Deep Learning Based Phase Reconstruction for Speaker Separation: A Trigonometric Perspective

Zhong-Qiu Wang, **Ke Tan**, DeLiang Wang

Perception and Neurodynamics Lab

The Ohio State University

THE OHIO STATE UNIVERSITY

Outline

- Introduction
- Iterative Phase Reconstruction
- Group Delay Based Phase Reconstruction
- Sign Prediction Network
- Experiments
- Conclusions

Introduction

- Significant progress has been made on monaural speech enhancement and multi-talker speaker separation
  - Deep learning and T-F masking based speech enhancement
  - Deep clustering (DC), permutation invariant training (PIT)
- Typically estimating real-valued masks for separation
  - Using the mixture phase for re-synthesis
  - Magnitude estimation can be dramatically improved using deep learning
- This study investigates magnitude based methods for phase reconstruction

Motivation - I

- Given a C-source time-domain mixture

$$y = \sum_{c=1}^{C} s^{(c)}$$

- And its STFT representation

$$Y_{t,f} = \sum_{c=1}^{C} S_{t,f}^{(c)} = \sum_{c=1}^{C} A_{t,f}^{(c)} e^{j\theta_{t,f}^{(c)}}$$

Geometric Constraint

- Assuming $C = 2$
- Assuming $\hat{A}_{t,f}^{(c)} = A_{t,f}^{(c)}$

Is there any closed-form solution for phase estimation?

Motivation - II

- It is reasonable to say yes as there are two equations with two unknowns

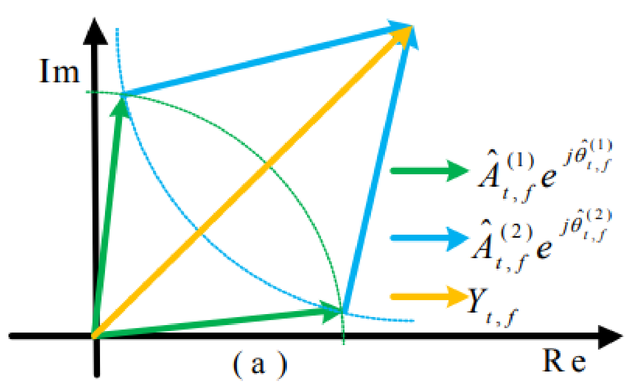
$$|Y_{t,f}|\cos(\angle Y_{t,f}) = \hat{A}_{t,f}^{(1)}\cos(\hat{\theta}_{t,f}^{(1)}) + \hat{A}_{t,f}^{(2)}\cos(\hat{\theta}_{t,f}^{(2)})$$
Real

$$|Y_{t,f}|\sin(\angle Y_{t,f}) = \hat{A}_{t,f}^{(1)}\sin(\hat{\theta}_{t,f}^{(1)}) + \hat{A}_{t,f}^{(2)}\sin(\hat{\theta}_{t,f}^{(2)})$$
Imaginary

- Phase-difference sign cannot be determined

$$\hat{\theta}_{t,f}^{(1)} = \angle Y_{t,f} \pm \arccos((|Y_{t,f}|^2 + \hat{A}_{t,f}^{(1)^2} - \hat{A}_{t,f}^{(2)^2})/(2|Y_{t,f}|\hat{A}_{t,f}^{(1)}))$$
$$\hat{\theta}_{t,f}^{(2)} = \angle Y_{t,f} \mp \arccos((|Y_{t,f}|^2 + \hat{A}_{t,f}^{(2)^2} - \hat{A}_{t,f}^{(1)^2})/(2|Y_{t,f}|\hat{A}_{t,f}^{(2)}))$$

- The absolute phase difference can be determined
- The potential phase solutions can be narrowed down to only two candidates !

