DEEP LEARNING FOR MONAURAL SPEECH SEPARATION

Po-Sen Huang, Minje Kim,

Mark Hasegawa-Johnson, Paris Smaragdis

ILLINOIS

Motivation

- Source separation is important for several real-world applications
  - Monaural speech separation is more difficult
- Previous approaches – linear models
  - Non-negative matrix factorization (NMF), probabilistic latent semantic indexing (PLSI)
  - Similar to a one layer linear network with non-negative weights and coefficients
- Representation
  - NMF based models – spectral representation
  - Deep learning models – learning optimal representation
- ➢ Explore deep learning models

Overview

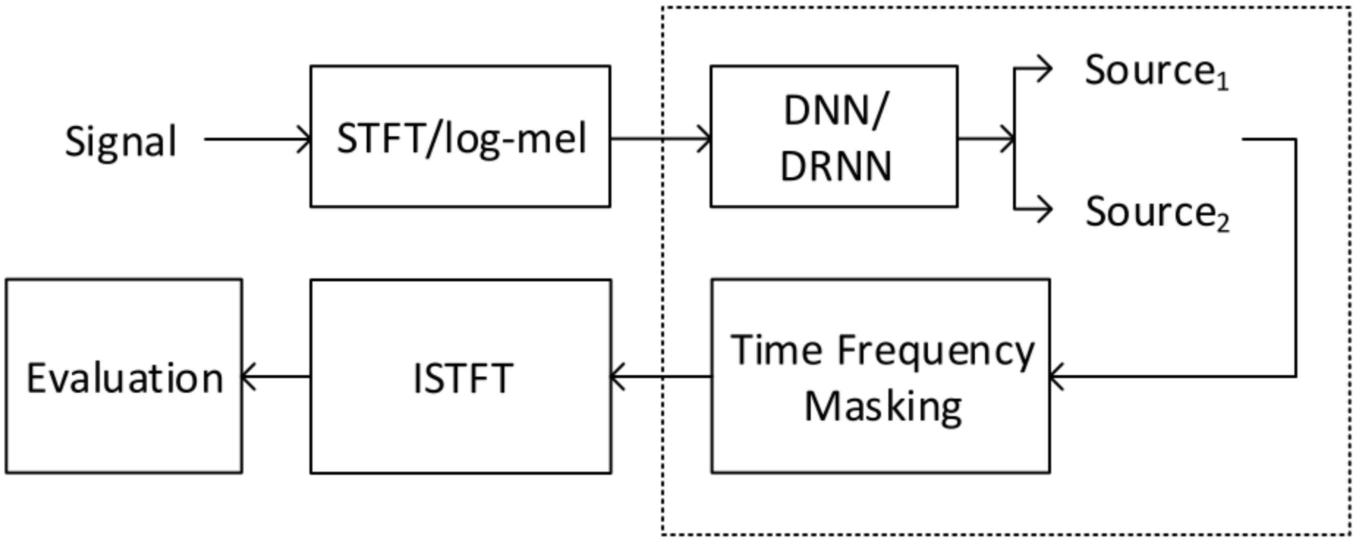
- Previous Work
- Proposed Framework
- Proposed methods
  - Model architecture
  - Joint time frequency masking
  - Discriminative training
- Experiment
- Demo
- Conclusion

Previous Work – Deep Learning

- Two-stage framework for predicting ideal binary mask [Narayanan and Wang, 2013]
  - First stage: Train one DNN per output dimension
  - Second stage: train another one layer perceptron or SVM for refinement
  - ➢ Impractical for high dimensional output
- Robust ASR [Mass et al. 2012]
  - Given noisy speech, train DRNN to predict clean speech
  - ➢ Suboptimal in the source separation scenario to model only one source

Proposed Framework

- Given spectral or log-mel features, use DNN or DRNN to predict spectral targets (multiple sources)
- Apply time-frequency masking
- ISTFT

