STCON System for the CHiME-8 Challenge

Anton Mitrofanov, Tatiana Prisyach, Tatiana Timofeeva, Sergei Novoselov, Maxim Korenevsky, Yuri Khokhlov, Artem Akulov, Alexander Anikin, Roman Khalili, Iurii Lezhenin, Aleksandr Melnikov, Dmitriy Miroshnichenko, Nikita Mamaev, Ilya Odegov, Olga Rudnitskaya, Aleksei Romanenko

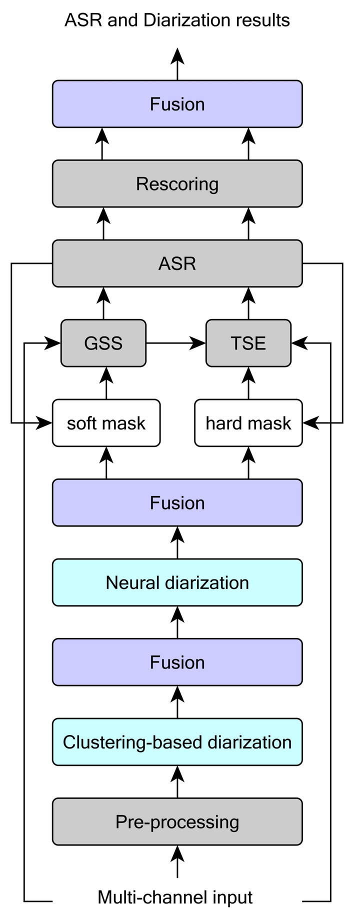
[1]STCON LLC., Kingdom of Saudi Arabia

Introduction

- Goal: to build accurate diarization and ASR system for multichannel conversations
- Data:
  - Four datasets: CHiME-6, DiPCo, Mixer 6 Speech, NOTSOFAR1
  - Different settings: dinner party, interview, office meeting
  - Different number of speakers (2–8) and microphones (7–35)
  - Very different session duration (from 6 min to over 2 hours)
- Main focus: generalization of a solution to all above factors of variability
- Metrics: time-constrained minimum-permutation WER (tcpWER, main), DER (auxiliary)

