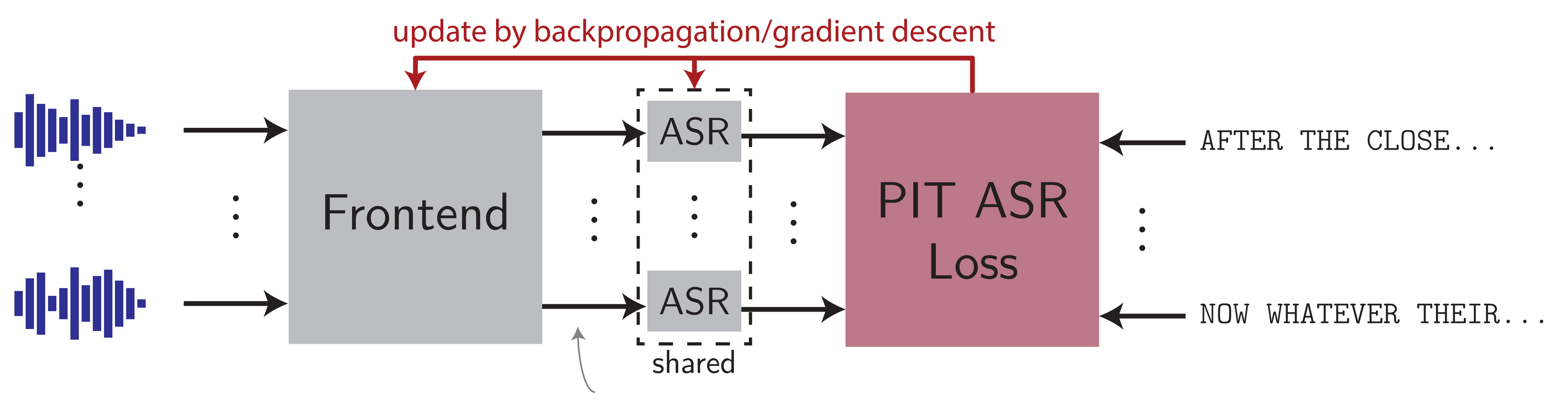
End-to-end Multi-speaker ASR with Independent Vector Analysis

Robin Scheibler[1], Wangyou Zhang[2], Xuankai Chang[3], Shinji Watanabe[3], Yanmin Qian[2] ([1]LINE, [2]SJTU, [3]CMU)

LINE

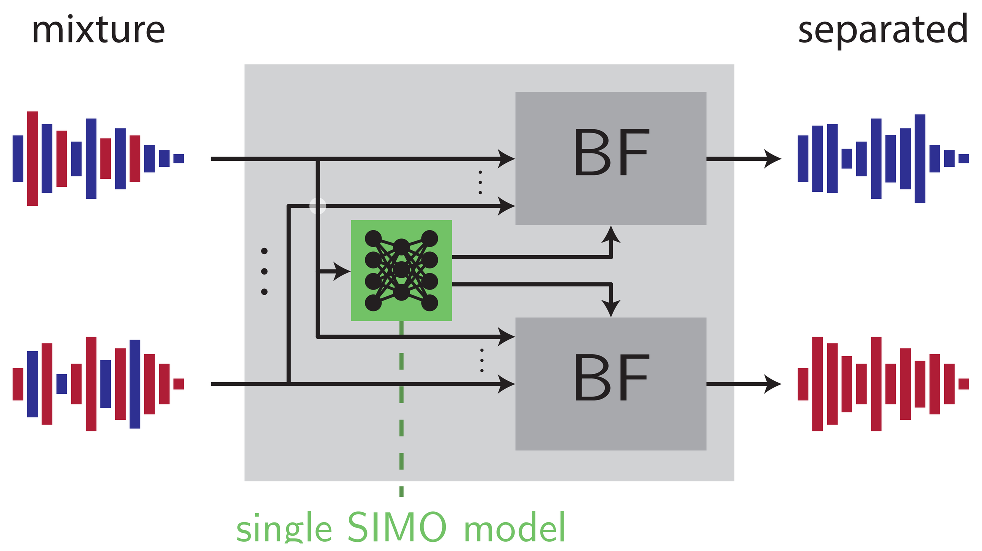
abstract—We develop an end-to-end system for multi-channel, multi-speaker automatic speech recognition. We propose a **frontend** for joint source separation and dereverberation based on the **independent vector analysis (IVA)** paradigm. The parameters from the ASR module and the frontend are optimized **jointly from the ASR loss**. We demonstrate competitive performance with previous systems using neural beamforming frontends with only one-ninth of the trainable parameter.

