Computer Speech and Language 97 (2026) 101901

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Recent trends in distant conversational speech recognition: A review of CHiME-7 and 8 DASR challenges☆

Samuele Cornell [a,*], Christoph Boeddeker [b], Taejin Park [c], He Huang [c], Desh Raj [d], Matthew Wiesner [e], Yoshiki Masuyama [f], Xuankai Chang [a], Zhong-Qiu Wang [g], Stefano Squartini [h], Paola Garcia [e], Shinji Watanabe [a]

[a] Carnegie Mellon University, USA
[b] Paderborn University, Germany
[c] NVIDIA, USA
[d] Meta, USA
[e] Johns Hopkins University, USA
[f] Mitsubishi Electric Research Laboratories, USA
[g] Southern University of Science and Technology, China
[h] Università Politecnica delle Marche, Italy

ARTICLE INFO

Keywords:
Robust automatic speech recognition
Meeting transcription
Speaker diarization
Microphone array processing
Speech separation
Multi-talker automatic speech recognition

ABSTRACT

The CHiME-7 and 8 distant speech recognition (DASR) challenges focus on multi-channel, generalizable, joint automatic speech recognition (ASR) and diarization of conversational speech. With participation from 9 teams submitting 32 diverse systems, these challenges have contributed to state-of-the-art research in the field. This paper outlines the challenges' design, evaluation metrics, datasets, and baseline systems while analyzing key trends from participant submissions. From this analysis it emerges that: (1) Most participants use end-to-end (e2e) ASR systems, whereas hybrid systems were prevalent in previous CHiME challenges. This transition is mainly due to the availability of robust large-scale pre-trained models, which lowers the data burden for e2e-ASR. (2) Despite recent advances in neural speech separation and enhancement (SSE), all teams still heavily rely on guided source separation, suggesting that current neural SSE techniques are still unable to reliably deal with complex scenarios and different recording setups. (3) All best systems employ diarization refinement via target-speaker diarization techniques. Accurate speaker counting in the first diarization pass is thus crucial to avoid compounding errors and CHiME-8 DASR participants especially focused on this part. (4) Downstream evaluation via meeting summarization can correlate weakly with transcription quality due to the remarkable effectiveness of large-language models in handling errors. On the NOTSOFAR-1 scenario, even systems with over 50% time-constrained minimum permutation WER can perform roughly on par with the most effective ones (around 11%). (5) Despite recent progress, accurately transcribing spontaneous speech in challenging acoustic environments remains difficult, even when using computationally intensive system ensembles.

☆ This article is part of a Special issue entitled: 'Multi-DSR' published in Computer Speech & Language.

* Corresponding author.
E-mail address: cornellsamuele@gmail.com (S. Cornell).

https://doi.org/10.1016/j.csl.2025.101901
Received 30 April 2025; Received in revised form 20 October 2025; Accepted 27 October 2025
Available online 1 November 2025

1. Introduction

Today, current state-of-the-art (SotA) automatic speech recognition (ASR) systems are well known to be capable of obtaining performance on par or superior to humans on several widely used benchmark datasets (Godfrey et al., 1992; Paul and Baker, 1992; Panayotov et al., 2015) featuring close-talk single speaker speech, mainly from telephone or audiobook recordings. However, far-field conversational speech recognition remains an arduous problem. For example, even in the AMI dataset (Carletta et al., 2005), developed in 2005 and featuring an office meeting scenario, SotA results in speaker-attributed word error rate (WER) are still above ∼20% (Kanda et al., 2022b; Cornell et al., 2024b; Park et al., 2024) without oracle diarization.

As such recent work (Szymański et al., 2020) has raised doubts if such widely used benchmark datasets are still adequate today, calling for a more comprehensive approach that can also, importantly, assess generalization capability instead of focusing only on one domain of interest. If benchmark evaluation persistently focuses on a narrow, fixed set of scenarios/domains, it inadvertently promotes techniques that are tailored to such domains, potentially compromising their robustness to domain shifts and their real-world effectiveness. In fact, in practical applications, domain shifts are inevitable due to the inherent unpredictability of real-world conditions and human behavior. Moreover, flexibility is often highly desired, e.g., being able to deal with different deployment platforms and services, recording hardware, and so on. This issue is more pronounced for distant ASR (DASR) for meeting scenarios, where the number of variability factors is significantly higher compared to close-talk ASR. In addition to acoustic conditions (e.g., variations in noise and reverberation), multi-party conversations (Li et al., 2014; Haeb-Umbach et al., 2019) introduce complexities in recording setups (e.g., array configurations and single versus multiple devices), speaker dynamics (e.g., static versus mobile participants), and contextual settings (e.g., informal versus formal environments). These factors are typically uncontrollable and also affect the speaking style and the rate of paralinguistic phenomena such as interruptions, fillers, stutters, repairs, backchannel responses and so on which in turn contribute to overlapped speech. This diversity of conditions, and the presence of overlapped speech, necessitate the need for specialized techniques, complicating the development of DASR systems for multi-speaker scenarios. For example, even reliable segmentation of speech in conversational and meeting scenarios itself presents a challenge. Segmentation errors in these contexts can severely affect ASR performance (Novitasari et al., 2022), a phenomenon often overlooked in "pure" ASR research. Many widely used ASR benchmark datasets in the literature assume oracle segmentation (Ardila et al., 2020; Panayotov et al., 2015; Chen et al., 2021; Wang et al., 2021; Hernandez et al., 2018), leading to overly optimistic result. Furthermore, most applications require speaker-attribution of the words, thus needing diarization together with ASR, adding another layer of significant complexity and difficulty. Despite these challenges, we argue that addressing the complexities of spontaneous "speech-in-the-wild" data also offers unique opportunities. Due to its richness, spontaneous conversational data can arguably be considered the next frontier in speech processing, particularly for the development of better "human-like" conversational agents (Défossez et al., 2024; Fang et al., 2024) that can engage in more nuanced, natural and context-aware communication with possibly multiple speakers. After all, dialog modeling and transcription, much like ASR and text-to-speech (TTS), are closely intertwined problems that can be viewed as complementary and drive progress in both domains (Défossez et al., 2024; Cornell et al., 2024a). Successfully tackling conversational speech opens up novel applications in healthcare (Zhang et al., 2021), government, education, customer service, and beyond, where flexibility and robustness in transcription systems are critical. Moreover, with the rise of large language models (LLMs), new possibilities for speech-enabled machine interactions have emerged, such as meeting summarization and retrieval (Zhong et al., 2021; Fang et al., 2024) to name just two. This paves the way for speech-enabled assistants that can better integrate into dynamic real-world environments.

1.1. Previous challenges and datasets

Over the past 25 years, the emergence of new datasets and challenges has been a catalyst for the advancement of research toward robust ASR. For the reader's convenience, we summarize them in Table 1 alongside some of their "high-level" characteristics.

In fact, at the turn of the millennia, Gaussian mixture model (GMM) hidden Markov model (HMM) ASR technology (Rabiner, 1989; Lee et al., 1990; Gales et al., 2008) reached maturity and inventions such as GMM universal background model (UBM) (Reynolds et al., 2000) opened the way for data-driven diarization. These advances raised interest in the automatic transcription of meeting scenarios captured with far-field microphone arrays. The first decade of the 2000s saw the release of the Santa Barbara Corpus, AMI (Carletta et al., 2005), ICSI (Janin et al., 2003) and CHIL (Mostefa et al., 2007) corpora. All of these feature real-world conversational speech captured with far-field microphone devices. In particular, among these, as mentioned, AMI is still widely used to this day. At the same time, research in the telephone speech domain continued, with the collection of Fisher (Cieri et al., 2004) contributing greatly in this latter direction. During these years, the National Institute of Standards and Technology (NIST) was instrumental in leading robust ASR and diarization research by organizing several Rich Transcription (RT) evaluation campaigns (Garofolo et al., 2002; Fiscus et al., 2007c; Garofolo et al., 2004; Fiscus et al., 2006b, 2007a) that ran from 2002 to 2009 almost every year. These RT evaluation campaigns challenged participants to create robust ASR and diarization systems in different domains, from telephone speech to broadcast news, and also meeting scenarios. Importantly, RT evaluation campaigns established foundations on standard scoring procedures, file formats (such as the rich transcription time-marked, segment time mark, conversation time mark etc.) and tools that are still widely used today. Together with the NIST RT campaigns, the Aurora challenges, which ran from the 2nd (2000) to the 5th (2006), were key in spurring research toward robust ASR during this period. Each challenge explored different aspects and scenarios focusing on single speaker and non-long-form input: hands-free telephone speech (Aurora-2 and 4), car scenarios (Aurora-3) and car and hands-free telephone scenarios together (Aurora-5).

Table 1
Notable robust ASR datasets and challenges since the 2000s. We categorize as "real-world" only datasets and challenges strictly featuring fully real-world recorded data.

	Year	Real world	Long form	Multi speaker	Far field	Multi domain
Santa Barbara (Du Bois et al., 2000)	2000	✓	✓	✓	✓	✓
Aurora2-6 (Hirsch and Pearce, 2000)	2000/06	✗	✗	✗	✓	✓
RT evaluations (Garofolo et al., 2002)	2002/09	✓	✓	✓	✓	✓
ICSI (Janin et al., 2003)	2003	✓	✓	✓	✓	✗
Fisher (Cieri et al., 2004)	2004	✓	✓	✓	✗	✗
AMI (Carletta et al., 2005)	2005	✓	✓	✓	✓	✗
Pascal (Cooke et al., 2010)	2006	✗	✗	✓	✗	✗
CHIL (Mostefa et al., 2007)	2007	✓	✓	✓	✓	✗
CHiME Corpus (Christensen et al., 2010)	2010	✗	✗	✗	✓	✗
Mixer 6 Speech (Brandschain et al., 2010)	2010	✓	✓	✓	✓	✗
CHiME-1 (Barker et al., 2013)	2011	✗	✗	✗	✓	✗
COSINE (Stupakov et al., 2012)	2012	✓	✓	✓	✓	✗
Sheffield Wargames (Fox et al., 2013)	2013	✓	✓	✓	✓	✗
REVERB (Kinoshita et al., 2013)	2013	✗	✗	✗	✓	✗
CHiME-2 (Vincent et al., 2013)	2013	✗	✗	✓	✓	✗
DIRHA (Cristoforetti et al., 2014)	2014	✗	✗	✗	✓	✗
CHiME-3 (Barker et al., 2017)	2015	✗	✗	✓	✗	✗
CHiME-4 (Vincent et al., 2016)	2015	✗	✗	✓	✗	✗
ASpIRE (Harper, 2015)	2015	✓	✓	✗	✓	✗
CHiME-5 (Barker et al., 2018)	2018	✓	✗	✓	✓	✗
VOiCES (Richey et al., 2018)	2018	✗	✗	✗	✓	✗
DiPCo (Van Segbroeck et al., 2019)	2019	✓	✓	✓	✓	✗
CHiME-6 (Watanabe et al., 2020)	2020	✓	✓	✓	✓	✗
Aishell-4 (Fu et al., 2021)	2020	✓	✓	✓	✓	✗
AliMeeting (Yu et al., 2022)	2020	✓	✓	✓	✓	✗
LibriCSS (Chen et al., 2020b)	2020	✗	✓	✓	✓	✗
(Grauman et al., 2022)	2022	✓	✓	✓	✓	✓
MISP (Wang et al., 2023b)	2022	✓	✓	✓	✓	✓
CHiME-7 DASR (Cornell et al., 2023)	2023	✓	✓	✓	✓	✓
CHiME-8 DASR (Cornell et al., 2024c)	2024	✓	✓	✓	✓	✓
CHiME-8 NOTSOFAR-1 (Vinnikov et al., 2024)	2024	✓	✓	✓	✓	✗
CHiME-8 MMCSG (Zmolikova et al., 2024)	2024	✓	✓	✓	✓	✗

Over the next decade, the transition to data-driven approaches accelerated dramatically, driven by the rise of deep learning and the increasing availability of data. Deep neural network (DNN)-based models started to replace traditional statistical models such as GMMs. This shift was particularly evident in automatic speech recognition (ASR) acoustic modeling. By the end of the 2010s, ASR in industry relied mainly on end-to-end frameworks (Prabhavalkar et al., 2023), including transducers (Graves, 2012), connectionist temporal classification (CTC) (Graves et al., 2006), and attention-based approaches (Chorowski et al., 2015), rather than the traditional hybrid DNN-HMM approach inherited from GMM-HMM.

During this period, the CHiME (computational hearing in multisource environments) challenges emerged as a "natural" successor to Aurora challenges played a key role in driving robust ASR research forward. The early CHiME challenges, up to CHiME-4, focused on single-speaker scenarios under controlled conditions with synthetic data. The motivation was simple, as synthetic data allowed for much less labor and cost for collecting data compared with recording and annotating real-world multi-speaker interactions. Moreover, synthetic data offered significant advantages such as precise control over signal-to-noise ratio (SNR) and the ability to disentangle the performance for specific components of transcription pipelines, such as front-end processing and acoustic modeling. CHiME-1 (Barker et al., 2013) and CHiME-2 (Vincent et al., 2013) concentrated on binaural data in domestic environment (Christensen et al., 2010). Then, CHiME-3 (Barker et al., 2015) and CHiME-4 (Vincent et al., 2016) expanded to more complex scenarios featuring multi-channel audio and more various noise environments. Similarly to these previous CHiME challenges, other notable efforts such as the REVERB Challenge (Kinoshita et al., 2013), DIRHA (Cristoforetti et al., 2014), and VOiCES (Richey et al., 2018) datasets also mainly focused on single-speaker scenarios again with fully or partially synthetic data.

The later CHiME-5 (Barker et al., 2018) and CHiME-6 (Watanabe et al., 2020) challenges moved beyond single-speaker data in artificial environments. Compared to CHiME-4 focusing on single-speaker reading speech, CHiME-5 and 6 feature a complex real-world multi-speaker scenario consisting of dinner party conversations recorded with far-field microphones. Another initiative in this direction was the 2015 ASpIRE challenge (Harper, 2015) which was focused on speech recognition in noisy/reverberant environments using recordings from the Mixer 6 corpus (Brandschain et al., 2010).

The focus on real-world data has recently gained momentum since 2020 with the release of even more datasets and challenges emphasizing spontaneous, multiparty conversational speech. This, again, has been driven by recent breakthroughs in self-supervised and weakly supervised training, along with increased availability of computing and training data as well as growing interest from industry as technology matures. The latter is also further fueled by the rapid advancements in LLMs which, as mentioned, enable more applications for conversational speech technologies such as meeting summarization. Prominent examples of this trend are

Ego4D (Grauman et al., 2022), a recently collected dataset of egocentric videos collected through smart glasses, conversations recordings on distant microphone arrays collected for MISP challenges (Wang et al., 2023b), office meeting datasets such as Aishell-4 (Fu et al., 2021) and the Alimeeting dataset (collected for M2MeT ICASSP 2022 challenge (Yu et al., 2022)).

The past two CHiME-7 and 8 challenges are also aligned with this research direction: the proposed CHiME-7 and 8 DASR challenges (Cornell et al., 2023, 2024c) (C7-8DASR), as well as the concurrent "twin" CHiME-8 NOTSOFAR-1 (Vinnikov et al., 2024) and CHiME-8 MMCSG (Zmolikova et al., 2024) challenges, all feature real-world meeting scenarios captured by far-field microphone arrays. A key distinguishing feature of the C7-8DASR challenges lies in their strong emphasis on multi-domain generalization and array-agnostic processing, as participant systems are evaluated across diverse scenarios with drastically different recording setups. This contrasts with CHiME-8 NOTSOFAR-1, which also focuses on meeting transcription but instead assume that knowledge about the deployment domain and the array device is available. The main motivation for having both challenges in CHiME-8 was to compare how "generalist" transcription systems compare with specialized ones and how the former can be adapted once the target domain and device are known. This question, in particular, will be addressed here in Section 5.3.

1.2. Main contributions

This work is an extension of our previous CHiME-7 DASR (C7DASR) (Cornell et al., 2023) and CHiME-8 DASR (C8DASR) (Cornell et al., 2024c) works, where we presented the motivations, datasets, rules, and baseline systems for C7-8DASR challenges. The primary objective of this paper is to provide a unified perspective on these two challenges by analyzing and comparing participant submissions, identifying key trends, and highlighting the most effective techniques across both iterations. Additionally, we draw comparisons with the concurrent CHiME-8 NOTSOFAR-1 challenge which focuses on a specific office-scenario setting with known microphone array configuration. This comparison is possible because CHiME-8 NOTSOFAR-1 was included as one of the four scenarios considered in C8DASR. This paper extends our previous work by offering significant novel contributions in the following areas:

- We analyze, summarize and discuss a total of 32 participant submissions from the C7-8DASR and CHiME-8 NOTSOFAR-1 challenges.
- We compare results across the past two CHiME challenges that focus on meeting transcription: C7DASR and subsequent C8DASR and CHiME-8 NOTSOFAR-1 twin challenges.
- In addition to WER-based metrics and diarization metrics, such as Jaccard error rate (Ryant et al., 2019), we also investigate downstream evaluation on meeting summarization on the NOTSOFAR-1 scenario.

1.3. Main findings

This paper presents several key findings from analyzing the submitted systems:

- Array and domain agnostic systems can achieve competitive performance. Cross comparison between CHiME-8 DASR and the concurrent NOTSOFAR-1 challenge (Section 5.3) revealed that generalizable, array-agnostic transcription systems can perform comparably to systems specifically tailored to known recording setups and deployment domains.
- Transition to end-to-end ASR with large-scale data pre-trained models (Section 4.4). Unlike CHiME-6, which predominantly featured hybrid ASR systems, nearly all CHiME-7/8 participants employ end-to-end ASR frameworks. This transition was enabled by the availability of robust large-scale pre-trained models (e.g., WavLM (Chen et al., 2022), Whisper (Radford et al., 2022)), which significantly reduced the data burden for training competitive end-to-end (e2e) systems.
- DNN-based front-end speech separation methods still struggle with far-field conversational speech. All participating teams relied on guided source separation (GSS) (Boeddeker et al., 2018) and, when a DNN-based front-end was used, it was used together with GSS for initialization or refinement (Section 4.1.1).
- Accurate diarization and, especially, speaker counting are critical. All top-performing systems employed diarization refinement via target-speaker voice activity detection (TS-VAD) techniques (Medennikov et al., 2020b) (Section 4.2). Accurate speaker counting in the initial diarization pass is crucial, as errors propagate catastrophically through the subsequent separation and recognition stages (Section 5.1). The CHiME-8 DASR challenge, in particular, drove innovations in robust speaker counting methods to handle diverse scenarios (Section 4.2.1).
- Weak correlation between transcription accuracy and summarization. Downstream evaluation via LLM-based meeting summarization in the NOTSOFAR-1 scenario (Section 5.3.1) revealed a weak correlation with transcription quality. This demonstrates the remarkable effectiveness of large language models in handling transcription errors and suggests that if precise transcription is not required, end-to-end meeting summarization may be a promising research direction.
- Generalization across scenarios remains difficult. Balancing performance across all scenarios proved challenging (Section 5.1). Systems optimized for CHiME-6 often performed worse on Mixer 6 and vice versa. This trade-off was evident even in the CHiME-8 DASR baseline systems, where re-tuning to accommodate NOTSOFAR-1 degraded performance on other scenarios.
- Multi-channel mechanisms remain, in general, under-explored. Most systems relied on simple ensembling techniques to fuse information across microphones rather than native multi-channel processing (Section 4.3). The effectiveness of sophisticated multi-channel diarization and separation techniques in complex conversational scenarios remains largely unexplored.
- Limited success of test-time adaptation and LLM integration. Several teams explored test-time adaptation (TTA) techniques for both diarization (Section 4.3) and ASR, as well as LLM-based post-processing (Section 4.4.4). However, these approaches yielded only marginal improvements despite significant computational overhead, raising questions about their practical value in real-world applications.

Table 2
Summary of the characteristics for the C7-8DASR challenges core scenarios. As the main focus is on generalization, the scenarios considered are highly diverse. NOTSOFAR-1 was included only in C8DASR.

Scenario	Setting	Number of speakers	Recording setup	Tot. Mics	Duration (min)
CHiME-6	Dinner party	4	6 linear arrays	24	~120 to 150
DiPCo	Dinner party	4	5 circular arrays	35	~33 to 47
Mixer 6 Speech	1-to-1 interview	2	10 heterogeneous devices	10	~25
NOTSOFAR-1	Office meeting	4–8	1 circular array	7	~10

1.4. Outline

This paper is structured as follows: in Section 2 we introduce the C7-8DASR challenges and discuss the motivations and goals of these two challenges in the context of the previous CHiME challenges. In Section 3 we describe in detail the two challenge datasets, evaluation tracks, and baseline systems. Section 4 presents an overview of the submissions of the participants and their design choices for the systems. In Section 5 we present, analyze and discuss the results of the two challenges with the goal of identifying common trends and techniques that appear to be particularly effective. C8DASR systems are also compared and discussed within the "twin" CHiME-8 NOTSOFAR-1 challenge, and meeting summarization as a potential downstream evaluation task is also considered. Finally in Section 6 we draw conclusions, outline the limitations of these two challenges, and propose directions for future research and evaluation campaigns.

2. DASR challenges motivation

The C7-8DASR challenges build on the previous CHiME-6 challenge and thus focus on far-field transcription of long-form meetings through ASR and diarization. These new challenges emphasize generalization as a core principle and introduce several innovations inspired by recent trends in the speech processing field, including the growing availability of pre-trained "foundation models" and open-source datasets.

The main goal is to encourage participants to explore new research directions beyond those tackled in the CHiME-6 challenge while, at the same time, keeping a strong emphasis on practical application scenarios.

2.1. Robustness and generalization

Compared to previous CHiME challenges, C7-8DASR significantly broaden the scope of evaluation, prioritizing generalization as the core principle. The C7-8DASR challenges aim to foster research into transcription systems that are robust to:

- changes in the recording setup/array topologies,
- variable meetings duration and with different number of speakers,
- varying acoustic conditions,
- linguistic and para-linguistic differences between diverse scenarios, such as "colloquial" dinner party conversations, office meetings and interviews.

This emphasis on variability mirrors real-world applications, where the ability to generalize across domains is essential and control on the recording setup (e.g. through proprietary devices) is not always possible as it is inherently less flexible and more expensive.

As such, C7-8DASR challenges "stress-test" the robustness and versatility of the participant's systems on 4 different (3 for C7DASR) datasets that feature spontaneous conversational speech and are highly diverse in recording setup, setting, duration, and number of speakers. These scenarios are CHiME-6, DiPCo, Mixer 6 Speech (MX6) and NOTSOFAR-1 (added in C8DASR). Their "high-level" characteristics are summarized in Table 2, while a detailed description for each of them is in Section 3.1. As can be seen, they are very diverse: on one hand, systems have to deal with multiple arrays and long meetings with 4 participants (CHiME-6); on the other, NOTSOFAR-1 has only a single array, features up to 8 participants, and a very short duration. This inter-scenario variability reflects real-world deployment scenarios and significantly complicates the diarization speaker counting task.

