



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

{naoyuki.kamo, naohiro.tawara, atsushi.ando, hrs.sato, marc.delcroix}@ntt.com

Abstract

We present a distant automatic speech recognition (DASR) system developed for the CHiME-8 DASR track. It consists of a diarization first pipeline. For diarization, we use end-to-end diarization with vector clustering (EEND-VC) followed by target speaker voice activity detection (TS-VAD) refinement. To deal with various numbers of speakers, we developed a new multi-channel speaker counting approach. We then apply guided source separation (GSS) with several improvements to the baseline system. Finally, we perform ASR using a combination of systems built from strong pre-trained models. Our proposed system achieves a macro tcpWER of 22.0 % on the dev set, which is a more than 60% relative improvement over the baseline.

Index Terms: Robust ASR, multi-talker ASR, speaker diarization, CHiME-8 DASR

1. Introduction

The distant automatic speech recognition (DASR) problem consists of identifying when each speaker speaks (diarization) and transcribing their speech (ASR) in conversations captured by distant microphones. Over the years, the CHiME challenge series has proposed tasks with increased levels of difficulty to measure progress in DASR, but was limited to recordings of up to four speakers. The CHiME-8 DASR [1, 2] track extends the difficulties of the previous editions by expanding the variety in the number of speakers per recording (up to eight), microphone array configurations, recording conditions, and speaking styles. Concretely, this is realized by requiring building a single DASR system, which can operate on four datasets, including dinner party recordings with four participants (CHiME 6 (CH6) [3] and DiPCO (DiP) [4]), two-speaker interviews (Mixer 6 (MX6) [5]) and a new corpus of business-like meetings called NOTSOFAR (NSF) [6].

Our contribution to the CHiME-8 DASR track consists of a diarization first pipeline [7], which builds on the system we proposed for CHiME-7 [8] and combines speaker diarization, speech enhancement (SE), and ASR as shown in Fig. 1. For the diarization, we extended our previously proposed end-to-end diarization with vector clustering (EEND-VC)-based diarization [9, 10] to include target-speaker voice activity detection (TS-VAD)-based refinement [11, 12]. Besides, we developed a novel multi-microphone speaker counting approach, which estimates the number of speakers via speaker embedding clustering per microphone and combines the result across all microphones. The speaker counting is crucial for CHiME-8 DASR task as there is great variety in the number of speakers per recording,

*Equal contribution

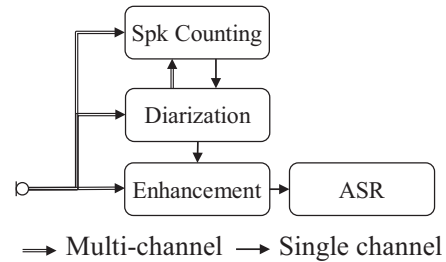


Figure 1: Proposed recognition system for DASR track.

and wrongly estimating the number of speakers can greatly impact diarization and ASR performance.

For SE, we made several key modifications to the baseline guided source separation (GSS). First, we propose a new rule for microphone subset selection, which is based on the envelope variance [13] and the speech clarity index C_{50} [14]. Besides, we refined the SE frontend by replacing the MVDR beamformer with the spatial-prediction multichannel Wiener filter (SP-MWF) [15, 16]. By doing so, we mainly aim to select a more effective reference microphone for beamforming, which is essential when dealing with distributed microphone arrays.

For ASR, we exploit the availability of strong pre-trained models, including Whisper, NeMo, and WavLM. First, we investigated fine-tuning Whisper and NeMo models on the CHiME-8 training data. We introduce a curriculum learning scheme to efficiently fine-tune Whisper on the very noisy CHiME-8 training data. In addition to the above models, we also developed a transducer-based ASR system, which uses WavLM as the front-end. This last model, although being much more computationally efficient, achieves comparable performance to the Whisper- and NeMo-based models. Finally, we perform N-best rescoring and system combination.

