CSPB: Conversational Speech Processing Benchmark for Self-supervised Speech Models

Zili Huang[1], Matthew Maciejewski[2], Leibny Paola Garcia[1,2], Shinji Watanabe[3], Sanjeev Khudanpur[1]

[1] Center for Language and Speech Processing, Johns Hopkins University,
[2] Human Language Technology Center of Excellence, Johns Hopkins University,
[3] Language Technologies Institute, Carnegie Mellon University,

Correspondence: hzili1@jhu.edu, khudanpur@jhu.edu

Abstract

Recent advances in self-supervised learning (SSL) have led to powerful speech representation models, yet their robustness in real-world conversational settings remains largely untested. Most existing benchmarks focus on clean, single-speaker, single-channel audio, failing to reflect the complexities of natural human interaction—where background noise, reverberation, and overlapping speech are the norm. To bridge these critical gaps, we present the Conversational Speech Processing Benchmark (CSPB), a new benchmark designed to assess the robustness of SSL speech models in realistic conversational scenarios. CSPB is constructed from four multi-party datasets—AMI, AliMeeting, MMCSG, and DiPCo—and supports both single-channel and multi-channel evaluation. By releasing CSPB as an open-source toolkit, we aim to establish a unified framework for evaluating and advancing robust, spatially-aware self-supervised speech models.

1 Introduction

Self-supervised learning (SSL) is a representation learning approach where the model learns from automatically generated labels derived from the data itself, without relying on manual annotation. It has achieved significant success across domains such as computer vision (Chen et al., 2020; He et al., 2022), natural language processing (Devlin et al., 2019; Peters et al., 2018), and speech processing (Baevski et al., 2020; Hsu et al., 2021a). The typical SSL pipeline consists of two stages. In the first stage, the model learns general-purpose, task-agnostic representations by solving pretext tasks such as masked prediction or contrastive learning. In the second stage, these representations are adapted to specific downstream tasks via supervised finetuning.

In the speech domain, SSL has become a foundational technique due to its versatility. SSL models serve as powerful backbones for finetuning on downstream tasks like ASR (Baevski et al., 2020), speaker identification (Fan et al., 2021), and speech separation (Chen et al., 2023). Moreover, their ability to encode rich phonetic and speaker information makes them ideal for use as speech tokens in large language models (LLMs) (Zhang et al., 2023; Borsos et al., 2023; Tang et al., 2024).

Due to the wide range of downstream tasks and evaluation protocols used in prior work, comparing SSL speech models in a fair and reproducible manner remains challenging. Standardized benchmarks are therefore essential for systematic evaluation. The SUPERB benchmark (Yang et al., 2021) was a landmark effort in this area, establishing a unified framework for testing SSL speech models on tasks ranging from content and semantics to speaker and paralinguistic traits. Its success spurred the development of benchmarks for speech in other languages (Evain et al., 2021; Shi et al., 2023), as well as for related modalities such as general audio (Turian et al., 2022) and music (Yuan et al., 2023).

However, a critical domain remains underexplored: **real-world conversational speech**. This domain introduces significant challenges rarely found in existing benchmarks, such as background noise, reverberation, and speaker overlap. While SUPERB includes relevant tasks like diarization and separation, its evaluation is often based on synthetic or controlled conditions. Consequently, it remains an open question how well current SSL models generalize to the acoustic and interactional complexities of natural, everyday conversations.

Moreover, most existing benchmarks are limited to single-channel speech, whereas real-world applications often involve multi-channel audio captured using microphone arrays. Despite a few recent efforts exploring this direction (Dimitriadis et al., 2023; Huang et al., 2024c; Yang et al., 2025), SSL from multi-channel input remains significantly underexplored. Progress in this area has been slow

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5878–5893

partially due to the absence of a standardized benchmark for systematic evaluation, and the lack of open-source multi-channel SSL models.

