PEKING UNIVERSITY

Advanced Data & Signal Processing Laboratory

Speaker-Aware Mixture of Mixtures Training for Weakly Supervised Speaker Extraction

*Zifeng Zhao, Rongzhi Gu, Dongchao Yang, Jinchuan Tian, Yuexian Zou**

ADSPLAB, School of Electronics and Computer Engineering, Peking University

Paper 96

Background

Over the decades, lots of efforts have been made to crack the cocktail-party problem. One direction is to extract target speech with the auxiliary of an enrollment utterance from the target speaker.

$$y = \sum_{j=1}^{J} s_j + n$$

$$\hat{s}_t = SpkExtr(y|e_t;\theta)$$

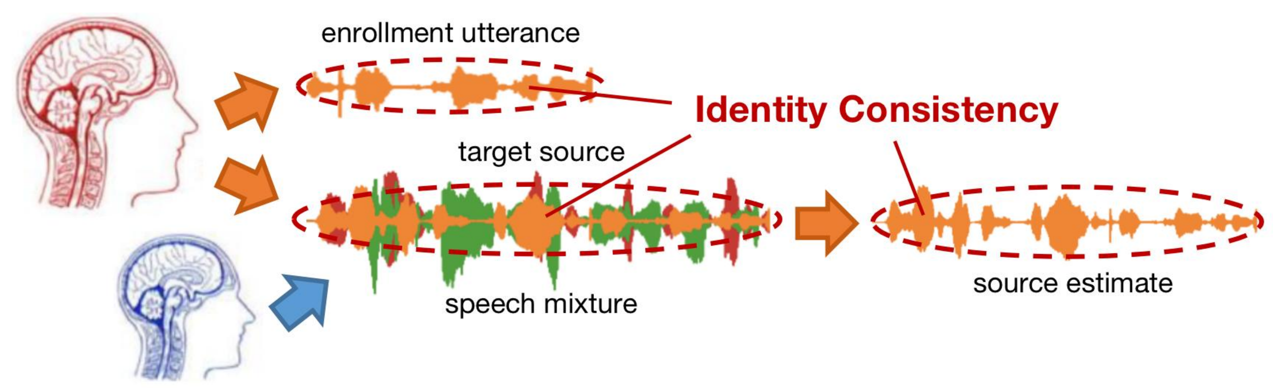
Motivation

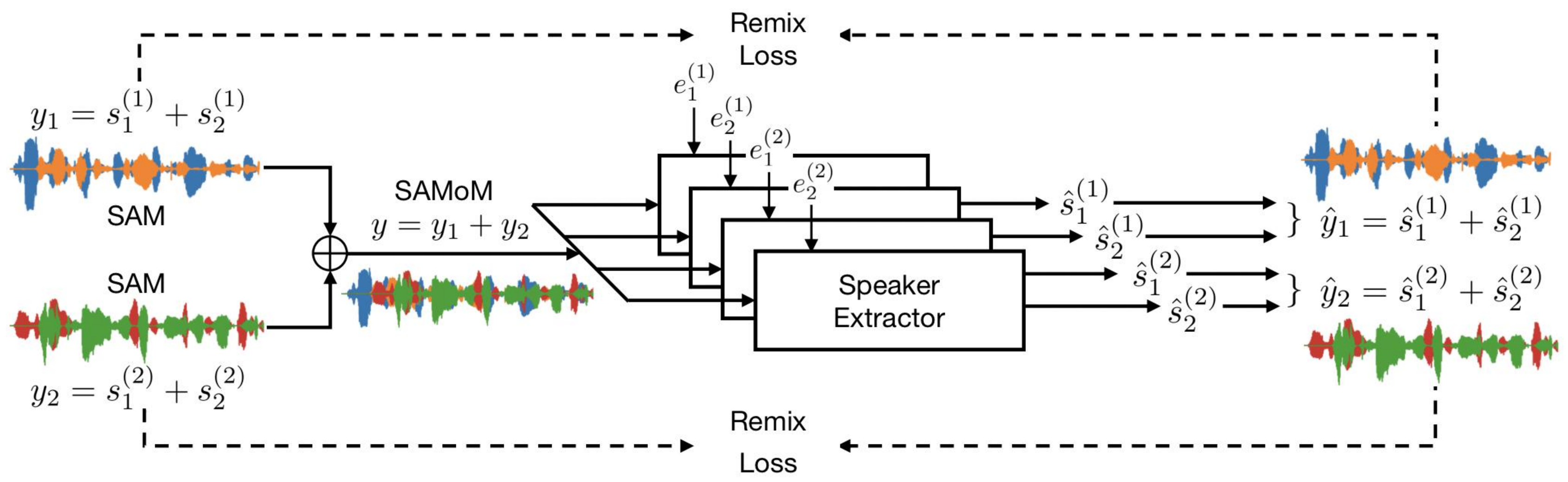
Dominant researches adopt supervised training for **target speaker extraction (TSE)**, while its mix-and-separate paradigm has two major drabacks:

- **need for clean corpus**: corpus with adequate clean utterances is required, serving as training ground truth as well as for simulating input mixtures
- **channel mismatch**: generalize poor in real-world scenarios, since there is usually a channel mismatch between the simulated data and the target domain

Methods

We propose **speaker-aware mixture of mixtures training (SAMoM)** as a wealy supervised learning framework for TSE task, by making advantages of the speaker identity consistency among target source, enrollment utterance and target estimate.





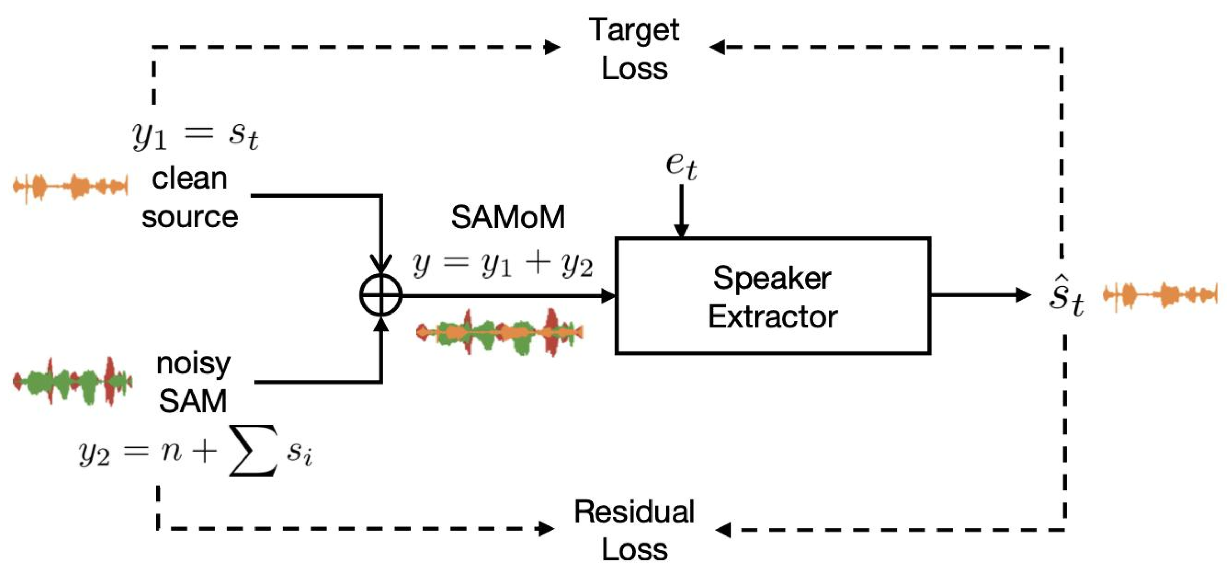
The proposed framework can be divided into 3 phases:

- **Mixture Generation**: mix up different speaker-aware mixtures (SAM) as the model's training input. SAM is a speech mixture but with speaker identities known and their enrollment utterances available
- **Target Speaker Extraction**: inform the model of target speakers' enrollment utterances, extract the target speech for each target speaker
- **SAM Remix**: remix target estimates according to identity consistency, so that the remixed mixtures approximate the original SAMs

Since SAMoM does not require any clean sources for training, it can adapt to the testing data through a weakly supervised fine-tuning, i.e. **domain adaptation (DA)**. This is helpful when channel mismatch exists.

Extension

SAMoM can be extended to a noisy setup for more general applications, but this may require a certain amount of single-speaker utterances as clean ground truths, turning into a semi-supervised paradigm.



