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]

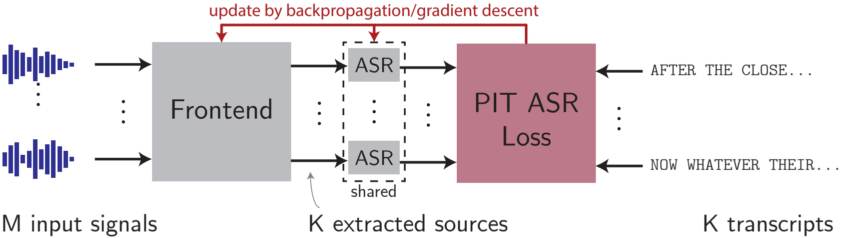
[1]LINE, [2]SJTU, [3]CMU

SLT2023

End-to-end Multispeaker ASR with Advanced Frontend

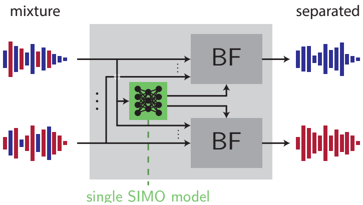
MIMO-Speech [Chang2019, Zhang2020, Zhang2021]

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



Conventional vs Independent Vector Analysis Frontend

Beamforming (e.g., MVDR)

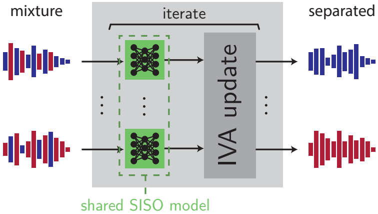


1. Masks: joint (SIMO)
2. Beamformers: one-by-one

Pro/Con

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

Neural IVA (this work)



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

Pro/Con

- + Flexible number of speakers
- + Stable IVA algo. [Nakashima2020]
- - Iterative

Our Contributions

1. Extension of IVA to overdetermined case:
    - Time-decorrelation Iterative Source Steering (**T-ISS**) [Nakashima2021]
    - 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

Experiment 1: Robustness to Noise Mismatch

clean : WSJ1
noise1: WSJ1 + CHiME3 (noise)
noise2: WSJ1 + TUT environ. sound

2 sources
Joint CTC-Attention
IVA 15 iterations

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

Number of frontend parameters

BF 23.15 M VS **IVA 2.57 M**

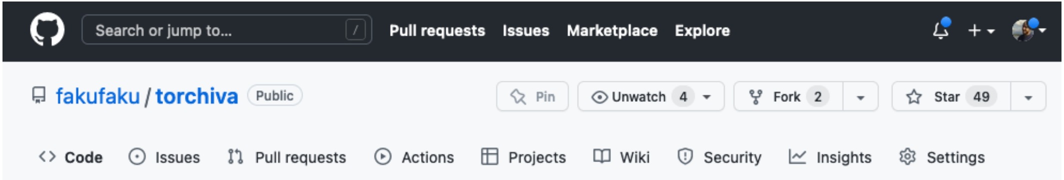
Experiment 2: IVA on Unseen Number of Speakers

Re-use model trained on **2-speakers** mixtures

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

Note: Neural BF cannot be applied due to SIMO mask model

torchiva: Pytorch Toolbox for IVA



```
stft = torchiva.STFT(n_fft=4096, hop_length=1024)
separator = torchiva.T_ISS(n_iter=10)

audio, fs = torchaudio.load("multichannel_mixture.wav")

X = stft(audio)
Y = separator(X)
y = stft.inv(Y)

torchaudio.save("separated_sources.wav", y, fs)
```

Summary

Summary

- IVA $=$ SISO neural model $+$ joint separation filter estimation
- joint training with ASR model
- torch IVA toolbox `https://git.linecorp.com/speechresearch/torchiva`

Advantage of IVA frontend in MIMO speech

- agnostic to # speakers/channels
- very robust to domain mismatch
- small model size (9x smaller)

References

Chang2019 Chang et al., **MIMO-SPEECH: End-to-End Multi-Channel Multi-Speaker Speech Recognition**, 2019, `https://arxiv.org/abs/1910.06522`

Zhang2020 W. Zhang et al., **End-to-End Far-Field Speech Recognition with Unified Dereverberation and Beamforming**, 2020, `https://arxiv.org/abs/2005.10479`

Zhang2021 Zhang et al., **End-to-End Dereverberation, Beamforming, and Speech Recognition with Improved Numerical Stability and Advanced Frontend**, 2021, `https://arxiv.org/abs/2102.11525`

Nakashima2021 Nakashima et al., **Joint Dereverberation and Separation with Iterative Source Steering**, 2021, `https://arxiv.org/abs/2102.06322`

Saijo2022 Saijo & Scheibler, **Independence-based Joint Dereverberation and Separation with Neural Source Model**, 2022, `https://arxiv.org/abs/2110.06545`