# Single Channel Speech Separation with Constrained Utterance Level Permutation Invariant Training Using Grid LSTM

**Chenglin Xu**[1,2], Wei Rao[2], Xiong Xiao[3], Eng Siong Chng[1,2] and Haizhou Li[2,4]

[1] School of Computer Science and Engineering, Nanyang Technological University (NTU),Singapore
[2] Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3] Microsoft Corporation, United States
[4] Department of Electrical and Computer Engineering, National University of Singapore, Singapore

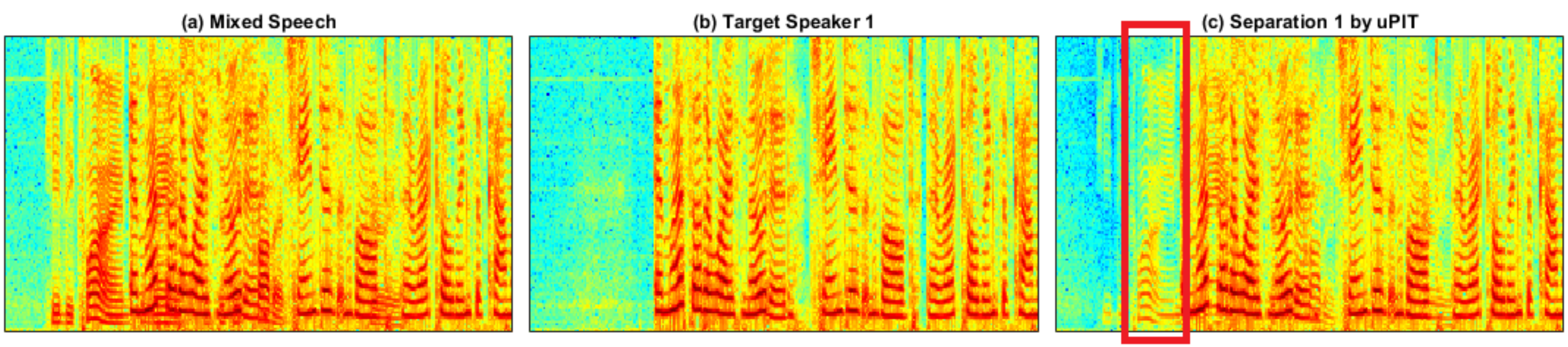
17 April, 2018

# Outline

1. Introduction
2. Methodology
3. Evaluation
4. Summary

# Introduction | Single channel speech separation with uPIT

- The performance of single channel speech separation has been significantly improved by deep learning based techniques, such as, deep clustering (DC) [1], deep attractor network (DANet) [2], utterance-level permutation invariant training (uPIT) [3], and so on.
- However, the state-of-the-art uPIT method runs into a *frame leakage* problem. (Frame leakage: Frames or time-frequency bins of speaker A are wrongly aligned to the output stream of speaker B, as shown in the red box of the figure.)



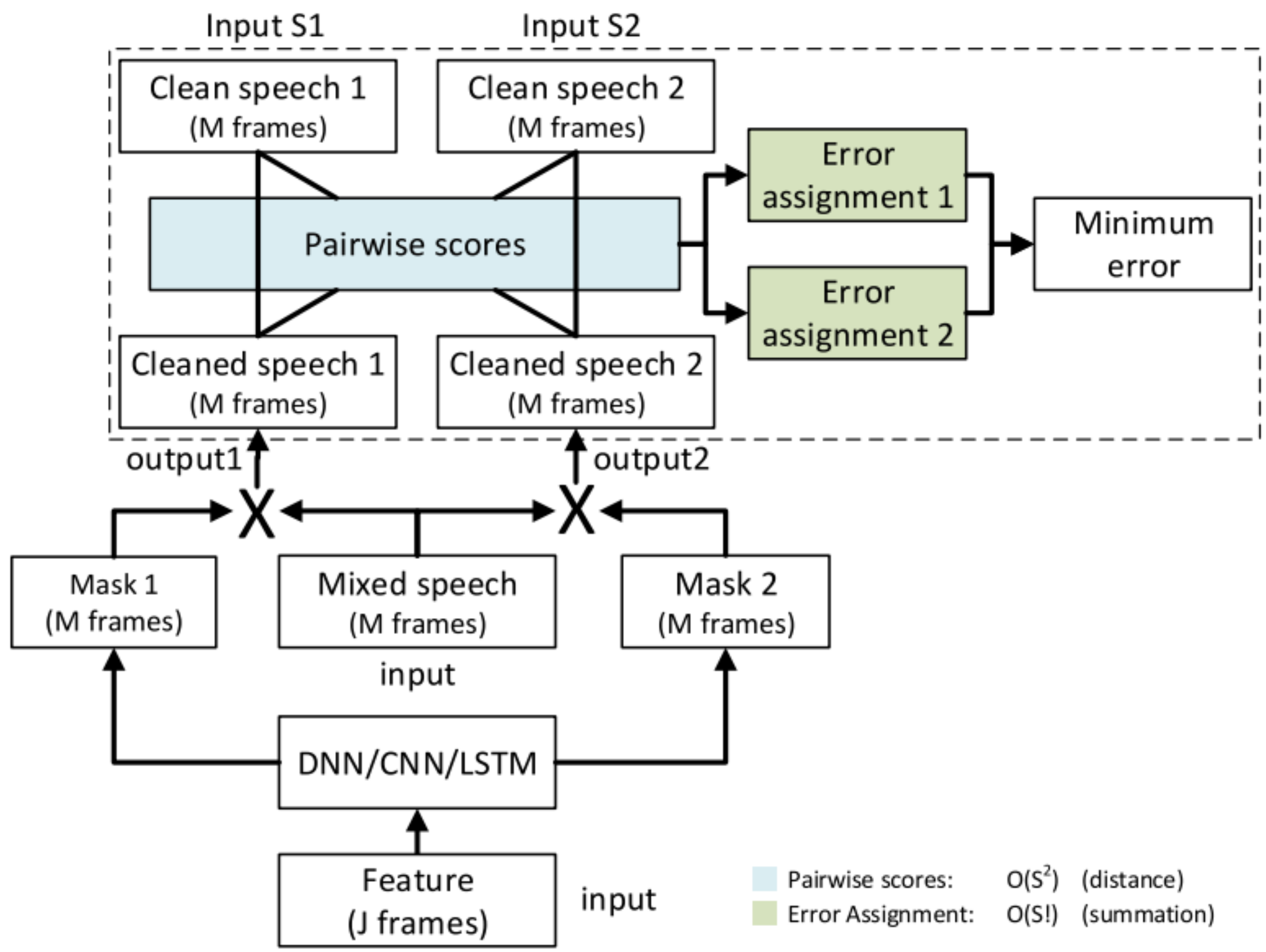

[1] J. R. Hershey, Z. Chen, J. L. Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation", in *Proc. ICASSP*, 2016, pp. 31-35

[2] Z. Chen, Y. Luo and N. Mesgarani, "Deep attractor network for single microphone speaker separation", in *Proc. ICASSP*, 2017

