ICASSP2020 Barcelona

Mask-dependent Phase Estimation for Monaural Speaker Separation

Zhaoheng Ni, Michael I Mandel

City University of New York

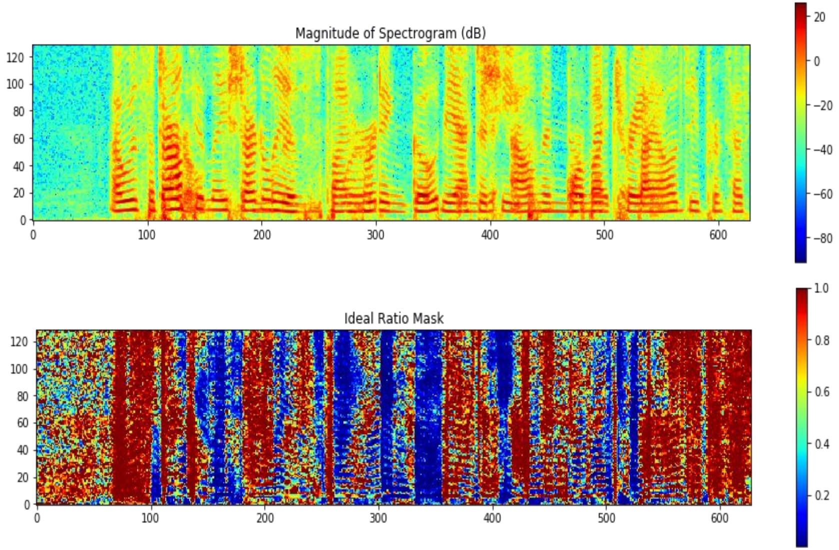
zni@gradcenter.cuny.edu mim@sci.brooklyn.cuny.edu

Agenda

- Introduction
- Model Architecture
- Mask-dependent PIT Criterion
- Weighted Phase Losses
- Experiments
- Results

Introduction

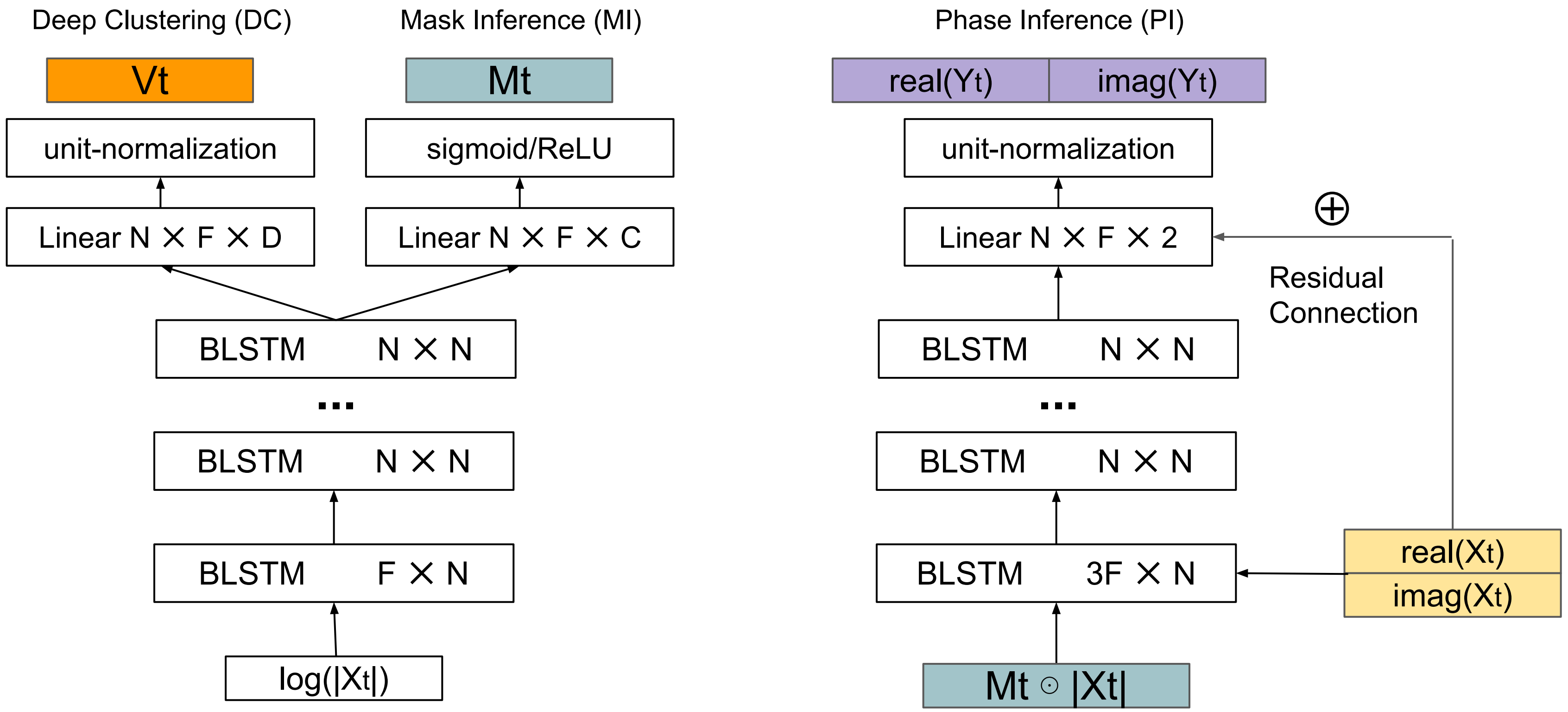
- Speaker Separation refers to the task of isolating speech in a multi-talker environment.
- Estimating the real-valued Time-Frequency Mask is a successful way to separate the speech from the mixture.



