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

- 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

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

• 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**: \( \mathcal{L}_{PIT} = \min_{\pi \in \Psi} \sum_{c=1}^C \|\hat{M}^{\pi(c)} \otimes |Y| - T_{0}^{[Y]} (|S^{(c)}| \otimes \cos(\angle S^{(c)} - \angle Y))\|_1 \)
- **Chimera++**: \( \mathcal{L}_{\text{chi++}} = \lambda \mathcal{L}_{DC,W} + (1 - \lambda) \mathcal{L}_{PIT} \)
- **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 $k = 1: K$

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

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)!

\[ \mathcal{L}_{\text{MSA}}^{\text{Enh1}}(\alpha) = \mathcal{L}_{\text{MSA}}(\alpha) = \sum_{c' \in \{c, \neg c\}} \| |Y| \otimes T_0^\alpha(\hat{R}(c')) - T_0^\alpha|Y|(|S(c')|) \|_1 \]

– 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 through MISI

\[ \mathcal{L}_{\text{MISI}}^{\text{Enh1}} = \sum_{c' \in \{c, \neg c\}} \| \text{iSTFT} (\hat{A}(c'), \hat{\theta}(c')(K)) - s(c') \|_1 \]
Group Delay Based Phase Reconstruction I

- Group delay (GD) is predictable from magnitudes
  
  \[ GD_{t,f}^{(c)} = \angle e^{j(\angle S_{t,f+1}^{(c)} - \angle S_{t,f}^{(c)})} \]

\[ L_{GD1} = \sum_{c' \in \{c, -c\}} \sum_{f=1}^{F-1} |S_{t,f+1}^{(c')}| (1 - \cos(\widehat{GD}_{t,f}^{(c')} - GD_{t,f}^{(c')})) / 2, \]

\[ L_{MSA(\alpha) + GD1} = L_{MSA(\alpha)} + L_{GD1} \]

- 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 $\hat{A}(c)$, $\hat{A}(-c)$ and $|Y|$ form a triangle at each T-F unit

$$\hat{\delta}(c') = |\arg e^{i(\hat{\theta}(c') - \angle Y)}| = \arccos(T\left(\frac{|Y|^2 + \hat{A}(c')^2 - \hat{A}(-c')^2}{2|Y|\otimes|\hat{A}(c')|}\right)),$$
  for $c'$ in $\{c, -c\}$

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

$$\hat{g}_{t,1}, ..., \hat{g}_{t,F} = \arg\max_{g_{t,1}, ..., g_{t,F}} \sum_{f=1}^{F-1} \sum_{c' \in \{c, -c\}} \cos \left( \hat{\theta}_{t,f+1}(g_{t,f+1}) - \hat{\theta}_{t,f}(g_{t,f}) - GD_{t,f}(c') \right)$$

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

- 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 $\angle Y - \hat{g} \otimes \hat{\delta}(-c)$
Sign Prediction Network I

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

- Loss computed on the resulting GD
  - $L_{GD2} = \sum_{c' \in \{c,-c\}} \sum_{t} \sum_{f=1}^{F-1} |S_{t,f+1}^{(c')}|(1 - \cos(\hat{\theta}_{t,f+1}^{(c')} - \hat{\theta}_{t,f}^{(c')} - GD_{t,f}^{(c')}))/2$

- Loss computed directly on the phase
  - $L_{phase} = \sum_{c' \in \{c,-c\}} \| \sum_{t} \sum_{f=1}^{F-1} |S_{t,f+1}^{(c')}|(1 - \cos(\hat{\theta}_{t,f+1}^{(c')} - \theta_{t,f}^{(c')}))/2 \|_1$

- Overall loss function
  - $L_{MSA(\alpha) + GD2} = L_{MSA(\alpha)} + L_{GD2}$
  - $L_{MSA(\alpha) + phase} = L_{MSA(\alpha)} + L_{phase}$
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
    $$L_{MISI-K}^{Enh3} = \sum_{c' \in \{c, -c\}} \| iSTFT(\hat{A}^{(c')}, \hat{\theta}^{(c')}(K)) - s^{(c')} \|_1$$
  - Frequency-domain loss, following Wang et al., 2018
    $$L_{MISI-K-MSA}^{Enh3} = \sum_{c' \in \{c, -c\}} \| \text{STFT} \left( iSTFT \left( \hat{A}^{(c')}, \hat{\theta}^{(c')}(K) \right) \right) - |S^{(c')}| \|_1$$
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 MISK.
- Training through MISK brings slight improvement on SI-SDRI, but not on PESQ.
  - Likely because $\mathcal{L}_{MISK-5}^{Enh}$ uses time-domain loss.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Models</th>
<th>Enhanced Phase?</th>
<th>SI-SDRI</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td>-</td>
<td>No</td>
<td>0.0</td>
<td>2.01</td>
</tr>
<tr>
<td>Chimera++</td>
<td>$\mathcal{L}_{chirp}$</td>
<td>No</td>
<td>11.9</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{L}_{PSA(0,1)}^{Enh}$</td>
<td>No</td>
<td>12.1</td>
<td>3.15</td>
</tr>
<tr>
<td></td>
<td>+MISK-5</td>
<td>Yes</td>
<td>12.5</td>
<td>3.17</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{L}_{PSA(0,5)}^{Enh}$</td>
<td>No</td>
<td>12.4</td>
<td>3.17</td>
</tr>
<tr>
<td></td>
<td>+MISK-5</td>
<td>Yes</td>
<td>12.9</td>
<td>3.19</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{L}_{PSA(-1,1)}^{Enh}$</td>
<td>No</td>
<td>12.4</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td>+MISK-5</td>
<td>Yes</td>
<td>12.9</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{L}_{PSA(-5,5)}^{Enh}$</td>
<td>No</td>
<td>12.7</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td>+MISK-5</td>
<td>Yes</td>
<td>13.3</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td>$\mathcal{L}_{MISK-5}^{Enh}$</td>
<td>No</td>
<td>11.1</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td>+MISK-5</td>
<td>Yes</td>
<td>14.4</td>
<td>3.43</td>
</tr>
<tr>
<td></td>
<td>+$\mathcal{L}_{MISK-5}^{Enh}$</td>
<td>Yes</td>
<td>15.0</td>
<td>3.38</td>
</tr>
</tbody>
</table>