In the remainder, we describe the different parts of our system, i.e., diarization and speaker counting in Section 2, SE in Section 3 and ASR in Section 4. We then present overall results and analysis in Section 6.

2. Diarization and speaker counting

Figure 2 shows our proposed diarization system, which consists of EEND-VC [9] that relies on multi-channel speaker counting, followed with TS-VAD-based refinement [12]. We explored two diarization pipelines, i.e., *DIA1*, which consists of only EEND-VC segmentation and clustering, and *DIA2*, which performs TS-VAD refinement on top of EEND-VC segmentation.

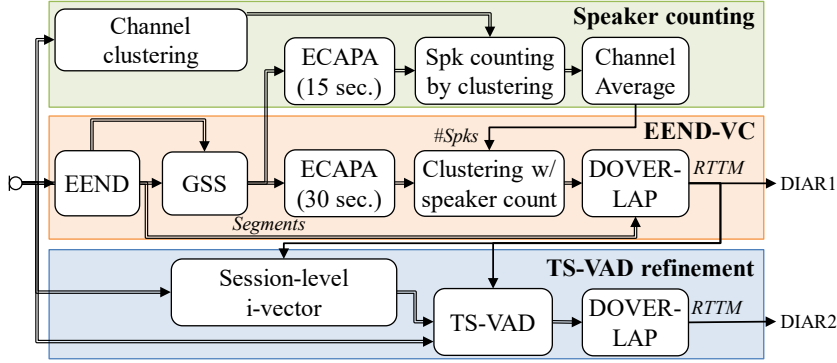


Figure 2: Proposed diarization and speaker counting system.

2.1. EEND-VC segmentation (DIA1)

The EEND-VC module is based on our CHiME-7 submission [8, 10], with some essential modifications. EEND-VC performs chunk-level segmentation to estimate the activity of each speaker in each chunk with EEND, where the maximum number of speakers in a chunk is set to $N^{\max} = 4$. Next, we aggregate the chunk-level segmentation results by clustering speaker embeddings computed for each speaker in each chunk with the ECAPA-TDNN model [17]. This process is performed for each channel, and then combined with diarization output voting error reduction + Overlap (DOVER-Lap) [18].

We made two main additions to our previous EEND-VC system. First, we apply GSS to each 30-sec chunk using the chunk-wise segmentation result obtained with EEND. We use the signals after GSS to compute cleaner speaker embeddings for speaker counting and clustering. We expect that GSS can reduce the interference speakers and noise, resulting in cleaner speaker embeddings. Here, we perform GSS using the segmentation obtained from each channel, which generates thus as many signals as channels. This allows us to generate multiple diarization outputs for DOVER-Lap, whose reliability may vary dynamically depending on the microphone used.

Second, we perform constrained spectral clustering [9] by setting the number of clusters to the number of speakers obtained from the speaker counting module, described in the next subsection.

Settings: We used the same configuration for EEND-VC as in our CHiME-7 submission [10, 8], which uses pre-trained WavLM-large [19] to obtain the input speech features. The model has approximately 324 million parameters. Different from our previous system, we use a chunk size of 30 seconds instead of 80 sec. We could obtain reliable embeddings with shorter chunks by using GSS to reduce the influence of the interference speakers. Reducing the chunk size is essential to allow dealing with recording with more than $N^{\max} = 4$ speakers with high overlap as observed in the NOTSOFAR data [6].

2.2. Multi-channel speaker counting

It is challenging in short sessions to count speakers since the number of segment-level speaker embeddings is limited. To solve this problem, we propose a speaker counting scheme that predicts the number of speakers from multi-channel signals. Channel clustering and GSS are employed to mitigate the influence of spatial characteristic differences across channels on speaker embeddings. The number of speakers is estimated in

each channel cluster, enabling to increase the number of embeddings, which leads to better speaker counting performance.

First, input channels are clustered to find nearby microphone groups in the input session. For this, we grouped channels using agglomerative hierarchical clustering based on inter-channel correlations from the initial fixed-length raw signals.