To bridge this gap, we introduce the **C**onversational **S**peech **P**rocessing **B**enchmark (CSPB), a benchmark designed to systematically evaluate the robustness of self-supervised speech models in realistic conversational settings. CSPB is constructed from four challenging multi-party conversation datasets—AMI (Kraaij et al., 2005), AliMeeting (Yu et al., 2022), MMCSG (Žmolíková et al., 2024) and DiPCo (Segbroeck et al., 2020)—and it supports both single-channel and multi-channel evaluation tracks across tasks including ASR, speaker diarization, and speech enhancement/separation. We hope our benchmark serves as a strong foundation for advancing future research on effective self-supervised speech models tailored to conversational speech. Our code is available at: https://github.com/HuangZiliAndy/CSPB.

2 Related Works

2.1 Robustness of Self-Supervised Speech Representations

Many self-supervised speech models, originally pretrained with clean speech, struggle to adapt to real-world scenarios that include noise, reverberation, and overlapping speech. Efforts to improve the robustness of self-supervised speech representations fall into three main categories: (1) *Data Augmentation*: These methods construct pairs of clean and noisy speech by augmenting clean recordings with noise and reverberation. The goal is to minimize differences in the embedding space using techniques like contrastive learning (Wang et al., 2022; Zhu et al., 2023) or domain adversary training (Huang et al., 2022). (2) *Domain Augmentation*: The diversity of the pretraining domain significantly influences the performance of SSL speech models. For instance, (Hsu et al., 2021b) shows that pretraining across multiple domains enhances generalization to new, unseen domains. Similarly, the robustness of WavLM (Chen et al., 2022) significantly improves after pretraining on more diverse datasets. (3) *Integration with Frontend Processing*: In the realm of speech processing, frontend techniques are frequently utilized for denoising and dereverberation. Several studies have explored the integration of these frontend methods with SSL backends to enhance model robustness. For example, the IRIS framework (Chang et al., 2022) combines a speech enhancement frontend with a WavLM-based end-to-end ASR backend to deliver robust single-channel ASR performance. Its multi-channel variant (Masuyama et al., 2023) further develops this concept by incorporating a WPD Beamformer, thus expanding the application to multi-channel scenarios.

2.2 Benchmarks for Evaluating Self-Supervised Speech Models

The rapid advancement of self-supervised speech models necessitates the development of a reliable and standardized evaluation protocol to facilitate consistent performance comparisons. SUPERB (Yang et al., 2021) was among the earliest efforts to address this need, providing a unified benchmark that evaluates SSL models across 10 downstream tasks covering four key aspects of speech: content, speaker characteristics, semantics, and paralinguistics. In SUPERB, the SSL encoder is kept frozen, and its extracted representations are evaluated by training lightweight downstream models under a standardized protocol. Building on this framework, the SUPERB-SG (Tsai et al., 2022) extends the evaluation to include more tasks related to semantics and generation. While both SUPERB and SUPERB-SG primarily target English datasets, similar benchmarking initiatives have emerged for other languages, such as LeBenchmark for French (Evain et al., 2021; Parcollet et al., 2024), IndicSUPERB for Indian languages (Javed et al., 2023), and multilingual efforts like ML-SUPERB (Shi et al., 2023).

Meanwhile, benchmarking work has been extended beyond the scope of self-supervised speech models. For example, HEAR (Turian et al., 2022) and MARBLE (Yuan et al., 2023) extend the scope to audio and music respectively. The SLUE series (Shon et al., 2022, 2023) integrates self-supervised speech models with text models to tackle a variety of spoken language understanding tasks. The Dynamic-SUPERB series (Huang et al., 2024b,a), MMAU (Sakshi et al., 2024), and AIR-Bench (Yang et al., 2024) evaluate the zero-shot and few-shot capabilities of Large Audio-Language Models through diverse instruction-following tasks.

