The CHiME-7 DASR Challenge: Distant Meeting Transcription with Multiple Devices in Diverse Scenarios

Samuele Cornell[1,2], Matthew Wiesner[3], Shinji Watanabe[2], Desh Raj[3], Xuankai Chang[2], Paola Garcia[3], Matthew Maciejewski[3], Yoshiki Masuyama[4], Zhong-Qiu Wang[2], Stefano Squartini[1], Sanjeev Khudanpur[3]

[1]Università Politecnica delle Marche, Italy [2]Carnegie Mellon University, USA
[3]Johns Hopkins University, USA [4]Tokyo Metropolitan University, Japan

Abstract

The CHiME challenges have played a significant role in the development and evaluation of robust automatic speech recognition (ASR) systems. We introduce the CHiME-7 distant ASR (DASR) task, within the 7th CHiME challenge. This task comprises joint ASR and diarization in far-field settings with multiple, and possibly heterogeneous, recording devices. Different from previous challenges, we evaluate systems on 3 diverse scenarios: CHiME-6, DiPCo, and Mixer 6. The goal is for participants to devise a single system that can generalize across different array geometries and use cases with no a-priori information. Another departure from earlier CHiME iterations is that participants are allowed to use open-source pre-trained models and datasets. In this paper, we describe the challenge design, motivation, and fundamental research questions in detail. We also present the baseline system, which is fully array-topology agnostic and features multi-channel diarization, channel selection, guided source separation and a robust ASR model that leverages self-supervised speech representations (SSLR).

Index Terms: CHiME challenge, speech recognition, multi-channel, speaker diarization, speech separation

1. Introduction

Despite significant leaps made in the last decade, reliable conversational speech recognition remains a significant challenge [1,2]. This is particularly true in meeting scenarios which are often constrained to use only distant array devices [1,3–6]. Besides difficulties due to noise and reverberation in far-field speech, in meeting scenarios a crucial challenge is the presence of overlapped speech, which may amount to more than 15% of total conversation [7]. A third issue are linguistic artifacts arising from informal settings or domain-specific words, an often disregarded characteristic in ASR research, as most commonly used benchmark datasets feature scripted or semi-scripted conversations [2].

As these difficulties arise from diverse factors, multi-talker distant conversational speech recognition requires a comprehensive approach for the development and integration of both front-end and ASR back-end techniques. Over the years, several challenges and corpora have played a significant role in advancing research in this field. These include, but are not limited to, ASpiRE [8], AMI [9], ICSI [10], the CHiME (1–4) challenges [11–14], the Sheffield Wargames corpus [15], and DIRHA [16]. More recently, we have seen VOiCES [17], DiPCo [18], LibriCSS [3], Alimeeting [19], Ego4D [20], and the recent CHiME 5 and 6 challenges [1,21].

Several of these datasets/challenges, such as DIRHA, VOiCES, and CHiME-4, focused primarily on acoustic robustness, thus featuring pre-segmented single speaker utterances in a noisy/reverberant environment captured by multiple microphones. The objective in this case was to foster research in robust ASR methods, ignoring possible issues that can arise from overlapped speech, wrong segmentation, and colloquial language. These datasets comprised either fully simulated or semi-real environments, where utterances come from an audiobook speech corpus such as LibriSpeech [22]. A notable exception in this regard is LibriCSS. It features a simulated, unsegmented full meeting scenario between multiple participants in semi-real conditions; it was obtained by playing LibriSpeech utterances via loudspeakers in a real room.

AMI, ICSI, and Sheffield Wargames were historically among the first projects that featured real unsegmented meeting-style scenarios recorded in far-field settings. Although these are more expensive to collect compared with simulated or semi-real datasets, they better reflect possible real-world applications. Recently, this line of research has seen renewed interest, and has resulted in the collection and creation of datasets such as CHiME-5/6, DiPCo, Alimeeting, and Ego4D. The progress made in ASR, diarization and speech separation is a significant contributing factor to this trend, and it suggests that reliable and robust transcription in these scenarios is within our reach in the coming years.