Results

Key results are listed as follows:

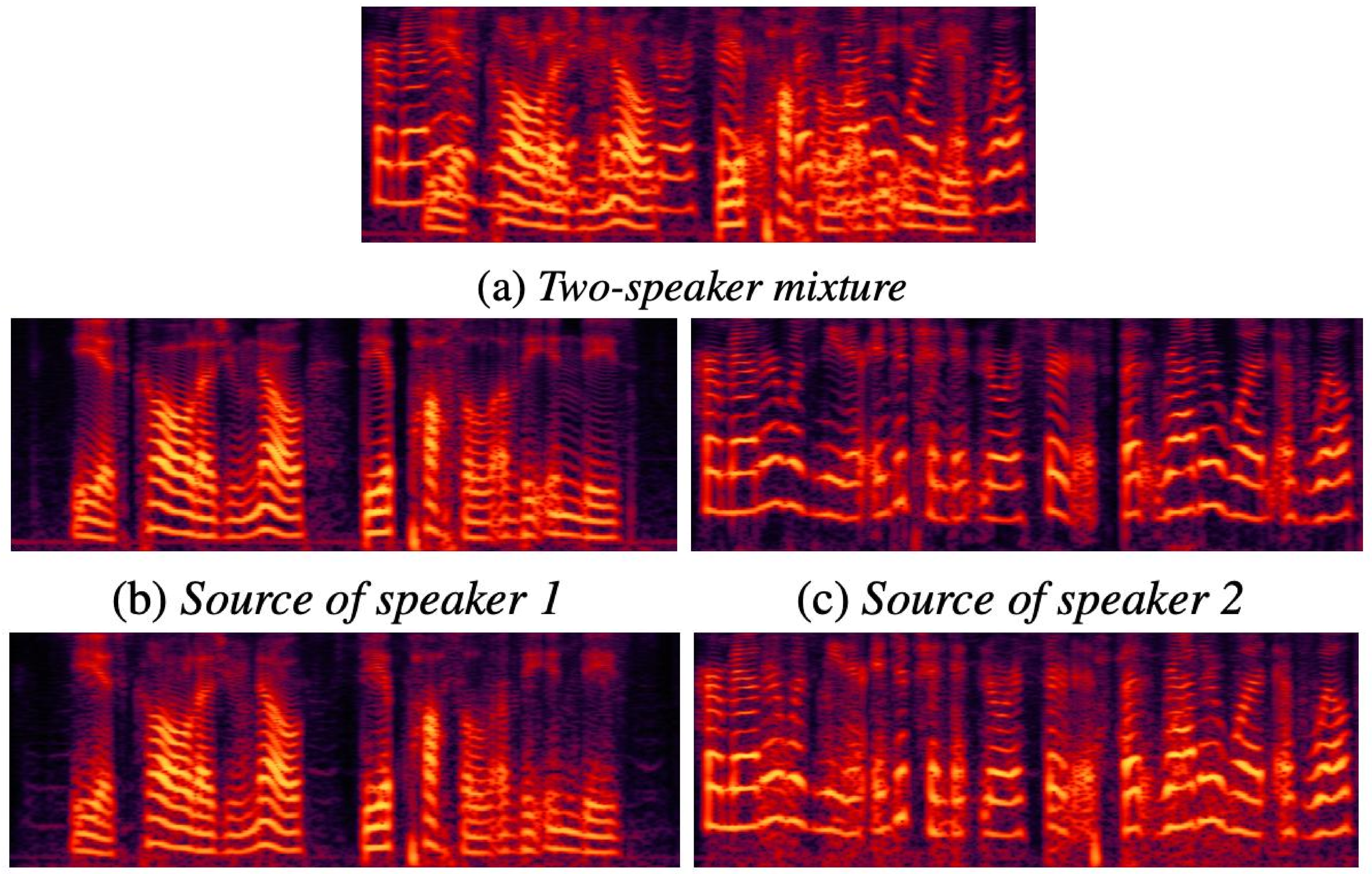
- Our weakly supervised manner achieves 11.06 dB SI-SDRi, which outperforms unsupervised MixIT and is close to the fully supervised baseline
- With domain adaptaion, SAMoM significantly outperforms supervised learning baseline in cross-domain evaluation

	SI-SDRi (dB)	SDRi (dB)	STOI	PESQ
sup SS	**13.40**	**13.82**	**0.92**	2.74
sup TSE	12.86	13.40	0.90	**2.75**
unsup MixIT	5.72	6.92	0.79	1.98
SAMoM	8.97	9.80	0.85	2.28
+Adaptation	**11.06**	**11.64**	**0.88**	**2.41**

Table 2: *Performance of different training methods for speech separation and speaker extraction on Libri2Mix.*

	SI-SDRi (dB)	SDRi (dB)	STOI	PESQ
sup TSE	1.99	2.65	0.68	1.77
+Adaptation	4.56	5.48	0.73	2.06
SAMoM	0.73	1.97	0.66	1.72
+Adaptation	**5.86**	**6.64**	**0.75**	**2.12**

Table 3: *Cross-domain evaluation on aishell1-2mix.*



(a) *Two-speaker mixture*

(b) *Source of speaker 1*

(c) *Source of speaker 2*

(d) *Estimate of speaker 1*

(e) *Estimate of speaker 2*