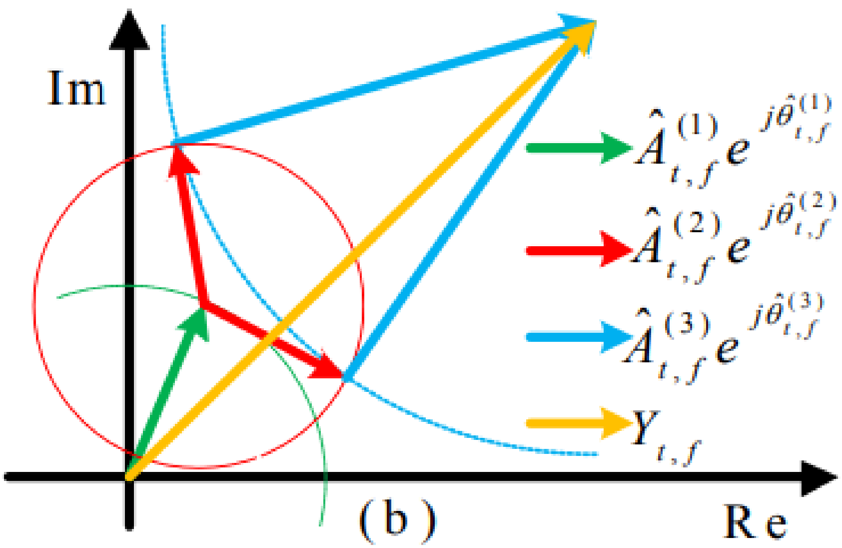


Motivation - III

- Solution: exploit inter T-F unit phase relations
  - Group delay
  - Instantaneous frequency
  - Phase consistency
- Propose three algorithms
  - Iterative phase reconstruction
  - Group delay based phase reconstruction
  - Sign prediction networks

