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

NTT Corporation, Japan

*Equal contribution

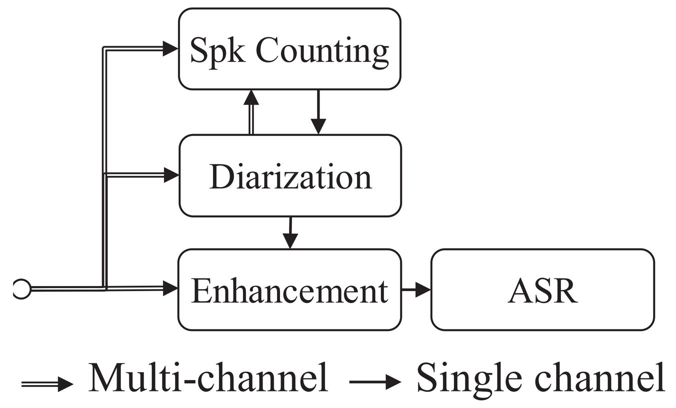
Abstract

Overall pipeline:

Propose a multi-channel multi-speaker DASR system extending NTT CHiME-7 task1 system

Following a pipeline similar to the CHiME-8 task1 baseline:

1) Diarization
2) Speech enhancement (SE) with guided source separation (GSS)
3) ASR



Main contributions:

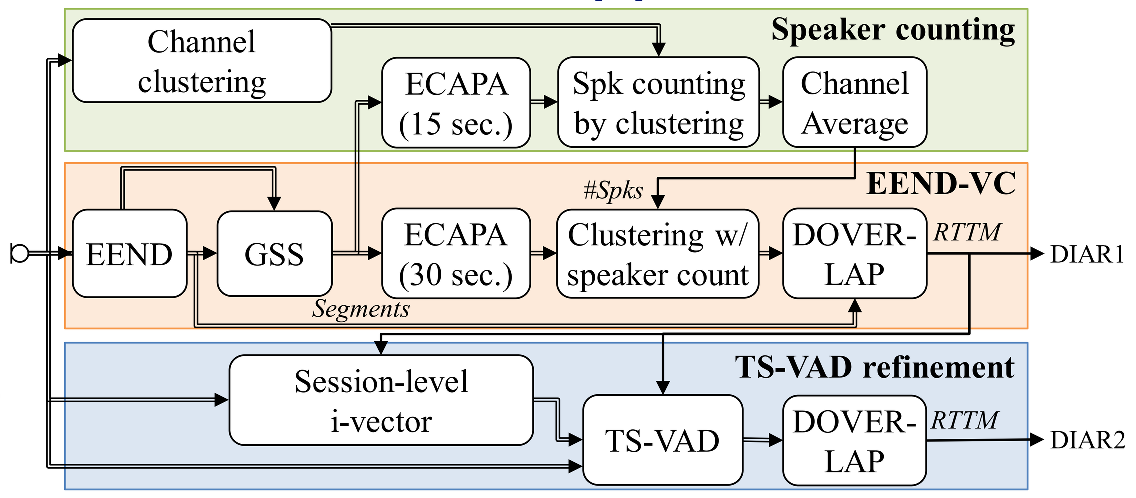
- **Diarization:**
  1. E2E neural diarization with vector clustering (EEND-VC)-based diarization and target-speaker voice activity detection (TS-VAD)-based refinement
  2. Novel multi-channel speaker counting approach
- **SE:** Modification on GSS: New rule for microphone subset selection and spatial-prediction multichannel Wiener filter (SP-MWF)
- **ASR:** Four strong ASR models and a Language Model (LM) for rescoring and system combination by ROVER
- **Overall:** Our system achieved 21.3% tcpWER on dev set and 57% relative reduction over the baseline system

Diarization

EEND-VC segmentation

- Perform EEND-VC to obtain chunk-level segmentation with speaker activity, setting a maximum of 4 speakers per chunk
- Extract speaker embeddings on each chunk with ECAPA-TDNN
- Three modifications from our EEND-VC system for CHiME7:

Diarization pipeline



- Apply GSS before embedding extraction to enhance speaker characteristic in embeddings
- Employ constrained spectral clustering instead of constraint AHC
- Reduce Chunk size from 80 to 30 sec to handle recordings with more than 4 speakers and high overlap.

