A Neural Network Alternative to Non-Negative Audio Models

PARIS SMARAGDIS[#*]

SHRIKANT VENKATARAMANI[#]

[#]UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

[*]ADOBE RESEARCH

ICASSP 2017

Introduction

- Supervised Single-channel Source Separation
  - Given a mixture of N sources

$$x(t) = \sum_i w_i s_i(t), \quad \text{where } w_i \in \mathbb{R} \text{ for } i = \{1, \ldots, N\}$$

  - Separate individual sources
  - Training data in the form of alternate unmixed recordings of the source.
- Non-negative Matrix Factorization (NMF)
- Objective: Develop a neural network alternative to NMF

Outline

- Non-negative Matrix Factorization (NMF)
- Non-negative Auto-encoder (NAE) equivalent to NMF
- Supervised source separation using NAE models
- Results

NMF

- NMF for matrices

$$\mathbf{X} = \mathbf{W}\mathbf{H} \qquad \mathbf{X} \in \mathbb{R}_{\geq 0}^{m \times n},\ \mathbf{W} \in \mathbb{R}_{\geq 0}^{m \times r},\ \mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times n},$$

- NMF is posed as a minimization problem

$$\begin{aligned} \underset{\mathbf{W},\mathbf{H}}{\text{minimize}} \quad & D(\mathbf{X}, \mathbf{W}\mathbf{H}) \\ \text{subject to} \quad & \mathbf{W} \geq 0, \mathbf{H} \geq 0. \end{aligned}$$

where ≥ 0 implies element-wise non-negativity

- Commonly used Cost functions

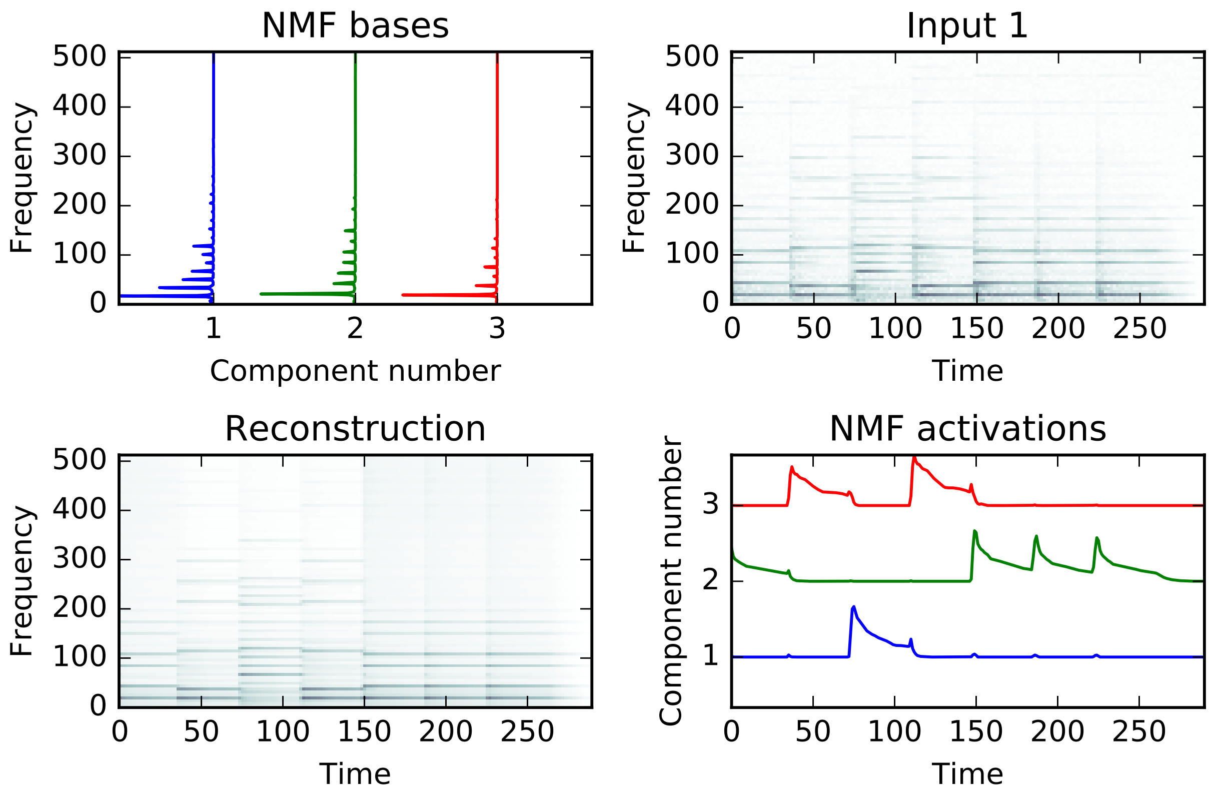
NMF: Piano example

- No cross-cancellations

- Part based decomposition

- Meaningful basis vectors.