2.2. Availability of large-scale pre-trained models

In the last 4 years, there has been growing research towards large-scale pre-trained "foundation" speech models, either trained via self-supervised learning (SSL) or weakly supervised objectives. These are increasingly available to speech researchers and have demonstrated exceptional effectiveness in numerous downstream applications (Yang et al., 2021). Notable examples are Whisper (Radford et al., 2022), OWSM (Peng et al., 2023), WavLM (Chen et al., 2022), HuBERT (Hsu et al., 2021) and wav2vec 2.0 (Baevski et al., 2020) to name just a few. Their success is so indisputable that in many speech processing applications nowadays the de facto standard approach is to use these as feature extractors or, more directly, via fine-tuning. This procedure allows to implicitly leverage, in a cost-effective way, their massive pre-training data, thus leading to a more robust system.

For monaural, pre-segmented, single-talker speech data, integrating these models is relatively straightforward. However, in the multi-channel, multi-speaker, long-form setting, fully exploiting their potential remains an open question.

Another significant advantage of allowing pre-trained models is the lower entry barrier for challenge participants, particularly those with limited resources. By relying on pre-trained models, training becomes faster and more efficient, eliminating the need for many training epochs. In many cases, modest amounts of fine-tuning data are sufficient along with few (1 or 2 epochs) training iterations on such data, making participation in the DASR challenges more accessible to a broader community compared to CHiME-6. For example, as also explained later in Section 2.5 the C7DASR baseline system is much faster to train compared to the CHiME-6 one. The full list of allowed pre-trained models is available on the challenge website.[1]

2.3. Leveraging open-source external datasets

CHiME-5 and 6 challenges had quite strict restrictions on the training material allowed, which prevented research in important directions. In C7-8DASR, the amount and diversity of allowed external data sources[2] are significantly expanded to include not only popular open-source speech datasets such as LibriSpeech (Panayotov et al., 2015) but also noise (Fonseca et al., 2021; Snyder et al., 2015) and room impulse responses (Jeub et al., 2009) datasets. This opens up the possibility of exploring also the use of techniques that heavily rely on synthetic data for pre-training such as neural speech separation and enhancement (SSE), neural speaker extraction (Žmolíková et al., 2019; Boeddeker et al., 2024), or end-to-end neural diarization (EEND) (Fujita et al., 2019b). C7-8DASR succeeded partially in this direction, as these techniques were explored by some participants, but with marginal improvements over guided source separation (GSS) (Boeddeker et al., 2018). Details are reported in Section 3.

2.4. More comprehensive evaluation for meeting transcription

Another limitation of the CHiME-6 challenge is that it used the concatenated minimum permutation word error rate (cpWER) (Watanabe et al., 2020) as the ranking metric. cpWER does not pose explicit requirements for participants to also produce reasonable utterance-level segmentation, as it only cares about speaker attribution. On the other hand, having reasonable segmentation at the utterance level is desirable in real-world applications since, among other things, it allows for quick audio retrieval for verification of the transcription. As such, in C7DASR we proposed a new metric called diarization-attributed WER (DA-WER) (Cornell et al., 2023) that computed the concatenated WER by using the speaker-assignment permutation that minimizes the diarization error rate (DER) rather than, as in cpWER, the permutation that directly minimizes the WER. Since C7DASR featured three scenarios, each with different duration and number of recordings and the focus was on probing robustness and generalization, the final ranking was given by macro-averaging DA-WER across all scenarios. DA-WER encouraged participants to also produce reasonable segmentation at the utterance level, while at the same time, in the C7DASR description paper (Cornell et al., 2023), we raised awareness of the fact that this was not an optimal solution and better metrics were needed.

Our call was answered the same year and, in the following C8DASR, we instead adopted the time-constrained cpWER (tcpWER), recently proposed in the Meeteval toolkit (von Neumann et al., 2023). tcpWER incorporates temporal constraints into Levenshtein distance computation by restricting substitutions or correct mappings between reference and hypothesis words to those within a specified time collar (e.g., 5 s). Utterances are segmented into word-level units, with segment lengths proportional to word length. The collar is set wide enough to account for potential inaccuracies in the word boundary estimation (in C8DASR, a generous collar of 5 s is employed). This temporal constraint makes tcpWER much more sensitive to segmentation errors than DA-WER. An alternative time-constrained measure is Asclite (Fiscus et al., 2006a), which includes a time penalty but disregards speaker attribution, making it unsuitable for certain applications. For this reason, and to be able to cross-compare between the two challenges, in Sections Section 5 we will use tcpWER instead of DA-WER even when analyzing C7DASR submissions.

2.5. Accessibility

Building a SotA diarization and robust ASR pipeline for meeting transcription is an inherently difficult task that requires a team to have expertise in different speech processing areas. As such, the entry bar for successfully participating in the CHiME-5 and CHiME-6 challenges was quite high. For example, running successfully the Kaldi (Povey et al., 2011) baseline of the CHiME-6 challenge could easily take more than a week if no large-scale computing infrastructure with many CPUs is available. As said, in C7-8DASR, at least this was reduced due to the possibility of leveraging large-scale pre-trained models. For example, the C7-8DASR ESPnet (Watanabe et al., 2018) baseline ASR, which uses WavLM as a feature extractor, can be trained in less than three days on 2 NVIDIA RTX 3090 GPUs. Moreover, both baselines use a GPU-accelerated implementation (Raj et al., 2023) of GSS that allows, in most computing infrastructures, to massively speed up the inference time compared to the CHiME-6 Kaldi baseline and the official implementation of GSS. In C8DASR, we also provided another baseline system based on the C7DASR NVIDIA NeMo team submission (Park et al., 2023c), which was built instead with the NeMo toolkit (Kuchaiev et al., 2019). This baseline also exploits pre-trained models for ASR and a GPU-based implementation of GSS. By offering two baseline systems that utilize two widely adopted speech processing toolkits, our goal was to make participation more accessible to a broader range of researchers.

[1] https://www.chimechallenge.org/challenges/chime8/task1/rules#external_models.

[2] https://www.chimechallenge.org/challenges/chime8/task1/data#external_datasets.

Table 3
C7-8DASR core datasets statistics overview. 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 (hh:mm)	Utts	Spk.	Sess.	sil (%)	1-spk (%)	ovl (%)
CHiME-6	train	40:05	79 967	32	16	22.6	52.7	24.7
	dev	4:27	7 437	8	2	13.1	43.4	43.5
	eval	5:12	11 028	8	2	21.3	52.0	26.7
DiPCo	train	1:12	1 379	8	3	8.3	72.0	19.6
	dev	1:31	2 294	8	2	7.4	61.9	30.6
	eval	2:36	3 405	16	5	9.4	65.7	24.9
Mixer 6	train calls	36:09	27 280	81	243	–	–	–
	train intv	26:57	29 893	77	189	–	–	–
	train	6:13	3 785	19	24	8.6	73.3	18.0
	dev	8:56	5 903	22	35	8.4	72.1	19.5
	eval	5:45	5 115	18	23	2.4	83.6	13.9
NOTSOFAR-1	train	14:43	101 301	14	379	6.0	62.3	31.7
	train_sc	53:43	139 913	14	526	5.9	62.4	31.7
	dev	13:25	24 238	11	130	15.6	67.7	16.7
	eval	16:29	38 662	12	160	5.6	64.7	29.6

Significant work was also done to make data preparation easier. C7-8DASR feature different core scenarios, each with its own annotation format and directory structure, potentially adding significant additional workload for the participants to parse and prepare each dataset. Thus, in C7DASR we provided a script that prepared all the data from the different scenarios to be available in a single common format which follows the convention of CHiME-6 and DiPCo. Expanding on this, in C8DASR, we developed a dedicated toolkit: `chime-utils`,[3] designed to streamline the process even more. With just a single line of code, `chime-utils` can automatically download and prepare CHiME-6, DiPCo, MX6 and NOTSOFAR-1 datasets. Additionally, `chime-utils` offers a suite of tools to facilitate integration with popular speech processing toolkits, including ESPnet (Watanabe et al., 2018), Kaldi (Povey et al., 2011), NeMo (Kuchaiev et al., 2019) and k2 (Povey, 2020) via Lhotse (Żelasko et al., 2021). It also includes scoring scripts that seamlessly interface with Meeteval (von Neumann et al., 2023) and Pyannote (Bredin et al., 2020) as well as various utilities including datasets statistics computation (Table 3 was produced using `chime-utils`). This effort aims to benefit the broader speech processing community beyond the CHiME challenges by making datasets like CHiME-6, DiPCo, MX6 and NOTSOFAR-1 more accessible and easier to experiment with.

3. DASR challenges description

3.1. Core scenarios analysis

In the following Sections 3.1.1–3.1.4, we describe the C7-8DASR core scenarios in detail, while here we discuss their high-level characteristics. In Table 3 we report a summary of their statistics including the percentage of overlapped speech, silence, and single-speaker speech. These statistics are computed by using the manual annotated ground-truth utterance-level segments. It is important to note that the characteristics between `dev` and `eval` splits can differ substantially for some scenarios, which may affect system development and generalization. For CHiME-6, the `dev` split has significantly higher overlapped speech ratio (43.5%) compared to `eval` (26.7%), while silence is lower (13.1% vs. 21.3%). These differences stem from the fact that different dinner party sessions can have very different interaction dynamics. For NOTSOFAR-1, the `dev` split has notably higher silence ratio (15.6%) and lower overlapped speech (16.7%) compared to `eval` (5.6% and 29.6% respectively). These distributional differences can affect hyperparameter tuning, but they also reflect real-world deployment challenges where conversational dynamics vary unpredictably across sessions and environments. In this sense, the `eval` set serves as a proxy for potentially unseen deployment scenarios, testing system robustness.

In Fig. 1 we report, for each scenario, the distribution for the mean duration of interpausal units (IPU), turns, pauses, gaps, interruptions and backchannels. These turn-taking events are defined in Nguyen et al. (2023, Figure 3) as:

- IPUs: single utterances i.e. continuous speech segments from a single speaker without significant internal pauses ($\leq$0.5 s).
- Pauses: pauses (>0.5 s) that occur within single utterances belonging to the same speaker.
- Gaps: pauses that instead occur within single utterances belonging to different speakers.
- Interruptions: overlapped speech where a second speaker begins talking near the conclusion of the first speaker's utterance.
- Backchannels: usually brief overlapping vocalizations (e.g., "mm-hmm", "yeah") produced by a second speaker within the primary speaker's utterance.
- Turns: span of contiguous utterances (with pauses in between) from the same speaker without interruptions.

[3] Available: https://github.com/chimechallenge/chime-utils.

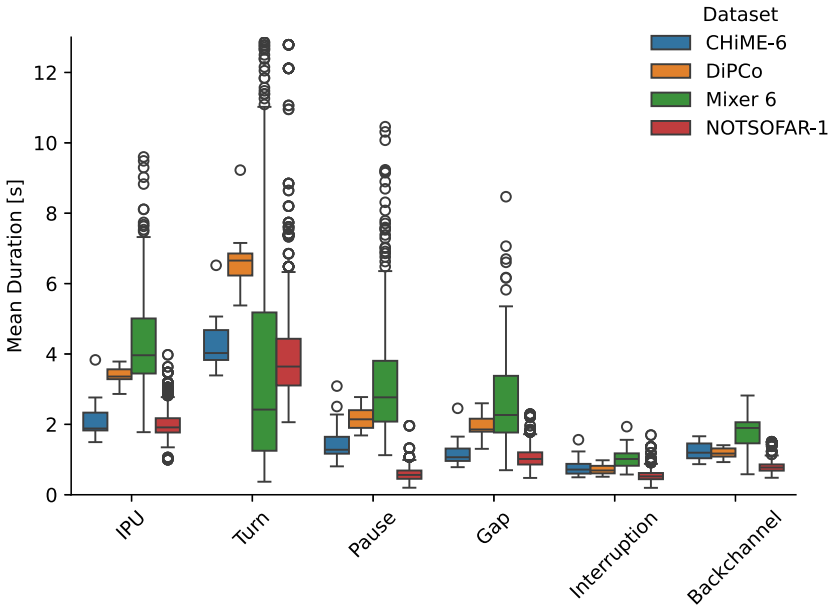

Fig. 1. Mean duration distribution of turn taking events for each C7-8DASR scenario as obtained on the whole `train`, `dev` and `eval` splits.

Again, we used the ground-truth annotation for each dataset and computed these metrics for each session, considering all `train`, `dev` and `eval` splits together. These metrics are complementary with what is observed in Table 3 and offer a different perspective. All 4 scenarios exhibit characteristics typical of spontaneous conversational speech, such as utterances on average around 2 to 4 s long and the presence of short interruptions and backchannel responses. Despite being short, these occur frequently and thus lead to a non-negligible percentage of total speech duration being overlapped as observed in Table 3. NOTSOFAR-1 and CHiME-6 tend to have very similar statistics across all event types, while among all scenarios, Mixer 6 is the one that exhibits the largest variance in gaps, turns, IPUs and pause durations. This is largely due to the fact that it consists of only two speaker interactions and thus there is more opportunity for a single speaker to produce longer speech segments without interruptions. Likewise, longer pauses can be obtained when having fewer speakers.

In Fig. 2 we instead analyze the signal-level amount of noise and interfering speech for each scenario. In detail, we report, for each scenario, the distribution of different signal-to-distortion ratio (SDR) (Vincent et al., 2006) statistics as computed for each utterance across all `train`, `dev` and `eval` splits. In detail, we used ground-truth diarization to compute the SDR for each utterance from each far-field device with respect to the close-talk microphone of the speaker of interest. Then, we compute for each utterance the maximum, the minimum, and the average SDR obtained across all microphones available and plot statistics over all each of the 4 datasets. SDR was computed across the whole utterance. Since datasets such as DiPCo present significant synchronization issues between far-field and close-talk signals, we used a filter with 4096 taps (instead of the default 512) for SDR. As this modification alone was insufficient to address the synchronization problems, we also computed SDR for different time-shifts of the reference close-talk signal. Specifically, we considered offsets of −8192, −6144, −4096, −2048, 0, +2048, +4096, +6144, and +8192 samples. The final SDR value for each utterance was determined by selecting the maximum SDR value obtained across these nine offset positions.

The 4 scenarios exhibit distinctly different distributions for these metrics. For example, DiPCo and NOTSOFAR-1 shows relatively consistent SDR across all devices (with close *min*, *max*, and *mean* SDR values across devices). On the other hand, for CHiME-6, being multi-room and multi-device has instead a significant disparity between *max* and *min* SDR obtainable for each utterance. This suggests that for most utterances, certain microphones or microphone subsets capture the speaker of interest with reduced interference and noise. Mixer 6 also presents a channel selection challenge, but with an even more pronounced gap of approximately 10 dB. This substantial difference comes from the varied placement of microphone devices. Some are positioned at considerable distances from speakers and may be relatively near noise sources such as ventilation systems.

In general, the mean SDR values in Fig. 2 and crucially their significant variance (observed for all scenarios in Fig. 2) are good indicators of the difficulty of these challenges and, more broadly, far-field multi-speaker ASR. The significant difference in channel quality emphasizes the importance of the research directions that DASR challenges aim to foster. Only a robust system capable of operating effectively across diverse acoustic environments and microphone configurations can tackle this issue, whereas ad hoc systems will fail.

3.1.1. CHiME-6

CHiME-6 (Watanabe et al., 2020) consists of recording "sessions" of different dinner parties between friends. Each dinner party is attended by 4 participants and takes place each time in a different home environment. Participants can move freely in the kitchen, living, and dining rooms. Thus, at a particular time, they may be scattered across different rooms (e.g., kitchen and living room).

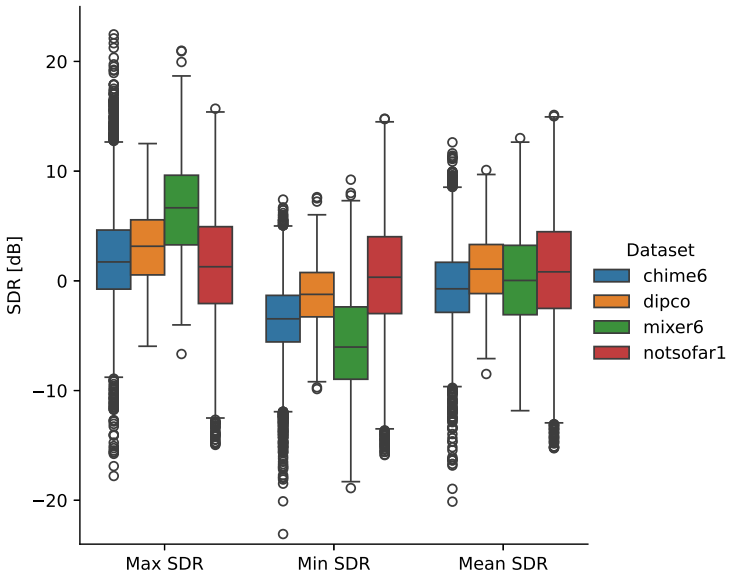

Fig. 2. Distribution of mean, maximum and minimum SDR as obtained across microphones for each utterance. We report statistics for each of the 4 scenarios separately.

In the home environment, there are also 6 Kinect array devices with 4 microphones each to capture the speech of the participants. Close-talk microphones are also made available for annotation and training purposes. However, these close-talk recordings still contain significant noise and cross-talk. Due to the particular "casual" dinner party setting, CHiME-6 is characterized by highly informal conversations with fast turn-taking and high environmental noise, for example, cooking or dining together.

It is worth mentioning that CHiME-6 and CHiME-5 share the exact same data, but with one important difference. CHiME-6 addresses the severe inter-array synchronization issue of CHiME-5 data. This synchronization issue was due to the compounding effect of packet losses and clock drift and led to inter-array signals misalignment of even several seconds in some instances. In CHiME-6 this misalignment is reduced to a few thousand samples on average as the data are resynchronized using the video from each Kinect (available only to CHiME-6 organizers). However, this misalignment is still quite significant as it amounts on average to several thousands of samples, thus posing a not-so-trivial challenge to the participants.

In particular, in the C7DASR and C8DASR challenges, we provided universal evaluation map (UEM) files to correct an issue of the previous CHiME-6 challenge. In CHiME-6, in each session, the very first minute was included by mistake in the evaluation. This contained a speaker enrollment section that was not annotated, but was scored nonetheless, thus leading to an overestimation of insertion errors.

CHiME-6 consists of a total of 24 recording sessions: 16 `train` (training), 2 `dev` (development), and 2 `eval` (evaluation). These three splits have no speakers in common, and each session is about 2 to 2 and a half hours long. In C7DASR, two training sessions (`S19` and `S20`) with no overlapping speaker-ids were moved to evaluation to increase the amount of evaluation data. C8DASR returns to the original CHiME-6 split as we did not observe, in C7DASR, this expansion providing any significant additional insight. Thus, in the remainder of this paper, we consider the `eval` split of CHiME-6 and C8DASR. This includes the results presented in Section 5 for the CHiME-6 scenario. Importantly, such choice allows us to consistently compare systems in CHiME-6, C7DASR, and C8DASR challenges.

3.1.2. Dinner Party Corpus (DiPCo)

The dinner party corpus (DiPCo) (Van Segbroeck et al., 2019) is directly inspired by CHiME-5/6 and, as the name implies, also features a dinner party scenario between 4 participants. Compared to CHiME-5/6, where dinner party takes place in a home environment with possibly multiple rooms, in DiPCo all recording sessions take place in a single room shared between all sessions. The participants' speech is captured using 5 far-field devices, each with a 7-mic circular array (6+1 microphone in the center) scattered throughout the room. Again, closed-talk on-speaker lapel microphone signals are also provided for annotations and training purposes. Compared to the binaural microphones used in CHiME-5/6 these have significantly less cross-talk and noise. However, they present severe misalignment with far-field microphone signals, with offsets exceeding 1000 samples. This misalignment is also often noncausal, meaning the close-talk signal can precede the far-field one.

DiPCo consists of 10 sessions originally partitioned into 5 for `dev` and 5 for `eval` without overlapping speaker-ids but, as said, with the same room and microphone placement shared across all recordings. C7DASR keeps this original split, which lacks an adaptation/training set. As a result, during C7DASR many participants used part of the `dev` to adapt/fine-tune their systems. As a

result, in C8DASR we made an explicit `train` split, also made by using part of the original `dev` set, again splitting to ensure that there are no overlapping speakers. Since the same room and mic placement is shared across all recordings, in some way, DiPCo can be considered a "best-case" scenario when it is possible to fine-tune/adapt a system with respect to a specific environment. This seldom happens for real-world applications, but it is reasonably possible in some domains (e.g. in-vehicle applications).

Compared to CHiME-6, DiPCo is characterized by arguably less noise and less overlapped speech (as is evident in Table 3) because the setting is more formal and the activities of the participants are more limited (e.g., no cooking). Moreover, it does not present the synchronization issues of CHiME-6, all the signals from all the arrays are sample-synchronized.

3.1.3. Mixer 6 speech (MX6)

The Mixer 6 Speech corpus (MX6) (Brandschain et al., 2010; Cornell et al., 2023) consists of 1425 recording sessions carried out between 2009 and 2010 among 594 native English speakers. Each recording session consists of mostly two-party conversations between an interviewer and a subject, and is divided into 4 parts: (i) entry questions, (ii) an interview, (iii) transcript reading, and (iv) a phone call.

The speech is captured by 10 different heterogeneous far-field recording devices scattered across the room where the session takes place. These include arrays (e.g. an Acoustimagic Array, with a single signal provided after internal processing), but also commercial single-microphone devices (RODE NT6) and even multi-purpose media recording devices (Panasonic Camcorder). As such, it presents a significantly different recording setup from CHiME-6 and DiPCo. Close-talk microphones are also available for the subject and interviewer. However, the subject lapel microphone has significant crosstalk in some sessions while a headset microphone is only available for the interviewer. The recording sessions are conducted in two different rooms called "LDC" and "HRM". In C7-8DASR, we consider only the interview portion for evaluation as it is the only multi-talker conversational speech portion in each session. The average duration is ∼14 min and is carried out by the interviewer in such a way as to elicit an informal conversation.

