

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

STCON LLC., Kingdom of Saudi Arabia

STCON system

ASR and Diarization results

Fusion

Rescoring

ASR

GSS

G-TSE

soft mask

hard mask

Fusion

Neural diarization

Fusion

Clustering-based diarization

Pre-processing

Multi-channel input

Input: Nbest
Output: Rescored hyps

Input: Enhanced audio
Output: Nbest

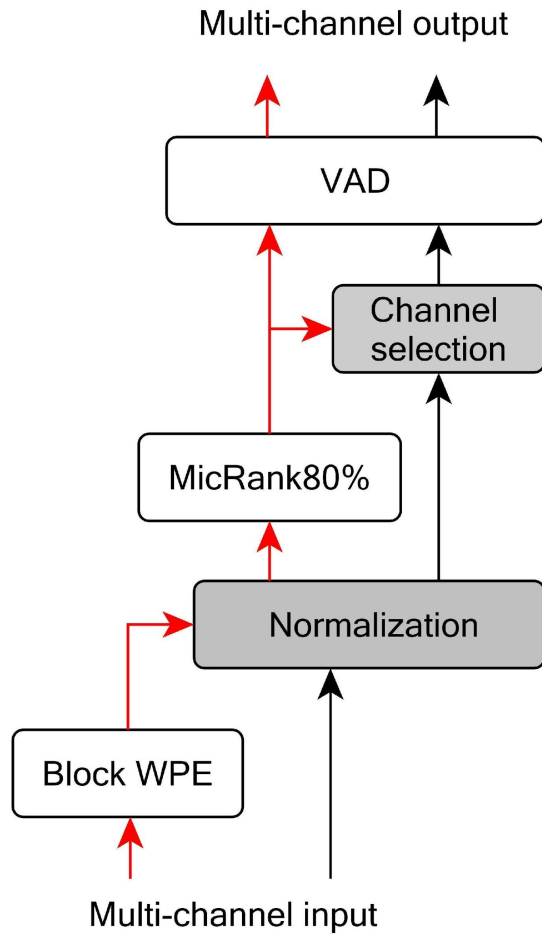
Input: Diarization RTTM, mask,
Multi-channel input
Output: Enhanced audio

Input: RTTM from clustering-based
diarization, pre-processed audio
Output: Diarization RTTM, mask

