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

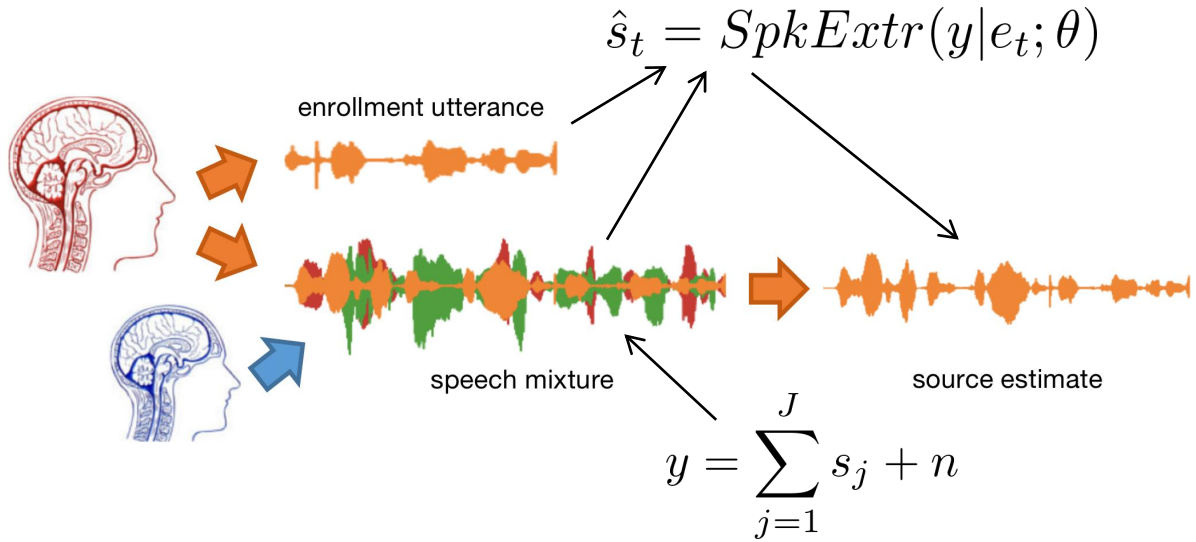
INTERSPEECH 2022

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

ADSPLAB, School of Electronics and Computer Engineering, Peking University

Background

- Cocktail-party Problem
  - speech sepataion (SS)
  - target speaker extracion (TSE)

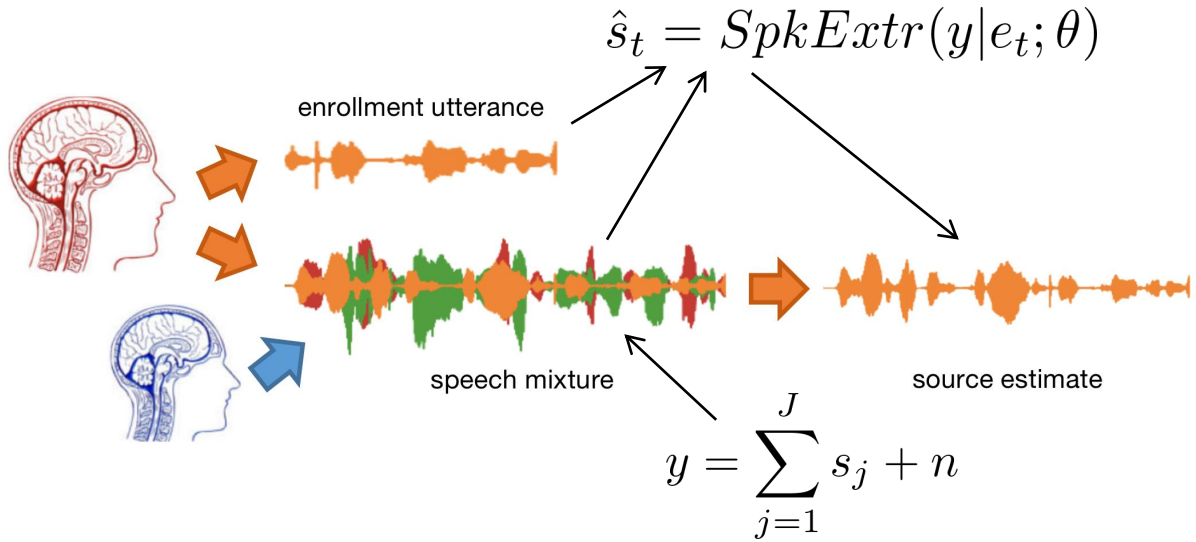


Motivation

- Drawbacks of mix-and-separate paradigm
  - need of clean corpus
  - channel mismatch

Background

- Cocktail-party Problem
  - speech sepataion (SS)
  - target speaker extracion (TSE)

