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], Jonathan Le Roux[2]

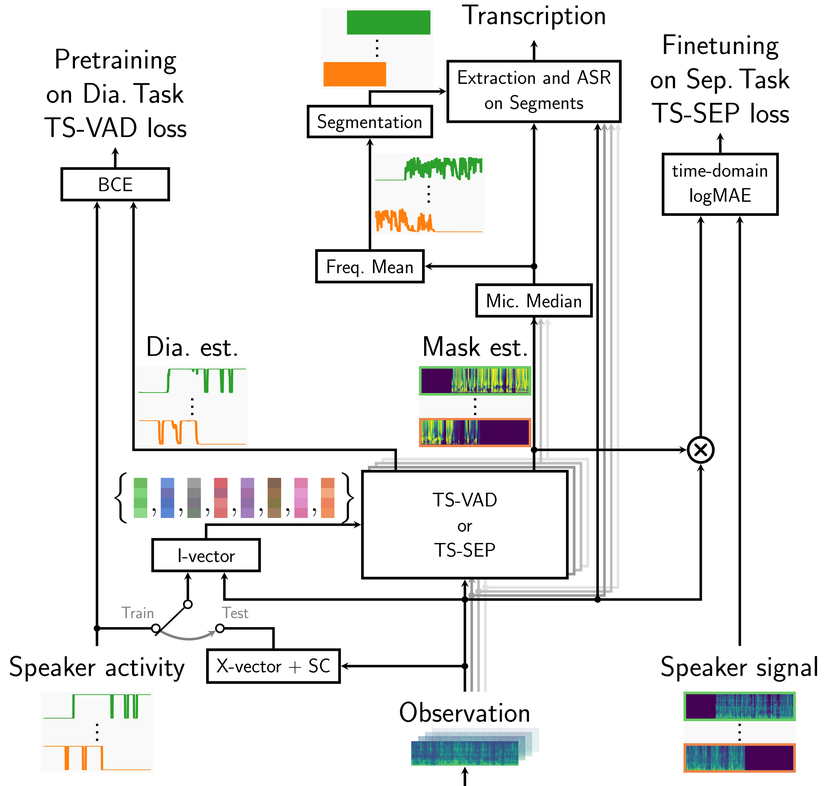
[1]Paderborn University, Germany, [2]Mitsubishi Electric Research Laboratories (MERL), USA

IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING

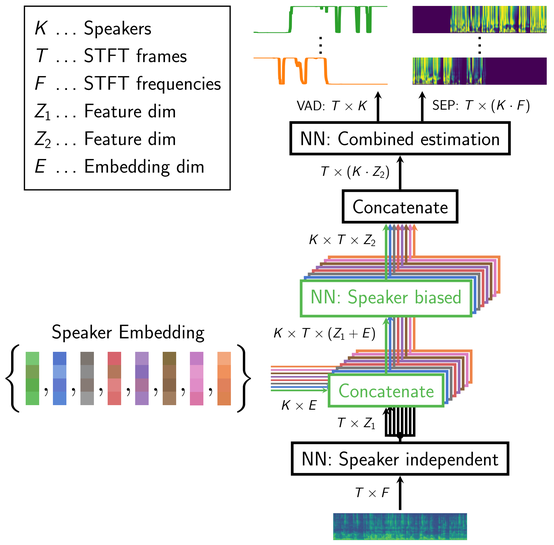
Highlights

- TS-SEP: Extension of TS-VAD from diarization to joint diarization and separation
- Extensive experimental evaluation
  - Segmentation analysis
    - Trade-off between DER and WER
    - WER prefers "overestimation"
  - Impact of multichannel information on system components
  - Comparison of extraction techniques: masking, beamforming, GSS
  - TS-SEP enables faster GSS
- SotA on LibriCSS
  - Single channel: ~~11.6 %~~ → 7.81 % cpWER
  - Multi channel: ~~5.9 %~~ → 5.36 % cpWER
  - TS-VAD: ~~11.2 %~~ → 5.77 % cpWER
- Open Source PyTorch implementation of TS-VAD and TS-SEP

