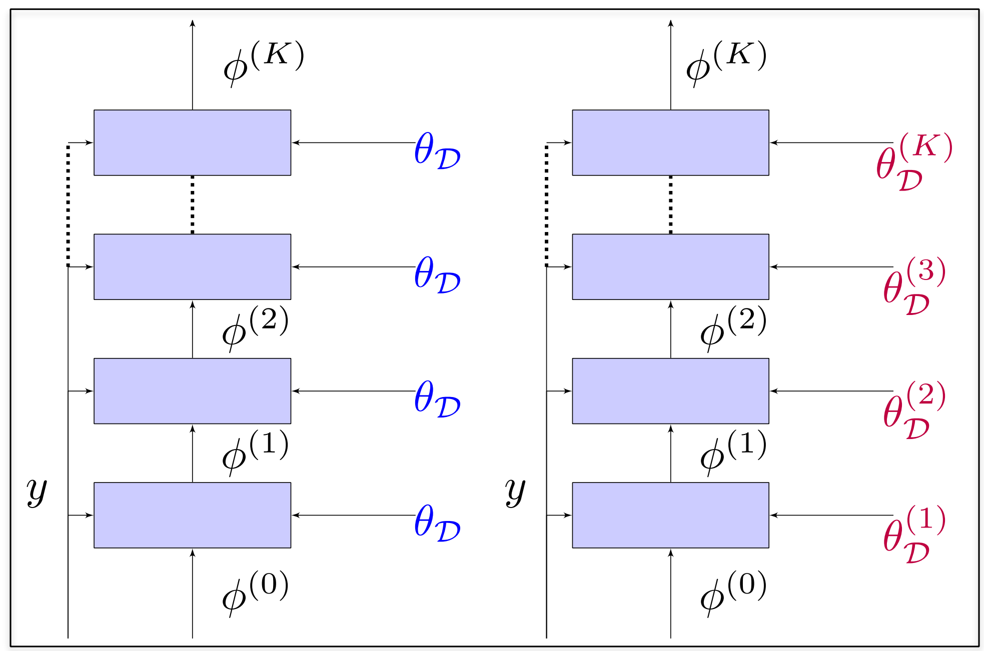
Iterative inference

For k=1:K,
Update $\phi^{(k)}$
using $\phi^{(k-1)}$, θ,
and data y

(3) Unfolded network



(4) Trained unfolded networks

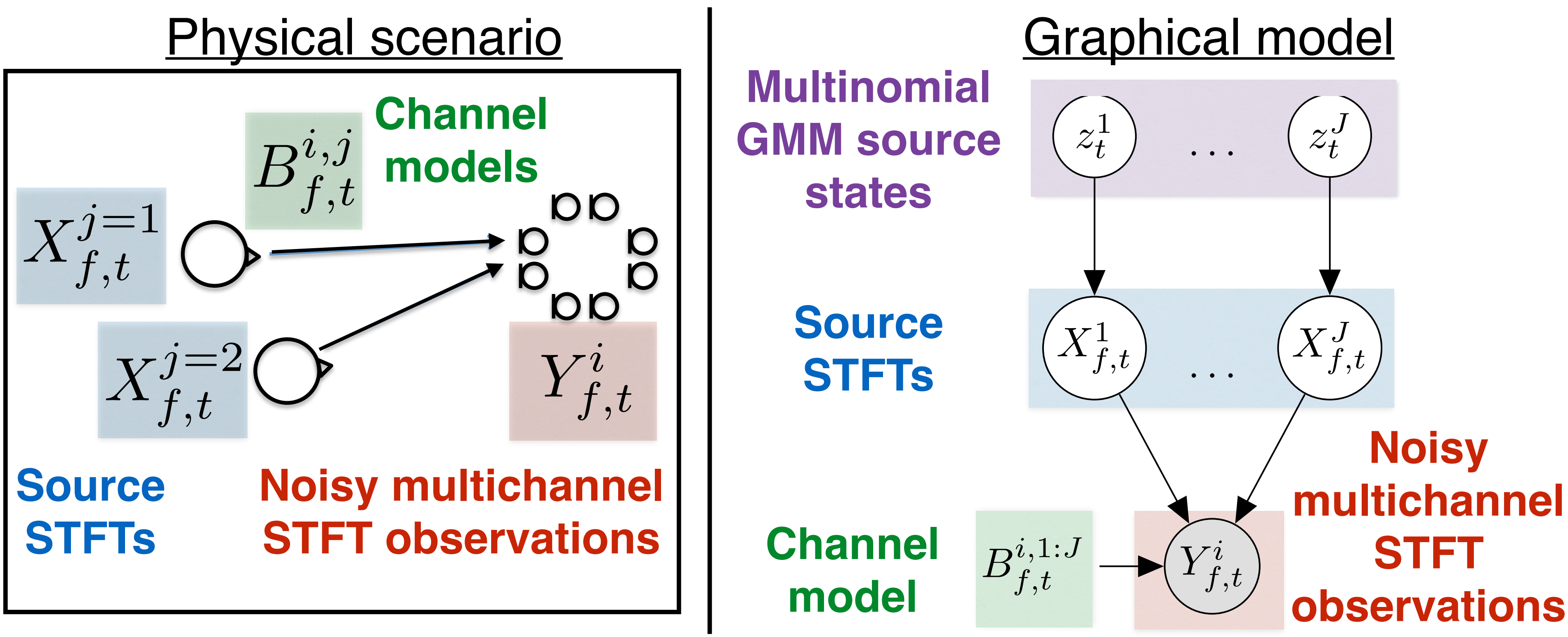


[1] Hershey, Le Roux, Weninger 2014

Outline

1. Deep unfolding overview
2. **Generative model: multichannel GMM**
3. Unfolding the multichannel GMM
4. Results

2. Generative model: multichannel GMM

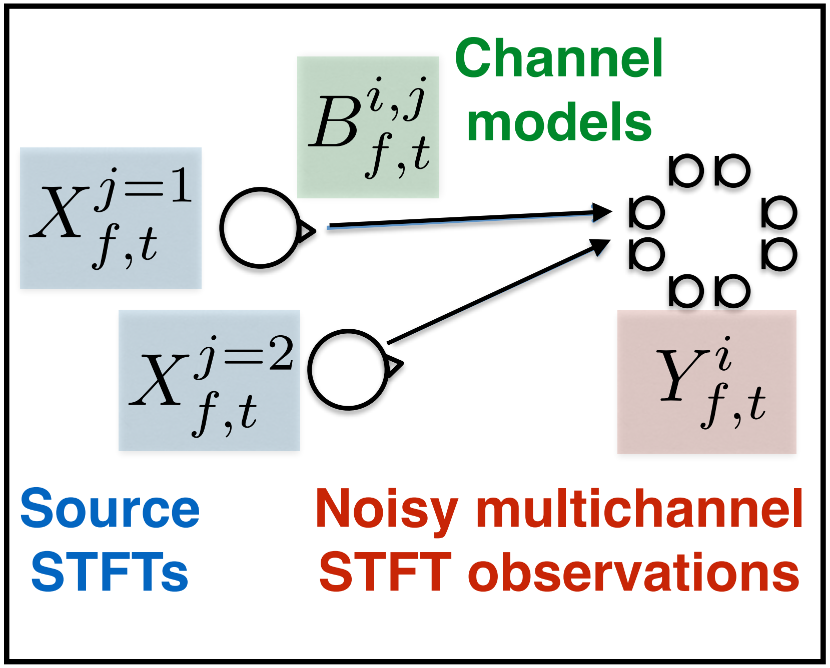


- Multichannel GMM (MCGMM): probabilistic model of complex-valued multichannel STFT [1]
  - GMM source models
  - Narrowband channel model