Note that the original MX6 release (Brandschain et al., 2010) was not transcribed other than a small section of read speech spoken by a single speaker. A portion (∼8.9 h) of the telephone call transcripts were released as part of the ASpIRE Challenge (Harper, 2015). A significant contribution of C7DASR, led by Matthew Wiesner and Desh Raj, was to produce annotations for 450 interview portions of the total 1425 conversations in order to obtain at least `dev` and `eval` splits with full annotation of respectively 41 and 23 sessions. The split is performed using the NACHOS toolkit[4] to ensure that there is no overlap between the interviewer, the subject, or the room between the two sets. In C8DASR, the development annotation is divided into `train` and `dev` (see Table 3) portion to provide participants with a small annotated training set for adaptation purposes (same rationale as done for DiPCo).

The annotation for the interviewer was obtained semi-automatically using Whisper (Radford et al., 2022) `large` on the close-talk interviewer lapel microphone followed by forced alignment using Kaldi (Povey et al., 2011) toolkit. Subsequently, this annotation is manually checked and corrected. Note that full conversation annotation for the remaining 975 sessions is still partial, with only the subject speech available. This data, together with the call portion, is made available for participants as an additional two training sets: `train_calls` and `train_intv`. Note that in Table 3 silence and overlapped speech statistics for these splits are not reported as the annotation is partial.

3.1.4. NOTSOFAR-1

The natural office talkers in settings of far-field audio recordings (NOTSOFAR-1) (Vinnikov et al., 2024) introduced in the CHiME-8 NOTSOFAR-1 challenge consists in 315 recordings of office meetings. Each meeting is very short (∼6 min) and is attended by 4 up to 8 speakers. The conversation is mediated by a professional actor whose job is to guide it around a certain topic. The same meeting is captured by up to 7 commercially available far-field array devices. These include 4 tabletop circular array devices with 7 microphones each and 3 linear array devices. For the latter, however, only monaural signals after in-device acoustic front-end processing (AFE) is made available. This presents some issues as the AFE pipeline could suppress some speakers, but nonetheless it is also a use-case that can happen in practical scenarios. Note that NOTSOFAR-1 data is shared between CHiME-8 DASR and CHiME-8 NOTSOFAR-1 challenges, and these single channel data was originally conceived for NOTSOFAR-1 challenge single channel track. In CHiME-8 DASR, the single-channel recordings are available only for training purposes (`train_sc` in Table 3), as the focus is strictly on multi-channel. Importantly, while each meeting is captured by many devices, in NOTSOFAR-1 the signals captured by each device are considered as a single, independent "session" in order to constrain participants to focus also on single-array application scenarios, which are arguably the most common. As such, the total number of sessions, in Table 3, is 1195 despite having only 315 unique meetings.

Again, as with the other corpora, on-person close-talk headset devices with low cross talk are made available for all `train` and `dev` splits. NOTSOFAR-1 shares similarities with AMI, but distinguishes itself by featuring more speakers and a more dynamic conversational setting. Transcriptions are of high quality, with word-level alignment provided, as well as useful metadata such as special tags for filler words. Transcriptions are obtained using a multi-judge system by using two different transcribers without any automatic aid, in order to reduce the risk of introducing bias in the annotation procedure.

[4] https://github.com/m-wiesner/nachos.

3.2. Rules

The main rationale of the C7-8DASR rules is to prevent specialized approaches that, e.g., use domain identification to tackle one scenario at a time or leverage a-priori information such array geometry and thus "by design" are limited in their generalization capability. A truly generalist, versatile system can almost always be fine-tuned or improved by exploiting some domain-specific knowledge (e.g. array geometry), while the contrary is far more challenging to realize. As such, participants are forbidden to use manual or automatic domain identification or any information on the recording setup (included the total number of microphones) or number of speakers. They need to produce a single array-topology agnostic system for all scenarios.

For training purposes, only data from the `train` splits of the core scenarios can be used together with the external data sources allowed. We encouraged participants to propose additional external pre-trained models and data sources during the first month of the challenge. Up to 8 pre-trained models and 5 external data sources as proposed by participants were added during C7DASR. A detailed description of the rules is available on the challenge website.[5]

3.3. Evaluation tracks

Since C7-8DASR challenges are designed to be a continuation of the CHiME-6 challenge, the evaluation setup is also inherited. Participants must perform joint ASR and diarization of different meeting scenarios using far-field recording devices and produce a transcription with speaker attribution and segmentation at the utterance level. As CHiME-6, C7DASR featured another optional "acoustic robustness" track, where oracle diarization information could be used. The motivation for such a track was to probe to what extent the error was due to diarization or, instead, to other components such as the ASR system and/or the particular acoustic front-end separation technique employed. In C8DASR, this latter optional track was removed, in order to keep the evaluation data fully blind for the newly added NOTSOFAR-1 scenario. Another important difference compared to CHiME-6 is that in all tracks no restriction is made on the language model used (LM), provided that it is trained only on the allowable training datasets. This was motivated by the fact that the CHiME-6-constrained LM track was too restrictive and practically forced many participants to adopt hybrid deep neural network (DNN)-HMM ASR models. In C8DASR, an additional, optional track was proposed that instead allowed also the use of some pre-trained external LLMs. This direction was motivated by recent work that found that LLMs are effective, in particular, for diarization post-processing (Park et al., 2023a; Wang et al., 2024b). However, this track received limited participation during the challenge, with only one team submitting and reporting very marginal improvements compared to the main track. On the other hand, in a recent work (Ogawa et al., 2024), it is shown how Llama 2 (Touvron et al., 2023) can be used effectively to improve transcription results on the C7DASR challenge. The authors found that by refining the ASR model's N-best list, Llama 2 can significantly improve results, primarily due to its ability to model long-context, inter-utterances dependencies.

Additionally, in the C8DASR and the CHiME-8 NOTSOFAR-1 challenges, a "jury award special mention" was introduced to recognize teams that developed efficient and practically viable transcription systems. This initiative aimed to encourage research into approaches that did not rely heavily on ensembles, iterative inference schemes, or test-time adaptation. Although this track also received limited interest, one team succeeded in creating a more practical system with excellent results. These findings will be discussed further in Section 5.

3.3.1. Annotation and text normalization

Participants' systems have to produce, for each core scenario, a JSON SEGment-wise Long-form Speech Transcription annotation (segLST) (von Neumann et al., 2023) file. An short example of the content of this file for the CHiME-6 scenario is given in Listing 1.

Listing 1: JSON SEGment-wise Long-form Speech Transcription annotation (segLST) for two utterances. Each utterance is a dictionary with several attributes.

```
{end_time":   76.340,
 start_time": 73.500,
 words":      let's do lunch",
 speaker":    P08",
 session_id": S02"},
{end_time":   76.560,
 start_time": 75.300,
 words":      okay",
 speaker":    P02",
 session_id": S02"}
```

For each transcribed utterance, the start, end, and an arbitrary (but consistent through the meeting) speaker-id tag must be inferred together with the words. The `session_id` is known and corresponds to the meeting session. The segLST format is inherited from the CHiME-5 challenge directly, but it differs from the fact that `end_time` and `start_time` are in seconds (instead of the `hh:mm:ss` format used in CHiME-5/6) to make the annotation easier to use.

[5] See https://www.chimechallenge.org/challenges/chime8/task1/rules.

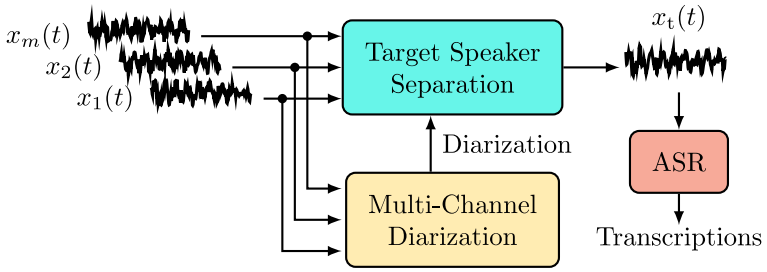

Fig. 3. ESPnet and NeMo baseline systems high-level overview. This same scheme was adopted by almost all C7-8DASR participants and top performing systems in the "twin" CHiME-8 NOTSOFAR-1 challenge.

Note that this format is also used for the annotation of the C7-8DASR core scenarios `train` and `dev` splits. As stated in Section 2.5, the annotation and the directory structure are made consistent across all scenarios to make the data easier to parse and prepare. For the `train` and `dev` split, additional entries may be present, such as `location` (for CHiME-6) or `word_alignment` (for NOTSOFAR-1) depending on the metadata available for each dataset. In detail, for `train` and `dev` splits, we make available to participants two separate JSON segLST annotations: one containing all the original dataset metadata and without text normalization, and another one, which has the same exact format (with only the base entries as in Listing 1) participants need to produce in inference and has text normalization applied.

Text normalization differs between C7DASR and C8DASR. C7DASR builds on CHiME-6 text normalization: punctuation is removed, and all characters are converted to lower case. In addition, nonverbal speech sounds such as "uhm", "umh" etc. are assigned to "hmmm" to be consistent across all datasets. In C8DASR, text normalization is more sophisticated and instead builds on the Whisper text normalization pipeline, but with crucial modifications. First, it was modified to be idempotent. The original Whisper text normalization gives inconsistent results if applied more than once on the same text. We argue that while this can be acceptable for data preprocessing, it is not ideal for scoring: text normalization should be a straightforward and well-defined mapping that can, in principle, also be applied manually. Secondly, we remove number normalization (i.e. "two dollars" will not be converted to 2$) so that the end result is more verbatim. Compared to C7DASR normalization, nonverbal speech sounds such as "uhm", "uhhh", "ah" are completely removed, and common abbreviations are expanded (e.g. "goin" to going).[6] This procedure ensures that the task is well defined: the text is more consistent among the scenarios. Moreover, it avoids biasing the final results by filler sounds ("uhm", "uhhh" etc.) which otherwise would be among the most common words. This same text normalization was adopted by the concurrent CHiME-8 NOTSOFAR-1 challenge, as the goal was to be able to compare systems across the two.

For these reasons, in the following sections, we will report all results (including those of C7DASR) always using C8DASR text normalization when scoring participant submissions.

3.4. Baseline systems

In C7DASR, as mentioned, a baseline system implemented with ESPnet (Watanabe et al., 2018) was provided.[7] In C8DASR, an additional baseline was added,[8] derived from the NVIDIA NeMo team submission to C7DASR. In C8DASR, the ESPnet baseline remained largely unchanged[9] except for some hyperparameters for the diarization component which were tuned to deal with the new NOTSOFAR-1 scenario.

The two baselines differ mainly in the diarization technique used. Both follow the pipeline illustrated in Fig. 3 consisting of multi-channel diarization followed by target speaker separation/extraction (TSE) via GSS (Boeddeker et al., 2018) and finally ASR on each separated utterance. This approach aligns with the strategies employed by the winning teams in the previous two CHiME-5 and CHiME-6 challenges (Arora et al., 2020; Chen et al., 2020a; Medennikov et al., 2020a), where GSS has been the de-facto standard approach for front-end speech separation due to its effectiveness.

3.4.1. Target speaker separation

Both baselines employ the same TSE pipeline consisting of channel selection using the envelope variance method (EV) (Wolf and Nadeu, 2014) followed by GSS.

The inclusion of EV-based channel selection was a novel feature introduced in the C7DASR baseline. The motivation for using channel selection was twofold. First, it allows for accelerating the inference process and/or saving GPU memory, as the particular GPU-based GSS implementation scales quadratically in the GPU memory requirements with the number of channels used. Secondly, it also allows for a marginal improvement in performance because microphones with lower quality signal are excluded. In our

[6] Some examples are available here: https://github.com/chimechallenge/chime-utils/blob/main/tests/test_normalizer.py.
[7] Available: https://github.com/espnet/espnet/blob/master/egs2/chime7_task1.
[8] Available: https://github.com/chimechallenge/C8DASR-Baseline-NeMo.
[9] Available: https://github.com/espnet/espnet/blob/master/egs2/chime8_task1.

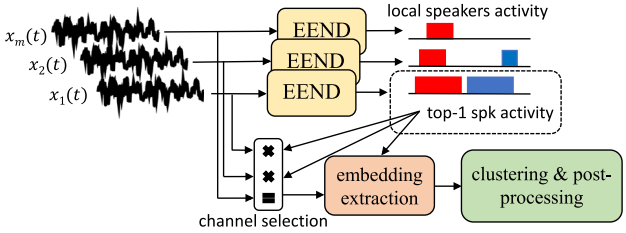

Fig. 4. ESPnet baseline diarization pipeline scheme.

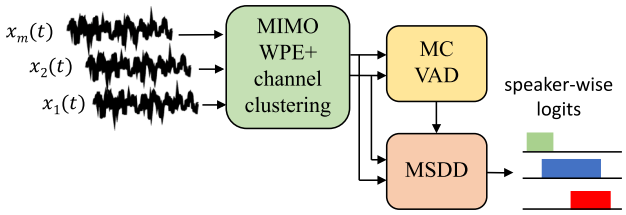

Fig. 5. NeMo baseline diarization pipeline scheme.

previous work (Cornell et al., 2023) introducing the C7DASR baseline, we found that there is an optimal trade-off between these two effects that occurs for C7DASR, when 80% of the microphones are retained. The choice of using EV instead of other channel selection measures (Wolf and Nadeu, 2014; Cornell et al., 2021) was dictated mainly because it is a simple and computationally efficient measure which has been found to be quite effective in the CHiME-5/6 scenario in a recent work (Cornell et al., 2021), even compared to more sophisticated data-driven methods or ASR-based measures.

GSS (Boeddeker et al., 2018) employs a spatial mixture model (SMM) (Ito et al., 2016) guided by diarization estimates to obtain an enhanced signal for an utterance. An SMM is fitted for each utterance (according to the diarization estimate) using observations from the utterance boundaries plus some context which can span several tens of seconds. The diarization estimate is used to indicate potential source activity. After this fitting, the model's posterior is used for mask-based beamforming. The minimum variance distortionless response (MVDR) (Capon, 1969) beamformer is used followed by a-posteriori maximum SNR channel selection (Erdogan et al., 2016) and blind analytic normalization (BAN) (Warsitz and Haeb-Umbach, 2007) post-filtering. As such, GSS does not use any training data, as it is simply fitted at inference time on each utterance. Thereby, it does not suffer from the domain mismatch issues often encountered with supervised DNN methods. As will be shown in Section 4, this same pipeline, comprising EV-based channel selection followed by GSS, was also adopted by almost all participants without significant modifications, further demonstrating its reliability and effectiveness.

3.4.2. Diarization

As said, the two baselines differ substantially in the diarization component employed. The ESPnet baseline diarization component is illustrated in Fig. 4 while the NeMo one is in Fig. 5.

The ESPnet baseline diarization component is built on top of the Pyannote (Bredin et al., 2020) diarization 2.1 pipeline (Bredin, 2023) which is suitably modified to deal with multiple microphone inputs. The latter consists of a local EEND model (Fujita et al., 2019b; Bredin and Laurent, 2021) and the pre-trained ECAPA-TDNN (Desplanques et al., 2020) speaker-id embedding model. The local EEND model predicts speech activation for up to 3 concurrent speakers within a context of 5 s and is applied to the whole meeting with a 0.5 s stride and to every microphone channel. We use this model to select, for each 5 s window, the channel with the highest total speech activity. This channel is then used to extract ECAPA-TDNN speaker-id embeddings with the overlap-aware mechanism proposed in Bredin and Laurent (2021) and Bredin (2023). To mitigate false alarms, the speaker activity selection decision is made after logit thresholding and median filtering applied to each microphone channel. This selection strategy is strongly needed as, especially in scenarios as CHiME-6, where diarization error rate (DER) between different microphones can vary by more than 10%. We also experimented with ensembling the diarization pipeline results across arrays using DOVER-Lap (Raj et al., 2021b). However we found this latter solution to have significantly worse performance on CHiME-6. Moreover, it is intrinsically significantly slower in inference, due to the fact that the whole pipeline must be run on every channel. In the proposed channel selection strategy instead, only the local EEND model is run on each channel, leading to a significant speed up, as computationally heavy operations, such as ECAPA-TDNN embedding extraction, are performed only for one channel. The local EEND model is fine-tuned on the CHiME-6 `train` set (using all far-field channels only) using for validation and early stopping the MX6 `dev` set. Diarization output post-processing is employed in order to avoid too long utterances or too short utterances, both of which will hurt ASR performance. In detail, we merge utterances from the same speaker that are 0.5 s seconds apart up to a maximum duration of 30 s and with a target duration of around 10 s. Utterances that are instead longer than 30 s are split. We tune hyper-parameters, such as clustering and local EEND model threshold, including this diarization post-processing, using the Optuna framework (Akiba et al., 2019) with

macro JER (over MX6, CHiME-6 and DiPCo dev sets) as the optimizing criteria and tree-structured Parzen estimator (Ozaki et al., 2020) as the optimization method. For the updated ESPnet C8DASR baseline these hyper-parameters are re-tuned by adding also a part of NOTSOFAR-1 train set (the dev set was not available when building the baseline).

The NeMo baseline diarization pipeline is based on a modified multi-scale diarization decoder (MSDD) technique (Park et al., 2022) which is an improved variant of TS-VAD. First, as a pre-processing step, multi-channel dereverberation is performed via MIMO-WPE (Yoshioka and Nakatani, 2012) applied on 40 s windows with 2 s overlap. Next, to reduce the number of channels, channel clustering is performed using normalized maximum eigengap spectral clustering (NME-SC) clustering (Park et al., 2019) on the spatial coherence matrix obtained from all microphone channels by averaging across the whole meeting. To identify speech and silent regions, a multi-channel VAD model (MC-VAD) is employed. This model is based on the pre-trained MarbleNet VAD model, which is then fine-tuned on a 3000 h mix of CHiME-6 train data and synthetic data obtained from VoxCeleb 1&2, LibriSpeech and MUSAN using the NeMo multi-speaker simulator (Park et al., 2023b). Such VAD model is applied on each clustered channel independently, with logits fused by a max operation over all channels for each timestep. Multi-resolution TitaNet (Koluguri et al., 2022) speaker-id embeddings are extracted on each speech segment detected by the VAD using 3 s, 1.5 s, and 0.5 s all with 50% overlap. The MSDD model is then initialized using NME-SC clustering over these multi-resolution TitaNet embeddings. To create multi-channel aware speaker-id embeddings, the TitaNet embeddings from each clustered channel belonging to the same segment are concatenated in this step. However, the channel with the lowest correlation is excluded. Then after a first diarization hypothesis, the MSDD model is applied on each clustered channel independently (thus not using multi-channel aware speaker embedding) with a context size of 15 s and an hop-size of 3 s and outputs a logit for each speaker each 0.05 s using input embedding scale interpolation (Park et al., 2023c). The final diarization output is derived, after logit thresholding, via majority voting over all clustered channels for each timestep. The particular MSDD model differs from the original MSDD model proposed in Park et al. (2022) as it employs a four layers Transformer network with an embedding size of 384 and feed-forward network hidden size of 2048 instead of long-short term memory networks. It is trained using the same 3000 h data employed for the VAD model consisting of CHiME-6 and simulated conversations. As in the ESPnet baseline, here diarization post-processing is also employed. Specifically, too short silences or utterances are suppressed, and onset and offset padding is applied on each utterance in order to avoid truncated words. Hyper-parameters were tuned using the Optuna framework on the C7DASR (C8DASR for the C8DASR challenge) dev set using DA-WER as the optimizing criteria.

3.4.3. Automatic speech recognition

Both baselines employ a "classical" single-channel ASR back-end; i.e. no target-speaker ASR or multi-channel features are integrated into the ASR model.

The ESPnet baseline ASR back-end is based on IRIS (Chang et al., 2022) and multi-IRIS (Masuyama et al., 2022) ASR back-ends. It integrates WavLM as a front-end for a Transformer (Vaswani et al., 2017) attention-based encoder–decoder model. WavLM large is employed and its weights are kept frozen. Instead, its internal representation is fed to the Transformer encoder through layer-wise learnable softmax weights, a technique which is commonly employed (wen Yang et al., 2021; Chang et al., 2022; Masuyama et al., 2022; Yang et al., 2024). The model consists of 12 encoder layers and 6 decoder layers for a total of only 30 M trainable parameters as the pre-trained WavLM representation is leveraged. It is trained using hybrid CTC/attention loss (Watanabe et al., 2017) on the full CHiME-6 train and MX6 train_intv and train_calls portions, using data from all microphone devices. Furthermore, data-augmented close-talk microphone data is created from both MX6 and CHiME-6. We follow the data augmentation pipeline developed for the Kaldi CHiME-6 baseline (Watanabe et al., 2020), where MUSAN (Snyder et al., 2015) and SLR26 (Ko et al., 2017) are employed to get noisy/reverberant signals from close-talk microphone utterances. To obtain satisfactory performance, we found critical to also include GSS-enhanced data (obtained using oracle diarization) from the CHiME-6 train set. The model is trained for 5 epochs, and the weights from the best 3 checkpoints are averaged. Adam is used for optimization with a peak learning rate of 0.0001. A warmup schedule of 40k steps is employed followed by linear decay. SpecAugment (Park et al., 2020) masking is performed during training on the WavLM representation after the learnable softmax operation. Hybrid CTC attention decoding is then performed in inference. We use a beam size of 10 and a CTC weight of 0.3.

The NeMo baseline ASR model is based instead on the pre-trained NeMo Conformer (Gulati et al., 2020) transducer XL model.[10] The pre-trained model is fine-tuned using CHiME-6 and MX6 train subsets data after pre-processing with GSS using ground-truth diarization. The model consists of 600 M parameters and is updated for 35 k steps with 128 batch size and a learning rate of 0.00001. For decoding, adaptive expansion search is employed (Kim et al., 2020) together with an external word-piece LM. The LM is based on byte-pair-encoding (BPE) tokens using SentencePiece (Kudo and Richardson, 2018) and is constructed using the KenLM toolkit (Heafield, 2011) by using text data from C7DASR train and dev sets.

4. Analysis of submitted systems

The two challenges attracted a total of 9 teams, in alphabetical order: BUT FIT (Karafiát et al., 2023), IACAS-Thinkit (Ye et al., 2023), NVIDIA-NeMo (Park et al., 2023c), NPU (Mu et al., 2023), NTT (Kamo et al., 2023, 2024), Paderborn University (Boeddeker et al., 2023), STCON (Prisyach et al., 2023; Mitrofanov et al., 2024), University of Cambridge (Deng et al., 2023) and USTC-NERCSLIP (Wang et al., 2023a). Two teams: STCON and NTT, participated in both the C7DASR and C8DASR challenges. Other teams, such as USTC-NERCSLIP, NPU and BUT participated instead only in the C7DASR and the CHiME-8 NOTSOFAR-1 challenge.

[10] Available: https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge.

Table 4
Summary of diarization and front-end components for C7-8DASR baseline and best submitted systems. Top panel (grey): baseline systems; middle panel (yellow): C8DASR submissions; bottom panel (pink): C7DASR submissions. Within each panel, participants systems are ordered according to macro-averaged tcpWER. We report only the best system for each team.

