TS-SEP: Joint Diarization and Separation Conditioned on Estimated Speaker Embeddings

Christoph Boeddeker[1,2], Aswin Shanmugam Subramanian[2], Gordon Wichern[2], Reinhold Haeb-Umbach[1], and Jonathan Le Roux[2]

[1]Paderborn University, Germany
[2]Mitsubishi Electric Research Laboratories (MERL), USA

2025 icassp HYDERABAD, INDIA
Celebrating Signal Processing

IEEE TRANSACTIONS ON
AUDIO, SPEECH AND
LANGUAGE PROCESSING

Highlight

- TS-SEP: Extension of TS-VAD [8] from diarization to joint diarization and separation
- SotA[1] on LibriCSS
  - Single channel: 7.81 % vs. 11.6 % cpWER
  - Multi channel: 5.36 % vs. 5.9 % cpWER
- Extensive experimental evaluation in paper
  - Activity estimation / Segmentation analysis
    - Trade-off between DER and WER
    - WER prefers "overestimation"
  - Impact of ignoring multichannel information on system components
  - Comparisons of extractions: masking, beamforming, GSS
  - TS-SEP enables faster GSS, with minimal performance impact
  - Literature comparison
- Open Source PyTorch implementation of TS-VAD and TS-SEP
  - `https://github.com/merlresearch/tssep`
  - Improved TS-VAD: 5.7 % vs. 11.2 % cpWER

[1] At time of publication

[8] I. Medennikov et al, "The STC system for the CHiME-6 challenge", CHiME 2020

Highlight

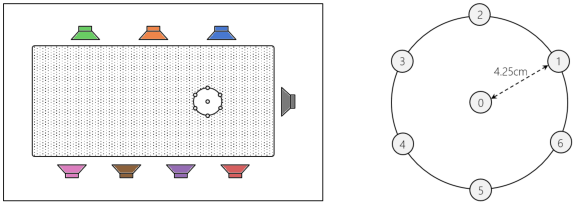
- TS-SEP: Extension of TS-VAD [8] from diarization to joint diarization and separation
- SotA[1] on LibriCSS
  - Single channel: 7.81 % vs. 11.6 % cpWER
  - Multi channel: 5.36 % vs. 5.9 % cpWER
- Extensive experimental evaluation in paper
  - Activity estimation / Segmentation analysis
    - Trade-off between DER and WER
    - WER prefers "overestimation"
  - Impact of ignoring multichannel information on system components
  - Comparisons of extractions: masking, beamforming, GSS
  - TS-SEP enables faster GSS, with minimal performance impact
  - Literature comparison
- Open Source PyTorch implementation of TS-VAD and TS-SEP
  - `https://github.com/merlresearch/tssep`
  - Improved TS-VAD: 5.7 % vs. 11.2 % cpWER

[1] At time of publication

[8] I. Medennikov et al, "The STC system for the CHiME-6 challenge", CHiME 2020

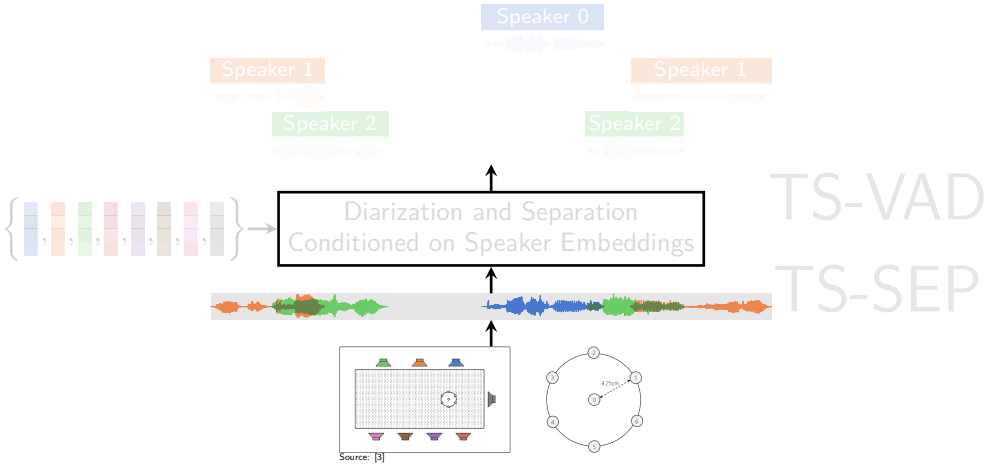
Motivation: Meeting transcription

- Meeting setup
  - Office meeting / informal conversations
  - Minutes to hours
  - Many speaker
  - Varying concurrent activity ($0 - K$ speakers)
- Applications:
  - Automatic protocolling
  - Smart home assistance

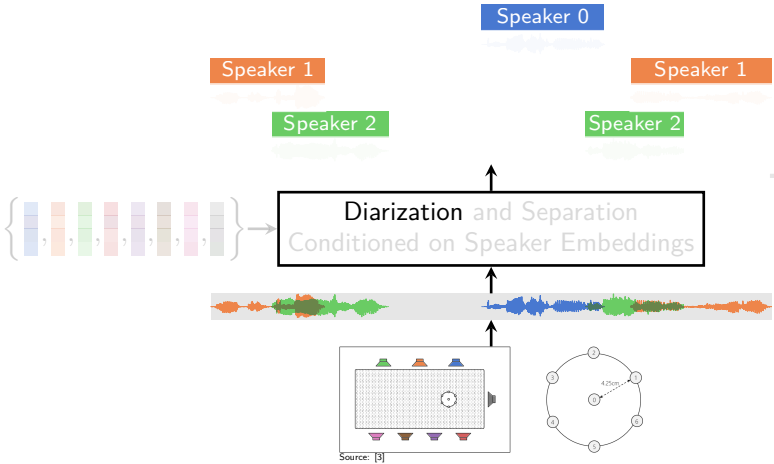


Source: [3] Z. Chen et al, "Continuous Speech Separation: Dataset and Analysis", ICASSP 2020

