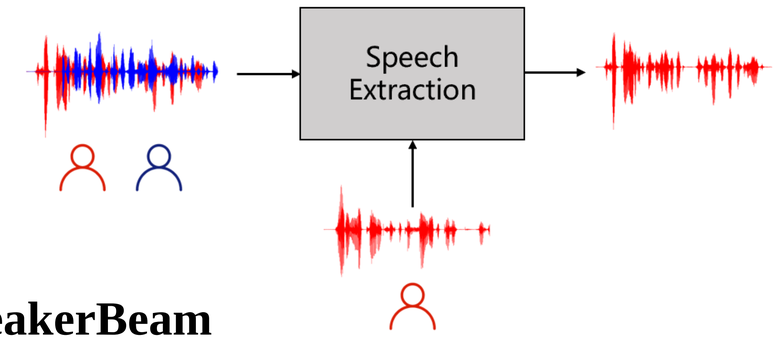
Multi-Channel Target Speech Extraction with Channel Decorrelation and Target Speaker Adaptation

Jiangyu Han, Xinyuan Zhou, Yanhua Long, Yijie Li

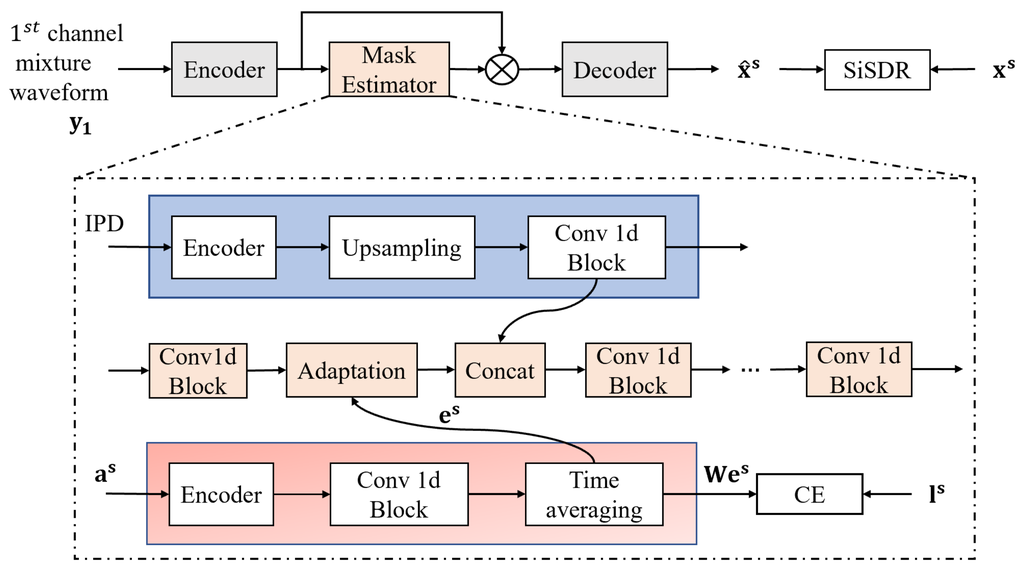
Shanghai Normal University
Unisound AI Technology Co., Ltd.

Introduction

Target Speech Extraction



TD-SpeakerBeam



Contribution:

Exploit multi-channel spatial information for TSE

- A parallel encoder structure
- A special target speaker adaptation
- A channel decorrelation mechanism

Experiments

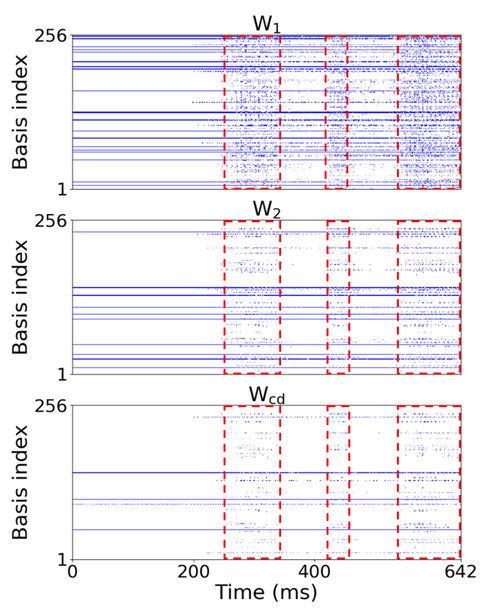
Configuration

- Dataset : spatialized reverberant WSJ0 2-mix
- The same network hyper-parameters as TD-SpeakerBeam
- IPD features: STFT, 32 msec window, 16 msec frame-shift
- Set $\alpha = 0.5$ to balance the loss tradeoff
- All experiments are about reverb against reverb

