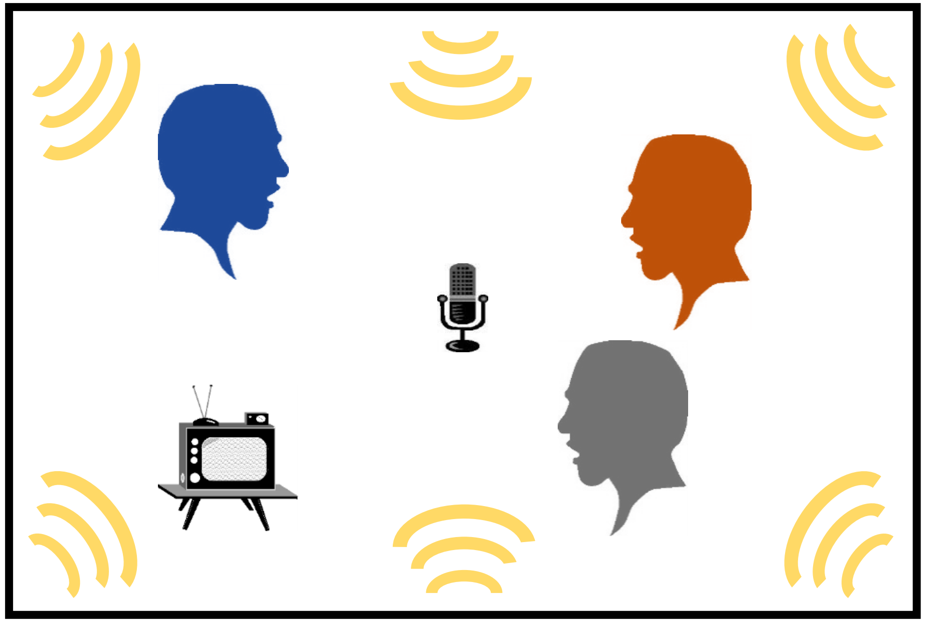
ALL-NEURAL ONLINE SOURCE SEPARATION, COUNTING, AND DIARIZATION FOR MEETING ANALYSIS

Thilo von Neumann[1,2], Keisuke Kinoshita[1], Marc Delcroix[1], Shoko Araki[1], Tomohiro Nakatani[1], Reinhold Haeb-Umbach[2]

[1] NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[2] Paderborn University, Department of Communications Engineering, Paderborn, Germany

Tackling Problems in Meeting Scenarios



Problems:

- multiple sources ⇒source separation
- number of sources not known ⇒source count estimation
- long recordings ⇒blockwise/online processing

NN-based Source Separation Methods

Permutation Invariant Training (PIT) [Kolbaek2017]

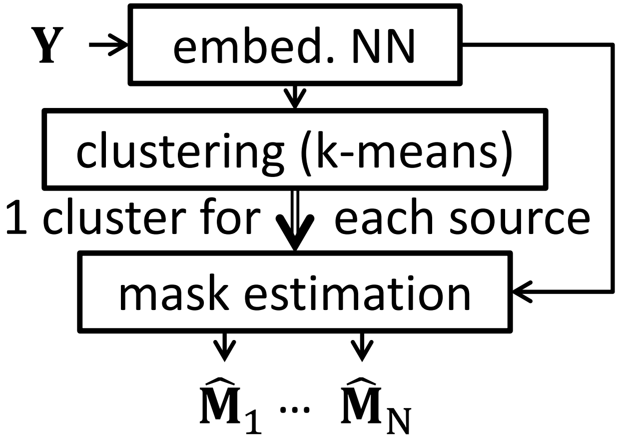
- **purely neural network based**
- model **structure depends on number of sources** to be estimated

~ **max number** of sources must be known **during training**

~ source counting

Deep Clustering (DC) **/ Deep Attractor Networks** (DAN) [Isik2016]/[Chen2017]

- **2 stages**: embedding + clustering

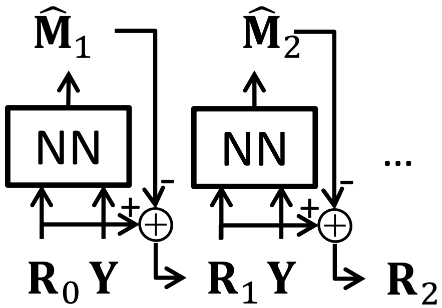


~ **number** of sources must be **known or estimated** separately

~ source counting

Recurrent Selective Attention Network (RSAN) [Kinoshita2018]