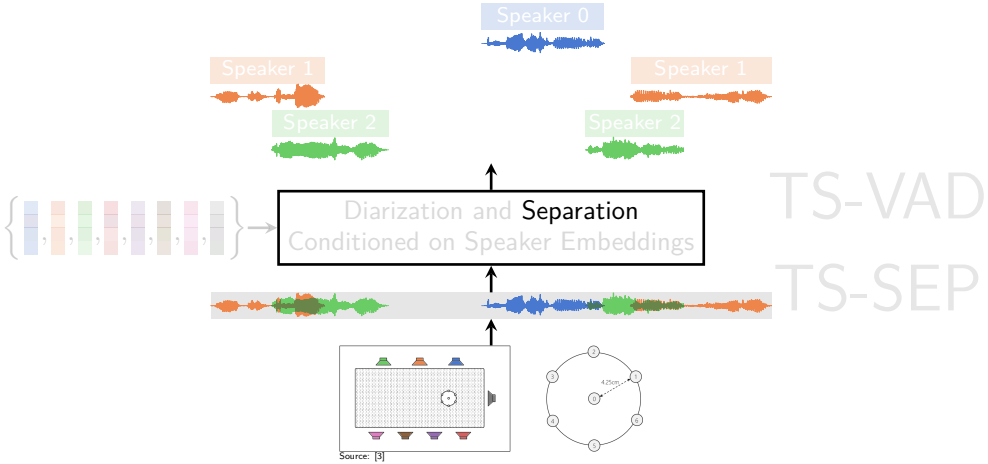
Technical terms



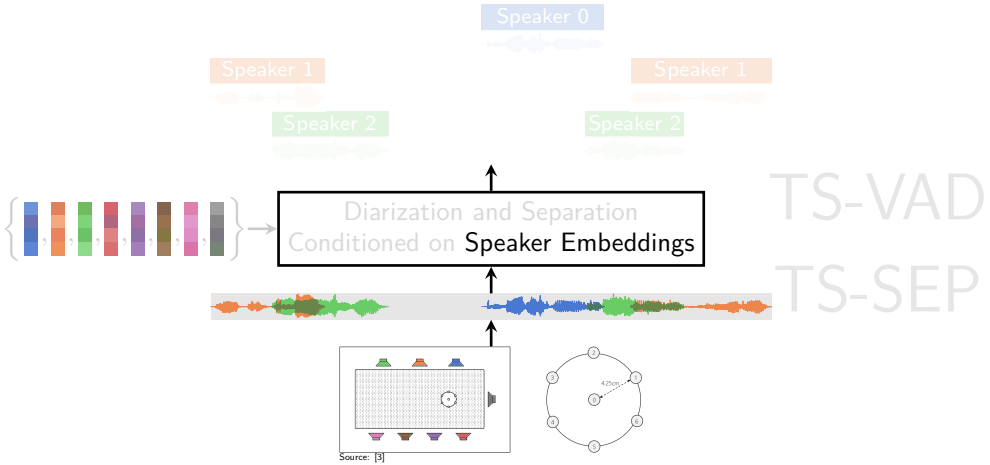
Technical terms



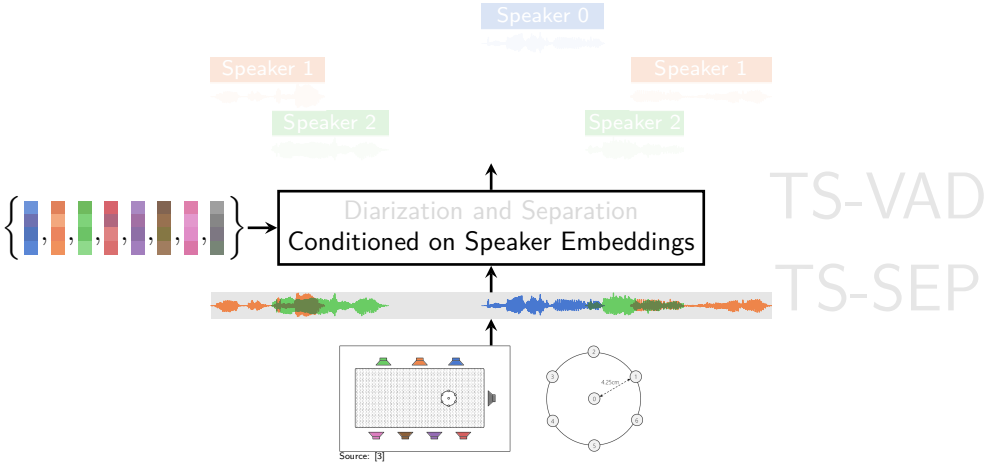
Technical terms



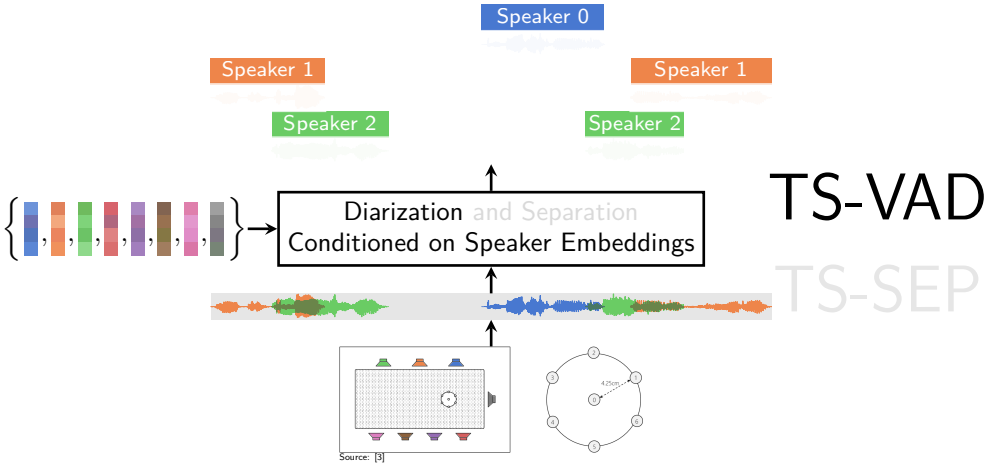
Technical terms



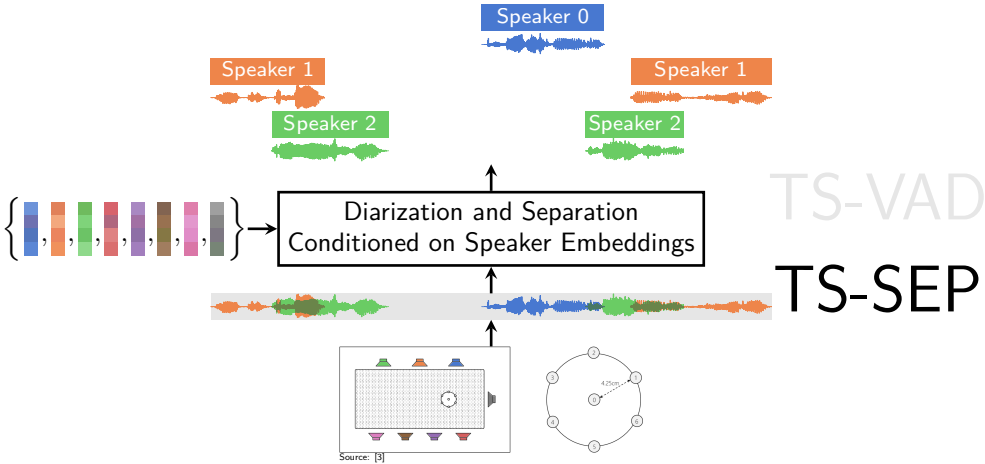
Technical terms



Technical terms