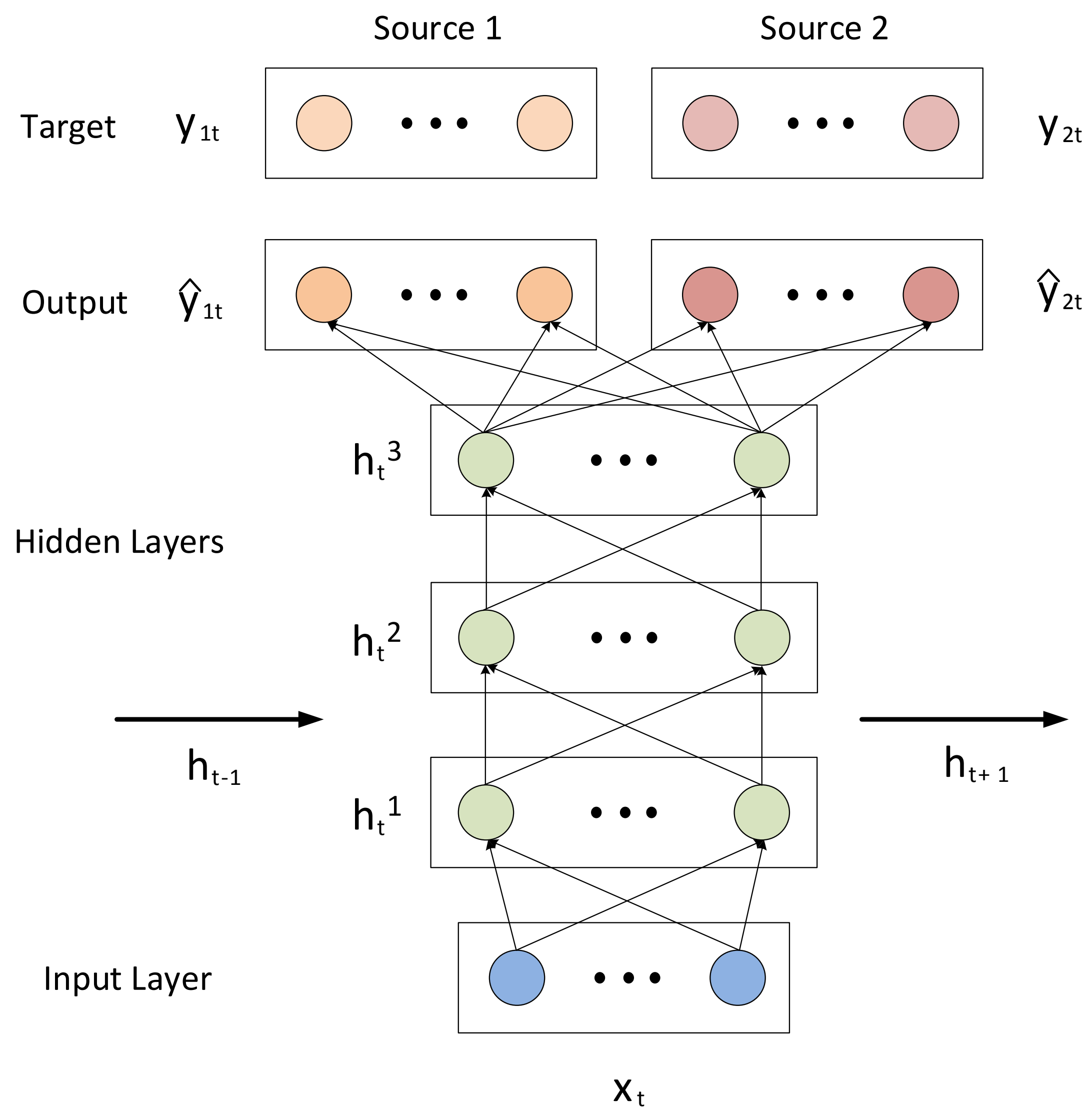


Proposed Methods

- Model architecture
- Joint time frequency masking
- Discriminative training

Model Architecture

- Jointly model two sources as targets
- Enable to use different features – log-mel, spectra
- Explore DNN, DRNN
- ➢ Time frequency masking



Time Frequency Masking

- Masking – enforce the constraints that sum of the predictions equals to the original mixture
- Binary mask

$$\mathbf{M_b}(f) = \begin{cases} 1 & |\hat{\mathbf{y}}_{\mathbf{1}_t}(f)| > |\hat{\mathbf{y}}_{\mathbf{2}_t}(f)| \\ 0 & \text{otherwise} \end{cases}$$

- Soft mask

$$\mathbf{M_s}(f) = \frac{|\hat{\mathbf{y}}_{\mathbf{1}_t}(f)|}{|\hat{\mathbf{y}}_{\mathbf{1}_t}(f)| + |\hat{\mathbf{y}}_{\mathbf{2}_t}(f)|}$$