[3] M. Kolbek, Dong Yu, Z.-H. Tan and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.25, No.10, pp.1901-1913, 2017

# Single channel speech separation with uPIT

- The uPIT baseline framework from [1]




[1] M. Kolbek, Dong Yu, Z.-H. Tan and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.25, No.10, pp.1901-1913, 2017

# Motivation

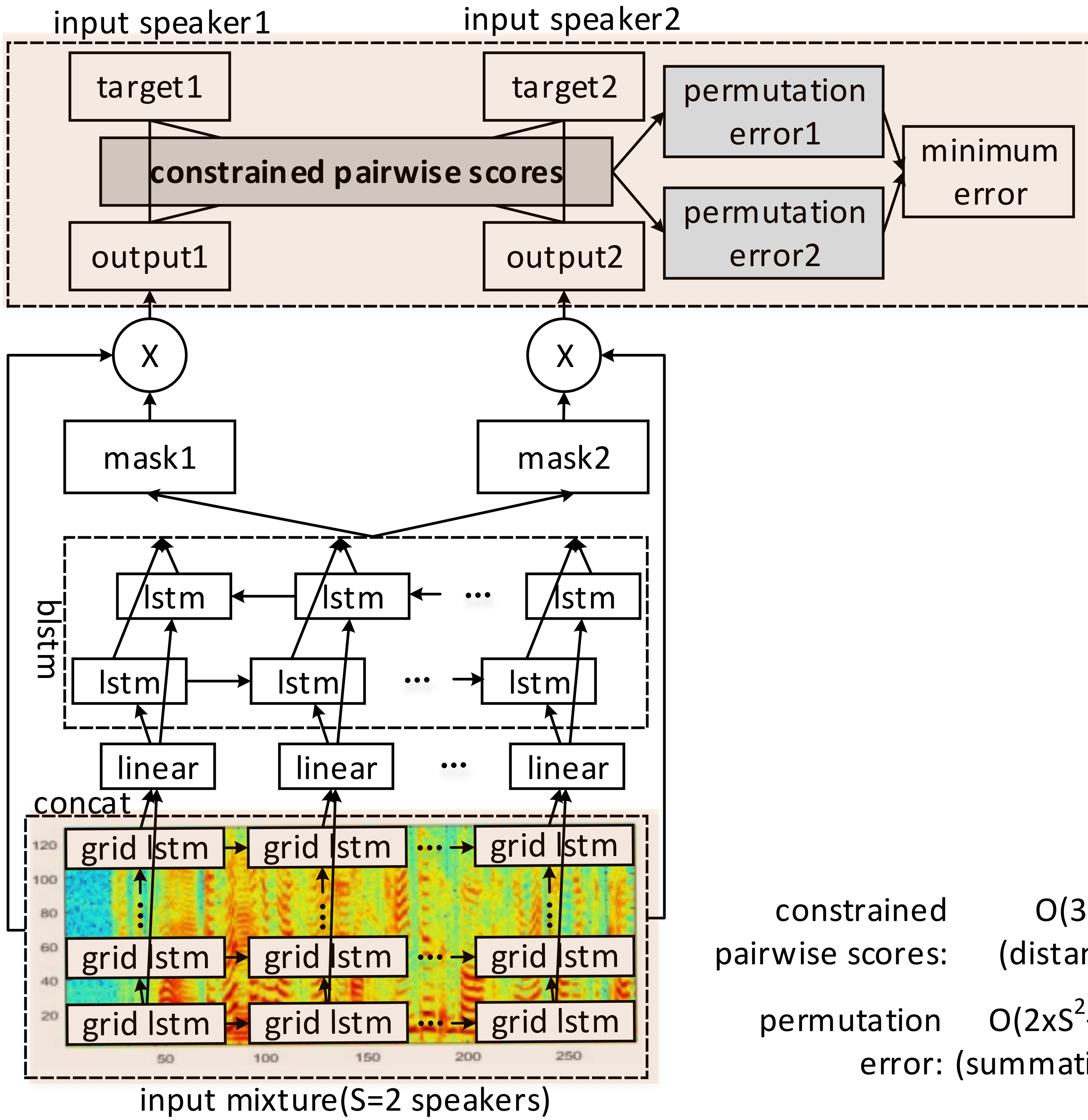
- **Constrain the objective using dynamic information**

  The dynamic information, e.g., the delta and acceleration, are used in the objective function to make the separation continuous across frames by using contextual information of several frames.

- **Capture temporal and spectral patterns simultaneously**

  Inspired by CASA method using heuristic rules, the grid LSTM is used to capture the heuristic patterns, e.g., common onset/offset, and learn corresponding temporal and spectral patterns from the magnitude spectrum both in time and frequency domain simultaneously.

# Methodology | cuPIT-Grid LSTM system




constrained pairwise scores: $O(3xS^2)$ (distance)

permutation error: $O(2xS^2+S!)$ (summation)

# cuPIT-Grid LSTM system

- **The objective function in uPIT baseline:**

$$J_{c,\phi_p(s)} = \frac{1}{T}\sum_{s=1}^{S}(||\hat{M}_s \odot |Y| - |X_{\phi_p(s)}| \odot cos(\theta_y - \theta_{\phi_p(s)})||_F^2)$$

$$\hat{p} = \arg\min_{p \in P} J_{c,\phi_p(s)}$$

$$J = J_{c,\phi_{\hat{p}}(s)}$$

- **The proposed constrained objective function (cuPIT):**

$$\begin{aligned} J_{c,\phi_p(s)} = \frac{1}{T}\sum_{s=1}^{S}(||\hat{M}_s \odot |Y| - |X_{\phi_p(s)}| \odot cos(\theta_y - \theta_{\phi_p(s)})||_F^2 \\ + w_D||f_D(\hat{M}_s \odot |Y|) - f_D(|X_{\phi_p(s)}| \odot cos(\theta_y - \theta_{\phi_p(s)}))||_F^2 \\ + w_A||f_A(\hat{M}_s \odot |Y|) - f_A(|X_{\phi_p(s)}| \odot cos(\theta_y - \theta_{\phi_p(s)}))||_F^2) \end{aligned}$$

$$f_D(v(t)) = \frac{\sum_{l=1}^{L} l \times (v(t+l) - v(t-l))}{\sum_{l=1}^{L} 2l^2}$$

$$\hat{p} = \arg\min_{p \in P} J_{c,\phi_p(s)}$$

$$J = J_{c,\phi_{\hat{p}}(s)}$$

# Experimental setup

- **Dataset**

  The WSJ0-2mix database* with the sampling rate at 8 kHz.
  - Training set: *20,000* utterances ≈ *30*h
  - Development set: *5,000* utterances ≈ *8*h
  - Test set: *3,000* utterances ≈ *5*h
- **Features**