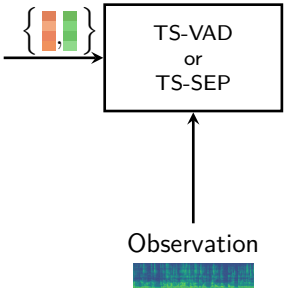
Technical terms



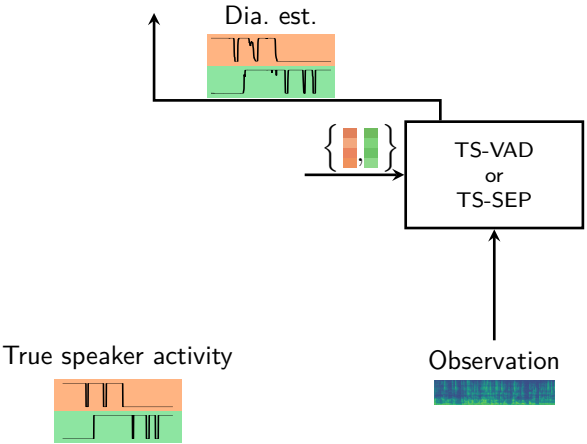
Training, Embedding and inference

Simplification for presentation

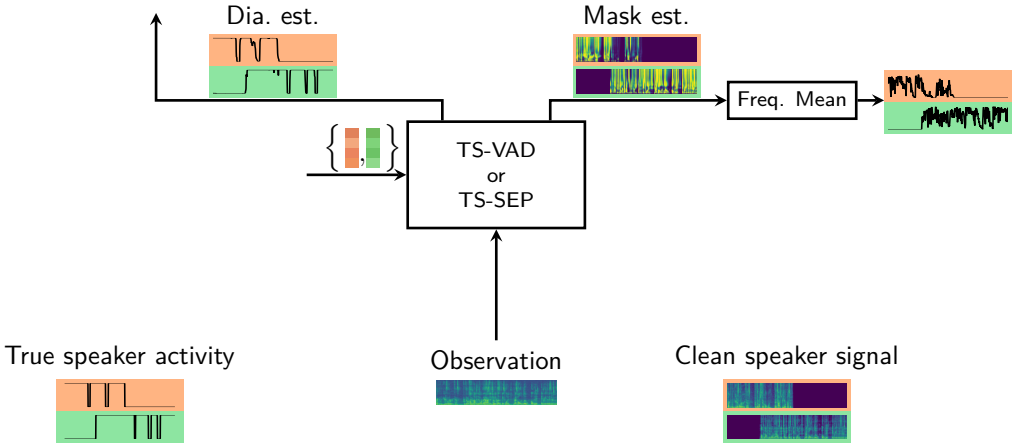
- We show 10 s excerpt from full 10 min recording
- We only show 2 of the 8 active speakers