Then, we compute speaker embeddings on the output of GSS applied in the EEND-VC stage using the ECAPA-TDNN model. For each microphone group, the number of speakers is estimated from all embeddings in the group using the approach proposed for normalized maximum eigengap spectral clustering (NMESC) [20].

Finally, the estimated speaker counts are integrated to obtain the session-level speaker counts \bar{c} , $\bar{c} = \lfloor \frac{1}{N} \sum_i n_i c_i \rfloor$, where c_i and n_i are estimated speaker counts and the number of embeddings in the i -th microphone group, N is the total number of embeddings in the session, and $\lfloor \cdot \rfloor$ is a rounding function to an integer. n_i serves as a weight factor that emphasizes the estimations from groups with more microphones, which are thus more reliable. The details of the proposed multi-channel speaker counting can be found in our subsequent paper [21].

Settings: We divided the output of GSS into 15-second segments to generate more samples. We used the first 120 seconds of the signals and a correlation threshold of 0.3 for channel clustering.

2.3. TS-VAD refinement (DIA2)

We apply memory-aware multi-speaker embedding with sequence-to-sequence architecture (NSD-MS2S) -based TS-VAD [12, 22] to refine the diarization results obtained with EEND-VC. NSD-MS2S was used in the top system for the CHiME-7 DASR task.

NSD-MS2S exploits session-level ivectors, obtained by averaging the segment-level ivectors belonging to the same speaker. It then combines these ivectors with local speaker embeddings derived from the input mixture and the local segmentation information. These combined embeddings are used to condition a conformer-based TS-VAD module. We use the same model configuration that was used in the CHiME-7 [22], but with a stronger initial diarization provided by EEND-VC.

Settings: We rely on the publicly available implementation of NSD-MS2S for the TS-VAD refinement¹. We used the default parameters, except that we used only a single deep interactive module block. The model has 5.80 million parameters.

¹<https://github.com/liyunlongaaa/NSD-MS2S>

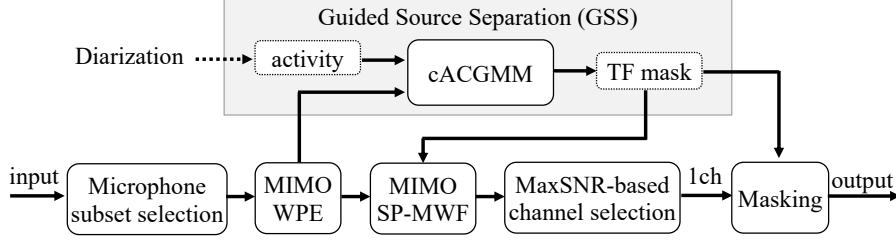


Figure 3: Our GSS-based SE frontend.

3. Speech enhancement (SE) front-end

Figure 3 shows our proposed GSS-based SE frontend, which basically follows the official SE frontend [23] implemented in the CHiME-8 DASR NeMo Baseline system [2, 24]. It consists of microphone subset selection, weighted prediction error (WPE)-based dereverberation [25, 26], time-frequency (TF) mask estimation using cACGMM [27] based GSS [23], and SE using a variant of the multichannel Wiener filter (MWF) followed by TF masking. Changes from the official SE frontend include improvement of microphone subset selection (Section 3.1) and replacement of the MVDR beamformer (Section 3.2).

3.1. Microphone subset selection

To select an effective subset of microphones for the SE frontend, we relied on two acoustic features: the envelop variance (EV) [13] (which is also used in the CHiME-8 DASR NeMo Baseline system) and a speech clarity index C_{50} [14]. The index C_{50} is defined as the ratio of the energy in the early phase (0 to 50 ms) to that in the late phase (more than 50 ms) of the room impulse response. We estimated C_{50} using the Brouhaha toolkit [28].