- **purely neural network based**
- **iterative** source extraction

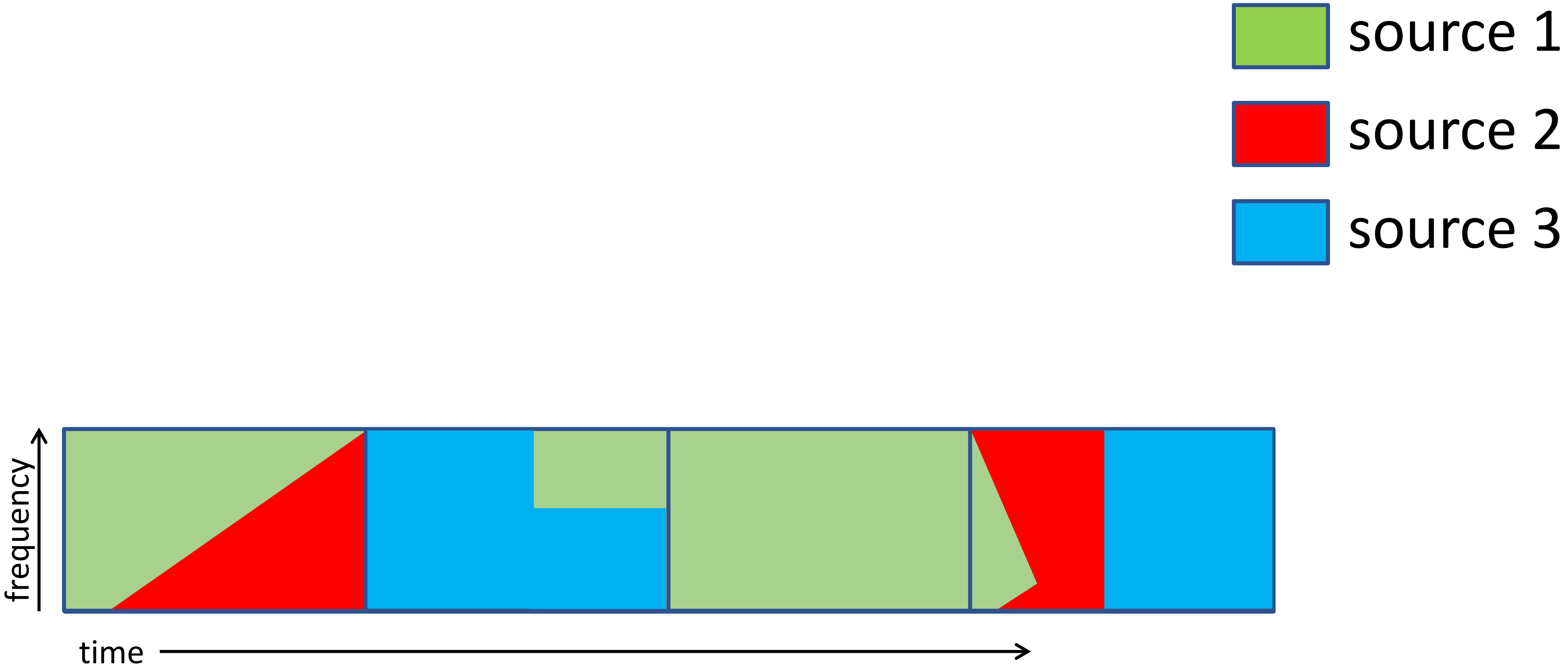


✓ source separation for **arbitrary number** of sources

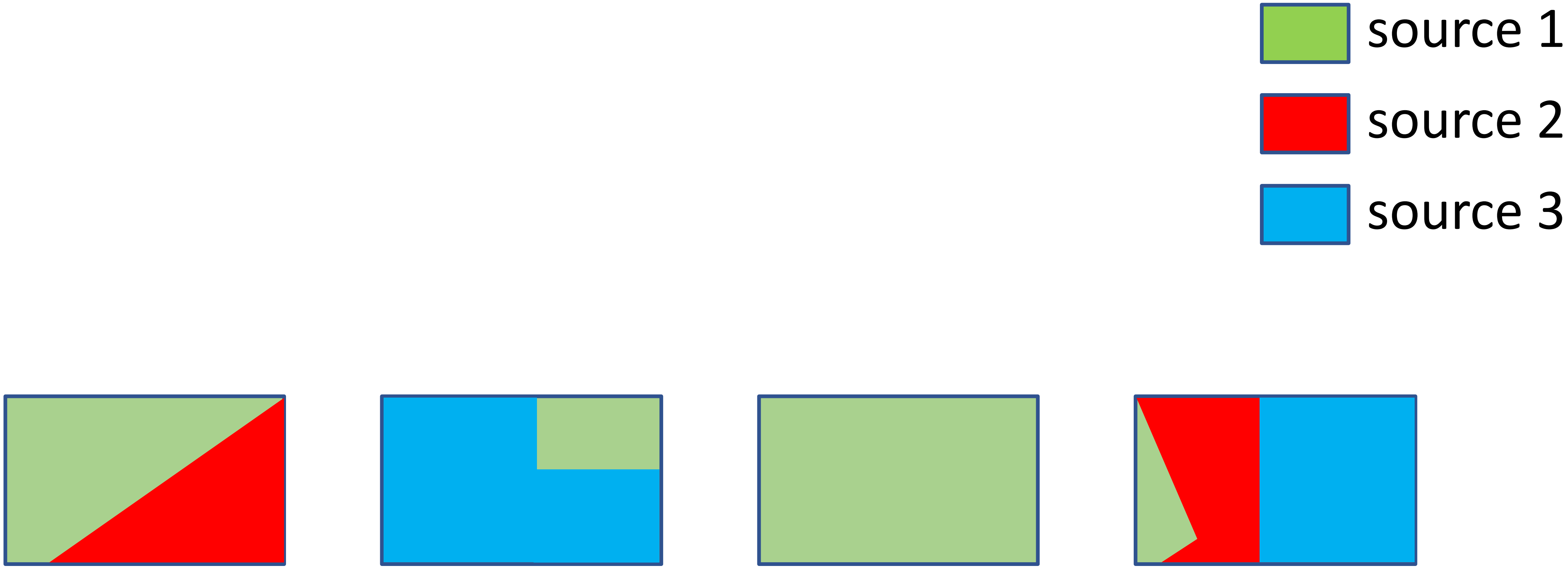
✓ source counting

✘ block-online processing

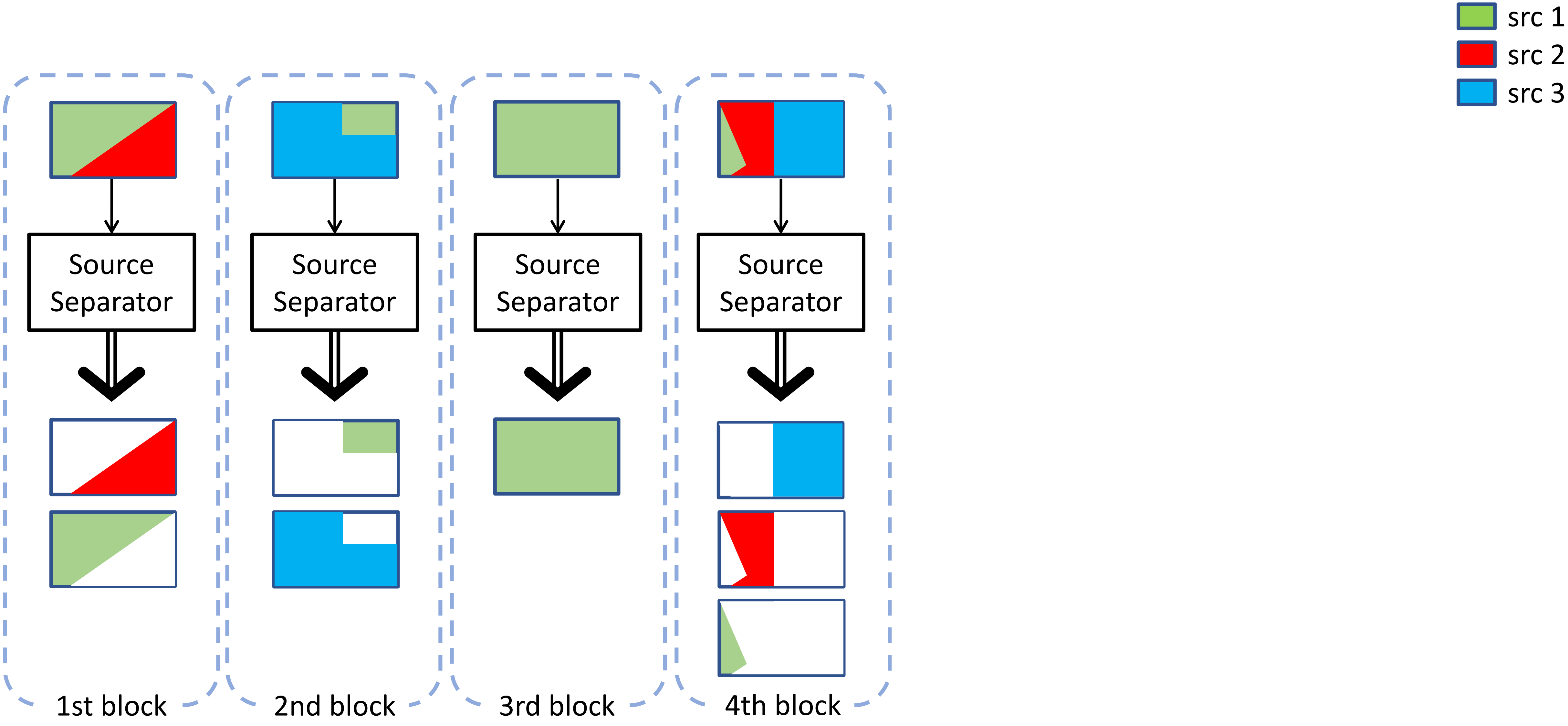
Block-Online Processing



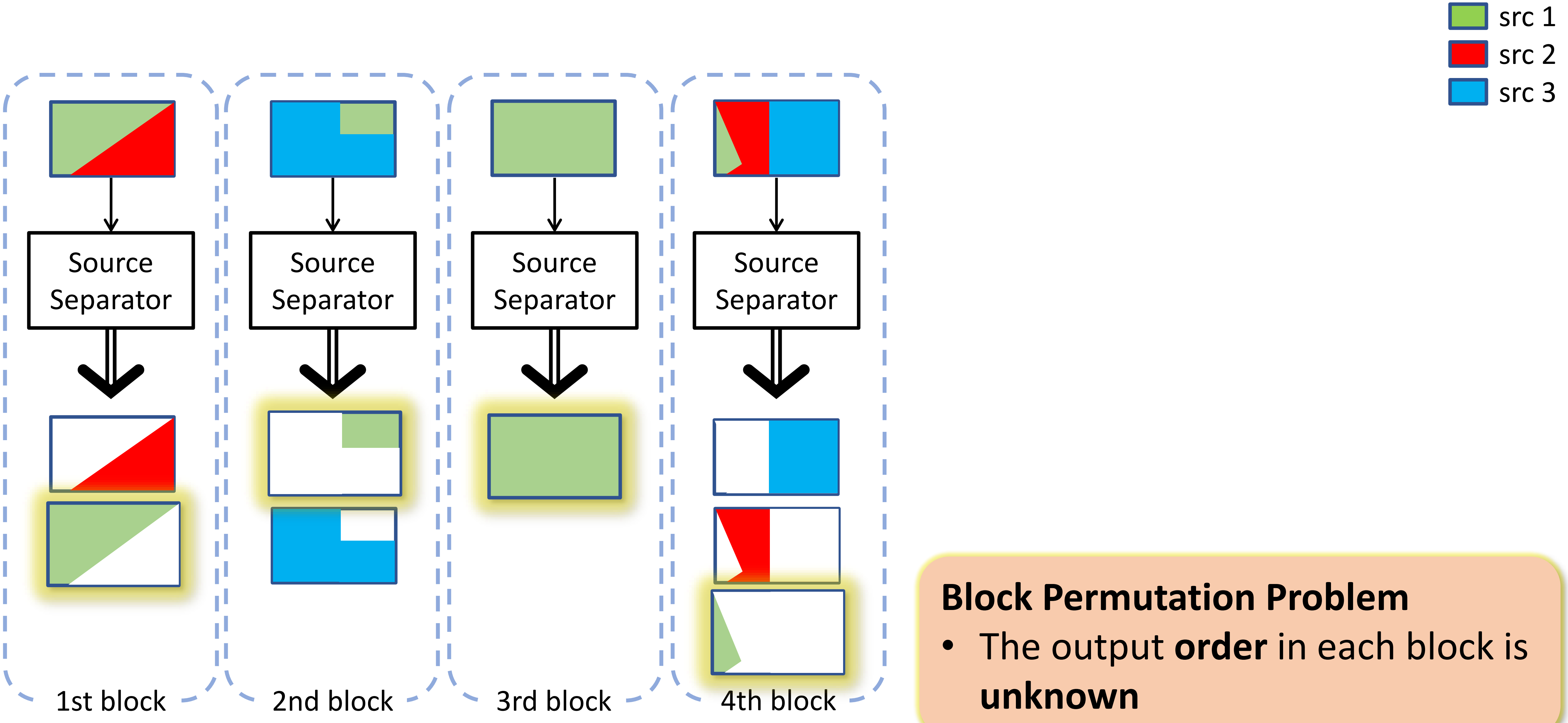
Block-Online Processing



Difficulties in Block-Online Processing



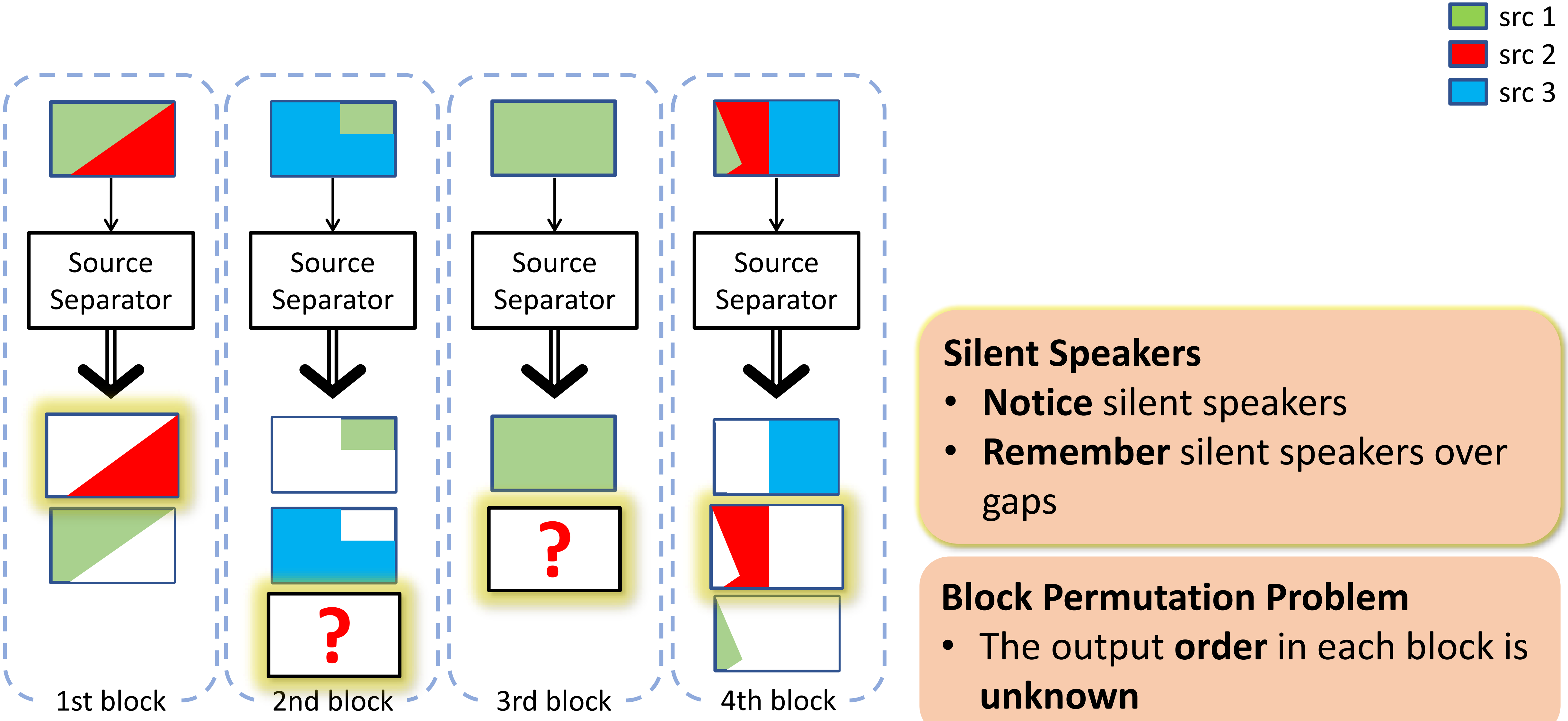
Difficulties in Block-Online Processing



Block Permutation Problem

- The output **order** in each block is **unknown**

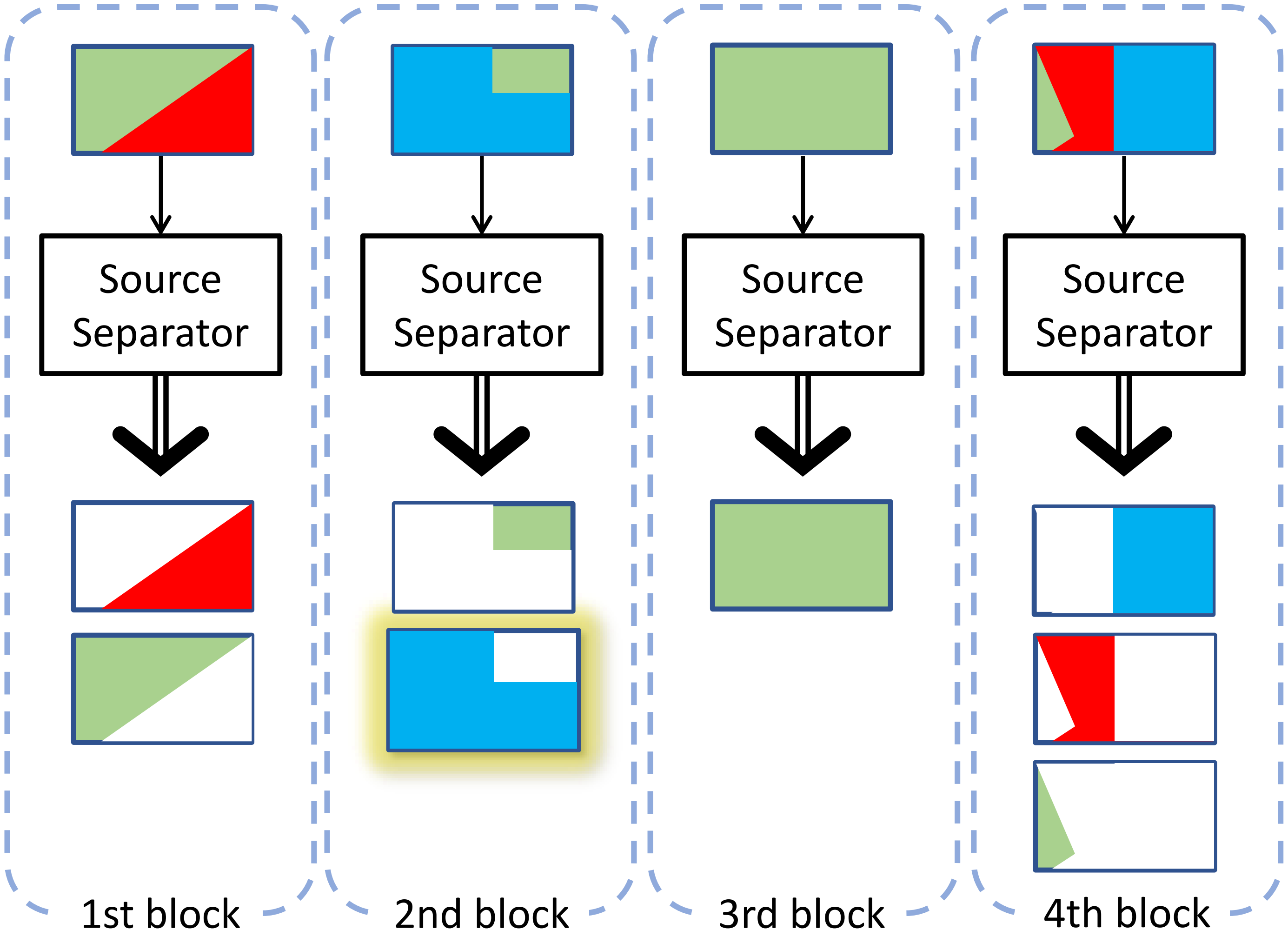
Difficulties in Block-Online Processing

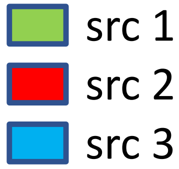


Silent Speakers
- **Notice** silent speakers
- **Remember** silent speakers over gaps

Block Permutation Problem
- The output **order** in each block is **unknown**

Difficulties in Block-Online Processing





New Speakers

- Detect **new speaker in each block**

Silent Speakers

- **Notice** silent speakers
- **Remember** silent speakers over gaps

Block Permutation Problem

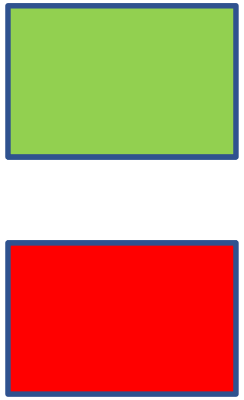
- The output **order** in each block is **unknown**

Recurrent Selective Attention Network

A fully Neural Network based Source Separation and Source Number Counting Approach

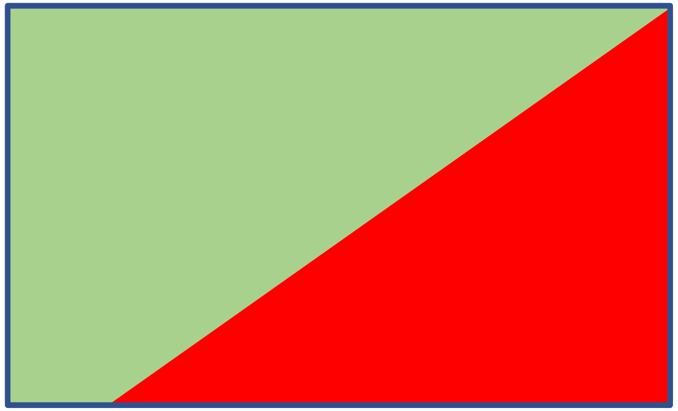
K. Kinoshita, L. Drude, M. Delcroix and T. Nakatani. “Listening to Each Speaker One by One with Recurrent Selective Hearing Networks.” *ICASSP* (2018).