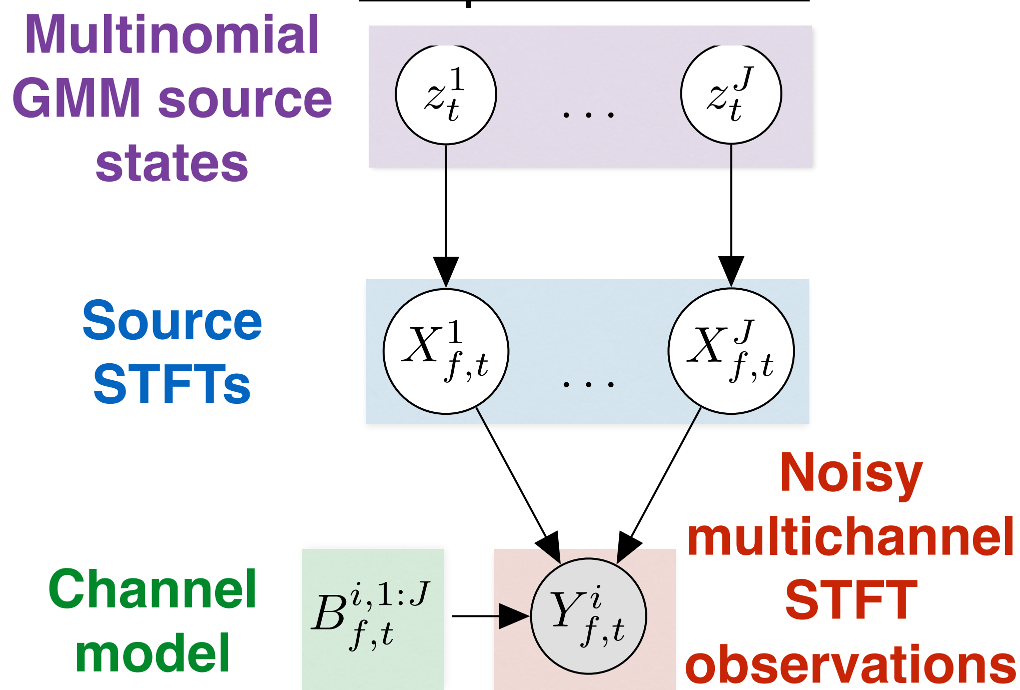
[1] Attias 2003

2. Generative model: multichannel GMM

<u>Physical scenario</u>



<u>Graphical model</u>



For k=1:K, estimate:

- Source GMM state probabilities
- Source means (complex STFTs)
- Channel model

<u>Iterative variational inference algorithm</u> [1]:

[1] Attias 2003

2. Generative model: multichannel GMM

Iterative variational inference algorithm [1]:

For k=1:K, estimate:

- Source GMM state prob.s
- Source means
- Channel model

Algorithm 1: Simplified variational EM algorithm for the MCGMM, where $\langle(\cdot)_t\rangle_t := \frac{1}{T}\sum_{t=1}^{T}(\cdot)_t$.