We measured EV and C_{50} for each microphone observation signal to rank the microphones. Let I_{EV} (resp. $I_{C_{50}}$) be the set of the top K microphones ranked by EV (resp. C_{50}), where $K = 65$ [%] in our setup. Let also $I = I_{EV} \cap I_{C_{50}}$ be the intersection of the two subsets. We selected the subset of microphones to pass to the subsequent SE frontend as follows:

- If $|I| \geq 15$, then we select I .
- If $|I| < 15$ and $I_{EV} \geq 15$, then we select I_{EV} .
- If $|I| < 15$ and $I_{EV} < 15$, then we select the set of the top 15 microphones ranked by EV.
- We use all microphones when there are fewer than 15.

In the aforementioned rule, we assumed that using at least 15 microphones helps improve SE performance. The value of 15 was adjusted among $\{5, 10, 15, 20\}$ using the dev set.

3.2. Mask-based MIMO source separation filter

As a source separation filter, the official SE frontend uses the mask-based MIMO MVDR beamformer with maximum SNR (MaxSNR)-based reference channel selection. We replaced this beamformer part with the so-called spatial-prediction multichannel Wiener filter (SP-MWF) [15, 16] given by

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

where $r \in \{1, \dots, M\}$ denotes the reference channel, $\mathbf{e}_r \in \mathbb{C}^M$ is the unit vector that selects the r -th microphone, $\mu \in \mathbb{R}_{\geq 0}$ is a hyperparameter (we set it to $\mu = 0$), and $\mathbf{R}_{\mathbf{x},f}$ and $\mathbf{R}_{\mathbf{n},f}$

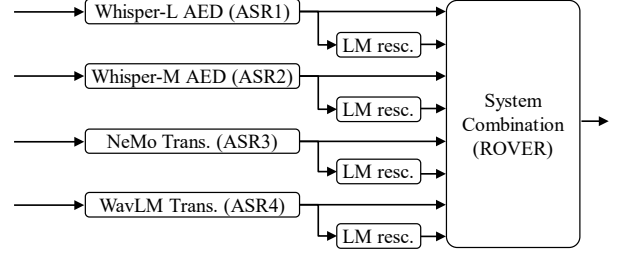


Figure 4: Proposed ASR backend.

are the target-source and noise covariance matrices at frequency bin f . The covariance matrices are estimated using the TF mask in the same way as in the baseline. Finally, unlike the baseline, we do not apply blind analytic normalization (BAN) postfilter.

4. ASR back-end

Figure 4 shows the schematic diagram of our ASR backend. We employed four ASR models: two attentional encoder-decoder (AED) models and two transducer models [29]. Each ASR model generated N-best hypotheses, which were accumulated and rescored by language models. Both the beam size and N-best size were set to 4. Then, the best hypothesis was determined among the original and rescored N-best hypotheses using recognizer output voting error reduction (ROVER) [30]. The architecture of our ASR models is described below.

4.1. Whisper Large v3 (ASR1)

We fine-tuned the Whisper Large V3 model [31] for the CHiME-8 task. The model has 1540M parameters, i.e., 32 Transformer encoder-decoder layers with 8 attention heads and a hidden dimension of 1280. The vocabulary size was 51,864 with the GPT-2 [32] byte-level BPE tokenizer.

The CHiME-8 training data is very noisy and thus unreliable for fine-tuning. To mitigate this issue, we proposed a curriculum learning scheme, which filters out utterances with a high character error rate (CER). Practically, during fine-tuning, we changed the target transcription of the cross-entropy loss to the self-generated decoding result when the CER exceeds 30%. In that case, we multiplied the loss by 1/1000th to reduce its impact. Otherwise, we used the ground-truth transcription as a reference. By computing the CER on the fly, we can adapt the number of training data as fine-tuning progresses, i.e., as the model becomes stronger, the CER decreases, and more difficult data can be reliably used.

4.2. Whisper Medium (ASR2)

We also utilized Whisper Medium English model [31] to initialize a Transformer-based encoder-decoder model. This model had approximately 770M parameters and consisted of a 24-layer Transformer encoder and decoder, each with 8 attention heads and a model width of 1024. The vocabulary size was 51,864 with the GPT-2 [32] byte-level BPE tokenizer. Despite having significantly fewer parameters than Whisper Large V3, it is pre-trained solely on English data, making it potentially more suitable for the CHiME-8 tasks.