Recurrent Selective Attention Network (RSAN)



source 1

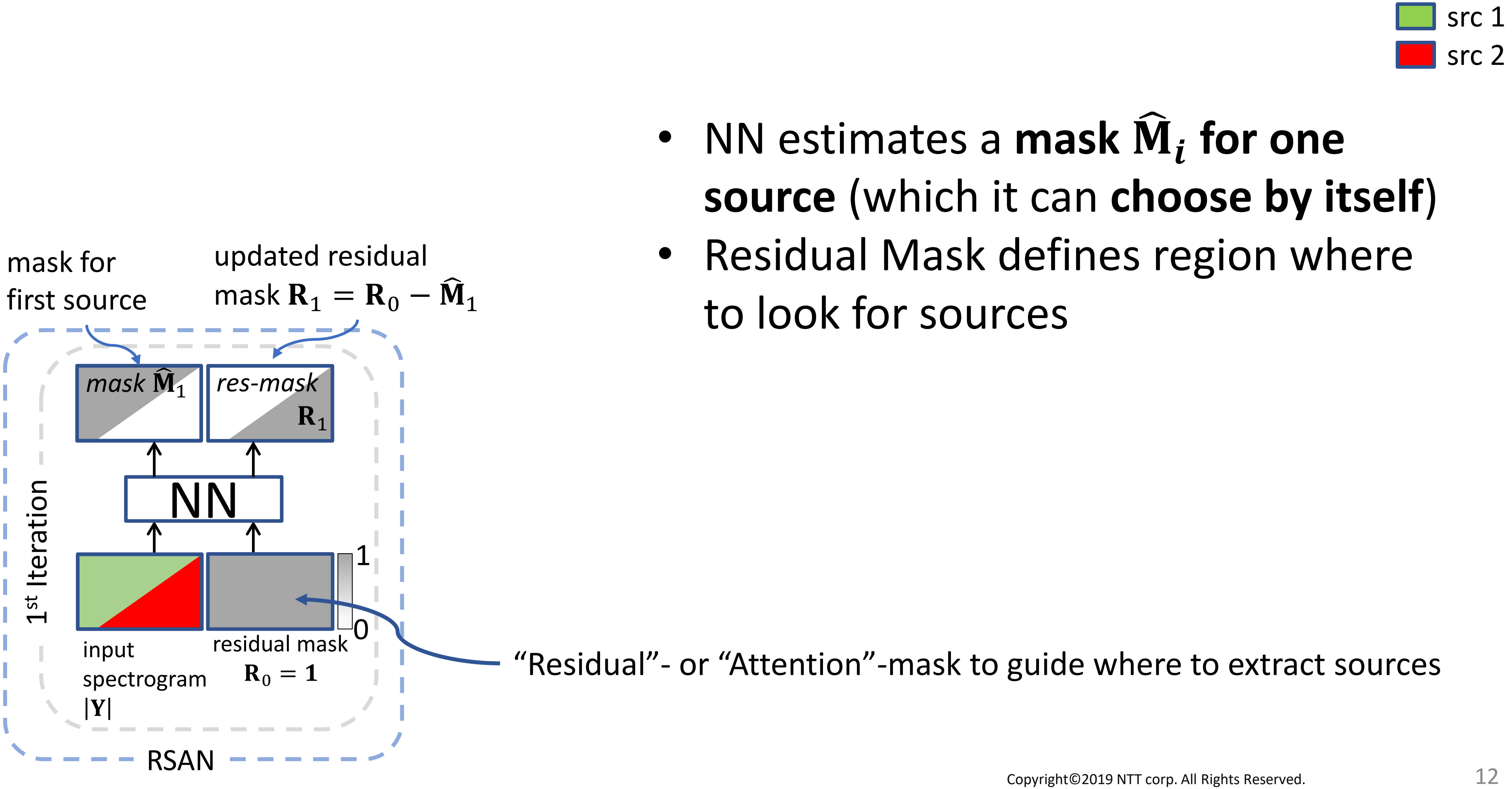
source 2



input spectrogram

$|\mathbf{Y}|$

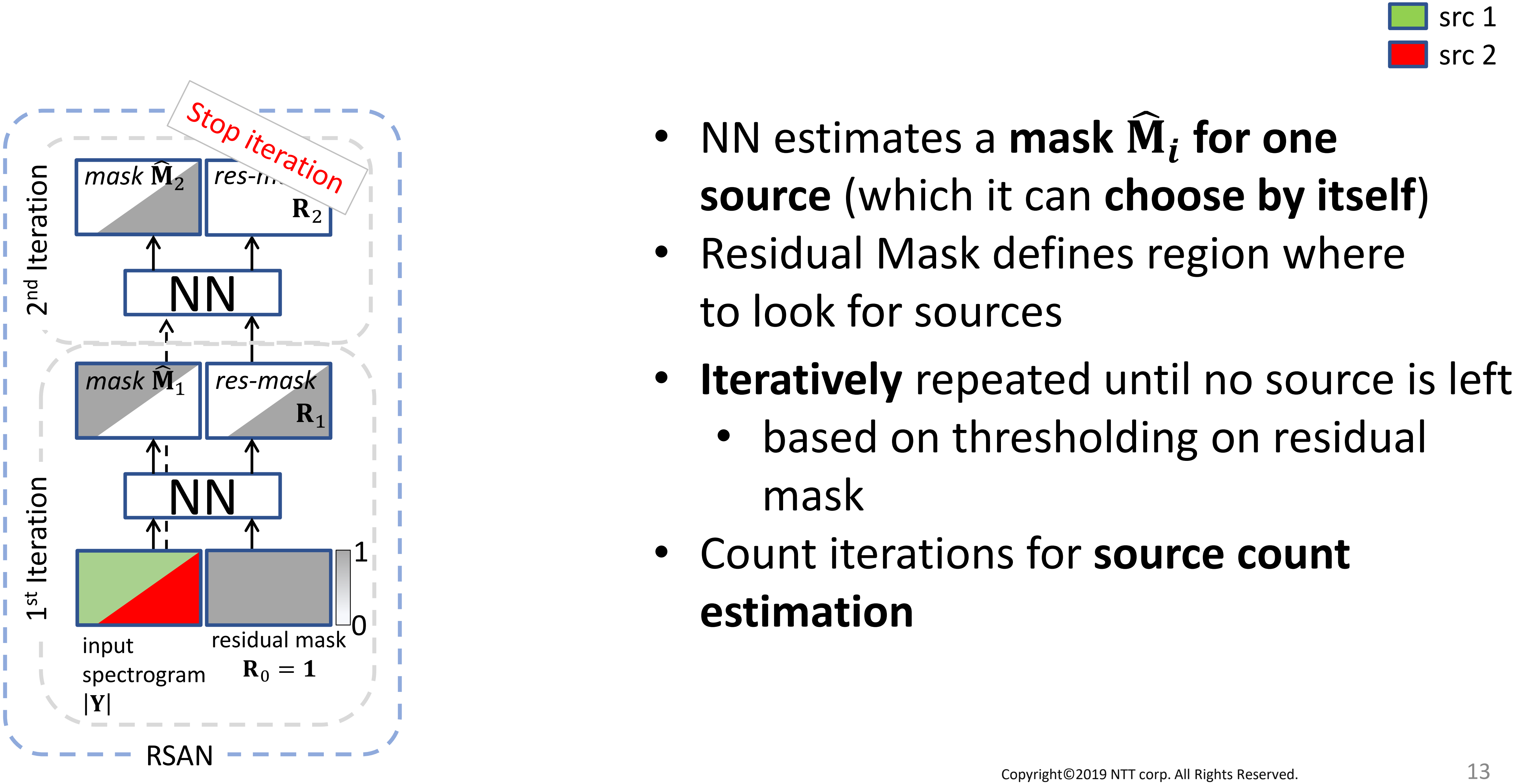
Recurrent Selective Attention Network (RSAN)



- NN estimates a **mask $\widehat{\mathbf{M}}_i$ for one source** (which it can **choose by itself**)
- Residual Mask defines region where to look for sources

"Residual"- or "Attention"-mask to guide where to extract sources

Recurrent Selective Attention Network (RSAN)

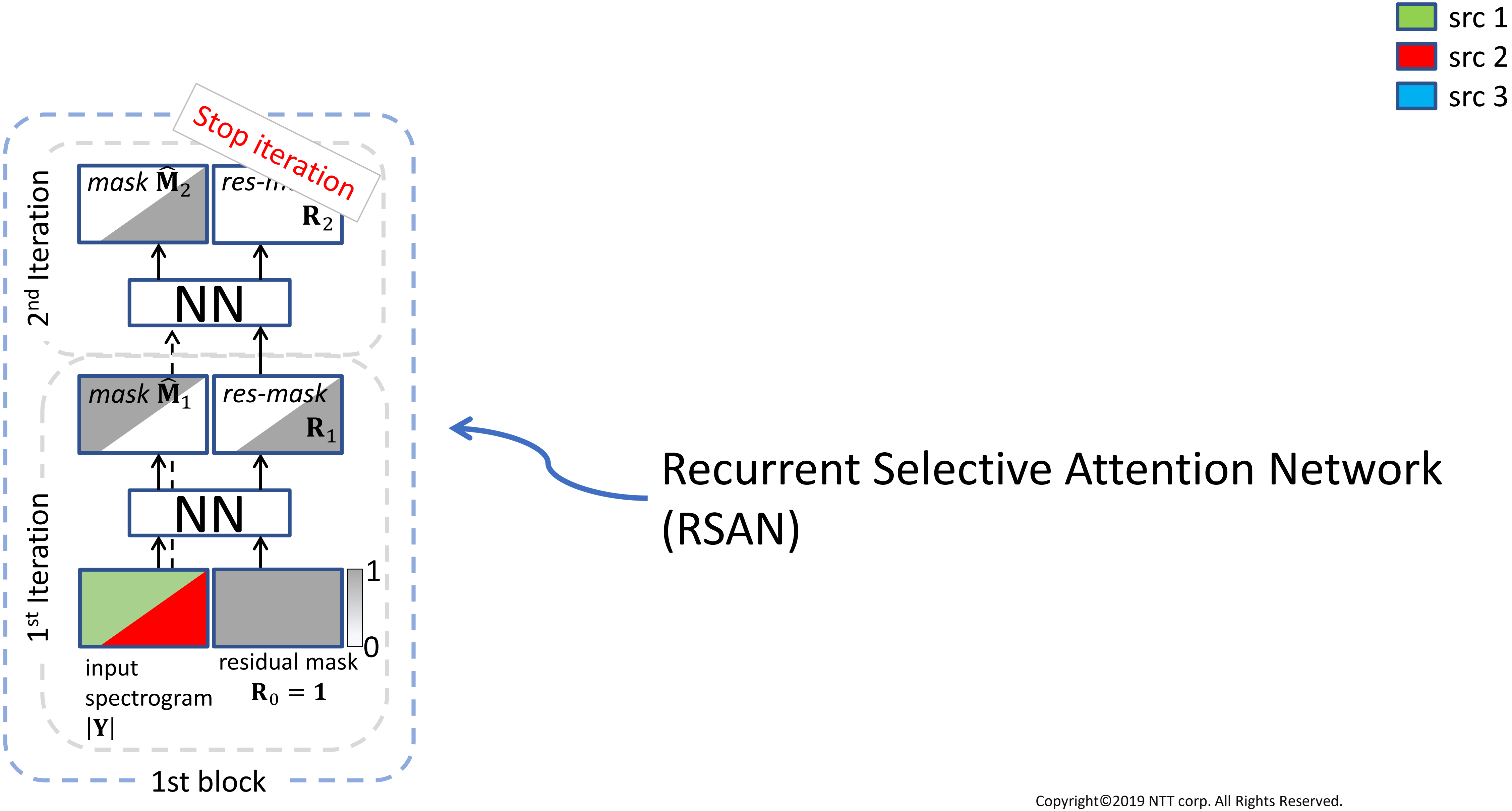


- NN estimates a **mask $\widehat{\mathbf{M}}_i$ for one source** (which it can **choose by itself**)
- Residual Mask defines region where to look for sources
- **Iteratively** repeated until no source is left
  - based on thresholding on residual mask
- Count iterations for **source count estimation**

Proposed Method

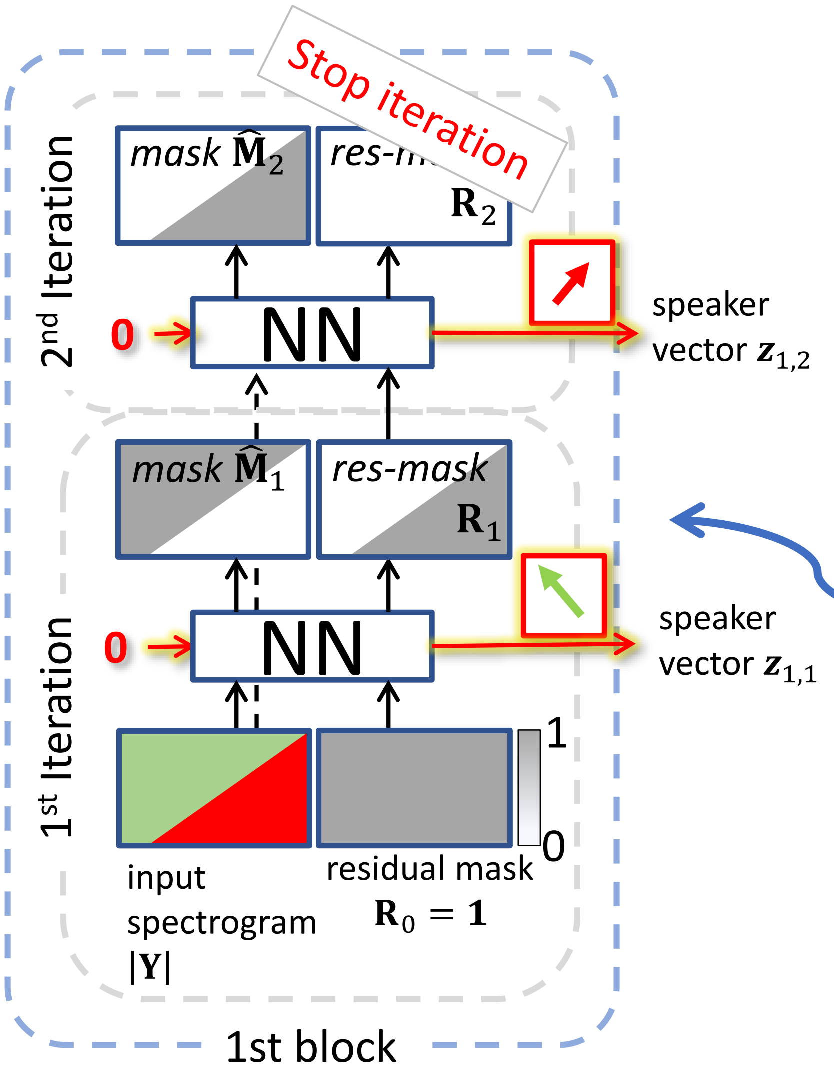
How to do block-online processing with RSAN