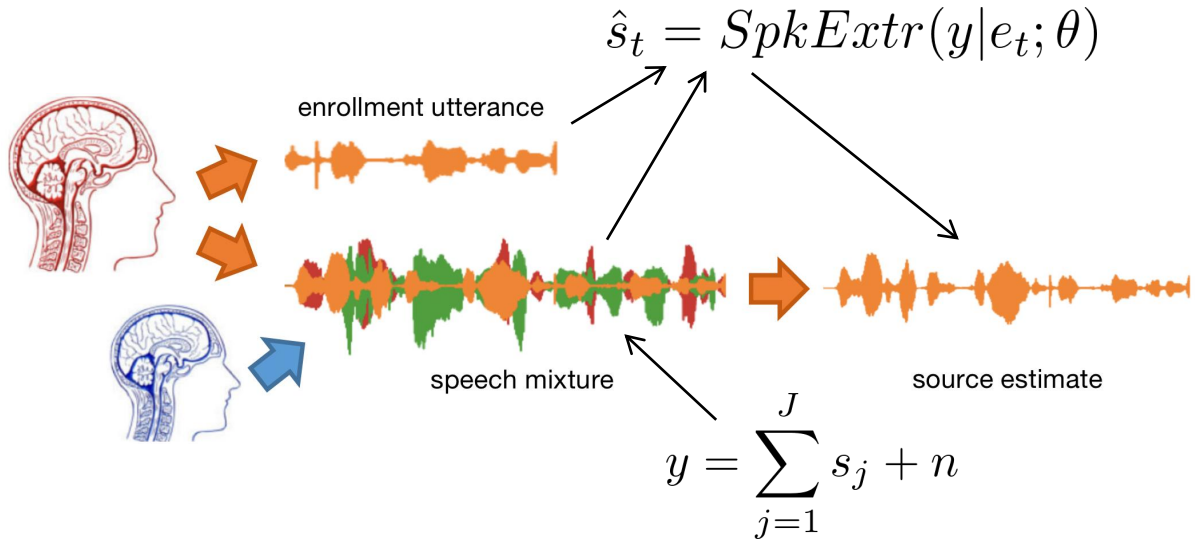


Motivation

- Drawbacks of mix-and-separate paradigm
  - need of clean corpus
  - channel mismatch

Background

- Cocktail-party Problem
  - speech sepataion (SS)
  - target speaker extracion (TSE)

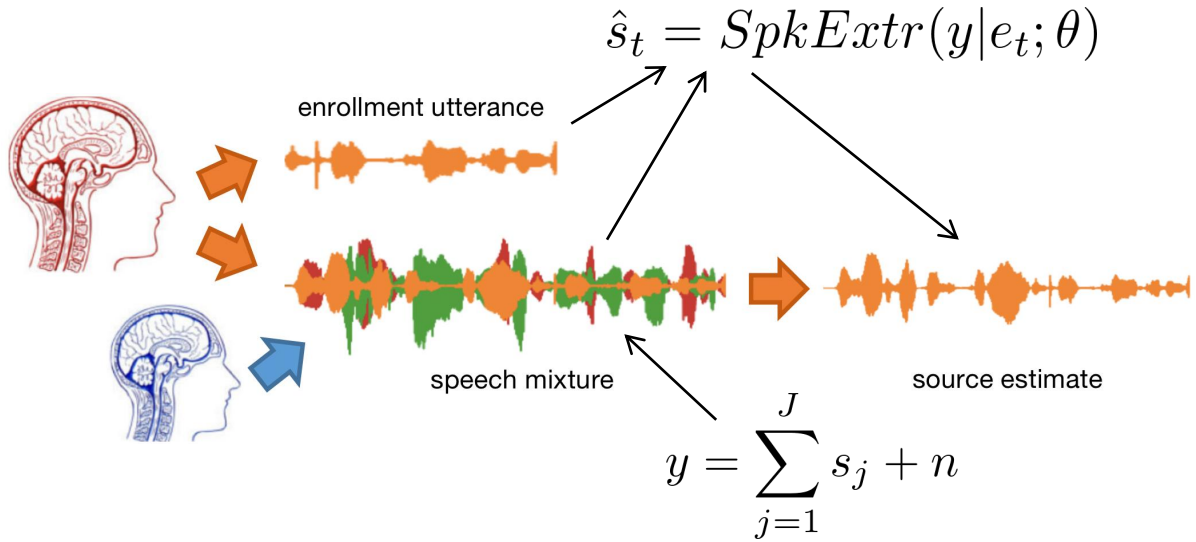


Motivation

- Drawbacks of mix-and-separate paradigm
  - need of clean corpus
  - channel mismatch

Background

- Cocktail-party Problem
  - speech sepataion (SS)
  - target speaker extracion (TSE)