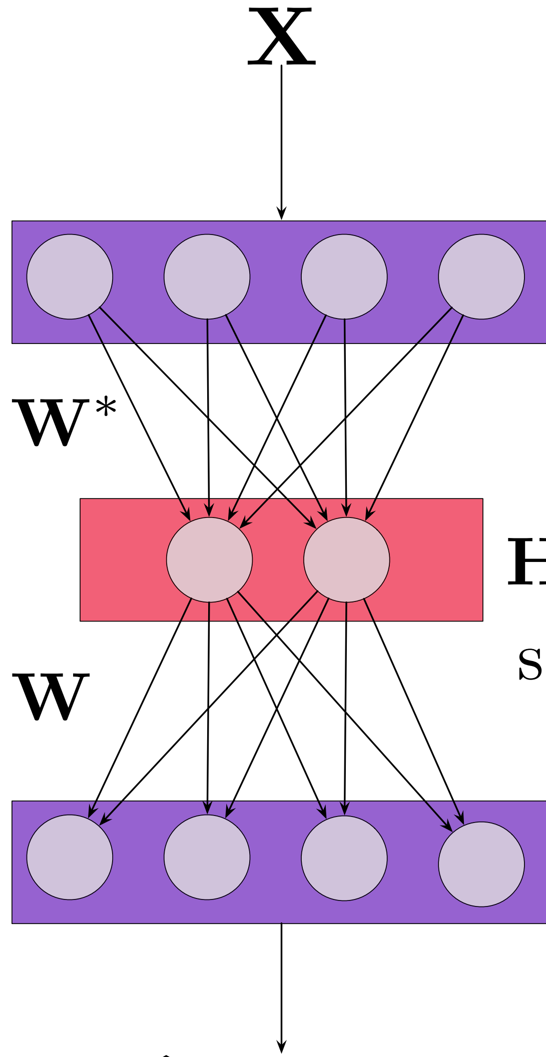
- Can be used as a model for supervised source separation.



$$\mathbf{D} = KL(\ \mathbf{X}\ ||\ \mathbf{WH}\ )$$

Towards an NMF neural network

- Autoencoder: Reconstructs the input at the output
  - Encoder: Input to Code
  - Decoder: Code to approximation of input