Proposed Method



Recurrent Selective Attention Network (RSAN)

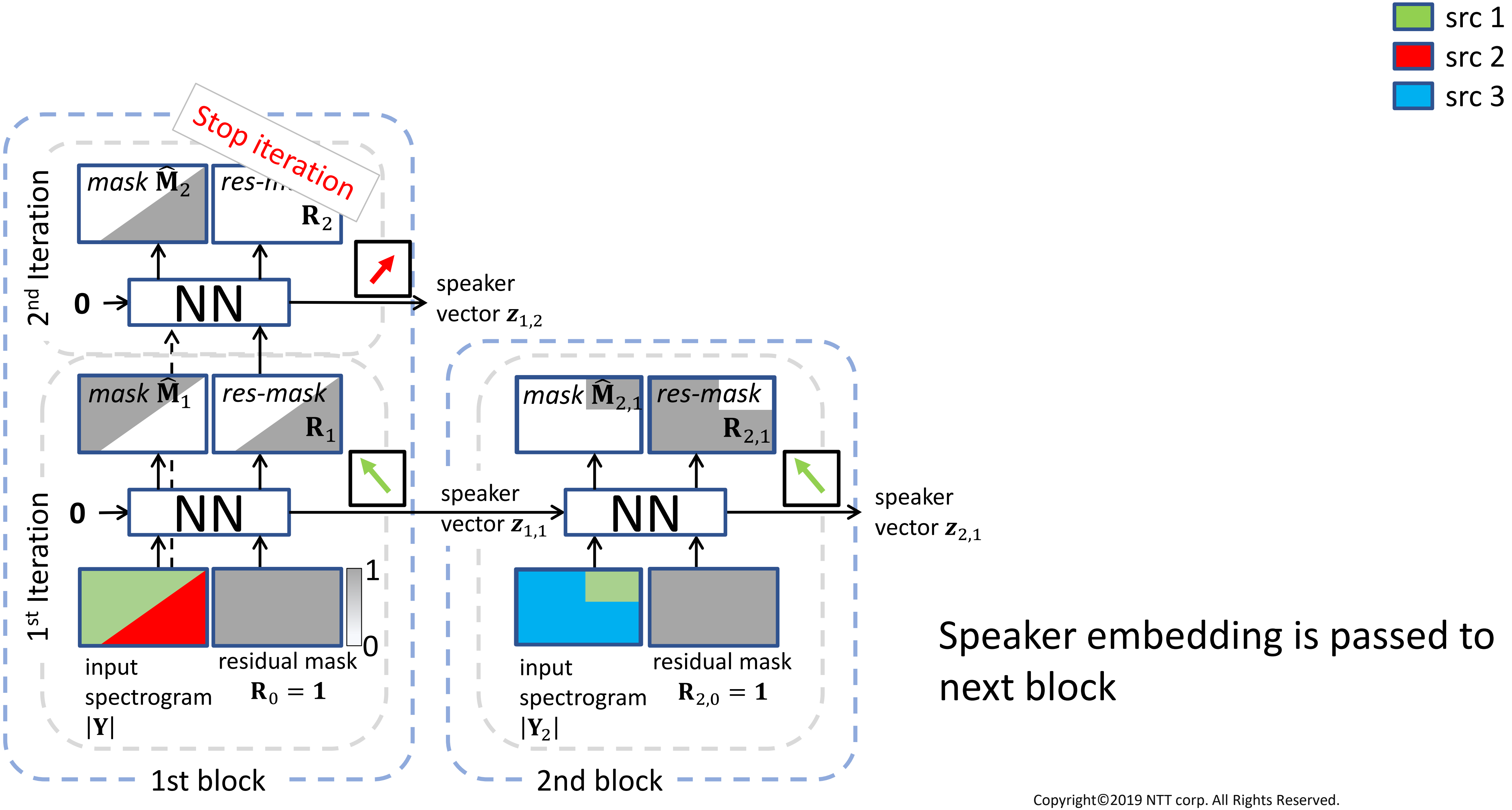
Proposed Method



Recurrent Selective Attention Network (RSAN)

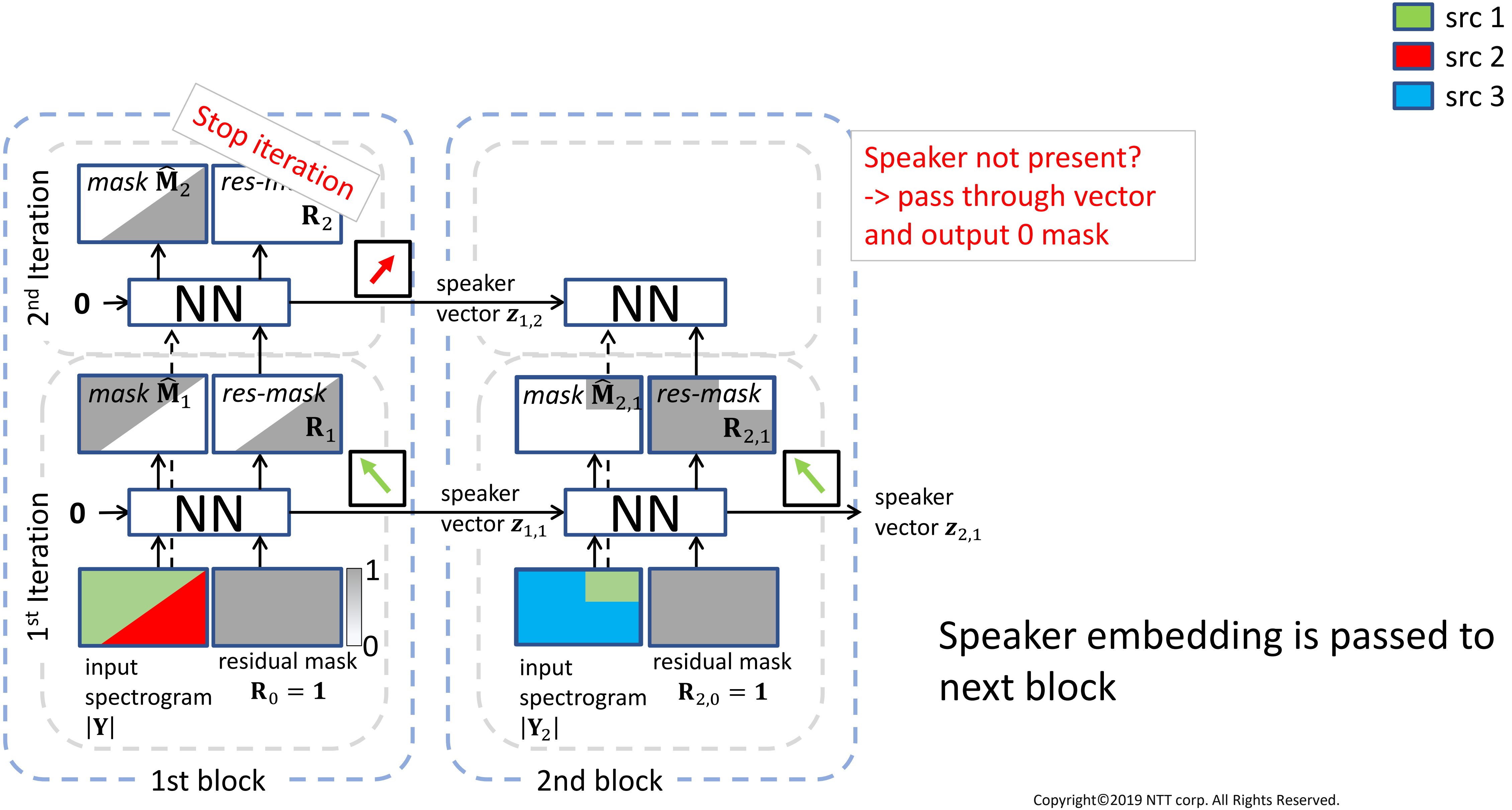
+ speaker embedding input & output

Proposed Method



Speaker embedding is passed to next block

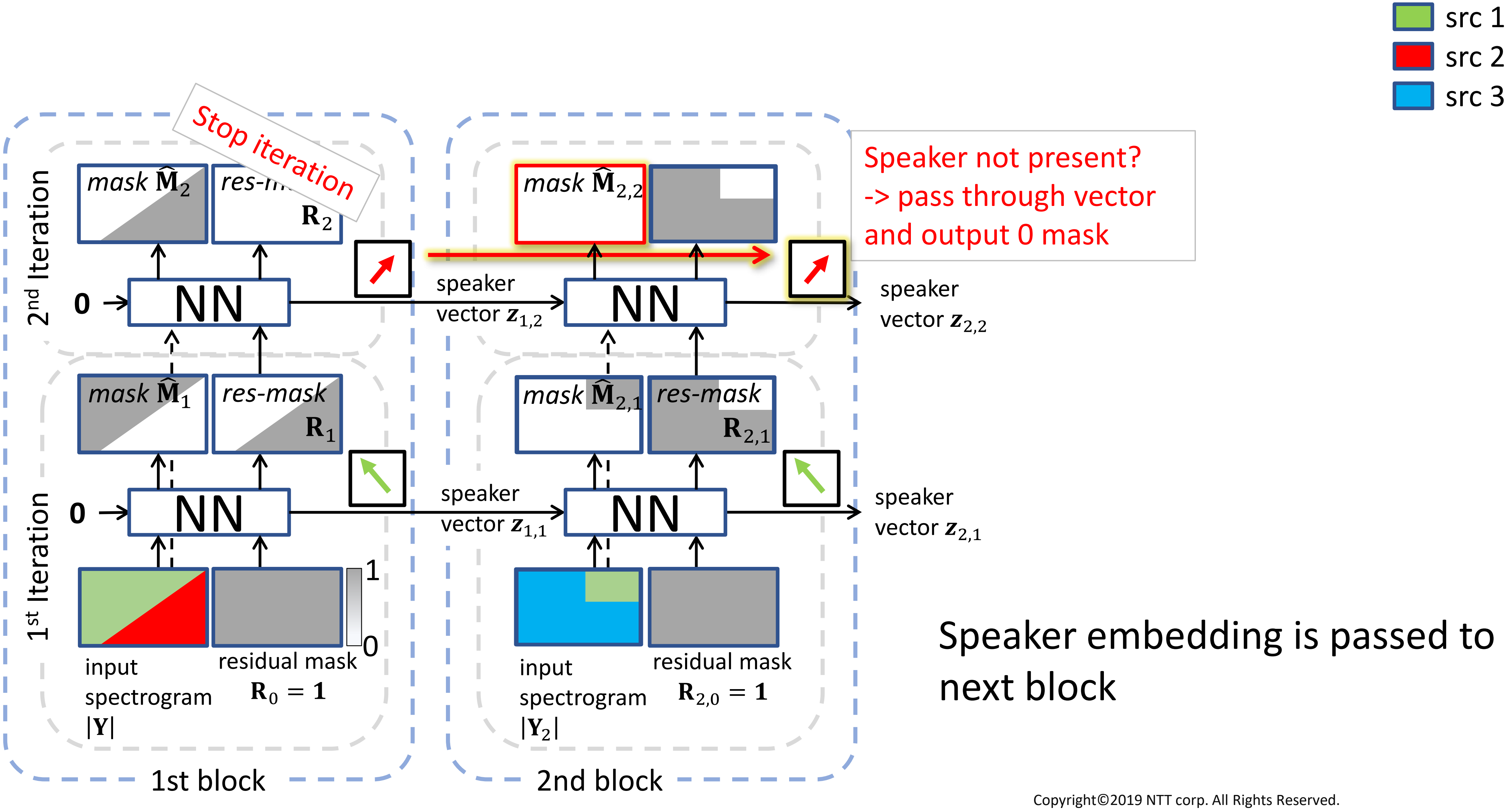
Proposed Method



Speaker not present?
-> pass through vector
and output 0 mask

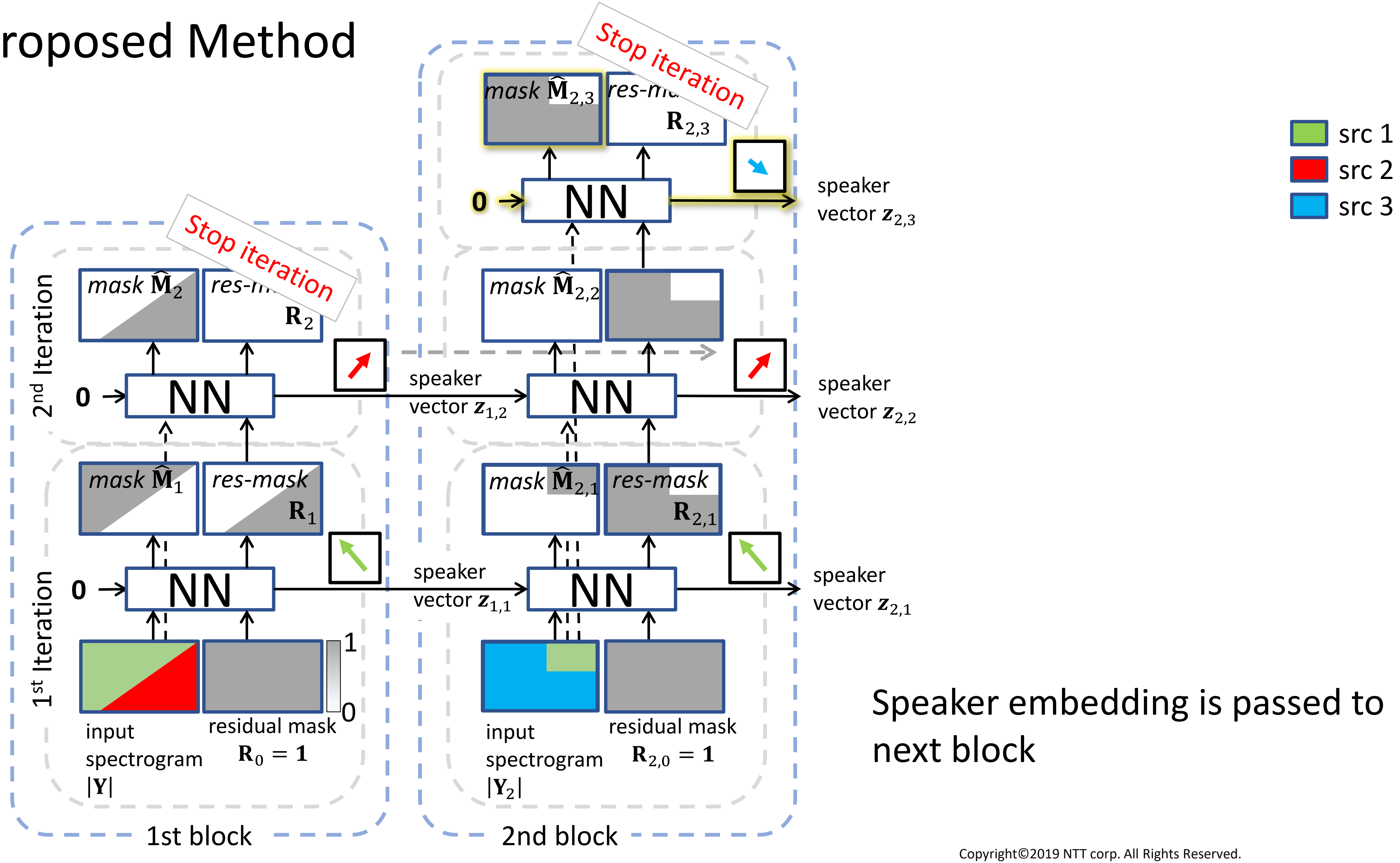
Speaker embedding is passed to next block