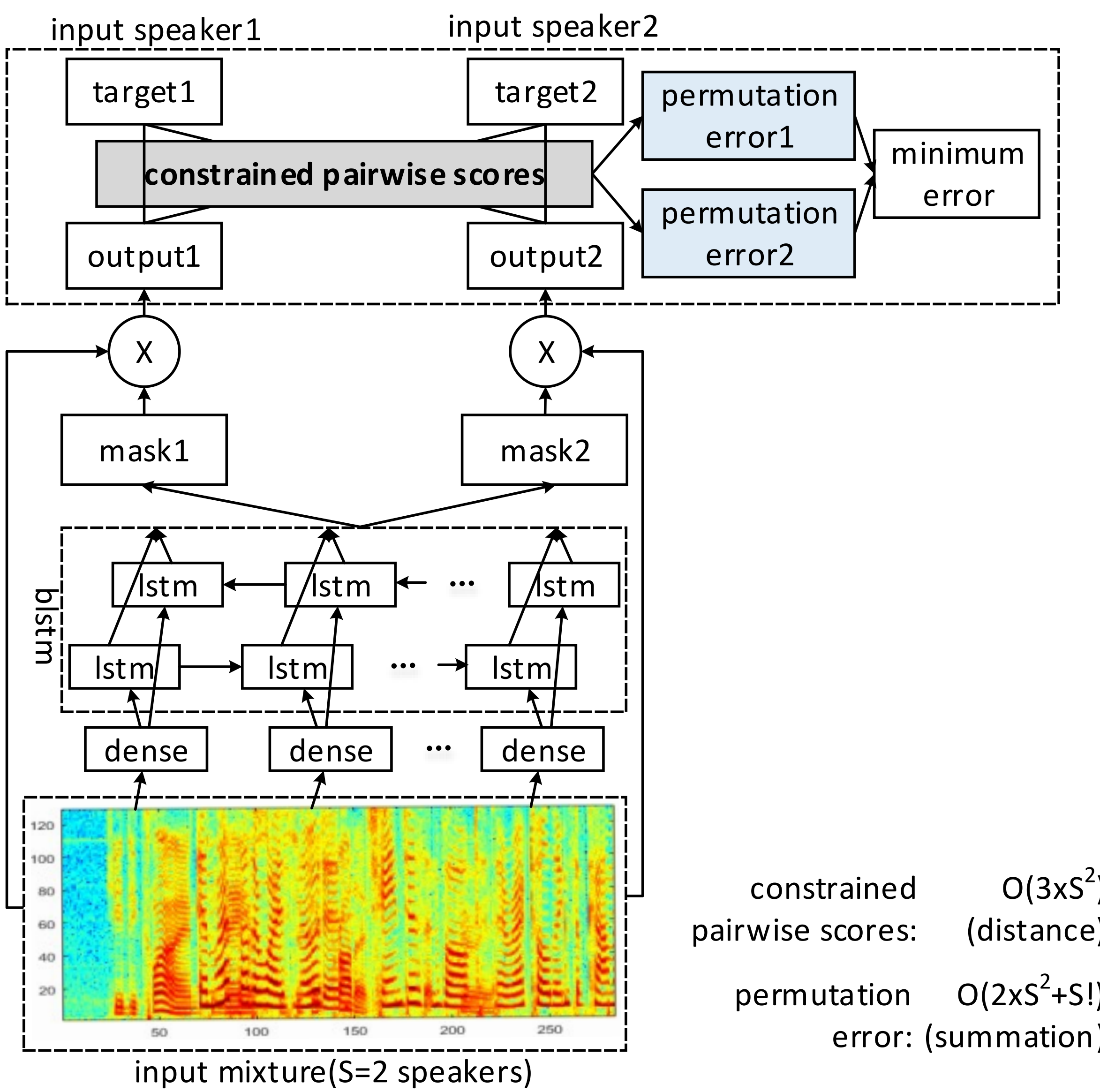
  *129*-dimensional spectral magnitude features computed by a STFT with a normalized square root of *32*ms length hamming window and *16*ms window shift.
- **Evaluation Metrics**
  - The global normalized signal-to-distortion ratio (GNSDR, same as “SDR improvement” in DC, DANet, uPIT baselines) using bss_eval toolbox [1].
  - Signal-to-interferences ratio (SIR).
  - Signal-to-artifacts ratio (SAR).

* Available at: http://www.merl.com/demos/deep-clustering

[1] Vincent, Emmanuel, Rémi Gribonval, and Cédric Févotte. "Performance measurement in blind audio source separation." *IEEE transactions on audio, speech, and language processing*14.4 (2006): 1462-1469.

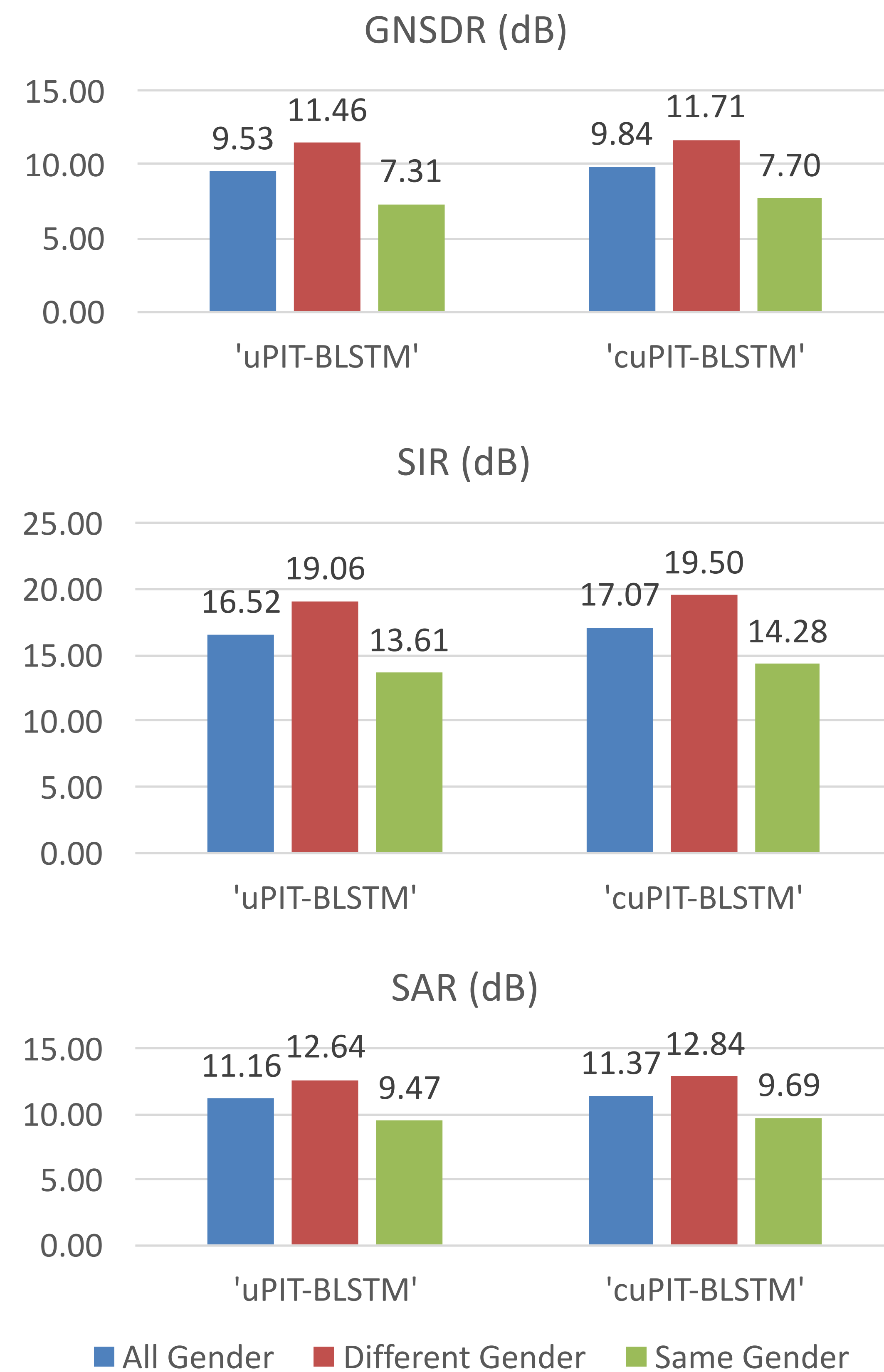
# Evaluation | Experiment 1: same network architecture as baseline