Despite these advances, benchmarking efforts specifically tailored to conversational speech remain limited. The benchmark work that is most relevant to our study is TS-SUPERB (Peng et al.,

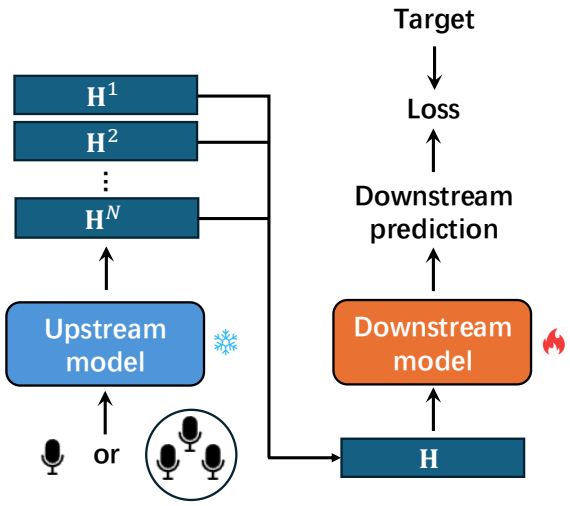


Figure 1: Evaluation protocol for the CSPB benchmark. Multi-level representations are extracted from the frozen upstream model and combined via a learnable weighted sum to form $\mathbf{H}$. The combined representation is fed into a lightweight, task-specific downstream model and optimized using the corresponding task-specific loss.

2025), however, it is based on synthetic datasets and focus on target-speaker speech processing.

3 Benchmark Design and Evaluation Protocol

We adopt the evaluation protocol introduced in the SUPERB (Yang et al., 2021) benchmark. As illustrated in Figure 1, given a single-channel input waveform $s \in \mathbb{R}^T$, we extract hidden representations $\mathbf{H}^{(i)} \in \mathbb{R}^{L \times D}$ from each layer $i \in \{1, \ldots, N\}$ of a pretrained self-supervised speech model, where L denotes the number of frames and D the feature dimension. These layer-wise representations are aggregated via a learnable weighted sum:

$$\mathbf{H} = \sum_{i=1}^{N} \alpha_i \mathbf{H}^{(i)} \quad (1)$$

where α_i are layer-specific scalar weights, normalized with a softmax: $\sum_i \alpha_i = 1$. During supervised finetuning, only the downstream head and the aggregation weights α_i are updated, while the pretrained SSL model remains frozen.

We extend the evaluation protocol to multi-channel input waveforms $s \in \mathbb{R}^{C \times T}$, where C denotes the number of microphone channels. The protocol assumes that the model extracts layer-wise hidden representations $\mathbf{H}^{(i)} \in \mathbb{R}^{L \times D}$ that are independent of C, i.e., the channel dimension is eliminated prior to aggregation. Under this assumption, multi-channel inputs are evaluated using the same aggregation mechanism and downstream heads as in the single-channel case.

To achieve this, we investigate two distinct approaches for processing multi-channel speech:

Hybrid Approach. This strategy combines a standard single-channel SSL model with a beamforming front end. First, beamforming exploits spatial information across microphone channels to suppress interference and aggregate the multi-channel waveform into a single enhanced signal. The single-channel SSL model then processes this enhanced signal to extract layer-wise hidden representations.

Multi-Channel SSL Model. In contrast, this approach utilizes a multi-channel SSL model that natively processes the raw multi-channel waveform. The model directly extracts hidden representations $\mathbf{H}^{(i)} \in \mathbb{R}^{L \times D}$ from the multi-channel input, bypassing the need for a separate beamforming step.

4 Datasets and Tasks

4.1 Datasets

We construct our benchmark using four publicly available conversational speech corpora: the Augmented Multi-party Interaction (AMI) corpus (Kraaij et al., 2005), the AliMeeting corpus (Yu et al., 2022), the Multi-modal Conversations in Smart Glasses (MMCSG) corpus (Žmolíková et al., 2024), and the Dinner Party corpus (Segbroeck et al., 2020) (DiPCo). These datasets span a diverse range of conversational scenarios, recording environments, and microphone configurations, providing a comprehensive foundation for evaluating model robustness. A statistical summary is available in Table 1.