4.3. NeMo Transducer (ASR3)

We adopted an official pre-trained NeMo transducer model, which had 644M parameters, and finetuned it using the CHiME-8 dataset. The NeMo transducer model consists of two-layer 2D convolutional neural networks (CNNs) followed by 24 fast-conformer blocks [33, 34]. The prediction and joint networks had a 640-dimensional long short-term memory (LSTM) and a 640-dimensional feed-forward network. The number of output units was 1025 byte pair encoding (BPE) tokens.

4.4. WavLM Transducer (ASR4)

We built another transducer-based ASR system that uses the weighted sum of WavLM [19] Transformer layers as input features. The ASR encoder has two 2D-CNN layers followed by 18 branchformer blocks [35]. The prediction and joint networks had two-layer 640-dimensional LSTMs and a 512-dimensional feed-forward network, respectively. We adopted 500 BPE tokens as output units. The total number of parameters was approximately 422M.

We conducted three-step training to build the ASR4 system sequentially: 1) partial parameters were trained using the CHiME&LibriSpeech&VoxCeleb datasets while freezing WavLM front-end, 2) all network parameters (including WavLM) were fine-tuned using the same data from the first step, and 3) we fine-tuned it using only the CHiME-8 data.

4.5. Language model

We built a Transformer-LM consisting of 35M parameters for LM rescoring. The LM has the vocabulary of 1000 BPE tokens. We pre-trained the LM using 1/10 of the LibriSpeech text dataset and then fine-tuned using the CHiME-8 train text dataset. At the inference, the LM uses 256 past rescored (re-ranked) 1-best tokens as the context (i.e., context carry-over) [36, 37].

5. Training data

Diarization: Each diarization model (EEND-VC and TS-VAD) was initially trained using simulated mixtures and then finetuned using the CHiME-8 training set [1]. The protocol for simulating mixtures basically followed the method that attempts to make utterance transitions natural [38], but the following modifications were made to have more similar statistics to the real data: i) we first considered turn-hold, turn-switch, and interruption to generate long-form audio, and then inserted backchannels afterward, ii) we directly sampled durations of silence/overlap between utterance from the real data instead of sampling from any fitted distribution, and iii) overlap durations in interruptions/backchannels were determined from absolute durations extracted from the real data (instead of relative ratios). We generated 1M and 91k 50-second mixtures of four speakers

Table 1: DER [%] (\downarrow) 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 | DIA1 + TS-VAD | 23.97 | 21.01 | 6.11 | 9.72 | 15.20 |

using LibriSpeech [39] for training EEND-VC and TS-VAD, respectively. We used 500 mixtures for validation in both cases. Each mixture was augmented using the simulated room impulse responses [40] and MUSAN noises [41].

ASR: We used 70 hours of CHiME-8 training data processed with GSS for the Oracle segmentation. We did not use train_call and train_intv in Mixer6 at this time because preliminary experiments showed minimal improvements with them. For the pre-training of the transducer-based ASR model with WavLM (ASR4), we also used LibriSpeech [39], and VoxCeleb1+2 [42] datasets in addition to the CHiME-8 described above. Note that the contrastive data selection algorithm [43, 8] was applied in an unsupervised fashion to the unlabeled VoxCeleb1+2 data, reducing its size by a quarter. We utilized the dev set for early stopping with a patience of 5 epochs.