- Apply the mask to predicted results

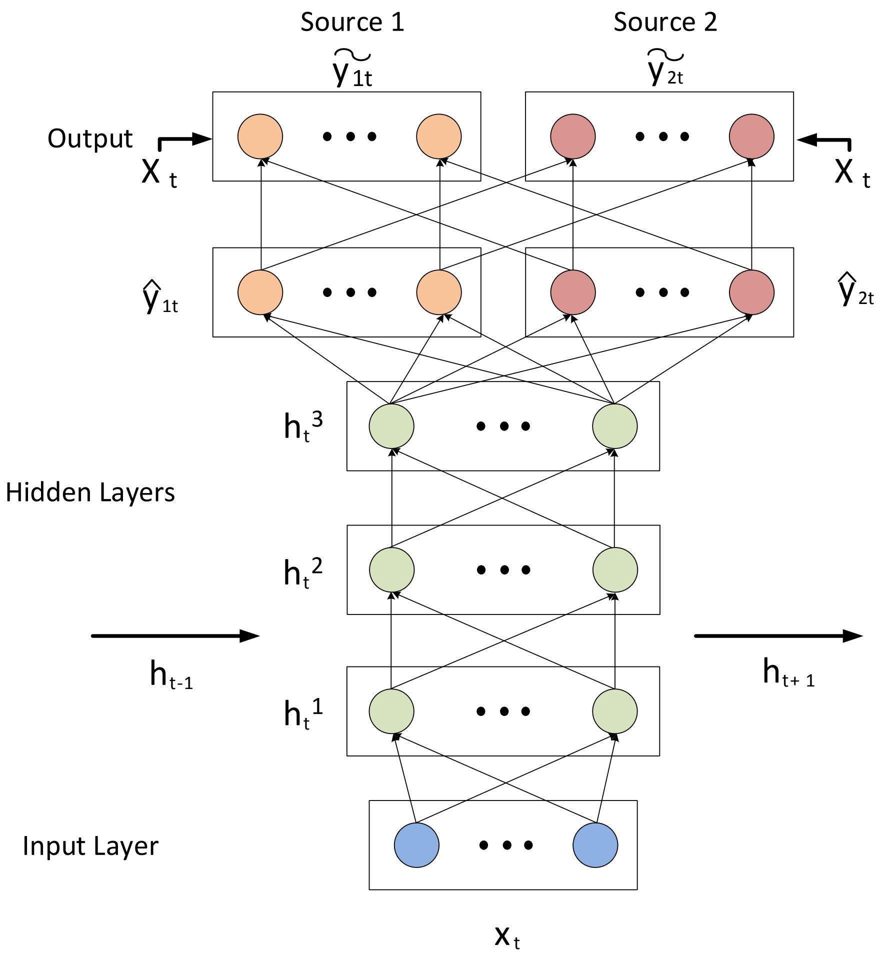
$$\hat{\mathbf{s}}_{\mathbf{1}_t}(f) = \mathbf{M}(f)\mathbf{X}_t(f)$$
$$\hat{\mathbf{s}}_{\mathbf{2}_t}(f) = (1 - \mathbf{M}(f))\,\mathbf{X}_t(f)$$

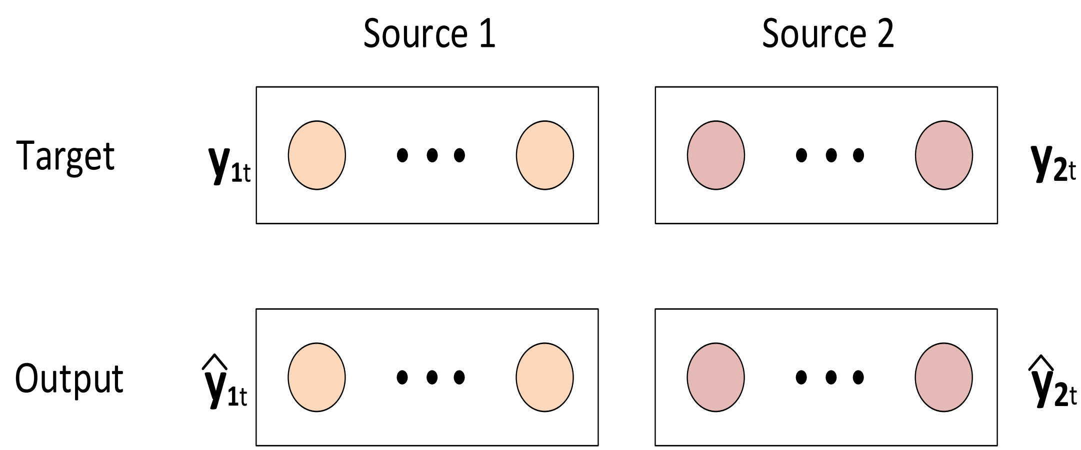
Joint Time Frequency Masking

- Viewed the masking operation as a deterministic layer
- Train the model with masking jointly

$$\tilde{\mathbf{y}}_{\mathbf{1}_t} = \frac{|\hat{\mathbf{y}}_{\mathbf{1}_t}|}{|\hat{\mathbf{y}}_{\mathbf{1}_t}| + |\hat{\mathbf{y}}_{\mathbf{2}_t}|} \odot \mathbf{X}_t$$

$$\tilde{\mathbf{y}}_{\mathbf{2}_t} = \frac{|\hat{\mathbf{y}}_{\mathbf{2}_t}|}{|\hat{\mathbf{y}}_{\mathbf{1}_t}| + |\hat{\mathbf{y}}_{\mathbf{2}_t}|} \odot \mathbf{X}_t$$





Discriminative training

- Minimize squared error

$$||\hat{\mathbf{y}}_{\mathbf{1}_t} - \mathbf{y}_{\mathbf{1}_t}||_2^2 + ||\hat{\mathbf{y}}_{\mathbf{2}_t} - \mathbf{y}_{\mathbf{2}_t}||_2^2$$