Why Estimating Phase?

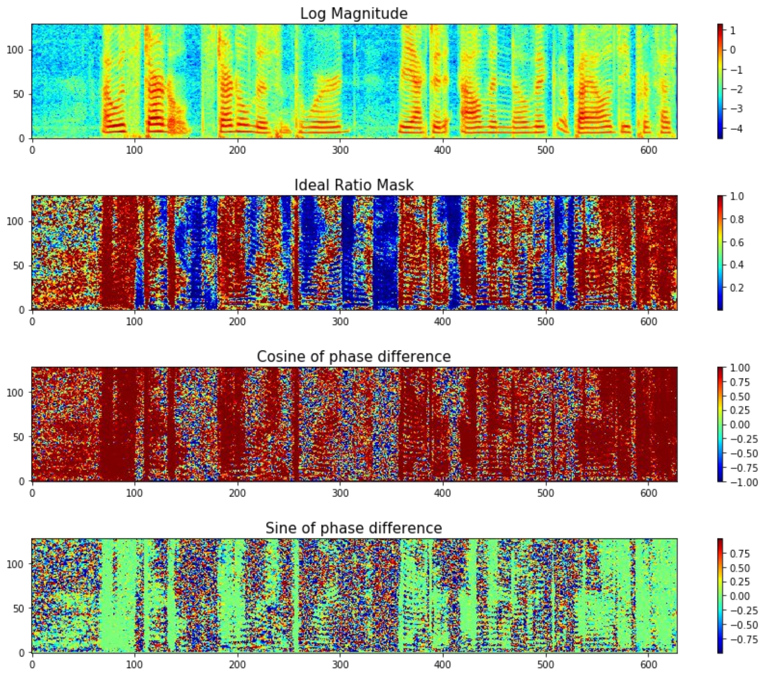
- Mask-based methods use the phase of the mixture speech fo iSTFT
  - $y = \mathrm{iSTFT}(X \odot \hat{M})$
- The phase error hence is unavoidable