SI-SDRI and PESQ on wsj0-2mix
## Experimental Results II

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

### SI-SDRi and PESQ on wsj0-2mix

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Models</th>
<th>Enhanced Phase?</th>
<th>SI-SDRi</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td>-</td>
<td>No</td>
<td>0.0</td>
<td>2.01</td>
</tr>
<tr>
<td>Chimera++</td>
<td>$\mathcal{L}_{\text{chi++}}$</td>
<td>No</td>
<td>11.9</td>
<td>3.12</td>
</tr>
<tr>
<td>Deep learning</td>
<td>$\mathcal{L}^{\text{Enh1}}_{\text{MSA}(5)}$</td>
<td>No</td>
<td>11.1</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td>+MISI-5</td>
<td>Yes</td>
<td>14.4</td>
<td>3.43</td>
</tr>
<tr>
<td></td>
<td>+$\mathcal{L}^{\text{Enh1}}_{\text{MISI−5}}$</td>
<td>Yes</td>
<td>15.0</td>
<td>3.38</td>
</tr>
<tr>
<td>Group delay</td>
<td>$\mathcal{L}^{\text{Enh2}}_{\text{MSA}(5)+\text{GD1}}$</td>
<td>Yes</td>
<td>13.6</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td>+$\mathcal{L}^{\text{Enh3}}_{\text{MSA}(5)+\text{GD2}}$</td>
<td>Yes</td>
<td>14.2</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td>+$\mathcal{L}^{\text{Enh3}}_{\text{MSA}(5)+\text{phase}}$</td>
<td>Yes</td>
<td>14.4</td>
<td>3.38</td>
</tr>
<tr>
<td></td>
<td>+MISI-5</td>
<td>Yes</td>
<td>15.0</td>
<td>3.44</td>
</tr>
<tr>
<td></td>
<td>+$\mathcal{L}^{\text{Enh3}}_{\text{MISI−5}}$</td>
<td>Yes</td>
<td>15.3</td>
<td>3.36</td>
</tr>
<tr>
<td></td>
<td>+$\mathcal{L}^{\text{Enh3}}_{\text{MISI−5−MSA}}$</td>
<td>Yes</td>
<td>15.2</td>
<td>3.45</td>
</tr>
</tbody>
</table>
Comparison with other studies

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

<table>
<thead>
<tr>
<th>Approaches</th>
<th>wsj0-2mix</th>
<th></th>
<th>wsj0-3mix</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SI-SDRi</td>
<td>SDRi</td>
<td>PESQ</td>
<td>SI-SDRi</td>
</tr>
<tr>
<td>Unprocessed</td>
<td>0.0</td>
<td>0.0</td>
<td>2.01</td>
<td>0.0</td>
</tr>
<tr>
<td>DC++</td>
<td>10.8</td>
<td>-</td>
<td>-</td>
<td>7.1</td>
</tr>
<tr>
<td>ADANet</td>
<td>10.4</td>
<td>10.8</td>
<td>2.82</td>
<td>9.1</td>
</tr>
<tr>
<td>uPIT-ST</td>
<td>-</td>
<td>10.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chimera++ (BLSTM)</td>
<td>11.2</td>
<td>11.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+MISI-5</td>
<td>11.5</td>
<td>11.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+WA-MISI-5</td>
<td>12.6</td>
<td>12.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+PhaseBook</td>
<td>12.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>conv-TasNet</td>
<td>14.6</td>
<td>15.0</td>
<td>3.25</td>
<td>11.6</td>
</tr>
<tr>
<td>Proposed (Sign prediction net, $\mathcal{L}_{MISI-5}^{Enh^3}$)</td>
<td>15.3</td>
<td>15.6</td>
<td>3.36</td>
<td>12.1</td>
</tr>
<tr>
<td>Proposed (sign prediction net, $\mathcal{L}_{MISI-5-MSA}^{Enh^3}$)</td>
<td>15.2</td>
<td>15.4</td>
<td>3.45</td>
<td>12.0</td>
</tr>
</tbody>
</table>
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