University of Illinois at Urbana-Champaign

Neural Network Alternatives to Convolutive Audio Models for Source Separation

Shrikant Venkataramani[*]

Cem Subakan[*]

Paris Smaragdis[#*]

[*]University of Illinois at Urbana Champaign

[#]Adobe Research

MLSP 2017

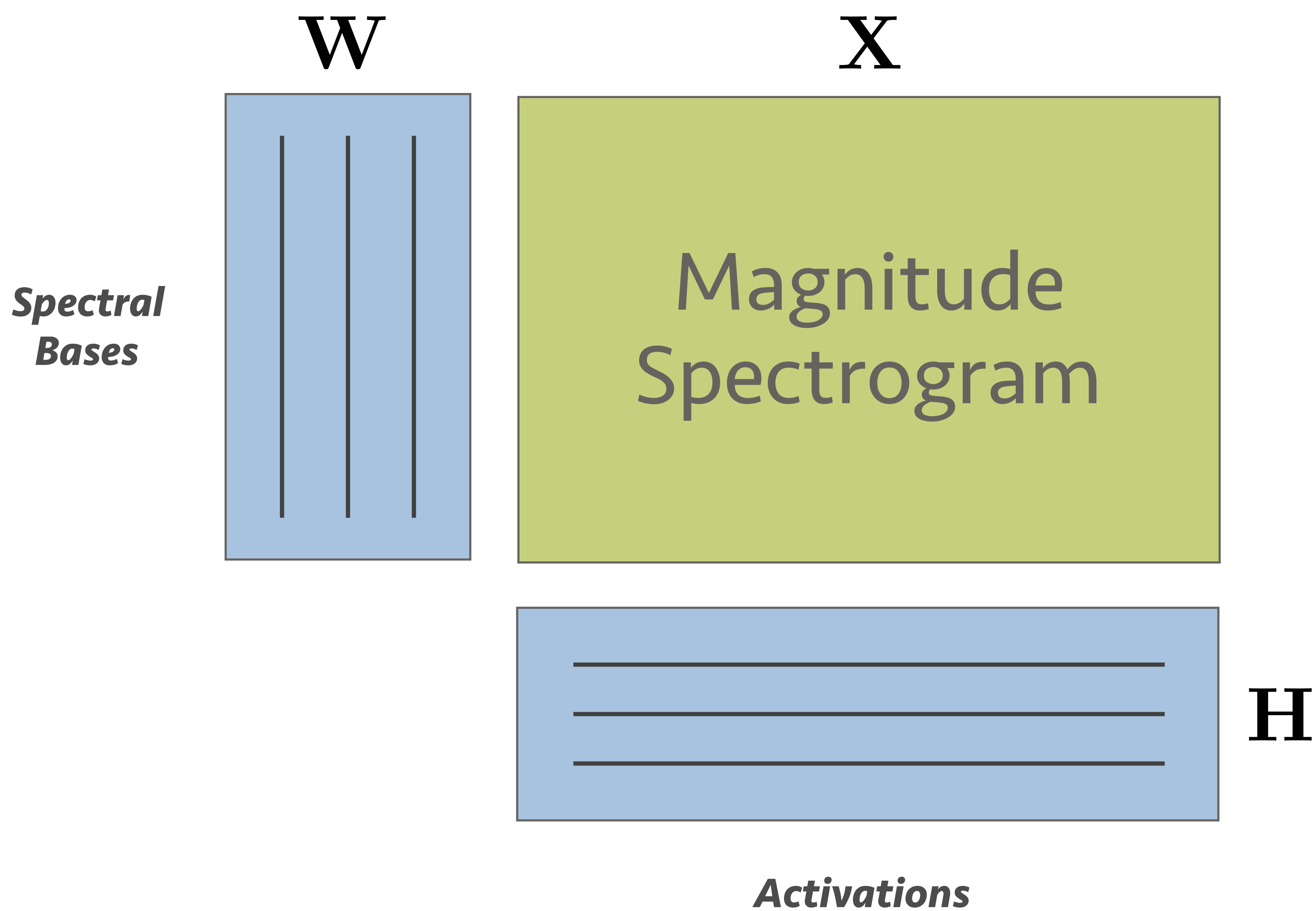
Motivation

- Supervised single channel source separation
  - Using models trained from clean sounds
- Dominant approach
  - Non-negative Matrix Factorization (NMF)
    - Interpretable, reusable models
- Non-negative Auto-encoder (NAE)
  - Interpreting NMF as a neural net
    - Reusable models with Significant improvements
- Modeling temporal dependencies in spectrograms
  - Incorporate temporal structure into NAE models
    - CNN’s, RNN’s, LSTM’s etc.

Learning an NMF model

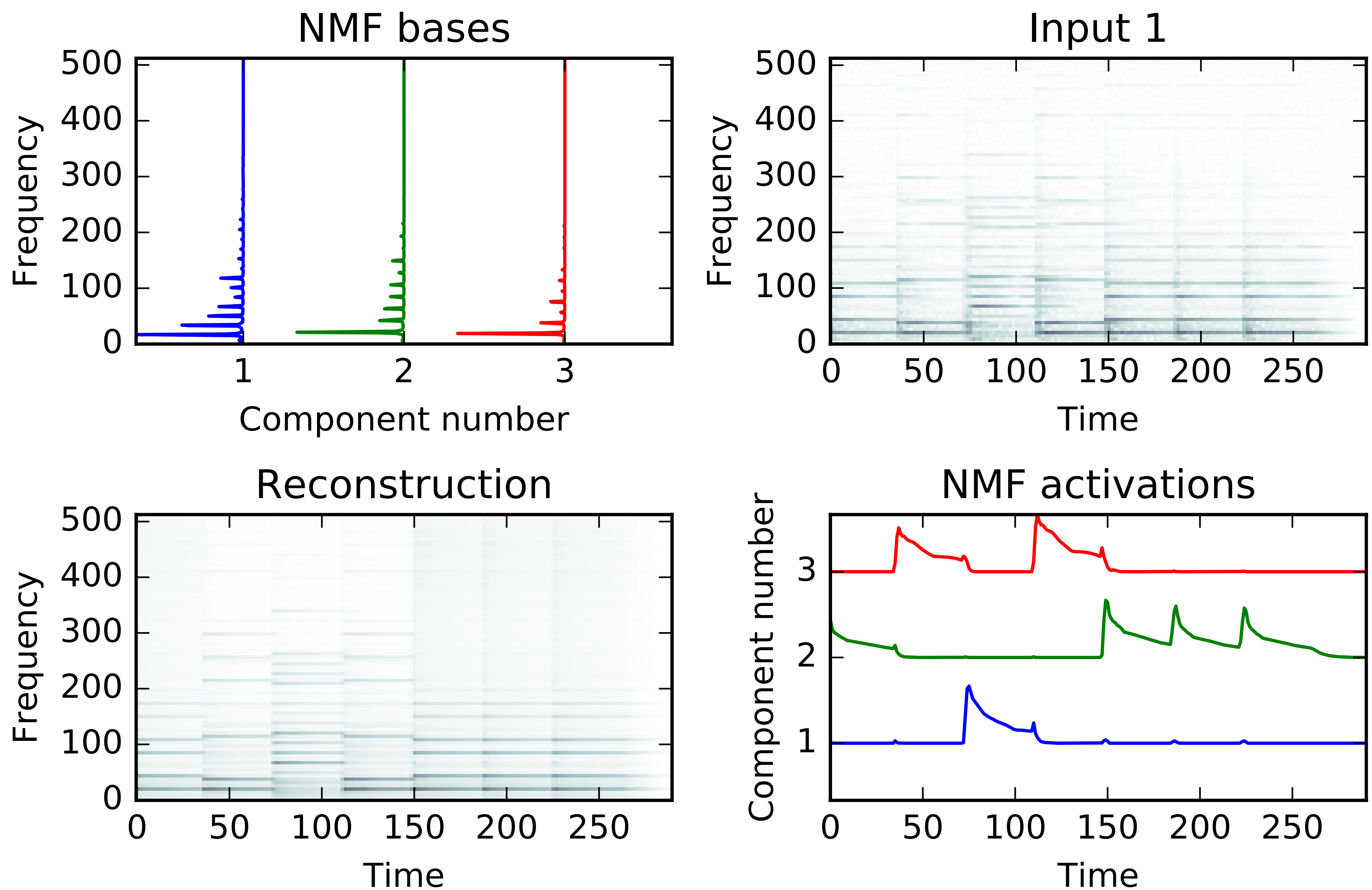
- Learning spectral bases from spectrograms.