The proposed CHiME-7 DASR challenge[1] follows this latter line of research on real, unsegmented meeting scenarios. Our main goal is to foster research in an important direction: how can we build systems that can generalize across a *wide range of real-world settings* and provide reliable ASR performance under adverse acoustic conditions? Additionally, we aim to incorporate recent progresses in *self-supervised learning* to address this problem as well as supervised DNN-based speech separation and enhancement (SSE), by allowing the use of external data and pre-trained models.

2. Motivation

Compared to the previous CHiME editions, CHiME-7 DASR features three main novelties, motivated by recent promising directions in speech processing.

2.1. Diverse Scenarios

To expand the breadth of evaluation conditions, we include two additional datasets (along with CHiME-6) — DiPCo [18], and Mixer 6 Speech [23]. The objective is to encourage research towards methods that (1) work well across different array topologies: linear (CHiME-6), circular (DiPCo), or heterogeneous (Mixer 6); (2) are capable of handling variable numbers of speakers in each session; (3) account for linguistic differ-

---

[1]CHiME-7 DASR website: chimechallenge.org/current/task1/index

ences between dinner party scenarios (CHiME-6 and DiPCo) versus interviews (Mixer 6); and (4) can effectively handle diverse acoustic conditions. Such variability reflects real use-cases where cross-domain generalization capability is desirable.

2.2. "Foundation" Models

Pre-trained speech models trained on large datasets are increasingly becoming available to the public. These include supervised models such as OpenAI Whisper [24], Nvidia JASPER [25] and QuartzNet [26], and self-supervised ones such as Wav2Vec 2.0 [27], HuBERT [28], and WavLM [29]. The latter, in particular, have proved effective in many downstream applications [30], and their integration with front-end speech enhancement has enabled state-of-the-art performance on benchmarks such as CHiME-4 [31, 32]. Leveraging these models offers unique opportunities as well as novel challenges. By allowing the use of pre-trained models, the CHiME-7 DASR task becomes more accessible to participants with limited resources, since training is often more efficient and quicker with such models. While SSLR models can be quite large, the features could be pre-extracted, thus enabling significant computational savings in the training phase. From a research perspective, it may be interesting to investigate how such models, which are trained in monoaural conditions, can make effective use of multi-channel audio and/or can be distilled into smaller models.

2.3. Open-Source Datasets

The allowable external data sources are expanded to include many commonly used open-source datasets (e.g. LibriSpeech [22] and FSD50k [33]). They facilitate the use of models for which, in prior challenges, necessary training data were lacking, but also enables the study of new research directions. For example, participants can create large synthetic datasets to train powerful deep neural network (DNN)-based speech separation and enhancement (SSE) [34, 35]. The structure of the challenge then encourages research into whether such models can effectively help improving real-world meeting transcription. This latter research direction also intersects with this year's "sister" CHiME-7 UDASE challenge, whose focus is on unsupervised domain adaptation for speech enhancement.

3. Challenge Tracks & Rules

The CHiME-7 DASR challenge features two tracks — a main track, and an optional sub-track, as described below.

3.1. Main Track

For the main track, we provide multi-channel recordings for the whole session without segmentation or speaker labels, and participants are required to generate time-marked, speaker-attributed transcripts, as shown in Fig. 1. For any session, let $\mathbf{r}_i = (\Delta, s, \mathbf{v})$ denote a time-marked, speaker-attributed transcript of a segment, where $\Delta = (t_{st}, t_{en})$ is a tuple containing start and end times, $s \in \mathbb{Z}^+$ is the speaker label[2], and $\mathbf{v} \in \Sigma^*$ represents the transcript for the segment, for some vocabulary Σ. A session is then completely defined by $\mathbf{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$, which we call the *reference*. Challenge participants are required to estimate the reference through generated *hypothesis*, denoted by $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_{\hat{N}}\}$. In this task estimated speaker labels $\hat{s}$

---

[2]Actual label names may be arbitrary, but can be mapped to the set of positive integers without loss of generality.
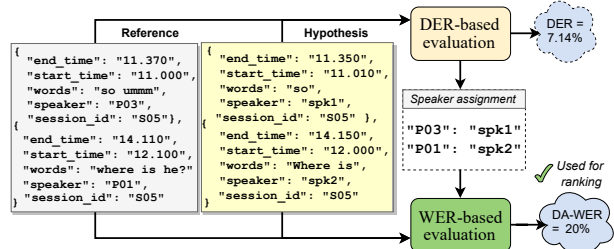

Figure 1: *Evaluation scheme for CHiME-7 DASR main track. Optimal speaker assignments are first determined using a DER-based evaluation, and then used to compute SA-WER, which is used as the final ranking metric.*

are not required to be identical to the reference labels. This flexibility poses a challenge in computing speaker-attributed word error rates (SA-WER) since, unlike earlier studies [36], the mapping between reference and estimated labels is not known.