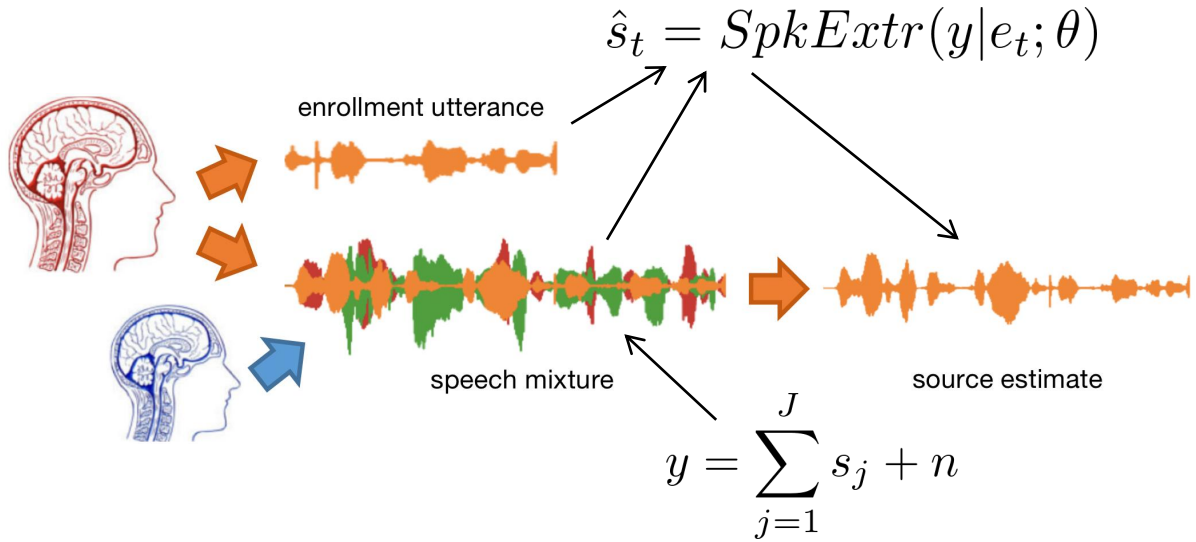


Motivation

- Drawbacks of mix-and-separate paradigm
  - need of clean corpus
  - channel mismatch

Background

- Cocktail-party Problem
  - speech sepataion (SS)
  - target speaker extracion (TSE)

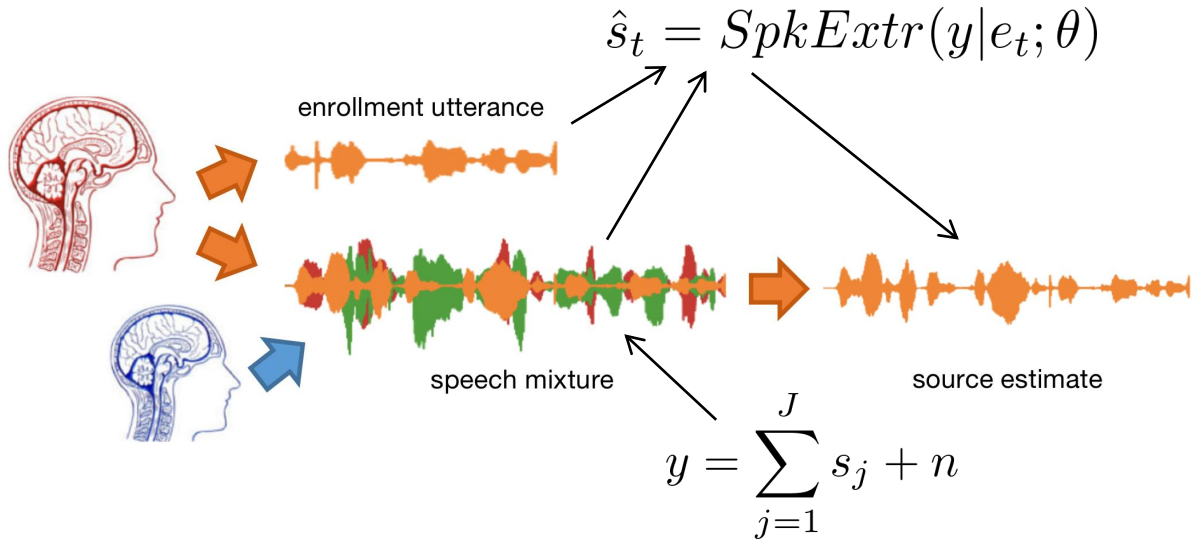


Motivation

- Drawbacks of mix-and-separate paradigm
  - need of clean corpus
  - channel mismatch

Background

- Cocktail-party Problem
  - speech sepataion (SS)
  - target speaker extracion (TSE)

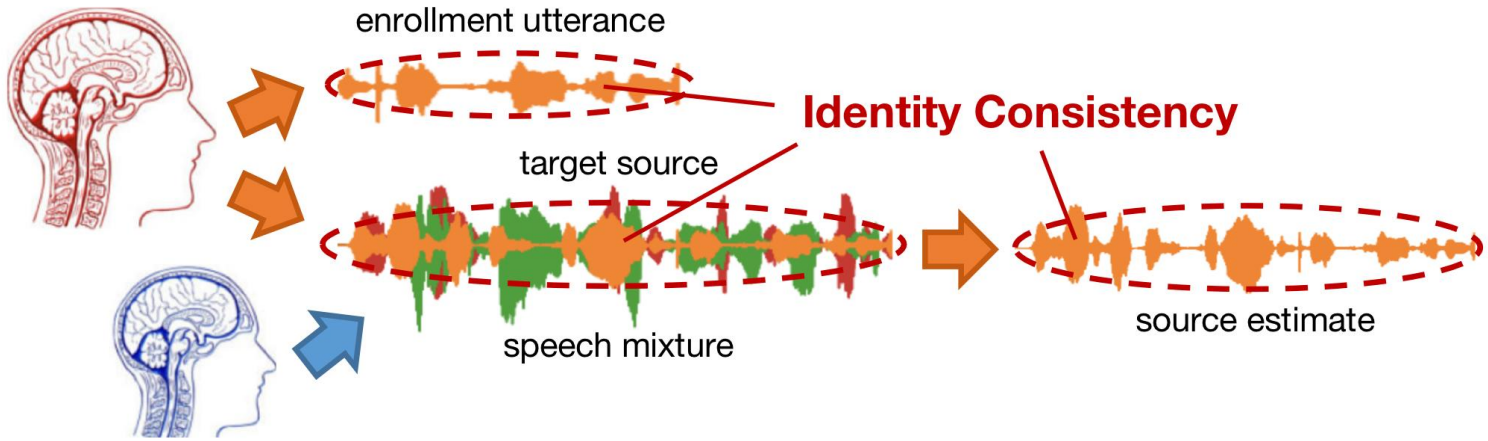


Motivation

- Drawbacks of mix-and-separate paradigm
  - need of clean corpus
  - channel mismatch