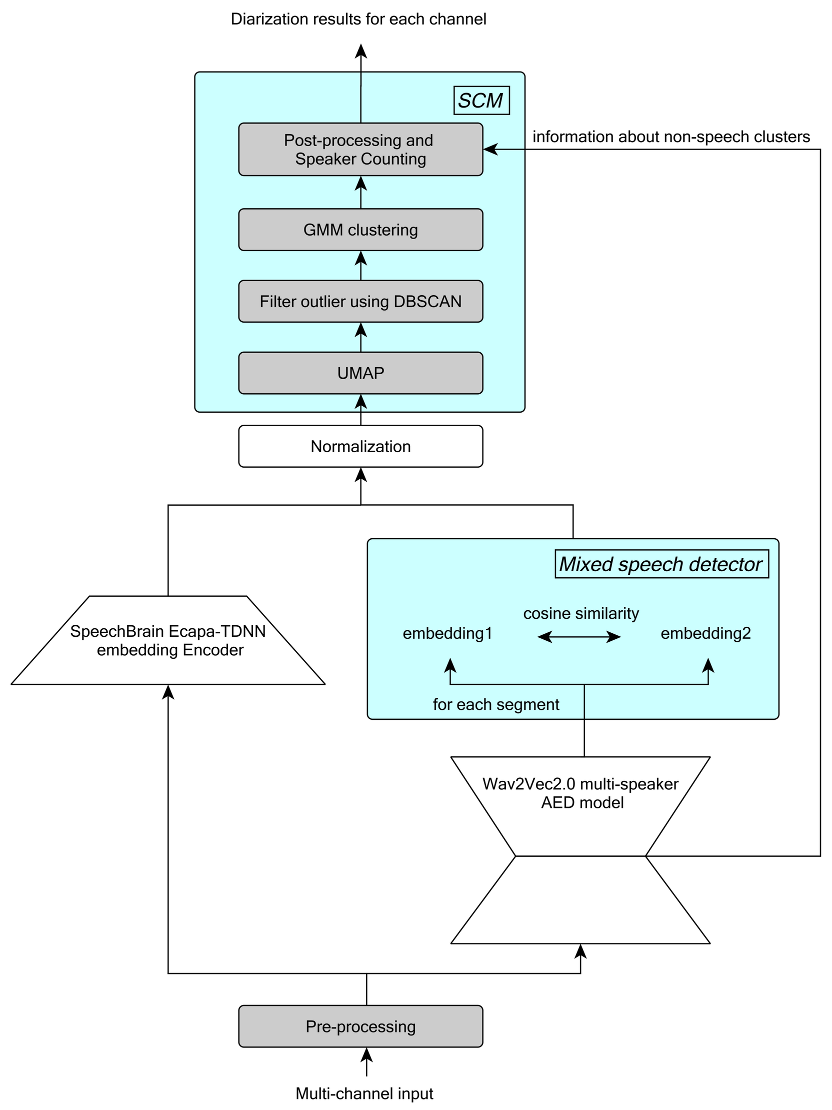
Pipeline Overview

- The pipeline follows the standard paradigm: Diarization - Source Separation - ASR
- It starts with **preprocessing** module which performs block WPE-dereverberation, suppressing knocks and clicks, volume normalization as well as channel selection with MicRank followed by Voice Activity Detection
- **Diarization** block consists of two stages, namely Clustering-based and Neural, both applied channel-wise and followed by DOVERLap fusion
- **Source Separation** block consists of two modules, namely Guided Source Separation followed by MVDR beamforming, and Target Speaker Extraction.
- Results of source separation are fed into multiple **ASR** models. ASR results can be used to update masks for source separation
- ASR results are optionally **re-scored** with Large LM and **fused** together to obtain the final system output



Clustering-based diarization (CBD)

- **Goals**:
  - to determine the correct number of speaker for each channel
  - to prepare initial segmentation for Neural Diarization
- **Mixed speech detector**: AED-model based on Wav2Vec2.0 XLS-R53
  - returns multiple speaker embeddings per chunk
  - clusterizes embeddings to detect overlapped speech
  - determines non-speech frames
- **SpeechBrain Ecapa-TDNN** speaker embeddings extractor
- **Speaker Counting Module (SCM)**
  - UMAP projection to low dimensionality (12)
  - DBSCAN clustering for outliers filtering
  - GMM-based clustering
  - Post-processing to remove non-speech clusters



- #speakers in session is determined by majority voting across session's channels
- diarization results from 12 different settings are **DOVERLap**-ed for each channel

Table 1. Clustering-based diarization results on devsets.

System	max_spk	DER / speaker count accuracy				
		chime6	dipco	mixer6	notsofar1	Avg
baseline	4	26.8	24.78	16.53	-	-
	8	36	26	24	-	-
single_orig*	8	25.3/0.87	23.7/1	16.3/0.91	20.0/0.86	20.6/0.88
single_wpe*	8	24.1/1	22.4/1	12.8/0.97	20.8/0.85	19.4/0.87
fusion	8	23.5/1	21.4/1	13.0/0.98	13.0/0.89	17.9/0.90

* The best of 6 systems with different parameters thr and VAD segments.

Neural diarization (ND)

- **Goal**: to improve diarization based on initial segmentation and estimated number of speakers
- **Approach**: using NSD-MS2S [1] model from the winner of CHiME-7
- **Synthetic dataset generation**:
  - using RIR classifier [2] to select RIRs similar to those in challenge data
  - generation of multichannel RIRs in selected room configurations
  - selection of background noises from challenge data
  - generation of multichannel reverberated and noisy conversations according to statistics of overlapping durations
- **Pretraining** of NSD-MS2S model on synthetic data
- **Fine-tuning** of NSD-MS2S model on challenge data with several modifications/filtering
- Diarization results from 8 different settings are **DOVERLap**-ed for each channel and then across channels

Table 2. Neural diarization results.