Model Architecture



Wang, Zhong-Qiu, Jonathan Le Roux, and John R. Hershey. "Alternative objective functions for deep clustering." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

Phase is Difficult to Estimate



The clean spectrogram, the IRM, cosine, and sine of phase difference between clean and noisy STFT (sample utterance: cv/s2(mix)/011a010g_0.16366_40pc0204_-0.16366.wav)

Mask-Phase PIT Criterion

$$L_{\text{PIT, MP}} = \min_{\pi \in P} \sum_{c} \left(\left\| \hat{M}_c \odot |X| - |S_{\pi(c)}| \right\|_F^1 - \sum_{t,f} \langle \hat{p}_{t,f}^{c}, p_{t,f}^{\pi(c)} \rangle \right)$$

Mc: magnitude mask for source c

|X|: speech mixture magnitude

|S|: clean speech magnitude

p: phase of the spectrogram

π: the permutation that has minimum **(magnitude + phase)** loss

Mask-dependent PIT Criterion

$$L_{\text{PIT, MD}} = \sum_c \left\| \hat{M}_c \odot |X| - |S_{\pi(c)}| \right\|_F^1 - \sum_{c,t,f} \langle \hat{p}^c_{t,f}, p^{\pi(c)}_{t,f} \rangle,$$

$$\pi = \arg\min_\pi \sum_c \left\| \hat{M}_c \odot |X| - |S_{\pi(c)}| \right\|_F^1$$

Mc: magnitude mask for source c

|X|: speech mixture magnitude

|S|: clean speech magnitude

p: phase of the spectrogram

π: the permutation that has minimum **magnitude** loss

Different weights to Phase Loss

- Phase of mixture is more similar to the more dominating clean speech (information from the mixture)
- The pattern of the phase is more random if the phases of two clean speech are in opposite directions. (phase cancellation)
- The difficulties of phase estimation for different T-F regions are thus different

Different weights to Phase Loss

- Magnitude Weighted Loss Function (MWL)
    - $$L_{\text{PI, MWL}} = -\sum_{c,t,f} (\gamma + M_{c,t,f}) \langle \hat{p}^{c}_{t,f}, p^{\pi(c)}_{t,f} \rangle$$
- Inverse Magnitude Weighted Loss Function (IMWL)
    - $$L_{\text{PI, I-MWL}} = -\sum_{c,t,f} (\gamma + M_{\neg c,t,f}) \langle \hat{p}^{c}_{t,f}, p^{\pi(c)}_{t,f} \rangle$$
- Joint Weighted Loss Function (Joint)
    - $$L_{\text{PI, Joint}} = -\sum_{c,t,f} \langle \hat{p}^{c}_{t,f}, p^{\pi(c)}_{t,f} \rangle \sum_{i} M_{i,t,f}$$

Experiments

- Dataset: WSJ0-2mix
    - 20,000 mixtures for training
    - 5,000 mixtures for validation (closed speaker condition)
    - 5,000 mixtures for testing (open speaker condition)
- 4 BLSTM layers, 600-dimension for each direction, 0.3 dropout rate
- Batch size: 16. Each sample has 400 consecutive frames.
- Feature: log STFT, 256 window size, 64 hop size,
- Adam optimizer with 1e-3 learning rate
- 100 epochs, early stops if validation loss stops improving for 10 epochs
- Evaluation metric: Signal-to-Distortion Ratio (SDR)

Onssen Library

We put our implementations along with other published methods to onssen library.

https://github.com/speechLabBcCuny/onssen

More models and reproduced scores have been or will be added soon:

- DPCL
- Chimera (++)
- MISI, end2end MISI
- TasNet, Conv-TasNet, DPRNN
- FurcaNet, FurcaNeXt

Results

Method	SDR	SI-SDR
Chimera++, MSA	10.5	-
+ tPSA [5]	11.5	11.2
+ MISI-5 [5]	11.8	11.5
+ WA-MISI-5 [9]	12.9	12.6
Phasebook, MISI-0 [16]	-	12.6
+ MISI-5 [16]	-	12.8
Chimera++(Encoder-BLSTM-Decoder) [17]	-	11.9
Sign prediction network [17]	15.6	15.3

Published SDR/SI-SDR improvements of different phase estimation methods on the open speaker condition (OSC) of the WSJ0-2mix dataset.