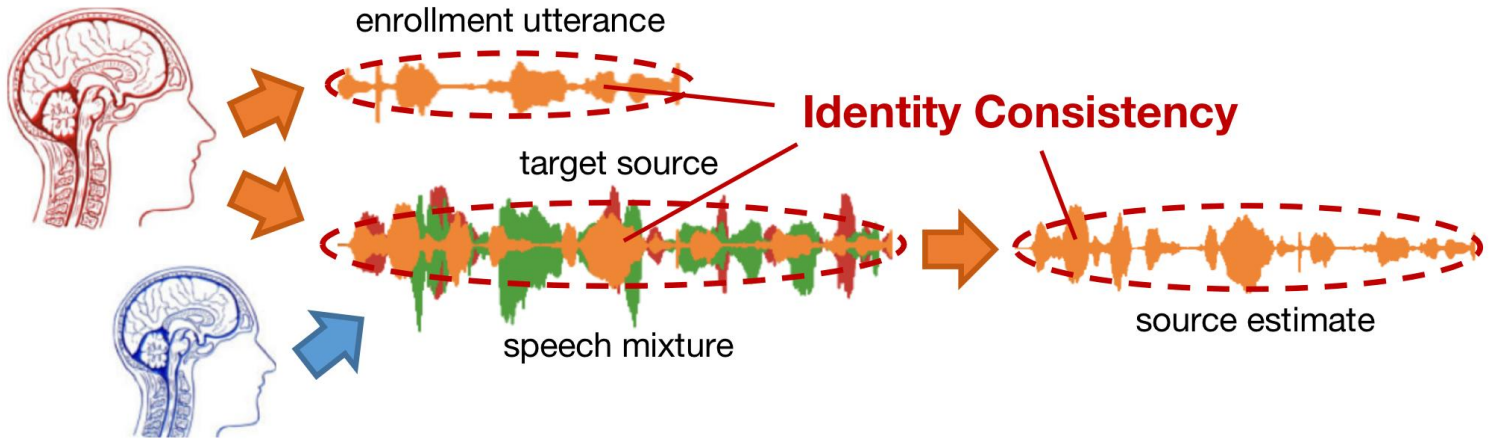
Speaker-Aware Mixture of Mixtures Training (SAMoM)

- **Intuition**
  - weakly supervised learning
  - premordial speech mixtures as training samples
  - speaker identity consistency among: target / enrollment / estimate



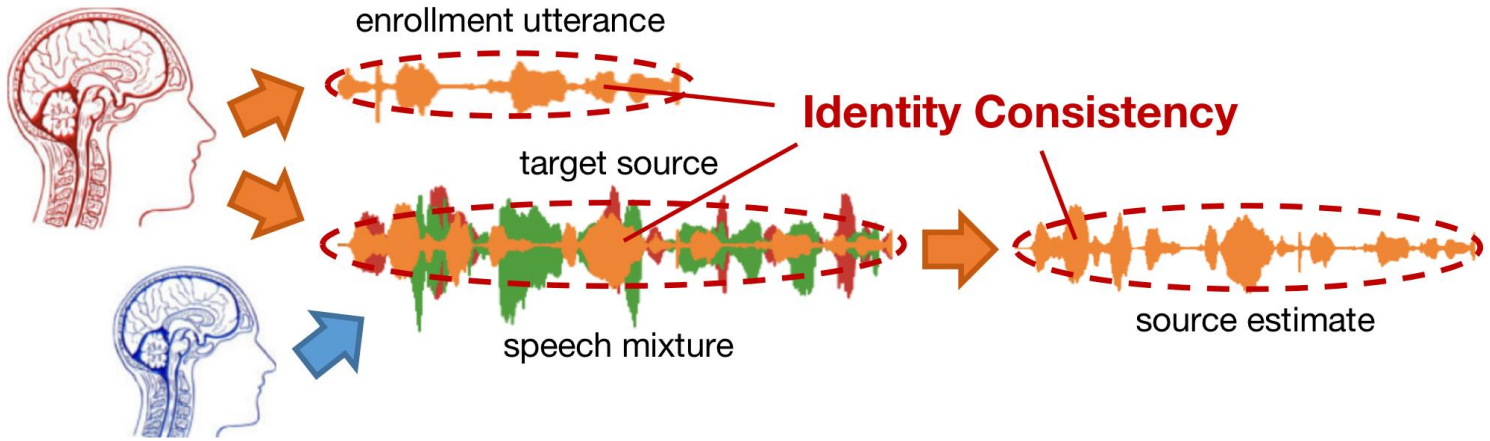
Speaker-Aware Mixture of Mixtures Training (SAMoM)