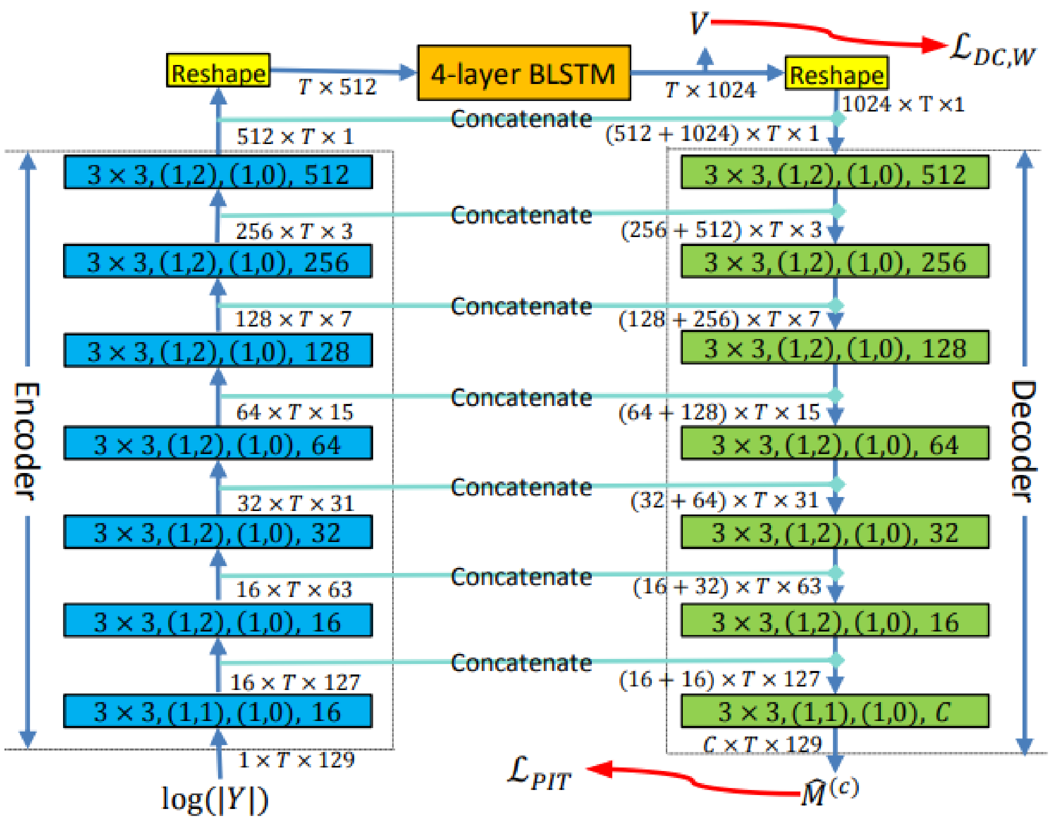
Motivation - IV

- What if $C > 2$?
  - Infinite number of phase solutions even if all the magnitudes are known
- Solution: *one-vs.-the-rest*
  - First use a chimera++ network to resolve the label permutation problem
  - Then train an enhancement network to further estimate the magnitudes of source c, and the remaining sources combined ($\neg c$) for phase reconstruction



Chimera++ Network

- DC loss: $\mathcal{L}_{DC,W} = ||V(V^TV)^{-1/2} - U(U^TU)^{-1}U^TV(V^TV)^{-1/2}||_F^2$
- PIT loss: $\mathcal{L}_{PIT} = \min_{\pi \epsilon \Psi} \sum_{c=1}^{C} \left\| \hat{M}^{\pi(c)} \otimes |Y| - T_0^{|Y|}\left(|S^{(c)}| \otimes \cos(\angle S^{(c)} - \angle Y)\right) \right\|_1$
- Chimera++: $\mathcal{L}_{chi++} = \lambda \mathcal{L}_{DC,W} + (1-\lambda)\mathcal{L}_{PIT}$
- 4-layer BLSTM with convolutional encoder-decoder structure



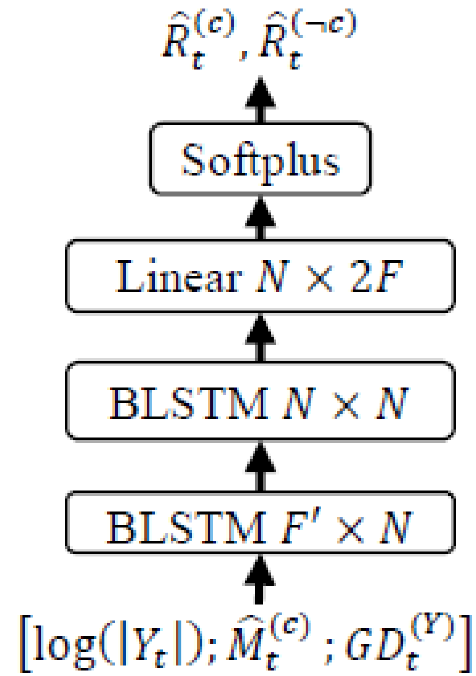
DNN Based Iterative Phase Reconstruction I

- Using estimated magnitudes and noisy phase to drive two-source multiple input spectrogram inverse (MISI)

For $k = 1{:}K$ **do**

- $\hat{s}^{(c')}(k) = \text{iSTFT}(\hat{A}^{(c')}, \hat{\vartheta}^{(c')}(k-1))$, for c' in $\{c, \neg c\}$;
- $\varepsilon(k) = y - \sum_{c' \in \{c, \neg c\}} \hat{s}^{(c')}(k)$;
- $\hat{\vartheta}^{(c')}(k) = \angle\text{STFT}(\hat{s}^{(c')}(k) + \varepsilon(k)/2)$, for c' in $\{c, \neg c\}$;

End

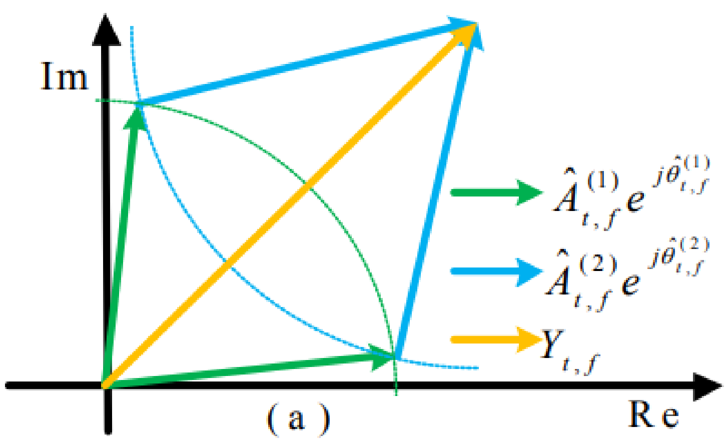


- Insight: the phase-difference signs could be resolved
  - The error distribution step can approximately satisfy the geometric constraint
  - Estimated magnitudes are sufficiently accurate
  - Only particular sign assignments lead to consistent phase structure