Data: Multichannel mixture STFT $Y_{1:F,1:T}$, sensor precision ψ_f, source parameters $\gamma_{1:F}^{1:J,1:Z}$, $\pi^{1:J,1:Z}$, initial channel estimate $B_{1:F}^{(0)}$

Result: Estimated source STFTs $\hat{X}_{1:F,1:T}^{1:J,(K)}$ and layer-wise intermediate variables

for $k = 1 : K$ **do**

Run E-step:

$$\bar{\gamma}_f^{j,z,(k)} = [B_f^{(k-1)}]_{:,j}^H \psi_f [B_f^{(k-1)}]_{:,j} + \gamma_f^{j,z,(k)} \tag{7}$$

$$\bar{\mu}_{f,t}^{j,z,(k)} = \frac{[B_f^{(k-1)}]_{:,j}^H \psi_f}{\bar{\gamma}_f^{j,z,(k)}} \left(Y_{f,t} - [B_f^{(k-1)}]_{:,\backslash j} \hat{X}_{f,t}^{\backslash j,(k-1)}\right) \tag{8}$$

$$L_t^{j,z,(k)} = \log \pi^{j,z} + \sum_f \log \frac{\gamma_f^{j,z,(k)}}{\bar{\gamma}_f^{j,z,(k)}} \ldots$$

$$\ldots + \sum_f \bar{\gamma}_f^{j,z,(k)} \left|\bar{\mu}_f^{j,z,(k)}\right|^2 \tag{9}$$

$$\bar{\pi}_t^{j,z,(k)} = \mathrm{softmax}\left(L_t^{j,1:Z,(k)}\right) \tag{10}$$

$$\hat{X}_{f,t}^{j,(k)} = \sum_z \bar{\pi}_t^{j,z,(k)} \bar{\mu}_{f,t}^{j,z,(k)} \tag{11}$$

Run M-step:

$$\hat{\Sigma}_f^{YX} = \left\langle Y_{f,t} (\hat{X}_{f,t}^{(k)})^H \right\rangle_t \tag{12}$$

$$[\hat{\Sigma}_f^{\hat{X}\hat{X}}]_{j,j} = \left\langle \sum_z \bar{\pi}_t^{j,z,(k)} \left(\frac{1}{\bar{\gamma}_f^{j,z,(k)}} + \left|\bar{\mu}_{f,t}^{j,z,(k)}\right|^2\right)\right\rangle_t \tag{13}$$

$$B_f^{(k)} = \hat{\Sigma}_f^{Y\hat{X}} \left(\hat{\Sigma}_f^{\hat{X}\hat{X}}\right)^{-1} \tag{14}$$

end

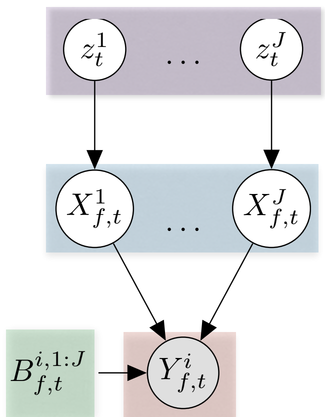
[1] Attias 2003

Outline

1. Deep unfolding overview
2. Generative model: multichannel GMM
3. **Unfolding the multichannel GMM**
4. Results

3. Unfolding the multichannel GMM

Graphical model

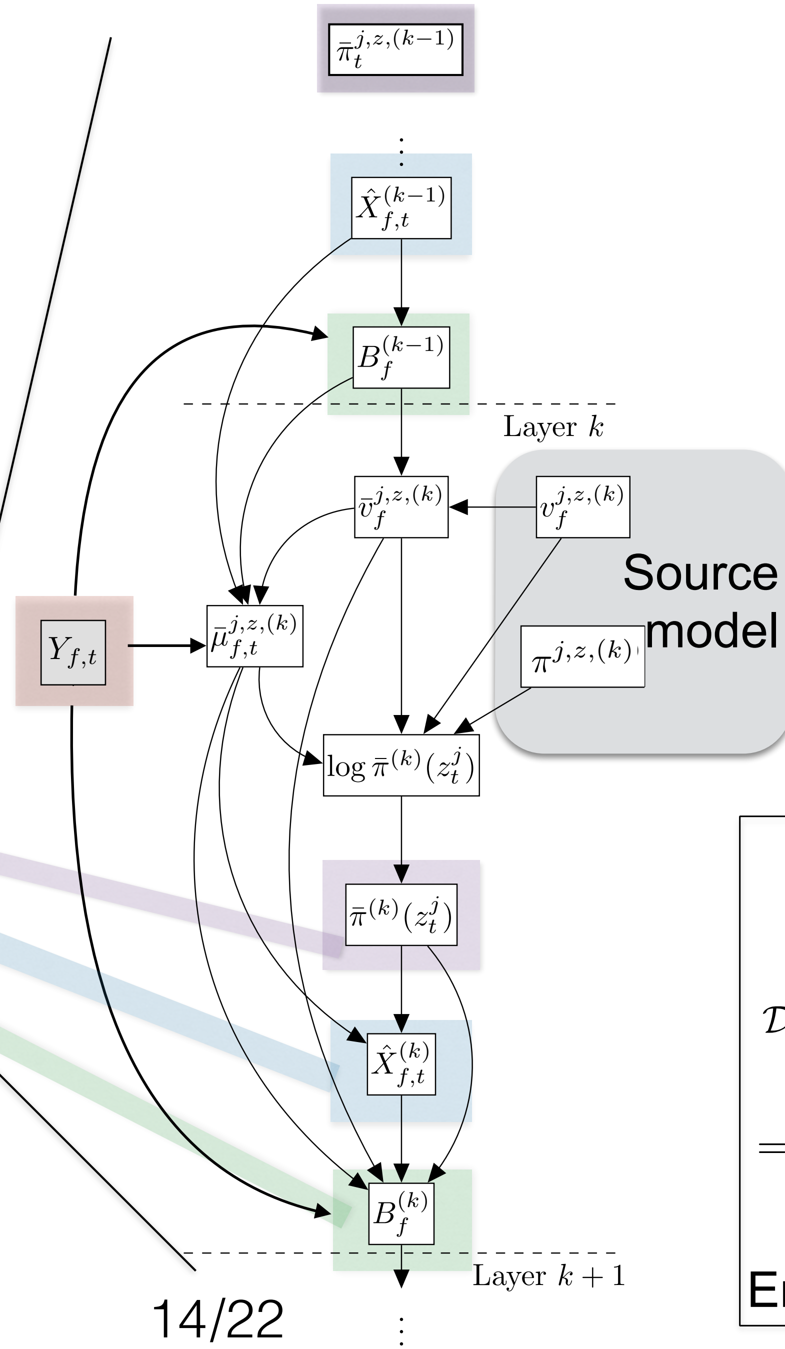


Iterative inference

For k=1:K, estimate:

- Source state probabilities
- Source means
- Channel model

One layer of unfolded network



Discriminative training:

Optimize

$v^{(k)}, \pi^{j,z,(k)}$

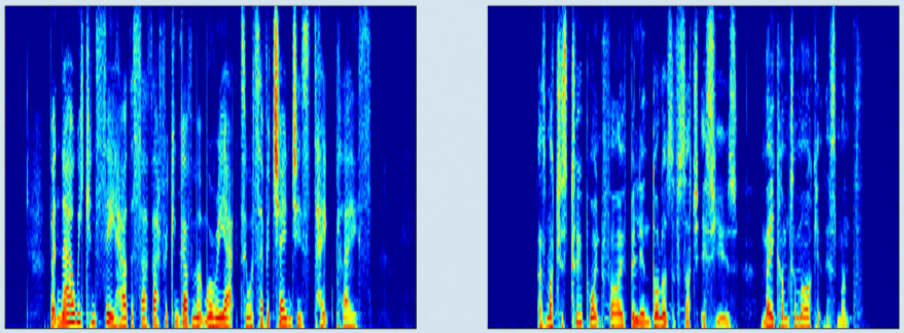
to minimize

$$\mathcal{D}_{ESR}(X_{f,t}, \hat{X}_{f,t}^{(K)}) = \sum_j \frac{\sum_{f,t} \left| \hat{X}_{f,t}^j - X_{f,t}^j \right|^2}{\sum_{f,t} \left| X_{f,t}^j \right|^2}$$

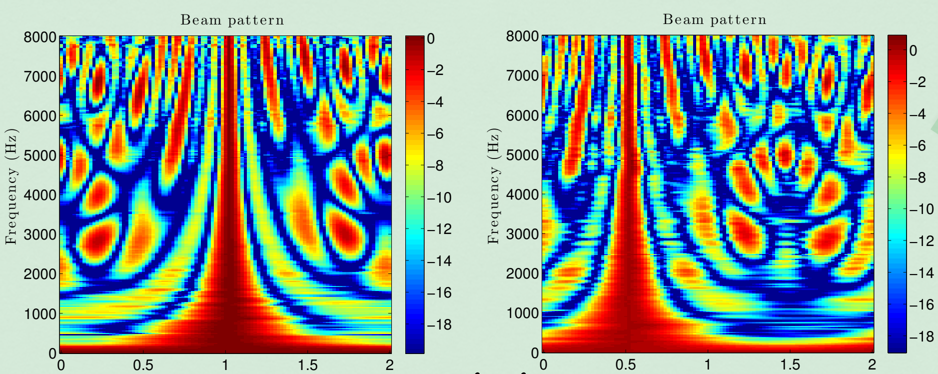
Error-to-source cost

3. Unfolding the multichannel GMM

- The unfolded network is perfectly interpretable!