- Enforce source to interference ratio

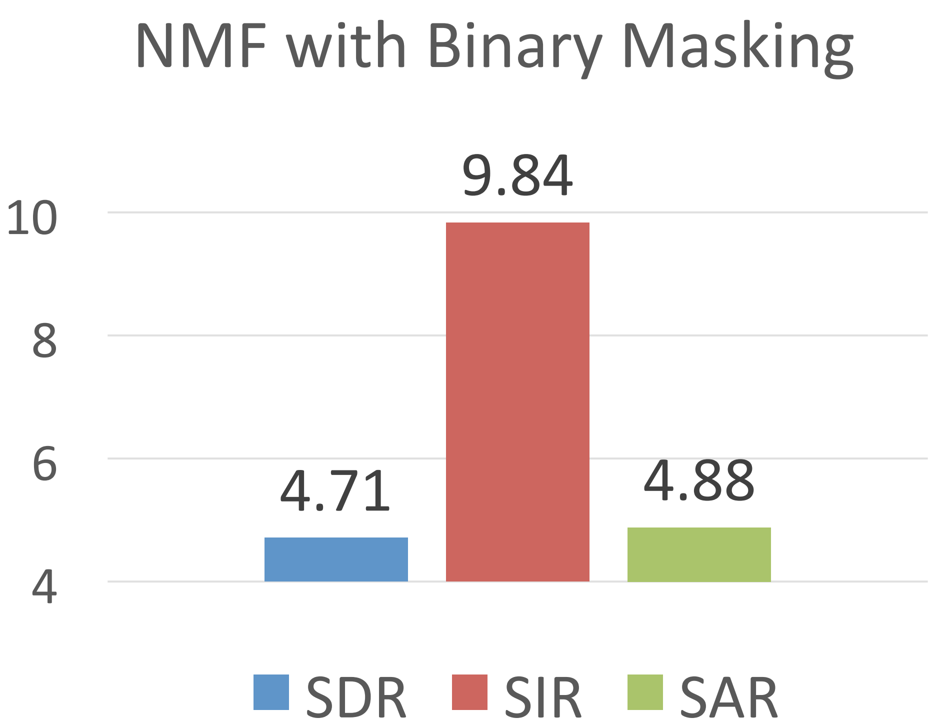
$$||\hat{\mathbf{y}}_{\mathbf{1}_t} - \mathbf{y}_{\mathbf{1}_t}||_2^2 - \gamma||\hat{\mathbf{y}}_{\mathbf{1}_t} - \mathbf{y}_{\mathbf{2}_t}||_2^2 + ||\hat{\mathbf{y}}_{\mathbf{2}_t} - \mathbf{y}_{\mathbf{2}_t}||_2^2 - \gamma||\hat{\mathbf{y}}_{\mathbf{2}_t} - \mathbf{y}_{\mathbf{1}_t}||_2^2$$

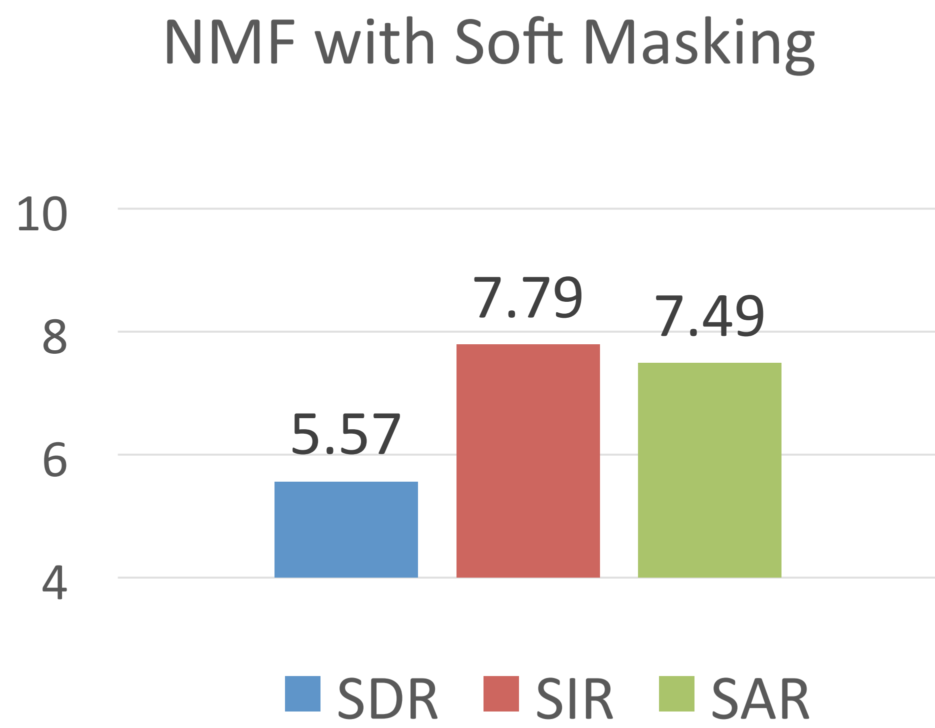
Experimental Setting

- TIMIT dataset
  - Mixed the speech from a male and a female speaker at 0 dB
- Circular shift to increase the variety of training samples
- Deep learning models
  - RELU
  - L-BFGS for optimization
  - Use 2 hidden layers with 150 hidden units
- BSS EVAL metric (SDR, SIR, SAR)
  - SIR - Suppression of interference
  - SAR - Artifacts introduced by the separation process
  - SDR - Overall performance