6. Experiments

We report a brief analysis of the different components of our system followed by the overall results in Section 6.4. We report results on the dev and eval sets of the CHiME-8 DASR task [1] in terms of diarization error rate (DER) computed with md-eval² and time-constrained minimum-Permutation Word Error Rate (tcpWER) computed with the MeetEval toolkit[44].³

6.1. Analysis I: Diarization and speaker counting

Table 1 shows the DER on dev set, without and with TS-VAD refinement, DIA1 and 2, respectively. Both systems greatly outperform the baseline. Note that we used a similar TS-VAD model as that of the top CHiME-7 system [22], but with a more powerful initialization with EEND-VC. Consequently, we achieved better performance with a single TS-VAD refinement pass, i.e., DERs of 24.0 % and 6.1 % on CHiME-6 and Mixer-6 dev set, compared to 25.8 % and 8.9 % with 4 diarization pass in the CHiME-7 top system [22].

Table 2 shows the channel-wise and microphone group-wise speaker counting accuracy, without and with group averaging. Our proposed system accurately estimated the number of speakers for CHiME-6, DiPCO, and Mixer 6 datasets. For NOTSOFAR, we achieved a lower accuracy of 58.2 % accuracy because there is more variability in the number of speakers, and the recordings are short. The proposed speaker-counting approach greatly outperforms the baseline in all conditions. Note that we report several ablation studies, such as the effectiveness of GSS in speaker counting, in our subsequent paper [21].

6.2. Analysis II: Speech enhancement

In preliminary experiments, we confirmed that our new microphone subset selection (Section 3.1) could improve the tcpWER

²<https://github.com/foundintranslation/Kaldi/blob/master/tools/sctk-2.4.0/src/md-eval/md-eval.pl>

³<https://github.com/fgnt/meeteval/tree/main>

Table 2: Speaker counting accuracy [%] (\uparrow) 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 |

Table 3: tcpWER [%] (\downarrow) 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 \times 6 +LM resc.) | 16.42 | 23.71 | 9.42 | 11.44 | 15.25 |

from 20.00 % to 19.41 % in the CHiME-6 dataset and from 31.43 % to 29.70 % in DiPCo when using the baseline NeMo ASR system with oracle diarization. Note that since Mixer 6 and NOTSOFAR have fewer than 15 microphones, all microphones were selected.

We also observed that replacing the separation filter and turning off the postfilter in the baseline system could improve the macro tcpWER by more than 0.3 %. When using ASR4 in Table 3 with oracle diarization, we observed that simply replacing the baseline separation filter with SP-MWF (where the BAN postfilter was still used) decreased the macro tcpWER by 0.18 %.

6.3. Analysis III: ASR

Table 3 shows the tcpWERs of our developed ASR backends when using Oracle diarization and our SE front-end. The tcpWERs of all systems were significantly improved compared to those of the baseline. Although the best single ASR system is the Whisper Large v3 (ASR1), the WavLM transducer (ASR4), which is the smallest system, achieved comparable performance. Note that we trained two versions of ASR1 and ASR4 with different training hyper-parameters, but only reported the best version in Table 3. ASR 5 consists of the combination of the six ASR systems using both 1-best and hypothesis obtained after LM rescoring.

6.4. Overall results on dev set

Table 4 compares the results of the three proposed DASR systems (NTT-1–3) with the baseline. The systems are sorted by increasing order of complexity.

- **NTT-1** is our most lightweight model, which uses only EEND-VC for diarization (without TS-VAD refinement), our proposed GSS front-end with microphone selection (SE1), and our most efficient ASR system (ASR4).
- **NTT-2** uses a stronger diarization (DIAR2) and ASR system (ASR1).
- **NTT-3** performs decoding with six ASR backends and system combination with ROVER as ASR5 in Table 3.