System	Diarization				Front-End	
	Segmentation	Spk-id Emb. Extr. & Clustering	Refinement	Multi-channel mechanism	Channel selection	Separation
ESPnet baseline	Pyannote segmentation	Pyannote diarization 2.1	– –	Top-1 channel VAD selection	EV	GSS
NeMo baseline	MarbleNet VAD	Multi-scale TitaNet & NME-SC	Transformer MSDD	logit max pooling & multi-channel TitaNet embeddings	EV	GSS
STCON	TDNN stats-based	Ecapa-TDNN & wav2vec 2.0 AED multi-speaker OSD with UMAP+DBSCAN+GMM clustering	NSD-MS2S	DOVER-Lap and TS-VAD posterior averaging	EV	GSS with neural refinement (G-TSE)
NTT	EEND-VC	EEND-VC & ECAPA-TDNN with NME-SC	NSD-MS2S	DOVER-Lap & channel clustering for speaker counting	EV & Brouhaha C_{50}	GSS with SP-MWF
IACAS-Thinkit	Pyannote segmentation	Pyannote diarization 2.1	TS-VAD	DOVER-Lap	–	Neural TSE for GSS init
USTC-NERCSLIP	Pyannote segmentation	ECAPA-TDNN + NME-SC	NSD-MS2S	NSD-MS2S posterior averaging	EV + virtual subarray SINR	GSS
NTT	EEND-VC	EEND-VC	–	DOVER-Lap	–	GSS
STCON	TDNN stats-based	ECAPA-TDNN + NME-SC	TS-VAD	TS-VAD posterior averaging	EV	GSS
University of Cambridge	Pyannote segmentation	Pyannote diarization 2.1	–	Top-3 channel VAD selection + DOVER-Lap	EV	GSS
NVIDIA NeMo	MarbleNet VAD	Multi-scale TitaNet & NME-SC	Transformer MSDD	logit max pooling & Multi-channel TitaNet embeddings	EV	GSS
NPU	Pyannote segmentation	Pyannote diarization 2.1	– –	Top-1 channel VAD selection	EV	GSS
Paderborn University	Pyannote segmentation	Pyannote diarization 2.1	TS-VAD + d-vector refinement	TS-VAD posterior averaging	–	GSS
BUT FIT	Pyannote segmentation	Pyannote diarization 2.1	– –	Top-1 channel VAD selection	EV	GSS

A total of 32 systems were submitted for the two challenges: 5 for C8DASR and 27 for C7DASR. Among the C7DASR systems, 13 were submitted to the optional oracle diarization track and 14 to the main track. Due to the sheer number of submitted systems and due to the fact that most of them differ only by hyper-parameter choices, we will restrict our analysis here by considering mostly the best systems submitted by each team with few notable exceptions which we deem particularly interesting.

We summarize baseline and participants systems diarization and front-end components in Table 4 and the back-end components (ASR and external LM) in Table 5. Each table is divided into three panels: in the top panel, for reference, are the baseline systems which have been described before in Section 3.4; in the middle one C8DASR submissions and in the last one C7DASR submissions. Within each panel, the team submissions are sorted in ascending order by tcpWER with best systems first. Challenge results will be summarized and discussed later in Section 5.

4.1. Front-end processing

In Table 4 participants system diarization techniques are broken down into:

- *Segmentation*: part of the diarization system used for VAD, overlapped speech detection (OSD), or multi-speaker segmentation. This includes EEND techniques (Fujita et al., 2019a,b; Kinoshita et al., 2021a,b), since these can be considered as a generalization of the VAD task (a VAD for each speaker).
- *Speaker-id embedding extraction and clustering*: technique used to extract speaker-id discriminative embeddings and clustering method used to determine the total number of speakers throughout the meeting.
- *Refinement*: diarization refinement techniques such as TS-VAD (Medennikov et al., 2020b) if any is used.
- *Multi-channel mechanism*: what mechanism/strategy is used to get predictions from multiple-microphones. A straightforward strategy is to just ensemble predictions from the same model applied to different channels.

With front-end processing, we instead denote the components used for microphone array processing, which has to be array agnostic in the two challenges due to the difference, across scenarios, in recording setups. Front-end processing is broken down into *Channel Selection* and *Speech Separation*. The two ESPnet and NeMo baselines, for example, employ EV and GSS for these two components.

Table 5
Summary of back-end components for C7-8DASR baseline and best submitted systems. Top panel: baseline systems (grey); middle panel: C8DASR submissions (yellow); bottom panel (next page): C7DASR submissions (pink). Within each panel, the participants' systems are ordered according to macro-averaged tcpWER. We report, for each team, only the best system configuration.

System	ASR							
	Features	Model	Criteria	Multi Channel	TTA	Spk Adapt	Ensembling	External LM
ESPnet Baseline	WavLM	Transformer	CTC+Att.	–	–	–	–	–
NeMo Baseline	Fbank	Conformer	Transducer	–	–	–	–	Word-piece N-gram
STCON	WavLM	Uconv-Conformer	CTC+Att.	–	STAR	–	N-best lattice fusion with MBR decoding	Word-piece 3-g + Transformer + AWD-LSTM
	WavLM	E-branchformer	CTC+Att.	–	STAR	–		
	WavLM	ZipFormer	CTC+Transducer	–	–	–		
	WavLM	MS TDNN-F	phoneme HMM	–	–	–		
NTT	Fbank	Whisper-M	Att.	–	–	–	ROVER	Transformer
	Fbank	Whisper-L	Att.					
	Fbank	NeMo Conformer	Transducer					
	WavLM	Transformer	Transducer					
IACAS-Thinkit	WavLM	Transformer	CTC	–	Pseudo label fine-tuning	–	–	Word-piece N-gram + Transformer (inter-utterance)
USTC-NERCSLIP	WavLM + ECAPA-TDNN + ConvTasNet	Conformer	CTC+Att	–	–	–	ROVER	–
	wav2vec 2.0 + ECAPA-TDNN + ConvTasNet	Conformer	CTC+Att					
NTT	WavLM	Conformer+S4	CTC+Att.	–	Pseudo label fine-tuning	–	ROVER	Transformer
	WavLM	Branchformer+S4	CTC+Att.					
	WavLM	Branchformer	Transducer					
STCON	WavLM	Uconv-Conformer	CTC+Att.	–	–	–	ROVER	Word-piece 3-g + Transformer + AWD-LSTM
	WavLM	E-branchformer	CTC+Att.					
	WavLM	ZipFormer	CTC+Transducer					
	WavLM	MS TDNN-F	phoneme HMM					
University of Cambridge	WavLM	–	CTC	–	SUTA	–	Two-pass decoder-only rescoring	Word-level N-gram
	WavLM	Transformer	CTC+Att.					
	WavLM	Transformer	CTC+Att. backwards					
	WavLM	Transformer	LS-Transducer					
NVIDIA NeMo	Fbank	Conformer	Transducer	–	–	–	ROVER	Word-piece N-gram
NPU	WavLM + cosIPD	Transformer	CTC+Att.	MFCCA+ CGCS+FGCS	– –	Implicit via GSS ch. input	ROVER –	Transformer –
Paderborn	WavLM	Transformer	CTC+Att.	–	–	–	–	–
BUT FIT	WavLM	Transformer	CTC+Att.	–	–	–	–	–

4.1.1. Neural speaker extraction/Separation

From Table 4 it is evident that all systems adopt GSS as the main separation component, thus following the same diarization→ GSS→ ASR pipeline scheme as adopted by the baseline systems and depicted in Fig. 3. As said, this same scheme was also used by the best systems in the past CHiME-6 challenge. Few participant systems try to improve upon GSS, confirming its efficacy overall. For example, the C7DASR IACAS-Thinkit and the C8DASR STCON systems complement GSS with DNN-based neural TSE. However, while the IACAS-Thinkit system uses the DNN-based TSE to initialize GSS, the C8DASR STCON instead uses it for refinement: the TSE model takes the GSS output as an additional input to provide the target speaker cue. In the latter case, improvements over a GSS-only baseline where possible only after fine-tuning with the ASR loss. These facts outline the robustness of GSS and its effectiveness despite its simplicity. It is also worth mentioning that, in their C8DASR system submission paper (Mitrofanov et al., 2024), the STCON team also reports that they experimented with a continuous speech separation (CSS) (Chen et al., 2020b) approach similar to the one employed in the CHiME-8 NOTSOFAR-1 challenge baseline system (Vinnikov et al., 2024). However, they found that such an approach was too brittle, as the CSS model was prone to introduce speaker confusion errors in complex situations with fast turn-taking dynamics and low SNR, such as in CHiME-6 scenario.

4.1.2. Channel selection

Similarly, regarding channel selection, almost all participants used the proposed baseline EV channel selection strategy. It was found to be quite effective despite its simplicity and has also the additional benefit of reducing inference time. The C8DASR NTT system combines it with the pre-trained Brouhaha (Lavechin et al., 2023) multi-task VAD model which can predict C_{50} speech clarity index, which measures the amount of reverberation. The C7DASR USTC-NERCSLIP instead combines it with a more complicated strategy: EV scores for each microphone are computed, then the channels are sorted and grouped into "virtual subarrays" of 5

microphones each. Notably, some teams, such as Paderborn University and NTT, did not use channel selection in their C7DASR submissions. However, NTT's next year C8DASR submission incorporated channel selection, indicating that this component is beneficial, particularly for excluding problematic channels. After all, as outlined in Fig. 2, Section 3.1, scenarios such as Mixer 6 and CHiME-6 inherently benefits from some explicit or implicit channel selection component due to their significant inter microphone channel SDR variability.

4.2. Diarization

Reliable diarization and speaker counting are crucial components for meeting transcription, even more so since, as said before, all participants relied on GSS. Consequently, significant effort was dedicated to improve diarization accuracy. The only exceptions are the C7DASR NPU and BUT FIT systems, which focused solely on improving the ASR component and used the ESPnet baseline diarization system without modifications. In C7DASR, most of the other teams, with the exception of NeMo (Park et al., 2023c), STCON (Prisyach et al., 2023), mainly built upon the C7DASR baseline diarization system by introducing effective modifications (Deng et al., 2023) and/or also a TS-VAD module for refinement (Boeddeker et al., 2023). The NTT system notably differs from all others as it does not rely on a classical segmentation plus clustering pipeline. Instead, it employs a modified end-to-end neural diarization with vector clustering (EEND-VC) (Kinoshita et al., 2021b,a) which uses WavLM features, resulting in a simple and streamlined diarization pipeline. A TS-VAD component appears in all top-performing C7-8DASR systems, confirming its effectiveness. Indeed, diarization performance heavily depends on TS-VAD's architecture and training data. C7DASR IACAS-Thinkit (Ye et al., 2023) and STCON teams use the original TS-VAD model (Medennikov et al., 2020b), while Paderborn University, USTC-NERCSLIP (Wang et al., 2023a) and NeMo explore different architectures and speaker embeddings for conditioning. Among these, the USTC-NERCSLIP NSD-MA-MSE TS-VAD model (He et al., 2023) proved to be very effective, leading to its adoption in subsequent C8DASR submissions by both STCON and NTT teams.

4.2.1. Improvements in the speaker counting components in C8DASR

In C8DASR, due to the increased difficulty of accurate speaker counting from the introduction of NOTSOFAR-1, both STCON (Mitrofanov et al., 2024) and NTT (Kamo et al., 2024) make substantial modifications regarding the core diarization component to make it more robust. For example, it is evident how the STCON team switched from a classical VAD+ECAPA-TDNN clustering pipeline for the initial diarization to a more complex one that can better deal with overlapped speech. The core component for their pipeline is a novel wav2vec 2.0 attention-based encoder–decoder (AED) model that extracts frame-wise speaker embeddings and can be used also to perform overlapped speech detection (OSD). On the other hand, the NTT system still relies on EEND-VC but adds robustness via a multi-channel speaker counting component that leverages ECAPA-TDNN speaker embeddings extracted after GSS.

4.3. Multi-channel mechanism, ensembling and test-time adaptation

As it is evident from Table 4, most systems rely on ensembling techniques in order to fuse information across multiple microphones. As such, the efficacy of native multi-channel diarization techniques such as Horiguchi et al. (2022) remains largely under-explored for complex conversational scenarios. A notable exception is the C7DASR NeMo system, which leverages multi-channel information in the speaker embedding clustering phase by concatenating as explained in Section 3.4.2.

C7DASR NTT and IACAS-Thinkit systems explored test time adaptation (TTA) in their diarization pipelines through iterative pseudo-labeling. However, according to the NTT technical report, this approach yielded only marginal improvements while introducing substantial computational overhead, raising questions about its practical value in real-world applications.

4.4. Automatic speech recognition

In Table 5 we present a taxonomy of participants' ASR techniques. These are summarized and broken down into the following:

- *Features*: input features used for the ASR model. These can include latent representation from pre-trained models such as WavLM.
- *Model*: main DNN architecture used e.g. Conformer.
- *Criteria*: ASR criteria used, such as CTC+Attention, Attention-only, connectionist temporal classification (CTC) (Graves et al., 2006) or transducer (Graves, 2012).
- *Multi-channel*: if the ASR model uses some technique to fuse information from multiple-channels.
- *Text-time adaptation (TTA)*: what technique is used for unsupervised TTA. These include simple iterative pseudo-labeling, STAR (Hu et al., 2024) or single-utterance test-time adaptation (SUTA) (Lin et al., 2022) for example.
- *Speaker adaptation*: what technique, if any, is used for target speaker adaptation e.g. if the ASR accepts a cue for the target speaker.
- *Ensembling*: ensembling technique used e.g. Recognizer Output Voting Error Reduction (ROVER) (Fiscus, 1997) or others to combine the output of multiple systems.
- *External LM*: external language model (LM) used during ASR decoding.

4.4.1. Input features, criteria and architecture

First, it can be noted that all participants leverage external pre-trained models in some way with WavLM being the most commonly used one. Some exceptions are the NeMo baseline and the C7DASR NeMo team submission which use the pre-trained NeMo Conformer-transducer model and the C8DASR NTT submission which also use Whisper (it was allowed in C8DASR to be consistent with the concurrent CHiME-8 NOTSOFAR-1 challenge). Among all submissions the C7DASR USTC-NERCSLIP is also peculiar in the fact that it leverages an interesting combination of features from WavLM, ECAPA-TDNN and even a pre-trained Conv-TasNet speech enhancement model (Luo and Mesgarani, 2019).

Regarding the ASR models architecture, most participants use Transformer or Conformer based ASR models and CTC+Att or transducer-based criteria. STCON C8DASR and C7DASR submissions also experiment with Kaldi-based DNN-HMM multi-stream TDNN-F (Han et al., 2019) and K2 Zipformer (Yao et al., 2023) with CTC plus pruned stateless transducer (Povey, 2020). The NTT C7DASR submission instead experimented with S4-based (Gu et al., 2021) layers to augment the Conformer and Branchformer (Peng et al., 2022) models. For the subsequent NTT C8DASR challenge, these S4 components were no longer incorporated into their approach, suggesting that their contribution to performance gains was marginal or debatable. Similarly to STCON, the University of Cambridge C7DASR submission (Deng et al., 2023) also includes alternative, more exotic, ASR criteria, such as a backward CTC+Attention, which operates on a right-to-left encoder representation, and the recently proposed label-synchronous neural transducer (Deng and Woodland, 2024). The rationale for this is that such diversity in criteria can enhance the effectiveness of ensemble techniques, as they can be complementary to each other. While the widespread use of multiple criteria and architectures for ensemble approaches makes direct comparisons challenging, the C7DASR IACAS-Thinkit WavLM-based CTC ASR system stands out for its notable simplicity and relative efficacy.

4.4.2. Ensembling and test-time adaptation techniques

Almost all submissions extensively use ensembling techniques from different ASR models, with ROVER being the most commonly used method. Notable exceptions are IACAS-Thinkit and BUT FIT submissions, as well as one submission from the NTT team for C8DASR, which focuses on lightweight inference and uses only the Whisper-M model.

Two teams use more complex ensembling strategies. The C8DASR STCON submission switches from ROVER as used for C7DASR to a more complicated fusion technique, involving lattice fusion and minimum Bayes risk decoding (Kumar and Byrne, 2004) while the University of Cambridge's C7DASR submission adopts a two-pass approach where hypotheses from a CTC-based model are rescored and combined using the decoders of three other models. However, the wide usage of ROVER suggests that, despite its simplicity, it is still competitive with these more refined methods.

Concerning ASR TTA techniques, we observe the same trend as for diarization. Few systems make use of these techniques (C8DASR STCON, C7DASR IACAS-Thinkit, NTT and University of Cambridge). However, when ablations are reported in their respective technical description papers (Kamo et al., 2023), they seem to offer small improvements despite their significant computational cost. This is why, for example, the C8DASR NTT system does not use TTA via pseudo-labeling anymore compared to their C7DASR system.

4.4.3. Multi-channel and target speaker information

The C7DASR NPU system (Mu et al., 2023) is the only one that tries to incorporate into the ASR model multi-channel information via spatial features such as cosine sine inter-channel phase differences (cosIPD) so that the model can still recover the original speech signal if the GSS enhancement is imperfect. No participants explicitly leveraged target-speaker information in either the C7DASR or C8DASR challenges. However, we argue that all participants ASR systems inherently possess some target-speaker capability due to their use of GSS as pre-processing. In fact, GSS-based TSE is far from perfect and, at least when looking at the baseline enhanced utterances, there are many instances where the background speakers are not well suppressed. Yet, the ASR model, if trained or fine-tuned on such data can learn that the target speaker is the one which is slightly more energetic and has the utterance "centered" regarding the segmentation. For this reason, as mentioned in Section 3.4.3, we found it crucial for both NeMo baseline and ESPnet baseline ASR components to include GSS-enhanced data into the training material. Similarly, all participants in the C7-8DASR challenges included GSS-enhanced data into the ASR training as specified in their technical reports.

4.4.4. Language modeling

Maybe, rather unsurprisingly, most of the best performing systems (C8DASR STCON and NTT, C7DASR IACAS-Thinkit, NTT and STCON) are the ones that also make use of neural LM models.

C8DASR had an optional sub-track where participants could leverage external pre-trained LLMs. The STCON C8DASR submission experimented with this direction and employed a fine-tuned Llama 2 7B model on top of their system. This model was used for hypothesis rescoring using a method similar to the one proposed by Ogawa et al. (2024), where the LLM is applied with inter-utterance context i.e. using past context from the same speaker previous utterances (up to 1024 tokens in total). However, they found that this approach yielded very marginal improvements in tcpWER. This limited gain may be attributed to their already robust ASR component, which incorporates multiple language models: a 3-g model alongside two neural models: one Transformer-based and another implemented using ASGD Weight-Dropped (AWD) long-short term memory (LSTM) (Merity et al., 2018) networks. Similarly, Ogawa et al. (2024) shows small but more significant improvements on C7DASR challenge data by using Llama 2 7B for inter-utterance N-best list ASR hypothesis rescoring of the NTT C7DASR system. In contrast, the C7DASR IACAS-Thinkit submission reports significant improvements by using an inter-utterance Transformer-based LM model. These observations suggest that the efficacy of integrating LLMs for ASR hypothesis post-processing remains debatable and requires further investigation, particularly when the ASR component is already strong, as this can lead to diminishing returns. In such cases, the computational resources required for LLM inference may outweigh performance improvements, especially in production environments where efficiency is crucial.

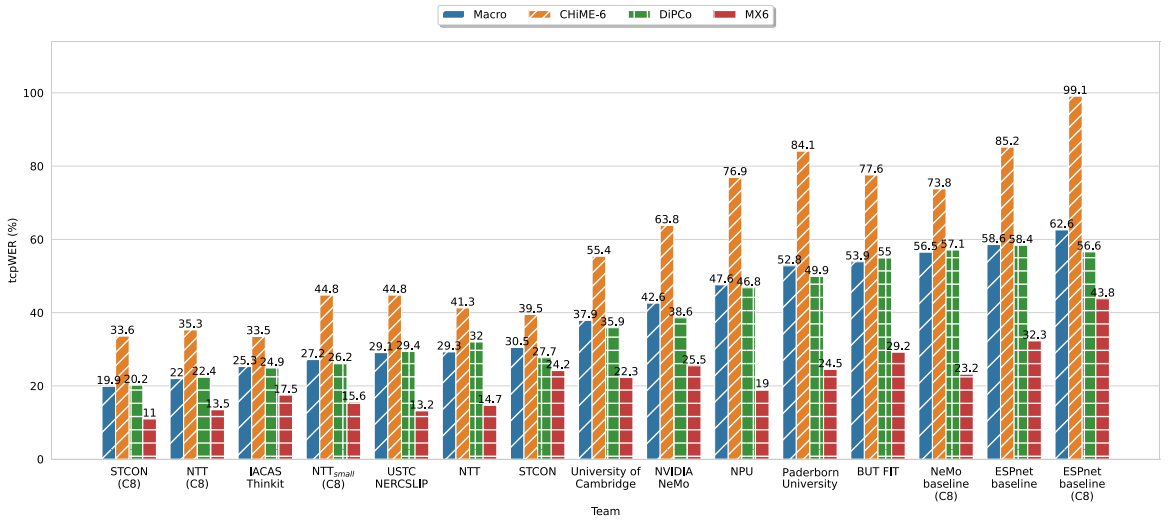

Fig. 6. tcpWER (%) for each C7DASR core scenarios (CHiME-6, DiPCo, MX6) as well its macro average (Macro) for both C7DASR and C8DASR systems. C8DASR systems are denoted by (C8).

5. Challenges results

In this section, we present and discuss the results of C7DASR and C8DASR. Due to the large number of submitted systems (up to 3 submissions per track were allowed for each team), we focus on and report results only for each team's best-performing system, as measured by scenario-wise macro-averaged tcpWER (as described in Section 3). One notable exception is the NTT submission to C8DASR, which includes a system designed with a greater emphasis on efficiency, aligned with the C8DASR jury award mechanism outlined in Section 3.3. This system is referred to as NTT$_{small}$ in the following discussion. Note that despite being among the most efficient and practically oriented systems in both challenges, it still had a real time factor (RTF) of 2.46 (but without any optimization). As highlighted in Sections 2.4 and 3.3, the results presented here are computed using the C8DASR text normalization strategy prior to scoring. Consequently, these results differ from those available on the official C7DASR challenge website,[11] which used a different normalization and metric (DA-WER). This choice was made to ensure the results are more directly comparable between C8DASR and C7DASR but also with the CHiME-8 NOTSOFAR-1 challenge, which uses the same ranking metric (tcpWER) and the same C8DASR text normalization protocol.