Multi-microphone speaker counting

- Employ Normalized maximum eigengap spectral clustering (NMESC) for speaker counting
- Apply following multi-channel speaker counting to perform NMESC using more embeddings
  1. Find channel groups using AHC based on inter-channel correlations
  2. Extract ECAPA-TDNN-based speaker embeddings from GSS outputs in each channel group
  3. Perform NMESC-based speaker counting on each channel group
  4. Integrate the group-wise speaker counting results by averaging them

Detailed speaker counting setting:

- GSS is calculated over 30-sec segments using time-stamp obtained by EEND
- Speaker embeddings are extracted from 15-sec segments of GSS output
- Channel clustering is performed on a 0.3 correlation threshold over the first 120 sec signal.

TS-VAD refinement

- NSD-MS2S is applied to refine diarization results from EEND-VC.
- Same model configuration as CHiME-7 winner, but with stronger initial diarization by EEND-VC

Speaker counting accuracy [%] (↑) on the dev set.

	CH6	DiP	MX6	NSF	Macro
Baseline (NeMo)	50.0	0.0	100.0	13.8	41.0
Channel-wise counting	95.5	84.3	99.7	48.5	82.0
Microphone group-wise counting	100.0	90.0	100.0	57.5	86.9
+ Group averaging	100.0	100.0	100.0	58.2	89.6

DER [%] (↓) on dev set computed with md-eval with a collar of 0.25 sec.

ID	Model	CH6	DiP	MX6	NSF	Macro
DIA0	Baseline (NeMo)	45.65	45.92	25.16	38.05	38.70
DIA1	EEND-VC w/ ECAPA	28.52	24.38	9.69	10.67	18.32
DIA2	+ TS-VAD	23.97	21.01	6.11	9.72	15.20

DIA1 : EEND-VC + Multichannel Speaker Counting

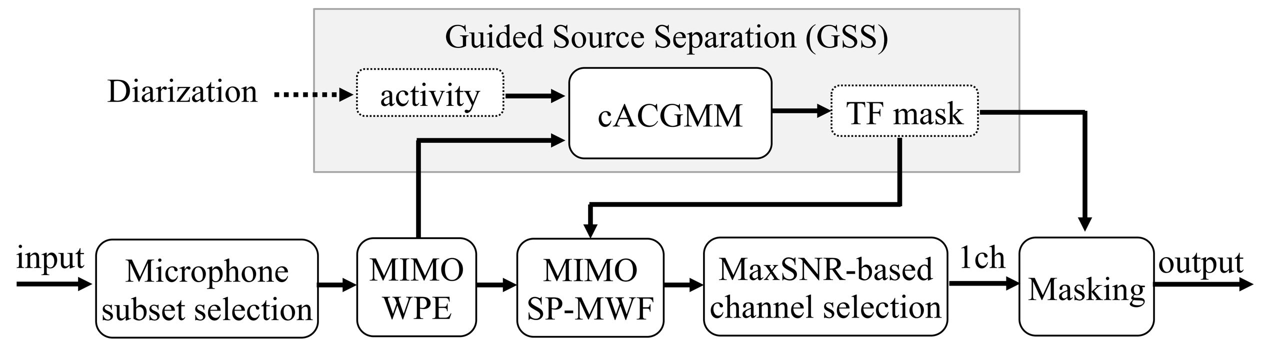
DIA2 : DIA1 + TS-VAD

Speech Enhancement (SE)

SE pipeline

Based on the official GSS in CHiME-8 NeMo baseline with modification:

- Microphone subset selection based on EV and C_{50}
- Using SP-MWF beamformer instead of MVDR beamformer
- Not applying blind analytic normalization (BAN) filter



Microphone subset selection

Using two features: the envelop variance (EV) and C_{50} (energy ratio of the early phase to the late phase of the room impulse response) estimated by Brouhaha toolkit.

Our selection rule:

- If the intersection of the top K microphones ranked by EV and C_{50} (K = 65%) contains at least 15 microphones, select them
- If fewer than 15 microphones intersect, but EV ranking has at least 15, select those
- If neither condition is met, select the top 15 microphones by EV
- Use all microphones if fewer than 15 are available

Spatial-prediction multichannel Wiener filter (SP-MWF)

- We replaced MVDR beamformer with SP-MWF

$$\mathbf{w}_f(r) = \frac{(\mathbf{e}_r^\top \mathbf{R}_{\mathbf{x},f}\mathbf{e}_r)\mathbf{R}_{\mathbf{n},f}^{-1}\mathbf{R}_{\mathbf{x},f}\mathbf{e}_r}{\mu \mathbf{e}_r^\top \mathbf{R}_{\mathbf{x},f}\mathbf{e}_r + \mathrm{Tr}(\mathbf{R}_{\mathbf{n},f}^{-1}\mathbf{R}_{\mathbf{x},f}\mathbf{e}_r\mathbf{e}_r^\top \mathbf{R}_{\mathbf{x},f})} \in \mathbb{C}^M,$$