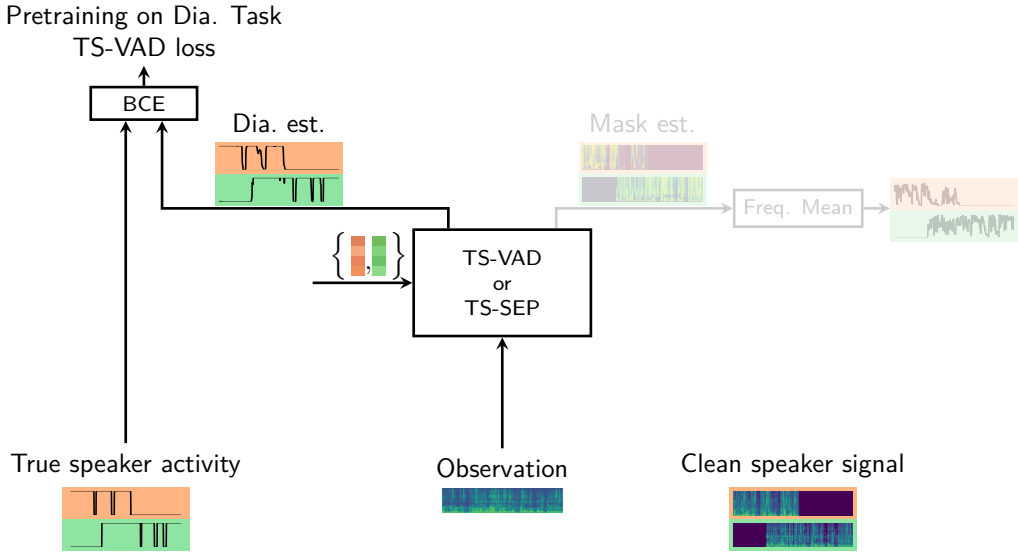
Training, Embedding and inference



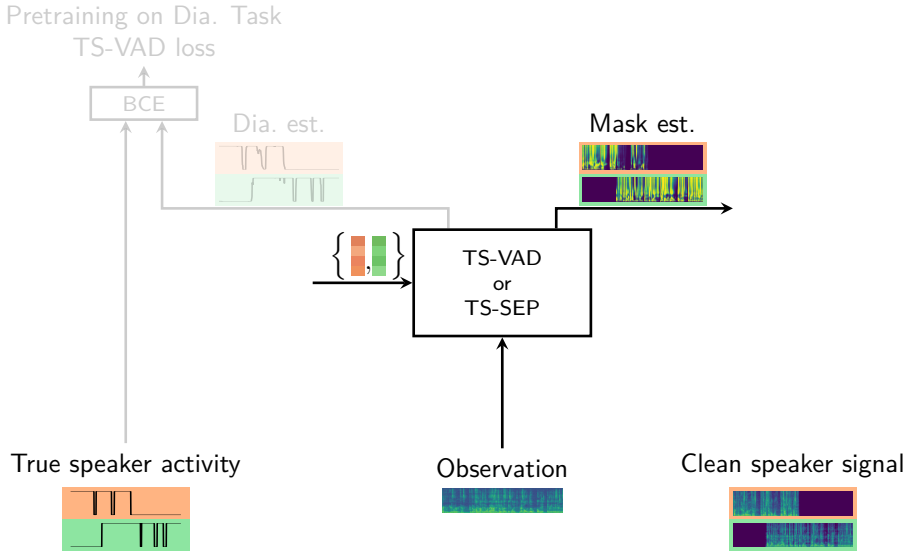
Training, Embedding and inference



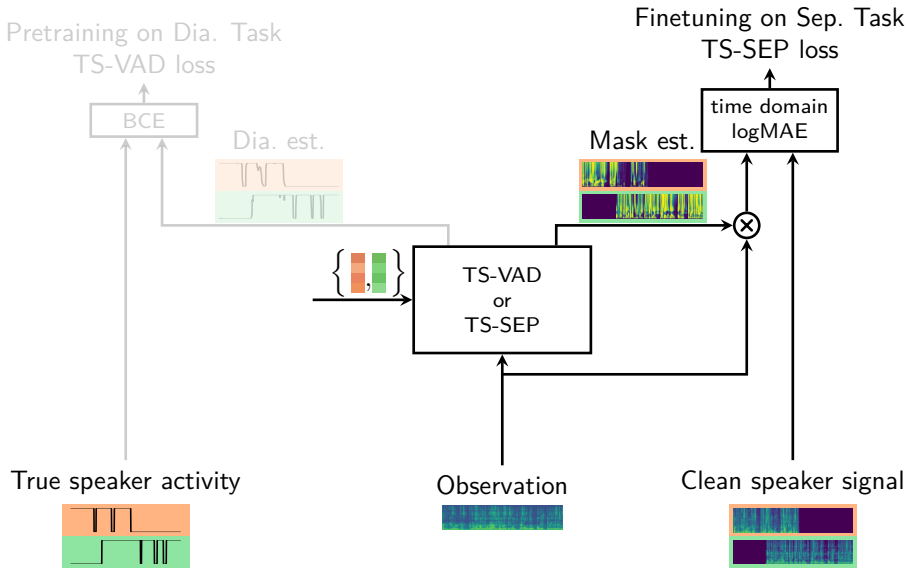
Training, Embedding and inference



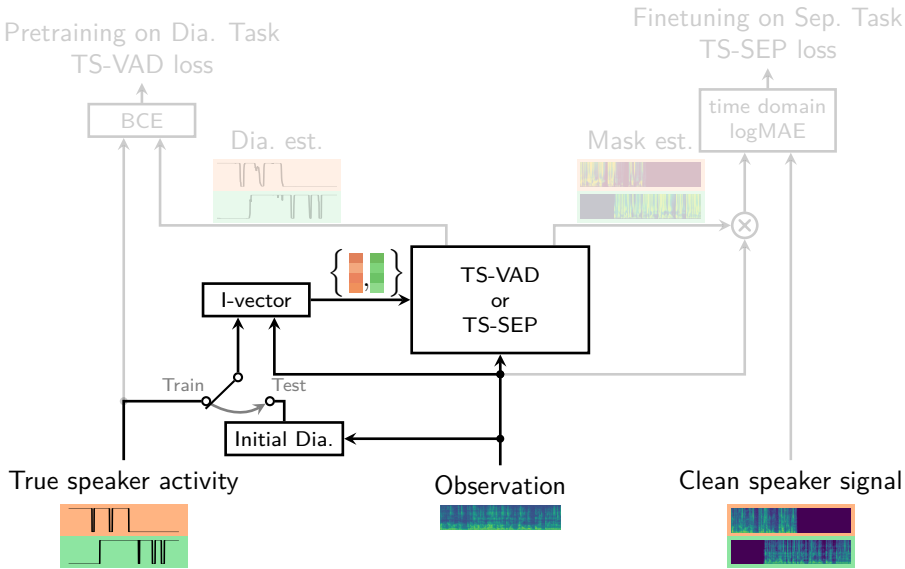
Training, Embedding and inference



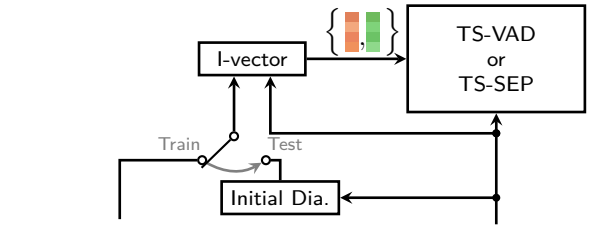
Training, Embedding and inference



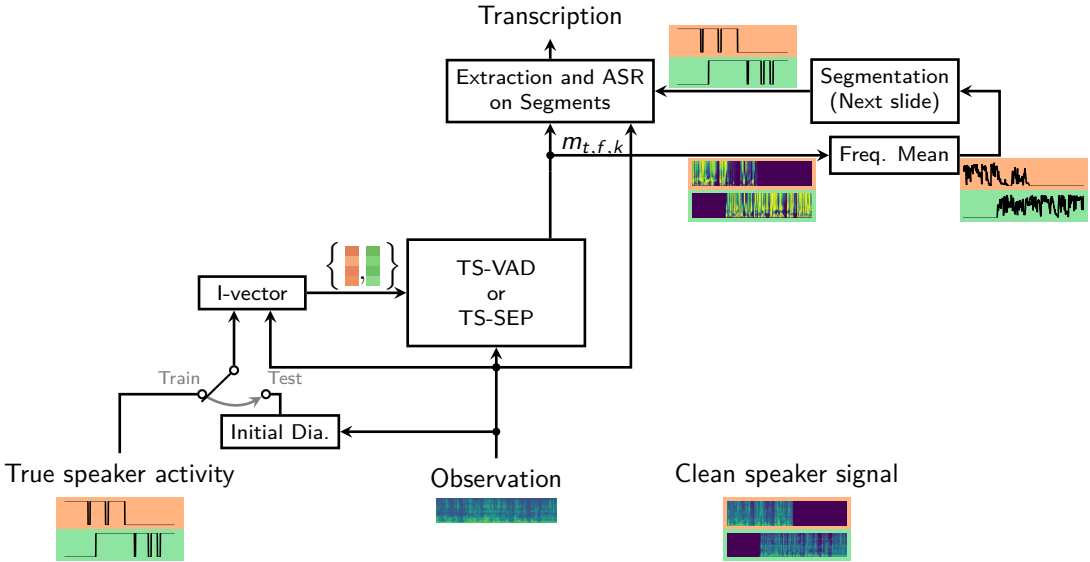