Our full system (NTT-3) achieves a relative improvement of 57 % over the baseline. Note that it also achieves a 40 %

Table 4: tcpWER [%] (\downarrow) on the dev set. The real-time factor (RTF) is computed on the NOTSOFAR dev set.

| ID | Diar | SE | ASR | CH6 | DiP | MX6 | NSF | Macro | RTF |
|-----------------|------|----|--------------|-------|-------|-------|-------|-------|------|
| Baseline (NeMo) | | | | 49.29 | 78.87 | 15.75 | 56.21 | 50.03 | - |
| NTT-1 | DIA1 | SE | ASR4 | 30.14 | 35.86 | 10.94 | 23.85 | 25.20 | 2.46 |
| NTT-2 | DIA2 | SE | ASR1 | 28.21 | 35.32 | 10.66 | 20.41 | 23.65 | 3.14 |
| NTT-3 | DIA2 | SE | ASR5 (ROVER) | 25.49 | 31.25 | 9.63 | 18.79 | 21.29 | 4.03 |

Table 5: tcpWER [%] (\downarrow) on the eval set.

| ID | Diar | SE | ASR | CH6 | DiP | MX6 | NSF | Macro |
|-----------------|------|----|--------------|-------|-------|-------|-------|-------|
| Baseline (NeMo) | | | | 73.80 | 57.10 | 23.20 | 72.00 | 56.50 |
| NTT-1 | DIA1 | SE | ASR4 | 44.80 | 26.20 | 15.60 | 22.10 | 27.20 |
| NTT-2 | DIA2 | SE | ASR1 | 38.70 | 25.00 | 14.90 | 18.30 | 24.30 |
| NTT-3 | DIA2 | SE | ASR5 (ROVER) | 35.30 | 22.40 | 13.50 | 16.80 | 22.00 |

relative tcpWER improvement over the DASR baseline, despite the fact that our system generalizes to more diverse recording conditions. NTT-3 is relatively complex and involves the combination of several ASR backends. However, lighter versions (NTT-1 and NTT-2) of our system can still achieve more than 50 % relative improvement compared to the DASR baseline.

We also provide RTF values computed on one A6000 GPU on the NOTSOFAR dataset, but these numbers are only indicative as our code has not been optimized for computational efficiency.

6.5. Results on Eval set

Table 5 shows the overall results on the evaluation set. Our proposed system significantly improved performance over the baseline for all four datasets. It achieved a macro tcp-WER improvement of 61 %. This system ranked second place in the DASR task. Interestingly, it also achieved the third-best performance on the NOTSOFAR task [45], although the system can handle various recording conditions and is not fine-tuned for the NOTSOFAR task.

7. References

- [1] S. Cornell, T. Park, S. Huang, C. Boeddeker, X. Chang, M. Maciejewski, M. Wiesner, P. Garcia, and S. Watanabe, “The chime-8 dasr challenge for generalizable and array agnostic distant automatic speech recognition and diarization,” *arXiv preprint arXiv:2407.16447*, 2024.
- [2] *CHiME-8 Task 1 - DASR*, <https://www.chimechallenge.org/current/task1>.
- [3] S. Watanabe, M. Mandel *et al.*, “CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings,” in *CHiME*, 2020, pp. 1–7.
- [4] M. V. Segbroeck, Z. Ahmed *et al.*, “DiPCo—Dinner Party Corpus,” in *Interspeech*, 2020, pp. 434–436.
- [5] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, “The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition,” in *LREC*, 2010, pp. 2441–2444.
- [6] A. Vinnikov, A. Ivry *et al.*, “NOTSOFAR-1 challenge: New datasets, baseline, and tasks for distant meeting transcription,” *arxiv:2401.08887*, 2024.
- [7] D. Raj, P. Denisov, *et al.*, “Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis,” in *SLT*, 2021, pp. 897–904.
- [8] N. Kamo, N. Tawara *et al.*, “NTT multi-speaker ASR system for the DASR task of CHiME-7 challenge,” in *CHiME*, 2023, pp. 45–50.