DNN Based Iterative Phase Reconstruction II

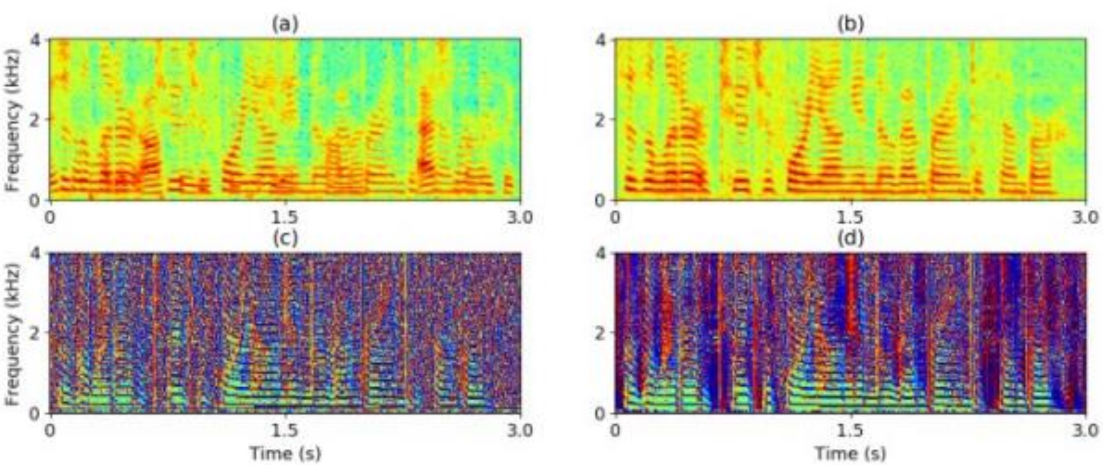
- Estimate the Spectral Magnitude Mask (SMM) !
  - $\mathcal{L}_{MSA(\alpha)}^{Enh1} = \mathcal{L}_{MSA(\alpha)} = \sum_{c' \in \{c, \neg c\}} \left\| |Y| \otimes T_0^{\alpha}(\hat{R}^{(c')}) - T_0^{\alpha|Y|}(|S^{(c')}|) \right\|_1$
  - Mask values need to be much larger than one
  - The two magnitudes can be long enough to support a valid triangle
  - Insight: magnitudes by estimated IRM, IBM and PSM cannot support a valid triangle as the masks sum up to one !
- Further train though MISI

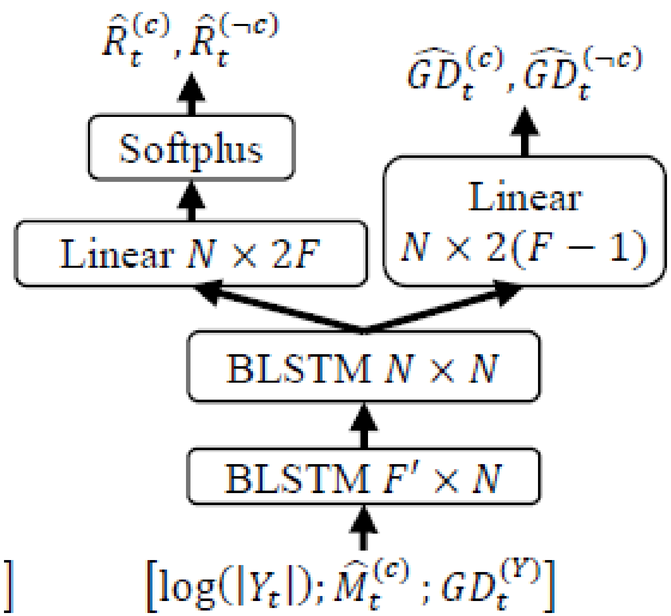
$$\mathcal{L}_{MISI-K}^{Enh1} = \sum_{c' \in \{c, \neg c\}} \left\| \mathrm{iSTFT}(\hat{A}^{(c')}, \hat{\vartheta}^{(c')}(K)) - s^{(c')} \right\|_1$$