5.1. Joint CHiME-7 and 8 DASR results

Since C8DASR includes an additional scenario compared to C7DASR (the NOTSOFAR-1 dataset), it allows for a comparative analysis of submitted systems across both challenges. Fig. 6 shows the tcpWER results for the core scenarios of C7DASR, along with their macro-average (the C8DASR ranking metric, as explained in Section 2.4), for systems submitted to both challenges. Across all systems and baselines, the tcpWER figures consistently reveal that the scenarios, ranked from most difficult to easiest, are CHiME-6, DiPCo, and MX6. This ranking aligns with expectations: MX6 involves only two participants in a relatively static interview setting, whereas DiPCo and CHiME-6 feature four speakers that dynamically interact in more complex environments. CHiME-6 is particularly challenging due to its multi-room setting and highly colloquial speech, resulting in significantly higher tcpWER figures for all participants. Even for the best performing systems, the tcpWER for CHiME-6 remains above 30%. This means that nearly one in three words is incorrectly transcribed, despite advances in ASR and diarization technology and the fact that almost every participant used ASR systems ensembles.

Achieving balanced performance across all three scenarios is a significant challenge. For instance, some systems, such as the C7DASR STCON submission, rank among the top C7DASR systems on CHiME-6 but shows significantly higher tcpWER on MX6. Conversely, systems like the C7DASR NPU submission achieve low tcpWER on MX6 but perform worse on CHiME-6. Others, such as the C8DASR submissions and the C7DASR IACAS-Thinkit submission, demonstrate more balanced performance overall. This difficulty underscores the main motivation behind these challenges: while it is relatively straightforward to optimize a system for a specific scenario like CHiME-6 or MX6, developing a system or technique capable of robust performance across all scenarios remains very challenging. Another example of this is the C7-8DASR ESPnet baseline: the C8DASR version, due to the fact that needs to handle the NOTSOFAR-1 scenario, performs worse than the C7DASR one on C7DASR core scenarios. In this sense, it is notable

[11] www.chimechallenge.org/challenges/chime7/task1/results.

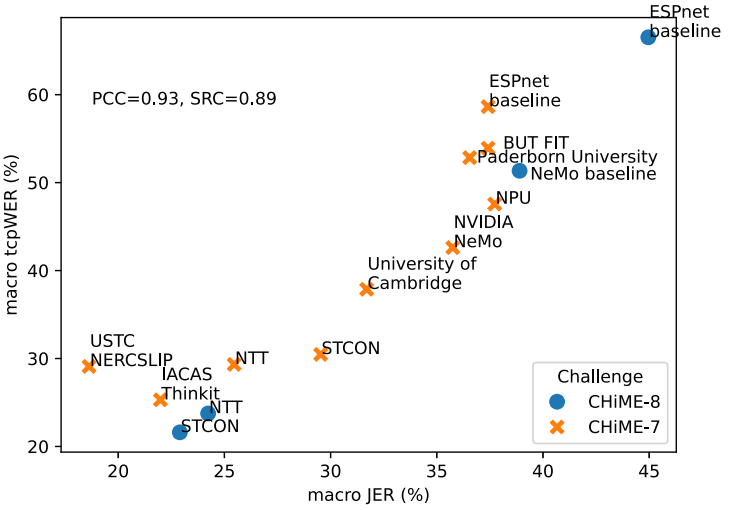

Fig. 7. tcpWER (%) vs. JER (%) for C7DASR and C8DASR systems (only top system for each team). Both are macro-averaged across C7DASR core scenario (CHiME-6, DiPCo, MX6). Pearson correlation coefficient (PCC) and Spearman rank correlation (SRC) are also reported.

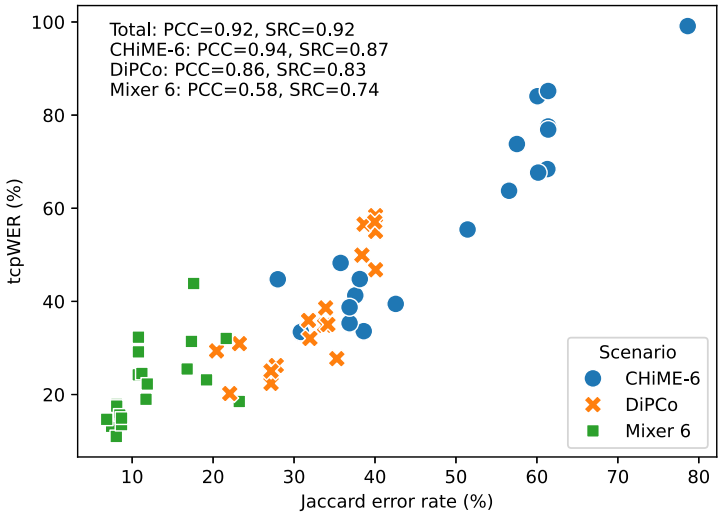

Fig. 8. tcpWER (%) vs. JER (%) for C7DASR and C8DASR systems (all submissions are included for a total of 22 systems). Figures are computed for each core scenario separately (CHiME-6, DiPCo, MX6). Pearson correlation coefficient (PCC) and Spearman rank correlation (SRC) are also reported for all the scenarios jointly (Total) and also separately for each scenario.

that STCON and NTT C8DASR submissions are still able to improve over C7DASR best systems even if they need to also tackle the new NOTSOFAR-1 scenario introduced in C8DASR.

In Fig. 7 macro-averaged tcpWER versus macro-averaged Jaccard error rate (JER) (Ryant et al., 2019) is reported for the same C7DASR and C8DASR systems. The macro-average is taken across all three C7DASR scenarios. JER is computed by considering a start and end-point collar of 250 ms, since the manual annotation of the utterance boundaries can have minor inconsistencies across the three datasets and is inherently imperfect due to human error (Fiscus et al., 2007b; Ryant et al., 2019).

In general, there is a clear correlation between JER and tcpWER (PCC = 0.93 and SRC = 0.89). This is largely expected. As mentioned in Section 3, all participants practically relied on GSS-based TSE, which requires accurate diarization to be effective. The best systems consistently have a macro JER around 20%–30%, while the worst over 40%. This is even more evident in Fig. 8 where we plot, respectively, JER versus tcpWER for each scenario (including the macro-average) and for all submitted systems (thus not only the best one for each team). The number of systems amounts to 22 (14 C7DASR systems plus 5 C8DASR and the three baseline systems). In Fig. 9 we instead report DER versus tcpWER in the same manner.

First, we observe that JER has slightly higher SRC and PCC than DER with respect to tcpWER. This is because JER better accounts for speaker counting errors. In a diarization → separation → ASR pipeline, speaker counting errors are particularly catastrophic as they compound through subsequent stages. If the number of speakers is estimated incorrectly in the first diarization pass, errors

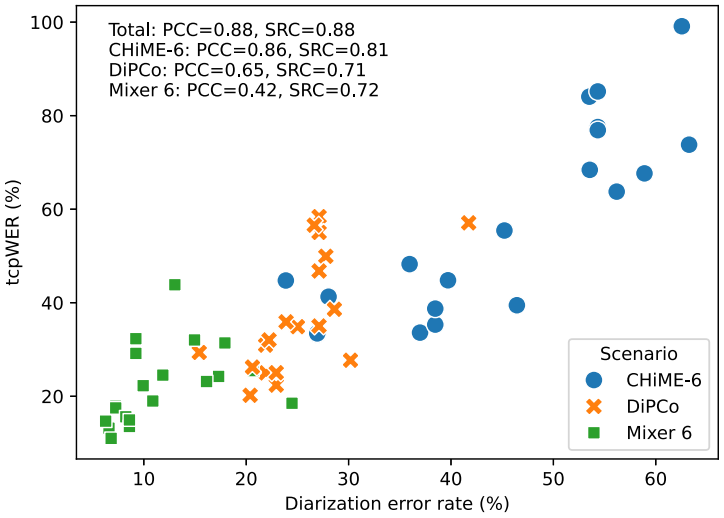

Fig. 9. tcpWER (%) vs. DER (%) for C7DASR and C8DASR systems (all submissions are included for a total of 22 systems). Figures are computed for each core scenario separately (CHiME-6, DiPCo, MX6). Pearson correlation coefficient (PCC) and Spearman rank correlation (SRC) are also reported for all the scenarios jointly (Total) and also separately for each scenario.

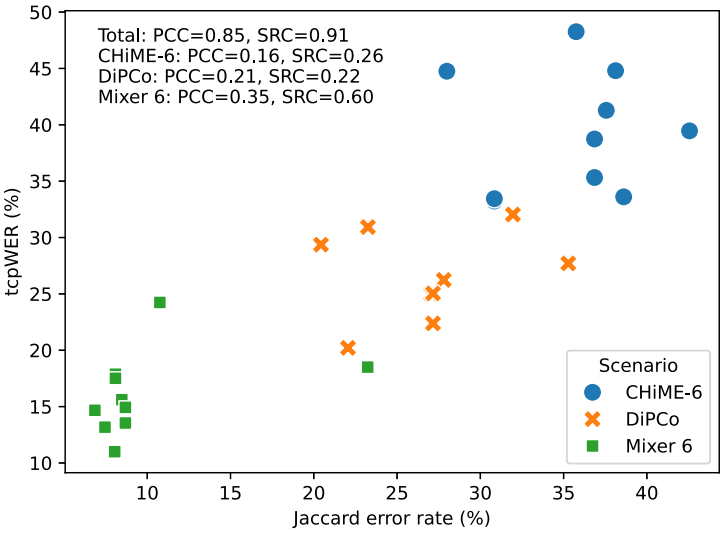

Fig. 10. tcpWER (%) vs. JER (%) for C7DASR and C8DASR systems (all submissions but only from the top-6 ranking teams). Figures are computed for each core scenario separately (CHiME-6, DiPCo, MX6). Pearson correlation coefficient (PCC) and Spearman rank correlation (SRC) are also reported for all the scenarios jointly (Total) and also separately for each scenario.

propagate through TS-VAD refinement (if employed), then GSS, and finally ASR. This compounding effect makes speaker counting errors more impactful than other types of diarization errors, even when those speakers have relatively short speech duration. This phenomenon was also observed in our previous work (Cornell et al., 2024c) when analyzing the C8DASR baseline results. There may be instances where DER values are similar between systems, but tcpWER differs substantially due to differences in speaker counting accuracy.

On the other hand, observing more closely Fig. 8, it also appears that for the top-4 ranking systems (STCON (C8), NTT (C8), STCON, NTT, USTC-NERCSLIP, IACAS-Thinkit) a lower JER does not always guarantee a lower tcpWER. That is, the correlation seems weaker for such top performing systems. Fig. 10 reports tcpWER versus JER only when considering such top-performing systems, confirming that indeed the correlation is weaker when scenarios are considered independently. Especially for the CHiME-6 and DiPCo scenarios (PCC and SRC around 20%).

This is, however, expected as such high performing systems do not make significant speaker counting errors and estimate correctly the number of speakers, especially on DiPCo and CHiME-6 scenarios. This is not true for Mixer 6 where even some high-performing systems overestimated the number of speakers (and, in fact, the SRC is lower than the one in Fig. 8 but still significant). On CHiME-6

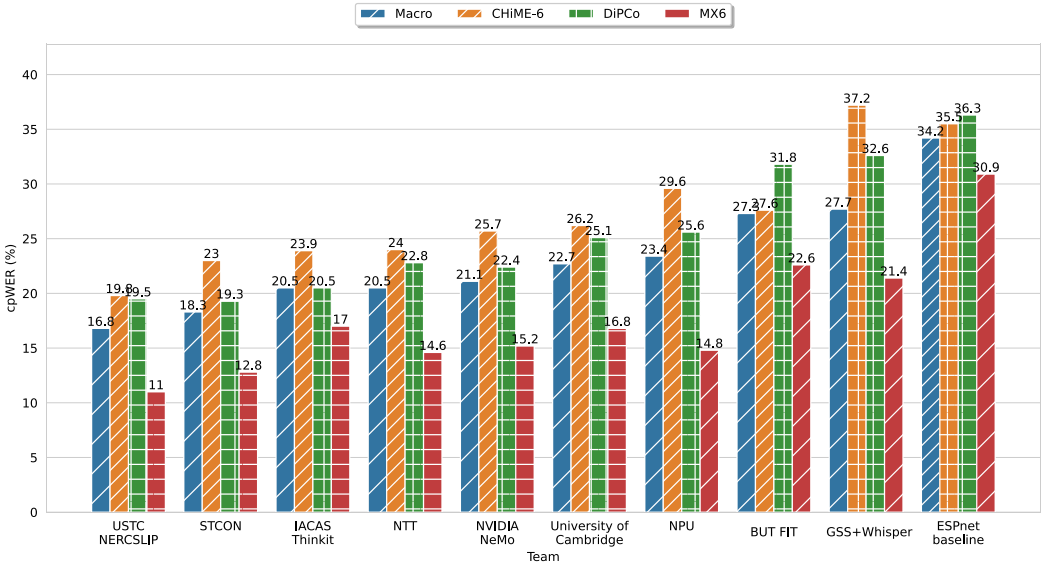

Fig. 11. tcpWER (%) for each C7DASR core scenario (CHiME-6, DiPCo, MX6) as well its macro average (Macro) for C7DASR acoustic robustness (oracle diarization) optional sub-track.

and DiPCo most JER errors for these systems arise from segmentation and are thus not very reflective of the final ASR performance. This phenomenon was already observed in the previous CHiME-6 challenge by the top-ranking team (STCON) on the non-oracle diarization track (Medennikov et al., 2020a,b). Among their system configuration, their best performing one was the one with worse DER as it had a higher false alarm rate. In fact, for ASR, it is better to combine short utterances from the same speaker into one and have less tight segmentation, as ASR can handle short pauses and usually benefits from having longer context. This approach sacrifices precise utterance-wise segmentation accuracy in favor of improved recognition performance, which is often the preferred outcome. Also recent work (Boeddeker et al., 2024) on joint TS-VAD/TSE and ASR, found that segment boundaries overestimation is beneficial for ASR even if DER is increased and performed some analysis. Such important post-processing step is also performed (as explained in Section 3.4.2) in the two baseline systems and also by all participants. These observations raise important questions on the relevance of diarization metrics like JER and DER when evaluating diarization systems for ASR applications. As shown, these metrics, past a certain threshold, may not always capture the impact of diarization quality on overall recognition performance and the resulting user experience. Therefore, if the application of the diarization system involves ASR, it is crucial to complement DER and JER with WER-based metrics.

In Fig. 11 we report the results of the C7DASR acoustic robustness sub-track, in which participants could use oracle diarization information. In Fig. 12 for the reader's convenience, we also report the absolute improvement vs. the main track (non oracle diarization, Fig. 6). Compared to previous plots, here we lack C8DASR systems as this track was not included in the C8DASR challenge. Note that this track was also optional: for example, the Paderborn University team did not submit, as they mainly focused on improving the diarization component. In this figure, as an additional reference, we also include the results for a baseline system (GSS+Whisper) that uses Whisper `large-v3` instead of the baseline ASR WavLM system with all other things equal.

In general, for all systems, we can observe a significant tcpWER improvement when using oracle diarization especially for the CHiME-6 scenario (Fig. 12). Instead, for DiPCo and MX6, at least for the best systems, the difference is much less pronounced except for STCON. Especially for MX6 which features an easier 2 speakers scenario. This suggests that much of the difficulty in the CHiME-6 scenario is still related to accurate speaker diarization and, in particular, speaker counting. For example, the STCON submission here ranks second with oracle diarization when considering only C7DASR systems. However, in Fig. 12 we can observe a significant difference between oracle and non oracle in macro-averaged tcpWER due to poor performance in MX6. This in turn was due to the fact that their diarization clustering component over-estimated the number of speakers in MX6 as it was over-tuned towards the CHiME-6 scenario (it is the best C7DASR system on CHiME-6 scenario in Fig. 6).

In contrast, their C8DASR system does not suffer from this issue and demonstrates greater robustness, thanks to some improvements which were explained in Section 4 such as the wav2vec 2.0-based frame-level speaker embedding model and the improved clustering pipeline. These observations are confirmed by Fig. 13 where we plot instead, for all C7-8DASR systems JER for each C7DASR core scenarios (CHiME-6, DiPCo, MX6) as well its macro average. It can be clearly seen that CHiME-6 is the scenario with the highest JER for all submissions followed by DiPCo. Also, C7DASR STCON system obtains higher JER on MX6 than the other top systems, and comparable to the C7DASR ESPnet baseline one, indicating sub-par diarization performance. It is evident the degradation in JER and thus in tcpWER (previous Fig. 6) for the two C8DASR baselines. This is because their parameters have been re-tuned to handle also the NOTSOFAR-1 scenario, and thus make significant speaker counting errors on the other scenarios. Again, this fact underscores the difficulty of achieving a good performance balance across all four scenarios.

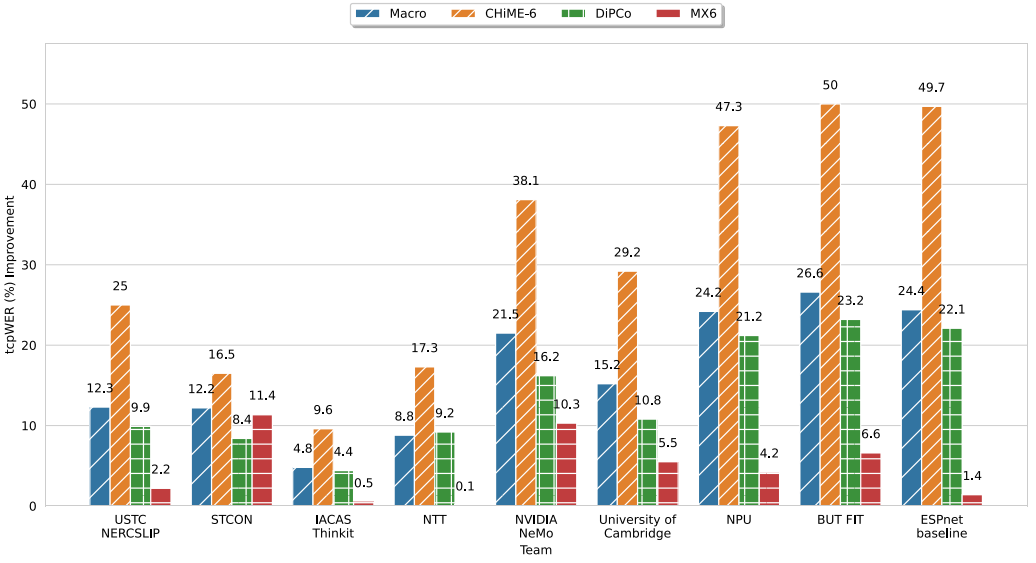

Fig. 12. tcpWER improvement (%) comparing the oracle diarization sub-track to the main track (non-oracle diarization, Fig. 6) across C7DASR core scenarios (CHiME-6, DiPCo, MX6) and their macro average.

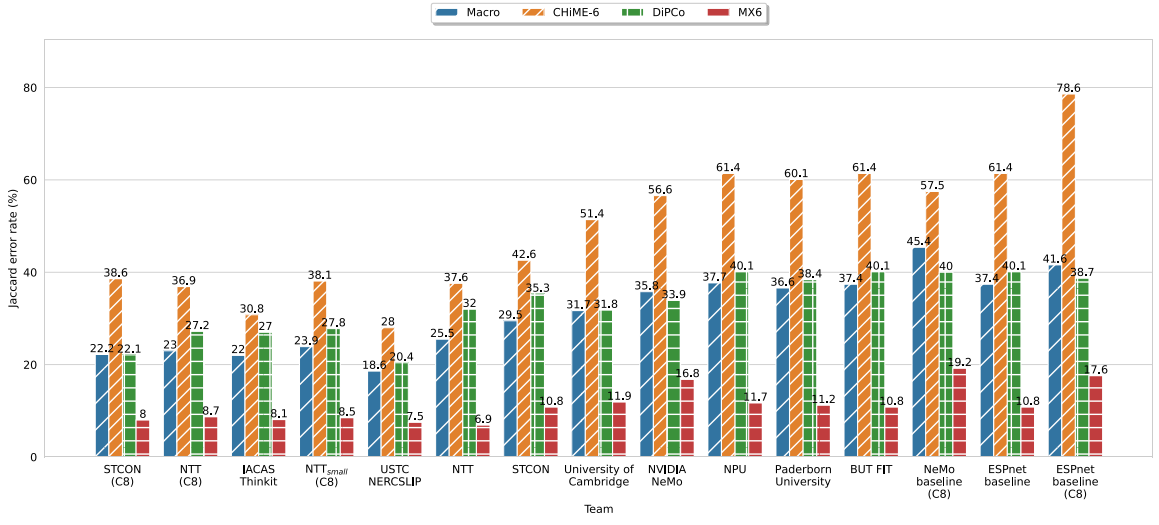

Fig. 13. JER (%) for each C7DASR core scenarios (CHiME-6, DiPCo, MX6) as well its macro average (Macro) for both C7DASR and C8DASR systems. C8DASR systems are denoted by (C8).

Similarly, USTC-NERCSLIP ranks first with oracle diarization, but its performance then degrades more compared to the IACAS-Thinkit submission. However, in this case, the degradation was mainly due to wrong word-level segmentation. In terms of DA-WER, the USTC-NERCSLIP team won and ranked first in the C7DASR challenge.[12] However, in Fig. 6, it is more penalized by tcpWER as it makes more utterance-wise timestamp errors. This appears to be the result of employing forced alignment during inference. In fact, this system uses a rather complex iterative inference procedure in which GSS is re-ran after a first inference pass, after using forced alignment with respect to ASR outputs in order to derive more precise word boundaries and thus speaker activity which could be re-used for GSS guidance. While this approach aims to enhance accuracy, it may introduces vulnerabilities: if the ASR fails to recognize certain words correctly, the forced alignment may produce outlier word segments, leading to timestamp errors and affecting the overall tcpWER. JER as observed in Fig. 13 is not affected, due to the fact that the collar mitigates such errors which consists mostly of single words. This phenomenon may be the reason why the CHiME-8 NOTSOFAR-1 USTC-NERCSLIP

[12] See leaderboard: https://www.chimechallenge.org/challenges/chime7/task1/results.