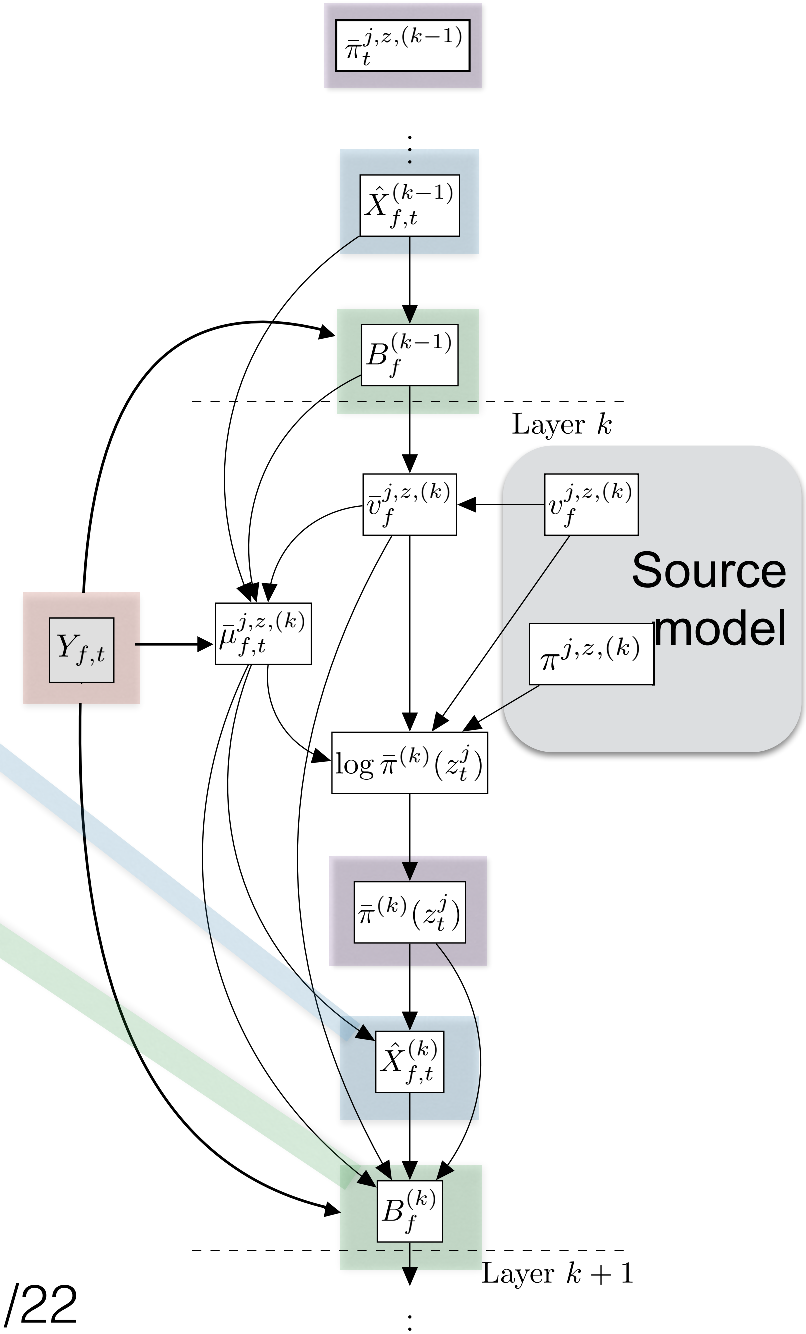


Source STFTs $X^{j}_{f,t}$



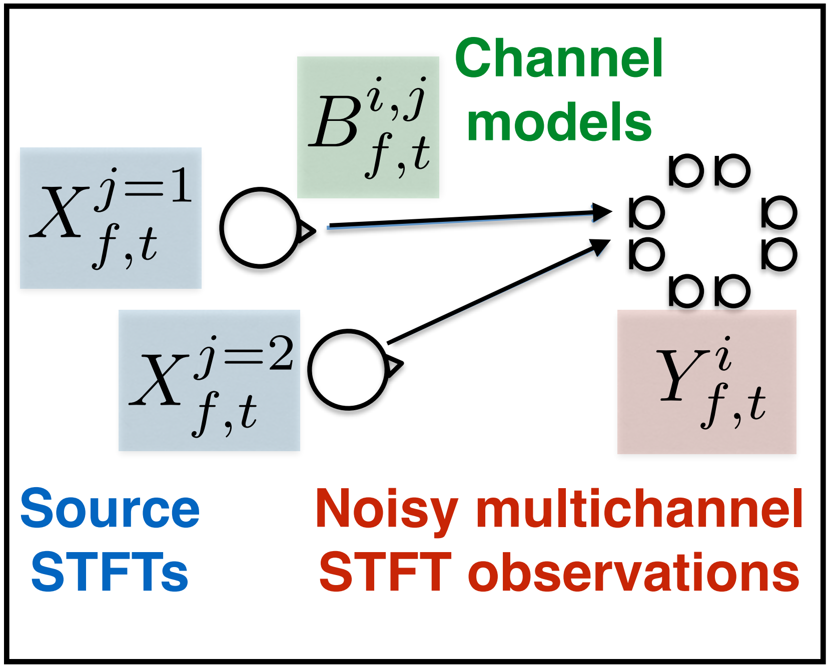
Channel models $B^{i,j}_{f,t}$

<u>One layer of unfolded network</u>

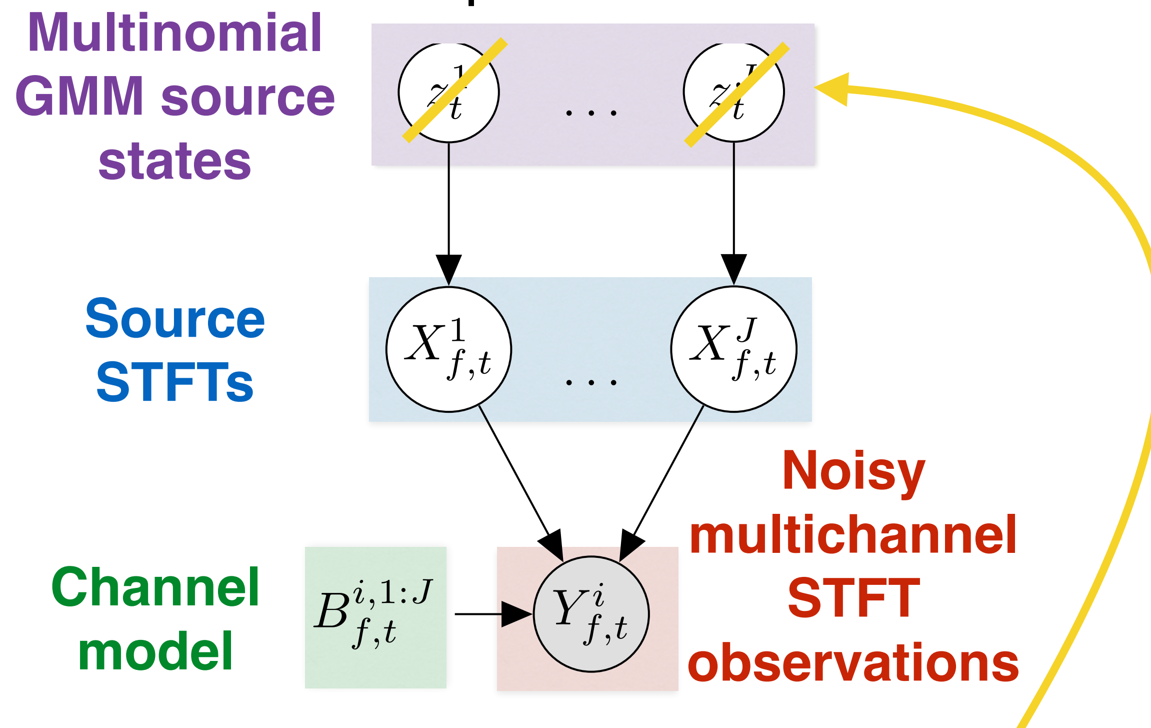


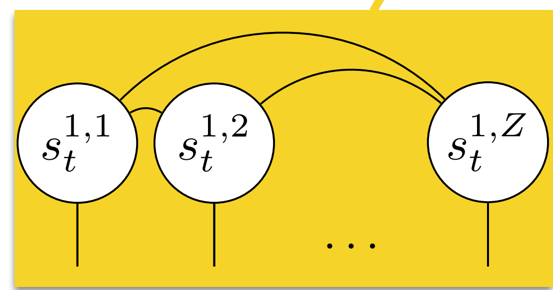