$$\mathbf{X} = \mathbf{W} \cdot \mathbf{H} \qquad \mathbf{X}, \mathbf{W}, \mathbf{H} \in \mathbb{R}^{+}$$



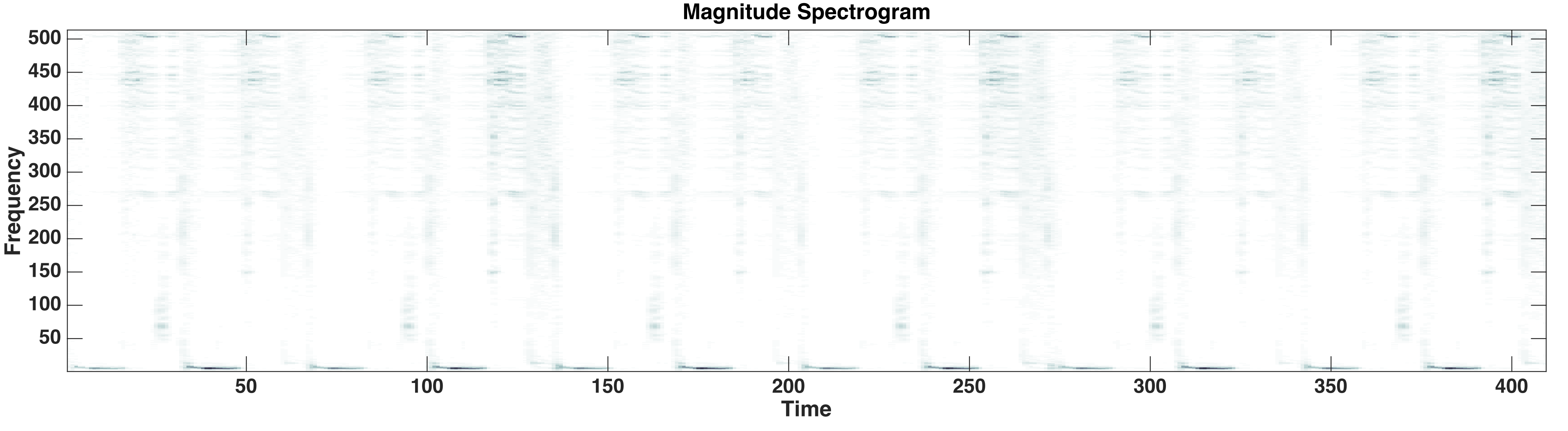
NMF in action

- Analyzing piano notes

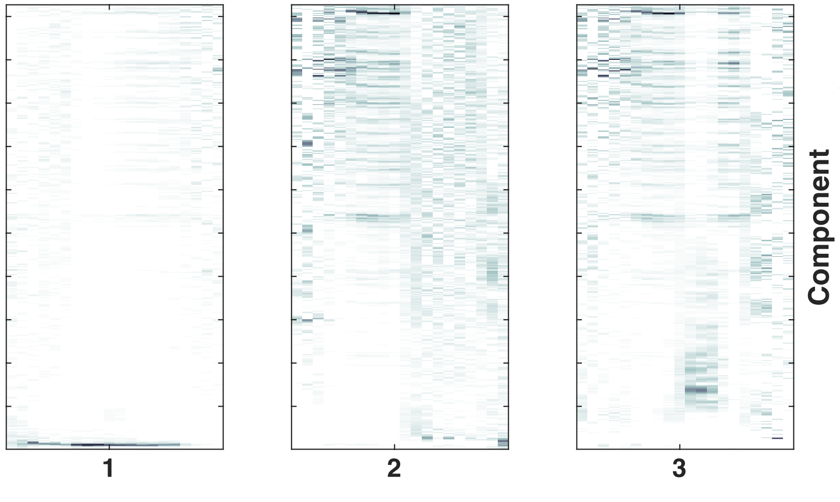


NMF for Non-stationary sounds

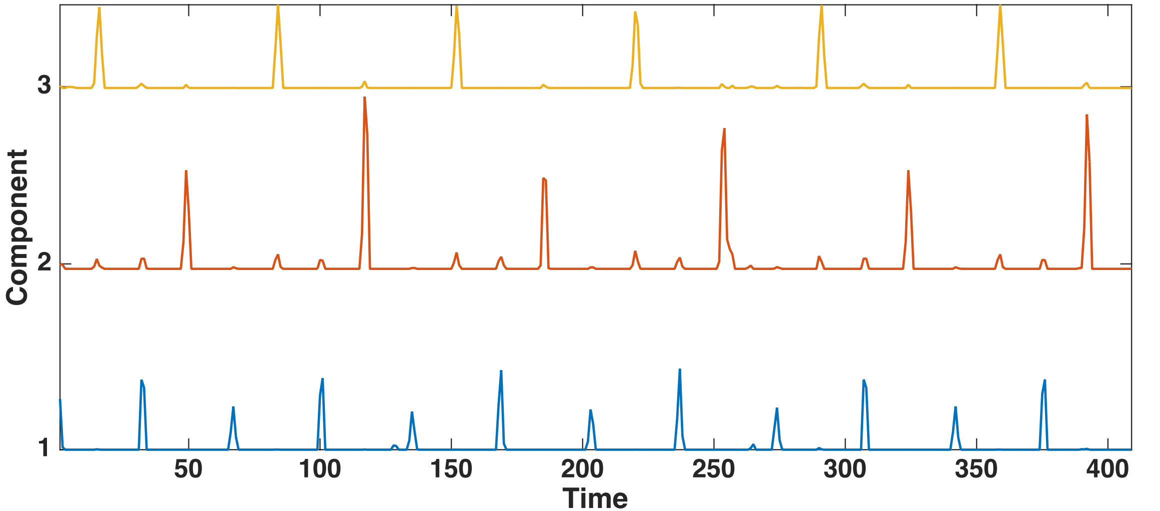
- Convolutive NMF
  - Modify spectral-bases to be matrices
    - Bases capture snippets from spectrogram
- Significant model changes
  - Difficult to model silences