Group Delay Based Phase Reconstruction I

- Group delay (GD) is predictable from magnitudes
  - $GD_{t,f}^{(c)} = \angle e^{j(\angle S_{t,f+1}^{(c)} - \angle S_{t,f}^{(c)})}$
  - $\mathcal{L}_{GD1} = \sum_{c' \in \{c, \neg c\}} \sum_t \sum_{f=1}^{F-1} |S_{t,f+1}^{(c')}| (1 - \cos(\widehat{GD}_{t,f}^{(c')} - GD_{t,f}^{(c')}))/2\,,$
  - $\mathcal{L}_{MSA(\alpha)+GD1}^{Enh2} = \mathcal{L}_{MSA(\alpha)} + \mathcal{L}_{GD1}$
- Key idea: find a sign assignment per T-F unit such that the resulting phase spectrums has GDs similar to the estimated GDs





]

Group Delay Based Phase Reconstruction II

- At run time, compute absolute phase difference based on the law of cosines assuming $\hat{A}^{(c)}$, $\hat{A}^{(\neg c)}$ and $|Y|$ form a triangle at each T-F unit

$$\hat{\delta}^{(c')} = |\angle e^{j(\hat{\theta}^{(c')} - \angle Y)}| = \arccos(\mathcal{T}(\frac{|Y|^2 + \hat{A}^{(c')2} - \hat{A}^{(\neg c')2}}{2|Y| \otimes |\hat{A}^{(c')}|})), \text{ for } c' \text{ in } \{c, \neg c\}$$

- Find a sign assignment per T-F unit, $\hat{g}_{t,f} \in \{-1,1\}$, that maximizes

$$\hat{g}_{t,1}, \dots, \hat{g}_{t,F} = \underset{g_{t,1},\dots,g_{t,F}}{\operatorname{argmax}} \sum_{f=1}^{F-1} \sum_{c' \in \{c, \neg c\}} \cos\left(\hat{\theta}_{t,f+1}^{(c')}(g_{t,f+1}) - \hat{\theta}_{t,f}^{(c')}(g_{t,f}) - \widehat{GD}_{t,f}^{(c')}\right)$$

where $\hat{\theta}_{t,f}^{(c)}(g_{t,f}) = \angle Y_{t,f} + g_{t,f}\hat{\delta}_{t,f}^{(c)}$ and $\hat{\theta}_{t,f}^{(\neg c)}(g_{t,f}) = \angle Y_{t,f} - g_{t,f}\hat{\delta}_{t,f}^{(\neg c)}$

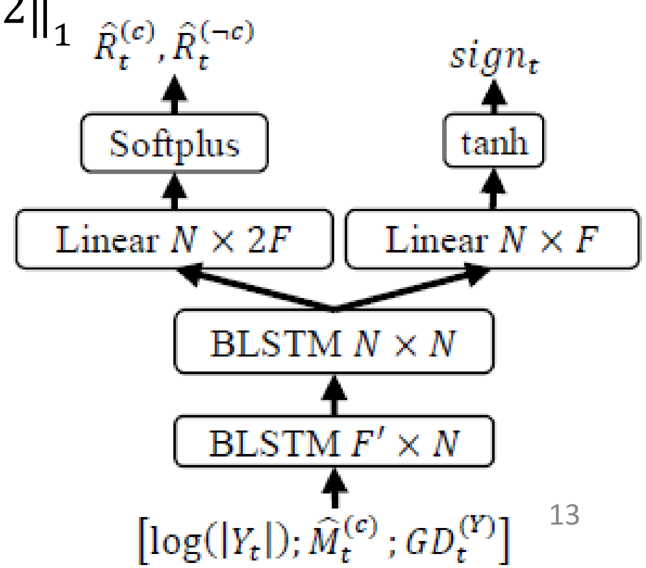
- Can be efficiently solved using dynamic programming per frame with time complexity $O(2^2 F)$
- Estimated phases are $\angle Y + \hat{g} \otimes \hat{\delta}^{(c)}$ and $\angle Y - \hat{g} \otimes \hat{\delta}^{(\neg c)}$