3. Unfolding the multichannel GMM

Physical scenario



Graphical model



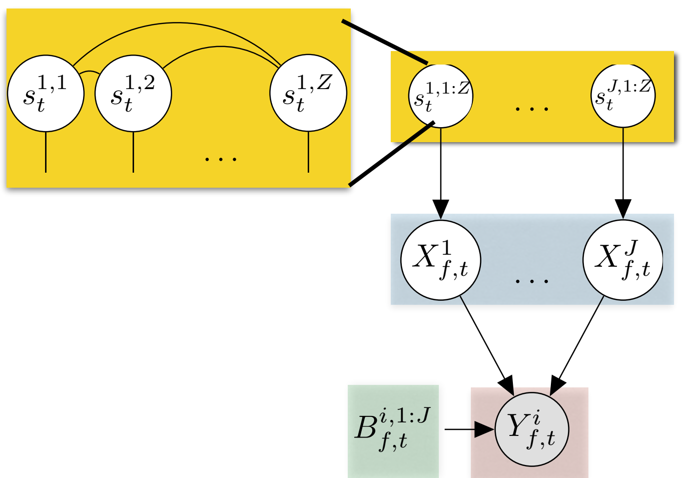


$$p(s = s^{1:Z}) = \exp\left\{s^T A s + c^T s\right\}$$

- Improve the generative model:
  Replace **multinomial source states** with
  one-hot binary Markov random field (MRF)