System overview



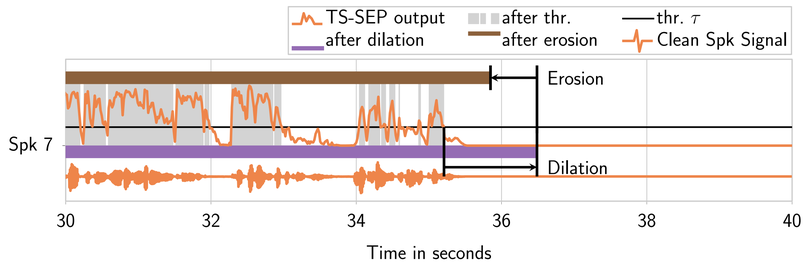
TS-VAD and TS-SEP structure



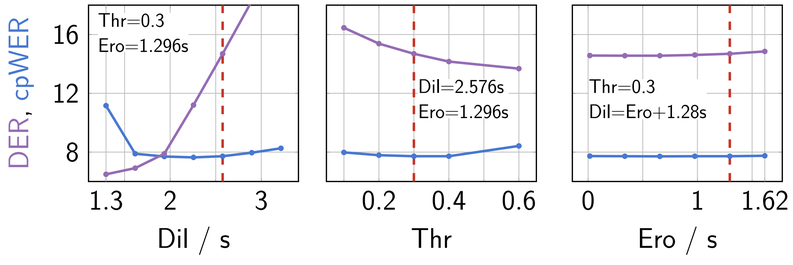
Datasets

- Train
  - jsalt2020_simulate
  - Simulated meeting data
  - From LibriCSS authors
- Test
  - LibriCSS
  - Meeting scenario
  - Re-recordings of LibriSpeech sentences
  - 60 recordings, 10 min each
  - Overlap ratios: 0 % (short/long pauses), 10 %, 20 %, 30 %, 40 %

Segmentation (Threshold, Dilation, Erosion)

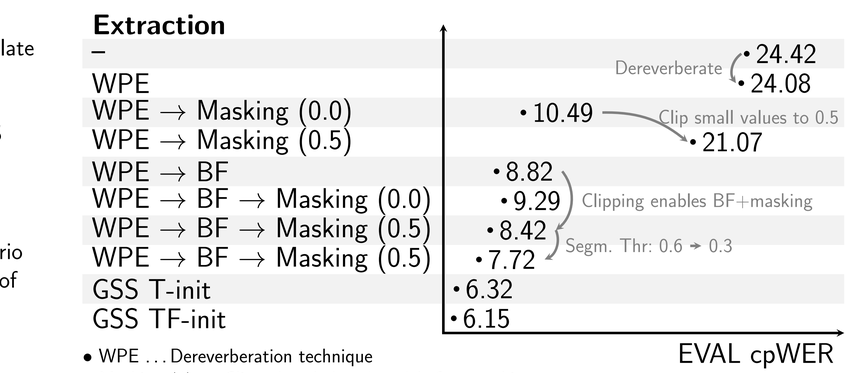


Impact of Segmentation



- cpWER ... concatenated minimum-permutation Word Error Rate
- DER ... Diarization Error Rate

Effect of Extraction techniques



- WPE ... Dereverberation technique
- Masking (ξ) ... Clip values below ξ to ξ before masking
- BF ... Beamforming (MVDR)
- GSS ... Guided Source Separation (includes WPE)
- T-init ... Classical GSS diarization init
- TF-init ... Init GSS with TF mask

Literature comparison (LibriCSS)

System (other)	Style	Single Ch.	cpWER
CSS with DOA Dia [58]	S → D → A	–	12.98
CSS with DOA Dia [58]	S → D → A	–	12.40 †
SMM [6]	D + S → A	–	5.9 †
CSS → SC [2]	S → D → A	–	12.7
Transcribe-to-Diarize [61]	E2E	✓	11.6
TS-VAD → Speakerbeam [12]	D → S → A	✓	18.8
TS-VAD → GSS [12]	D → S → A	–	11.2
SC → GSS [62]	D → S → A	–	12.12
System (mixed)			
SC [62] → GSS → WavLM	D → S → A	–	13.85
System (our)			
TS-VAD → GSS → Base	D → S → A	–	6.70
TS-VAD → WavLM	D → A	✓	9.26
TS-VAD → GSS → WavLM	D → S → A	–	**5.77**
TS-SEP → Mask. → Base	D + S → A	✓	11.61
TS-SEP → GSS → Base	D + S → A	–	6.42
TS-SEP → Mask. → WavLM	D + S → A	✓	**7.81**
TS-SEP → GSS → WavLM	D + S → A	–	**5.36**

[2] D. Raj et al., "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis", SLT 2021.
[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.
[58] Z.-Q. Wang and D. Wang et al., "Localization based sequential grouping for continuous speech separation,", *ICASSP*.
[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.
[62] D. Raj et al., "GPU-accelerated guided source separation for meeting transcription,", Interspeech 2023.

Work supported by DFG under project no. 448568305.