We solve this problem using mechanisms employed in the evaluation of diarization systems, as shown in Fig. 1. Let $\Psi : \mathbb{Z}^+ \to \mathbb{Z}^+ \cup \{\phi\}$ denote a mapping from the reference labels to the estimated labels. We use the Hungarian method [37] to obtain the optimal mapping $\hat{\Psi}$ that minimizes the diarization error rate (DER) between the reference segments and hypothesized segments, i.e.,

$$\hat{\Psi} = \arg\min_{\Psi} \text{DER}(\mathbf{R}, \mathbf{H}), \quad (1)$$

where the DER is calculated with a 250 ms collar, following the setup in CHiME-6. Given $\hat{\Psi}$, we compute this newly proposed diarization attributed WER (DA-WER) metric as

$$\text{DA-WER}(\mathbf{R}, \mathbf{H}) = \frac{\sum_s \mathcal{L}(\frown \mathbf{R}^s, \frown \mathbf{H}^{\hat{\Psi}(s)})}{\sum_s |\frown \mathbf{R}^s|}, \quad (2)$$

where $\mathbf{R}^s$ denotes segments in $\mathbf{R}$ with speaker s, $\mathcal{L}$ is the Levenshtein distance, and the $\frown$ operator concatenates segment transcripts of its operand. Final ranking of systems will be based on DA-WER macro-averaged across all scenarios. The macro-average operation is to encourage participants to develop a model whose performance is consistent across all three scenarios. Note that this evaluation is different from the concatenated minimum-permutation WER (cpWER) based ranking in the CHiME-6 challenge, since it requires participants to also produce estimates of the utterances boundaries. Utterance boundaries estimation are important in many applications, e.g. for double checking the transcripts. It also differs from the Asclite multi-dimensional Levenshtein metric [38], which can account for segmentation errors when the time-based cost is used. Contrary to this latter, DA-WER does not require word-level alignment, which is quite time-consuming to double-check for ground-truth purposes. DA-WER instead offers a simple way to encourage participants to produce reasonable segmentation at the utterance level, which is sufficient in most applications. Future work is needed to devise more principled metrics for joint ASR and diarization.

3.2. Optional Sub-Track

Similar to the CHiME-6 challenge, we include an optional sub-track where participants can use oracle diarization. The goal of this sub-track is to assess the acoustic front-end and ASR performance and disentangle it from the impact of diarization. The segmentation and speaker labels are known and are freely usable by participants, reducing the problem to an ASR-only task.

3.3. Rules

Detailed description of the task rules are available in the CHiME-7 DASR website[3]. Hereafter we briefly summarize them.

- Only some external data and pre-trained models are allowed. Participants are encouraged to propose additional ones in the first month.
- Close-talk microphones cannot be used for inference.
- *Automatic domain identification is prohibited.* This includes the use of a-priori information such as array topology (including number of channels). *A single system must be submitted for all three scenarios.*
- Automatic channel selection is allowed and encouraged.
- *Self-supervised adaptation/training is allowed*, as long as it is performed for each evaluation session independently i.e. as it would happen in a real-world deployment.
- Participants may use all available data (including external data) for language model training, but cannot use large language models (LM) such as BERT [39]. This because we want participants to focus mainly on acoustic robustness.

4. Datasets

As explained in Section 2, CHiME-7 DASR is composed of three different scenarios/datasets: CHiME-6, DiPCo, and Mixer 6 Speech. For the purpose of this challenge, several changes have been made to these aforementioned data sources. In particular, we partially re-annotated Mixer 6 and we harmonized the transcripts across each of the three by using the same convention for common non-verbal speech sounds such as "mhm", "mmh", "hmm", which are rather frequent. Dataset statistics are summarized in Table 1. The annotation comes in the form of JSON files (one for each session), in the style of CHiME-6 (see Listing 1). Hereafter, we describe each scenario in more detail.