AMI and AliMeeting are multi-party **meeting** corpora, where each session involves 2 to 5 participants engaged in spontaneous or scenario-driven discussions. Each session was recorded simultaneously with individual headset microphones and far-field circular arrays of 8 microphones, allowing comparative analysis between close-talk and distant multi-channel recordings.

DiPCo also captures multi-party conversational speech with both headset and far-field microphones. However, it is set in a more naturalistic, home-like environment. Sessions include background

Dataset	Scenarios	Hours	# Sessions	# Speakers	Overlap (%)	Lang
AMI	Meeting	98	168	3–5	13.6	en
AliMeeting	Meeting	126	237	2–4	27.8	zh
MMCSG	Conversation	26	530	2	11.7	en
DiPCo	Dinner Party	5	10	4	27.9	en

Table 1: Statistics of the conversational speech corpora used in CSPB.

activities such as food gathering and music playback, introducing additional acoustic challenges. The 5 far-field microphone arrays, each with 7 microphones, are strategically placed throughout the room to simulate various **smart home devices**.

In contrast, MMCSG offers a unique perspective focused on **wearable devices**. It captures naturalistic conversations between two speakers, with one speaker wearing Aria smart glasses equipped with embedded microphone arrays. Unlike the other three datasets, the microphone array maintains a nearly fixed distance from the glasses wearer (SELF) while being significantly farther from the conversation partner (OTHER), resulting in a distinct spatial and acoustic configuration.

4.2 Tasks

4.2.1 Automatic Speech Recognition (ASR)

The ASR task involves transcribing speech into text and serves as a direct measure of an SSL model's ability to **encode fine-grained phonetic information**. Our ASR setup follows the same configuration as SUPERB (Yang et al., 2021), although the evaluation data used in CSPB is considerably more challenging. SSL representations are passed through a lightweight downstream model followed by a linear projection layer to predict character-level posteriors, and the model is trained using the CTC loss. We report word error rate (WER) for the English datasets AMI, MMCSG and DiPCo, and character error rate (CER) for the Mandarin dataset AliMeeting.

4.2.2 Speaker Diarization (SD)

The speaker diarization task addresses the "who speaks when" problem, and evaluates the SSL model's ability to **encode speaker-discriminative information and speaker-change cues over time**. Given an unsegmented audio as input, the diarization task predicts speaker turn boundaries and groups segments belonging to the same speaker into clusters. In contrast to the speaker diarization setup in Yang et al. (2021), which evaluates diarization performance on 2-speaker synthetic mixtures, we perform long-form speaker diarization on the entire meeting using the pyannote (Kinoshita et al., 2021; Bredin, 2023) pipeline. The pipeline consists of three stages. (1) Local neural speaker segmentation (2) Local speaker embedding extraction (3) Global agglomerative clustering. Following prior work (Han et al., 2025), we use SSL-derived representations to train an EEND-based (Fujita et al., 2019) segmentation model in the first stage. To remove confounding effects from speaker embedding quality and clustering, stages (2) and (3) are replaced with reference-based optimal speaker assignment, thereby isolating the contribution of SSL representations to local speaker segmentation. Diarization error rate (DER) is adopted as the evaluation metric.

4.2.3 Speech Enhancement and Separation (SE and SS)

Unlike audio recorded in controlled, single-speaker environments, conversational speech is often corrupted by background noise, reverberation, and overlapping speakers. To assess how well SSL models handle these realistic conditions, we define a single task that jointly addresses denoising, dereverberation, and speech separation. This task evaluates an SSL model's ability to preserve information relevant for **disentangling background noise and overlapping speech sources.**

Since real-world datasets typically lack clean reference signals, we simulate a partially overlapping multi-channel dataset using Pyroomacoustics (Scheibler et al., 2018) to construct a supervised finetuning set. The simulation procedure is summarized in Algorithm 1. During supervised finetuning, the model learns to reconstruct individual clean speech signals from multi-talker reverberant mixtures.