Proposed Method



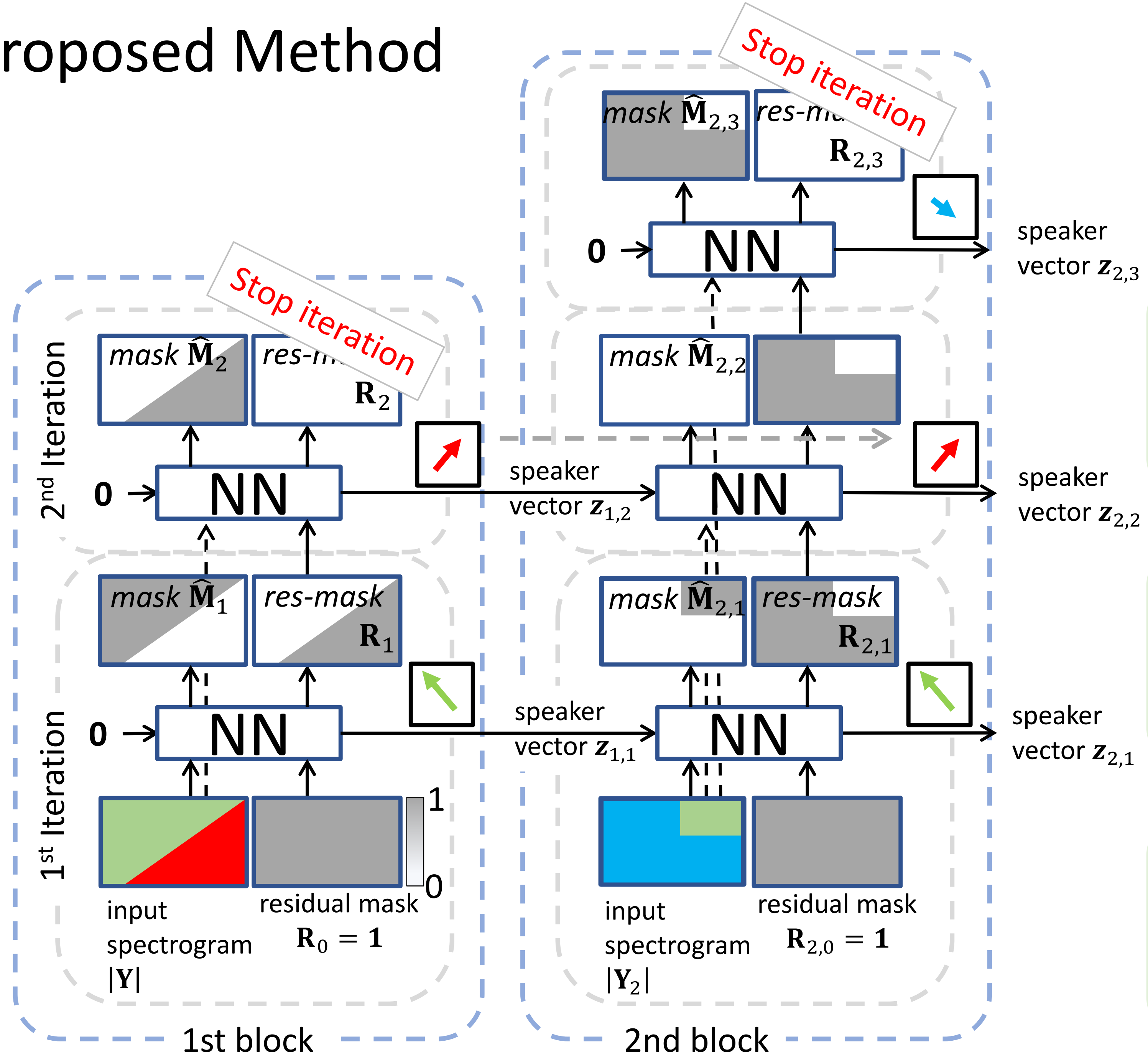
Speaker embedding is passed to next block

Proposed Method



Speaker embedding is passed to next block

Proposed Method

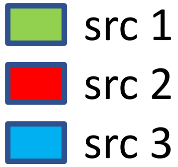


New Speakers
Can estimate **new speaker vector in each block**

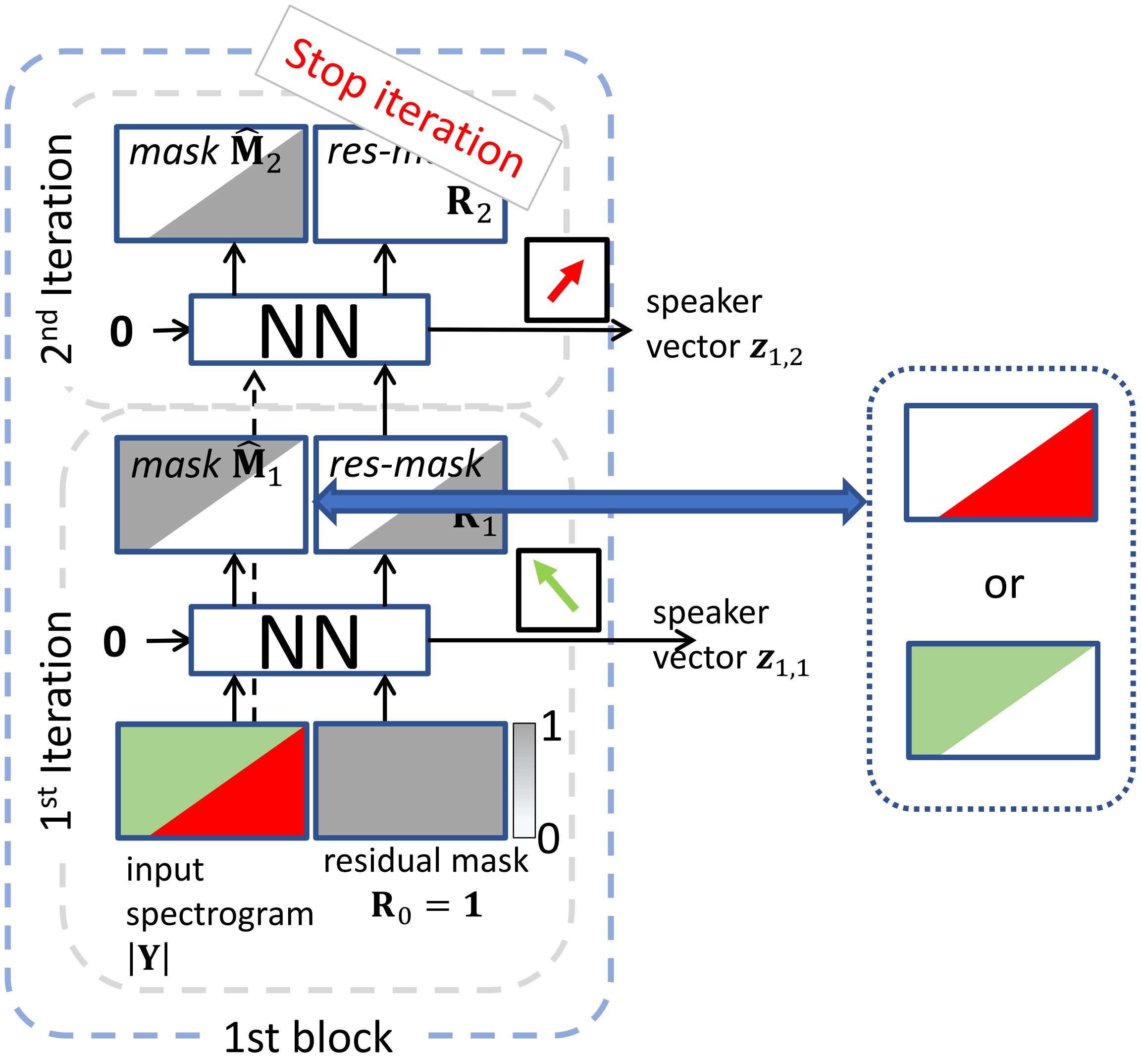
Silent Speakers
Notices silent speakers **and estimates 0** mask