Table 1: *CHiME-7 DASR data overview for the three scenarios. For the CHiME-6 data, we show statistics for the original annotation, as well as a force-aligned (FA) annotation. We report the number of utterances, speakers, and sessions, as well as silence (sil), single-speaker speech (1-spk) and overlapped speech (ovl) ratios over the total duration.*

Scenario	Split	Size (h)	Utts	Spk.	Sess.	sil (%)	1-spk (%)	ovl (%)
CHiME-6	train	30:57	67615	28	14	22.6	53.5	23.9
	dev	4:27	7437	8	2	12.5	43.7	43.8
	eval	10:21	20683	12	4	19.9	50.9	29.2
CHiME-6 (FA)	train	30:57	78640	28	14	38.8	50.2	11.0
	dev	4:27	11020	8	2	24.1	54.2	21.7
	eval	10:21	81890	12	4	38.8	49.2	12.0
DiPCo	dev	2:43	3673	16	5	7.6	66.5	25.8
	eval	2:36	3405	16	5	9.2	65.8	25.0
Mixer 6	train calls	36:09	27280	81	243	–	–	–
	train intv	26:57	29893	77	189	–	–	–
	dev	14:75	14863	27	59	3.1	76.2	20.7
	eval	5:45	5115	18	23	2.4	83.6	13.9

4.1. CHiME-6

The CHiME-6 dataset features real dinner parties with 4 participants in a home environment, usually across different rooms. Due to the particular setting, it features informal speech and low signal-to-noise ratio (SNR). Recordings from binaural microphones worn by each speaker are provided along with distant speech captured by 6 Kinect array devices with 4 microphones each for a total of 24 microphones. Two annotations are provided [1], one manual and another obtained via forced

alignment (FA)[4]. In CHiME-7 DASR, we retain the CHiME-6 dataset in its original form, with the exception that two sessions were moved from the original `train` set to the `eval` set. This in order to have a bigger `eval` set, with more variety in acoustical conditions, compared to the CHiME-6 challenge. Importantly, we can still reasonably compare results with the previous challenge ones, on the dev set and the original portion from `eval`. We also provide universal evaluation map (UEM) files to exclude the speaker enrollment from evaluation, a feature missing from the previous iteration which incorrectly scored this portion as silence.

```
1  ''end_time'': ''76.340'',
2  ''start_time'': ''73.500'',
3  ''words'': ''Let's do lunch. [laughs]'',
4  ''speaker'': ''P08'',
5  ''ref'': ''U02'',
6  ''location'': ''kitchen'',
7  ''session_id'': ''S02''
```

Listing 1: *Sample annotation for the CHiME-6 dataset.*

4.2. DiPCo

DiPCo consists of 10 recorded dinner party sessions, each between 4 speakers, recorded on 5 far-field devices each with a 7-mic circular array (six plus one microphone at the center). Per-speaker close-talk microphones are provided, and have less cross-talk than the corresponding mics from CHiME-6 . Additionally, each session has a lower duration (15–47 mins) and is recorded in the same single room. Importantly, the additional DiPCo data allows us to evaluate participants's submissions in a "best-case scenario": where the acoustic conditions between training and testing are reasonably matched.

4.3. Mixer 6 Speech

The Mixer 6 Speech set (LDC catalog no. LDC2013S03) features a very different but relevant multi-speaker, multi-channel scenario. It consists of 594 distinct native English speakers participating in a total of 1425 sessions, each lasting approximately 45 minutes. Sessions include an interview between an "interviewer" and a "subject" (15 min.), a telephone call (10 min.), and prompt reading. Each session is recorded using 14 microphones of varying styles, placed in various locations around one of 2 different rooms (LDC and HRM).

We make use of the interview and call portions of 450 of the 1425 sessions for which we have collected human-annotated transcriptions. The remaining audio in the 450 sessions, for which no human annotation is available, can be used by the participants, for instance, to support semi/self-supervised training. The 450 sessions are split into `train`, `dev`, and `eval` partitions with no speaker overlap. We only use 13 of the channels in the challenge as the SNR of the 14th channel is too low to be useful (or intelligible).