Components



Activations



Non-negative Auto-encoder

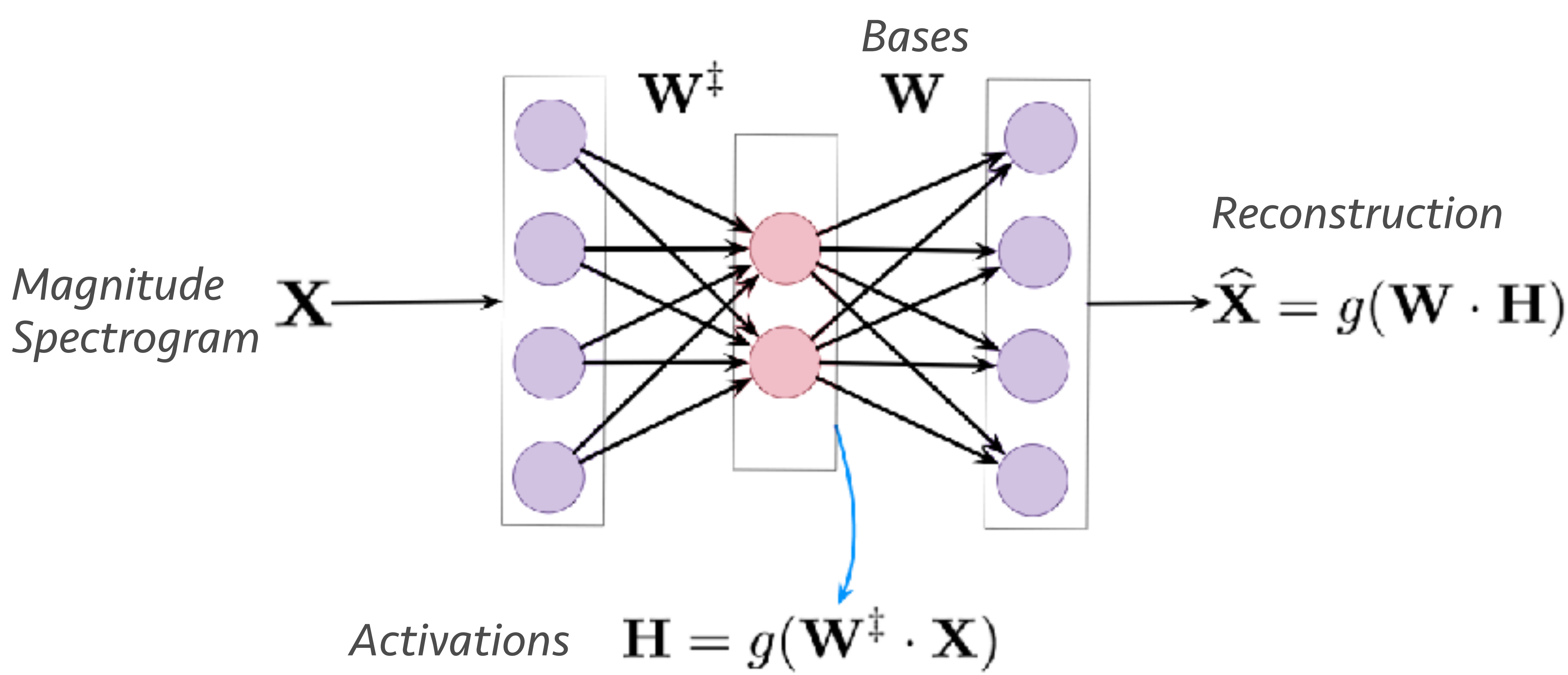
- Interpret NMF as a neural network

NMF — *Non-negative Auto-encoder (NAE)*

$$\mathbf{X} = \mathbf{W} \cdot \mathbf{H}$$

$$\mathbf{H} = g(\mathbf{W}^{\ddagger}\ \mathbf{X}) \ ; \quad \widehat{\mathbf{X}} = g(\mathbf{W}\ \mathbf{H})$$