3. Unfolding the multichannel GMM

Graphical model

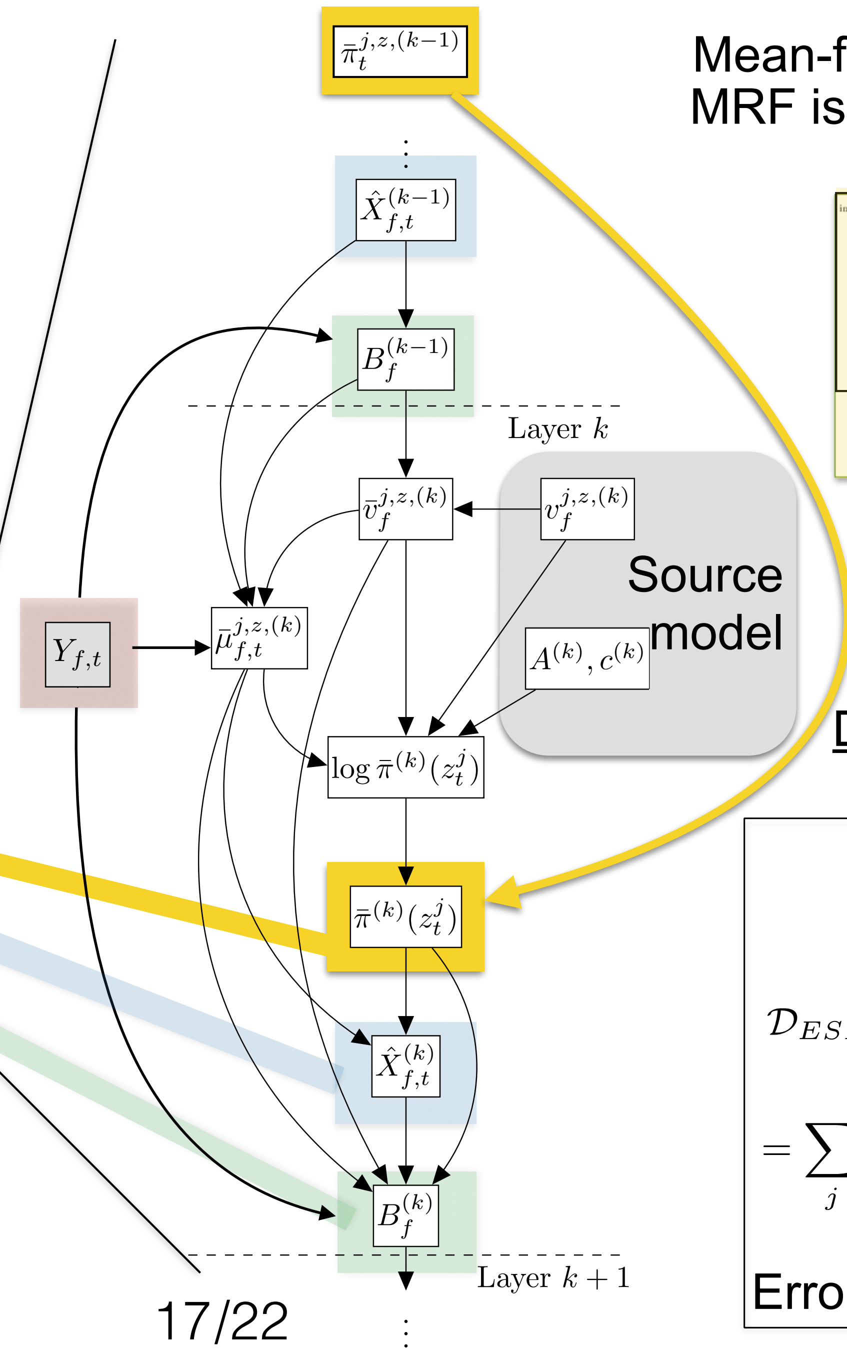


Iterative inference

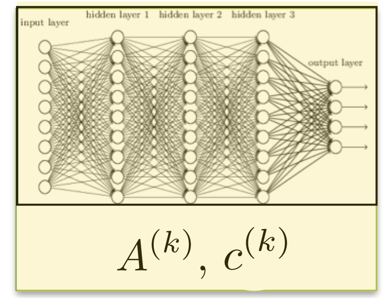
For k=1:K, estimate:

- Source state probabilities
- Source means
- Channel model

One layer of unfolded network



Mean-field inference in MRF is a deep sigmoid network



Discriminative training:

Optimize

$v^{(k)}, A^{(k)}, c^{(k)}$

to minimize

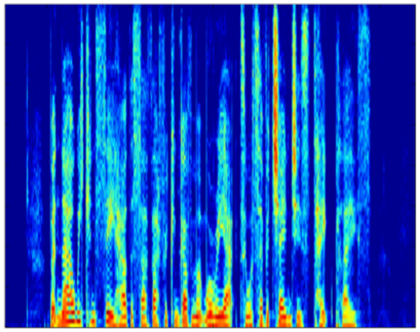
$$\mathcal{D}_{ESR}(X_{f,t}, \hat{X}^{(K)}_{f,t}) = \sum_j \frac{\sum_{f,t} \left|\hat{X}^j_{f,t} - X^j_{f,t}\right|^2}{\sum_{f,t} \left|X^j_{f,t}\right|^2}$$

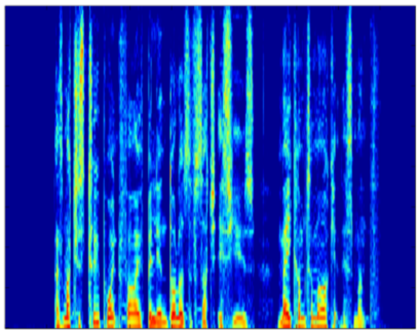
Error-to-source cost

Outline

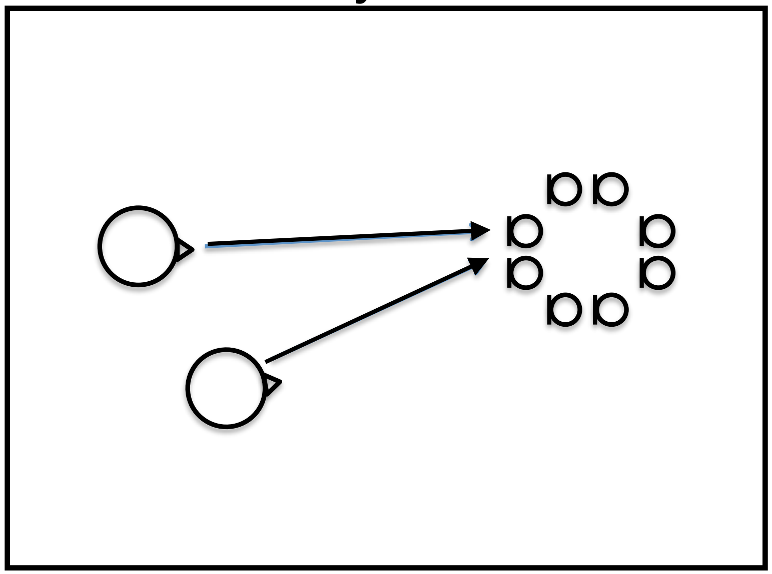
1. Deep unfolding overview
2. Generative model: multichannel GMM
3. Unfolding the multichannel GMM
4. **Results**

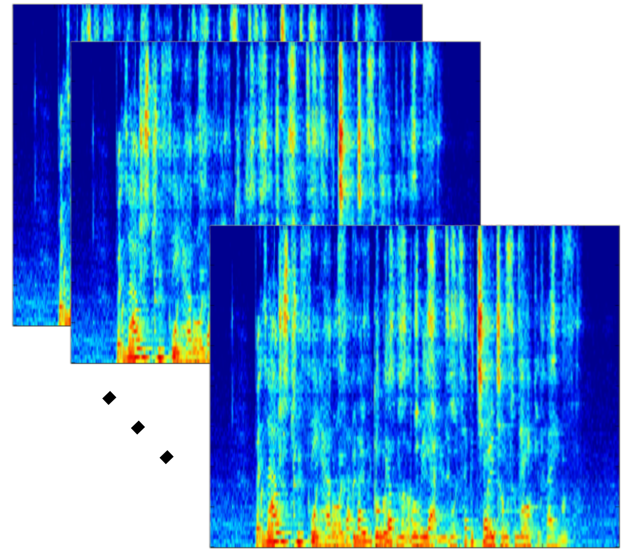
Dataset: overlapping* REVERB challenge [1]





8ch array scenario





- 20dB SNR stationary background noise.
- T60 times up to 700ms (realistic and hard!)
- Source 1 to source 2 power ratio between -15dB and +15dB.
- Training set: 15763 files of 6-10 seconds each, 6 different rooms.
- Validation set: 65 files of 6-10 seconds each, 3 different rooms.
- Evaluation set: 1435 files of 6-10 seconds each, 3 different rooms.