Results

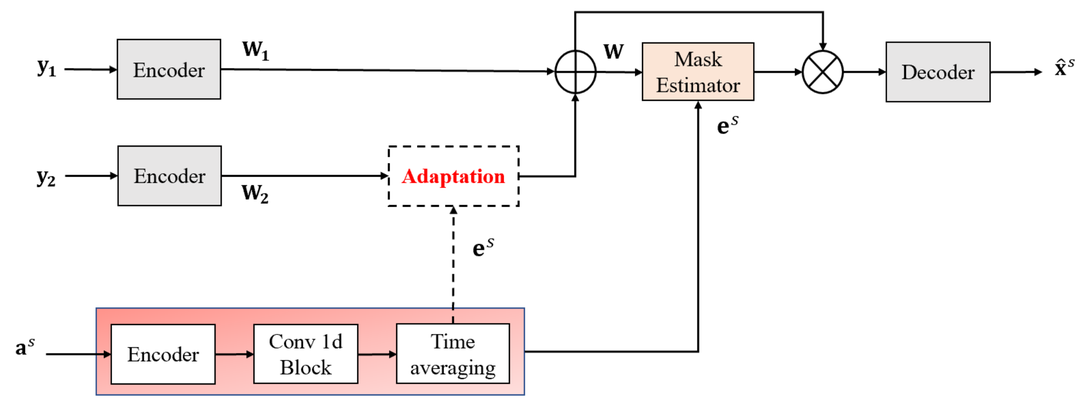
System	IPD	Adapt	SDR	SiSDR
(1) TD-SpkBeam (our)	✓	-	11.57	11.07
(2) Parallel (our)	-	-	12.43	11.91
(3)	-	✓	12.73	12.20
(4) CD	-	-	**12.87**	12.34
(5)	-	✓	**12.87**	**12.35**
(6)	✓	✓	12.55	12.01
(7) CC	-	✓	12.66	12.13

- Adaptation on the parallel encoded mixture is effective
- CD results 0.44/0.43 dB gains over the simple parallel encoder
- Adaptation on CD output **does not** bring any performance gains
- IPD + CD **degrade** the performance
- CD is better than CC

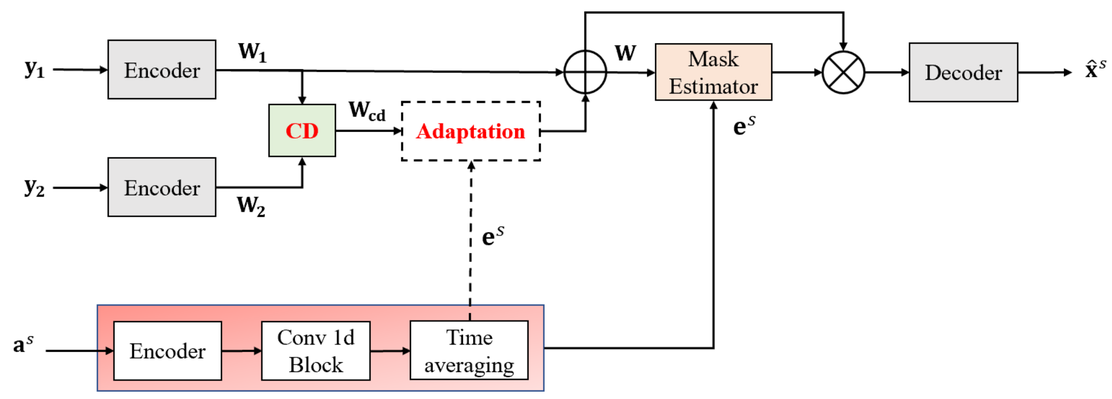


Proposed

Parallel Encoder



Channel Decorrelation



Correlation

$$\mathbf{W_i} = [\mathbf{w_i}^1, \mathbf{w_i}^2, ..., \mathbf{w_i}^N]^\mathbf{T}, i = 1, 2 \quad (1)$$

$$\phi_{1,2}^j = \frac{\langle \mathbf{w_1}^j, \mathbf{w_2}^j \rangle}{\|\mathbf{w_1}^j\|_2 \|\mathbf{w_2}^j\|_2}, j = 1, 2, ..., N \quad (2)$$

$$\boldsymbol{\phi}_{\mathbf{1,2}} = [\phi_{1,2}^1, \phi_{1,2}^2, ..., \phi_{1,2}^N]^\mathbf{T} \quad (3)$$

Decorrelation

$$\mathbf{a} = [1, 1, ..., 1]^\mathbf{T} \quad (4)$$

$$p_{1,2}^j = \frac{e^{\phi_{1,2}^j}}{e + e^{\phi_{1,2}^j}}, j = 1, 2, ..., N \quad (5)$$

$$\mathbf{p_{1,2}} = [p_{1,2}^1, p_{1,2}^2, ..., p_{1,2}^N]^\mathbf{T}$$

$$\mathbf{s_{1,2}} = \mathbf{a} - \mathbf{p_{1,2}}$$
$$\mathbf{S_{1,2}} = [\mathbf{s_{1,2}}, \mathbf{s_{1,2}}, ..., \mathbf{s_{1,2}}] \quad (6)$$

$$\mathbf{W_{cd}} = \mathbf{W_2} \odot \mathbf{S_{1,2}} \quad (7)$$

Conclusions

- The adaptation of the parallel encoded representation is very effective
- Channel decorrelation gives 0.44/0.43 improvements over parallel encoder
  - inter-channel differential spatial information is effectively exploited
  - speaker adaptation on CD output does not bring any performance gains
- Parallel encoder with CD significantly improved the TD-SpeakerBeam
  - 11.57/11.07 -> 12.87/12.35

References

[1] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, “Improving speaker discrimination of target speech extraction with time-domain speakerbeam,” in Proc. ICASSP. IEEE, 2020, pp. 691–695.

[2] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation,” in Proc. ICASSP. IEEE, 2018, pp. 1–5.

Code: https://github.com/jyhan03/channel-decorrelation