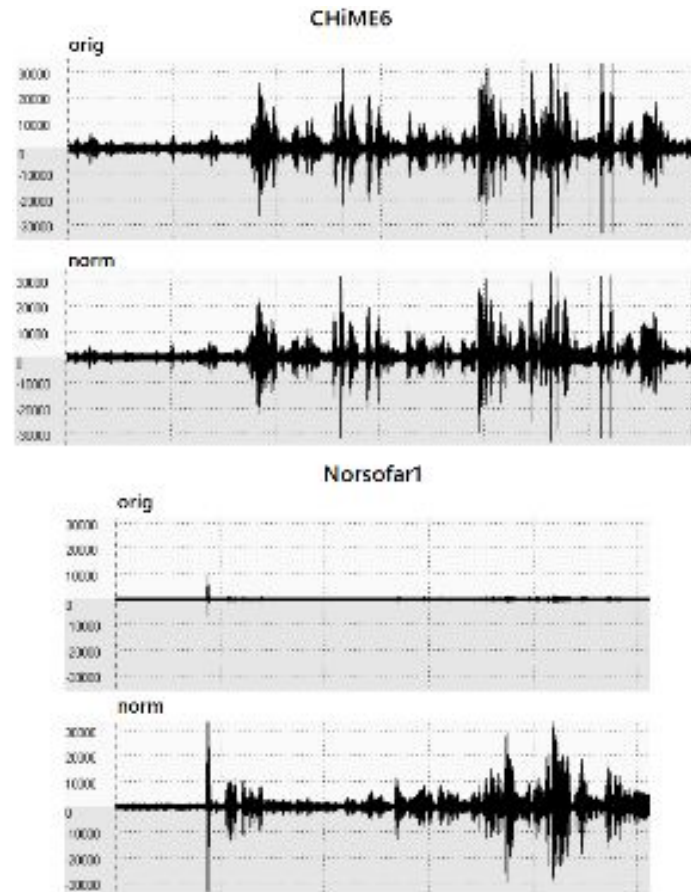
WPE + normalization +
Micrank80 + VAD

Input: VAD segments, pre-processed audio
Output: Diarization RTTM

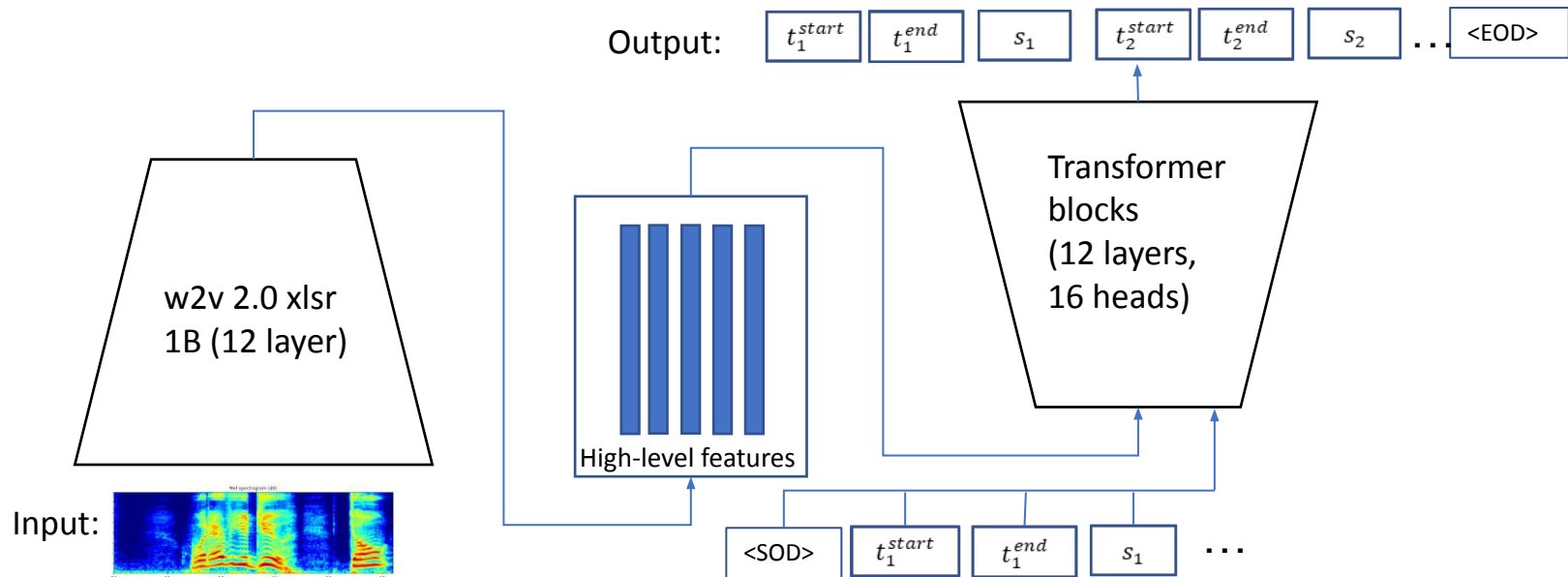
Pre-processing



Normalization:



AED-based multiple speaker extractor



overlapped speech processing

$\langle SOD \rangle$ start-of-decoding token

$\langle EOD \rangle$ end-of-decoding token

t_i^{end} - end timestamp token for i-th speaker

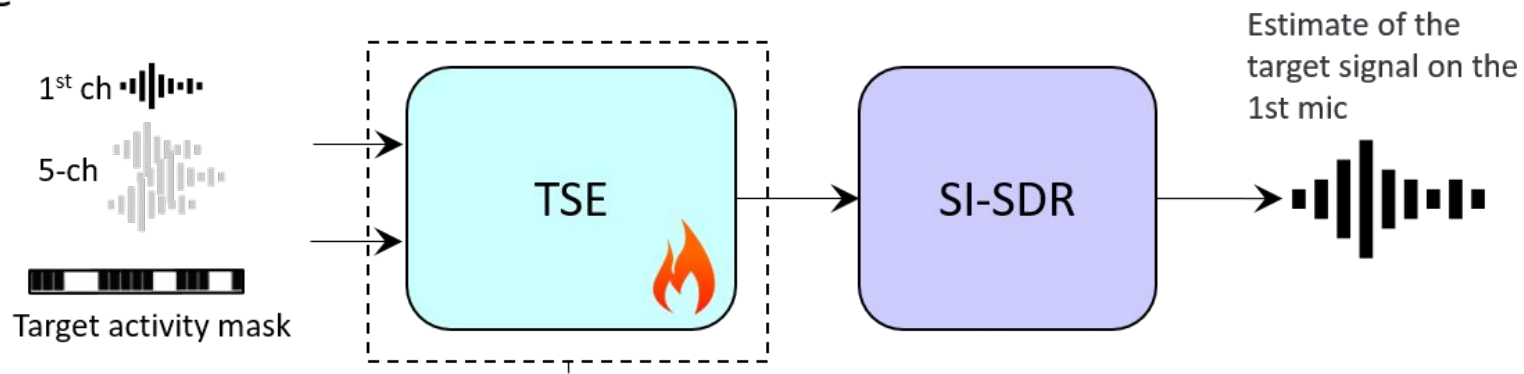
t_i^{start} - start timestamp token for i-th speaker