$$g(x) = \max(x, 0) \text{ or } |x| \text{ or } ln(1 + e^{x})$$

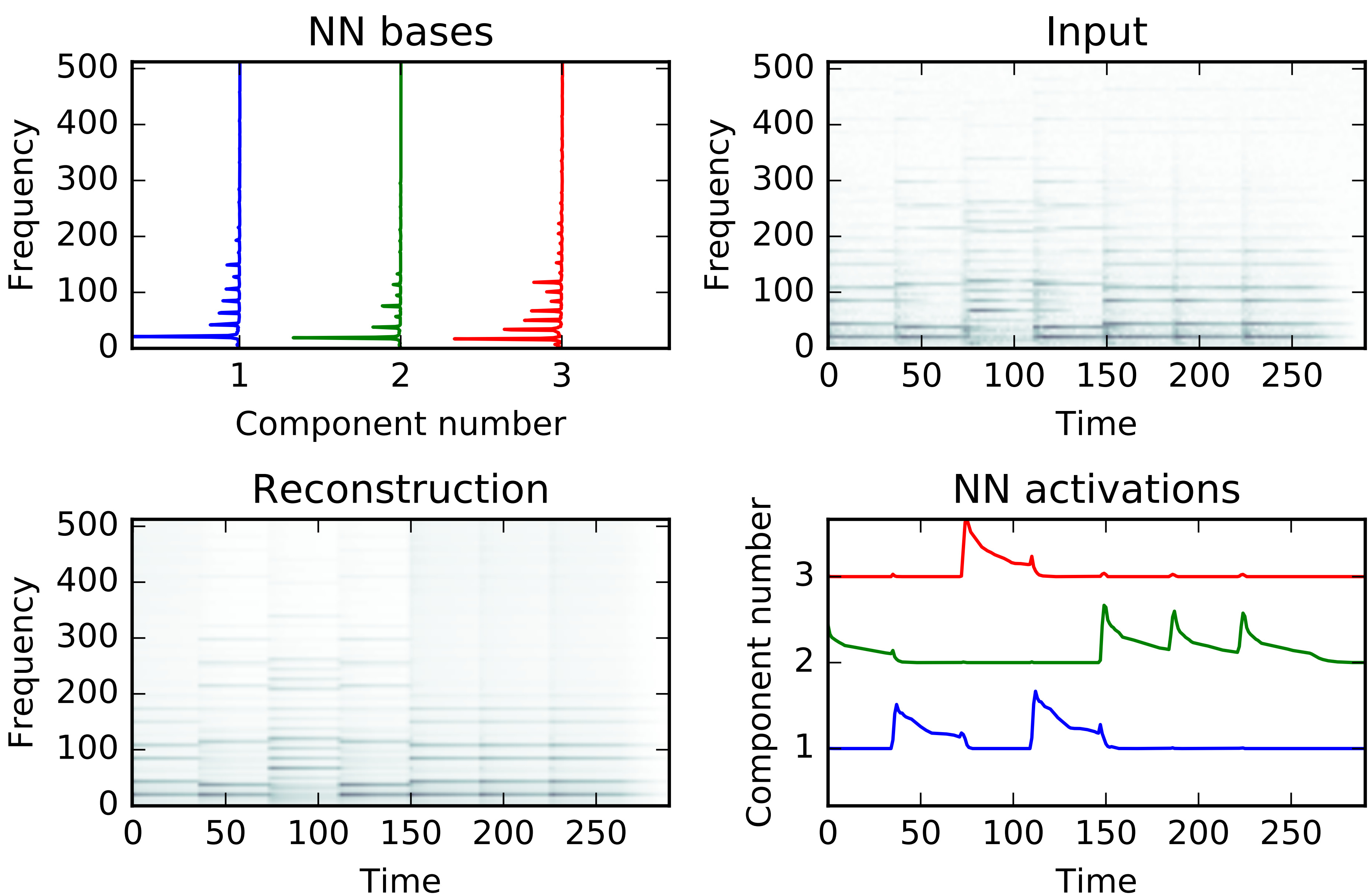


NAE in action

- Bases can take negative values

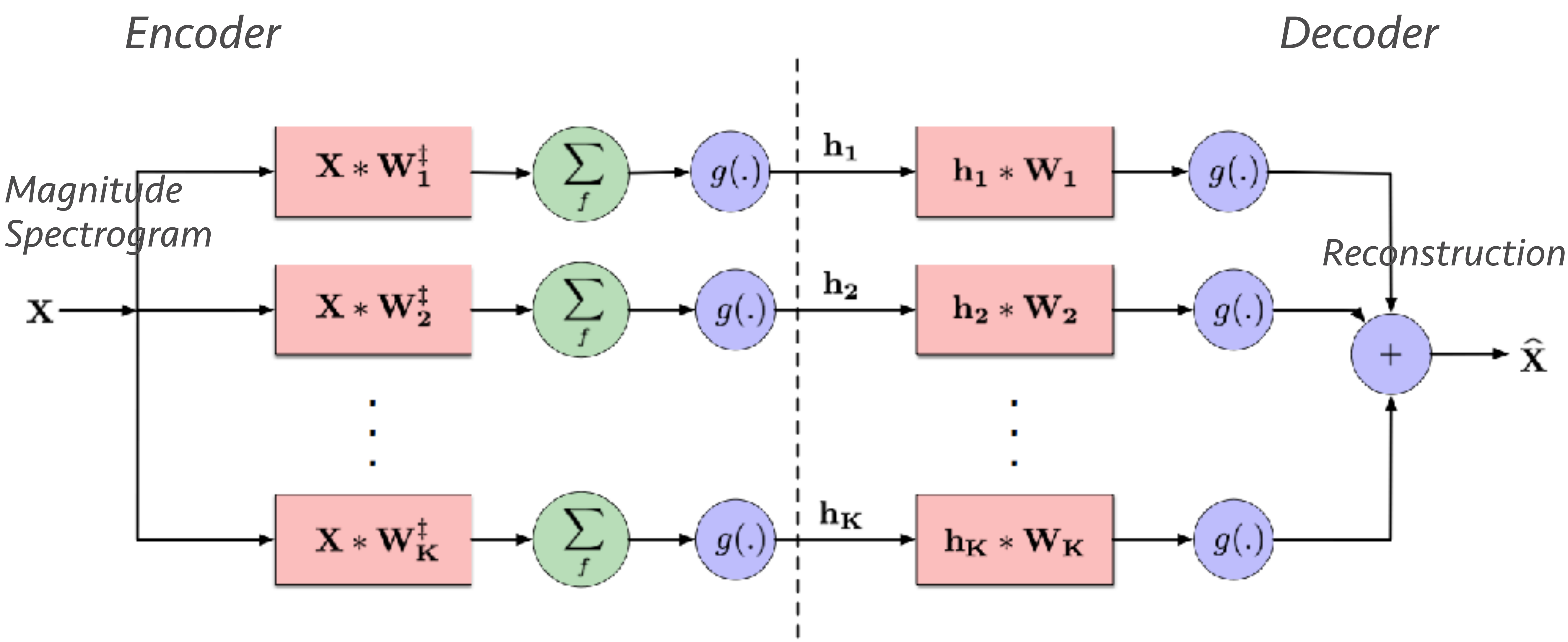
$$KL(\mathbf{X}||g(\mathbf{W}\cdot\mathbf{H}))+\lambda||\mathbf{H}||_1$$

$$g(x)=ln(1+e^x)$$



Convolutive models

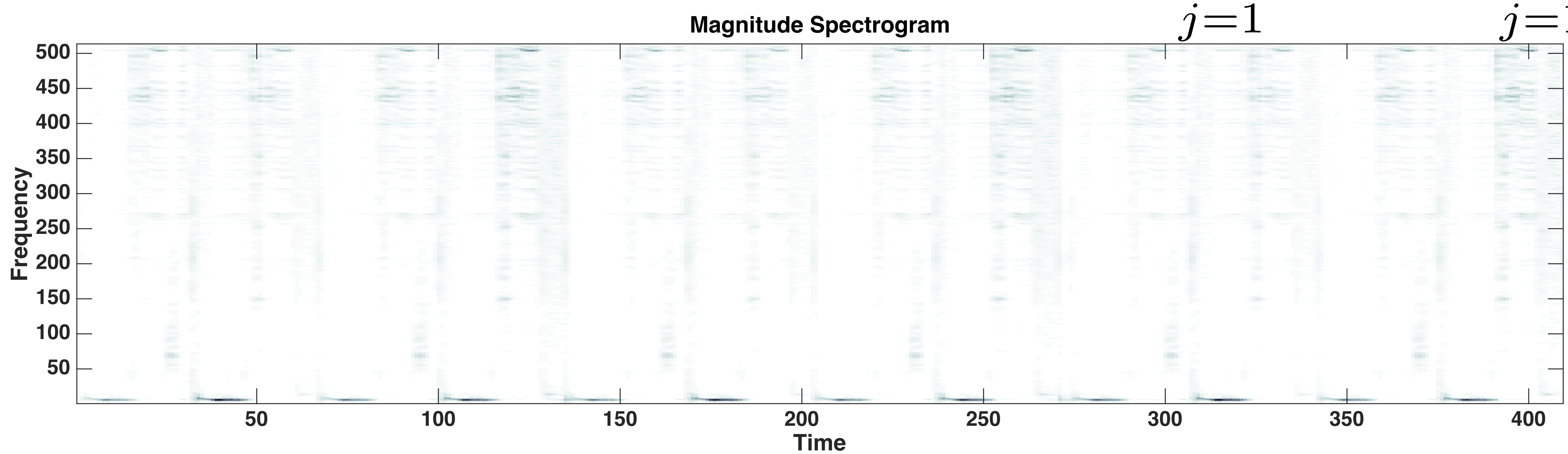
- Cross-frame patterns in spectrograms
  - CNN's naturally deal with sequences
    - Spectro-Temporal models
- CNN-CNN auto-encoder (CCAE)