Human-annotated time-marks and transcripts are available only for the "subject", who is usually the dominant speaker. For the "interviewer" segments, preliminary transcripts were generated by running Whisper [40] large inference on the lapel microphone (`eval`), or via manual annotation (`dev`). Interviewer transcripts were then force-aligned with the audio from the interviewer's lapel microphone (for `dev`) and/or manually corrected (for `eval`) to get the ground-truth annotations. Information about the splits is shown in Table 1. Note that `eval` was

recorded in the (HRM) room; this new room condition tests robustness to shifts in the recording room, while performance on dev represents a more matched scenario. Close-talk channels are not provided to challenge participants for eval.

5. Baseline Description

The CHiME-7 DASR baseline model is shown in Fig. 2. For the main track, it features an *array topology agnostic* speech processing pipeline composed of multi-channel diarization, channel selection, guided source separation (GSS) [41], and monaural ASR.
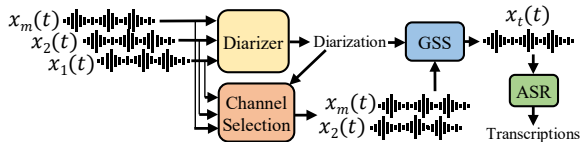
Figure 2: *Block scheme for the CHiME-7 DASR baseline. The diarizer module is omitted in the acoustic robustness sub-track.*

For the optional sub-track, the diarizer component is omitted since participants are provided oracle segmentation. This baseline system, along with the data preparation, is implemented within ESPNet2 [42] as a recipe[5]. We now describe each component in detail.

5.1. Front-End Processing

The baseline system for the CHiME-6 challenge comprised GSS for front-end processing [41], with array topology aware choices such as the use of outer microphones. Since CHiME-7 DASR forbids the a-priori knowledge of array geometry, we instead use automatic channel selection using an envelope variance (EV) approach as proposed in [43]. This measure, while simple, has been found to be effective on CHiME-6 data [44] where it compared favourably with more elaborated data-driven approaches. Channel selection is used to select the most promising subset of microphones for each utterance for later processing via GSS. We use a GPU-accelerated implementation of GSS [45] that allows to significantly speed up the inference time (e.g. up to 300x on CHiME-6 development data with two NVIDIA A100 40GB GPUs).

5.2. Diarization

During preliminary experiments with the Pyannote [46] diarization pipeline[6] on CHiME-6, we found that diarization error rate (DER) between microphones belonging to the same array may vary by up to 10%. This agrees with similar observations made during the CHiME-6 challenge, where teams employed channel fusion methods to subvert these effects [47]. At the same time, low speaker variability in CHiME-6 and the lack of train segmentation for the interviewer segments in Mixer-6 prohibit methods such as EEND-VC [48] and suggest the use of embedding-based approaches, unless one wants to create extensive multi-channel synthetic data.

For these reasons, here we opt for a simple but reasonably effective approach built over the Pyannote [46] diarization system, which is very popular and has been proven to be successful on a wide variety of scenarios. Our approach is depicted in Figure 3. In detail, instead of channel ensembling, we instead
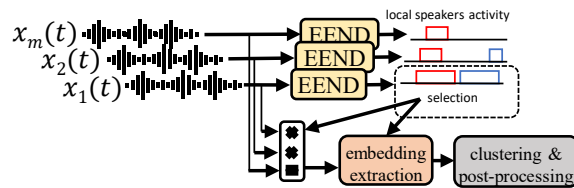
[5]espnet/tree/chime7task1_diar/egs2/chime7_task1/diar_asr1
[6]pyannote/speaker-diarization

Figure 3: *Baseline multi-channel diarization pipeline, built over the Pyannote speaker diarization pipeline.*

leverage the local end-to-end diarization (EEND) [49,50] module in the pipeline [46] for channel selection. The maximum number of local speakers is fixed to 3 and we use 5 s segments. Only this latter is ran over all channels and, assuming it is robust against false alarms, we select the channel which has the most speech activity as the best channel. The rest of the pipeline (including embedding extraction and clustering, which make the bulk of computation) is then run only over this selected microphone channel, leading to a significant speed up versus running the whole pipeline over all channels. This approach has obviously its limitations, e.g. for two concurrent speakers it may be preferable to select a different microphone for each. Despite this limitation, we found it to work very well on the dev set, significantly outperforming all channels ensembling via DOVER-Lap [51] (40% versus 52% DER on CHiME-6 dev set). This said, this simple approach can be easily extended in the future to use the top-K channels instead of just one in the clustering phase.

