ICANN19
28th International Conference on Artificial Neural Networks

Deep Recurrent Neural Networks with Nonlinear Masking Layers and Two-Level Estimation for Speech Separation

Jiantao Zhang* and Pingjian Zhang
South China University of Technology, 510006 Guangzhou, China
Contact Author: *, 1277472231@qq.com

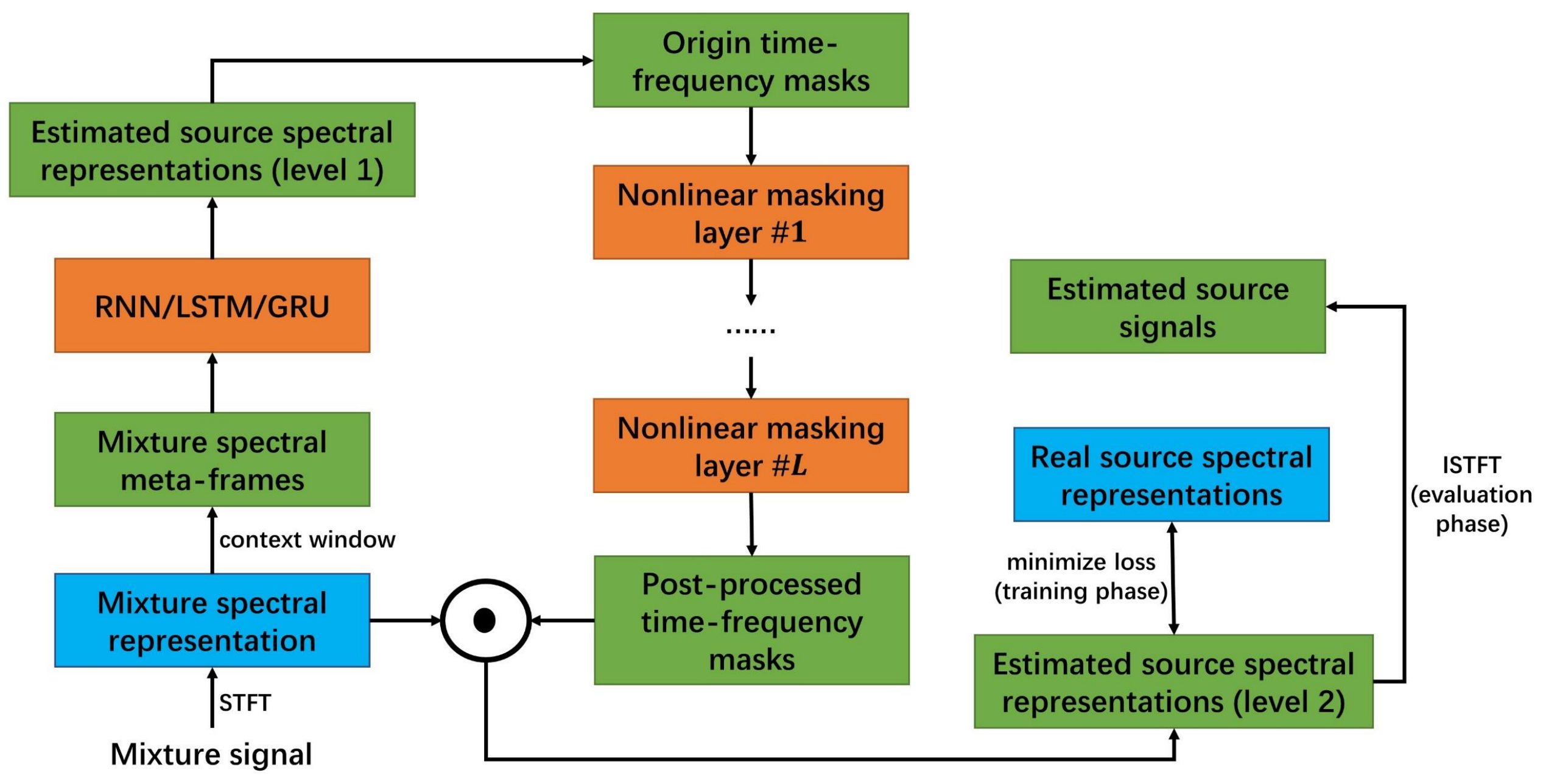
Introduction

The goal of speech separation is to separate a specific target speech from some background interferences and it has been treated as a signal processing problem traditionally. Recently, deep neural networks (DNNs) have played an increasingly important role in this field. In our study, deep RNNs with nonlinear masking layers and two-level estimation are proposed for speech separation.