- **Intuition**
  - weakly supervised learning
  - premordial speech mixtures as training samples
  - speaker identity consistency among: target / enrollment / estimate



Speaker-Aware Mixture of Mixtures Training (SAMoM)

- **Intuition**
  - weakly supervised learning
  - premordial speech mixtures as training samples
  - speaker identity consistency among: target / enrollment / estimate



Speaker-Aware Mixture of Mixtures Training (SAMoM)

- **Methods**
  - **STEP 1**: Input Generation

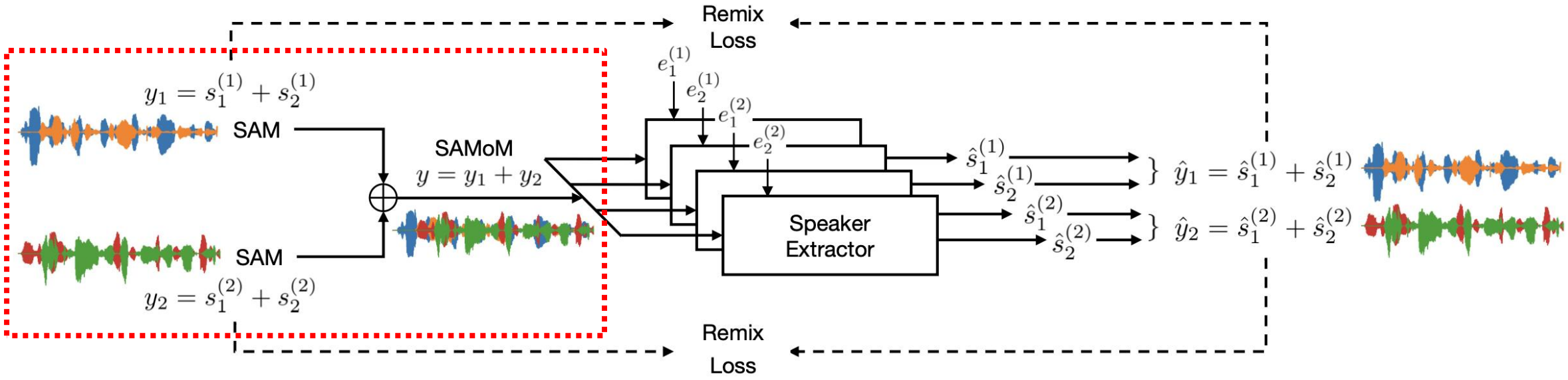


Figure 2: *The proposed SAMoM training framework.*

Speaker-Aware Mixture of Mixtures Training (SAMoM)

- **Methods**
  - **STEP 2**: Targe Speaker Extraction

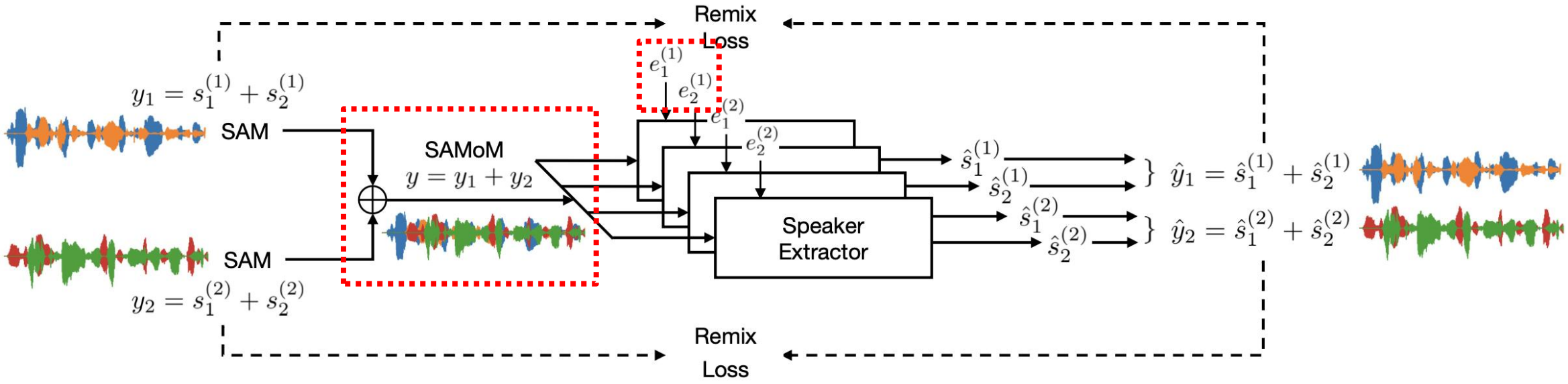


Figure 2: *The proposed SAMoM training framework.*

Speaker-Aware Mixture of Mixtures Training (SAMoM)