CCAE in action

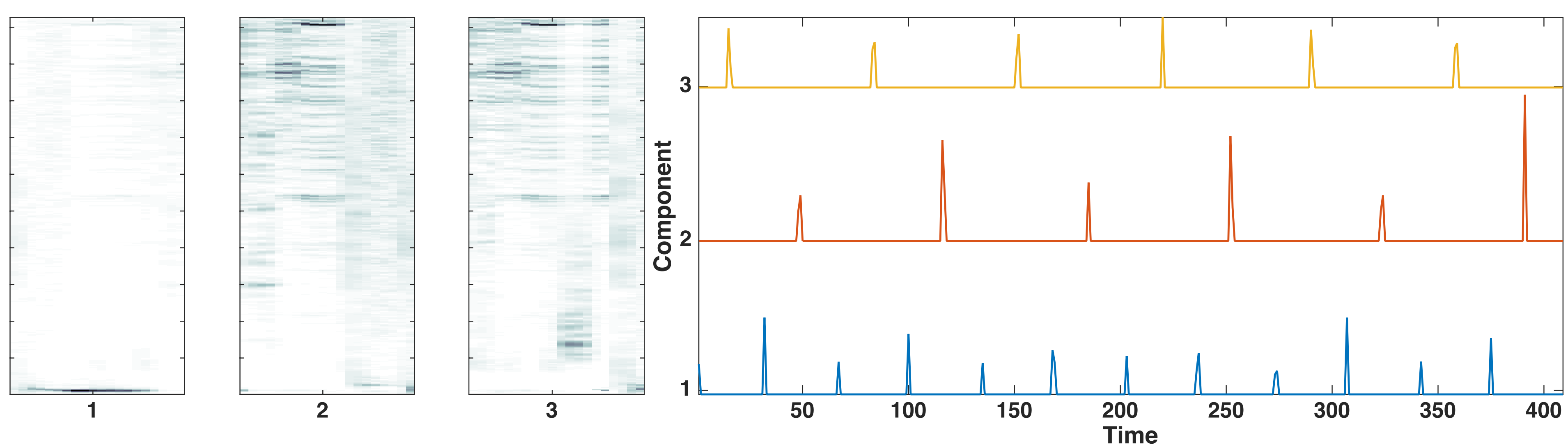
- Encoder acts as a matched filter
  - Bases allow negative values
    - Models silence easily

$$g(x) = ln(1 + e^x) \qquad KL(\mathbf{X}||\widehat{\mathbf{X}}) + \lambda \sum_{j=1}^{K} |\mathbf{h_j}| + \mu \sum_{j=1}^{K} ||\mathbf{W_j}||_1$$



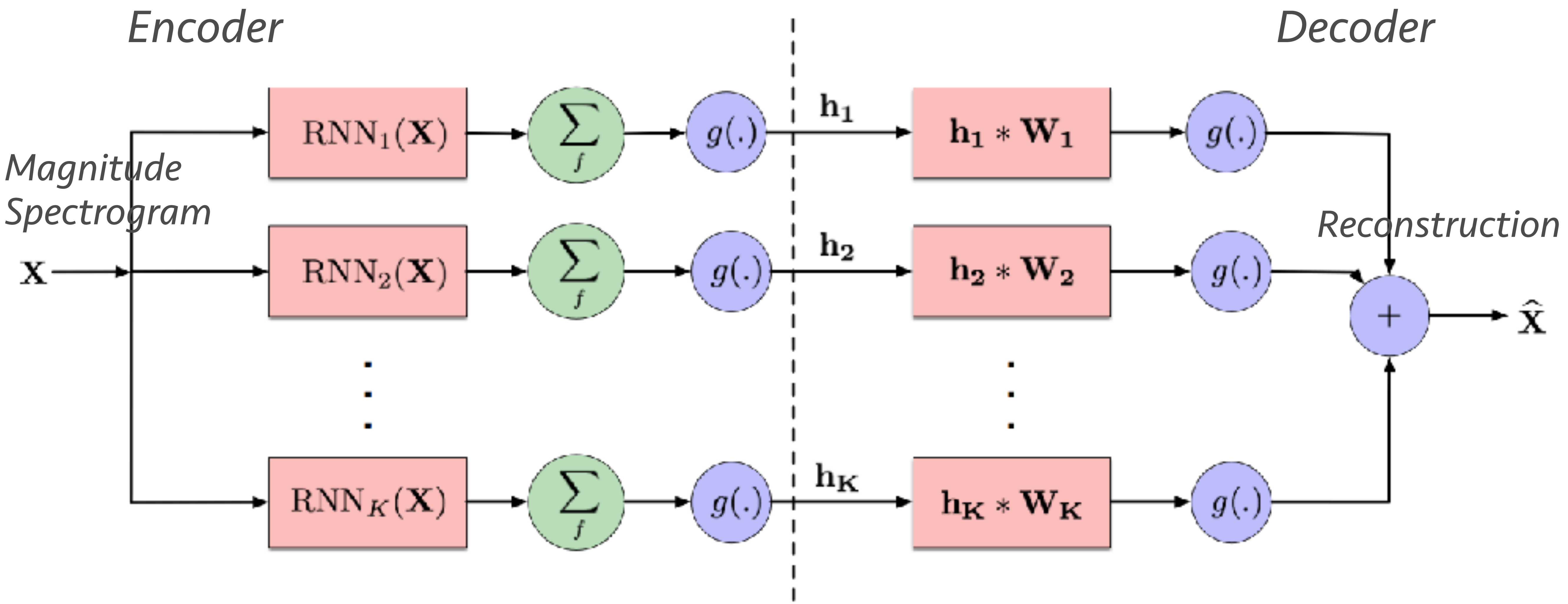
Components

Activations



Extensions

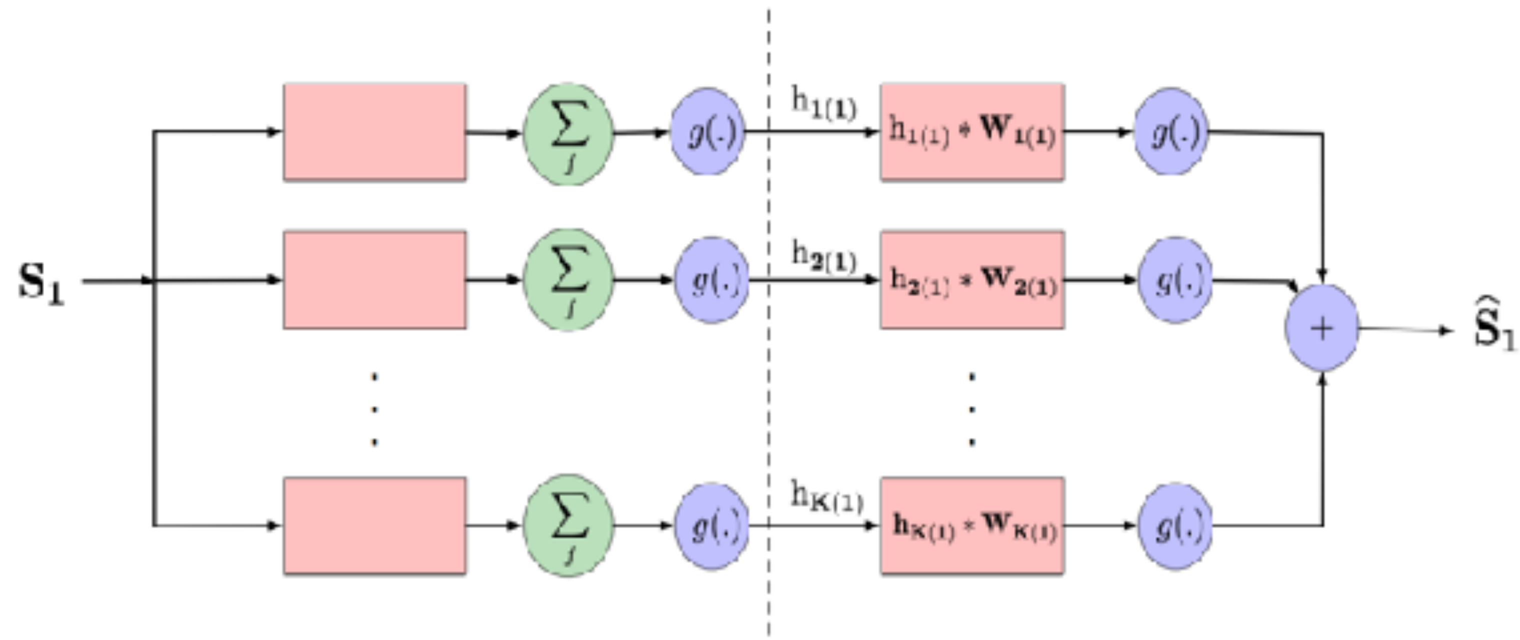
- Difficult to extend NMF models
  - Easy to extend neural nets
- Encoder acts as a matched filter
  - Inverse of FIR is IIR
    - Use RNN's (LSTM's) in the encoder
- RNN-CNN auto-encoder (RCAE)