MIMO-Speech Paradigm [1, 2, 3]



- jointly train frontend and ASR model
- use non-parallel data, i.e., mixture/transcript
- demonstrate good ASR and separation performance

Neural Beamforming (MVDR, WPD, ... [3])



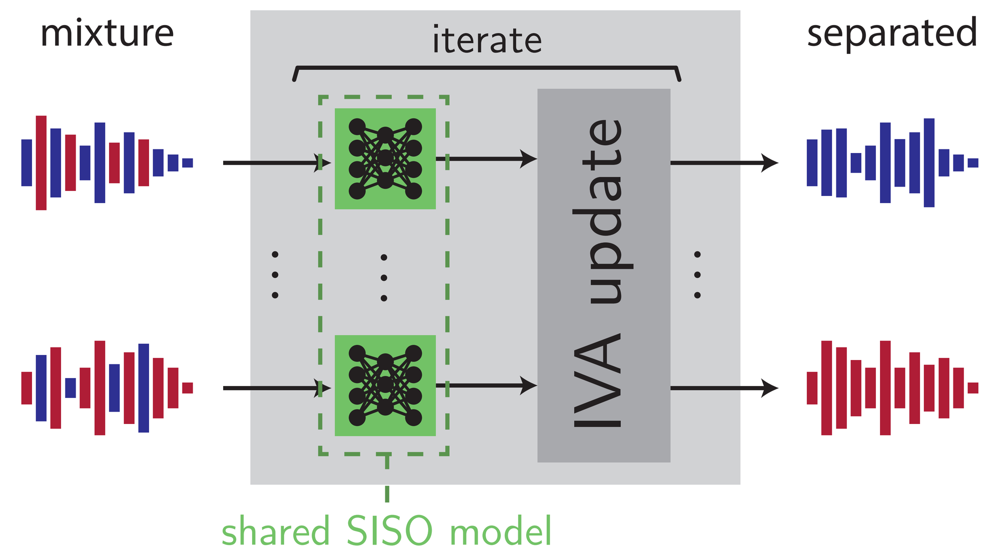
1. Masks: joint (SIMO)
2. Beamformers: one-by-one

Contributions

1. Extension of IVA to overdetermined case:
   - Time-decorrelation Iterative Source Steering (**T-ISS**) [6]
   - T-ISS with neural source model [Saijo2022]
   - **New: overdetermined (more mics than sources)**
2. Joint training of neural IVA frontend and ASR
   - Integration into ESPnet MIMO-Speech
   - Demonstrate **robustness** to noise mismatch
   - Demonstrate **flexible** number of speakers

Neural IVA Frontend

Overview [5, 7]



1. Masks: one-by-one (SISO)
2. Beamformers: joint

T-ISS updates for IVA

Cost function derived from maximum likelihood estimation

$$\mathcal{L}_+(\mathbf{W}) = \sum_{kn} \underbrace{u_{kn}(\mathbf{Y}_k)}_{\text{mask}} |\mathbf{w}_k^H \mathbf{x}_{fn}|^2 - 2 \log |\det \mathbf{W}|$$

where $\mathbf{W}$ contains beamforming filters in its rows.
Update $\mathbf{W}$ with **stable** rank-1 updates [6]