Training, Embedding and inference



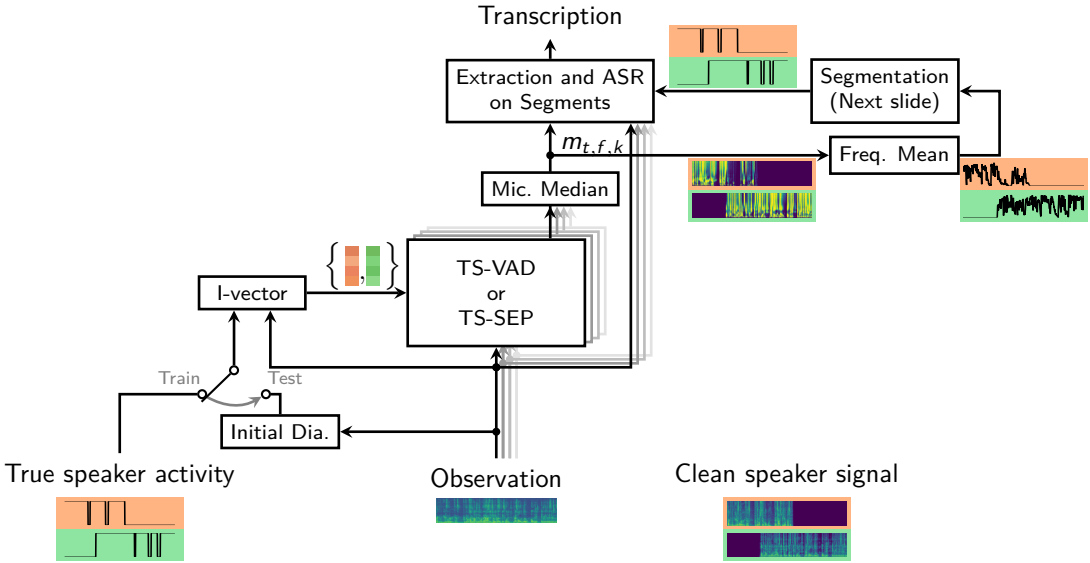
Training, Embedding and inference



Training, Embedding and inference

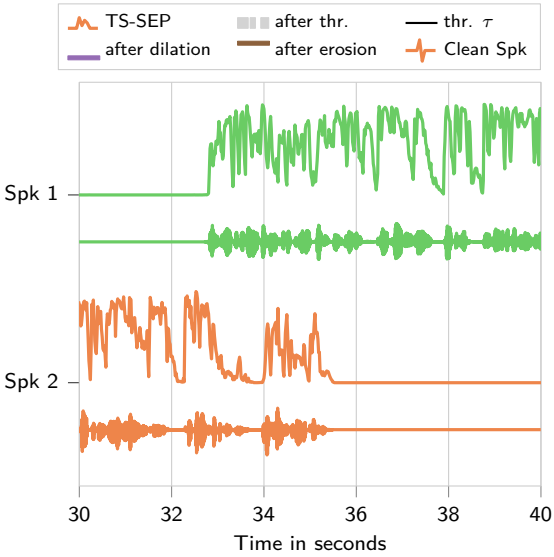


Training, Embedding and inference



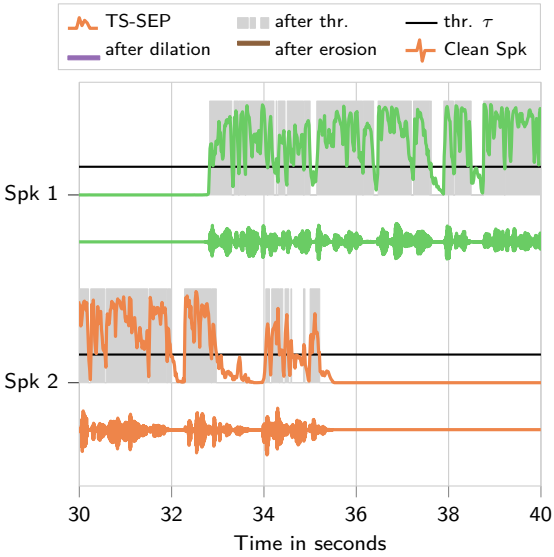
Activity estimation / Segmentation

- Simple thresholding not enough
- Closing (morphological operation)
  - Close gaps with dilation $\rightarrow$ erosion