Block Permutation Problem
Estimated signals **always in the same order**

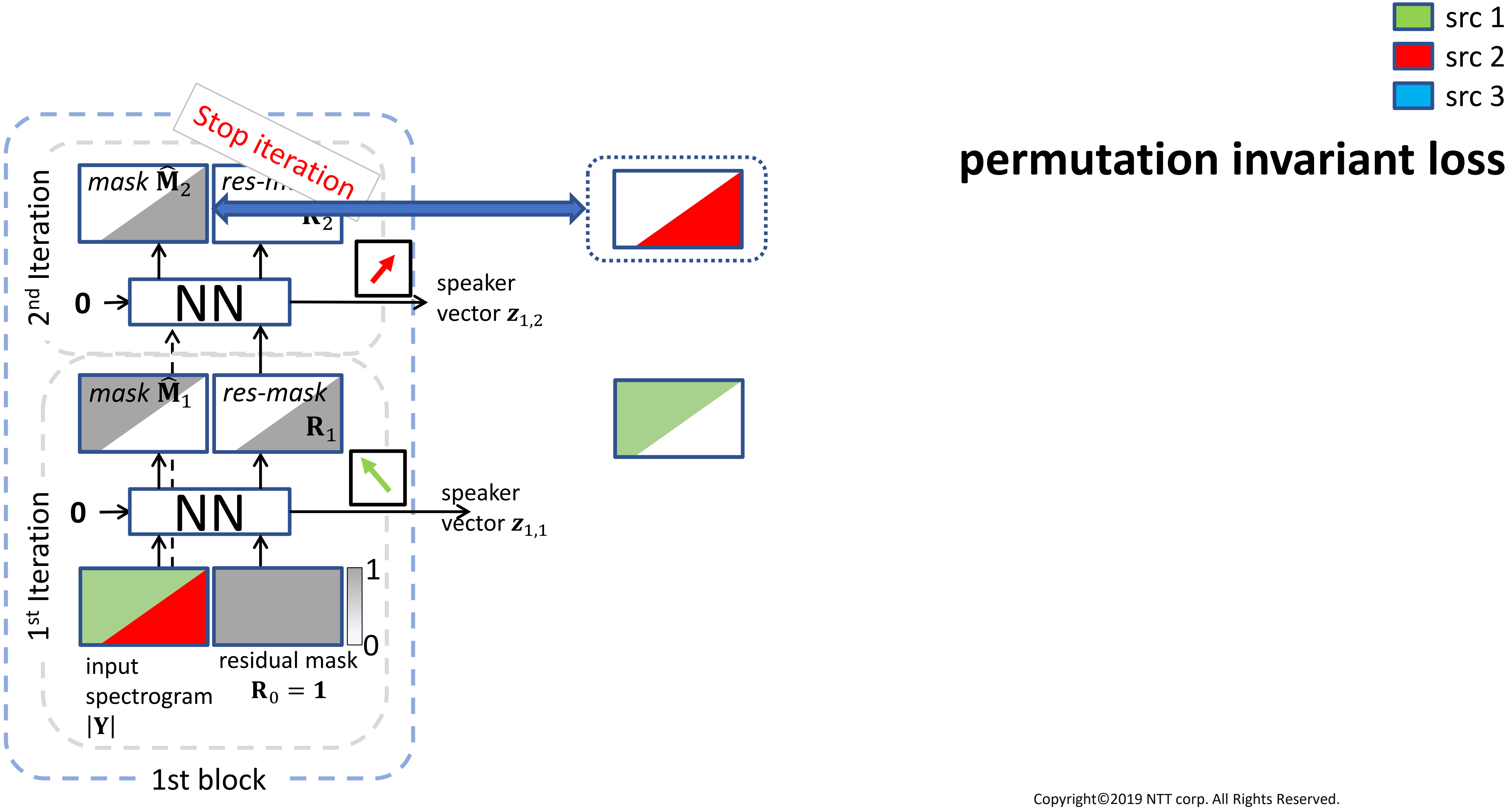
Proposed Method: Loss



permutation invariant loss

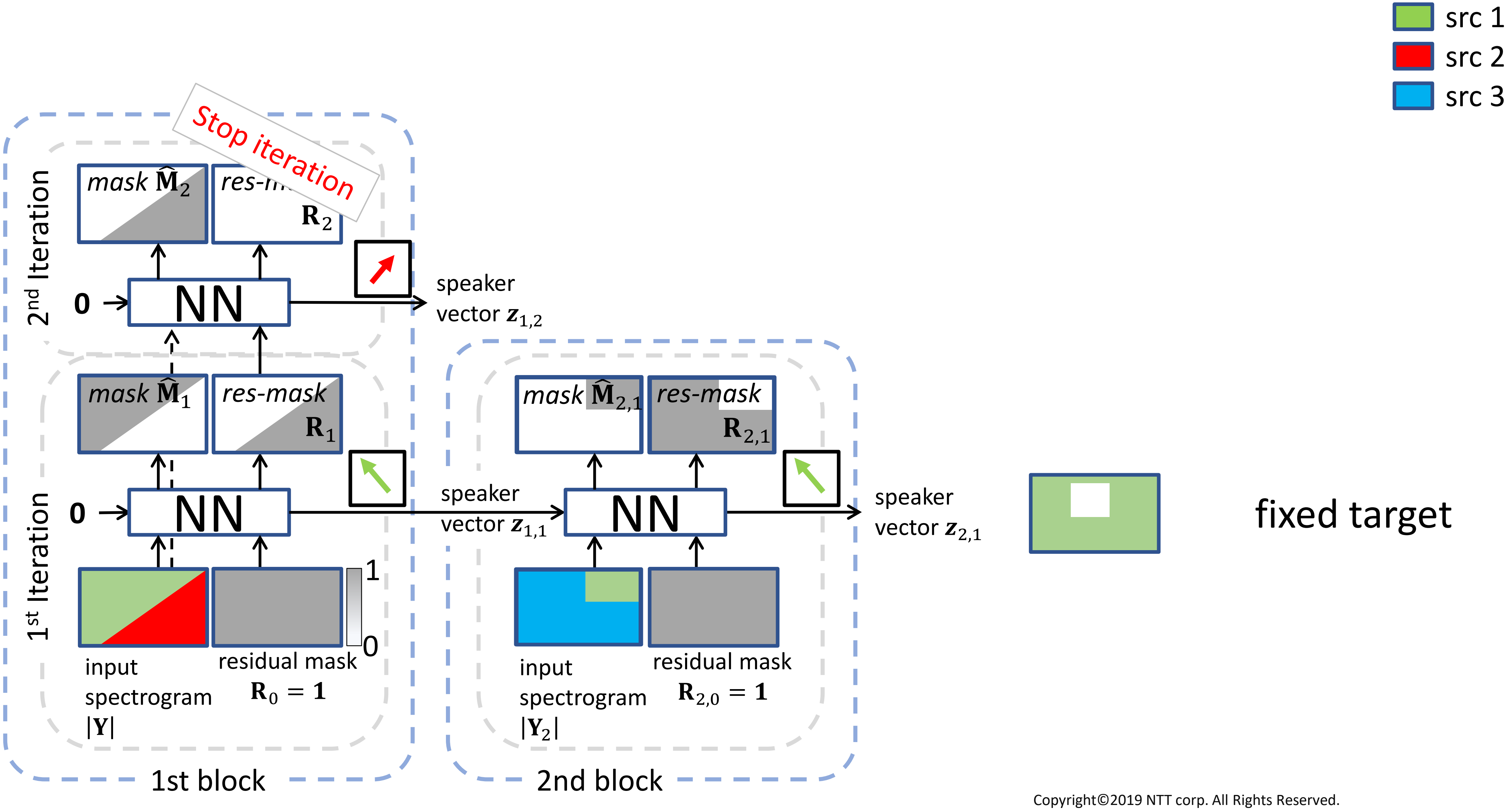


Proposed Method: Loss

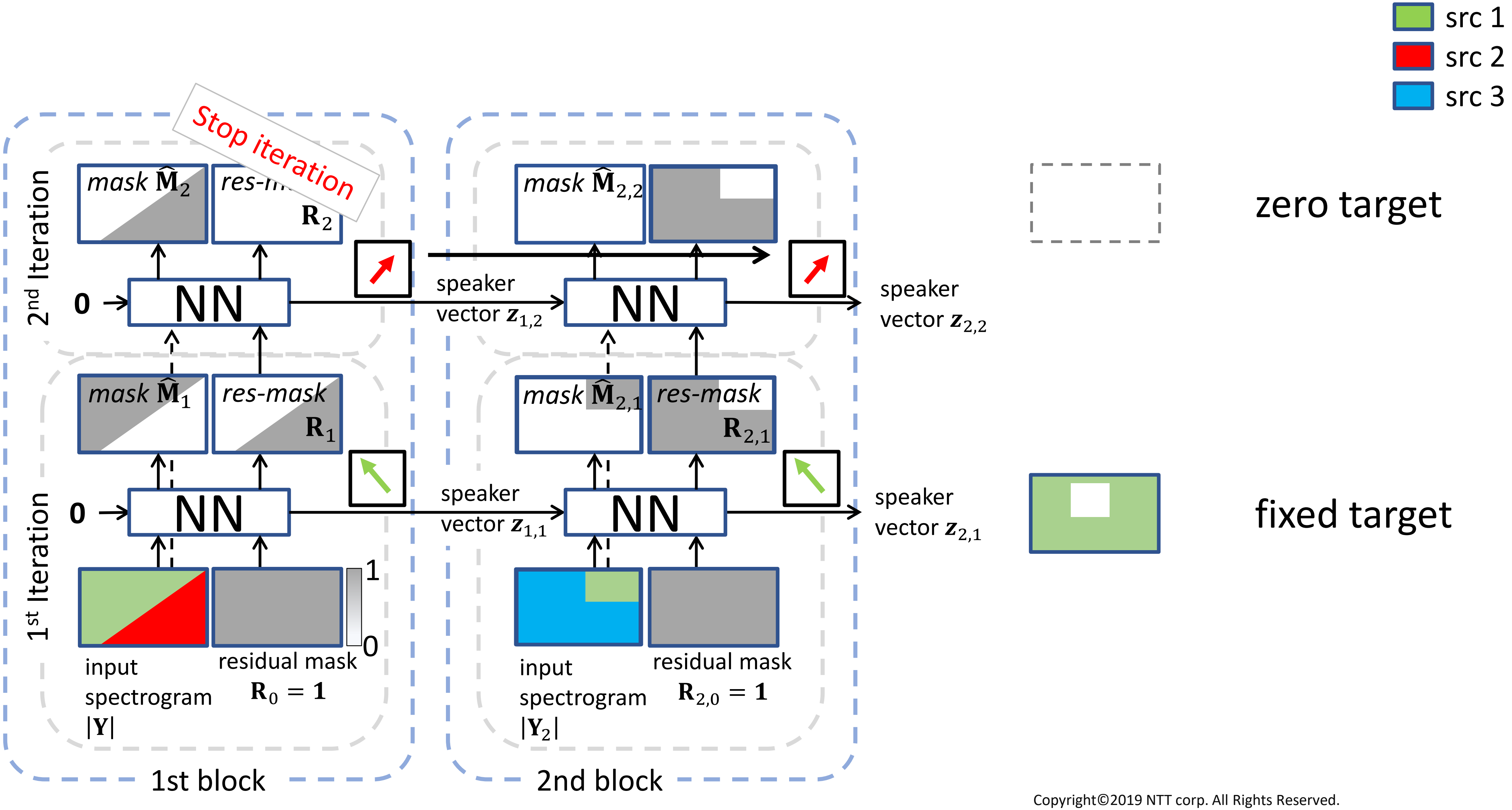


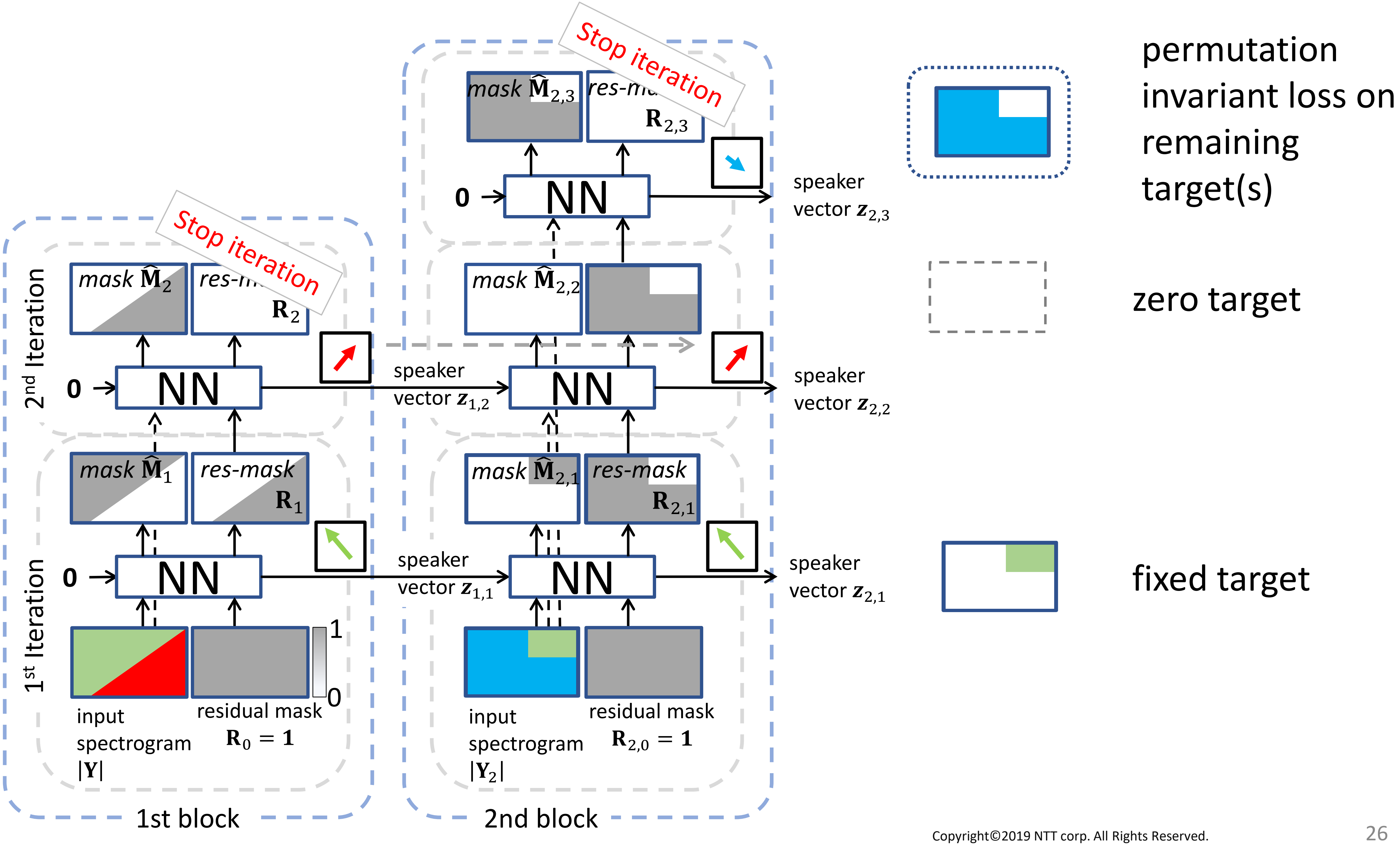
permutation invariant loss

Proposed Method: Loss



Proposed Method: Loss





Stop iteration
mask $\hat{\mathbf{M}}_{2,3}$
res-ma
$\mathbf{R}_{2,3}$
0
NN
speaker vector $\boldsymbol{z}_{2,3}$
Stop iteration
2nd Iteration
mask $\hat{\mathbf{M}}_2$
res-m
$\mathbf{R}_2$
0
NN
speaker vector $\boldsymbol{z}_{1,2}$
mask $\hat{\mathbf{M}}_{2,2}$
NN
speaker vector $\boldsymbol{z}_{2,2}$
mask $\hat{\mathbf{M}}_1$
res-mask $\mathbf{R}_1$
mask $\hat{\mathbf{M}}_{2,1}$
res-mask $\mathbf{R}_{2,1}$
1st Iteration
0
NN
speaker vector $\boldsymbol{z}_{1,1}$
NN
speaker vector $\boldsymbol{z}_{2,1}$
1
0
input spectrogram $|\mathbf{Y}|$
residual mask $\mathbf{R}_0 = \mathbf{1}$
input spectrogram $|\mathbf{Y}_2|$
residual mask $\mathbf{R}_{2,0} = \mathbf{1}$
1st block
2nd block
permutation invariant loss on remaining target(s)
zero target
fixed target

Proposed Method - Experiments

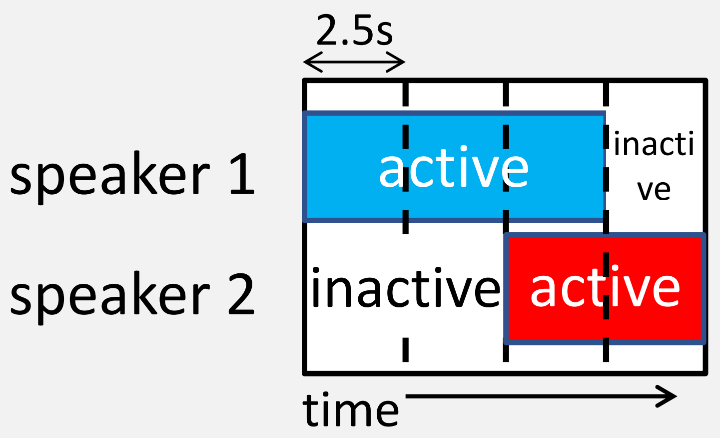
Data

- Artificially mixed from wsj
- 1 or 2 speakers
- Blocks of 2.5s
- Random activity pattern
- ≥ 1 speaker active in first block