Sign Prediction Network I

- The GD based method is hard to be trained end-to-end
- Predict the sign using DNN
  - $\hat{\theta}^{(c)} = \angle Y + sign \otimes \hat{\delta}^{(c)}$
  - $\hat{\theta}^{(\neg c)} = \angle Y - sign \otimes \hat{\delta}^{(\neg c)}$

Two phases are on different sides of mixture phase

- Loss computed on the resulting GD
  - $\mathcal{L}_{GD2} = \sum_{c' \in \{c, \neg c\}} \sum_t \sum_{f=1}^{F-1} |S_{t,f+1}^{(c')}| (1 - \cos(\hat{\theta}_{t,f+1}^{(c')} - \hat{\theta}_{t,f}^{(c')} - GD_{t,f}^{(c')}))/2$
- Loss computed directly on the phase
  - $\mathcal{L}_{phase} = \sum_{c' \in \{c, \neg c\}} \left\| |S^{(c')}| \otimes (1 - \cos(\hat{\theta}^{(c')} - \theta^{(c')}))/2 \right\|_1$
- Overall loss function
  - $\mathcal{L}_{MSA(\alpha)+GD2}^{Enh3} = \mathcal{L}_{MSA(\alpha)} + \mathcal{L}_{GD2}$
  - $\mathcal{L}_{MSA(\alpha)+phase}^{Enh3} = \mathcal{L}_{MSA(\alpha)} + \mathcal{L}_{phase}$



Sign Prediction Network II

- Train through 0 or K iterations of MISI
  - Starting from estimated magnitude $\hat{A}^{(c)}$ and estimated phase $\hat{\theta}^{(c)}$, following Le Roux *et al*., 2019.
  - Time-domain loss

$$\mathcal{L}_{MISI-K}^{Enh3} = \sum_{c' \in \{c, \neg c\}} \left\| \mathrm{iSTFT}(\hat{A}^{(c')}, \hat{\vartheta}^{(c')}(K)) - s^{(c')} \right\|_1$$

  - Frequency-domain loss, following Wang *et al*., 2018

$$\mathcal{L}_{MISI-K-MSA}^{Enh3} = \sum_{c' \in \{c, \neg c\}} \left\| \left| \mathrm{STFT}\left(\mathrm{iSTFT}\left(\hat{A}^{(c')}, \hat{\vartheta}^{(c')}(K)\right)\right) \right| - \left| S^{(c')} \right| \right\|_1$$

Experimental Setup

- Open wsj0-2mix and wsj0-3mix
  - Speaker-independent
  - 30 h training, 10 h validation, 5 h testing
- Evaluation metrics
  - SDRi (dB)
  - SI-SDRi (dB)
  - PESQ

Experimental Results I

- Estimating SMM is more suitable than estimating PSM for MISI
- Training through MISI brings slight improvement on SI-SDRi, but not on PESQ
  - Likely because $\mathcal{L}_{MISI-5}^{Enh1}$ uses time-domain loss

SI-SDRi and PESQ on wsj0-2mix

Approaches	Models	Enhanced Phase?	SI-SDRi	PESQ
Unprocessed	-	No	0.0	2.01
Chimera++	$\mathcal{L}_{chi++}$	No	11.9	3.12
Deep learning based iterative phase reconstruction	$\mathcal{L}_{PSA(0,1)}^{Enh1}$	No	12.1	3.15
	+MISI-5	Yes	12.5	3.17
	$\mathcal{L}_{PSA(0,5)}^{Enh1}$	No	12.4	3.17
	+MISI-5	Yes	12.9	3.19
	$\mathcal{L}_{PSA(-1,1)}^{Enh1}$	No	12.4	3.21
	+MISI-5	Yes	12.9	3.24
	$\mathcal{L}_{PSA(-5,5)}^{Enh1}$	No	12.7	3.21
	+MISI-5	Yes	13.3	3.24
	$\mathcal{L}_{MSA(5)}^{Enh1}$	No	11.1	3.27
	+MISI-5	Yes	14.4	3.43
	$+\mathcal{L}_{MISI-5}^{Enh1}$	Yes	15.0	3.38