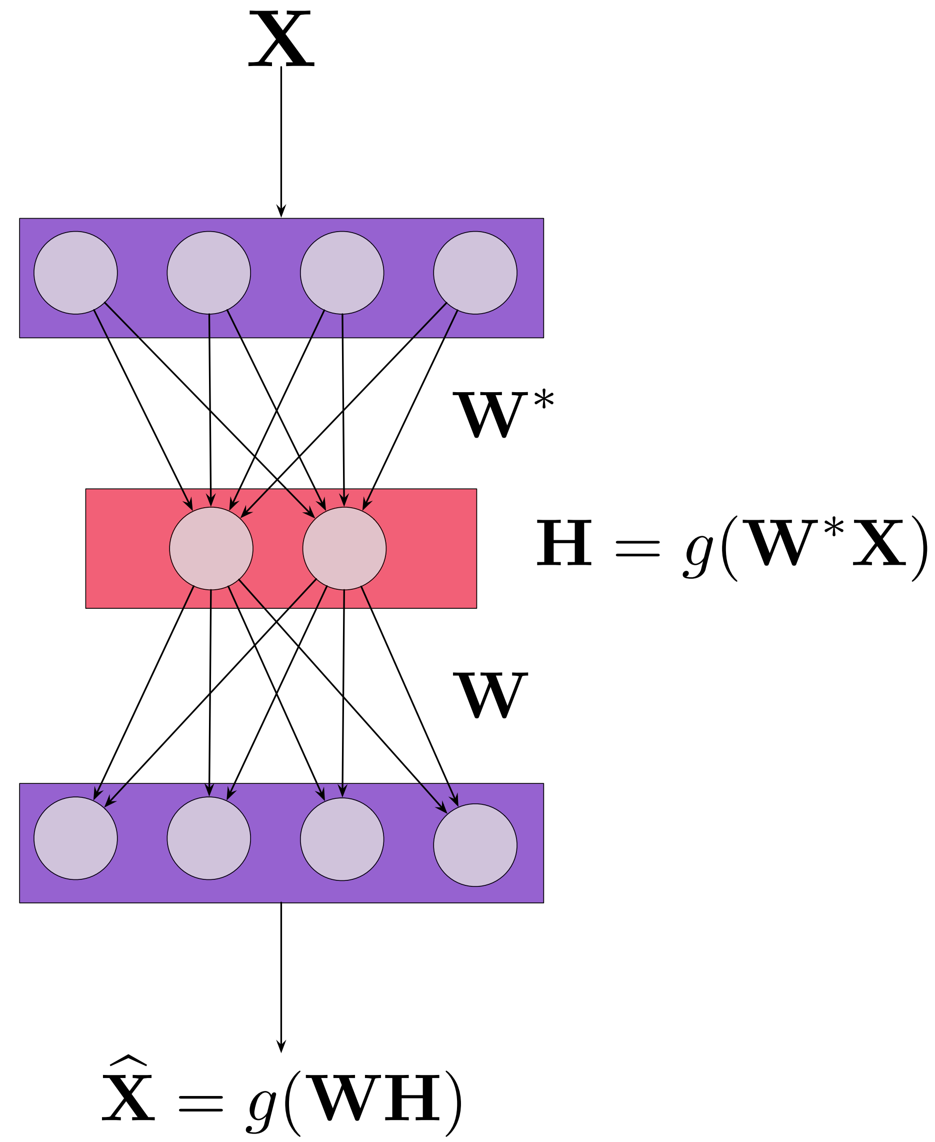


$$\mathbf{H} = \mathbf{W}^*\mathbf{X}$$
$$\text{s.t.}\,\mathbf{H} \geq 0$$

$$\widehat{\mathbf{X}} = \mathbf{W}\mathbf{H}\,\text{s.t.}\,\mathbf{W} \geq 0$$

Towards an NMF neural network

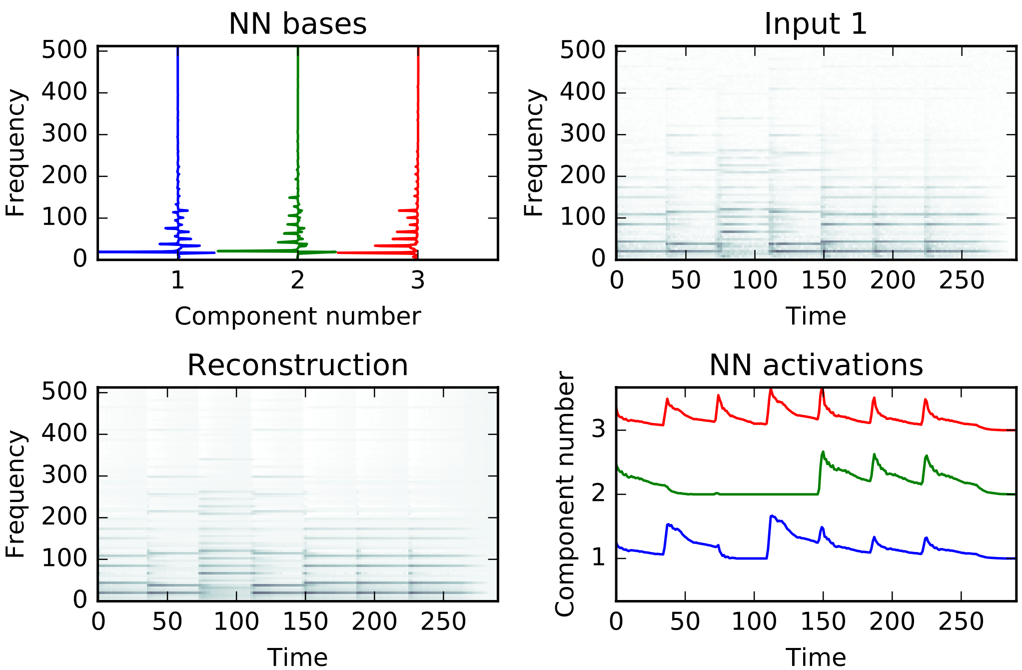
- Autoencoder: Reconstructs the input at the output
  - Encoder: Input to Code
  - Decoder: Code to approximation of input
- g(x) = max(x,0)
  or ln(1 + exp(x))
  or |x|
  mapping to the space of positive real nos.