In the baseline recipe, we fine-tune the local EEND model of the Pyannote pipeline using all CHiME-6 train set (by using all far-field channels only) and use for validation Mixer 6 dev set; only this latter is used because Mixer 6 is the most acoustically different dataset among the three scenarios from CHiME-6, so it made sense to use it as validation, to enforce early-stopping and preventing overfitting. To make the output of the diarization more suitable for ASR we post-process the diarization output by merging segments from the same speaker which are 0.5 seconds apart. The pre-trained model is made available via HuggingFace[7].

5.3. Automatic Speech Recognition

For the baseline ASR, we directly take the model from [31,32], which consists of a hybrid CTC/Attention transformer [52] encoder-decoder ASR model with WavLM-based features and has been proven very effective on noisy-reverberant data. It is trained on the full CHiME-6 train (including the binaural microphones) and the Mixer 6 train. In addition, we used the augmentation scheme as described in the CHiME-6 baseline [1], which leveraged close-talk microphones and external datasets for RIRs and noises, namely SLR26 [53] and MUSAN [54]. We also used GSS-enhanced data obtained from the whole CHiME-6 training set. Both these addition were found critical to reach state-of-the-art performance. The training takes approximately 18 hours using two A100 GPUs on a DGXA100 machine. This model is made available via HuggingFace[8]

[7]popcornell/pyannote-segmentation-chime6-mixer6
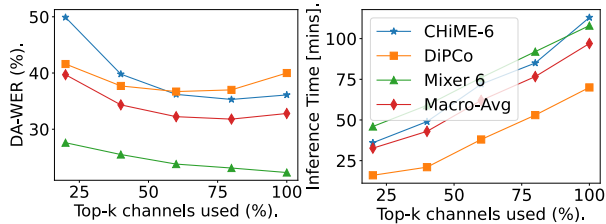[8]popcornell/chime7_task1_asr1_baseline

Figure 4: *Effect of EV-based channel selection with GSS on WER performance (left) and inference time (right), on the* dev *set. We show scenario-wise results and the macro-average.*

6. Results & Discussion

6.1. Front-End Processing

As mentioned in Sec. 5.1, a key difference between the front-end module in CHiME-7 DASR, compared with the CHiME-6 challenge, is the use of automatic channel selection. In Fig. 4, we show WER results and inference time on the dev sets for different selection ratios when the ASR is trained using data containing GSS performed with all channels.

We can observe two trends. First, the time employed to perform selection+GSS appears to be approximately linear with the number of microphones used. Secondly, the error rate generally decreases as the number of microphones used increases. Thus there appears to be a trade-off between computational complexity and performance. In particular, the best performance for CHiME-6 and DiPCo is achieved by using the top 80%, while Mixer 6 will benefit from the use of all microphones. This could be related to the fact that Mixer 6 is less noisy and all channels might contribute significantly to improving the GSS result. In the baseline we always use the top 80% channels for GSS for both training and inference, as this seems to maximize the performance overall.

6.2. Diarization

In Table 2 we report the results achieved by our proposed diarization system in terms of DER and Jaccard error rate (JER) with 250 ms collar. We can see that, overall, the system, despite being fine-tuned only on CHiME-6 data, generalizes well to the other scenarios.

Table 2: *CHiME-7 DASR diarization baseline results. We show scenario-wise results as well as the macro-average.*

Scenario	Dev		Eval	
	DER	JER	DER	JER
CHiME-6	40.0	51.1	56.3	62.5
DiPCo	29.8	41.4	27.9	40.9
Mixer 6	16.6	22.8	9.3	11.0
Macro	28.8	38.5	31.2	38.2

CHiME-6 remains by far the most difficult scenarios among the three for what regards diarization, due to the challenging acoustical conditions and frequent overlapped speech.

6.3. Overall Results

In Table 3 we present DA-WER figures obtained on the three CHiME-7 DASR scenarios, for both sub-track (sub) and main-track (main), together with their macro averages. We compare our baseline model against Whisper [24] large — note that

Whisper cannot be used by participants as it was not included among the allowed pre-trained models (its training data may contain our evaluation). Nevertheless, it serves as an interesting and strong point of comparison.