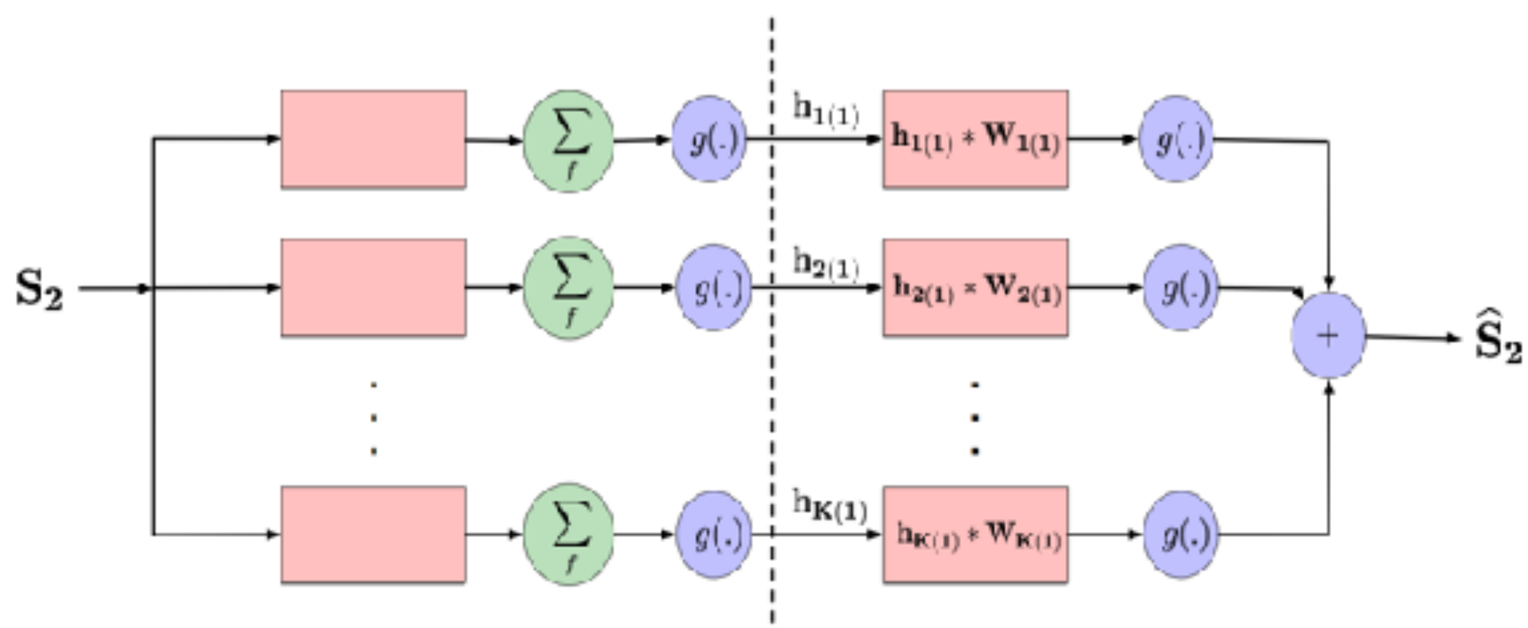
CAE Source separation

- Estimate models for each source

Training data for source 1

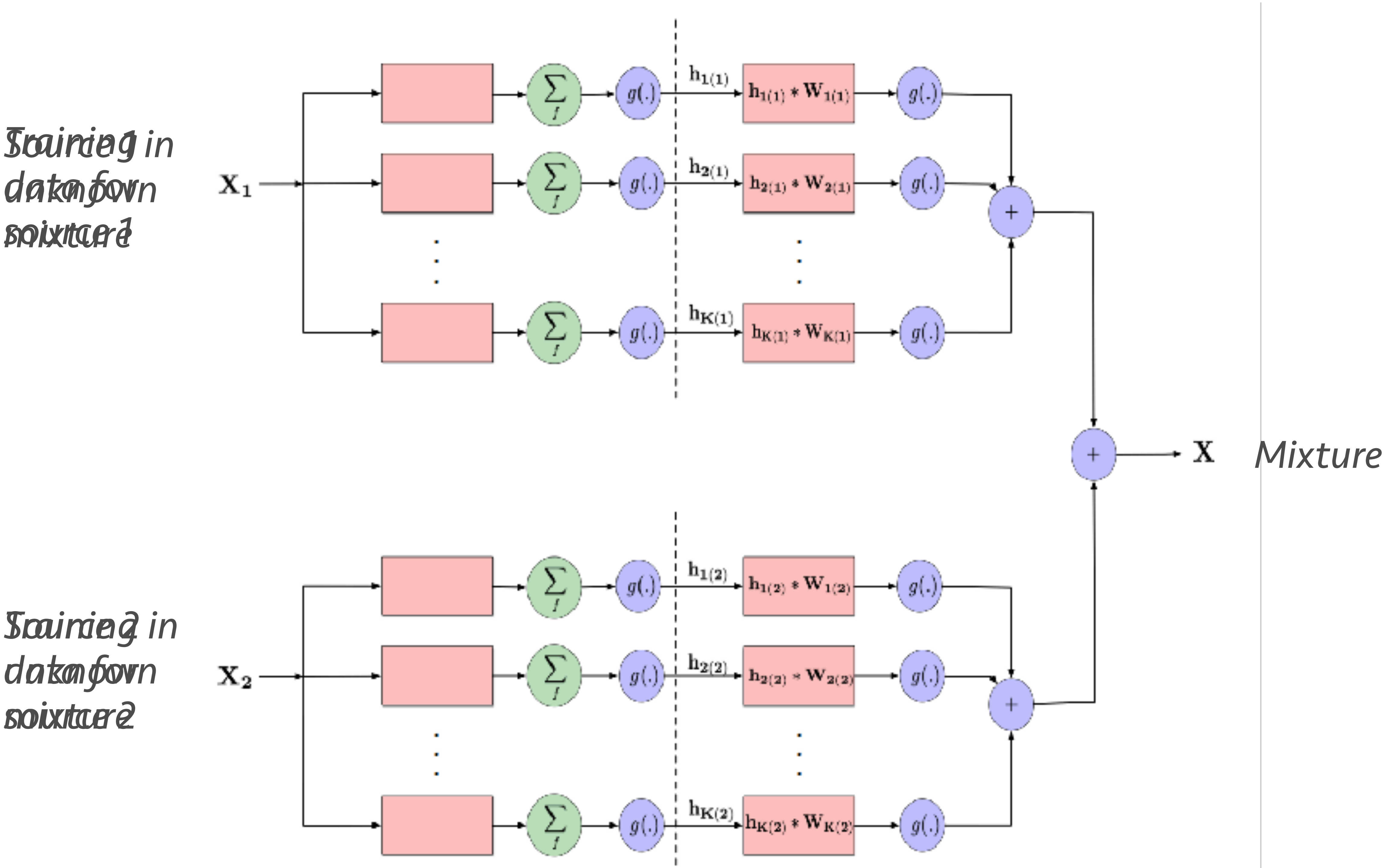


Training data for source 2



CAE Source separation

- Estimate the contribution of the sources in the unknown mixture using trained models



CAE Source separation

- Goal: Estimating network inputs (source spectrograms)
  - Given the source models

$$\mathbf{X} = \mathbf{X_1} + \mathbf{X_2}$$

  - Gradient-descent/back-propagation to train the network

- Spectrograms to sources