Piano Example

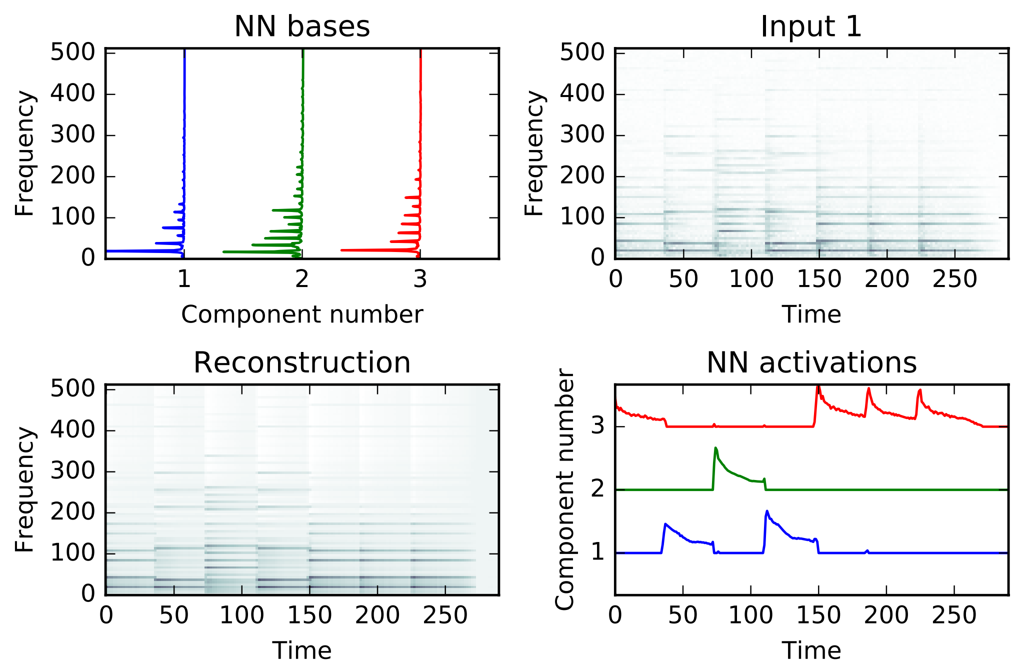
Without sparsity

$$\mathbf{D} = KL(\ \mathbf{X}\ ||\ g(\mathbf{WH})\ )$$



With Sparsity

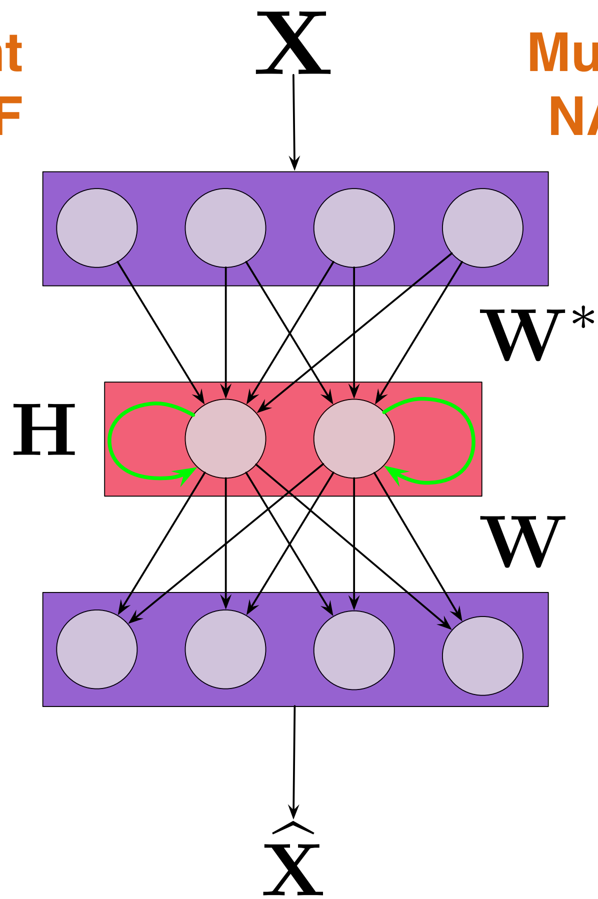
$$\mathbf{D} = KL(\ \mathbf{X}\ ||\ g(\mathbf{WH})\ ) + ||\mathbf{H}||_1$$



