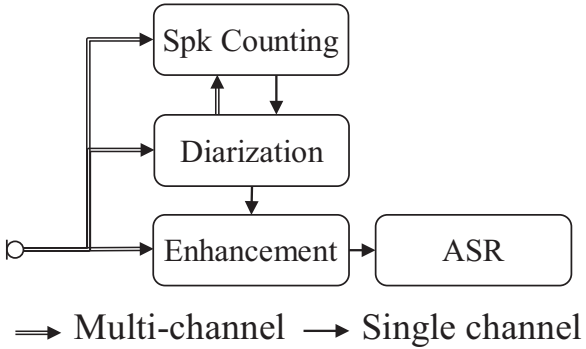
# NTT Multi-Speaker ASR System for the DASR Task of CHiME-8 Challenge

Naoyuki Kamo*, Naohiro Tawara*, Atsushi Ando, Takatomo Kano, Hiroshi Sato, Rintaro Ikeshita, Takafumi Moriya, Shota Horiguchi, Kohei Matsuura, Atsunori Ogawa, Alexis Plaquet, Takanori Ashihara, Tsubasa Ochiai, Masato Mimura, Marc Delcroix, Tomohiro Nakatani, Taichi Asami, Shoko Araki

# Overall pipeline

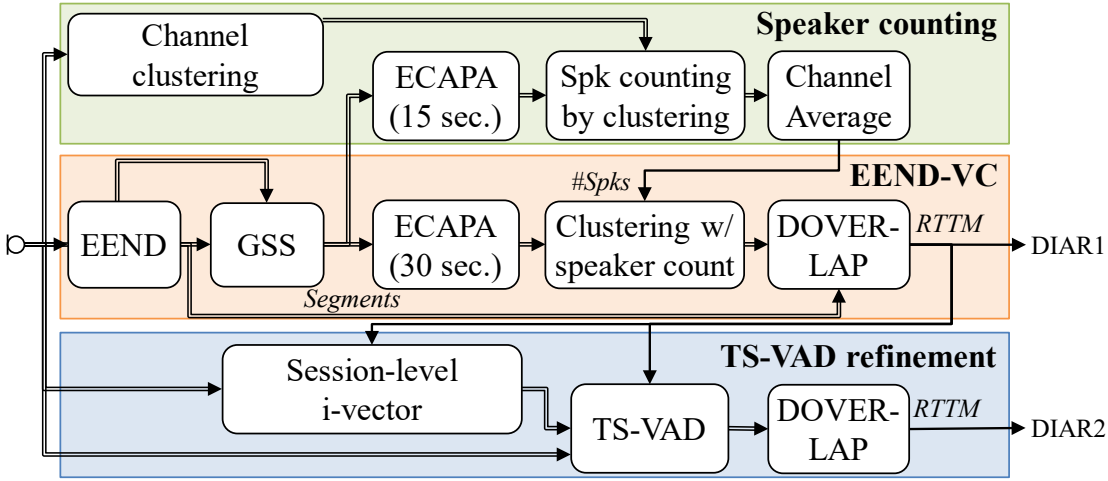



Similar pipeline to baseline:

1) Diarization
   - E2E neural diarization with vector clustering(EEND-VC) and with TS-VAD refinement
2) Multi channel speech enhancement
   - Guided source separation(GSS) in official baseline with several modifications
3) ASR
   - Four ASR models combination with LM rescoring

We achieve 22.0 % tcpWER for task1 and 59 % of relative improvement over the NEMO baseline, and 16.8 % tcpWER for NOTSOFAR1 scenario

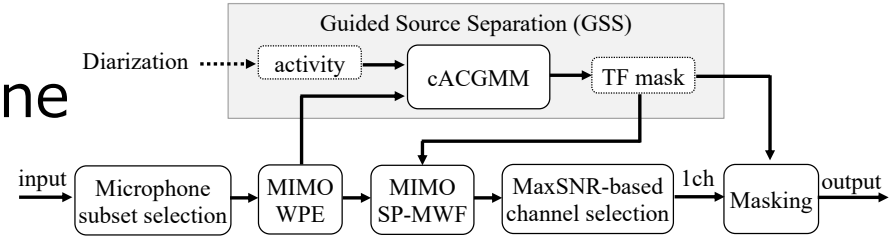
# Diarization




- Consists of three parts
  - Segmentation by EEND-VC
    - EEND diarization on 30 sec. chunks
      + GSS
      + speaker embedding clustering
  - Speaker counting by clustering
    - Different from EEND-VC, specific for counting
    - Clustering across microphone channels
  - TS-VAD (NSD-MS2S [G. Yang, 2024]) refinement
    - Same model as CHiME7 winner
- DOVER-LAP for combination of multi-channel results