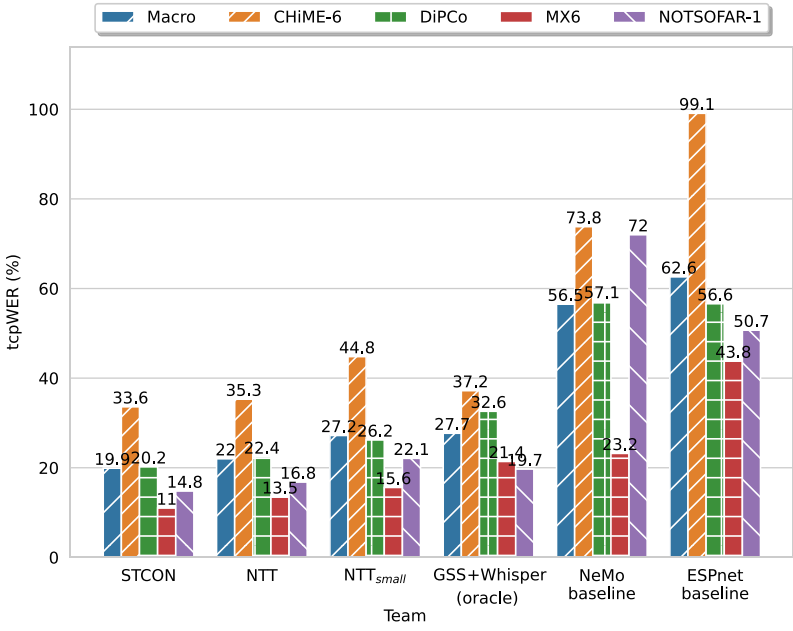

Fig. 14. C8DASR results. We report tcpWER (%) for each C8DASR core scenario (CHiME-6, DiPCo, MX6 and NOTSOFAR-1) as well as its macro average (Macro).

submission (Niu et al., 2024), which is very similar to their C7DASR submission in terms of pipeline structure, no longer employs this approach.

5.2. CHiME-8 DASR results

In Fig. 14 we plot the final results for the C8DASR challenge. As mentioned, the challenge saw limited participation from 3 teams, also due to the fact that many participants split between the two CHiME-8 DASR and NOTSOFAR-1 challenges. An oracle-based system, consisting of oracle diarization, GSS and Whisper `large-v3` GSS+Whisper (oracle) is also added in the figure as a reference.

The STCON system achieved the best overall performance, with the NTT system closely following. Both systems have fairly well-balanced performance in all four evaluation scenarios. Interestingly, overall, the tcpWER for NOTSOFAR-1 is, for these two systems, only slightly higher than the one obtained on MX6, indicating very high robustness in speaker counting capability, at least for relatively high SNR environments. The NTT$_{small}$ system overall does not exhibit much degraded performance despite not using any diarization refinement and only a single ASR model.

The two baseline systems perform significantly worse, primarily due to incorrect speaker counting, as previously discussed in the context of C7-8DASR JER results shown in Fig. 13. It should be noted that the two baselines display contrasting behaviors on the NOTSOFAR-1 dataset. The NeMo baseline exhibits very weak performance in this scenario, as it tends to underestimate the number of speakers (Cornell et al., 2024c). The ESPnet baseline instead performs better on NOTSOFAR-1 but then underestimates the number of speakers in the CHiME-6 scenario. Since both baselines rely on classical speaker counting strategies, they exhibit a "zero-sum game" behavior when faced with such diverse scenarios. Improvements in speaker counting for the CHiME-6 scenario often lead to a deterioration in performance for NOTSOFAR-1 and vice-versa, due to their contrasting session-level characteristics: CHiME-6 has 2+ hour sessions with 4 speakers, while NOTSOFAR-1 has ∼10 min meetings with 4–8 speakers.

As mentioned in Section 3.3, C8DASR featured an additional optional track where several pre-trained LLM models were also allowed. As briefly mentioned in Section 3, only the STCON team participated in this sub-track with an approach based on Llama-2 and similar to Ogawa et al. (2024). However, as mentioned, this resulted in very limited performance gains: a reduction of only 0.5% in absolute macro-tcpWER compared to their main track submission.

In general, the best-performing systems (STCON and NTT) are remarkable in being superior even to the GSS+Whisper (oracle) system (which uses oracle diarization). This includes even the NTT$_{small}$ system, despite the fact that it does not use any ASR ensembling technique and no diarization refinement.

5.3. Joint CHiME-8 DASR and NOTSOFAR-1 challenges results

Some of the C7DASR participants did not participate in C8DASR but only in the "twin" time-concurrent CHiME-8 NOTSOFAR-1 challenge that, as mentioned, focused exclusively on the NOTSOFAR-1 scenario. In C8DASR, this latter is instead one of the four

Table 6
Summary of diarization and front-end components for CHiME-8 NOTSOFAR-1 challenge multi-channel track. Top panel (grey): baseline system; bottom panel: participant submissions. Within each panel, the participants' systems are ordered according to macro-averaged tcpWER. C7-8DASR baselines are omitted here. STCON and NTT systems are the same across NOTSOFAR-1 and CHiME-8 challenges. C8DASR submissions are in yellow .

Team	Diarization				Front-End	
	Segmentation	Spk-id Extr. & Clustering	Refinement	Multi-Channel Mechanism	Channel Selection	Separation
NOTSOFAR-1 Baseline	Whisper word timestamps	TitaNet + NME-SC	–	–	–	Conformer CSS + MVDR
USTC	Conformer CSS/OSD + JDS + Enhanced Whisper word timestamps	ResNet-221 + NME-SC	NSD-MS2S + JDS	–	–	WPE + Conformer CSS/OSD + GSS + JDS refinement
STCON (C8)	TDNN stats-based	ECAPA-TDNN & wav2vec 2.0 AED multi-speaker OSD with UMAP+DBSCAN+GMM clustering		DOVER-Lap and TS-VAD posterior averaging	EV	GSS with neural refinement (G-TSE)
NTT (C8)	EEND-VC	EEND-VC & ECAPA-TDNN with NME-SC	NSD-MS2S	DOVER-Lap & channel clustering for speaker counting	EV & Brouhaha C_{50}	GSS with SP-MWF
NPU	Silero + Whisper word timestamps	ResNet293 + NME-SC + speaker merging	–	–	–	WavLM Conformer CSS + MVDR
NAIST	Whisper word timestamps	TitaNet + NME-SC	–	–	–	WPE + Conformer CSS + MVDR
BUT/JHU	WavLM multi-channel local EEND	ECAPA-TDNN + NME-SC	–	TAC-layers	–	–

scenarios participants should address (together with CHiME-6, DiPCo and MX6). As mentioned in Section 1, the main motivation for having both challenges was to compare across the two. The idea was to assess how "generalist", recording setup agnostic solutions can measure up with respect to domain-specific ones, considering that the latter can use a-priori information about the array configuration and the deployment domain.

CHiME-8 DASR and CHiME-8 NOTSOFAR-1 joint results are in Fig. 15 where we plot, for each system, the tcpWER on the NOTSOFAR-1 scenario. For NOTSOFAR-1, we report only the best systems from the multi-channel track and not those from the single-channel one, which, as expected, have slightly worse performance. Note that the values are different from those reported in the official NOTSOFAR-1 challenge results, as we accumulate error statistics rather than average tcpWER across the sessions as done in the CHiME-8 NOTSOFAR-1 challenge. The reason is that averaging tcpWER per session biases the overall figure on the easier sessions, which have fewer words and also fewer participants. However, it is worth mentioning that the overall system rankings remain consistent between both computation methods, with differences in absolute tcpWER values of only 1–2%.[13] The CHiME-8 NOTSOFAR-1 challenge had 5 teams participating, in alphabetical order: Blue Sky Wave Riders, BUT JHU (Polok et al., 2024a), NAIST (Hirano et al., 2024), NPU (Huang et al., 2024) and USTC-NERCSLIP (Niu et al., 2024).

In Table 6, we summarize the front-end and diarization techniques employed by participants to aid in the analysis of results, following a similar approach to that used in Table 4 for the C7-8DASR challenges. For the reader's convenience, the characteristics of the C8DASR STCON and NTT submissions are also included here, marked with a (C8) tag. C8DASR baselines are omitted, as they have already been discussed in ample detail.

Notably, compared to C7-8DASR, the approaches in this context are more varied. They are divided among three main pipelines: diarization → GSS → ASR (adopted by USTC-NERCSLIP, STCON, and NTT), separation → ASR → diarization (NOTSOFAR-1 baseline, NAIST and NPU), and a simpler diarization → ASR pipeline, as used by the BUT JHU submission. The former, as said, has been the prevalent approach for CHiME-6 and CHiME-7 DASR challenges; the second is similar to Yoshioka et al. (2019), Raj et al. (2021a), Kanda et al. (2022a) and here is adopted mainly because it is the approach that the NOTSOFAR-1 challenge baseline uses (Vinnikov et al., 2024). The latter uses a Conformer-based CSS model, which is applied without VAD on the entire session. After CSS, Whisper `large-v2` is employed for recognition and word-level segmentation, as it can predict word-level timestamps. Such segmentation is then used to extract TitaNet embeddings which are then clustered with NME-SC to obtain also speaker-attribution. NPU (Huang et al., 2024) and NAIST (Hirano et al., 2024) teams adopt the NOTSOFAR-1 baseline CSS model. NPU modifies it by using WavLM as a feature extractor, while NAIST uses MIMO-WPE before CSS to enhance dereverberation capabilities. For the backend, NPU used Whisper `large-v2` but fine-tuned it using AdaLoRA (Zhang et al., 2023). NAIST instead uses K2 Zipformer with WavLM `large` as the frontend, reporting a great speed up in inference time over the baseline and better performance.

The USTC-NERCSLIP submission (Niu et al., 2024), also uses the same CSS component as the NOTSOFAR-1 baseline, but only within its diarization sub-module and for diarization purposes as they extend it to perform also overlapped speech detection jointly with separation. The system has some similarities with their C7DASR submission and in fact heavily relies on NSD-MS2S TS-VAD diarization refinement followed by GSS. GSS is initialized by a neural TSE model, similarly to the C7DASR IACAS-Thinkit system but with the difference that here this neural TSE model is integrated with the NSD-MS2S TS-VAD in a single joint diarization-separation

[13] Results with NOTSOFAR-1 tcpWER averaging: https://www.chimechallenge.org/challenges/chime8/task2/results#notsofar-and-dasr-results-supplementary.

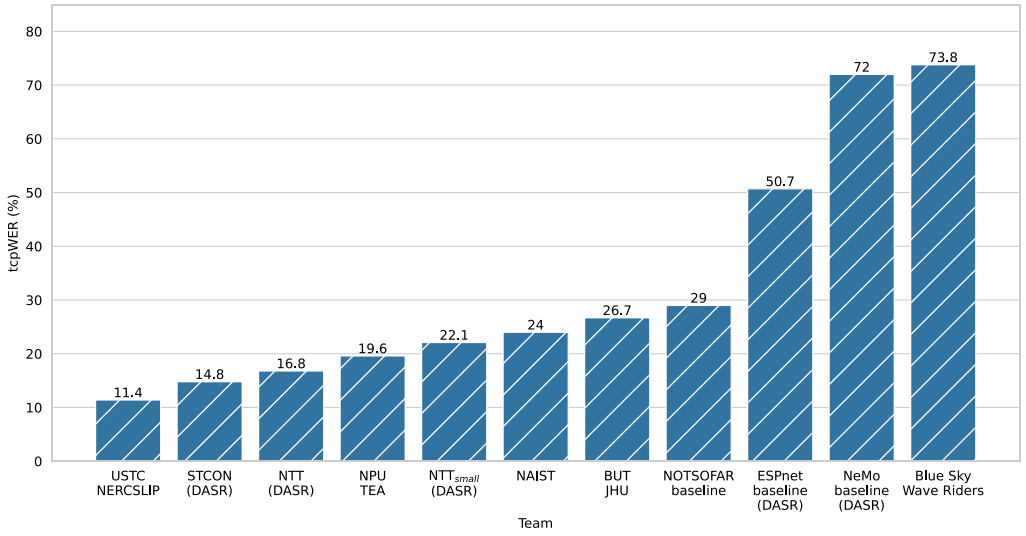

Fig. 15. tcpWER (%) for CHiME-8 NOTSOFAR-1 (multi-channel track) and DASR challenges systems on the NOTSOFAR-1 scenario.

model (JDS). The back-end is a heavily modified version of Whisper (both `large-v2` and `large-v3` are used, as the final system is an ensemble), which is augmented/modified with RoPE positional encodings (Heo et al., 2025), a mixture of expert component from You et al. (2021) and even WavLM `large` features.

The BUT-JHU (Polok et al., 2024a) system instead significantly departs from the baseline system by not relying at all on explicit separation. This system uses a target speaker augmented Whisper model (Polok et al., 2024b), which can perform TSE implicitly, without relying on techniques such as GSS. For diarization, they use a WavLM-based local EEND model (Han et al., 2025) similar to the one employed in the Pyannote diarization pipeline 2.1 (and thus the ESPnet baseline).

Regarding the three C8DASR systems, we observe that they compare favorably with the CHiME-8 NOTSOFAR-1 submissions, despite also addressing three additional scenarios, two of which are arguably more challenging, as it is evident from tcpWER figures in Fig. 14. More broadly, all top-three systems utilize GSS and diarization refinement through the NSD-MS2S TS-VAD model, reaffirming the effectiveness and versatility of these techniques. Furthermore, this highlights, again, the superior robustness of GSS compared to current CSS or purely neural TSE methods, even, crucially, when the domain and array configuration is known and it is not ad-hoc. As such, there is a clear need for further research in this area, particularly by focusing on self/semi-supervised learning techniques (Wisdom et al., 2020; Bando et al., 2024).

Among the C8DASR systems, NTT$_{small}$ is particularly noteworthy, as it does not employ any ensembling techniques (unlike most other submissions, except BUT JHU) or diarization refinement. The BUT JHU system also shows promise due to its streamlined pipeline. However, it falls short of the performance achieved by other submissions. One limitation is that its separation is handled implicitly within the ASR system, limiting its ability to fully leverage multi-channel information.

5.3.1. System evaluation on meeting summarization for the NOTSOFAR-1 scenario

The NOTSOFAR-1 dataset, described in Section 3.1.4, differs from the other scenarios by focusing on structured office meetings centered on specific topics. This feature makes it well-suited for meeting summarization tasks. In contrast, the other three C8DASR scenarios represent purely conversational, speech-in-the-wild settings without particular structure, making them less appropriate for gauging downstream summarization performance. As mentioned in the introduction, summarization is an important downstream task for joint diarization and ASR which has recently gained even more traction due to the rise of LLMs. It is thus interesting to evaluate how much the quality of meeting summaries can depend on accurate transcription and correct speaker attribution. In numerous practical settings, such as routine office meetings, perfect verbatim transcription may be unnecessary, while concise, accurate meeting summaries remain highly desirable. For these applications, traditional ASR and diarization metrics derived from WER (e.g. tcpWER, cpWER) or DER/JER may inadequately reflect actual user experience and over-estimate the amounts of errors. This is especially true for spontaneous conversational speech where it is common to find contraptions, hesitations, false starts, filler words, and repetitions that may be counted as errors but have minimal impact on the overall meaning (Kim et al., 2021). This is particularly relevant considering that the LLM summarization process can correct or compensate for certain inaccuracies in transcription and speaker attribution, producing useful summaries despite imperfect input. On the other hand, text summarization has other issues such as the fact that the evaluation is difficult due to the inherent subjectivity of what constitutes a “good” summary (Lee et al., 2024).

For this evaluation, we used the Gemini 2.0 Flash (Georgiev et al., 2024) LLM to generate summaries for each NOTSOFAR-1 session starting from each submitted transcript. We prompt the LLM to generate summaries of approximately 200 words. We selected this 200 words limit because NOTSOFAR-1 meetings are relatively short (averaging 6 min as shown in Table 2, Section 3.1)

with ground-truth transcriptions averaging ~1600 words. This summary length was considered a reasonable trade-off between compression and content preservation.

The following prompt was used, where `{transcription}` was replaced with the actual system transcription:

```
You are provided with the transcription of a meeting
involving a minimum of two to maximum eight participants.

Please create a comprehensive summary (approximately 200 words)
of the provided meeting transcription. Your summary should capture
all key points, decisions, action items, and significant exchanges.
The transcription has speakers labeled as [speaker 1],
[speaker 2] and so on.
Your summary must maintain these speaker attributions,
clearly indicating who said what, proposed ideas, or took on action items.
Meeting transcription: {transcription}
Summary (write only the requested summary hereafter):
```

This prompt encourages the LLM to preserve speaker attributions in the summaries, which is a highly desired feature in most real-world applications. Each system transcript was pre-processed using the C8DASR scoring text normalization procedure (Section 3.3.1). Then, a further processing step was performed to add speaker-id tags at each utterance in addition to the recognized words, i.e.:

```
[speaker 1] ok [speaker 2] but yeah [speaker 1] yes what i was
suggesting that we [speaker 3] it is fine so if you think that we can
move forward.
```

We used the convention of `[speaker 1]` being the first speaker to speak in the whole meeting, `[speaker 2]` the second, and so on. We did not include start and end timestamps for each utterance as it is not critical for meeting summarization. Note that these speaker-id tags might be misaligned between reference and hypothesis transcriptions due to the fact that the hypothesis diarization usually contains errors. To address this issue, we used tcpWER to derive optimal reference vs. hypothesis speaker-id assignments for each speaker and re-assign the hypothesis speaker-id tags based on this optimal matching with respect to the ground truth transcript. This step ensures that the ground truth derived summaries and the hypothesis transcripts summaries have aligned speaker-id labels and thus our summarization evaluation is well defined. Note that this step still accounts for missed and false alarm speakers. False alarm speakers would still be present in the hypothesis transcript and would be assigned tags incrementally as `[speaker N]` with $N > S$ where S is the number of speakers in both the reference and hypothesis transcriptions. We acknowledge that this speaker alignment approach has potential limitations. The optimal permutation is computed at the transcript level using tcpWER, rather than at the summary level across all possible permutations. While the latter would be theoretically preferable, it is computationally prohibitive: for a meeting with 6 speakers, this would require 720 (6!) metric evaluations per summary. The proposed approach represents a practical trade-off between theoretical optimality and computational feasibility. Moreover, we also experimented with using alternative labeling schemes such as sequential letters A, B, C (in randomized order) and even the pseudonyms that participants used in the NOTSOFAR-1 challenge (e.g. Walter, Melissa etc.). However, we did not observe an appreciable difference in the results, suggesting that the LLM relies primarily on content rather than speaker label formatting for summarization.

To measure downstream meeting summarization performance, we employ complementary metrics to minimize potential evaluation biases. These include general LLM prompt-based approaches, specialized LLM-based methods, and n-gram based techniques, which are detailed below.

- G-Eval (Liu et al., 2023): a SotA automatic meeting summarization technique that leverages an LLM (OpenAI GPT-4o (OpenAI, 2024)) and prompt engineering in order to evaluate the summarization performance. G-Eval is summary reference free and uses only the system output transcript and the original ground truth transcription. It measures:
  - *Coherence*: assesses how well-structured and logically connected the summary is, evaluating whether information flows naturally and maintains clear relationships between ideas.
  - *Consistency*: measures if the summary contains only factual information entailed in the original transcript.
  - *Relevance*: evaluates how well the summary captures the most important and salient information from the original meeting transcript.
  - *Fluency*: assesses the linguistic quality of the summary, including grammatical correctness, natural language use, and readability.

  G-Eval scores are natural numbers in the range $[1, 5]$ ($[1, 3]$ for *fluency*), with higher values indicating better performance.

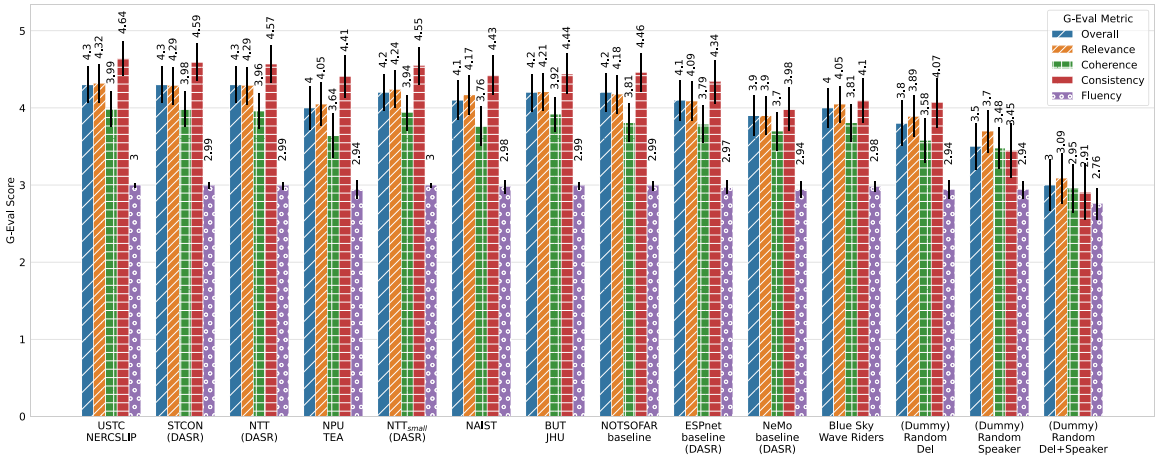

Fig. 16. G-Eval scores for summarization downstream task on the NOTSOFAR-1 scenario for each team best performing system in terms of tcpWER. The *overall* score here is the average of *coherence*, *relevance*, and *consistency* metrics. Error bars indicate standard error for each metric.

- UniEval (Zhong et al., 2022): A unified evaluation framework that fine-tunes pre-trained LM models (which are T5-based (Raffel et al., 2020)) and evaluates text generation across multiple dimensions. Among other tasks, it is also specifically designed for summarization and, as G-Eval, also evaluates *coherence*, *conciseness*, *relevance*, and *fluency* of the system summary. Unlike G-Eval, UniEval also uses the reference summary on top of the original meeting transcription to calculate these scores. UniEval scores are normalized real numbers in the range $[0, 1]$, with higher values indicating better performance.
- ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004): standard lexical overlap metrics that measure unigram, bigram, and longest common subsequence matches between the generated and reference summaries, respectively. For every ROUGE metric, here we report only the F1 scores.

As mentioned, meeting summarization suffers from the fact that the evaluation procedure is inherently "noisy" due to its intrinsic subjective nature. To mitigate this issue and ensure more reliable evaluation, we employed multiple reference summaries and multiple system hypothesis in order to be able to compute standard error bounds. As reference summaries, for each NOTSOFAR-1 session, we generated 8 different summaries obtained by using ground-truth transcription by varying the seed at each Gemini 2.0 flash inference run. These are needed for UniEval and ROUGE metrics. Similarly, from each submitted system transcript, we generate 8 different summaries using same process. This allows us to consider, for UniEval and ROUGE metrics, 64 hypothesis and reference combinations per meeting summary. Instead, since G-Eval is reference free, we only consider 8 different evaluation scores per meeting summary. This is also reasonable, as, in fact, this latter is also more expensive to run.