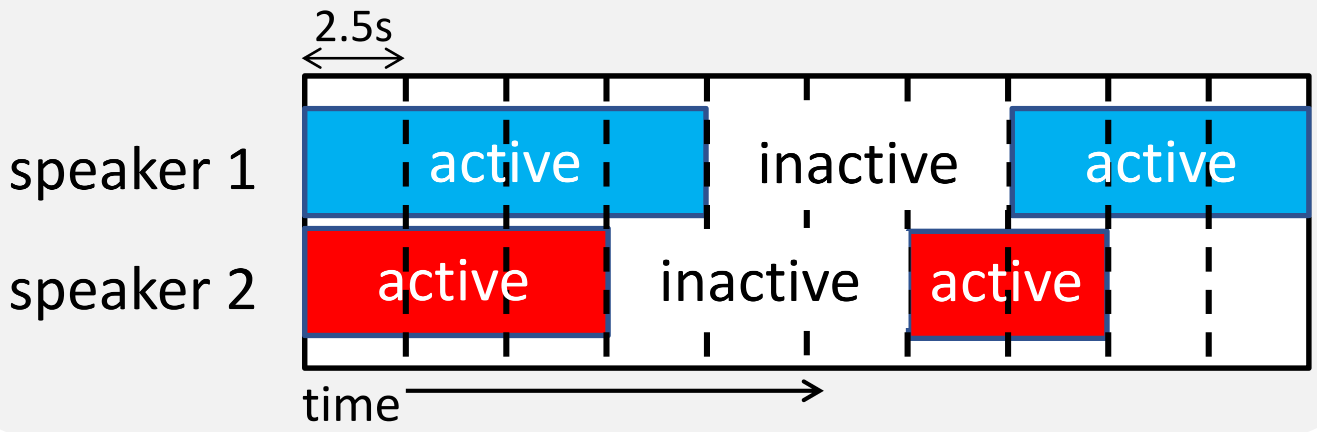
Training Data
- 4 blocks / 10s

Test Data
- 12 blocks / 30s
- ~5 hours
- Considerably longer than training data

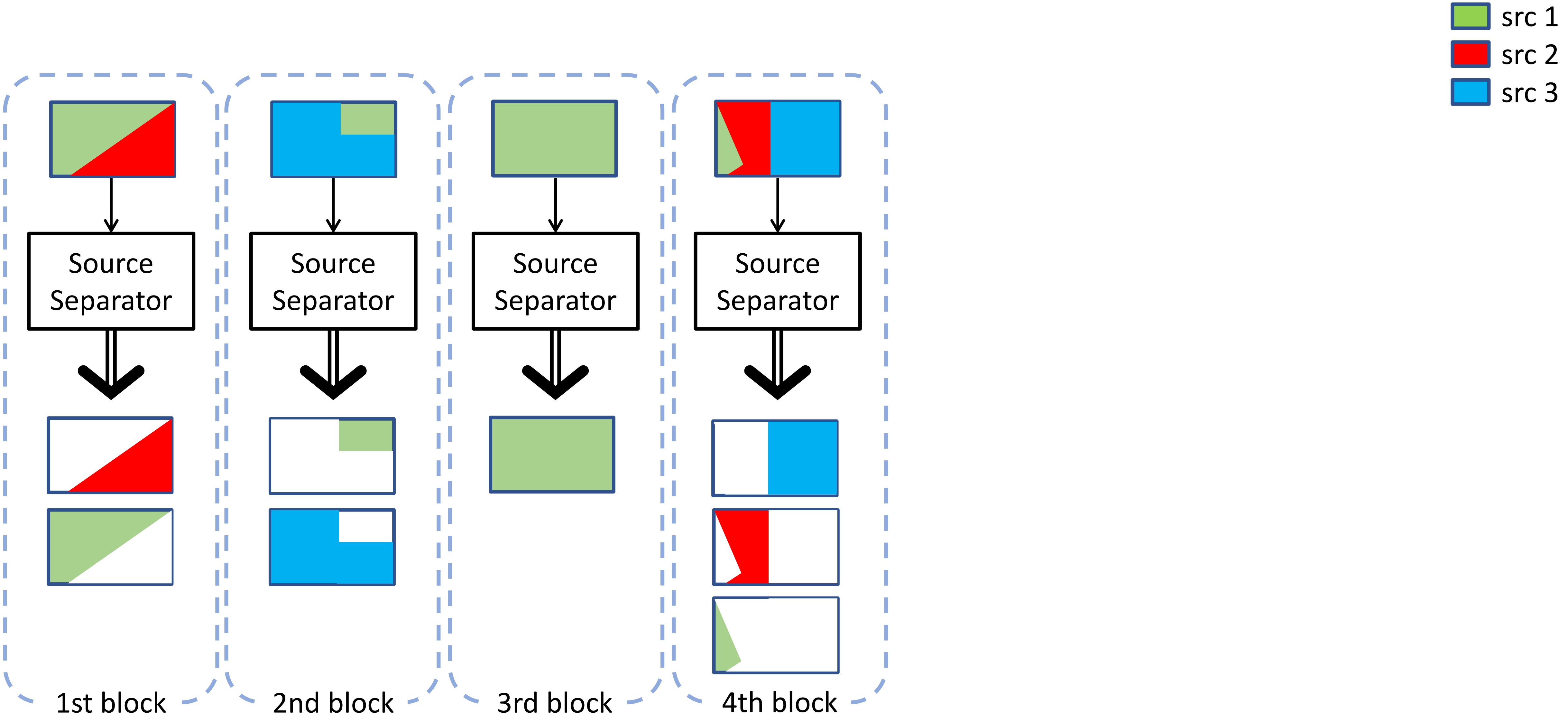
Example Train Mixture



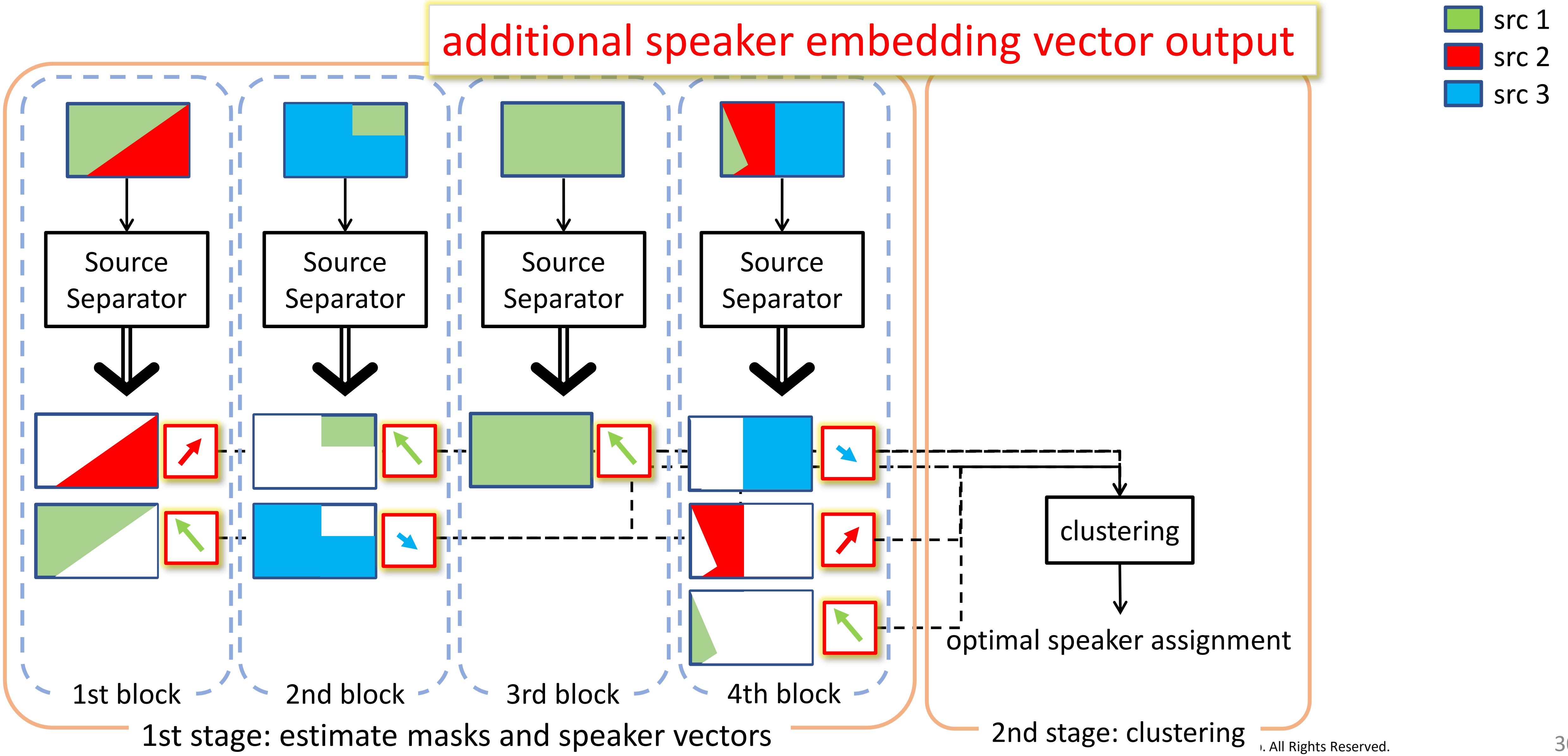
Example Test Mixture

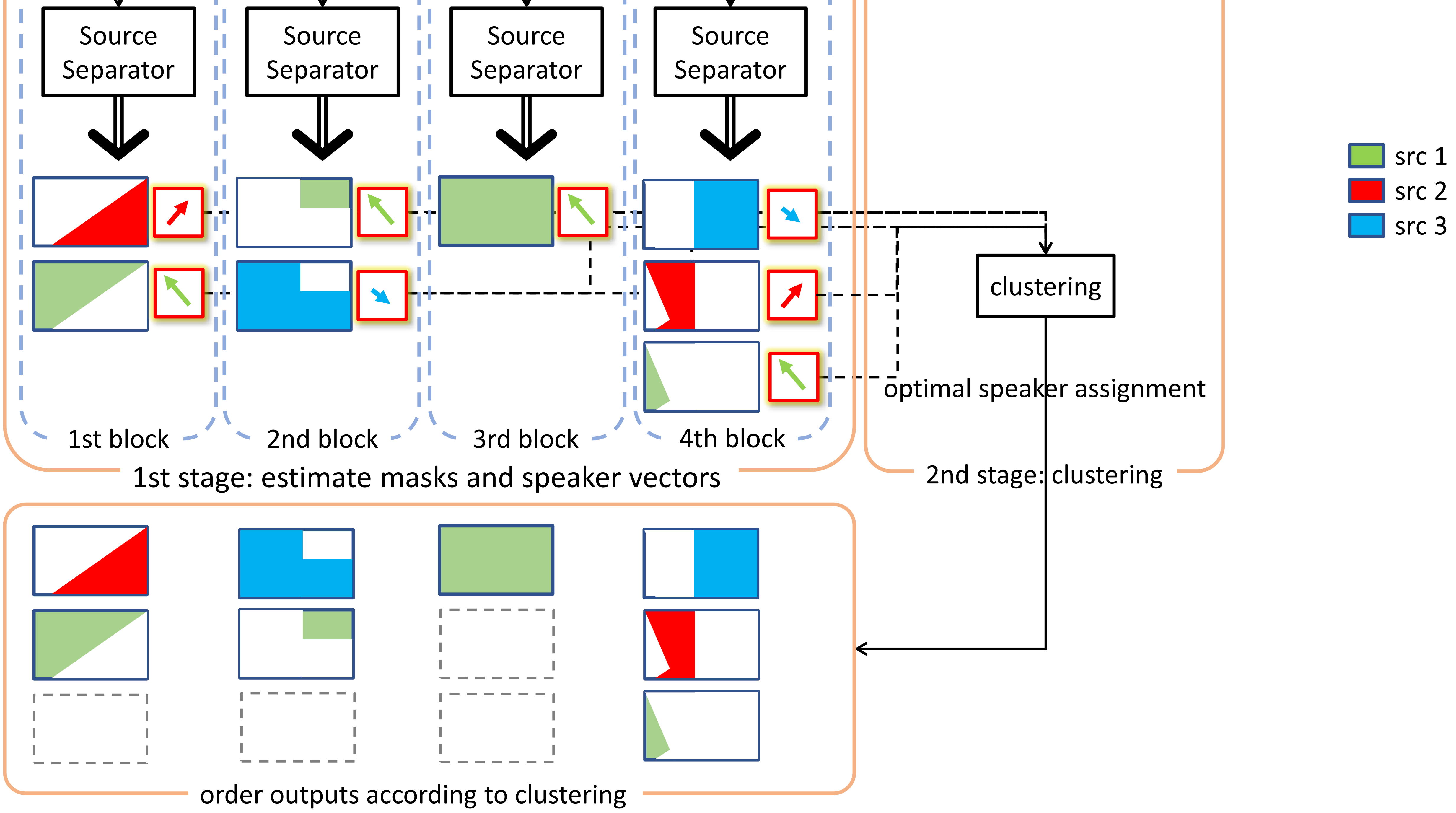


Conventional 2-stage Method: Clustering of Speaker Characteristics



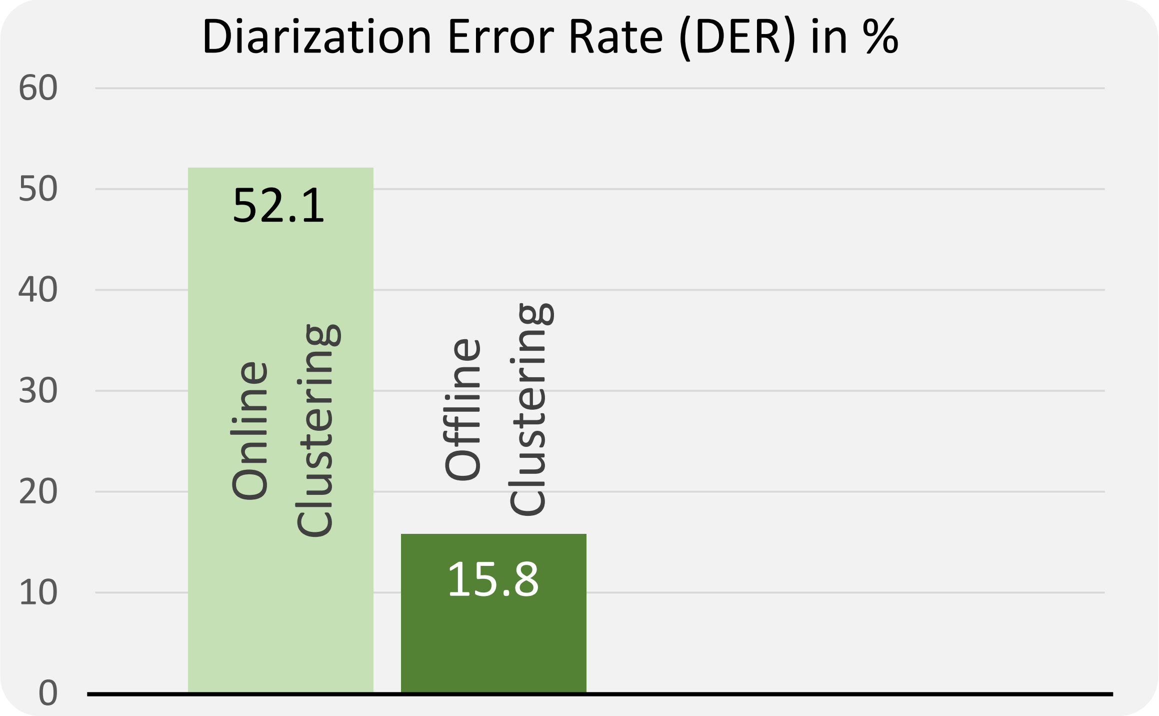
Conventional 2-stage Method: Clustering of Speaker Characteristics