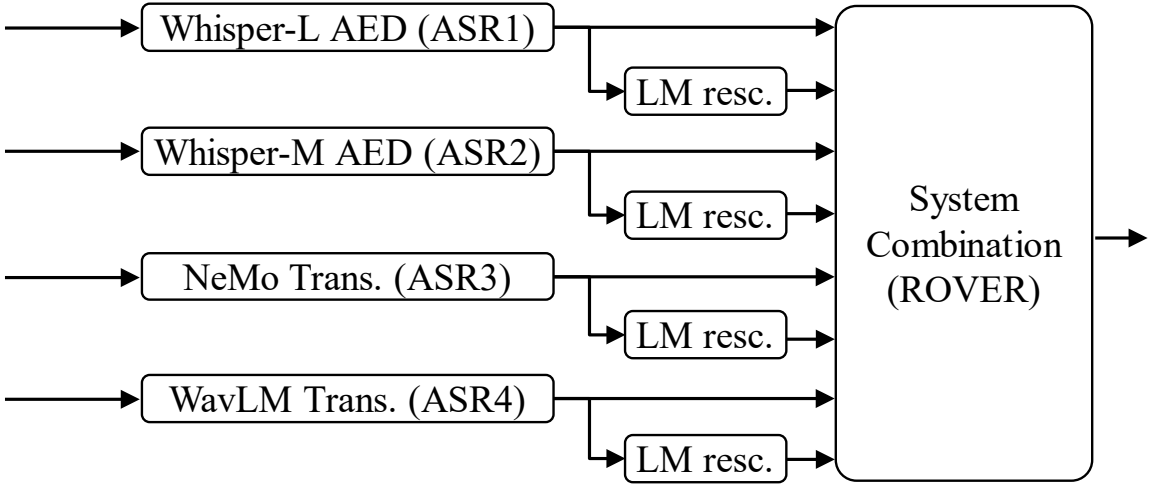
# Multi channel speech enhancement

- Based on GSS in Nemo baseline
- Several modifications




| | Nemo baseline | NTT system |
|---|---|---|
| Microphone subset selection | Top 80% mics based on EV (Envelop variance) | New rule based on EV and $C_{50}$ (Room acoustic estimation) |
| Beamformer | Souden MVDR | Spatial-prediction multichannel wiener filter (SP-MWF) [J. Benesty, 2008] |
| Post processing | Applying BAN(Blind Analytic normalization) & TF-mask | Without BAN & TF-mask |

# ASR




- Build four ASR models
  - Finetuning Whisper,  Nemo baseline, etc.
- Apply LM rescoring for each ASR models
- Using ROVER for system combination

# Results

We submitted three systems:

| ID | Diar | ASR | NOTSOFAR1 tcpWER | Macro tcpWER | RTF |
|---|---|---|---|---|---|
| **Baseline** | NeMo | - | 61.0 | 53.2 | - |
| **NTT-1** | DIA1 (EEND-VC) | ASR4 (Transducer) | 22.1 | 27.2 | 2.46 |
| **NTT-2** | DIA2 (EEND-VC + TS-VAD) | ASR1 (Whisper Large) | 18.3 | 24.3 | 3.14 |
| **NTT-3** | DIA2 (EEND-VC + TS-VAD) | ASR5 (ROVER) | 16.8 | 22.0 | 4.03 |

# Thank you for listening!

**Diarization:**

Naohiro Tawara, Atsushi Ando, Shota Horiguchi, Alexis Plaquet, Marc Delcroix

**Speech enhancement:**

Hiroshi Sato, Rintaro Ikeshita, Tsubasa Ochiai, Tomohiro Nakatani, Shoko Araki,

Naoyuki Kamo

**ASR:**

Takatomo Kano, Takafumi Moriya, Kohei Matsuura,
Atsunori Ogawa, Takanori Ashiharam, Masato Mimura, Taichi Asami