Our model architecture is identical to Hung et al. (2022), we first extract SSL representations from the input mixture. These representations are then concatenated with the STFT magnitude and passed through the downstream model to predict time-frequency masks for each source.

Let $S_i(t, f)$ and $X(t, f)$ denote the STFTs for clean source i and the mixture, and let $\hat{M}_i(t, f)$

be the predicted mask for source i. The estimated STFT magnitude for source i is computed as:

$$|\hat{S}_i(t,f)| = \hat{M}_i(t,f) \cdot |X(t,f)|$$

where $|\cdot|$ denotes the magnitude of the STFT.

We then compute the permutation invariant L1 loss between the predicted and reference magnitudes:

$$\mathcal{L}_{\text{PIT}} = \min_{\pi \in \mathcal{P}} \frac{1}{N} \sum_{i=1}^{N} \left\| |\hat{S}_i| - |S_{\pi(i)}| \right\|_1$$

where $\mathcal{P}$ is the set of all permutations over N sources. Our experiments are conducted under a two-speaker setup, i.e., $N = 2$.

At evaluation time, we adopt the **utterance group-based evaluation protocol** proposed by Kanda et al. (2021). An utterance group is defined as a set of utterances connected by overlapping speaker regions. We focus on utterance groups involving no more than two speakers and apply our mask-based speech enhancement model to perform inference on these segments.

Since ground-truth clean reference signals are not available in real conversational scenarios, we evaluate the enhanced speech using a fixed joint CTC/attention-based end-to-end ASR model (Kim et al., 2017), which is pretrained only on close-talk microphone recordings. WER or CER is reported as the evaluation metric.

5 Experimental Setup

5.1 Upstream Models and Baselines

We evaluate a diverse set of upstream models, which are organized into three categories. A detailed summary of all models is provided in Table 10.

Traditional Baseline As a baseline, we use 80-dimensional log Mel-filterbank (FBANK) features, one of the most common acoustic features in speech processing. They are extracted using a 25ms frame length and a 10ms frame shift.

Self-Supervised Models We evaluate 9 self-supervised speech models in our benchmark. These include 8 prominent single-channel models from the **wav2vec 2.0**, **HuBERT**, and **WavLM** families, and one multi-channel model, **UniX-Enc**. Our UniX-Enc implementation is largely based on the original work of Huang et al. (2024c), with a few key modifications detailed in Appendix D.

Task	Architecture	Hidden Size	# params
ASR	2-layer BLSTM	512	11.6M
DIAR	2-layer BLSTM	512	11.6M
SE / SS	3-layer BLSTM	256	5.3M

Table 2: Downstream model architectures and parameter sizes for each task.

Supervised Models Finally, to provide a powerful point of comparison, we include models from the supervised **Whisper** family. Although Whisper is trained in a fully supervised manner and supports a limited set of downstream tasks, it is widely adopted as a speech encoder in prior work (Chu et al., 2023; Tang et al., 2024), making it a relevant reference for our evaluation. Since Whisper models use an encoder-decoder architecture while our evaluation protocol is designed for encoder-only models, we utilize only the encoder component during evaluation.

5.2 Implementation Details

The downstream model architecture for the three tasks is summarized in Table 2. To ensure that benchmark results reflect the quality of the underlying SSL representations, we employ lightweight downstream models across all tasks. Specifically, we use BLSTMs with task-specific configurations in terms of hidden size and number of layers. Although we also experimented with shallow Transformers, they yielded inferior performance, likely due to their inherent data-hungry nature.

We finetune all models using a fixed learning rate, selected by grid search over $\{1e^{-3}, 1e^{-4}, 1e^{-5}\}$. We use the AdamW optimizer for ASR and SE/SS tasks, and Adam for the SD task. A batch size of 32 is used consistently across all tasks. Finetuning is performed on a single GPU, and the best checkpoint is selected based on development set performance.