- **Methods**
  - **STEP 2**: Targe Speaker Extraction

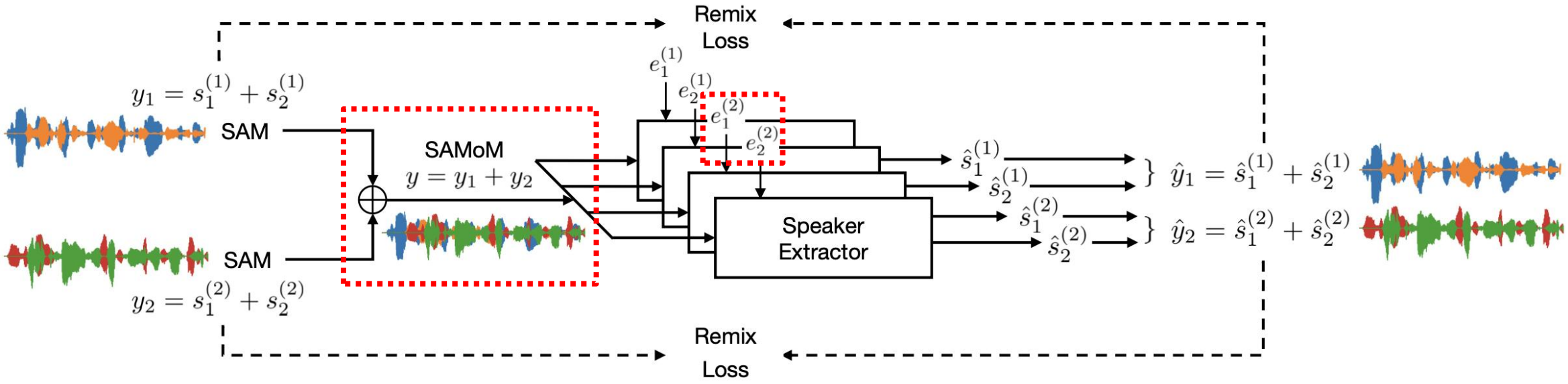


Figure 2: *The proposed SAMoM training framework.*

Speaker-Aware Mixture of Mixtures Training (SAMoM)

- **Methods**
  - **STEP 2**: Targe Speaker Extraction

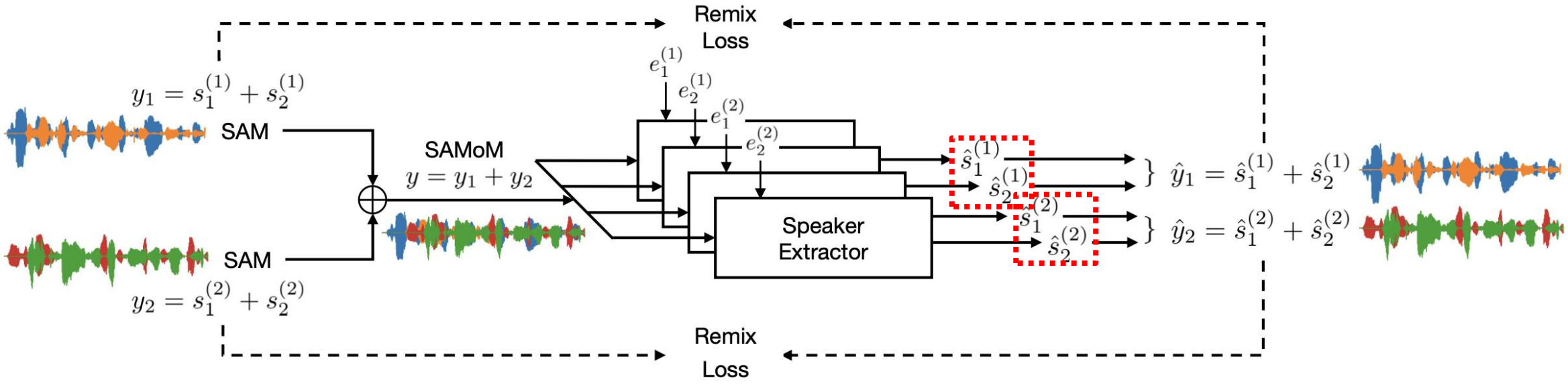


Figure 2: *The proposed SAMoM training framework.*

Speaker-Aware Mixture of Mixtures Training (SAMoM)