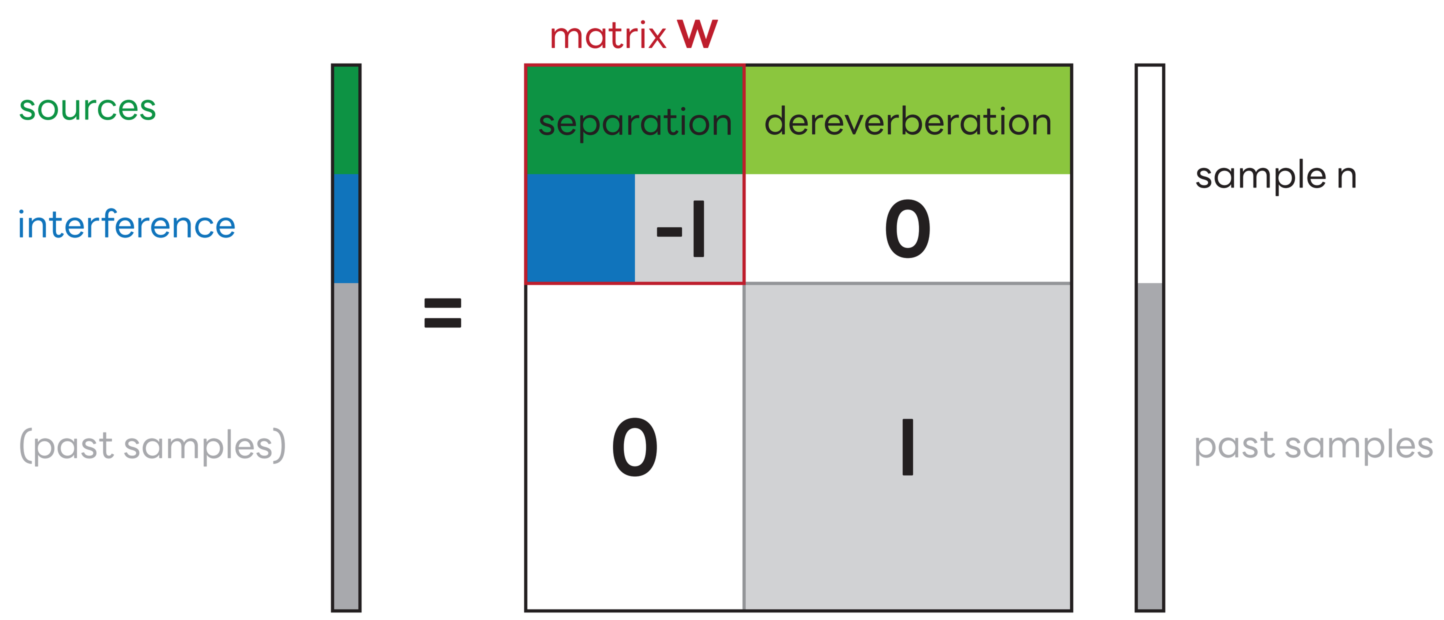
$$\mathbf{v} \leftarrow \arg\min_{\mathbf{v}} \mathcal{L}_+(\mathbf{W} - \tilde{\mathbf{v}}\mathbf{w}_k^H)$$
$$\mathbf{W} \leftarrow \mathbf{W} - \mathbf{v}\mathbf{w}_k^H$$

BF
- + Non-iterative
- - Stability issues (matrix inv.)
- - Brittle mask estimation

IVA
- + Flexible # speakers
- + Stable T-ISS algo. [6]
- - Iterative
- - Only for # source = # mics

Extension to # mics > # sources (new)

Parameterization of the Demixing Matrix



Apply different updates for
- separation (ISS, dark green)
- Dereverberation (ISS, light green)
- interference (IP, blue)

Experimental Validation

ASR Model

We use joint CTC/Attention encoder-decoder with $\hat{\mathbf{Y}}_k$ being **80-dim. log-Mel filterbank** features.

$$\mathbf{O}_k = \text{MVN-LMF}(\hat{\mathbf{Y}}_k),\ \mathbf{H}_k = \text{Enc}(\mathbf{O}_k),$$
$$\hat{\mathbf{R}}_k^{(\text{ctc})} = \text{CTC}(\mathbf{H}_k),\quad \hat{\mathbf{R}}_k^{(\text{dec})} = \text{AttentionDec}(\mathbf{H}_k),$$

The ASR loss is $\mathcal{L}_{asr} = \alpha\mathcal{L}_{ctc} + (1-\alpha)\mathcal{L}_{dec}$.

Number of frontend parameters

BF 23.15 M VS **IVA** 2.57 M

Datasets

Label	Speech	Noise
clean	WSJ1	(none)
noise1	WSJ1	CHiME3
noise2	WSJ1	TUT

Training (ESPnet)

- Adam optimizer
- Init. learning rate 1
- Warm-up 25000

Robustness Experiment

			WER (%) ↓		SIR (dB) ↑	
Test set	Train	Matched	BF	IVA	BF	IVA
WSJ1 clean	clean	O	9.57	**9.16**	13.9	**16.8**
WSJ1 + noise1	clean	X	17.12	**12.48**	12.3	**15.6**
	noise1	O	**11.40**	11.80	**14.7**	14.4
WSJ1 + noise2	clean	X	31.36	**14.55**	6.3	**13.7**
	noise1	X	15.17	**14.75**	10.0	**12.3**

Unseen Number of Speakers Experiment

Sources	Train	WER ↓	SIR ↑
3	clean	17.80 %	10.2 dB
	noise1	16.19 %	9.9 dB
4	clean	33.06 %	5.8 dB
	noise1	30.44 %	6.1 dB

References

[1] Chang et al., ASRU, 2019.
[2] Zhang et al., INTERSPEECH, 2020.
[3] Zhang et al., ICASSP, 2021.
[4] Scheibler & Ono, ICASSP, 2020.
[5] Scheibler & Togami, ICASSP, 2020.
[6] Nakashima et al., ICASSP, 2021.
[7] Saijo & Scheibler., INTERSPEECH, 2022.