- [9] K. Kinoshita, M. Delcroix, and N. Tawara, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," in *Interspeech*, 2021, pp. 3565–3569.
- [10] N. Tawara, M. Delcroix, A. Ando, and A. Ogawa, "NTT speaker diarization system for CHiME-7: Multi-domain, multi-microphone end-to-end and vector clustering diarization," in *ICASSP*, 2024, pp. 11 281–11 285.
- [11] I. Medennikov, M. Korenevsky *et al.*, "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," in *Interspeech*, 2020, pp. 274–278.
- [12] G. Yang, M. He *et al.*, "Neural speaker diarization using memory-aware multi-speaker embedding with sequence-to-sequence architecture," in *ICASSP*, 2024, pp. 11 626–11 630.
- [13] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170–180, 2014.
- [14] Organización Internacional de Normalización, *ISO 3382-1: Acoustics - Measurement of Room Acoustic Parameters. Part 1: Performance Rooms*. ISO, 2009.
- [15] J. Benesty, J. Chen, and Y. Huang, "Noncausal (frequency-domain) optimal filters," *Microphone Array Signal Processing*, pp. 115–137, 2008.
- [16] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, 2011.
- [17] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020, pp. 3830–3834.
- [18] D. Raj, P. Garcia, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "DOVER-Lap: A method for combining overlap-aware diarization outputs," in *SLT*, 2021, pp. 881–888.
- [19] S. Chen, C. Wang *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [20] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2020.
- [21] N. Tawara, A. Ando, S. Horiguchi, and M. Delcroix, "Multi-channel speaker counting for EEND-VC-based speaker diarization on multi-domain conversation," *Submitted to ICASSP'25*, 2025.
- [22] R. Wan, M. He *et al.*, "The USTC-NERCSLIP systems for CHiME-7 challenge," in *CHiME*, 2023, pp. 13–18.
- [23] C. Boeddecker, J. Heitkaemper, J. Schmalenstroerer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME*, 2018, pp. 35–40.
- [24] T. J. Park, H. Huang *et al.*, "The CHiME-7 challenge: System description and performance of NeMo team's DASR system," in *CHiME*, 2023, pp. 57–62.
- [25] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.
- [26] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.
- [27] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *EUSIPCO*, 2016, pp. 1153–1157.
- [28] M. Lavechin, M. Métais *et al.*, "Brouhaha: Multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation," in *ASRU*, 2023.
- [29] A. Graves, "Sequence Transduction with Recurrent Neural Networks," in *ICML Workshop on Representation Learning*, 2012.
- [30] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," *ASRU*, pp. 347–354, 1997.
- [31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023, pp. 28 492–28 518.
- [32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [33] A. Gulati, J. Qin *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020, pp. 5036–5040.
- [34] D. Rekesch, N. R. Koluguri *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," in *ASRU*, 2023.
- [35] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-Branchformer: Branchformer with enhanced merging for speech recognition," in *SLT*, 2023, pp. 84–91.
- [36] A. Ogawa, N. Tawara, M. Delcroix, and S. Araki, "Lattice rescoring based on large ensemble of complementary neural language models," in *ICASSP*, 2022, pp. 6517–6521.
- [37] A. Ogawa, N. Kamo, K. Matsuura, T. Ashihara, T. Moriya, T. Kano, N. Tawara, and M. Delcroix, "Applying LLMs for rescoring N-best ASR hypotheses of casual conversations: Effects of domain adaptation and context carry-over," *arXiv:2406.18972*, 2024.
- [38] N. Yamashita, S. Horiguchi, and T. Homma, "Improving the naturalness of simulated conversations for end-to-end neural diarization," in *Odyssey*, 2022, pp. 133–140.
- [39] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.
- [40] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017, pp. 5220–5224.
- [41] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv:1510.08484*, 2015.
- [42] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.
- [43] Z. Lu, Y. Wang, Y. Zhang, W. Han, Z. Chen, and P. Haghani, "Unsupervised data selection via discrete speech representation for ASR," in *Interspeech*, 2022, pp. 3393–3397.
- [44] T. von Neumann, C. Boeddecker, M. Delcroix, and R. Haeb-Umbach, "MeetEval: A toolkit for computation of word error rates for meeting transcription systems," in *Proc. 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023)*, 2023, pp. 27–32.
- [45] *CHiME-8 Task 2 - NOTSO FAR Results*, <https://www.chimechallenge.org/current/task2/results#notsofar-and-dasr-results-supplementary>.