For the DiPCo dataset, however, the limited amount of labeled data makes it difficult to achieve reasonable performance using the above configuration. We therefore adopt a domain-adaptive finetuning strategy: rather than initializing from a vanilla SSL model and a randomized downstream head, we utilize a model pretrained on the AMI dataset as a starting point. We then conduct further finetuning on DiPCo using a reduced learning rate selected from $\{3e^{-5}, 1e^{-5}, 3e^{-6}\}$. Since DiPCo does not provide a training split, we finetune on the development set and report evaluation results using the final checkpoint.

6 Experimental Results

6.1 Main result

Our benchmark supports both single-channel and multi-channel evaluation, with results summarized in Table 3 and Table 4, respectively. Since evaluation metrics vary in scale across tasks, we apply min-max normalization to enable fair comparison. For metrics where lower values indicate better performance (e.g., WER, DER), the normalized score is defined as:

$$F_{norm} = 100.0 * \frac{F_{worst} - F}{F_{worst} - F_{best}}$$

where F is the raw metric score, and F_{best} and F_{worst} are the best and worst system performances, computed over all channel configurations (1, 2, 4, and ALL channels). After normalization, a higher F_{norm} score indicates better performance.

For example, in AMI ASR, the best-performing system—ALL-channel BeamformIt + WavLM Large—achieves a WER of 32.7%, while the worst-performing system—the single-channel FBANK baseline—yields a WER of 89.3%; thus, $F_{best} = 32.7$, $F_{worst} = 89.3$. We compute the normalized score F_{norm} for each task and dataset, and report the average across all datasets in Tables 3 and 4.

The results in Table 3 reveal a clear performance hierarchy among the different upstream models. All SSL models dramatically outperform the traditional FBANK baseline, and a distinct trend emerges based on their pretraining data. Models trained on the clean speech corpora (e.g., wav2vec 2.0, HuBERT) without any data augmentation clearly under-perform their counterpart such as HuBERT Base Robust MGR and WavLM Base.[1] This trend culminates with WavLM Large, which was pretrained on massive and diverse datasets with heavy augmentation. It stands as the top-performing SSL model, achieving the best scores in ASR (90.3), SE/SS (85.2), and the highest Overall Score (84.0). Compared to similarly sized SSL models such as WavLM Base+, the single-channel UniX-Enc delivers competitive performance in SD and SE/SS but falls behind in ASR.

The Whisper family demonstrates exceptional performance, challenging the best SSL models. Notably, Whisper Small (Overall: 75.5) significantly outperforms all similarly sized SSL models. Whisper Medium achieves the second-best overall score and significantly stronger speaker diarization performance. This overall strong performance likely stems from the model's exposure to massive amounts of in-the-wild speech data sourced from the web, which offers significantly greater acoustic and linguistic diversity than the curated datasets typically used for SSL pretraining.

Multi-channel results are presented in Table 4. For each downstream task, we evaluate performance with 2, 4, and all channels, with channel selection details summarized in Table 11. As described in Section 3, for single-channel SSL models (HuBERT, WavLM, and the Whisper family), we equip them with a BeamformIt frontend, whereas for the multi-channel SSL model (UniX-Enc), hidden representations are extracted directly from the multi-channel inputs.

As shown in Table 4, increasing the number of microphone channels generally leads to improved performance, with the trend being most consistent for ASR. The relative ranking of upstream models remains largely stable across channel conditions. BeamformIt + WavLM Large achieves the best overall performance across 2-, 4-, and all-channel settings, while the Whisper family delivers the strongest speaker diarization results.

For UniX-Enc, we report two variants: *Channel Positional Encoding–Frozen (CPE-F)* and *Trainable (CPE-T)*. During pretraining, UniX-Enc is exposed to multi-channel data from diverse microphone topologies, where channel positional encodings (CPE) are ill-defined due to arbitrary microphone orderings. During supervised finetuning, however, the microphone topology is fixed and consistent across all samples, making CPE semantically meaningful. This setting enables us to assess the benefit of adapting the CPE in UniX-Enc: we either keep it frozen (CPE-F) or allow it to be updated (CPE-T). Our results show that CPE-T consistently outperforms CPE-F.