- **Constrained uPIT (cuPIT) *vs.* baseline uPIT**



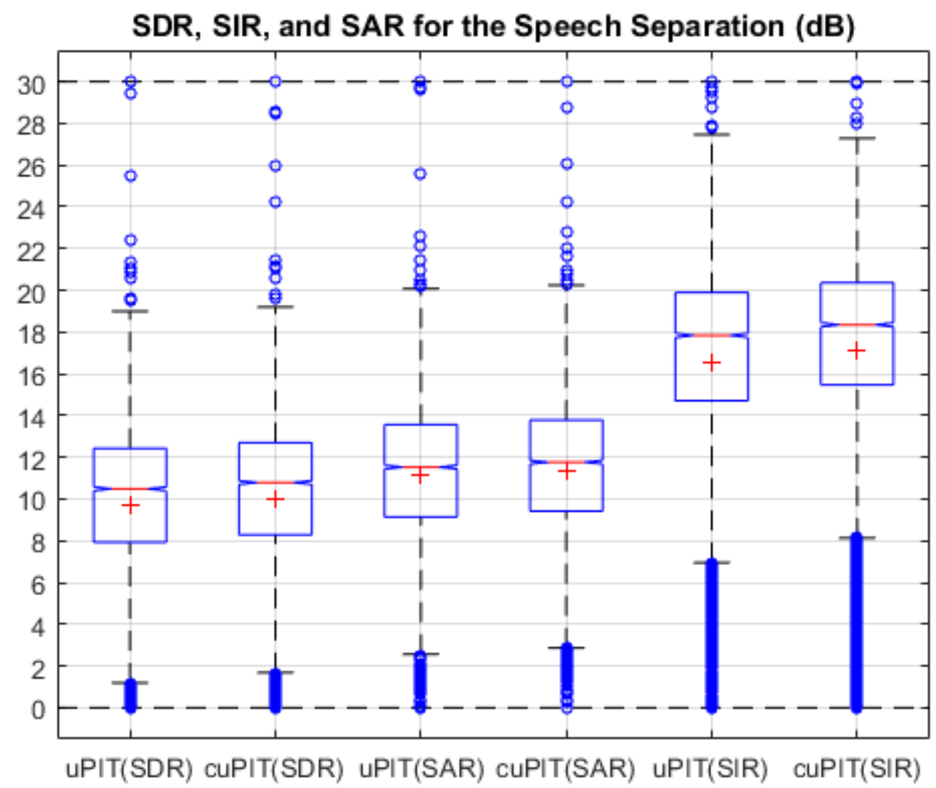

constrained pairwise scores: $O(3xS^2)$ (distance)

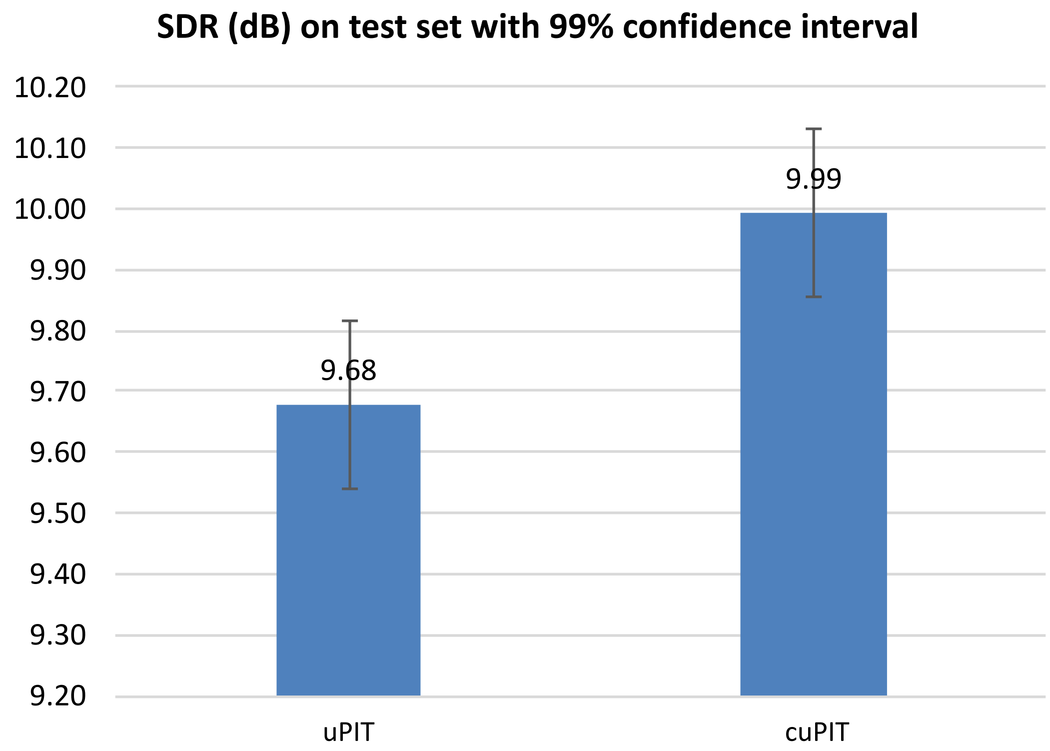
permutation error: $O(2xS^2+S!)$ (summation)