- **Methods**
  - **STEP 3**: Mixture Remix

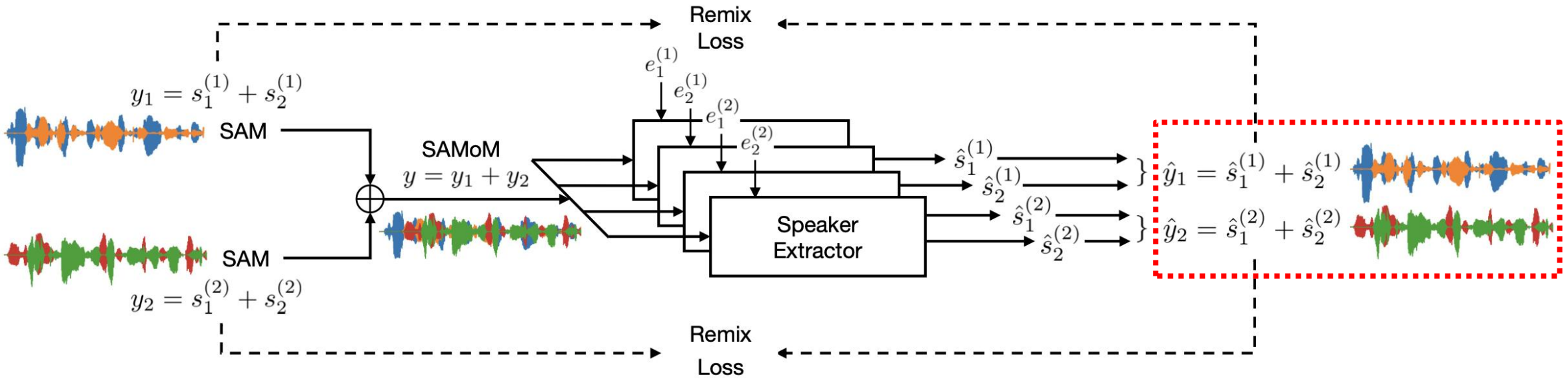


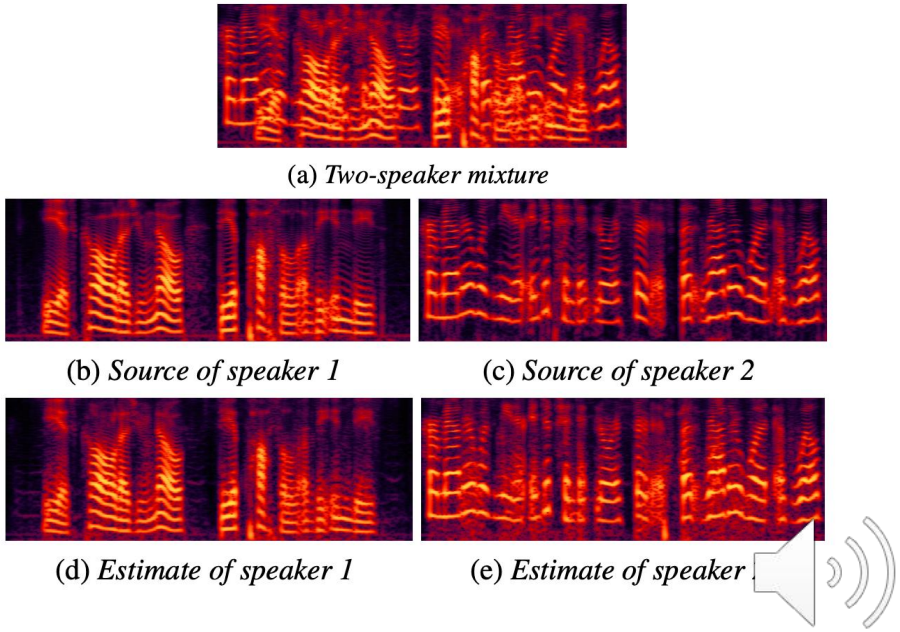
Figure 2: *The proposed SAMoM training framework.*

Experiments

- **Exp1: Proposed VS Baselines**
  - **Data**
    - ✓ **trainset**: Libri2Mix (8kHz)
    - ✓ **testset**: Libri2Mix (8kHz)
  - **Models**
    - ✓ **SS**: Conv-TasNet
    - ✓ **TSE**: TD-SpeakerBeam
  - **Training Methods**
    - ✓ fully supervised baselines (for SS and TSE)
    - ✓ unsupervised MixIT (for SS)
    - ✓ weakly supervised SAMoM (for TSE)
    - ✓ domain adaptation with SAMoM (for TSE)

	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.*



(a) *Two-speaker mixture*

(b) *Source of speaker 1*

(c) *Source of speaker 2*

(d) *Estimate of speaker 1*

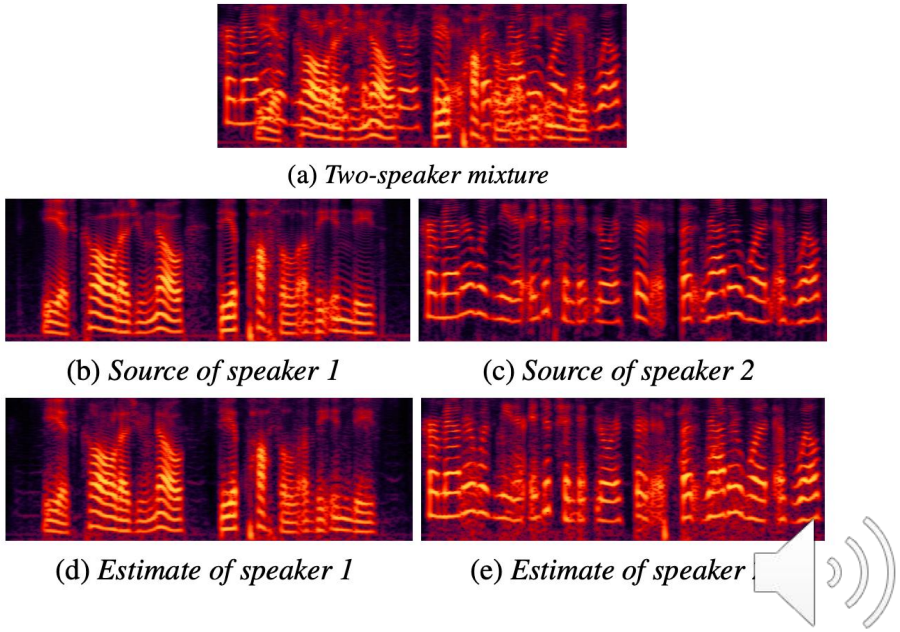
(e) *Estimate of speaker*

Experiments

- **Exp1: Proposed VS Baselines**
  - **Data**
    - ✓ **trainset**: Libri2Mix (8kHz)
    - ✓ **testset**: Libri2Mix (8kHz)
  - **Models**
    - ✓ **SS**: Conv-TasNet
    - ✓ **TSE**: TD-SpeakerBeam
  - **Training Methods**
    - ✓ fully supervised baselines (for SS and TSE)
    - ✓ unsupervised MixIT (for SS)
    - ✓ weakly supervised SAMoM (for TSE)
    - ✓ domain adaptation with SAMoM (for TSE)