Baseline - NMF

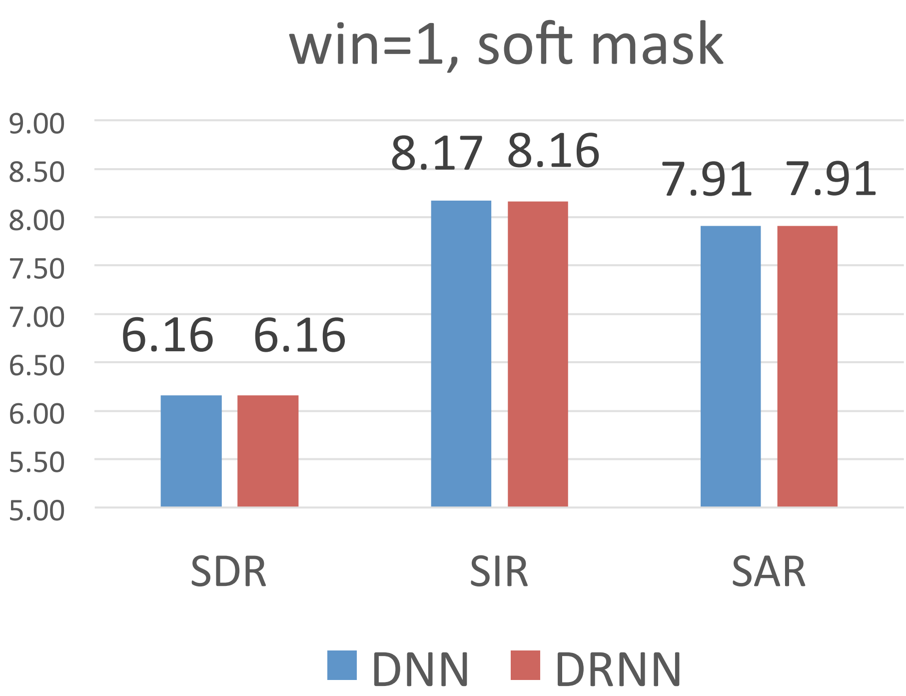
- 512-point STFT
- Generalized KL-divergence metric