# Experiment 1: same network architecture as baseline

- **Constrained uPIT *vs.* baseline uPIT**



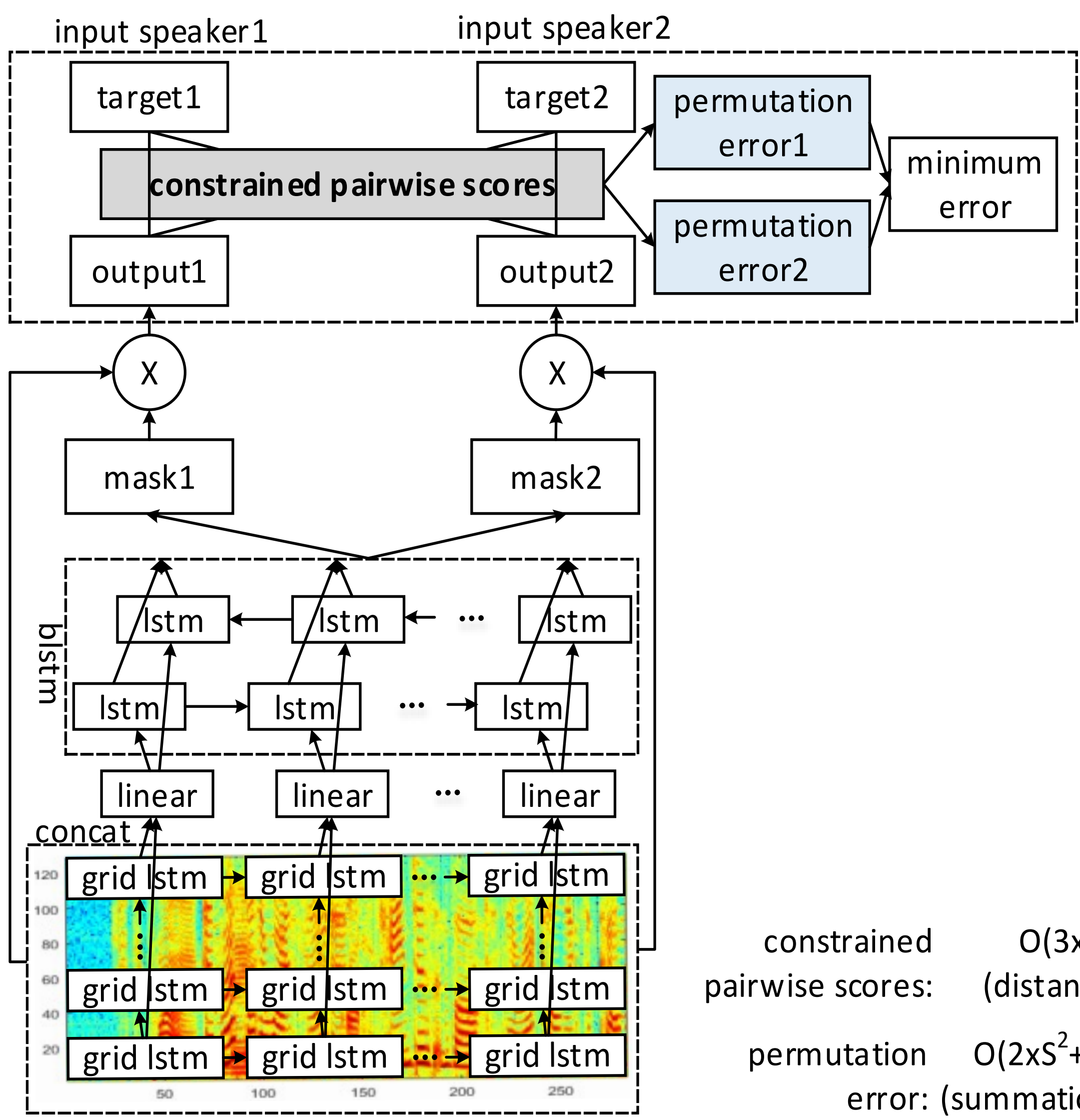




Paired t-test on test set of SDR result: statistically significant

# Evaluation

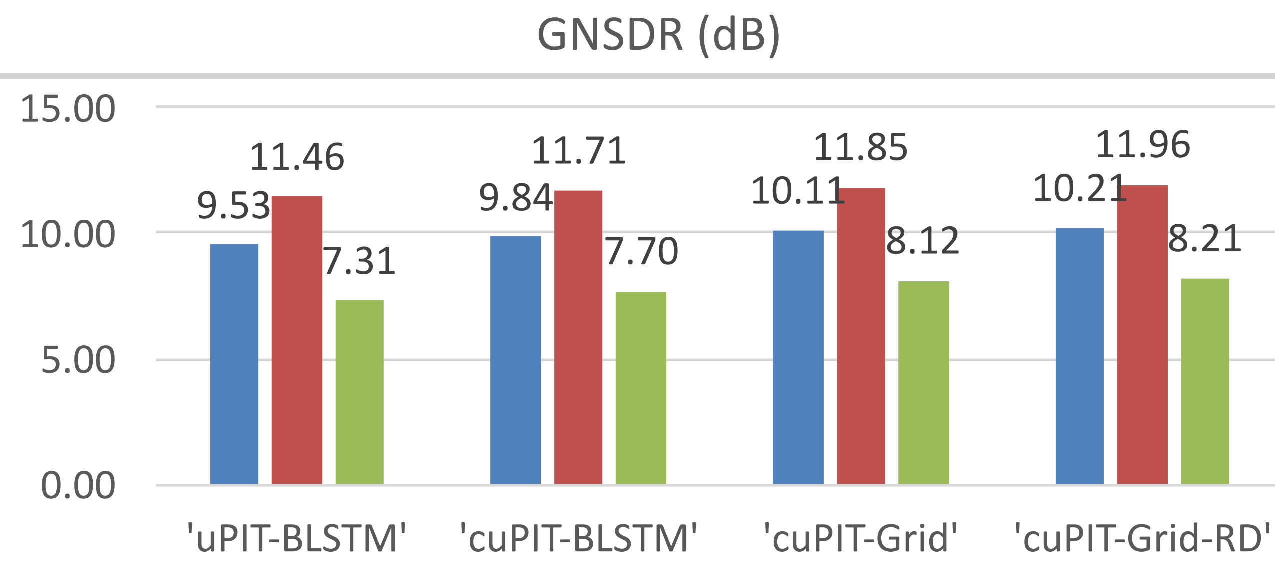
## Experiment 2:

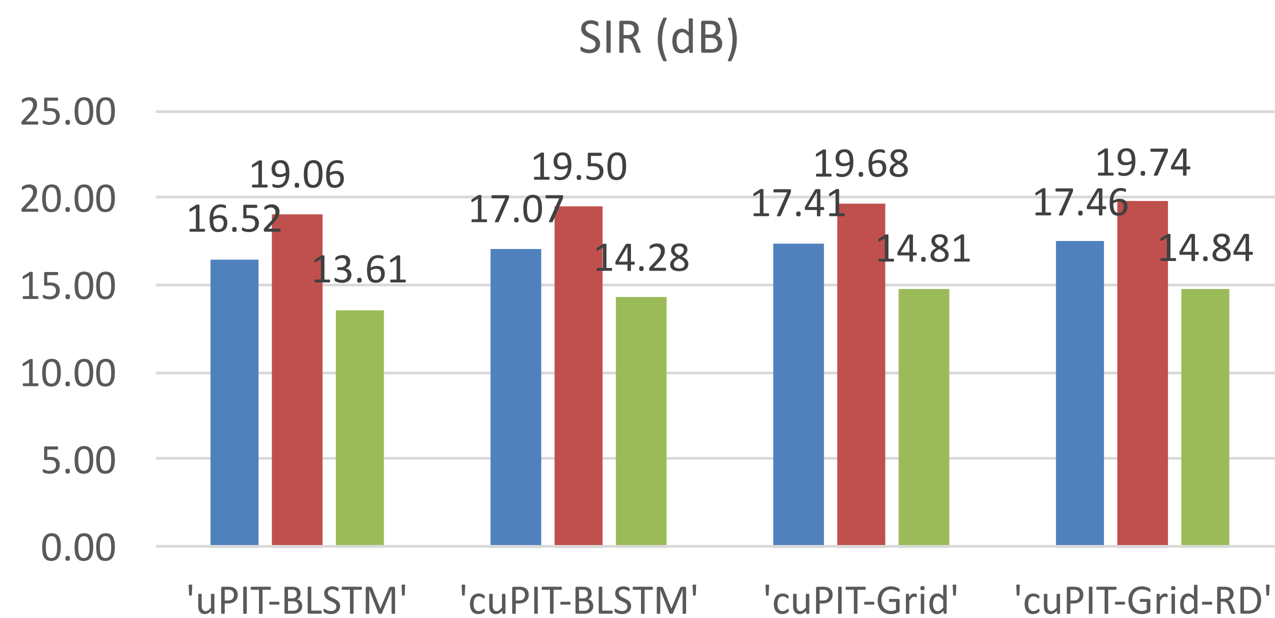
## Grid LSTM with cuPIT objective

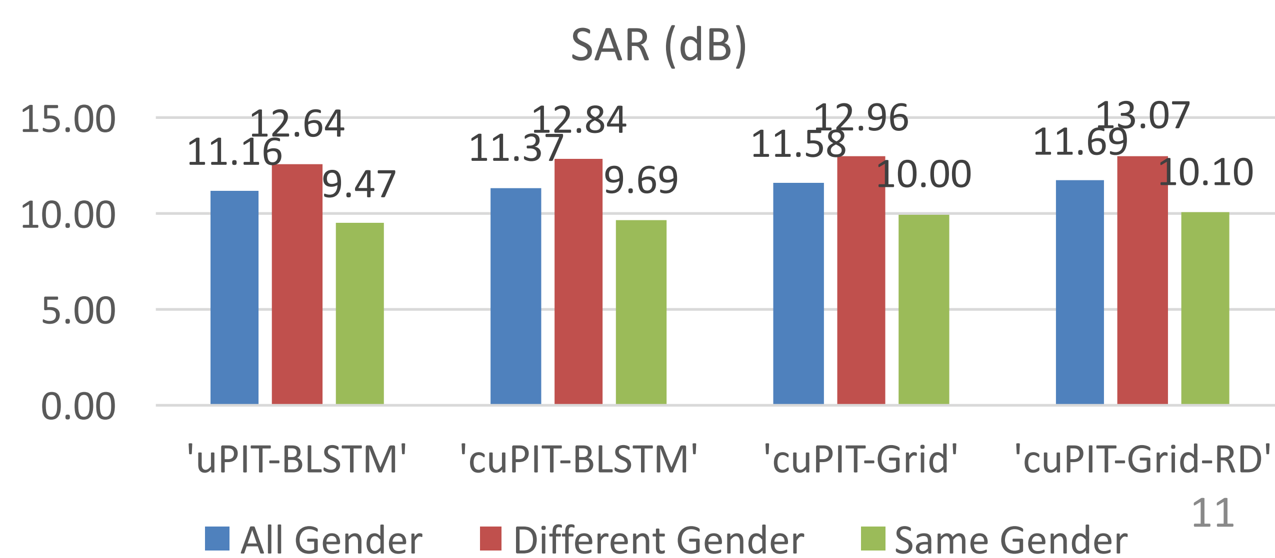



constrained pairwise scores: $O(3xS^2)$ (distance)

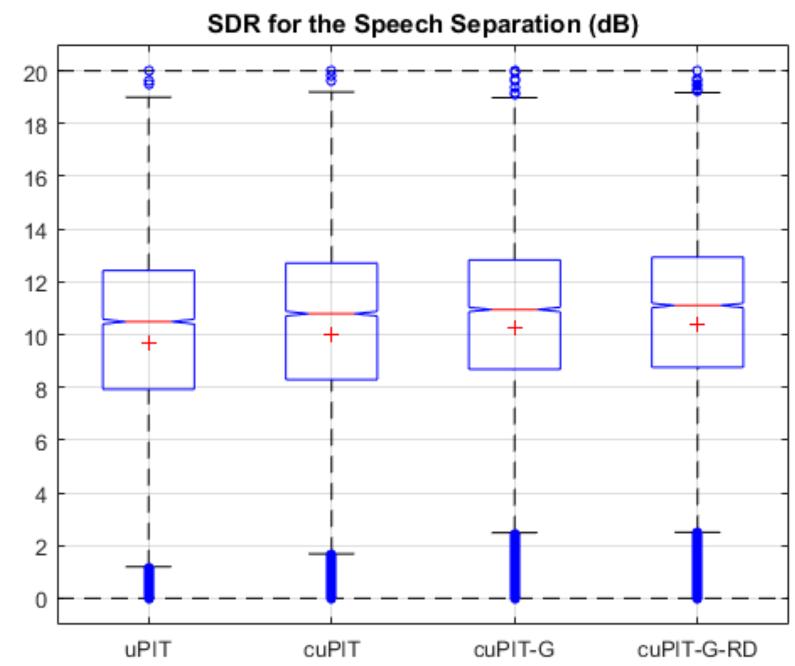
permutation error: $O(2xS^2+S!)$ (summation)