Objectives

- To obtain the level 1 estimated sources via the RNN and use them to form the original deterministic time-frequency (T-F) masks.
- To correct and enhance the original masks (SMM, IRM, etc.) via the nonlinear masking layer, i.e., to form the post-processed nonlinear masks.
- To improve the overall quality of speech separation via the post-processed nonlinear masks, i.e., to obtain the level 2 estimated sources.

Two-Level Estimation


Origin time-frequency masks
Estimated source spectral representations (level 1)
Nonlinear masking layer #1
......
Nonlinear masking layer #L
RNN/LSTM/GRU
Estimated source signals
Mixture spectral meta-frames
context window
Real source spectral representations
ISTFT (evaluation phase)
minimize loss (training phase)
Mixture spectral representation
Post-processed time-frequency masks
Estimated source spectral representations (level 2)
STFT
Mixture signal

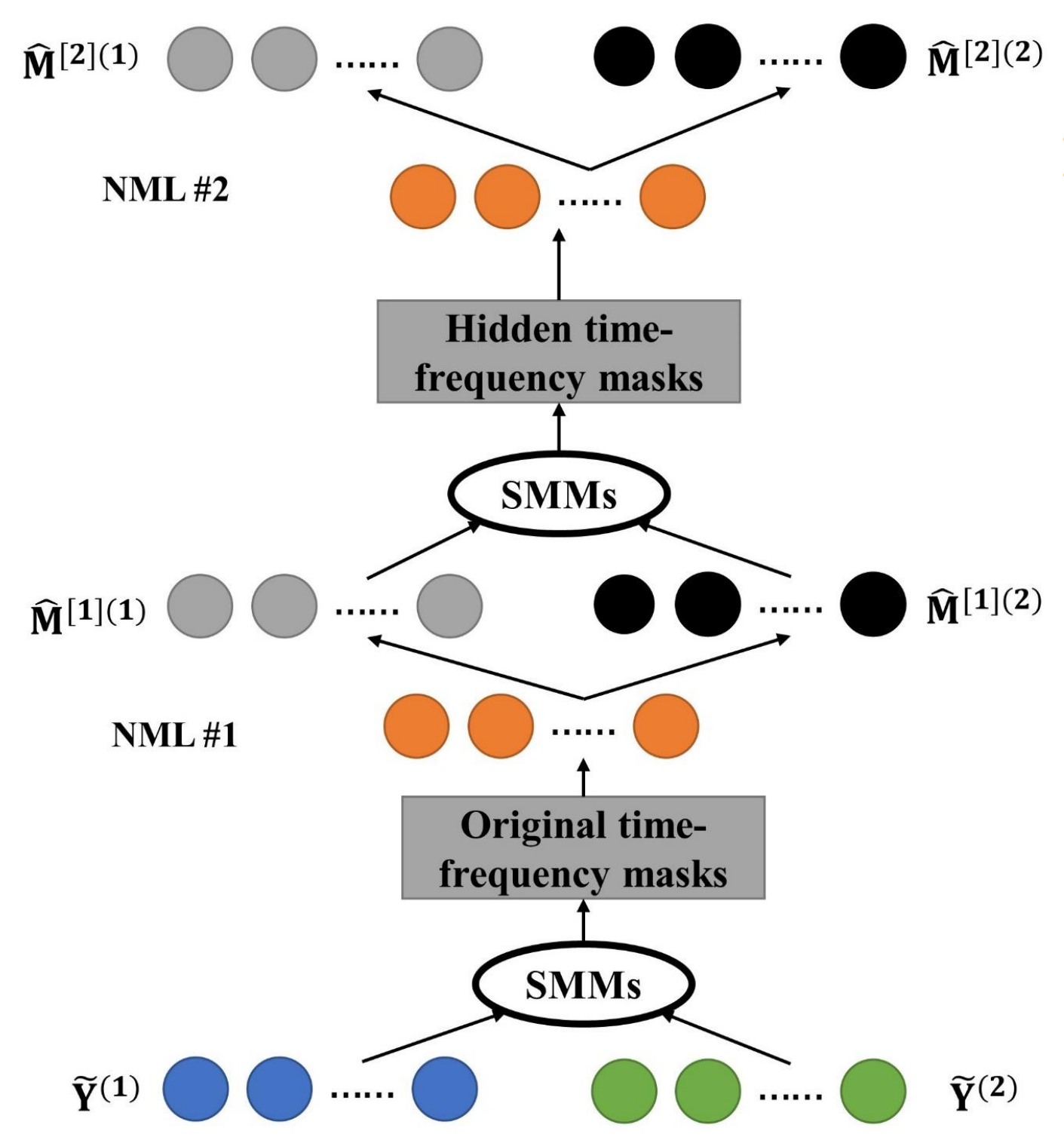
- ***Learning a simple mapping-based model***

