EFFICIENT INTEGRATION OF FIXED BEAMFORMERS AND SPEECH SEPARATION NETWORKS FOR MULTI-CHANNEL FAR-FIELD SPEECH SEPARATION

Zhuo Chen, Takuya Yoshioka, Xiong Xiao, Jinyu Li, Michael L. Seltzer, Yifan Gong
Microsoft AI and Research, USA

1. Introduction

- The neural network based speech separation largely boosts the state-of-the-art in recent years. However it still suffers from different limitations:
  1. The single channel separation system has performance limitation and is vulnerable to reverberation
  2. The traditional beamforming entirely relies on the spatial information for separation, with limited performance achieved
- To solve the multi-speaker separation & recognition in a more general environment, we proposed a multi-channel based system, where the overlapped speech is processed in three steps:
  1. Pre-process with a fixed set of beamformer
  2. Predict the best beam for each mixing speaker
  3. Generate the final separation with single channel separation net for each selected beam

2. System architecture

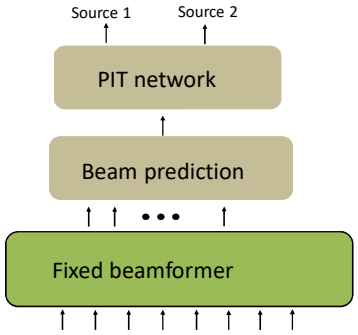


Figure 1. *System Overview*

Beamformer

- Pre-enhance the mixture from spatial clues
- Fixed beamformer with hand designed beam pattern, uniformly sample the space

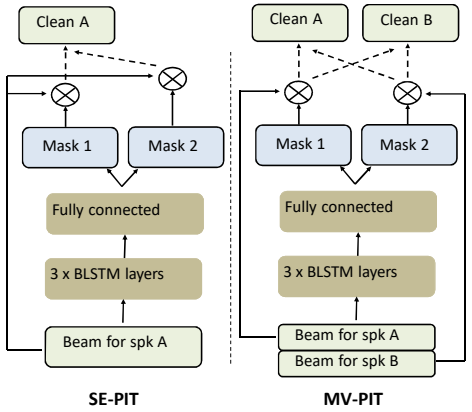
Beam prediction network

- Predict the best beam for each speaker
- Best beam refers to the beam with the best signal to distortion ratio for each speaker, which is not necessarily the closet beams to speakers
- The signals from all beams and inter-microphone phase difference are used as input
- A N-hot vector is estimated per frame, indicting the best beam for each speaker

Separation network

- Further separate the beamformed signal to acquire final result
- When the speakers are close, the beamforming is not sufficient for separation, the permutation invariant training network (PIT) is applied to avoid the permutation confusion
- For better speech recognition performance, the mask-based beamforming could be further applied with the estimated mask per speaker

3. Two variations on PIT network



Speech enhancement PIT

- Only reconstruct the target speaker from each beam
- Similar to the traditional speech enhancement network
- Permutation invariant training is still applied to overcome the permutation ambiguity when beamforming is not effective

Multi-view PIT

- Use all selected beams together, recovering all speaker simultaneously
- The different beams provides complementary information
- Assuming there are at most 2 speakers, can easily generalize to more speakers

Evaluation

Data

Training

- 3 mixing conditions: FO, PO, SD
- 40 hours artificial mixed data, from 103 speakers, at 16kHz sampling rate
- Image method for RIR generation

Testing

- 1 hour of each mixing condition from 44 novel speakers

	FO	PO	SD
PIT	3.82	2.83	2.34
IRM	6.66	6.97	6.79
IRM-OB	10.53	10.87	10.78
OMVDR	9.86	9.66	9.46
MV-PIT-OB	8.00	9.27	8.5
SE-PIT-OB	6.78	7.96	7.31
MV-PIT-PB	6.19	7.33	6.5
SE-PIT-PB	5.99	7.28	6.38
OB	4.2	4.01	4.00
PB	2.72	2.57	2.39
ORI	-1.56	-1.45	0.04

Table 2. *Signal to Distortion ratio result*

Network

Beam Prediction

- 3 x 300 Bi-directional LSTM network
- Input feature: Concatenation between linear spectrogram of 12 beams and inter-microphone phase difference

Separation

- 3 x 1024 Bi-directional LSTM network followed by two fully connect layer with sigmoid activation.
- Input feature: linear spectrogram with utterance normalization

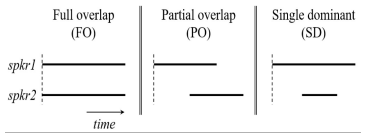


Figure 2. *3 mixing types*

Baseline

- **PIT**: Single channel PIT network
- **IRM**: Single channel ideal ratio mask
- **IRM-OB**: Ideal ratio mask with oracle beam selection
- **OMVDR**: Minimum variance distortion-less response beamformer with covariance from ground truth(clean sources)
- **MV-PIT-OB**: MV-PIT with oracle beam selection
- **MV-PIT-PB**: MV-PIT with beam prediction
- **SE-PIT-OB**: SE-PIT with oracle beam selection
- **SE-PIT-PB**: SE-PIT with beam prediction
- **PB**: Predicted beam from
- **OB**: Ideal ratio mask on channel 0
- **ORI**: single channel attractor network

Conclusion

- Both proposed systems outperform the single channel system and single channel ideal ratio mask system.
- The MV-PIT consistently outperforms the SE-PIT, showing beams provide complementary information.
- The beam prediction network leads around 2dB degradation than the oracle beam selection.
- Comparison with oracle MVDR and IRM-OB indicates further improvement room, e.g. joint training, mask based MVDR etc.