- Overestimate activity (dilation $>$ erosion)
  - More likely include start and end, matches ASR training



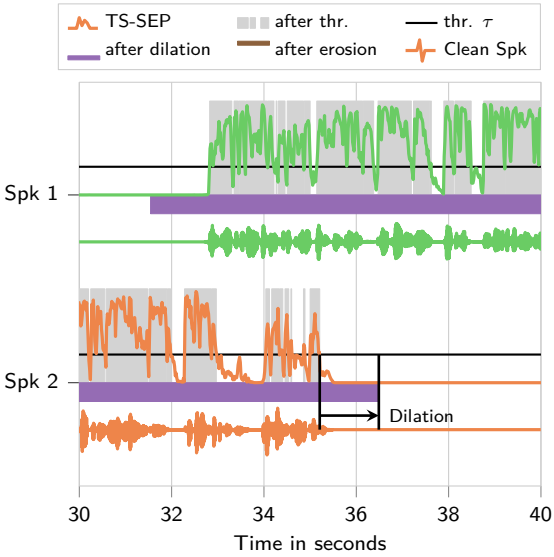
Activity estimation / Segmentation

- Simple thresholding not enough
- Closing (morphological operation)
  - Close gaps with dilation $\rightarrow$ erosion
- Overestimate activity (dilation $>$ erosion)
  - More likely include start and end, matches ASR training