	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.*



(a) *Two-speaker mixture*

(b) *Source of speaker 1*

(c) *Source of speaker 2*

(d) *Estimate of speaker 1*

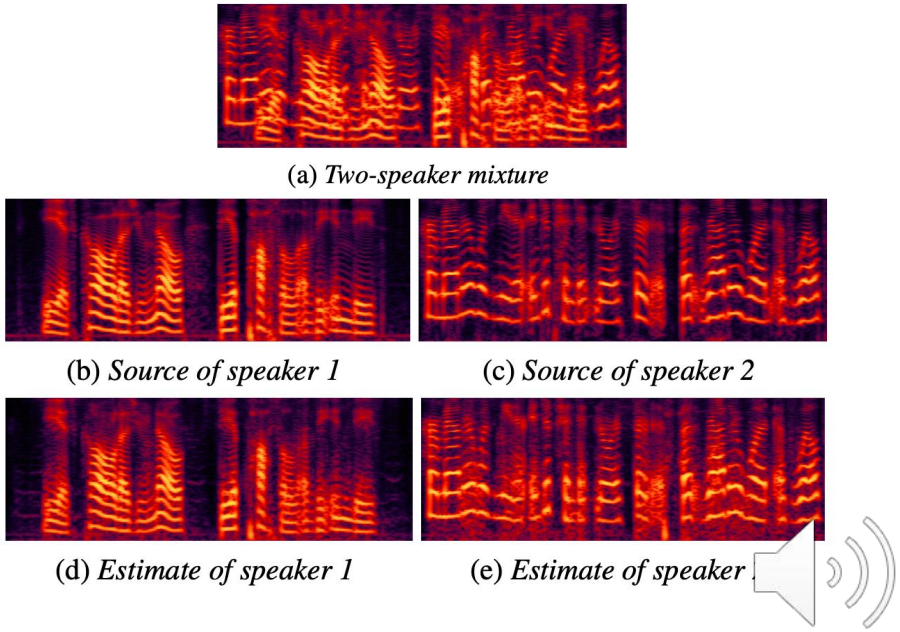
(e) *Estimate of speaker*

Experiments

- **Exp1: Proposed VS Baselines**
  - **Data**
    - ✓ **trainset**: Libri2Mix (8kHz)
    - ✓ **testset**: Libri2Mix (8kHz)
  - **Models**
    - ✓ **SS**: Conv-TasNet
    - ✓ **TSE**: TD-SpeakerBeam
  - **Training Methods**
    - ✓ fully supervised baselines (for SS and TSE)
    - ✓ unsupervised MixIT (for SS)
    - ✓ weakly supervised SAMoM (for TSE)
    - ✓ domain adaptation with SAMoM (for TSE)

	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.*



(a) *Two-speaker mixture*

(b) *Source of speaker 1*

(c) *Source of speaker 2*

(d) *Estimate of speaker 1*

(e) *Estimate of speaker*

Experiments

- **Exp2: Cross-Domain Evaluation**
  - **Data**
    - ✓ **trainset**: Libri2Mix (8kHz)
    - ✓ **testset**: aishell1-2mix (8kHz)
  - **Models**
    - ✓ **SS**: Conv-TasNet
    - ✓ **TSE**: TD-SpeakerBeam
  - **Training Methods**
    - ✓ fully supervised baselines (for TSE)
    - ✓ weakly supervised SAMoM (for TSE)
    - ✓ domain adaptation with SAMoM (for TSE)

	Libri2Mix / test set	*aishell1-2mix* / eval set
#Speakers	40	60
#Utterances	3000	2500
Hours	11	2.08
Language	English	Chinese

Table 1: *A comparison between the test set of Libri2Mix and the evaluation set of aishell1-2mix.*

	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.*

Experiments

- **Exp2: Cross-Domain Evaluation**
  - **Data**
    - ✓ **trainset**: Libri2Mix (8kHz)
    - ✓ **testset**: aishell1-2mix (8kHz)
  - **Models**
    - ✓ **SS**: Conv-TasNet
    - ✓ **TSE**: TD-SpeakerBeam
  - **Training Methods**
    - ✓ fully supervised baselines (for TSE)
    - ✓ weakly supervised SAMoM (for TSE)
    - ✓ domain adaptation with SAMoM (for TSE)

	Libri2Mix / test set	*aishell1-2mix* / eval set
#Speakers	40	60
#Utterances	3000	2500
Hours	11	2.08
Language	English	Chinese

Table 1: *A comparison between the test set of Libri2Mix and the evaluation set of aishell1-2mix.*

	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.*

Thanks for Your Attention !

That's All

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

ADSPLAB, School of Electronics and Computer Engineering, Peking University