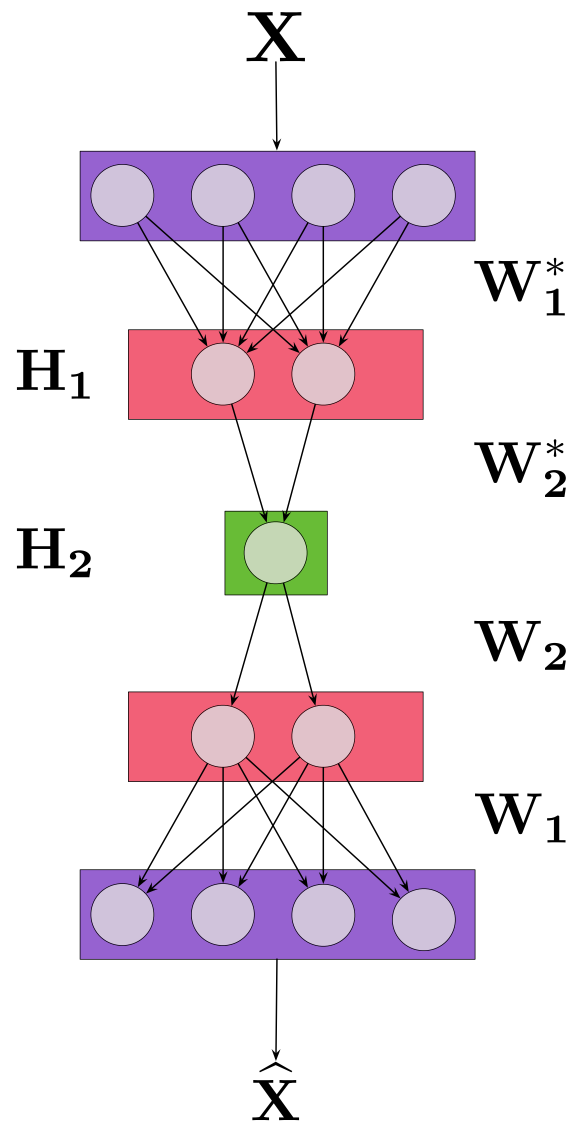
Why is this a good idea?

- Allows for several extensions over regular NMF

Recurrent NAE-NMF



Multi-layer NAE-NMF



Supervised source separation

- Learn representative bases for all the sources.
  - Autoencoder training on unmixed training examples gives representative matrices Ws and Wn.

- The spectrogram of the mixture is the sum of spectrograms of the sources.

$$\mathbf{X}_m = \mathbf{S} + \mathbf{N} = g(\mathbf{W}_s\mathbf{H}_s) + g(\mathbf{W}_n\mathbf{H}_n)$$

Thus,

$$\mathbf{X}_m^T = g(\mathbf{H}_s^T\mathbf{W}_s^T) + g(\mathbf{H}_n^T\mathbf{W}_n^T)$$

An output neural network with inputs: $\mathbf{W}_s^T, \mathbf{W}_n^T$ and output: $\mathbf{X}_m^T$

Supervised source separation

- Solve the minimization problem for Hs and Hn

$$\underset{\mathbf{H}_s,\mathbf{H}_n}{\text{minimize}} \quad KL(\ \mathbf{X}_m \ ||\ g(\mathbf{W}_s\mathbf{H}_s) + g(\mathbf{W}_n\mathbf{H}_n)\ )$$