Table 3: *CHiME-7 DASR sub-track results obtained using the best configuration from development set (80% channels GSS). Whisper results are reported for reference.*

ASR model	Scenario	Dev		Eval	
		DA-WER (%)		DA-WER(%)	
		sub	main	sub	main
Baseline	CHiME-6	32.6	62.4	35.5	77.4
	DiPCo	**33.5**	56.6	36.3	54.7
	Mixer 6	**20.2**	**22.5**	28.6	**33.7**
	Macro	**28.8**	47.2	33.4	55.3
Whisper	CHiME-6	**30.9**	**58.4**	36.6	74.0
	DiPCo	34.5	52.5	35.7	49.7
	Mixer 6	21.2	23.7	**25.2**	35.8
	Macro	**28.8**	44.8	32.5	53.2

Overall, our baseline ASR model compares favorably with Whisper large especially in the acoustic robustness sub-track where the final macro-average DA-WER scores are very close. In the main track the difference is more pronounced, and Whisper appears slightly more robust to segmentation errors, due to its significantly larger training data.

By comparing sub and main columns in Table 3, we can observe that diarization has a big impact on the final recognition score. We discovered that the permutation was always assigned correctly in these results during DA-WER computation, yet, sub-optimal segmentation and mis-attributed sentences degraded considerably the performance, especially on CHiME-6. On the other hand, we can see how both Whisper and the baseline system fail, even with oracle diarization, to reach satisfactory performance suitable for real-world applications, even on the arguably easier Mixer 6 scenario.

7. Conclusions

In this paper, we introduced the CHiME-7 DASR challenge, which builds upon the past CHiME-6 and extends it to multiple scenarios which feature diverse array topologies, number of participants, acoustical conditions and linguistic differences, to test meeting transcription systems cross-domain generalization capability. Other novelties regard the use of a novel DA-WER metric which also accounts implicitly for diarization accuracy and the allowance for pre-trained "foundation" models and popular external datasets. Despite the efforts in developing a solid baseline system, the proposed challenge scenario still poses significant obstacles in achieving accurate transcriptions; this is also demonstrated by results with Whisper large and oracle diarization. Our hope is that participants will devise novel and more effective ways to address this important problem.

8. Acknowledgements

9. References

[1] S. Watanabe, M. Mandel *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *CHiME Workshop*, 2020.

[2] P. Szymański, P. Żelasko *et al.*, "WER we are and WER we think we are," in *Findings of EMNLP*, 2020.

[3] Z. Chen, T. Yoshioka *et al.*, "Continuous speech separation: Dataset and analysis," in *IEEE ICASSP*, 2020.

[4] D. Raj, P. Denisov *et al.*, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *IEEE SLT*, 2021.

[5] S. Araki, N. Ono *et al.*, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer," in *IEEE ICASSP*, 2018.

[6] T. Gburrek, J. Schmalenstroeer *et al.*, "Informed vs. blind beamforming in AD-HOC acoustic sensor networks for meeting transcription," in *IWAENC*, 2022.

[7] S. Cornell, M. Omologo *et al.*, "Overlapped speech detection and speaker counting using distant microphone arrays," *Computer Speech & Language*, vol. 72, 2022.

[8] M. Harper, "The automatic speech recognition in reverberant environments (aspire) challenge," in *IEEE ASRU*, 2015.

[9] J. Carletta, S. Ashby *et al.*, "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*, 2005.

[10] A. Janin, D. Baron *et al.*, "The ICSI meeting corpus," in *IEEE ICASSP*, 2003.

[11] J. Barker, E. Vincent *et al.*, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, 2013.

[12] E. Vincent, J. Barker *et al.*, "The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes," in *IEEE ASRU*, 2013.

[13] J. Barker, R. Marxer *et al.*, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE ASRU*, 2015.

[14] E. Vincent, S. Watanabe *et al.*, "The 4th CHiME speech separation and recognition challenge," 2016.

[15] C. Fox, Y. Liu *et al.*, "The Sheffield wargames corpus," in *InterSpeech*, 2013.

[16] L. Cristoforetti, M. Ravanelli *et al.*, "The DIRHA simulated corpus." in *LREC*, 2014.