Experimental Results II

- Group delay based method is not as good as MISI
  - But gets clear improvement over $\mathcal{L}^{Enh1}_{MSA(5)}$
  - Phase consistency might be more important for monaural phase estimation
- Sign prediction net obtains SI-SDRi similar to MISI
  - Avoids STFT/iSTFT iterations
  - $\mathcal{L}^{Enh3}_{MSA(5)+phase}$ slightly better than $\mathcal{L}^{Enh3}_{MSA(5)+GD2}$
- $\mathcal{L}^{Enh3}_{MISI-5-MSA}$ better than $\mathcal{L}^{Enh3}_{MISI-5}$ on PESQ, but slightly worse on SI-SDRi
  - PESQ is largely computed based on magnitude

SI-SDRi and PESQ on wsj0-2mix

Approaches	Models	Enhanced Phase?	SI-SDRi	PESQ
Unprocessed	-	No	0.0	2.01
Chimera++	$\mathcal{L}_{chi++}$	No	11.9	3.12
Deep learning based iterative phase reconstruction	$\mathcal{L}^{Enh1}_{MSA(5)}$	No	11.1	3.27
	+MISI-5	Yes	14.4	3.43
	$+\mathcal{L}^{Enh1}_{MISI-5}$	Yes	15.0	3.38
Group delay based phase reconstruction	$\mathcal{L}^{Enh2}_{MSA(5)+GD1}$	Yes	13.6	3.39
Sign prediction network	$\mathcal{L}^{Enh3}_{MSA(5)+GD2}$	Yes	14.2	3.39
	$\mathcal{L}^{Enh3}_{MSA(5)+phase}$	Yes	14.4	3.38
	+MISI-5	Yes	15.0	3.44
	$+\mathcal{L}^{Enh3}_{WA}$	Yes	14.6	3.36
	$+\mathcal{L}^{Enh3}_{MISI-5}$	Yes	15.3	3.36
	$+\mathcal{L}^{Enh3}_{MISI-5-MSA}$	Yes	15.2	3.45

Comparison with other studies

- State-of-the-art results were obtained on wsj0-2mix and 3mix at the time of submission, especially on PESQ

Approaches	wsj0-2mix			wsj0-3mix		
	SI-SDRi	SDRi	PESQ	SI-SDRi	SDRi	PESQ
Unprocessed	0.0	0.0	2.01	0.0	0.0	1.66
DC++	10.8	-	-	7.1	-	-
ADANet	10.4	10.8	2.82	9.1	9.4	2.16
uPIT-ST	-	10.0	-	-	7.7	-
Chimera++ (BLSTM)	11.2	11.5	-	-	-	-
+MISI-5	11.5	11.8	-	-	-	-
+WA-MISI-5	12.6	12.9	-	-	-	-
+PhaseBook	12.8	-	-	-	-	-
conv-TasNet	14.6	15.0	3.25	11.6	12.0	2.50
Proposed (Sign prediction net, $\mathcal{L}_{MISI-5}^{Enh3}$)	**15.3**	**15.6**	3.36	**12.1**	**12.5**	2.64
Proposed (sign prediction net, $\mathcal{L}_{MISI-5-MSA}^{Enh3}$)	15.2	15.4	**3.45**	12.0	12.3	**2.77**

Concluding Remarks

- We have proposed three algorithms to resolve the sign ambiguity in phase estimation
- Deep learning based magnitude estimation can clearly help phase estimation
- The geometric constraint affords a mechanism to narrow down the potential solutions of phase, and could play a fundamental role in future research on phase estimation

Thanks