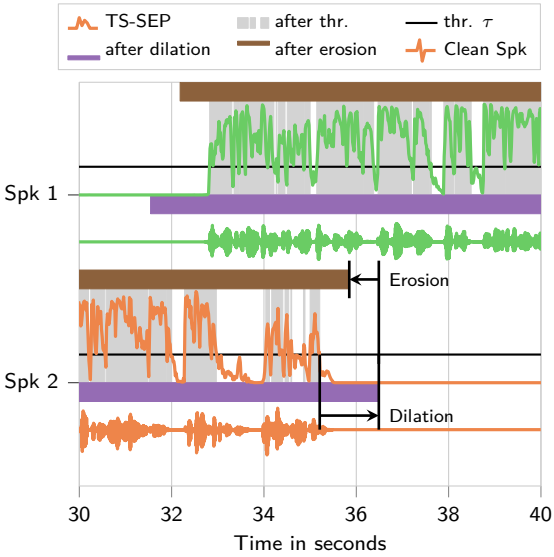
Activity estimation / Segmentation

- Simple thresholding not enough
- Closing (morphological operation)
  - Close gaps with dilation → erosion
- Overestimate activity (dilation > erosion)
  - More likely include start and end, matches ASR training



Activity estimation / Segmentation

- Simple thresholding not enough
- Closing (morphological operation)
  - Close gaps with dilation → erosion
- Overestimate activity (dilation > erosion)
  - More likely include start and end, matches ASR training

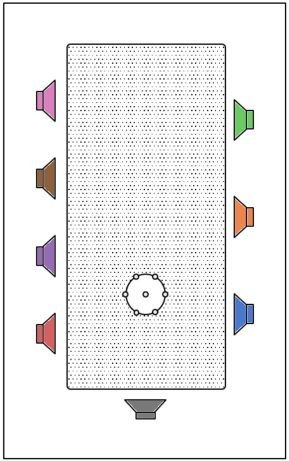


Extraction

- Details and experiments in paper
- Masking
- Mask based Beamforming (MVDR) followed by Thresholded Masking
- Guided Source Separation [7]
  - Diarization as guide to estimate masks with spatial mixture model (SMM)
  - TS-SEP enables initialization with tf-masks

[7] C. Boeddeker et al, "Front-end processing for the CHiME-5 dinner party scenario", CHiME 2018

Data: LibriCSS



Source: [3]

- Test: LibriCSS [3]
  - Meeting scenario
  - Re-recordings of LibriSpeech sentences
  - 60 recordings, 10 min each
  - Overlap ratio: 0 % (short/long pauses), 10 %, 20 %, 30 %, 40 %
- Train: Simulated meeting from LibriCSS authors ([1])
  - Simulated instead of rerecorded (reverberated with image method)
- Metric: Concatenated minimum-Permutation Word Error Rate (cpWER)

[3] Z. Chen et al, "Continuous Speech Separation: Dataset and Analysis", ICASSP 2020

[1] github.com/jsalt2020-asrdiar/jsalt2020_simulate

Experiments

Ref.	System	Comment	Single Ch.	cpWER
[6]	Spatial Mixture Model	SotA	-	5.9
[12]	SC ➔ TS-VAD ➔ GSS	Best TS-VAD	-	11.2
	SC ➔ TS-VAD ➔ GSS	Reimplementation	-	5.77
	SC ➔ TS-SEP ➔ GSS	Proposed	-	5.36
[61]	Transcribe-to-Diarize	Single Ch. SotA	✓	11.6
[12]	SC ➔ TS-VAD ➔ Speakerbeam	Best Single Ch. TS-VAD	✓	18.8
	SC ➔ TS-SEP ➔ Masking	Proposed	✓	7.81

[6] C. Boeddeker et al, "An Initialization Scheme for Meeting Separation with Spatial Mixture Models", Interspeech 2022
[12] M. Delcroix et al, "Speaker activity driven neural speech extraction", ICASSP 2021
[61] N. Kanda et al, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR", ICASSP 2022

Experiments

Ref.	System	Comment	Single Ch.	cpWER
[6]	Spatial Mixture Model	SotA	-	5.9
[12]	SC ➔ TS-VAD ➔ GSS	Best TS-VAD	-	11.2
	SC ➔ TS-VAD ➔ GSS	Reimplementation	-	5.77
	SC ➔ TS-SEP ➔ GSS	Proposed	-	5.36
[61]	Transcribe-to-Diarize	Single Ch. SotA	✓	11.6
[12]	SC ➔ TS-VAD ➔ Speakerbeam	Best Single Ch. TS-VAD	✓	18.8
	SC ➔ TS-SEP ➔ Masking	Proposed	✓	7.81

[6] C. Boeddeker et al, "An Initialization Scheme for Meeting Separation with Spatial Mixture Models", Interspeech 2022
[12] M. Delcroix et al, "Speaker activity driven neural speech extraction", ICASSP 2021
[61] N. Kanda et al, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR", ICASSP 2022

Experiments

Ref.	System	Comment	Single Ch.	cpWER
[6]	Spatial Mixture Model	SotA	-	5.9
[12]	SC ➔ TS-VAD ➔ GSS	Best TS-VAD	-	11.2
	SC ➔ TS-VAD ➔ GSS	Reimplementation	-	5.77
	SC ➔ TS-SEP ➔ GSS	Proposed	-	5.36
[61]	Transcribe-to-Diarize	Single Ch. SotA	✓	11.6
[12]	SC ➔ TS-VAD ➔ Speakerbeam	Best Single Ch. TS-VAD	✓	18.8
	SC ➔ TS-SEP ➔ Masking	Proposed	✓	7.81