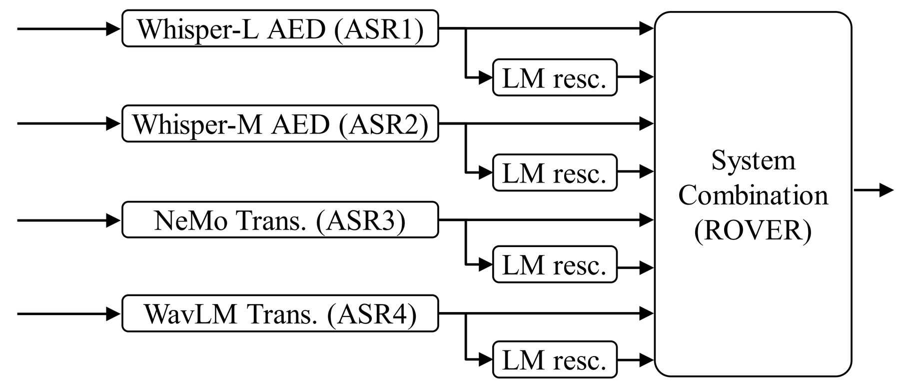
r: reference microphone selected by MaxSNR-based reference channel selection

μ = 0 in our implementation

ASR

ASR pipeline

1. Generate hypotheses from four end-to-end ASR models
2. Perform LM rescoring
3. System combination by ROVER



ASR models and LM

- **ASR1**: Whisper Large v3
  - 32-layer Transformer encoder and decoder, 1.54B parameters
  - When the decoded sample CER over 30%, we skip update parameters.
- **ASR2:** Whisper medium.en
  - 24-layer Transformer encoder and decoder, 770M parameters
- **ASR3**: NeMo Transducer
  - Encoder: 2-layer 2D CNN + 24 FastConformer blocks
  - Decoder: 1-layer LSTM with 640 cells + 640-dim. Joiner, 1024 BPE tokens, 644M params
- **ASR4**: WavLM Transducer
  - Encoder: WavLM preencoder + 2-layer 2D CNN + 18 Branchformer blocks
  - Decoder: 2-layer LSTM with 640 cells + 512-dim. Joiner, 500 BPE tokens, 422M params
- **Rescoring**: Transformer LM
  - 512-att-dims, 2048-MLP-dims, 16-layers, 1000-BPE-tokens, 68M parameters
  - Using 256 past rescored (re-ranked) 1-best tokens as the context in N-best rescoring

Training data

70 hours of CHiME-8 training data processed with GSS for the Oracle segmentation

tcpWER [%] (↓) on the dev set with oracle diarization and SE front-end.

ID	Model	CH6	DiP	MX6	NSF	Macro
ASR0	NeMo Trans. (Baseline)	19.78	31.01	10.61	17.95	19.84
ASR1	Whisper-L AED	17.80	26.29	10.43	13.05	16.89
ASR2	Whisper-M AED	19.81	27.15	11.16	13.57	17.92
ASR3	NeMo Trans.	20.30	28.33	11.25	14.33	18.55
ASR4	WavLM Trans.	19.76	27.52	10.79	13.23	17.82
ASR5	ROVER (ASR × 6 +LM resc.)	16.42	23.71	9.42	11.44	15.25

Overall results & Discussions

tcpWER [%] (↓) on the dev and eval sets. The real-time factor(RTF) is computed on the NOTSOFAR dev set.

				dev					eval					
ID	Diar	SE	ASR	CH6	DiP	MX6	NSF	Macro	CH6	DiP	MX6	NSF	Macro	RTF
Baseline	NeMo	-	-	49.3	78.9	15.8	56.2	50.0	56.5	75.8	19.4	61.0	53.2	-
NTT-1	DIA1	SE	ASR4	30.1	35.9	10.9	23.9	25.2	44.8	26.2	15.6	22.1	27.2	2.46
NTT-2	DIA2	SE	ASR1	28.2	35.3	10.7	20.4	23.7	38.7	25.0	14.9	18.3	24.3	3.14
NTT-3	DIA2	SE	ASR5 (ROVER)	25.5	31.3	9.6	18.8	21.3	35.3	22.4	13.5	16.8	22.0	4.03

We proposed three versions of our DASR system, each with a different computational complexity.
They achieve between 49% and 59 % of relative tcpWER improvement over the NEMO baseline.