PIT Criterion	Mask Activation	Phase Loss	SDR
MP	ReLU	MWL	11.0
MP	Sigmoid	MWL	11.5
MD	Sigmoid	I-MWL	12.0
MD	Sigmoid	MWL	12.6
MD	Sigmoid	Joint	13.0
MD	Sigmoid	Joint,$\alpha = 0.5$	13.6

SDR improvements of the proposed method with different settings on the OSC of the WSJ0-2mix dataset. PIT criteria Mask+Phase (MP) and Mask-dependent (MD). Phase losses magnitude-weighted loss (MWL), inverse magnitude-weighted loss (I-MWL), and joint weighted loss (Joint).

References

- John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 31–35.
- Hakan Erdogan, John R Hershey, ShinjiWatanabe, and Jonathan Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 708–712.
- Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, “Permutation invariant training of deep models for speakerindependent multi-talker speech separation,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 241–245.
- Yi Luo, Zhuo Chen, John R Hershey, Jonathan Le Roux, and Nima Mesgarani, “Deep clustering and conventional networks for music separation: Stronger together,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 61–65.
- Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, “Alternative objective functions for deep clustering,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 686–690.
- Donald S Williamson, Yuxuan Wang, and DeLiang Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 24, no. 3, pp. 483–492, 2016.
- Donald S Williamson and DeLiang Wang, “Speech dereverberation and denoising using complex ratio masks,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5590–5594.
- David Gunawan and Deep Sen, “Iterative phase estimation for the synthesis of separated sources from single-channel mixtures,” IEEE Signal Processing Letters, vol. 17, no. 5, pp. 421–424, 2010.
- Zhong-Qiu Wang, Jonathan Le Roux, DeLiang Wang, and John R Hershey, “End-to-end speech separation with unfolded iterative phase reconstruction,” arXiv preprint arXiv:1804.10204, 2018.
- Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, “Looking to listen at the cocktail party: A speakerindependent audio-visual model for speech separation,” arXiv preprint arXiv:1804.03619, 2018.
- Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, “The conversation: Deep audio-visual speech enhancement,” arXiv preprint arXiv:1804.04121, 2018.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, “Automatic differentiation in pytorch,” in NIPS-W, 2017

References

- Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” in Proceedings of the 14th python in science conference, 2015, pp. 18–25.
- Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel, “mir eval: A transparent implementation of common mir metrics,” in In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR. Citeseer, 2014.
- Jonathan Le Roux, Gordon Wichern, Shinji Watanabe, Andy Sarroff, and John R Hershey, “Phasebook and friends: Leveraging discrete representations for source separation,” IEEE Journal of Selected Topics in Signal Processing, 2019.
- Zhong-Qiu Wang, Ke Tan, and DeLiang Wang, “Deep learning based phase reconstruction for speaker separation: A trigonometric perspective,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 71–75.
- Jonathan Le Roux, JR Hershey, A Liutkus, F St¨oter, STWisdom, and H Erdogan, “Sdr–half-baked or well done?,” Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA, Tech. Rep, 2018.
- Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey, “Single-channel multi-speaker separation using deep clustering,” arXiv preprint arXiv:1607.02173, 2016.
- Yoshua Bengio, J´erˆome Louradour, Ronan Collobert, and Jason Weston, “Curriculum learning,” in Proceedings of the 26th annual international conference on machine learning. ACM, 2009, pp. 41–48.
- Yuzhou Liu and DeLiang Wang, “Divide and conquer: A deep casa approach to talker-independent monaural speaker separation,” arXiv preprint arXiv:1904.11148, 2019.

Thank you very much!