Solved by training the output neural network

- Reconstruct the sources

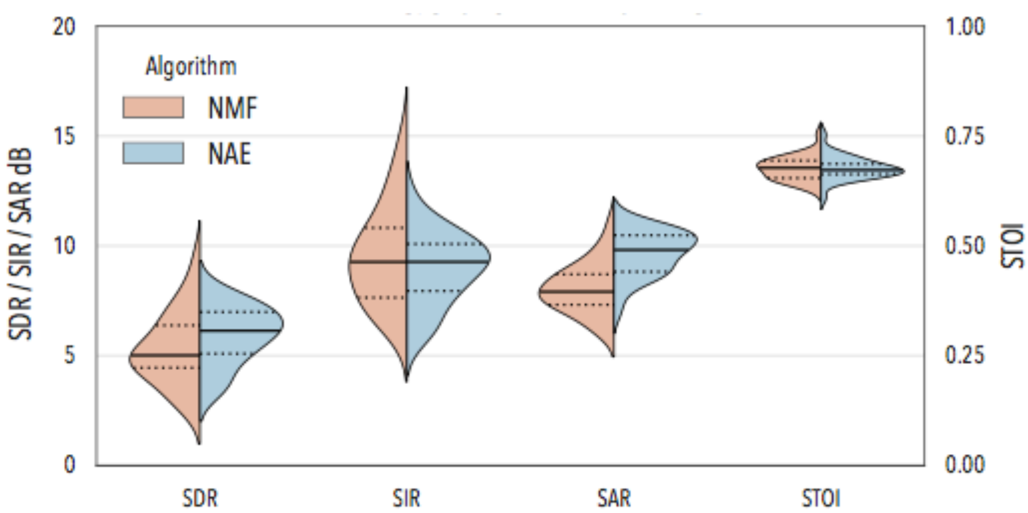
$$\hat{s}_i[n] = \text{STFT}^{-1}\left(\frac{g(\mathbf{W}_i\mathbf{H}_i)}{\sum_{i\in\{s,n\}} g(\mathbf{W}_i\mathbf{H}_i)} \odot \mathbf{X}_m \odot e_m^{j\cdot\mathbf{\Phi}_m}\right) \ \text{for } i \in \{s,n\}$$

where $\mathbf{\Phi}_m$ represents the phase of the mixture

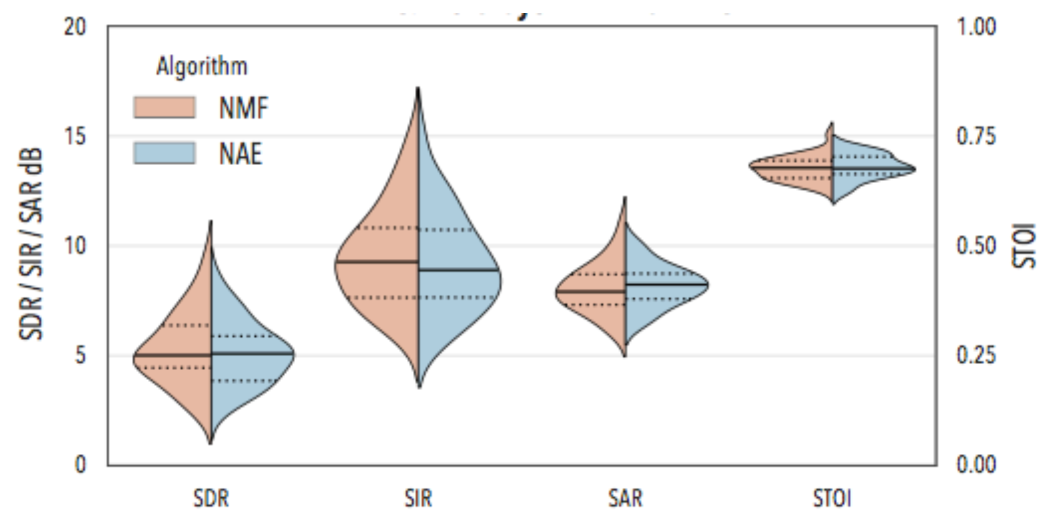
STFT^{-1} represents the overlap and add STFT operation