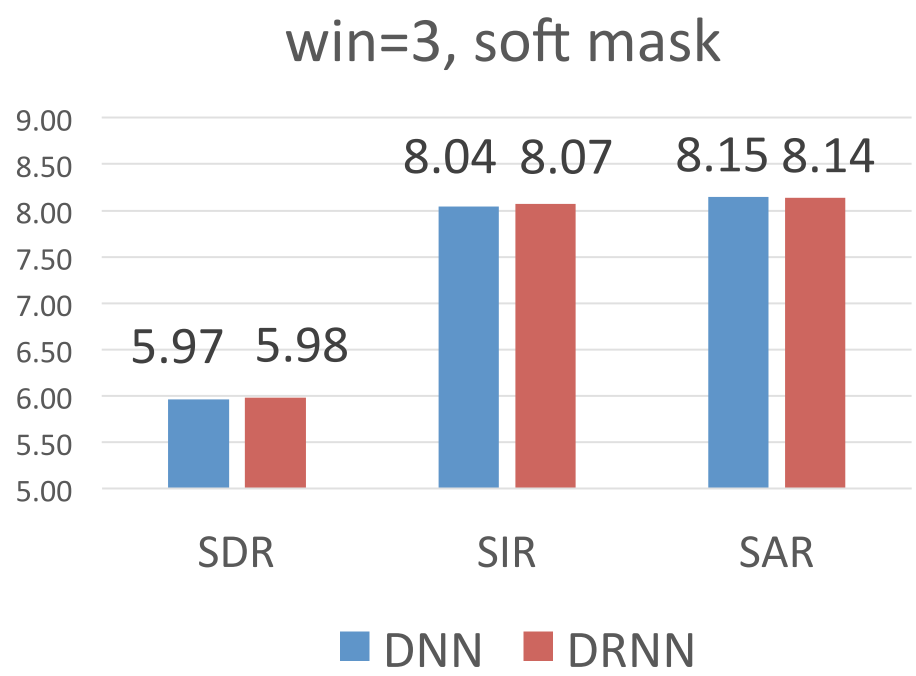




Experimental Results – DNN vs. DRNN

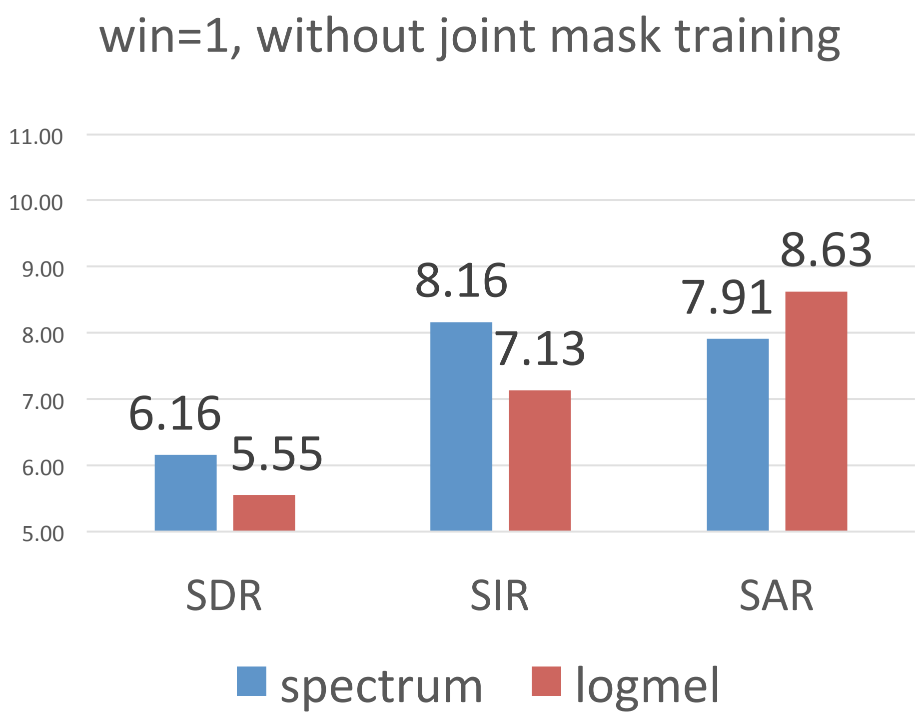
- Spectral features
- No significant difference

