Source-Aware Context Network for Single-Channel Multi-speaker Speech Separation

Zeng-Xi Li[1], Yan Song[1], Li-Rong Dai[1], Ian McLoughlin[2]

[1]NELSLIP, University of Science and Technology of China, China [2]School of Computing, University of Kent, UK

Abstract

- Conventional deep learning based approaches may encounter difficulties in speaker-independent single-channel multi-speaker speech separation.
  - Partly due to the *label permutation problem*.
- We propose a novel source-aware context network:
  - Explicitly inputs speech sources as well as mixture signal.
  - The permutation order of outputs can be easily determined without any additional post-processing.
- A Multi-time-step Prediction Training (MPT) strategy is proposed to address the mismatch between training and inference stages.

Label Permutation Problem

- Conventional deep learning based methods commonly cast multi-speaker separation as a multi-class regression problem. In two-speaker situation:

$$\hat{\mathbf{x}}_{1,t}, \hat{\mathbf{x}}_{2,t} = H\left(\mathbf{y}_{t+F}, \cdots, \mathbf{y}_{t-P}\right)$$

- During training, the error between targets $[\mathbf{x}_{1,t}, \mathbf{x}_{2,t}]$ and outputs $[\hat{\mathbf{x}}_{1,t}, \hat{\mathbf{x}}_{2,t}]$ needs to be computed for backpropagation. However, it is unknown in advance whether the outputs order is $[\hat{\mathbf{x}}_{1,t}, \hat{\mathbf{x}}_{2,t}]$ or $[\hat{\mathbf{x}}_{2,t}, \hat{\mathbf{x}}_{1,t}]$, given only input $\mathbf{y}$

Source-Aware Context Network

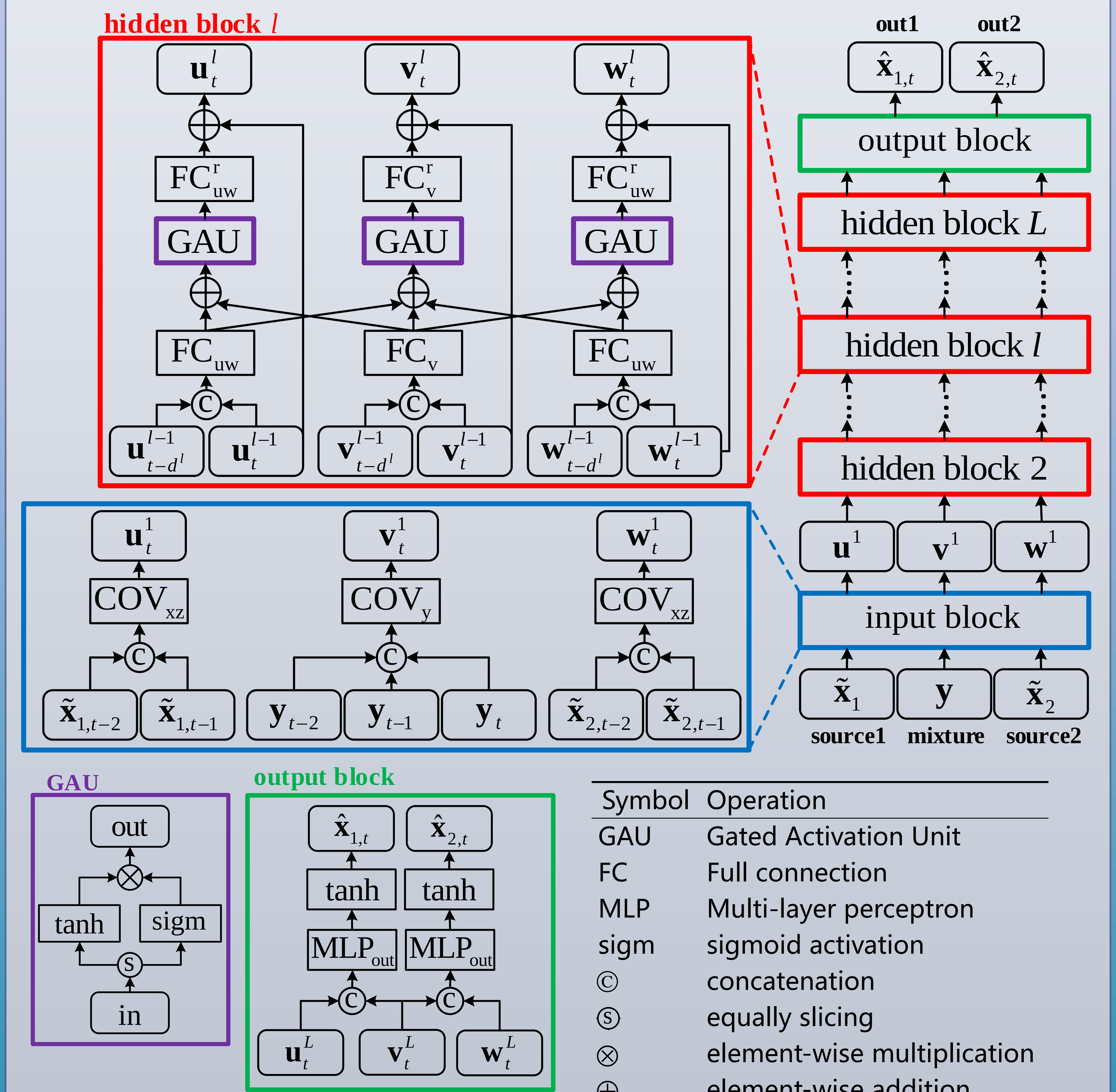
- Our proposed model simultaneously and recursively estimates two sources by modeling the conditional distribution of current sources' spectra, given past sources' spectra and mixture spectra:

$$\hat{\mathbf{x}}_{1,t} \sim p\left(\mathbf{x}_{1,t} \mid \mathbf{x}_{1,t-1}, \cdots, \mathbf{x}_{1,t-P}; \mathbf{x}_{2,t-1}, \cdots, \mathbf{x}_{2,t-P}; \mathbf{y}_t, \cdots, \mathbf{y}_{t-P}\right)$$

$$\hat{\mathbf{x}}_{2,t} \sim p\left(\mathbf{x}_{2,t} \mid \mathbf{x}_{2,t-1}, \cdots, \mathbf{x}_{2,t-P}; \mathbf{x}_{1,t-1}, \cdots, \mathbf{x}_{1,t-P}; \mathbf{y}_t, \cdots, \mathbf{y}_{t-P}\right)$$

- An overview of the proposed network $\mathbf{G}$

$$\hat{\mathbf{x}}_{1,t}, \hat{\mathbf{x}}_{2,t} = G\left(\tilde{\mathbf{x}}_{1,t-1}, \cdots, \tilde{\mathbf{x}}_{1,t-P}; \tilde{\mathbf{x}}_{2,t-1}, \cdots, \tilde{\mathbf{x}}_{2,t-P}; \mathbf{y}_t, \cdots, \mathbf{y}_{t-P}\right)$$



Symbol	Operation
GAU	Gated Activation Unit
FC	Full connection
MLP	Multi-layer perceptron
sigm	sigmoid activation
Ⓒ	concatenation
Ⓢ	equally slicing
⊗	element-wise multiplication
⊕	element-wise addition

- Without additional operations, the outputs order is determined in advance – just the same as input sources.
- $\mathbf{G}$ does not require future mixture spectra during inference

Multi-time-step Prediction Training

- To alleviate the mismatch between training and inference stages.
- At the first time step $t' = t$, source inputs are all clean spectra. Then at each time step t', outputs $\hat{\mathbf{x}}_{k,t'}$ are fed back as inputs $\tilde{\mathbf{x}}_{k,t'}$ to replace original clean spectra $\mathbf{x}_{k,t'}$ for the next time step $t'+1$
- This procedure repeats S times recursively, generating a sequence of estimated source spectra $\hat{\mathbf{x}}_{k,t'}$
- MSE across all time steps is used for training:

$$L = \frac{1}{FS}\sum_{s=0}^{S-1}\sum_{k=1}^{K}\left\|\mathbf{x}_{k,t+s} - \hat{\mathbf{x}}_{k,t+s}\right\|_2^2$$

Experiments

- Dataset: WSJ0-2mix
  - training set: 30 hours from WSJ0 si_tr_s.
  - validation set: 10 hours from WSJ0 si_tr_s, closed condition (CC).
  - test set: 5 hours from WSJ0 si_dt_05 and si_et_05, 16 unseen speakers, open condition (OC).
- Experimental Results
  - SDR improvements (dB) for different step numbers in MPT

Step number	SDR Imp. CC	SDR Imp. OC
1	-3.0	-2.4
5	6.7	6.9
10	7.1	7.4
30	8.8	9.0
60	**9.3**	**9.5**
90	9.2	9.2
120	9.0	9.0

  - SDR improvements (dB) and approximate model size comparisons of different methods

Method	Model Size (million)	SDR Imp. CC	SDR Imp. OC
Oracle NMF	-	5.1	-
CASA	-	2.9	3.1
DPCL	6.3	6.5	6.5
DPCL+	10.6	-	9.4
PIT-CNN-51\51	-	7.6	7.5
uPIT-BLSTM-AM	46.4	9.0	8.7
uPIT-BLSTM-PSM	46.4	9.4	9.4
DANet-6 anchor-LSTM	-	-	9.0
uPIT-LSTM-PSM	65.7	7.0	7.0
Source-aware context network	7.2	9.3	9.5