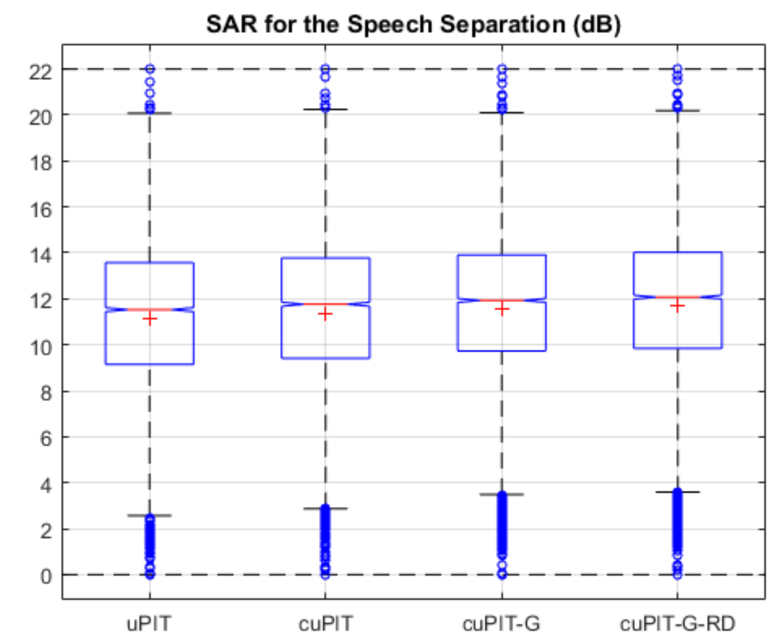


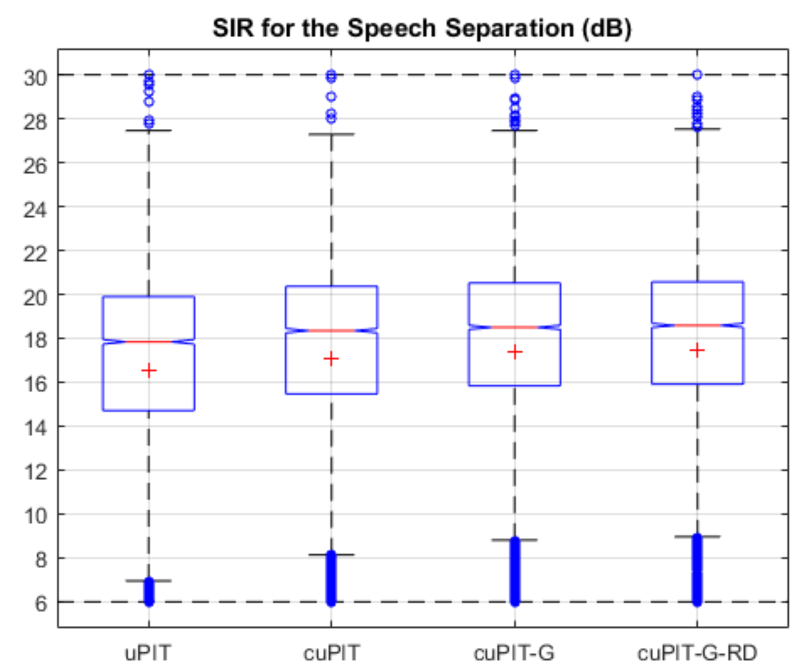



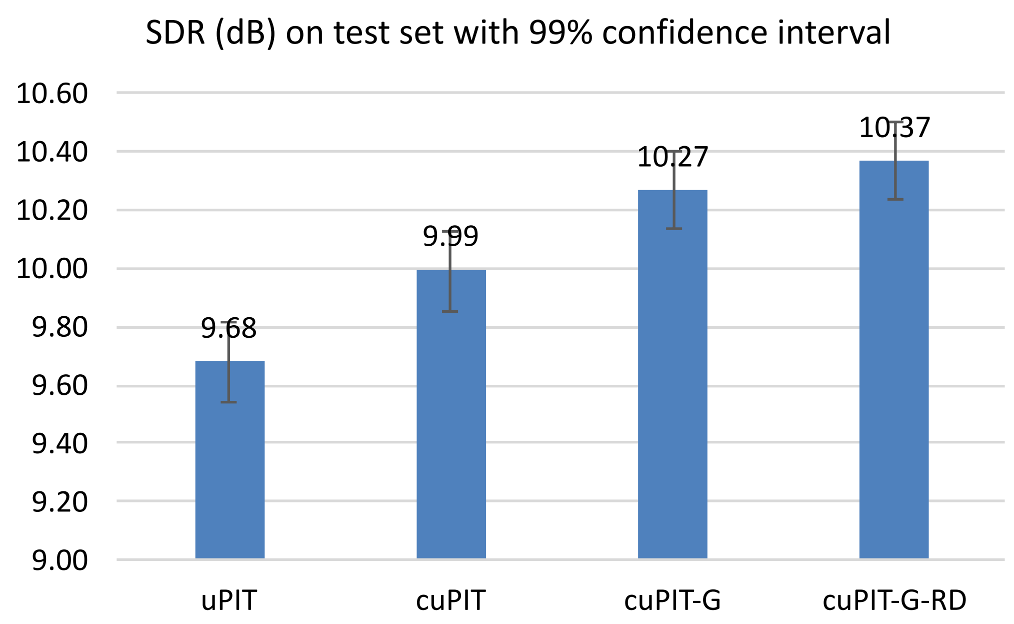



# Evaluation | Experiment 2: Grid LSTM with cuPIT objective




Paired t-test on test set of SDR result: statistically significant.

## • Comparisons with state-of-the-art methods

**DC [1]:** The mixture is projected into an embedding space, where time-frequency bins belonging to the same speaker are grouped into a cluster using k-means to form a binary mask used to separate the speakers from the mixture signal.

**DC+ [2]:** The cluster stage is connected with the embedding learning network to do end-to-end mask estimation.

**DANet [3]:** Attractor points, which attract the time-frequency bins corresponding to each target speaker, are created in the embedding space. The network is trained in end-to-end to estimate the masks, which are used to separate the mixture signal.

**PIT-DNN [4]:** The magnitude approximation masks are estimated in end-to-end by using a permutation invariant training with context expansion in inputs and calculating the cost using DNN.

**PIT-CNN [4]:** The magnitude approximation masks are estimated in end-to-end by using a permutation invariant training using CNN.

**uPIT-BLSTM [5]:** The magnitude approximation masks are estimated in end-to-end by using an utterance level permutation invariant training to solve the label ambiguity problem in training and inference stage.

[1] J. R. Hershey, Z. Chen, J. L. Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation", in *Proc. ICASSP*, 2016, pp. 31-35
[2] Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., & Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. arXiv preprint arXiv:1607.02173.
[3] Z. Chen, Y. Luo and N. Mesgarani, "Deep attractor network for single microphone speaker separation", in *Proc. ICASSP*, 2017
[4] D. Yu, M. Kolbek, Z.-H. Tan and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation", in Proc. ICASSP, 2017
[5] M. Kolbek, Dong Yu, Z.-H. Tan and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.25, No.10, pp.1901-1913, 2017

- **Comparisons with state-of-the-art methods**

| Method | Opt Assign (GNSDR, dB) | | Def Assign (GNSDR, dB) | |
|---|---|---|---|---|
| | Dev Set | Test Set | Dev Set | Test Set |
| DC [1] | - | - | 5.9 | 5.8 |
| DC+ [2] | - | - | - | 9.4 |
| DANet [3] | - | - | - | 9.6 |
| PIT-DNN [4] | 7.3 | 7.2 | 5.7 | 5.2 |
| PIT-CNN [4] | 8.4 | 8.6 | 7.7 | 7.8 |
| uPIT-BLSTM [5] | 10.9 | 10.8 | 9.4 | 9.4 |
| uPIT-BLSTM* | 10.8 | 10.7 | 9.6 | 9.5 |
| cuPIT-BLSTM | 11.1 | 11.0 | 10.0 | 9.8 |
| cuPIT-Grid | 11.2 | 11.2 | 10.2 | 10.1 |
| cuPIT-Grid-RD | **11.3** | **11.3** | **10.3** | **10.2** |
| IRM | 12.4 | 12.7 | 12.4 | 12.7 |
| IPSM | 14.9 | 15.1 | 14.9 | 15.1 |

**Opt Assign:** realign the output streams by using target speaker's speech to show the upper bound without frame leakage.