System	Data type	DER							AVG
		chime6		dipco		mixer6		notsofar1	
		dev	eval	dev	eval	dev	eval	dev	
CBD fusion	orig&wpe	23.5	29.6	21.4	17.3	13.0	7.5	13.0	17.9
Best single ND finetune	wpe	11.7	15.2	13.3	10.2	7.4	4.4	8.1	10.0
ND fusion	orig&wpe	10.8	14.8	13.8	10.0	7.1	4.3	7.9	9.8

Source Separation

- **Basic approach: Guided Source Separation (GSS)** with GPU acceleration [3]
  - Using soft weights from ND improves ASR accuracy and reduces the number of GSS EM-iterations
  - The second pass of the GSS uses the same soft weights, but multiplies them by hard VAD masks based on the recognition results from the first pass
- **Alternative approach: Guided Target Speaker Extraction (G-TSE)**
  - Two multichannel architectures were used: SpatialNet [4] and TF-GridNet [5]

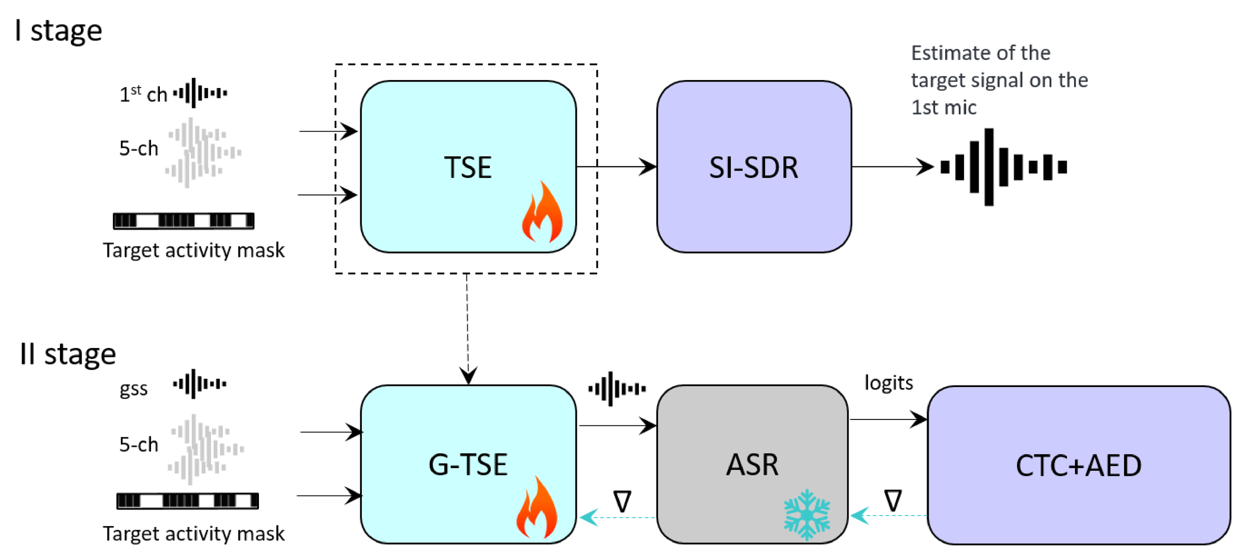
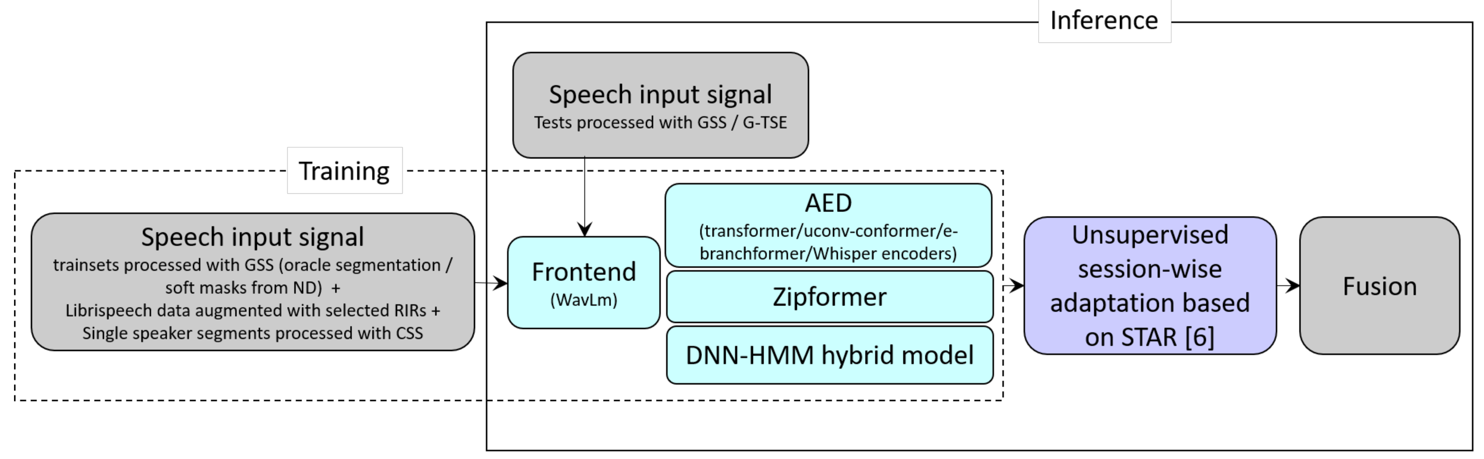