Thanks to Michael Mandel for generating this dataset during JSALT 2015 in Seattle

[1] Kinoshita et al. 2013

Implementation

- Everything implemented in Matlab using Bespoke Network Toolbox (BeNToBox) [1]

- Speaker- and gender-independent GMM source model trained with maximum-likelihood on WSJCAM0 [2]

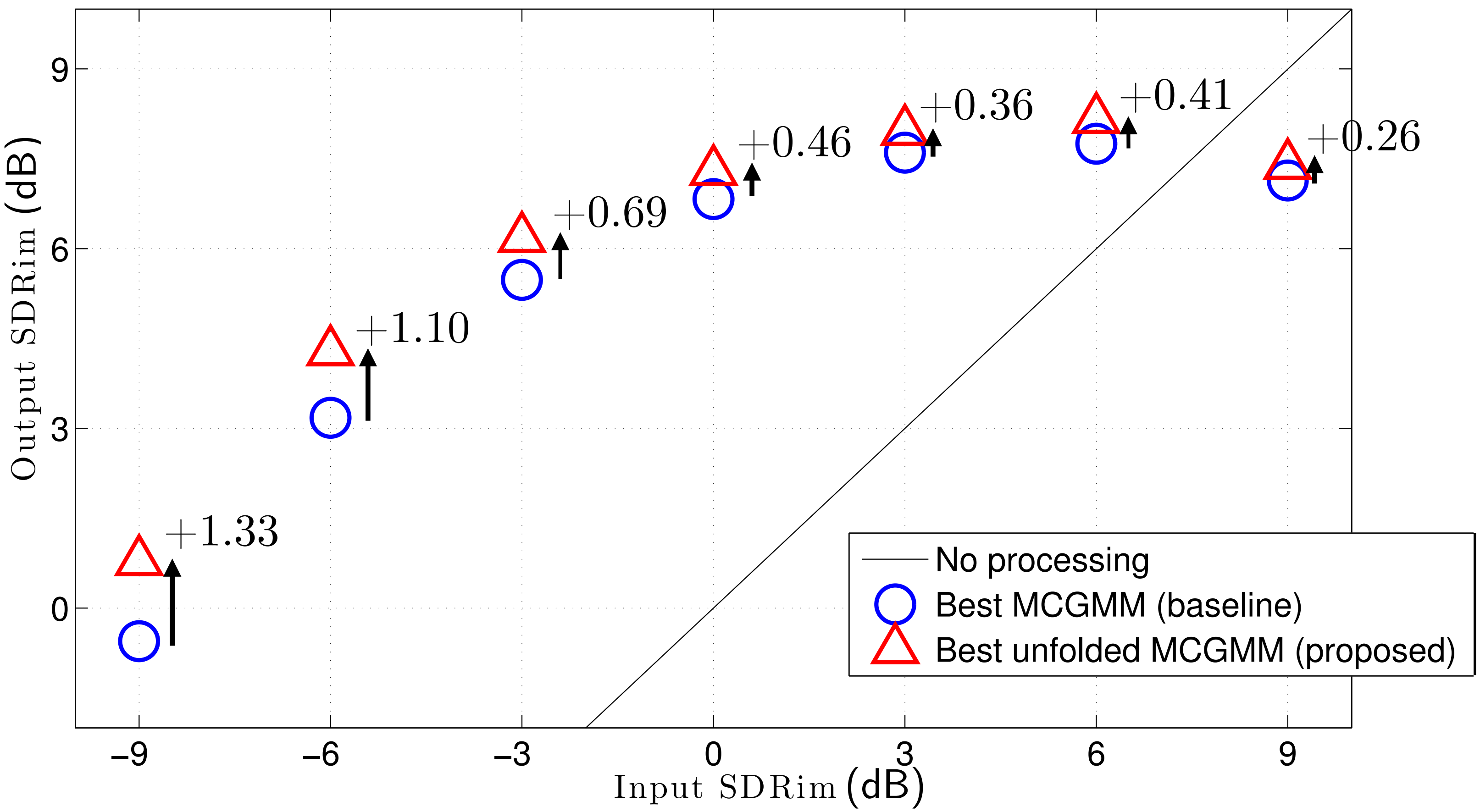
- “Warm up” with 10 untrained generative model layers

- Discriminative training: incremental layer-wise training on single GPU with stochastic gradient descent with momentum

[1] github.com/stwisdom/bentobox

[2] Robinson et al. 1995

Source-to-distortion ratio of source spatial images (SDR) [1]



# gen. layers + # unf.layers	Mean SDR (dB)
No proc.	-0.06
10+0	5.88
15+0	5.94
10+1	6.02
10+2	6.23
10+3	6.32
10+4	**6.37**

- **Baseline**: 10 or 15 MCGMM variational inference iterations
- **Proposed**:10 MCGMM iterations + *K* trained unfolded layers, (for *K*=1,2,3, or 4)

[1] `bss_eval_images` *by Vincent et al.*

Conclusion

Results:

- We used a new technique, deep unfolding, to convert variational inference for a generative model, the multichannel GMM (MCGMM) [1], into a deep network
- The resulting network has meaningful and interpretable activation functions and directly processes complex-valued multichannel frequency domain
- Improvements to the generative model manifest in the unfolded network
- Discriminative training improves performance over the original generative model

Future work:

- Integrate with ASR systems
- Recurrent and convolutional layers
- Unfold other generative models

[1] Attias 2003

Thank you!
Questions?

Code and supplementary materials:

http://www.merl.com/demos/deep-MCGMM