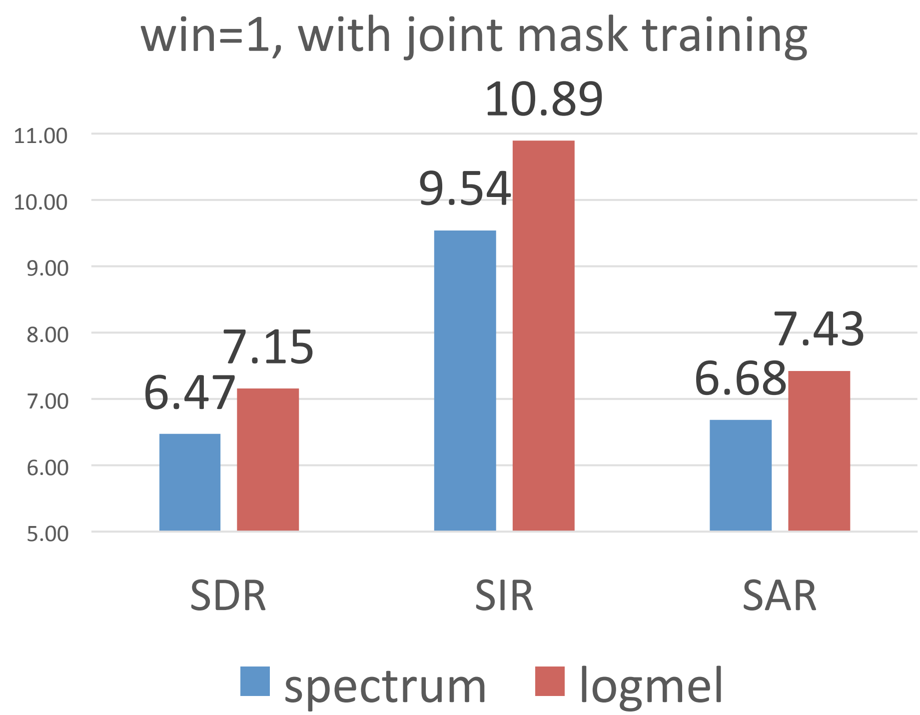




Experimental Results – Features, Joint Mask Training

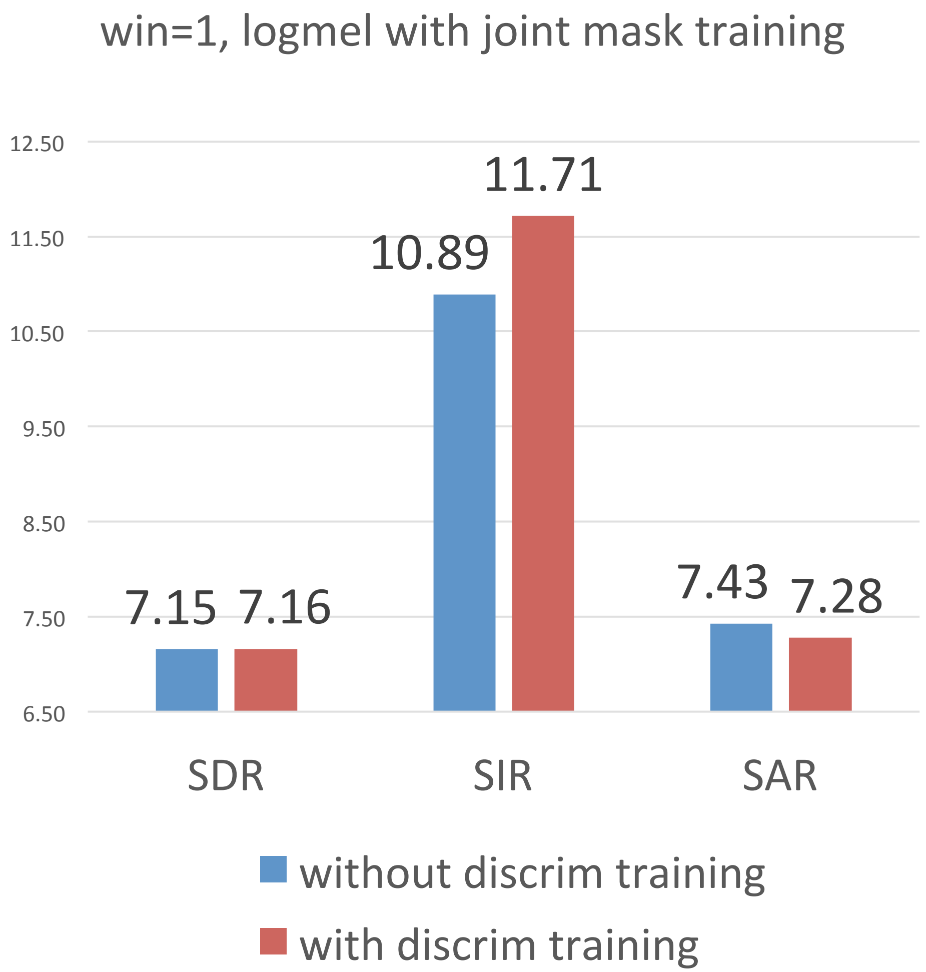
- Log-mel features perform better with joint mask training