s_i - i-th speaker label token

$i = [1:3]$

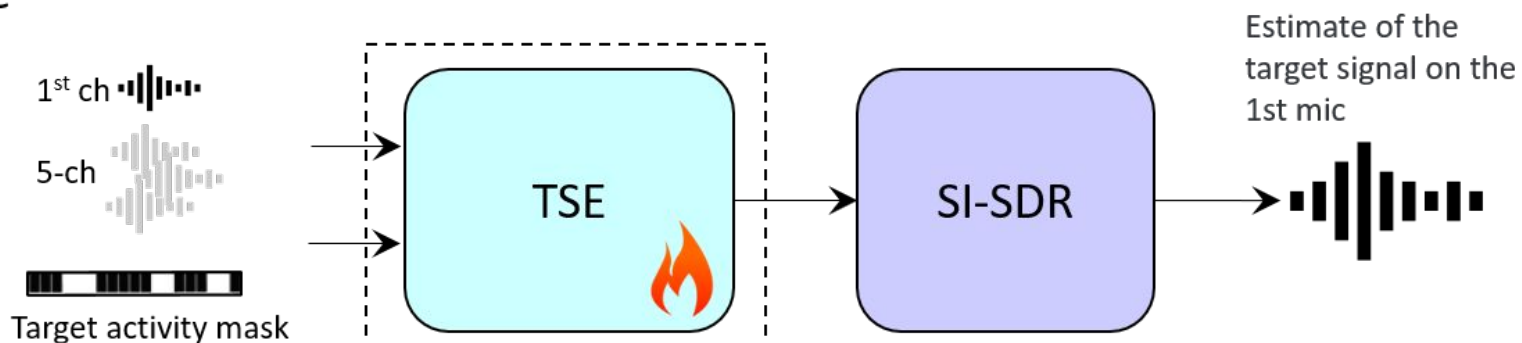
Target Speaker Extraction (TSE)

I stage

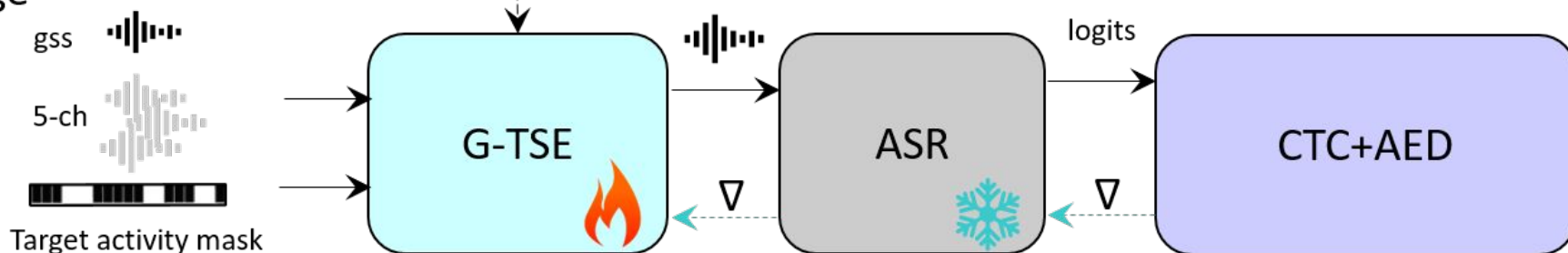


TSE+ASR joint training

I stage



II stage



Thank you for your attention!