[6] C. Boeddeker et al, "An Initialization Scheme for Meeting Separation with Spatial Mixture Models", Interspeech 2022
[12] M. Delcroix et al, "Speaker activity driven neural speech extraction", ICASSP 2021
[61] N. Kanda et al, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR", ICASSP 2022

Experiments

Ref.	System	Comment	Single Ch.	cpWER
[6]	Spatial Mixture Model	SotA	-	5.9
[12]	SC ➔ TS-VAD ➔ GSS	Best TS-VAD	-	11.2
	SC ➔ TS-VAD ➔ GSS	Reimplementation	-	5.77
	SC ➔ TS-SEP ➔ GSS	Proposed	-	5.36
[61]	Transcribe-to-Diarize	Single Ch. SotA	✓	11.6
[12]	SC ➔ TS-VAD ➔ Speakerbeam	Best Single Ch. TS-VAD	✓	18.8
	SC ➔ TS-SEP ➔ Masking	Proposed	✓	7.81

[6] C. Boeddeker et al, "An Initialization Scheme for Meeting Separation with Spatial Mixture Models", Interspeech 2022
[12] M. Delcroix et al, "Speaker activity driven neural speech extraction", ICASSP 2021
[61] N. Kanda et al, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR", ICASSP 2022

Experiments

Ref.	System	Comment	Single Ch.	cpWER
[6]	Spatial Mixture Model	SotA	-	5.9
[12]	SC ➔ TS-VAD ➔ GSS	Best TS-VAD	-	11.2
	SC ➔ TS-VAD ➔ GSS	Reimplementation	-	5.77
	SC ➔ TS-SEP ➔ GSS	Proposed	-	5.36
[61]	Transcribe-to-Diarize	Single Ch. SotA	✓	11.6
[12]	SC ➔ TS-VAD ➔ Speakerbeam	Best Single Ch. TS-VAD	✓	18.8
	SC ➔ TS-SEP ➔ Masking	Proposed	✓	7.81

[6] C. Boeddeker et al, "An Initialization Scheme for Meeting Separation with Spatial Mixture Models", Interspeech 2022
[12] M. Delcroix et al, "Speaker activity driven neural speech extraction", ICASSP 2021
[61] N. Kanda et al, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR", ICASSP 2022

Experiments

Ref.	System	Comment	Single Ch.	cpWER
[6]	Spatial Mixture Model	SotA	-	5.9
[12]	SC ➔ TS-VAD ➔ GSS	Best TS-VAD	-	11.2
	SC ➔ TS-VAD ➔ GSS	Reimplementation	-	5.77
	SC ➔ TS-SEP ➔ GSS	Proposed	-	5.36
[61]	Transcribe-to-Diarize	Single Ch. SotA	✓	11.6
[12]	SC ➔ TS-VAD ➔ Speakerbeam	Best Single Ch. TS-VAD	✓	18.8
	SC ➔ TS-SEP ➔ Masking	Proposed	✓	7.81
Since publication:				
[Niu25]	DCF-DS	Single Ch. SotA	✓	4.43
[Boe24]	SC ➔ TS-SEP ➔ GSS ➔ SLR	Follow-up work	-	3.45
[Tah24]	SSND	SotA	-	3.22

[Tah24] H. Taherian et al, "Multi-channel conversational speaker separation via neural diarization", TASLP 2024
[Boe24] C. Boeddeker et al, "Once more Diarization: Improving meeting transcription systems through segment-level speaker reassignment", Interspeech 2024
[Niu25] S. Niu et al, "DCF-DS: Deep Cascade Fusion of Diarization and Separation for Speech Recognition under Realistic Single-Channel Conditions", preprint

Highlight

- TS-SEP: Extension of TS-VAD from diarization to joint diarization and separation
- SotA on LibriCSS
  - Single channel: 7.81 % vs. 11.6 % cpWER
  - Multi channel: 5.36 % vs. 5.9 % cpWER
- Extensive experimental evaluation in paper
  - Activity estimation / Segmentation analysis
    - Trade-off between DER and WER
    - WER prefers "overestimation"
  - Impact of ignoring multichannel information on system components
  - Comparisons of extractions: masking, beamforming, GSS
  - TS-SEP enables faster GSS, with minimal performance impact
  - Literature comparison
- Open Source PyTorch implementation of TS-VAD and TS-SEP
  - `https://github.com/merlresearch/tssep`
  - Improved TS-VAD: 5.7 % vs. 11.2 % cpWER

Thank you for your attention!

Questions?
boeddeker@nt.upb.de

Highlight

- TS-SEP: Extension of TS-VAD from diarization to joint diarization and separation
- SotA on LibriCSS
  - Single channel: 7.81 % vs. 11.6 % cpWER
  - Multi channel: 5.36 % vs. 5.9 % cpWER
- Extensive experimental evaluation in paper
  - Activity estimation / Segmentation analysis
    - Trade-off between DER and WER
    - WER prefers "overestimation"
  - Impact of ignoring multichannel information on system components
  - Comparisons of extractions: masking, beamforming, GSS
  - TS-SEP enables faster GSS, with minimal performance impact
  - Literature comparison
- Open Source PyTorch implementation of TS-VAD and TS-SEP
  - `https://github.com/merlresearch/tssep`
  - Improved TS-VAD: 5.7 % vs. 11.2 % cpWER

Thank you for your attention!

Questions?
boeddeker@nt.upb.de