  - Inversion using mixture phase

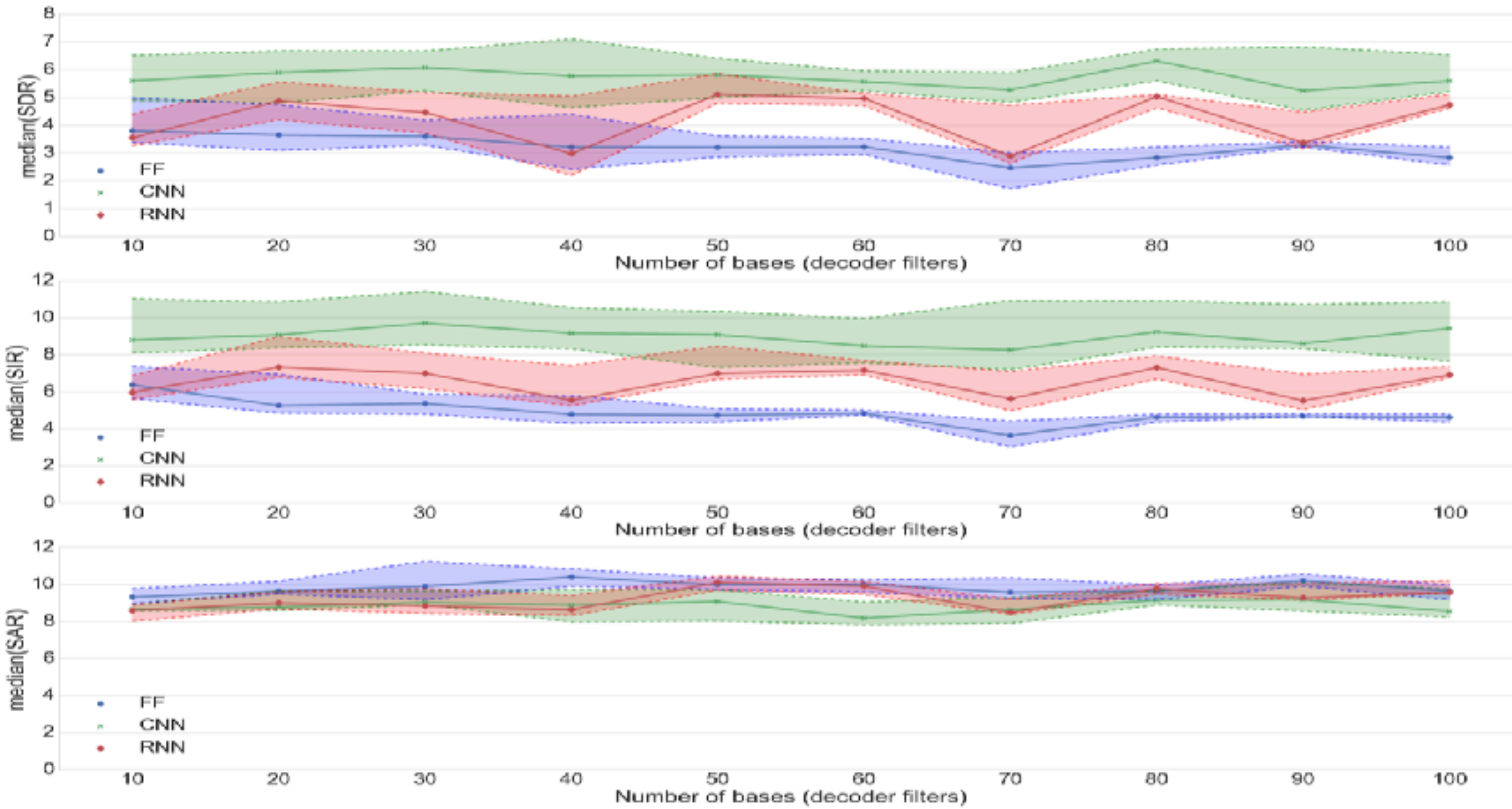
$$x_i(t) = \mathrm{STFT}^{-1}\left(\frac{\mathbf{X_i}}{\sum_i \mathbf{X}_i} \odot \mathbf{X} \odot e^{i\Phi_m}\right) \text{ for } i \in \{1, 2\}$$

Evaluation

- Two-speaker mixtures
  - Training data ~ 15-20 seconds
  - Test data: Single sentence mixture at 0 dB
  - Evaluated for 10 pairs of speakers
- Evaluation metrics
  - BSS_eval metrics (SDR, SIR, SAR)
- Compared CCAE and RCAE versions
  - NAE models as baseline
  - Parameters
    - Decomposition rank