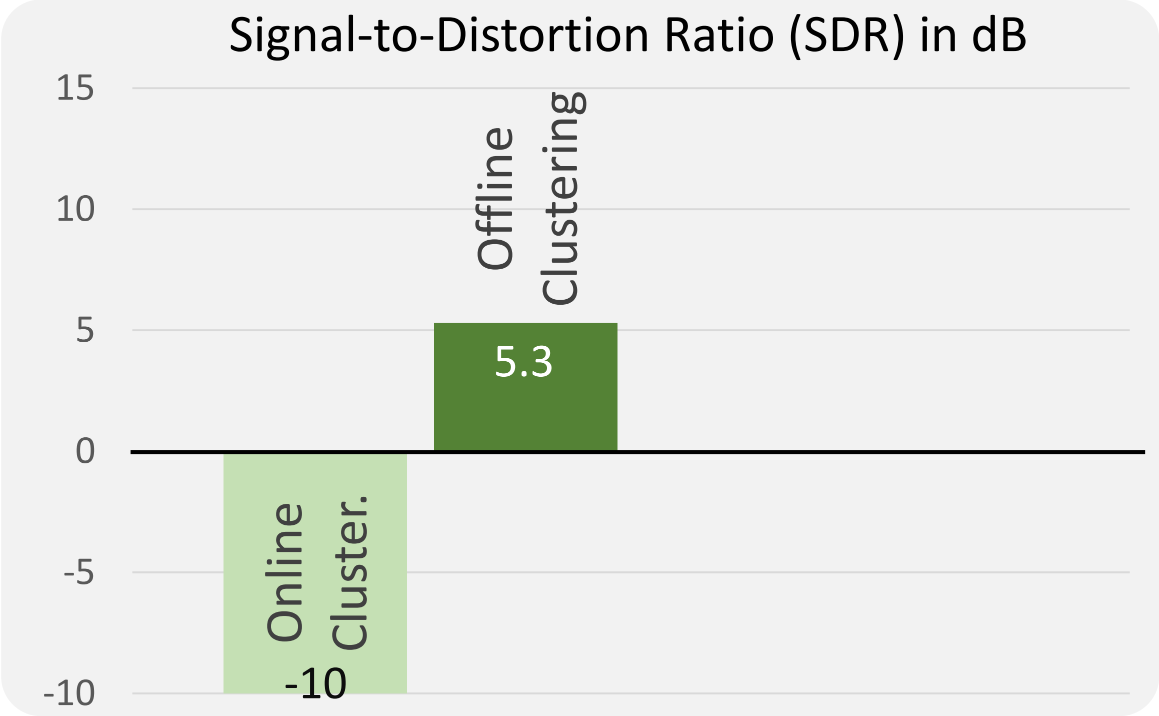




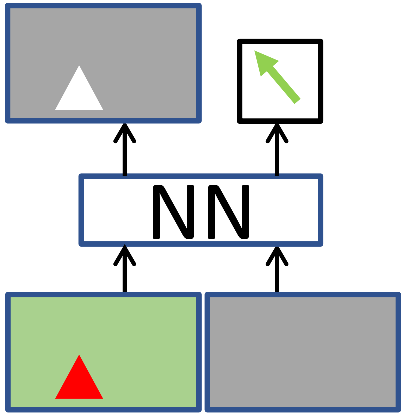
Source Separator
Source Separator
Source Separator
Source Separator
clustering
src 1
src 2
src 3
optimal speaker assignment
1st block
2nd block
3rd block
4th block
1st stage: estimate masks and speaker vectors
2nd stage: clustering
order outputs according to clustering

Evaluation on 12-block (30s) Mixtures





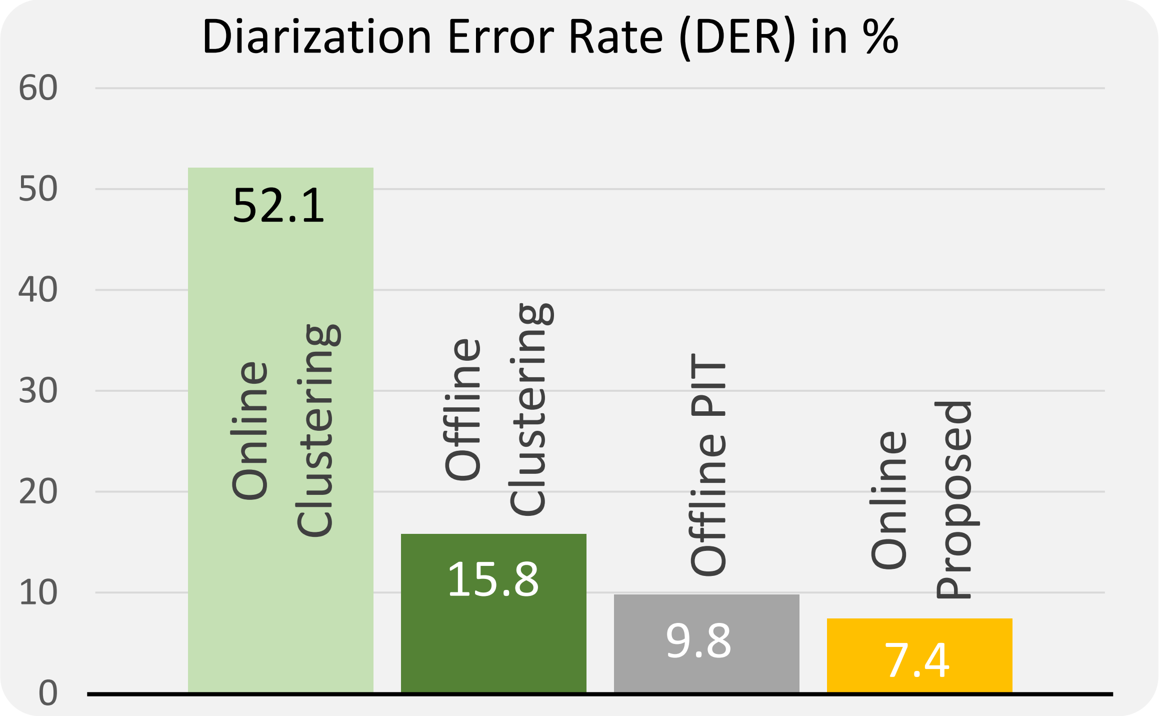
RSAN with **Speaker Embedding** + clustering

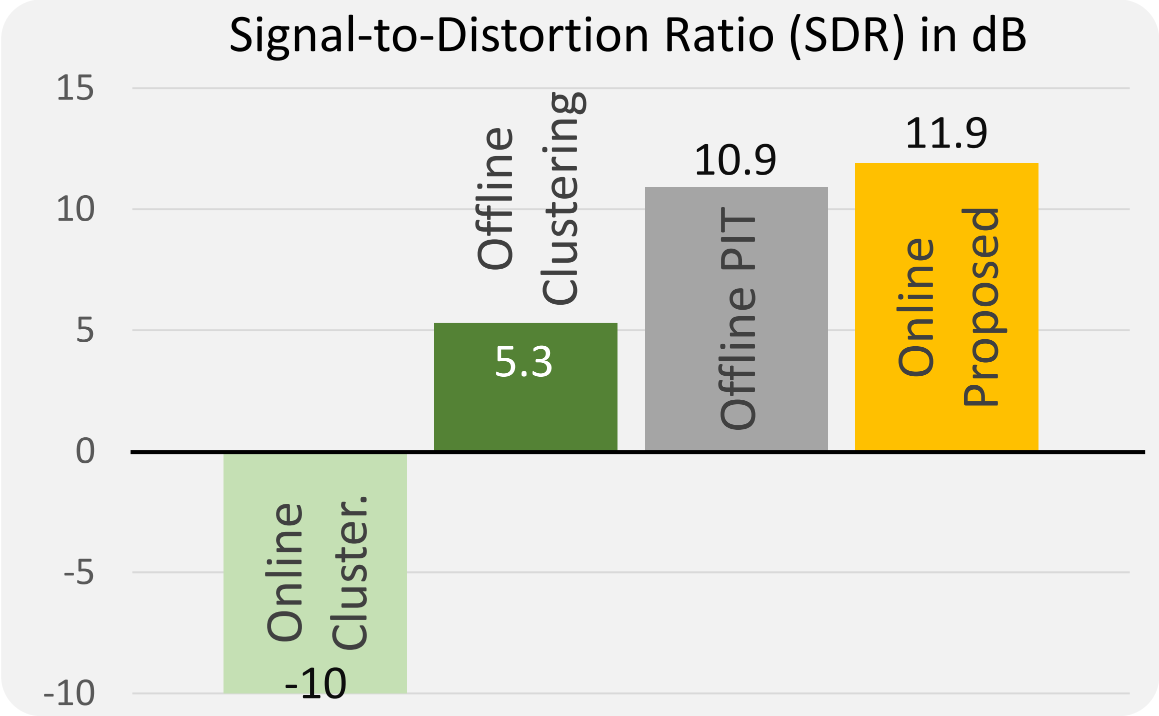


Online Clustering: leader-follower clustering

Offline Clustering: hierarchical clustering given the oracle #speakers

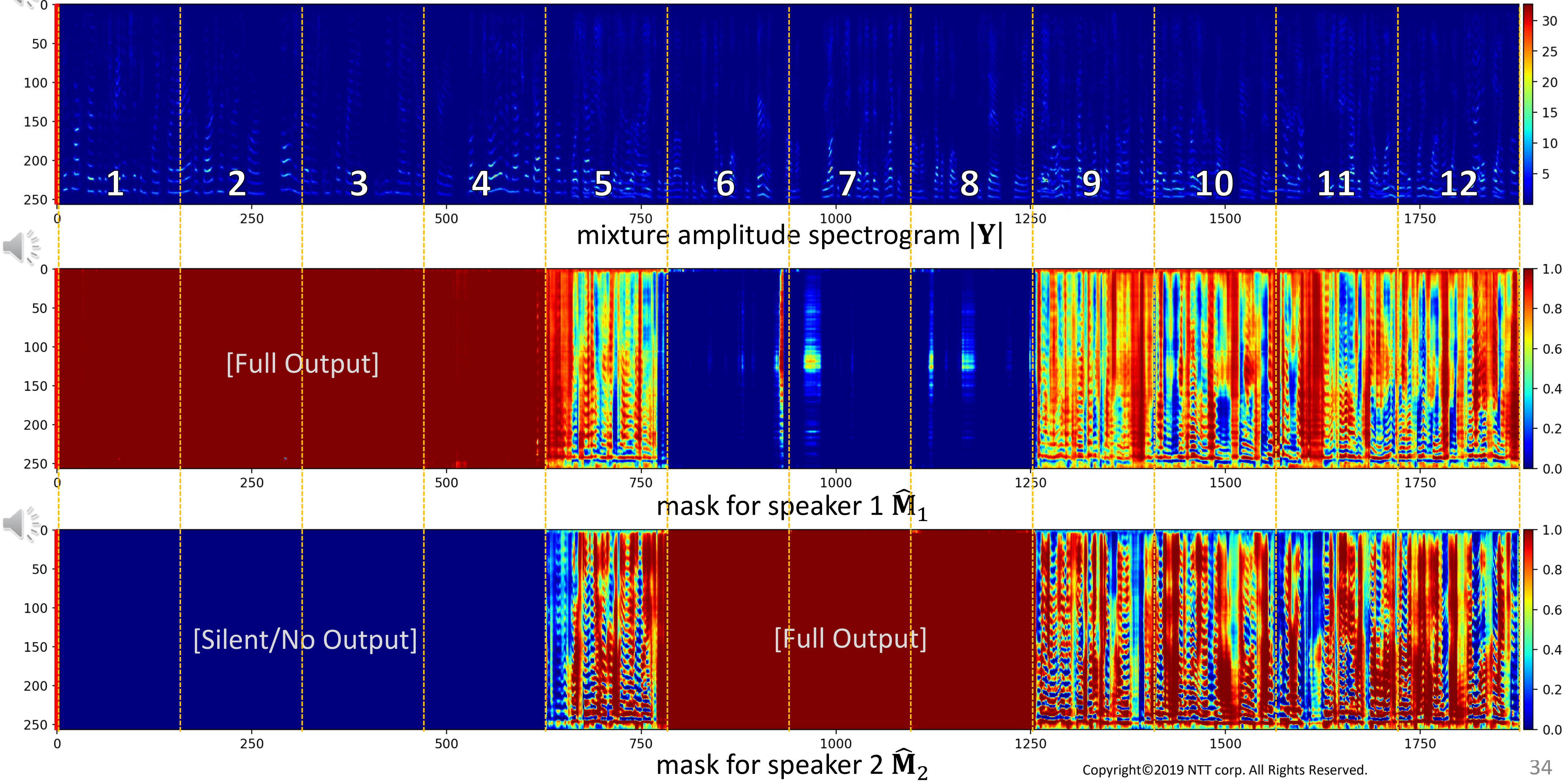
Evaluation on 12-block (30s) Mixtures





- **Proposed Method generalizes well to an unseen number of blocks**
  - Trained on 4 blocks, evaluated on 12 blocks
- **Proposed Method outperforms the other approaches**

Example: 12-block (30s) Mixture



Summary

Problems in Meeting Scenarios:

⇒ source separation

⇒ source count estimation

⇒ blockwise/online processing

Solved:

✓ RSAN

- ✓ iterative source extraction
- ✓ count of iterations

✓ Proposed **all-neural** block-online method based on RSAN

- ✓ generalizes well to unseen number of blocks

Thank you for your attention!