[17] C. Richey, M. A. Barrios *et al.*, "Voices obscured in complex environmental settings (VOiCES) corpus," in *InterSpeech*, 2018.

[18] M. Van Segbroeck, A. Zaid *et al.*, "DiPCo – dinner party corpus," *ArXiv*, 2019.

[19] F. Yu, S. Zhang *et al.*, "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *IEEE ICASSP*, 2022.

[20] K. Grauman, A. Westbury *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022.

[21] J. Barker, S. Watanabe *et al.*, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *InterSpeech*, 2018.

[22] V. Panayotov, G. Chen *et al.*, "LibriSpeech: an ASR corpus based on public domain audio books," in *IEEE ICASSP*, 2015.

[23] L. Brandschain, D. Graff *et al.*, "The mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *LREC*, 2010.

[24] A. Radford, J. W. Kim *et al.*, "Robust speech recognition via large-scale weak supervision," *ArXiv*, 2022.

[25] J. Li, V. Lavrukhin *et al.*, "Jasper: An end-to-end convolutional neural acoustic model," in *InterSpeech*, 2019.

[26] S. Kriman, S. Beliaev *et al.*, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *IEEE ICASSP*, 2019.

[27] A. Baevski, Y. Zhou *et al.*, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[28] W.-N. Hsu, B. Bolte *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.

[29] 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, 2022.

[30] S.-w. Yang, P.-H. Chi *et al.*, "SUPERB: Speech processing universal performance benchmark," in *InterSpeech*, 2021.

[31] X. Chang, T. Maekaku *et al.*, "End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation," in *InterSpeech*, 2022.

[32] Y. Masuyama, X. Chang *et al.*, "End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation," 2022.

[33] E. Fonseca, X. Favory *et al.*, "FSD50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, 2021.

[34] N. Kanda, J. Wu *et al.*, "VarArray meets t-SOT: Advancing the state of the art of streaming distant conversational speech recognition," *ArXiv*, 2022.

[35] T. Yoshioka, X. Wang *et al.*, "VarArray: Array-geometry-agnostic continuous speech separation," in *IEEE ICASSP*), 2022.

[36] N. Kanda, J. Wu *et al.*, "Streaming multi-talker ASR with token-level serialized output training," in *InterSpeech*, 2022.

[37] I. H. Toroslu and G. Üçoluk, "Incremental assignment problem," *Information Sciences*, vol. 177, 2007.

[38] J. G. Fiscus, J. Ajot *et al.*, "Multiple dimension Levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech." in *LREC*. Citeseer, 2006, pp. 803–808.

[39] M.-W. C. Jacob Devlin and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[40] A. Radford, J. W. Kim *et al.*, "Robust speech recognition via large-scale weak supervision," *ArXiv*, 2022.

[41] C. Boeddeker, J. Heitkaemper *et al.*, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME5 Workshop*, 2018.

[42] S. Watanabe, T. Hori *et al.*, "ESPnet: End-to-end speech processing toolkit," in *InterSpeech*, 2018.

[43] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, 2014.

[44] S. Cornell, A. Brutti *et al.*, "Learning to rank microphones for distant speech recognition," in *InterSpeech*, 2021.

[45] D. Raj, D. Povey, and S. Khudanpur, "GPU-accelerated guided source separation for meeting transcription," in *InterSpeech*, 2023.

[46] H. Bredin, R. Yin *et al.*, "Pyannote. audio: neural building blocks for speaker diarization," in *IEEE ICASSP*, 2020.

[47] A. Arora, D. Raj *et al.*, "The JHU multi-microphone multi-speaker ASR system for the CHiME-6 challenge," in *The CHiME-6 Workshop*, 2020.

[48] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *IEEE ICASSP*, 2021.

[49] Y. Fujita, N. Kanda *et al.*, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[50] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *InterSpeech 2021*, 2021.

[51] D. Raj, L. P. Garcia-Perera *et al.*, "DOVER-Lap: A method for combining overlap-aware diarization outputs," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 881–888.

[52] A. Vaswani, N. Shazeer *et al.*, "Attention is all you need," in *NIPS*, 2017.

[53] T. Ko, V. Peddinti *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE ICASSP*. IEEE, 2017, pp. 5220–5224.

[54] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *ArXiv*, 2015.