In Fig. 16 we report the G-Eval metrics for each NOTSOFAR-1 submitted system, including C8DASR and those submitted in the CHiME-8 NOTSOFAR-1 challenge. We only included the best system in terms of tcpWER for each team and order the teams in the plot based on their tcpWER ranking to be consistent with previous Fig. 15. For reference, we include results for several "dummy" submissions with artificially introduced errors in the ground-truth transcription:

- *Random Del*: each word in the ground-truth transcription is randomly removed with 50% probability. The resulting tcpWER is 49.83%.
- *Random Speaker*: speaker IDs in the ground-truth transcription are randomly reassigned for each utterance. The resulting tcpWER is 114%.
- *Random Del+Speaker*: a combination of both error types: speaker IDs are randomly reassigned for each utterance, and each word is randomly removed with 50% probability. The resulting tcpWER is 101%. This value is lower than the Random Speaker condition because fewer words are incorrectly attributed to different speakers, as many are now counted only as deletions. In tcpWER, speaker-word misattributions count as two errors: once as a deletion and once as an insertion.

We can observe that, despite the evident tcpWER difference observed in Fig. 15, for which, the best system (USTC-NERCSLIP) had a tcpWER ~11% while the worst ~70%, the G-Eval metrics show reduced variation between submissions. That is, even systems for which roughly only one out of two words is correctly recognized and speaker-attributed (e.g. ESPnet baseline) appear to generate summaries which are roughly on-par with systems for which ~80% of the words are correctly recognized and attributed (e.g. NPU TEA). In fact, most differences between systems in Fig. 15 can be considered to be not significant enough since they fall within one standard error range. Only when the amount of errors are extremely severe as in the Random Del and Random Del+Speaker systems an appreciable degradation in summarization metrics can be observed. Among the G-Eval metrics, *fluency* is the one that is most consistent across all submissions, including the "dummy" systems and virtually shows no variation. This same trend is also observed when computing ROUGE scores which are reported in Fig. 17. Crucially G-Eval consistently assigns a high *fluency* score to

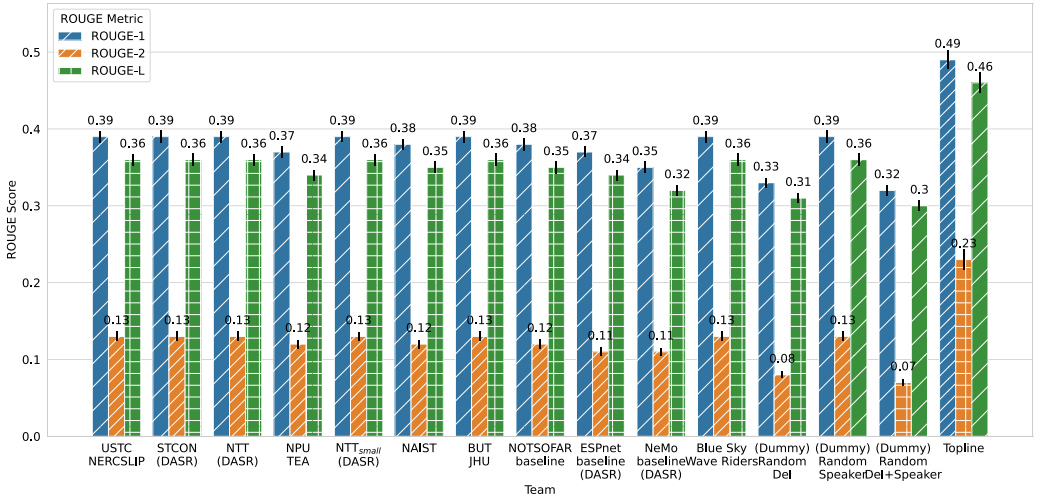

Fig. 17. ROUGE-1, ROUGE-2 and ROUGE-L F1-scores for summarization downstream task on the NOTSOFAR-1 scenario for each team best performing system in terms of tcpWER. Error bars indicate standard error for each metric. The "topline" system on the right ("Reference vs. Reference") shows ROUGE scores between reference summaries generated from ground-truth transcriptions with different seeds serves as an approximate upper bound for evaluation.

Table 7
Pearson correlation coefficients (PCC) and Spearmank rank correlation (SRC) between tcpWER and G-Eval, UniEval and ROUGE F1 scores metrics.

Summarization metric	PCC	SRC
G-Eval Overall	−0.51	−0.5
G-Eval relevance	−0.46	−0.45
G-Eval consistency	−0.54	−0.55
G-Eval coherence	−0.27	−0.28
G-Eval fluency	−0.22	−0.13
UniEval overall	−0.15	−0.18
UniEval relevance	−0.11	−0.16
UniEval consistency	−0.07	−0.10
UniEval coherence	−0.12	−0.20
UniEval fluency	0.01	−0.01
ROUGE-1 F1 score	−0.34	−0.36
ROUGE-2 F1 score	−0.28	−0.32
ROUGE-L F1 score	−0.33	−0.35

all systems even for the Random Del and Random Del+Speaker "dummy" systems. This is due to the fact that all our summaries were generated using the same Gemini Flash 2.0 model, and modern SotA LLMs have become increasingly good at generating linguistically well-formed output, thus saturating these fluency summarization metrics. It is also possible that there is a potential systemic bias where LLM evaluators tend to rate LLM-generated text highly on fluency metrics (Liu et al., 2023) generated via Gemini Flash 2.0. However, upon inspection of the generated summaries, we found no noticeable issues and all inspected summaries appeared to be well written. Table 7 reports PCC and SRC of tcpWER vs. G-Eval, ROUGE and also UniEval scores. These correlation coefficients are computed considering each submitted system and each session independently for a total of 1760 samples. For this study we did not consider including the "dummy" systems with artificially generated errors. These figures indicate that only G-Eval has a moderate negative correlation with tcpWER with *consistency* being the most correlated. This can be expected since it quantifies if the summary contains only factually correct information that was contained in the original transcript, thus penalizing hallucinated content and speaker mis-attributions due to transcription errors. For ROUGE and even more so for UniEval metrics, the correlation is much weaker. To give a reader a sense of the weakness of the correlation between tcpWER and G-Eval, in Fig. 18 we report also a scatter plot for all submitted systems (top: all systems, bottom: top 5 systems) of tcpWER vs. G-Eval *overall* metric. Each data-point corresponds to one NOTSOFAR-1 session in the evaluation set.

These results underscore the challenges of evaluating transcription systems through meeting summarization, as the use of LLMs introduces an additional black-box component that can significantly bias and influence the results. When an application requires reliable or even acceptable transcription quality, meeting summarization may not be a good proxy evaluation task due to its high robustness to transcription errors. Among the metrics we have analyzed, only G-Eval showed moderate correlation with transcription errors. Thus, despite promising advantages in handling text normalization and common linguistic artifacts in spontaneous speech,

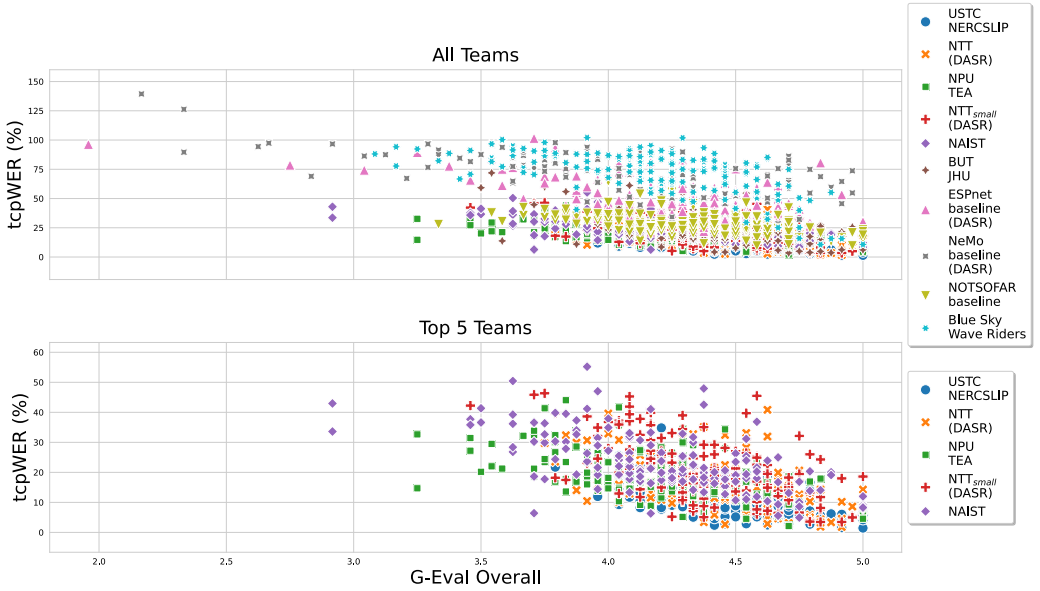

Fig. 18. Scatter plot of tcpWER (%) vs. G-Eval Overall scores for all submitted systems on the NOTSOFAR-1 scenario. Top: all systems; Bottom: top-5 systems by tcpWER. Each point represents one NOTSOFAR-1 evaluation session.

this same robustness makes summarization currently sub-optimal as a proxy evaluation task for transcription quality. At least until better summarization metrics or ad-hoc evaluation frameworks are developed for dialogue-oriented summarization, though G-Eval already shows improved correlation compared to legacy metrics like UniEval and ROUGE. For example, in our preliminary experiments, we found it crucial to include explicit speaker attribution instructions in the LLM summarization prompt; without these, the produced summaries were too generic, and no appreciable correlation was observed across all summarization metrics, including G-Eval. We hope that this observation and, more generally, the summarization evaluation framework adopted in this work can be a valuable contribution towards improving dialogue summarization evaluation, but further research in this direction remains essential.

From a more optimistic perspective, these findings can also be interpreted in another way. The near-consistent performance across all systems and metrics demonstrates that current LLM-based summarization approaches are remarkably powerful in recovering from transcription errors and generating accurate summaries. This suggests that some degree of decoupling between transcription and effective summarization exists, challenging traditional assumptions about the dependency chain in speech processing pipelines. As such, if the application at hand does not strictly require accurate transcription, these findings could also motivate research toward fully end-to-end spoken meeting summarization. An ad-hoc end-to-end approach could perhaps even offer more contextually aware summarization capabilities as the raw audio is much richer in content and may offer additional cues that are unavailable in transcripts. An example is speech prosody indicating surprise, emphasis, or hesitation. This promising research direction has been enabled by recent advances in multimodal audio and text LLM models (SpeechLMs) (Cui et al., 2024; Peng et al., 2024) and is already supported by models such as Gemini, GPT-4o, and Llama 3 (Fang et al., 2024). However, practical challenges remain, particularly regarding reliable long-context modeling to effectively capture extended meeting interactions and keyfacts. These problems are amplified for SpeechLMs as the sequence length increases dramatically. As a consequence, computational demands (and thus cost) also increase substantially, presenting a critical barrier to widespread adoption of these technologies in many practical settings.

6. Conclusions

6.1. Summary of challenges outcomes and findings

In this work, we described the motivations, design, and outcomes of the recent CHiME-7 and CHiME-8 distant automatic speech recognition (DASR) challenges. The core objective was to promote the development of joint ASR and diarization transcription systems capable of generalizing across diverse scenarios by testing their performance on multiple datasets: CHiME-6, DiPCo, Mixer 6, and NOTSOFAR-1 (with NOTSOFAR-1 added in the CHiME-8 challenge). These datasets encompass significant variations in acoustic environments, recording setups, meeting durations, participant numbers, and linguistic and paralinguistic speaking styles. A total of 9 teams participated in the two challenges, submitting 33 different transcription systems. Our analysis of the techniques employed by participants spans both front-end components (including diarization and speech separation) and back-end ASR components, revealing several key trends.

First, the transition towards e2e ASR systems, facilitated by the availability of large-scale pretrained models, represents a significant evolution from the previous CHiME-6 challenge. Second, guided source separation (GSS) remains a crucial component in most high-performing systems, as evidenced by submissions across both the CHiME-7/8 DASR and CHiME-8 NOTSOFAR-1 multi-channel track challenges. Few participants attempted to integrate neural-based speech separation methods, as challenges persist in reliably handling the complexities of real-world conversational environments.

Third, cross-comparison of results between CHiME-8 DASR and CHiME-8 NOTSOFAR-1 challenges revealed that generalizable array-agnostic systems can achieve comparable or competitive results versus ad-hoc systems developed for specific narrow scenarios or known microphone array configurations. This is a significant finding, as more flexible systems are highly desirable for real-world applications.

Fourth, with few exceptions, most high-performing systems heavily relied on diarization refinement via target-speaker voice activity detection (TS-VAD) techniques. These approaches, together with the availability of large-scale pre-trained models, were the primary drivers of performance improvement. The improved TS-VAD techniques such as the NSD-MA-MSE model from USTC-NERCSLIP can be considered a significant outcome of the CHiME-7 DASR challenge. Finally, accurate speaker counting remains crucial, as errors propagate catastrophically through the rest of the pipeline. In this regard, novel methods were developed during the CHiME-8 DASR challenge by NTT and STCON teams, featuring improved clustering techniques and overlapped speech handling in the diarization component. However, developing systems more robust to speaker counting errors remains an important and practical research direction. As discussed in Section 5.1, such errors have catastrophic downstream effects in typical pipelined approaches. Few works have addressed this issue, for example by making TS-VAD robust to such speaker counting errors (Wang et al., 2024a).

As an additional contribution, we explored the suitability of meeting summarization as a downstream evaluation task for meeting transcription using NOTSOFAR-1 data. We outlined an evaluation framework where we used an LLM to generate summaries from system transcriptions while ensuring and measuring correct speaker attribution. Our results indicate that recent summarization metrics, such as G-Eval, show moderate correlation with transcription accuracy. However, we also found that contemporary LLMs demonstrate remarkable ability in handling transcription errors and inferring missing information. This resilience suggests that summarization is not a suitable proxy evaluation task for applications where precise transcription accuracy is crucial and, instead, justify the research direction of direct e2e meeting summarization, when verbatim transcription is not needed.

6.2. Limitations and future research directions

Despite the contributions of the CHiME-7 and CHiME-8 DASR challenges, there are several limitations that warrant further discussion. These limitations provide valuable insights for future research and the design of more effective evaluation campaigns.

First, the datasets used in the challenges, while diverse, may not fully capture the complexities of all real-world conversational environments. The scenarios considered are still limited in scope, and moreover, they only consider the English language. More effort should be done in the future to also consider the possibility of multilingual meeting transcription that can also be robust to code switching so that this technology can benefit a larger amount of people.

Second, most importantly, to truly gauge system generalization, future benchmarks should consider having at least some fully blind evaluation scenarios for which the domain is not known a-priori by the participants. In the C7-8DASR challenges, all the evaluation scenarios were already known to the participants and this could have biased the development of submitted systems. On the other hand, in real-world applications, some deployment conditions/domains are not known a-priori. Future challenges should include evaluation scenarios that are completely hidden from participants until submission time to better simulate real-world deployment conditions.

Third, while our downstream evaluation via meeting summarization provided some valuable insights, several limitations remain. Our speaker alignment approach (using tcpWER-based optimal permutation before summary generation) represents a practical trade-off given computational constraints. Moreover, developing more robust summarization evaluation metrics that better capture the nuances of multi-speaker dialogue remains an open challenge. However, this will require specialized data collection efforts with human-annotated reference summaries that explicitly ground key facts, decisions, and action items. Finally, while the participants' submissions show promise, as e.g. best submissions on CHiME-8 DASR challenge were even able to surpass Whisper `large-v3` applied to oracle diarization and GSS, most rely on rather impractical ensembling techniques. While some mechanisms were introduced in the CHiME-8 DASR challenge to spur research towards more efficient and practical approaches, the resulting systems are still far from being practical. For example, among all submissions in both C7-8DASR challenges, NTT_{small} is the one that achieved the best compromise between performance and efficiency, by avoiding ensembling and diarization refinement techniques. However, as reported by the authors, the RTF was more than 2, even if this could be brought down by ad-hoc optimizations. Thus, future evaluation campaigns should also try to devise ways to better encourage the practicality of developed techniques.

CRediT authorship contribution statement

Samuele Cornell: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Christoph Boeddeker:** Writing – review & editing, Supervision, Software. **Taejin Park:** Writing – review & editing, Software. **He Huang:** Software. **Desh Raj:** Writing – review & editing, Validation, Supervision, Software, Data curation. **Matthew Wiesner:** Writing – review & editing, Validation, Supervision, Data curation. **Yoshiki Masuyama:** Writing – review & editing, Validation, Software. **Xuankai Chang:** Writing – review & editing, Software. **Zhong-Qiu Wang:** Writing – review & editing, Software. **Stefano Squartini:** Writing – review & editing, Supervision. **Paola Garcia:** Writing – review & editing, Validation, Supervision. **Shinji Watanabe:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Samuele Cornell reports financial support was provided by Oak Ridge Institute for Science and Education. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank the CHiME Steering Committee members Marc Delcroix, Michael Mandel, and Jon Barker for their invaluable help, support, and guidance in organizing these challenges. S. Cornell was supported by the IC Postdoctoral Research Fellowship Program via Oak Ridge Institute for Science and Education, United States through an agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence, United States.

Data availability

All data and code is open source. Participants submissions are available in the CHiME website.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2623–2631.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G., 2020. Common voice: A massively-multilingual speech corpus. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4218–4222.

Arora, A., Raj, D., Subramanian, A.S., Li, K., Ben-Yair, B., Maciejewski, M., Żelasko, P., Garcia, P., Watanabe, S., Khudanpur, S., 2020. The JHU multi-microphone multi-speaker ASR system for the CHiME-6 challenge. In: CHiME Workshop.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. In: NeurIPS.

Bando, Y., Nakamura, T., Watanabe, S., 2024. Neural blind source separation and diarization for distant speech recognition. In: Interspeech.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third CH-iME speech separation and recognition challenge: Dataset, task and baselines. In: IEEE ASRU.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2017. The third 'CHiME'speech separation and recognition challenge: Analysis and outcomes. Comput. Speech & Lang. 46, 605–626.

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2013. The PASCAL CHiME speech separation and recognition challenge. Comput. Speech & Lang. 27.

Barker, J., Watanabe, S., Vincent, E., Trmal, J., 2018. The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines. In: Interspeech.

Boeddeker, C.B., Cord-Landwehr, T., Neumann, T.v., Haeb-Umbach, R., 2023. Multi-stage diarization refinement for the chime-7 DASR scenario. In: CHiME Workshop. pp. 51–56.

Boeddeker, C., Heitkaemper, J., Schmalenstroeer, J., Drude, L., Heymann, J., Haeb-Umbach, R., 2018. Front-end processing for the CHiME-5 dinner party scenario. In: CHiME5 Workshop.

Boeddeker, C., Subramanian, A.S., Wichern, G., Haeb-Umbach, R., Le Roux, J., 2024. TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings. IEEE/ACM Trans. Audio, Speech, Lang. Process. 32, 1185–1197.

Brandschain, L., Graff, D., Cieri, C., Walker, K., Caruso, C., Neely, A., 2010. The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition. In: LREC.

Bredin, H., 2023. Pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In: Interspeech. ISCA, pp. 1983–1987.

Bredin, H., Laurent, A., 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In: Interspeech.

Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.-P., 2020. Pyannote. audio: neural building blocks for speaker diarization. In: Proc. of ICASSP.

Capon, J., 1969. High-resolution frequency-wavenumber spectrum analysis. IEEE 57 (8).

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al., 2005. The AMI meeting corpus: A pre-announcement. In: International Workshop on Machine Learning for Multimodal Interaction.

Chang, X., Maekaku, T., Fujita, Y., Watanabe, S., 2022. End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation. In: Interspeech.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., et al., 2021. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In: Interspeech.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al., 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. 16 (6).

Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y., Wu, J., Xiao, X., Li, J., 2020b. Continuous speech separation: Dataset and analysis. In: Proc. of ICASSP.

Chen, H., Zhang, P., Shi, Q., Liu, Z., 2020a. Improved guided source separation integrated with a strong back-end for the chime-6 dinner party scenario.. In: Interspeech. pp. 334–338.

Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. In: Proc. NeurIPS.

Christensen, H., Barker, J., Ma, N., Green, P.D., 2010. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In: Eleventh Annual Conference of the International Speech Communication Association.

Cieri, C., Miller, D., Walker, K., 2004. The Fisher Corpus: A resource for the next generations of speech-to-text.. In: LREC. Vol. 4, pp. 69–71.

Cooke, M., Hershey, J.R., Rennie, S.J., 2010. Monaural speech separation and recognition challenge. Comput. Speech & Lang. 24 (1), 1–15.

Cornell, S., Brutti, A., Matassoni, M., Squartini, S., 2021. Learning to rank microphones for distant speech recognition. arXiv preprint arXiv:2104.02819.

Cornell, S., Darefsky, J., Duan, Z., Watanabe, S., 2024a. Generating data with text-to-speech and large-language models for conversational speech recognition. In: SynData4GenAI Workshop.

Cornell, S., Jung, J.-w., Watanabe, S., Squartini, S., 2024b. One model to rule them all? Towards end-to-end joint speaker diarization and speech recognition. In: Proc. of ICASSP. pp. 11856–11860.

Cornell, S., Park, T., Huang, S., Boeddeker, C., Chang, X., Maciejewski, M., Wiesner, M., Garcia, P., Watanabe, S., 2024c. The CHiME-8 DASR challenge for generalizable and array agnostic distant automatic speech recognition and diarization. In: CHiME Workshop.