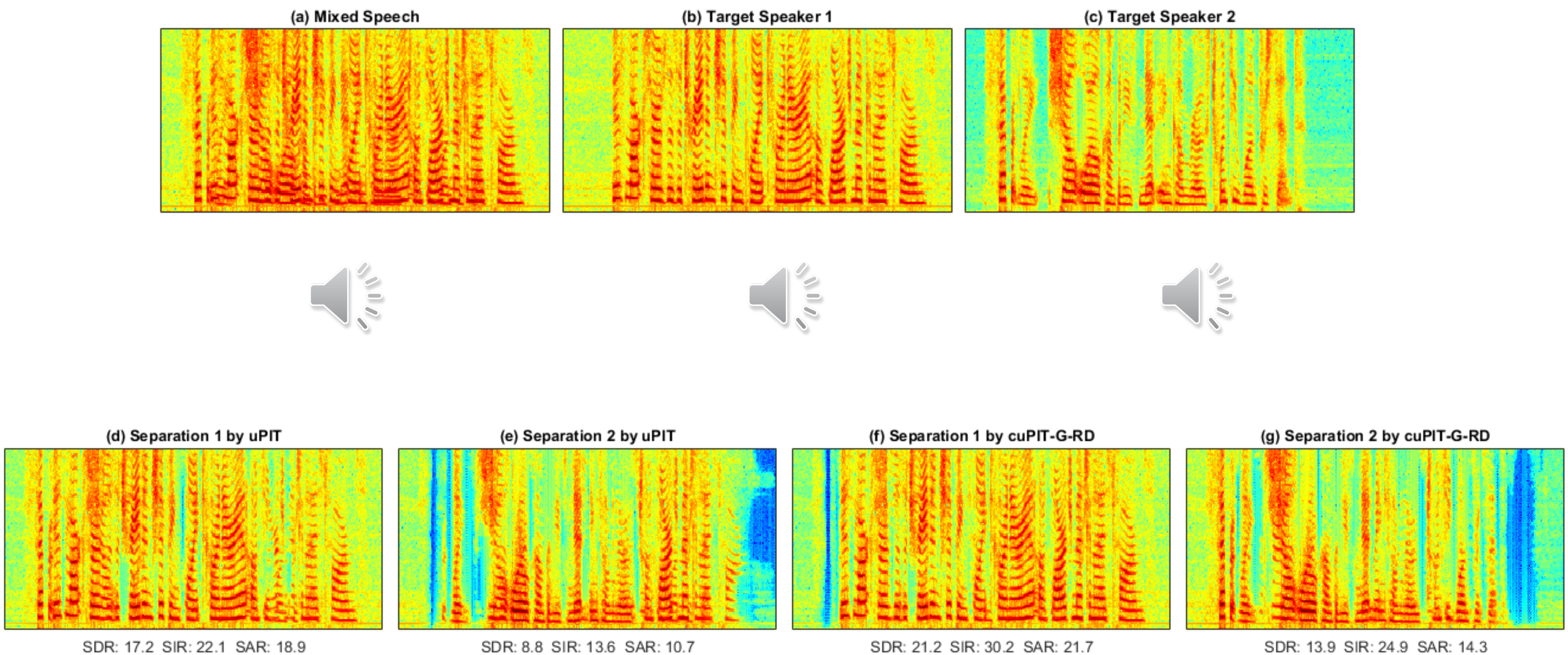
**Def Assign:** default output streams from the system without realignment.

**uPIT-BLSTM*:** Our reimplementation of uPIT-BLSTM baseline.

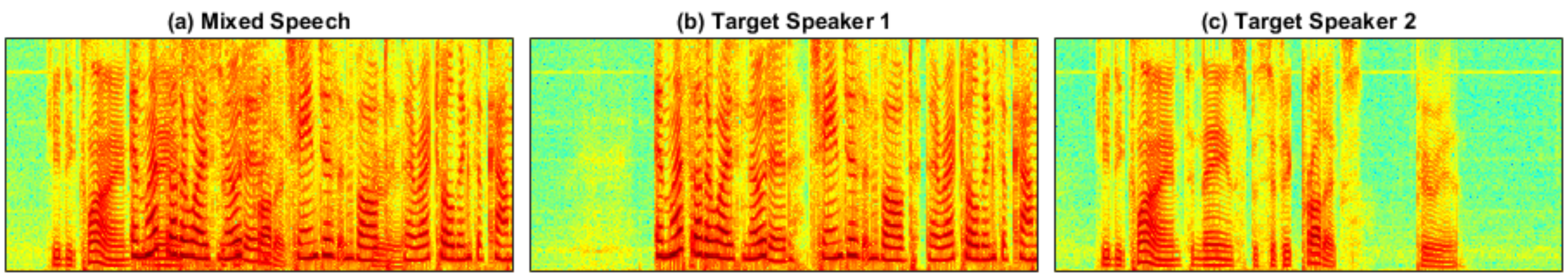
[1] J. R. Hershey, Z. Chen, J. L. Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation", in *Proc. ICASSP*, 2016, pp. 31-35
[2] Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., & Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. arXiv preprint arXiv:1607.02173.
[3] Z. Chen, Y. Luo and N. Mesgarani, "Deep attractor network for single microphone speaker separation", in *Proc. ICASSP*, 2017
[4] D. Yu, M. Kolbek, Z.-H. Tan and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation", in Proc. ICASSP, 2017
[5] M. Kolbek, Dong Yu, Z.-H. Tan and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.25, No.10, pp.1901-1913, 2017

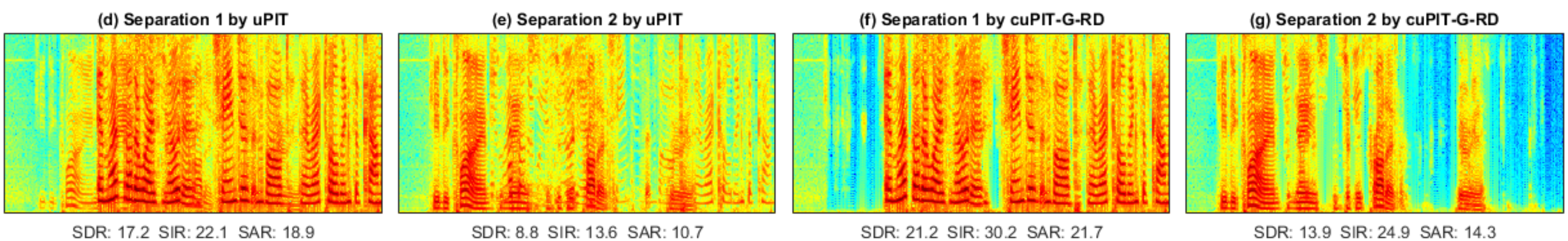
# Evaluation | Demo

**Example: two female speakers' mixture** ('050a050i_2.1935_421c020b_-2.1935')



**Example: male-female speakers' mixture** ('441c020m_2.4506_447o030z_-2.4506')



# Summary

- We propose a constrained cost function in uPIT by using dynamic information to solve the frame leakage problem.

- We further propose to use a Grid LSTM to learn temporal and spectral patterns from the time and frequency domain of the mixture signal simultaneously.

- The proposed method achieves better results than the current state-of-the-art uPIT method.

Thank you!