First of all, we construct a simple mapping-based model via using the recurrent neural network, where “mapping-based” means directly mapping the mixture to the sources. The sources here are called level 1 estimated sources and used for construct original deterministic T-F masks.

- ***Stacking multiple nonlinear masking layers***

After that, multiple nonlinear masking layers are stacked together and accept the original T-F masks to output the nonlinear post-processed masks. These nonlinear masks are used to obtain the level 2 estimated sources.

Nonlinear Masking Layer (NML)

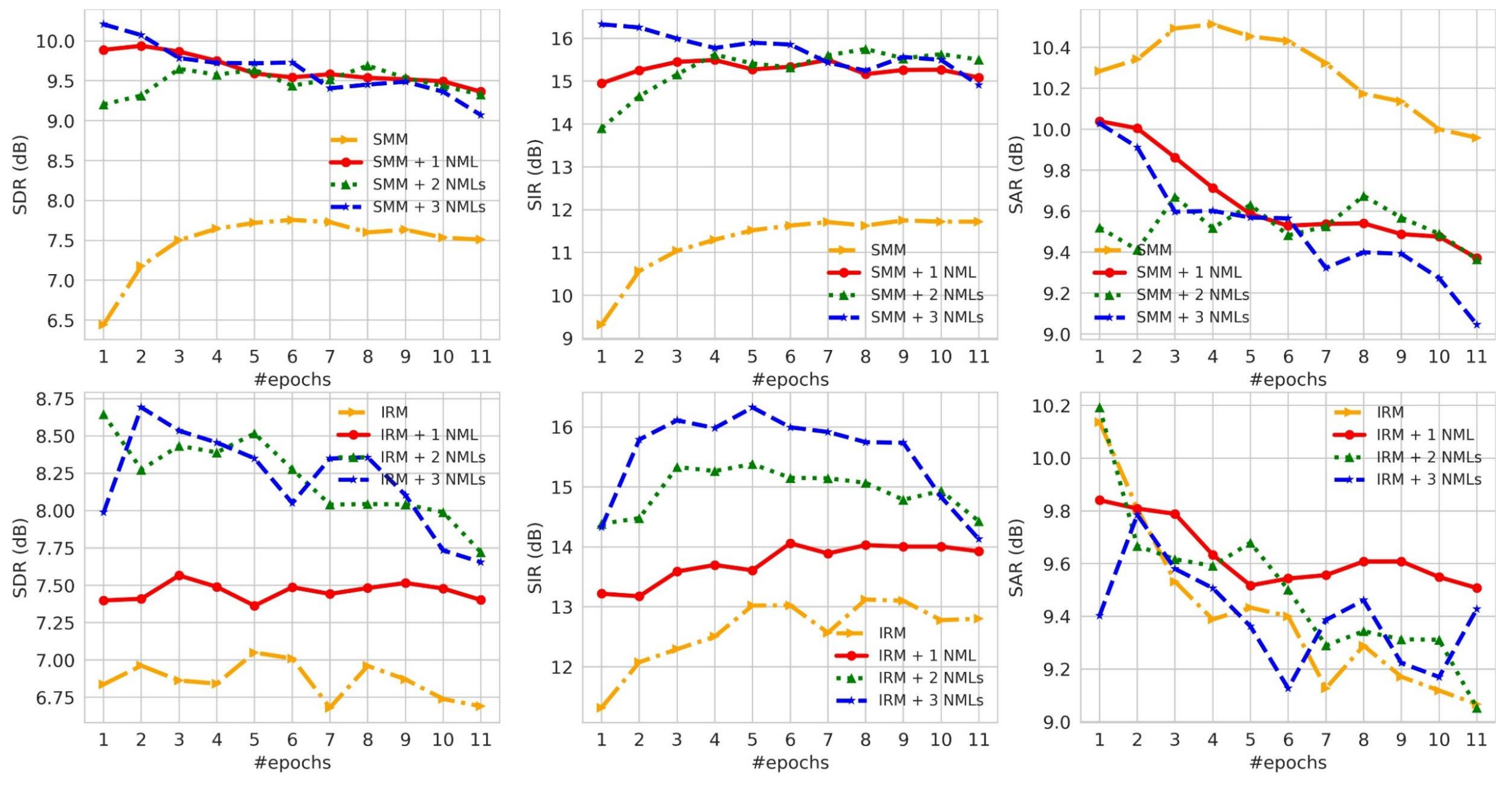

$\hat{\mathbf{M}}^{[2](1)}$
$\hat{\mathbf{M}}^{[2](2)}$
NML #2
Hidden time-frequency masks
SMMs
$\hat{\mathbf{M}}^{[1](1)}$
$\hat{\mathbf{M}}^{[1](2)}$
NML #1
Original time-frequency masks
SMMs
$\tilde{\mathbf{Y}}^{(1)}$
$\tilde{\mathbf{Y}}^{(2)}$

$$\hat{\mathrm{M}}^{[l](i)} = \begin{cases} g\left(\mathbf{W}^{[l]}\frac{\alpha^{(i)}|\tilde{\mathbf{Y}}^{(i)}|}{\sum_{j=1}^{N}\alpha^{(j)}|\tilde{\mathbf{Y}}^{(j)}|} + \mathbf{b}^{[l]}\right) & l = 1 \\ g\left(\mathbf{W}^{[l]}\frac{\alpha^{(i)}|\hat{\mathbf{M}}^{[l-1](i)}|}{\sum_{j=1}^{N}\alpha^{(j)}|\hat{\mathbf{M}}^{[l-1](j)}|} + \mathbf{b}^{[l]}\right) & 1 < l \leq L \end{cases}$$