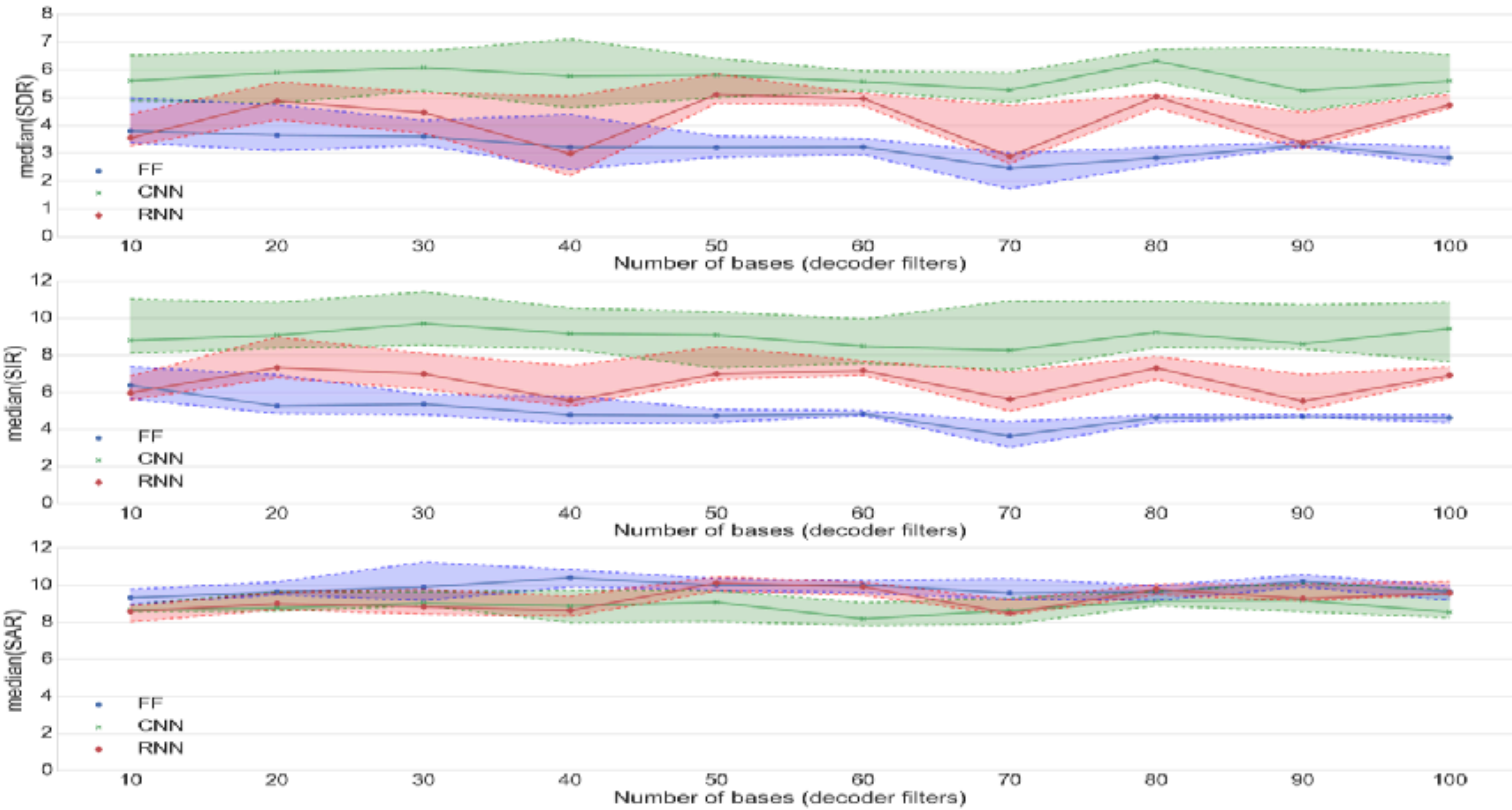
Separation Results

- NAE vs CCAE vs RCAE
  - Filter width = 8 samples
  - Filter height = 512 samples
- Best performance setting: K = 80



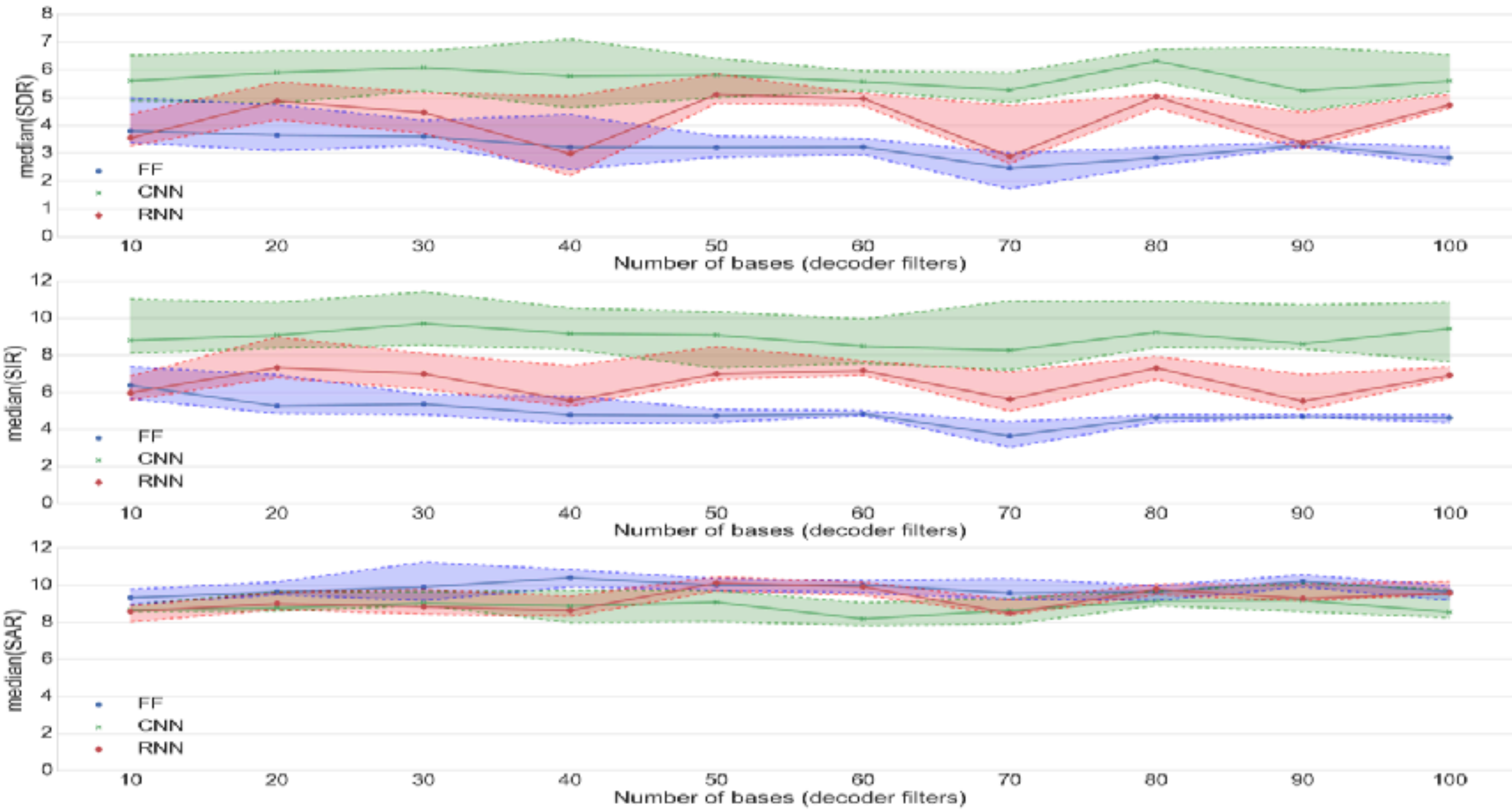
Separation Results

- CCAE models are significantly better
  - Inter-quartile range is higher
- Significant improvement in SIR
  - SAR values comparable



Separation Results

- Median performance almost constant
  - CCAE models are robust to choice of decomposition rank
- RCAE models better than NAE models
  - Not as good as CCAE models



Conclusions

- An alternative to convolutive basis decompositions
  - CNN's allow network to learn spectro-temporal patterns
- CAE models superior to NAE models
  - Significant improvement in separation performance
- Easily generalizable to novel convolutive models and architectures
  - RCAE models and other possible extensions
- Code available on GitHub
  - https://github.com/ycemsubakan/sourceseparation_nn

THANK YOU