Cornell, S., Wiesner, M., Watanabe, S., Raj, D., Chang, X., Garcia, P., Masuyama, Y., Wang, Z.-Q., Squartini, S., Khudanpur, S., 2023. The CHiME-7 DASR Challenge: Distant meeting transcription with multiple devices in diverse scenarios. In: CHiME Workshop.
Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M., Maragos, P., 2014. The DIRHA simulated corpus.. In: LREC.
Cui, W., et al., 2024. Recent advances in speech language models: A survey. arXiv preprint arXiv:2410.03751.
Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., Zeghidour, N., 2024. Moshi: a speech-text foundation model for real-time dialogue. arXiv preprint arXiv:2410.00037.
Deng, K., Woodland, P.C., 2024. Label-synchronous neural transducer for adaptable online E2E speech recognition. IEEE/ACM Trans. Audio, Speech, Lang. Process..
Deng, K., Zheng, X., Woodland, P.C., 2023. The University of Cambridge system for the CHiME-7 DASR Task. In: Proc. CHiME 2023. pp. 73–76.
Desplanques, B., Thienpondt, J., Demuynck, K., 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In: Interspeech.
Du Bois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A., Martey, N., 2000. Santa Barbara corpus of spoken American English. CD-ROM. Phila.: Linguist. Data Consort..
Erdogan, H., Hershey, J.R., Watanabe, S., Mandel, M.I., Le Roux, J., 2016. Improved MVDR beamforming using single-channel mask prediction networks. In: Interspeech. pp. 1981–1985.
Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., Feng, Y., 2024. Llama-omni: Seamless speech interaction with large language models. arXiv preprint arXiv:2409.06666.
Fiscus, J.G., 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In: Proc. ASRU.
Fiscus, J.G., Ajot, J., Garofolo, J.S., 2007a. The rich transcription 2007 meeting recognition evaluation. In: Proc. Multimodal Tech. Percept. Hum.. pp. 373–389.
Fiscus, J.G., Ajot, J., Garofolo, J.S., 2007b. The Rich Transcription 2007 meeting recognition evaluation. In: International Evaluation Workshop on Rich Transcription. Springer, pp. 373–389.
Fiscus, J.G., Ajot, J., Radde, N., Laprun, C., et al., 2006a. Multiple dimension Levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech.. In: LREC. Citeseer, pp. 803–808.
Fiscus, J.G., Doddington, G., Le, A., Sanders, G., Przybocki, M., Pallett, D., 2007c. 2003 NIST rich transcription evaluation data LDC2007s10. Web Download. Phila.: Linguist. Data Consort..
Fiscus, J.G., Radde, N., Garofolo, J.S., Le, A., Ajot, J., Laprun, C., 2006b. The rich transcription 2005 spring meeting recognition evaluation. In: Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2. Springer, pp. 369–389.
Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X., 2021. FSD50k: An open dataset of human-labeled sound events. IEEE/ACM Trans. Audio, Speech, Lang. Process. 30.
Fox, C., Liu, Y., Zwyssig, E., Hain, T., 2013. The Sheffield wargames corpus. In: Interspeech.
Fu, Y., Cheng, L., Lv, S., Jv, Y., Kong, Y., Chen, Z., Hu, Y., Xie, L., Wu, J., Bu, H., et al., 2021. AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. In: Interspeech.
Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., Watanabe, S., 2019a. End-to-end neural speaker diarization with permutation-free objectives. In: Interspeech 2019. pp. 4300–4304.
Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., Watanabe, S., 2019b. End-to-end neural speaker diarization with self-attention. In: Proc. of ASRU. IEEE, pp. 296–303.
Gales, M., Young, S., et al., 2008. The application of hidden Markov models in speech recognition. Found. Trends® Signal Process. 1 (3), 195–304.
Garofolo, J.S., Fiscus, J.G., Laprun, C.D., 2004. The Rich Transcription 2004 Spring Meeting Recognition Evaluation. US Department of Commerce, National Institute of Standards and Technology.
Garofolo, J.S., Fiscus, J.G., Martin, A.F., Pallett, D.S., Przybocki, M.A., 2002. NIST rich transcription 2002 evaluation: A preview.. In: LREC.
Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al., 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.
Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development. In: Proc. of ICASSP.
Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al., 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR.
Graves, A., 2012. Sequence transduction with recurrent neural networks. In: ICML.
Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: ICML.
Gu, A., Goel, K., Ré, C., 2021. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396.
Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al., 2020. Conformer: Convolution-augmented transformer for speech recognition. In: Interspeech.
Haeb-Umbach, R., Watanabe, S., Nakatani, T., Bacchiani, M., Hoffmeister, B., Seltzer, M.L., Zen, H., Souden, M., 2019. Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. IEEE Signal Process. Mag. 36 (6), 111–124.
Han, J., Landini, F., Rohdin, J., Silnova, A., Diez, M., Burget, L., 2025. Leveraging self-supervised learning for speaker diarization. In: Proc. of ICASSP.
Han, K.J., Prieto, R., Ma, T., 2019. State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions. In: Proc. of ASRU. IEEE, pp. 54–61.
Harper, M., 2015. The automatic speech recognition in reverberant environments (aspire) challenge. In: IEEE ASRU.
He, M.-K., Du, J., Liu, Q.-F., Lee, C.-H., 2023. Ansd-ma-mse: Adaptive neural speaker diarization using memory-aware multi-speaker embedding. IEEE/ACM Trans. Audio, Speech, Lang. Process. 31, 1561–1573.
Heafield, K., 2011. KenLM: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 187–197.
Heo, B., Park, S., Han, D., Yun, S., 2025. Rotary position embedding for vision transformer. In: European Conference on Computer Vision. Springer, pp. 289–305.
Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., Esteve, Y., 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In: Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20. Springer, pp. 198–208.
Hirano, Y., Nguyen, M., Azuma, K., Saragih, J.M., Sakti, S., 2024. The NAIST system for the CHiME-8 NOTSOFAR-1 task. In: CHiME Workshop. pp. 59–63.
Hirsch, H.-G., Pearce, D., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop. ITRW.
Horiguchi, S., Takashima, Y., Garcia, P., Watanabe, S., Kawaguchi, Y., 2022. Multi-channel end-to-end neural diarization with distributed microphones. In: Proc. of ICASSP. pp. 7332–7336.
Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio, Speech, Lang. Process. 29.
Hu, Y., Chen, C., Yang, C.-H.H., Qin, C., Chen, P.-Y., Chng, E.S., Zhang, C., 2024. Self-taught recognizer: Toward unsupervised adaptation for speech foundation models. arXiv preprint arXiv:2405.14161.
Huang, K., Li, Y., Wang, Z., Wang, H., Rao, W., Sun, Z., Tang, Z., Huang, S., Wang, Y., Yu, T., et al., 2024. The NPU-tea system for the CHiME-8 NOTSOFAR-1 challenge. In: CHiME Workshop. pp. 45–48.

Ito, N., Araki, S., Nakatani, T., 2016. Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing. In: 2016 24th European Signal Processing Conference. EUSIPCO, IEEE, pp. 1153–1157.
Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al., 2003. The ICSI meeting corpus. In: Proc. of ICASSP.
Jeub, M., Schafer, M., Vary, P., 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In: 2009 16th International Conference on Digital Signal Processing. IEEE, pp. 1–5.
Kamo, N., Tawara, N., Ando, A., Kano, T., Sato, H., Ikeshita, R., Moriya, T., Horiguchi, S., Matsuura, K., Ogawa, A., et al., 2023. NTT multi-speaker ASR system for the DASR task of CHiME-7 challenge. In: CHiME Workshop.
Kamo, N., Tawara, N., Ando, A., Kano, T., Sato, H., Ikeshita, R., Moriya, T., Horiguchi, S., Matsuura, K., Ogawa, A., et al., 2024. NTT multi-speaker ASR system for the DASR task of CHiME-8 challenge. In: CHiME Workshop.
Kanda, N., Wu, J., Wang, X., Chen, Z., Li, J., Yoshioka, T., 2022a. VarArray meets t-SOT: Advancing the state of the art of streaming distant conversational speech recognition. ArXiv.
Kanda, N., Xiao, X., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2022b. Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR. In: Proc. of ICASSP. IEEE, pp. 8082–8086.
Karafiát, M., Veselỳ, K., Szöke, I., Mošner, L., Beneš, K., Witkowski, M., Barchi, G., Pepino, L., 2023. BUT CHiME-7 system description. In: CHiME Workshop.
Kim, S., Arora, A., Le, D., Yeh, C.-F., Fuegen, C., Kalinli, O., Seltzer, M.L., 2021. Semantic distance: A new metric for ASR performance analysis towards spoken language understanding. In: Interspeech.
Kim, J., Lee, Y., Kim, E., 2020. Accelerating RNN transducer inference via adaptive expansion search. IEEE Signal Process. Lett. 27, 2019–2023.
Kinoshita, K., Delcroix, M., Tawara, N., 2021a. Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech. In: Interspeech.
Kinoshita, K., Delcroix, M., Tawara, N., 2021b. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. In: Proc. of ICASSP.
Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Maas, R., et al., 2013. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In: Proc. of WASPAA. IEEE, pp. 1–4.
Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017. A study on data augmentation of reverberant speech for robust speech recognition. In: Proc. of ICASSP. IEEE, pp. 5220–5224.
Koluguri, N.R., Park, T., Ginsburg, B., 2022. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In: Proc. of ICASSP. IEEE, pp. 8102–8106.
Kuchaiev, O., et al., 2019. NeMo: a toolkit for building AI applications using neural modules. In: Proc. Systems for ML Worshop, NeurIPS.
Kudo, T., Richardson, J., 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71.
Kumar, S., Byrne, B., 2004. Minimum bayes-risk decoding for statistical machine translation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. pp. 169–176.
Lavechin, M., Métais, M., Titeux, H., Boissonnet, A., Copet, J., Rivière, M., Bergelson, E., Cristia, A., Dupoux, E., Bredin, H., 2023. Brouhaha: multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation. In: Proc. of ASRU. IEEE, pp. 1–7.
Lee, K.-F., Hon, H.-W., Reddy, R., 1990. An overview of the SPHINX speech recognition system. IEEE Trans. Acoust. Speech Signal Process. 38 (1), 35–45.
Lee, Y., Yun, T., Cai, J., Su, H., Song, H., 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3941–3960.
Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio, Speech, Lang. Process. 22 (4), 745–777.
Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81.
Lin, G.-T., Li, S.-W., Lee, H.-y., 2022. Listen, adapt, better WER: Source-free single-utterance test-time adaptation for automatic speech recognition. In: Interspeech.
Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C., 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634.
Luo, Y., Mesgarani, N., 2019. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. 27 (8), 1256–1266.
Masuyama, Y., Chang, X., Cornell, S., Watanabe, S., Ono, N., 2022. End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation.
Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., Timofeeva, T., Mitrofanov, A., Andrusenko, A., Podluzhny, I., et al., 2020a. The STC system for the CHiME-6 challenge. In: CHiME Workshop.
Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., Timofeeva, T., Mitrofanov, A., Podluzhny, I., Romanenko, A., et al., 2020b. Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario. In: Interspeech. pp. 274–278.
Merity, S., Keskar, N.S., Socher, R., 2018. Regularizing and optimizing LSTM language models. In: International Conference on Learning Representations.
Mitrofanov, A., Prisyach, T., Timofeeva, T., Novoselov, S., Korenevsky, M., Khokhlov, Y., Akulov, A., Anikin, A., Khalili, R., Lezhenin, I., et al., 2024. STCON system for the CHiME-8 challenge. In: CHiME Workshop.
Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S.M., Tyagi, A., Casas, J.R., Turmo, J., Cristoforetti, L., Tobia, F., et al., 2007. The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. Lang. Resour. Eval. 41, 389–407.
Mu, B., Guo, P., Wang, H., Li, Y., Li, Y., Zhou, P., Chen, W., Xie, L., 2023. The NPU system for DASR task of chime-7 challenge. In: CHiME Workshop. pp. 63–66.
von Neumann, T., Boeddeker, C., Delcroix, M., Haeb-Umbach, R., 2023. MeetEval: A toolkit for computation of word error rates for meeting transcription systems. In: CHiME Workshop.
Nguyen, T.A., Kharitonov, E., Copet, J., Adi, Y., Hsu, W.-N., Elkahky, A., et al., 2023. Generative Spoken Dialogue Language Modeling. Trans. Assoc. Comput. Linguist..
Niu, S., Wang, R., Du, J., Yang, G., Tu, Y., Wu, S., Qian, S., Wu, H., Xu, H., Zhang, X., et al., 2024. The ustc-nercslip systems for the chime-8 notsofar-1 challenge. In: CHiME Workshop.
Novitasari, S., Fukuda, T., Kurata, G., 2022. Improving ASR robustness in noisy condition through VAD integration.. In: Interspeech. pp. 3784–3788.
Ogawa, A., Kamo, N., Matsuura, K., Ashihara, T., Moriya, T., Kano, T., Tawara, N., Delcroix, M., 2024. Applying LLMs for rescoring n-best asr hypotheses of casual conversations: Effects of domain adaptation and context carry-over. In: CHiME Workshop.
OpenAI, 2024. GPT-4o System Card. URL: https://cdn.openai.com/gpt-4o-system-card.pdf.
Ozaki, Y., Tanigaki, Y., Watanabe, S., Onishi, M., 2020. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference. pp. 533–541.
Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. LibriSpeech: an ASR corpus based on public domain audio books. In: Proc. of ICASSP.
Park, T.J., Dhawan, K., Koluguri, N., Balam, J., 2023a. Enhancing speaker diarization with large language models: A contextual beam search approach. arXiv preprint arXiv:2309.05248.
Park, T.J., Han, K.J., Kumar, M., Narayanan, S., 2019. Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. IEEE Signal Process. Lett. 27, 381–385.

Park, T.J., Huang, H., Hooper, C., Koluguri, N., Dhawan, K., Jukic, A., Balam, J., Ginsburg, B., 2023b. Property-aware multi-speaker data simulation: A probabilistic modelling technique for synthetic data generation. In: Interspeech.
Park, T.J., Huang, H., Jukic, A., Dhawan, K., Puvvada, K.C., Koluguri, N., Karpov, N., Laptev, A., Balam, J., Ginsburg, B., 2023c. The CHiME-7 Challenge: System description and performance of nemo team's DASR system. In: CHiME Workshop.
Park, T.J., Koluguri, N.R., Balam, J., Ginsburg, B., 2022. Multi-scale speaker diarization with dynamic scale weighting. In: Interspeech.
Park, T., Medennikov, I., Dhawan, K., Wang, W., Huang, H., Koluguri, N.R., Puvvada, K.C., Balam, J., Ginsburg, B., 2024. Sortformer: Seamless integration of speaker diarization and asr by bridging timestamps and tokens. IEEE/ACM Trans. Audio, Speech Language Process. (in preparation).
Park, D.S., Zhang, Y., Chiu, C.-C., Chen, Y., Li, B., Chan, W., Le, Q.V., Wu, Y., 2020. Specaugment on large scale datasets. In: Proc. of ICASSP. IEEE, pp. 6879–6883.
Paul, D.B., Baker, J., 1992. The design for the wall street journal-based CSR corpus. In: Speech and Natural Language Workshop.
Peng, Y., Dalmia, S., Lane, I., Watanabe, S., 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In: International Conference on Machine Learning. PMLR, pp. 17627–17643.
Peng, Y., Tian, J., Yan, B., Berrebbi, D., Chang, X., Li, X., Shi, J., Arora, S., Chen, W., Sharma, R., et al., 2023. Reproducing Whisper-style training using an open-source toolkit and publicly available data. In: Proc. of ASRU. IEEE, pp. 1–8.
Peng, J., et al., 2024. A survey on speech large language models. arXiv preprint arXiv:2410.18908.
Polok, A., Klement, D., Han, J., Sedlacek, S., Yusuf, B., Maciejewski, M., Wiesner, M., Burget, L., 2024a. BUT/JHU system description for chime-8 NOTSOFAR-1 challenge. In: CHiME Workshop.
Polok, A., Klement, D., Wiesner, M., Khudanpur, S., Černockỳ, J., Burget, L., 2024b. Target speaker ASR with whisper. In: ICASSP 2025. (in preparation).
Povey, D., 2020. k2. URL: https://github.com/k2-fsa/k2 (Accessed 19 December 2024) https://github.com/k2-fsa/k2.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The kaldi speech recognition toolkit. In: Proc. of ASRU.
Prabhavalkar, R., Hori, T., Sainath, T.N., Schlüter, R., Watanabe, S., 2023. End-to-end speech recognition: A survey. arXiv preprint arXiv:2303.03329.
Prisyach, T., Khokhlov, Y., Korenevsky, M., Mitrofanov, A., Timofeeva, T., Odegov, I., Nasretdinov, R., Lezhenin, I., Miroshnichenko, D., Karelin, A., et al., 2023. STCON system for the CHiME-7 challenge. In: CHiME Workshop.
Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.
Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust speech recognition via large-scale weak supervision. ArXiv.
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (140), 1–67.
Raj, D., Denisov, P., Chen, Z., Erdogan, H., Huang, Z., He, M., Watanabe, S., Du, J., Yoshioka, T., Luo, Y., et al., 2021a. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. In: IEEE SLT.
Raj, D., Garcia-Perera, L.P., Huang, Z., Watanabe, S., Povey, D., Stolcke, A., Khudanpur, S., 2021b. DOVER-Lap: A method for combining overlap-aware diarization outputs. In: Proc. of SLT. IEEE, pp. 881–888.
Raj, D., Povey, D., Khudanpur, S., 2023. GPU-accelerated guided source separation for meeting transcription. In: Interspeech.
Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. 10 (1–3), 19–41.
Richey, C., Barrios, M.A., Armstrong, Z., Bartels, C., Franco, H., Graciarena, M., Lawson, A., Nandwana, M.K., Stauffer, A., van Hout, J., et al., 2018. Voices obscured in complex environmental settings (VOiCES) corpus. In: Interspeech.
Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, M.L., 2019. The second DIHARD diarization challenge: Dataset, task, and baselines. In: Interspeech.
Snyder, D., Chen, G., Povey, D., 2015. MUSAN: A music, speech, and noise corpus. ArXiv.
Stupakov, A., Hanusa, E., Vijaywargi, D., Fox, D., Bilmes, J., 2012. The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments. Comput. Speech & Lang. 26 (1), 52–66.
Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J., Carmiel, Y., 2020. WER we are and WER we think we are. In: Findings of EMNLP.
Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
Van Segbroeck, M., Zaid, A., Kutsenko, K., Huerta, C., Nguyen, T., Luo, X., Hoffmeister, B., Trmal, J., Omologo, M., Maas, R., 2019. DiPCo – dinner party corpus. ArXiv.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: NIPS.
Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M., 2013. The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes. In: IEEE ASRU.
Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio, Speech, Lang. Process. 14 (4), 1462–1469.
Vincent, E., Watanabe, S., Barker, J., Marxer, R., 2016. The 4th chime speech separation and recognition challenge. In: CHiME Workshop.
Vinnikov, A., Ivry, A., Hurvitz, A., Abramovski, I., Koubi, S., Gurvich, I., Peer, S., Xiao, X., Elizalde, B.M., Kanda, N., et al., 2024. NOTSOFAR-1 challenge: New datasets, baseline, and tasks for distant meeting transcription. In: Interspeech.
Wang, R., He, M., Du, J., Zhou, H., Niu, S., Chen, H., Yue, Y., Yang, G., Wu, S., Sun, L., et al., 2023a. The USTC-NERCSLIP systems for the CHiME-7 DASR challenge. In: CHiME Workshop.
Wang, Q., Huang, Y., Zhao, G., Clark, E., Xia, W., Liao, H., 2024b. DiarizationLM: Speaker diarization post-processing with large language models. In: Interspeech.
Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., Dupoux, E., 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 993–1003.
Wang, Z., Wu, S., Chen, H., He, M.-K., Du, J., Lee, C.-H., Chen, J., Watanabe, S., Siniscalchi, S., Scharenborg, O., Liu, D., Yin, B., Pan, J., Gao, J., Liu, C., 2023b. The multimodal information based speech processing (misp) 2022 challenge: Audio-visual diarization and recognition. In: Proc. of ICASSP. pp. 1–5.
Wang, D., Xiao, X., Kanda, N., Yousefi, M., Yoshioka, T., Wu, J., 2024a. Profile-error-tolerant target-speaker voice activity detection. In: Proc. of ICASSP. IEEE, pp. 11906–11910.
Warsitz, E., Haeb-Umbach, R., 2007. Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. Audio, Speech, Lang. Process. 15 (5), 1529–1539.
Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., Ochiai, T., 2018. ESPnet: End-to-end speech processing toolkit. In: Interspeech.
Watanabe, S., Hori, T., Kim, S., Hershey, J.R., Hayashi, T., 2017. Hybrid CTC/Attention architecture for end-to-end speech recognition. IEEE J. Sel. Top. Signal Process. 11 (8), 1240–1253.
Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., et al., 2020. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In: CHiME Workshop.
Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R.J., Wilson, K., Hershey, J.R., 2020. Unsupervised sound separation using mixture invariant training. arXiv preprint arXiv:2006.12701.
Wolf, M., Nadeu, C., 2014. Channel selection measures for multi-microphone speech recognition. Speech Commun. 57.

Yang, S.-w., Chang, H.-J., Huang, Z., Liu, A.T., Lai, C.-I., Wu, H., Shi, J., Chang, X., Tsai, H.-S., Huang, W.-C., et al., 2024. A large-scale evaluation of speech foundation models. IEEE/ACM Trans. Audio, Speech, Lang. Process..

wen Yang, S., Chi, P.-H., Chuang, Y.-S., Lai, C.-I.J., Lakhotia, K., Lin, Y.Y., Liu, A.T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., tik Lee, K., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., yi Lee, H., 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In: Interspeech 2021. pp. 1194–1198. http://dx.doi.org/10.21437/Interspeech.2021-1775.

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I.J., Lakhotia, K., Lin, Y.Y., Liu, A.T., Shi, J., Chang, X., Lin, G.-T., et al., 2021. SUPERB: Speech processing universal performance benchmark. In: Interspeech.

Yao, Z., Guo, L., Yang, X., Kang, W., Kuang, F., Yang, Y., Jin, Z., Lin, L., Povey, D., 2023. Zipformer: A faster and better encoder for automatic speech recognition. In: International Conference on Learning Representations.

Ye, L., Lu, H., Cheng, G., Chen, Y., Shang, Z., Li, X., 2023. The IACAS-thinkit system for chime-7 challenge. In: CHiME Workshop.

Yoshioka, T., Abramovski, I., Aksoylar, C., Chen, Z., David, M., Dimitriadis, D., Gong, Y., Gurvich, I., Huang, X., Huang, Y., et al., 2019. Advances in online audio-visual meeting transcription. In: Proc. of ASRU. IEEE, pp. 276–283.

Yoshioka, T., Nakatani, T., 2012. Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. IEEE Trans. Audio, Speech, Lang. Process. 20 (10), 2707–2720.

You, Z., Feng, S., Su, D., Yu, D., 2021. SpeechMoE: Scaling to large acoustic models with dynamic routing mixture of experts. In: Interspeech.

Yu, F., Zhang, S., Fu, Y., Xie, L., Zheng, S., Du, Z., Huang, W., Guo, P., Yan, Z., Ma, B., et al., 2022. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In: Proc. of ICASSP.

Żelasko, P., Povey, D., Trmal, J., Khudanpur, S., et al., 2021. Lhotse: a speech data representation library for the modern deep learning ecosystem. arXiv preprint arXiv:2110.12561.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., Zhao, T., 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In: International Conference on Learning Representations.

Zhang, L., Negrinho, R., Ghosh, A., Jagannathan, V., Hassanzadeh, H.R., Schaaf, T., Gormley, M.R., 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 3693–3712.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., Han, J., 2022. Towards a unified multi-dimensional evaluator for text generation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 2023–2038.

Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Hassan, A., Celikyilmaz, A., Liu, Y., Qiu, X., et al., 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5905–5921.

Žmolíková, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., Černockỳ, J., 2019. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. IEEE J. Sel. Top. Signal Process. 13 (4), 800–814.

Zmolikova, K., Merello, S., Kalgaonkar, K., Lin, J., Moritz, N., Ma, P., Sun, M., Chen, H., Saliou, A., Petridis, S., et al., 2024. The CHiME-8 MMCSG Challenge: Multi-modal conversations in smart glasses. In: CHiME Workshop.