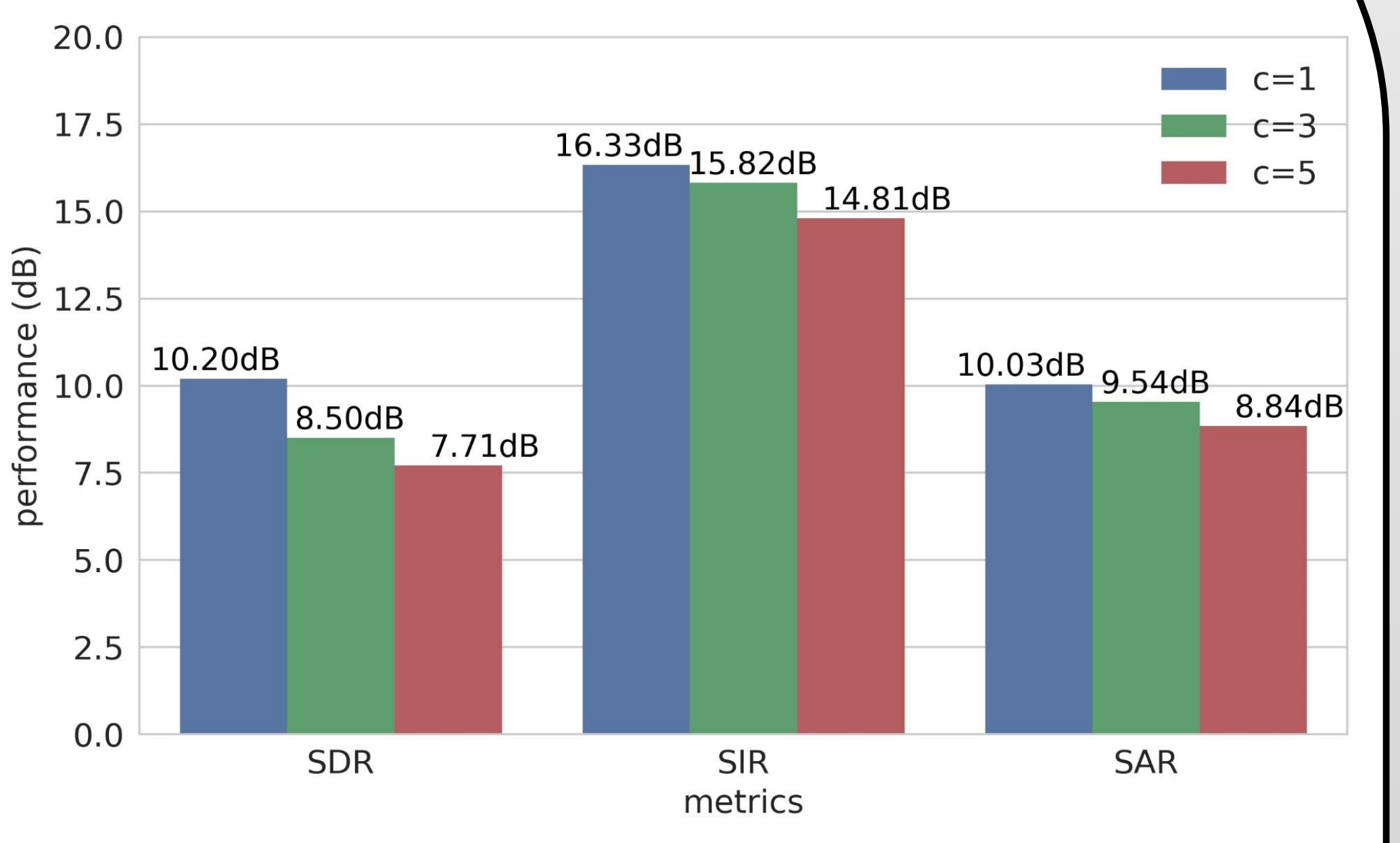
- N: the number of sources
- L : the number of NMLs
- $\mathbf{W}^{[l]}, \mathbf{b}^{[l]}$: the trainable parameters of the lth NML.
- $\tilde{\mathbf{Y}}^{(i)}$: the ith level 1 estimated source
- $\hat{\mathbf{M}}^{[l](i)}$: the post-processed mask for the ith source in the lth NML.
- $g(*)$: the activation function
- $\alpha^{[i]}$: the control parameter for the ith source

- The input features are the spectral magnitudes of the level 1 estimation if $l = 1$.
- By contrast, the input features are the hidden post-processed masks from the previous nonlinear masking layer if $1 < l \leq L$.
- In general, $L \geq 1$, but in particular, we say $L = 0$ means that there are no nonlinear masking layers, i.e., only the original masks.