Table 3. Dev/eval tcpWER comparison of GSS and G-TSE results.

system	chime	dipco	mixer6	notsofar1
2-pass GSS	25.9/37.1	32.0/22.6	11.8/13.8	21.4/-
G-TSE	25.5/36.7	32.1/22.7	11.6/13.5	21.2/-

ASR

- The set of **ASR models** is mainly the same as in CHiME-7



Rescoring and fusion

- In the Unconstrained LM track the N-best rescoring was applied to the numerous recognition results from different version of Source Separation and ASR models
- Model for rescoring: finetuned non-istructive Llama2-7B
- Data for finetuning: texts from CHiME-8 training data and Librispeech
- Rescored/original N-best lists were converted to the lattices and lattice fusion was applied to the set of results selected based on average tcpWER over devsets

Results and conclusions

The **results** of our system on CHiME-8 DASR Task are presented in the table:

dev tcpWER,%					eval tcpWER,%				
chime6	dipco	mixer6	notsofar1	Avg	chime6	dipco	mixer6	notsofar1	Avg
Constrained LM track									
22.8	29.0	10.1	19.1	20.2	33.6	20.2	11.0	14.8	19.9
Unconstrained LM track									
22.5	28.4	9.8	18.7	19.9	33.1	19.9	10.9	14.6	19.6

References

[1] https://github.com/liyunlongaaa/NSD-MS2S.

[2] Y. Khokhlov, T. Prisyach, A. Mitrofanov, D. Dutov, I. Agafonov, T. Timofeeva, A. Romanenko, and M. Korenevsky, "Classification of room impulse responses and its application for channel verification and diarization," in *INTERSPEECH*, 2024, p. to appear.

[3] D. Raj, D. Povey, and S. Khudanpur, "Gpu-accelerated guided source separation for meeting transcription," *arXiv:2212.05271*, 2022.

[4] C. Quan and X. Li, "Spatialnet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *arXiv:2307.16516*, 2023.

[5] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, pp. 1–15, 01 2023.

[6] Y. Hu, C. Chen, C.-H. H. Yang, C. Qin, P.-Y. Chen, E. S. Chng, and C. Zhang, "Self-taught recognizer: Toward unsupervised adaptation for speech foundation models," *arXiv:2405.14161*, 2024.