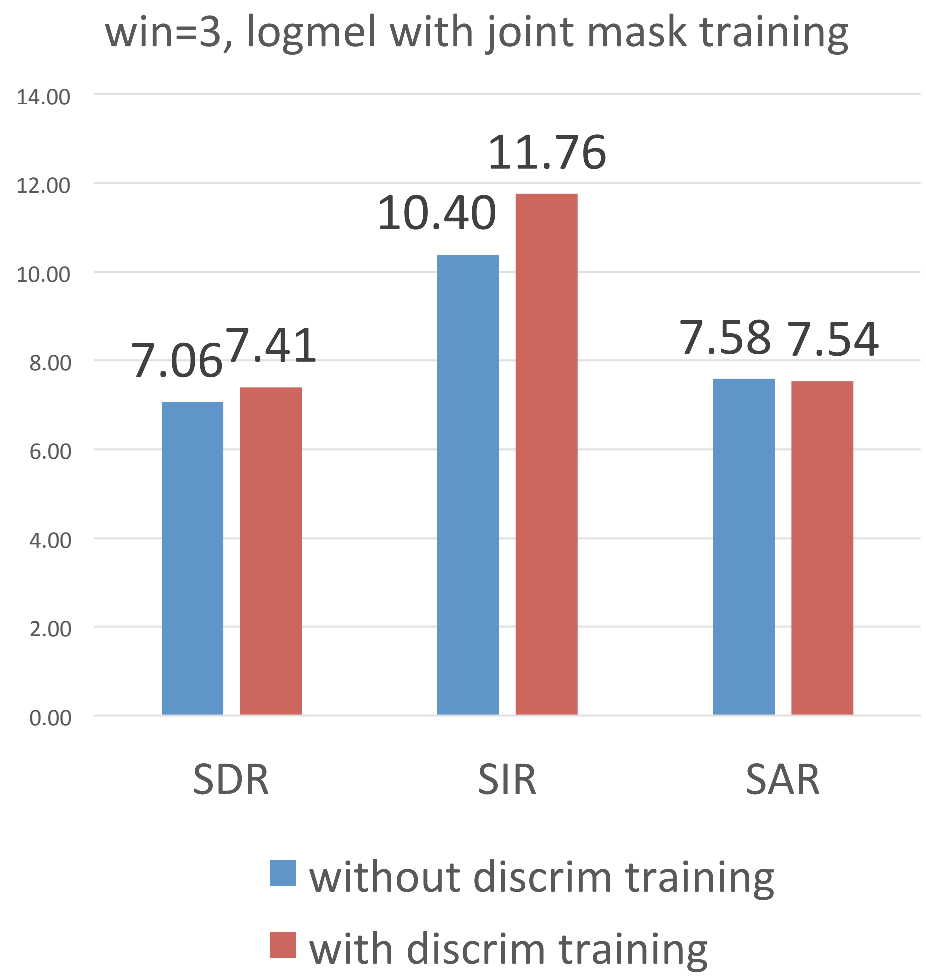




Experimental Results – Discriminative Training

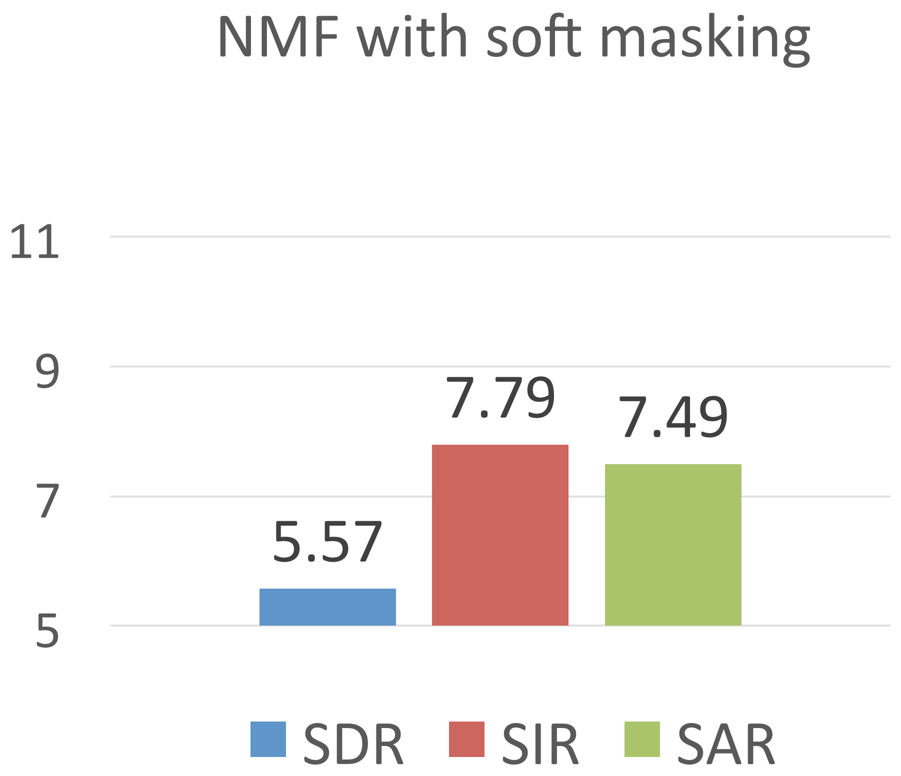
- Discriminative training provides extra regularization

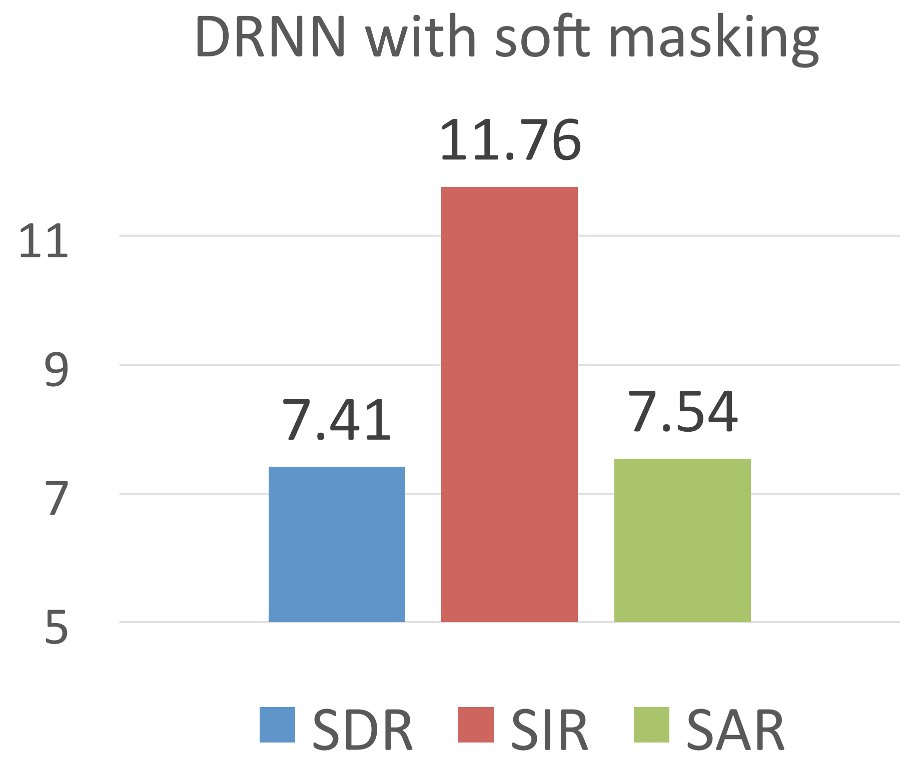




Summary

- Comparison between NMF and DRNN with log-mel, joint mask training, and discriminative training objective





DEMO

Conclusion

- Propose using deep learning models for monaural speech separation.
  - Propose the joint optimization of a soft masking function and deep learning models
  - Discriminative training criterion to further improve the SIR
- Overall, our proposed models achieve 3.8~4.9 dB SIR gain compared to the NMF baseline
- Future work
  - Explore longer temporal information with neural networks
  - Apply many other applications such as robust ASR

Thank you!

https://sites.google.com/site/deeplearningsourceseparation

MIR1K GNSDR

Train on two singers and test on other 17 singers