Experiments



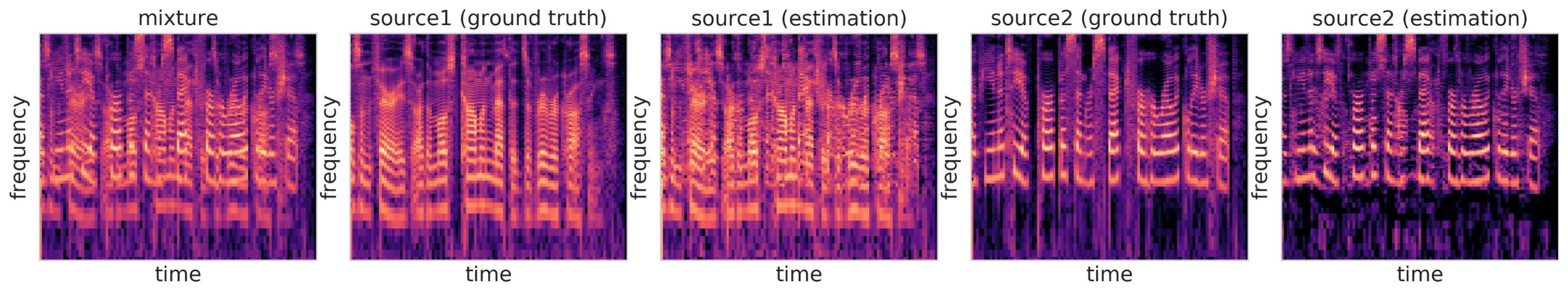
OM	#NMLs	$\alpha^{[l](1)}$	$\alpha^{[l](2)}$	SDR	SIR	SAR
None	None	None	None	6.18	8.97	8.02
SMM	$L = 0$	None	None	7.76	11.75	**10.51**
SMM	$L = 1$	$\alpha^{[1](1)} = 1.0$	$\alpha^{[1](2)} = 1.0$	9.69	15.50	10.03
SMM	$L = 2$	$\alpha^{[1](1)} = 1.0$ $\alpha^{[2](1)} = 1.5$	$\alpha^{[1](2)} = 1.0$ $\alpha^{[2](2)} = 1.5$	9.94	15.75	9.68
SMM	$L = 3$	$\alpha^{[1](1)} = 1.0$ $\alpha^{[2,3](1)} = 1.5$	$\alpha^{[1](2)} = 1.0$ $\alpha^{[2,3](2)} = 1.5$	**10.20**	**16.33**	10.03
IRM	$L = 0$	None	None	7.05	13.12	10.13
IRM	$L = 1$	$\alpha^{[1](1)} = 1.0$	$\alpha^{[1](2)} = 1.0$	7.57	14.06	9.84
IRM	$L = 2$	$\alpha^{[1](1)} = 1.0$ $\alpha^{[2](1)} = 1.5$	$\alpha^{[1](2)} = 1.0$ $\alpha^{[2](2)} = 1.5$	8.64	15.38	10.19
IRM	$L = 3$	$\alpha^{[1](1)} = 1.0$ $\alpha^{[2,3](1)} = 1.5$	$\alpha^{[1](2)} = 1.0$ $\alpha^{[2,3](2)} = 1.5$	8.69	**16.33**	10.27



- The NML accelerates the training procedure of the model especially for SDR and SIR.
- The models with multiple NMLs achieve much better SDRs and SIRs than those with original masks.
- The effect of using the NML is not necessarily better than that of not using the NML for SAR.

- “OM” denotes the type of the original T-F masks.
- Both SDRs and SIRs are improved with the increasement of the number of NMLs no matter what kind of original mask is used, by contrast, SARs maintain relatively stable.
- The model with SMMs followed by 3 NMLs obtains both best SDR and SIR.

- Increasing the size of the context window (c = 1, 3, 5) may harm the performance of the model due to overfitting possibly.



- An utterance example: the spectrograms of the real sources are compared to those of the estimated sources.
- The model “RNN + SMMs + 3 NMLs” generates relatively good results since the spectral representations of the estimated sources are quite closed to those of the real sources.

Other contact Information: JiantaoChaserLLer (WeChat)