Results

NMF vs shallow- NAE
Rank = 20

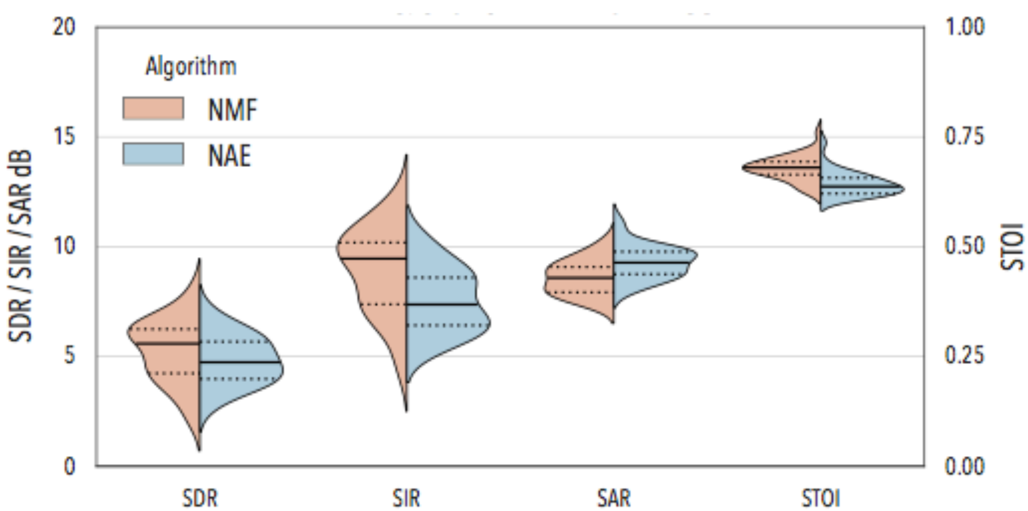


NMF vs multilayer NAE
Rank = 20

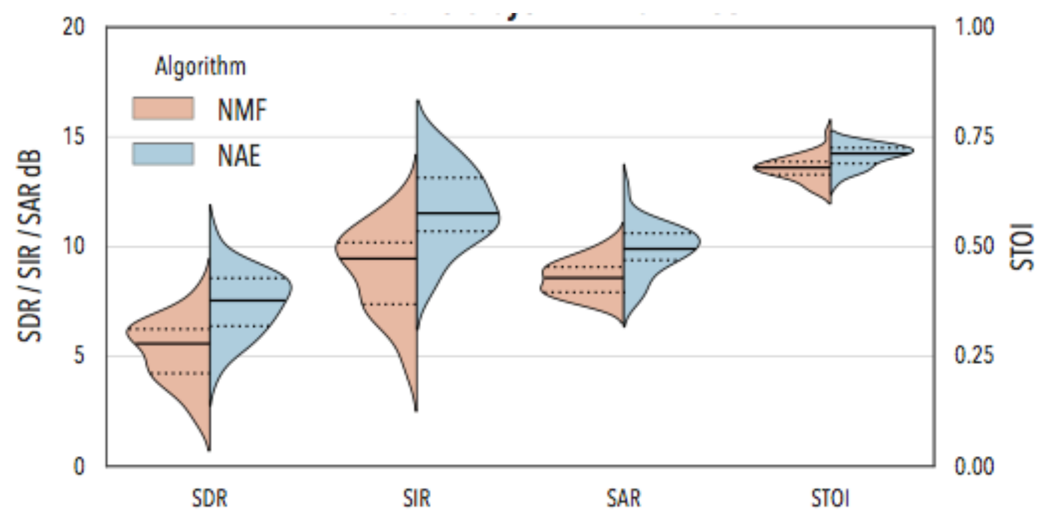


Results

NMF vs shallow- NAE
Rank = 100



NMF vs multilayer NAE
Rank = 100



Conclusion

- Non-negative Auto-encoder (NAE) audio models equivalent to NMF
  - Easily generalizable
- Separation Performance
  - Shallow NAE models equivalent to NMF
  - Multilayer NAE models outperform NMF by ~ 1.5 dB (SDR)
- Future work
  - Alternate neural net architectures for NAE
  - Towards an end-to-end neural net for source separation.

THANK YOU

Demo

	Ground truth	NMF (SDR = 6.05 dB)	Two layer NN (SDR = 5.4 dB)	Four layer NN (SDR = 7.1 dB)
Source 1 (Male)				
Source 2 (Female)