Overall, UniX-Enc slightly underperforms the BeamformIt + WavLM Base+ system, with the primary performance gap arising from ASR. This disparity likely stems from the limited availability of multi-channel pretraining data; despite aggregating all available open-source multi-channel corpora, the total pretraining data amounts to approximately 0.7k hours, which is significantly less than the 94k hours used to pretrain WavLM Base+.

[1] HuBERT Base Robust MGR is continually trained from HuBERT Base using LibriSpeech augmented with MUSAN (Snyder et al., 2015) noise, Gaussian noise, and reverberation. WavLM Base incorporates additive noise and interference during pretraining, jointly performing masked speech prediction and speech denoising tasks.

Upstream	ASR	SD	SE/SS	Overall Score
FBANK	0.0	6.0	0.0	2.0
wav2vec 2.0 Base	31.1	46.1	35.9	37.7
wav2vec 2.0 Large	35.0	51.1	33.0	39.7
HuBERT Base	37.0	45.1	38.3	40.2
HuBERT Base Robust MGR	39.1	50.1	43.8	44.3
HuBERT Large	52.1	49.3	50.9	50.8
WavLM Base	42.5	55.8	51.0	49.8
WavLM Base+	61.0	65.0	69.2	65.1
WavLM Large	**90.3**	76.5	**85.2**	**84.0**
UniX-Enc	50.1	69.8	67.7	62.6
Whisper Base	60.6	68.4	51.2	59.6
Whisper Small	77.9	85.2	69.4	75.5
Whisper Medium	85.7	**96.0**	72.5	82.0

Table 3: Min–max normalized scores for ASR, SD, and SE/SS (**single-channel track**). Higher values indicate better performance.

Front-End	Upstream	ASR			SD			SE/SS			Overall Score		
		2ch	4ch	All	2ch	4ch	All	2ch	4ch	All	2ch	4ch	All
BeamformIt	FBANK	3.0	4.2	5.3	7.9	1.8	8.2	31.0	33.2	39.3	14.0	13.1	17.6
BeamformIt	HuBERT Base	50.0	54.8	57.8	47.5	48.6	48.9	57.0	64.5	64.5	51.5	55.9	57.1
BeamformIt	HuBERT Large	65.0	68.6	70.7	52.7	55.8	54.6	69.9	70.4	70.3	62.5	64.9	65.2
BeamformIt	WavLM Base	55.5	60.8	62.9	58.5	56.7	60.6	69.7	76.0	79.3	61.2	64.5	67.6
BeamformIt	WavLM Base+	76.5	78.6	81.3	67.2	69.6	64.5	79.2	81.5	85.9	74.3	76.6	77.2
BeamformIt	WavLM Large	**97.7**	**99.3**	**100.0**	77.5	78.2	75.8	**96.2**	**98.3**	**100.0**	**90.5**	**91.9**	**91.9**
UniX-Enc (CPE-F)		57.2	59.0	59.5	72.3	72.8	75.4	76.6	79.0	83.0	68.7	70.3	72.6
UniX-Enc (CPE-T)		60.8	62.2	63.4	75.4	76.9	80.2	79.0	84.1	81.2	71.7	74.4	74.9
BeamformIt	Whisper Small	76.0	77.9	79.7	90.1	85.2	86.3	80.5	77.6	87.3	82.2	80.2	84.4
BeamformIt	Whisper Medium	80.6	82.9	84.6	**97.9**	**96.8**	**96.9**	80.3	79.7	87.3	86.2	86.4	89.6

Table 4: Min-max normalized scores for ASR, SD, and SE/SS (**multi-channel track**) across 2-, 4-, and all-channel settings. Higher values indicate better performance.

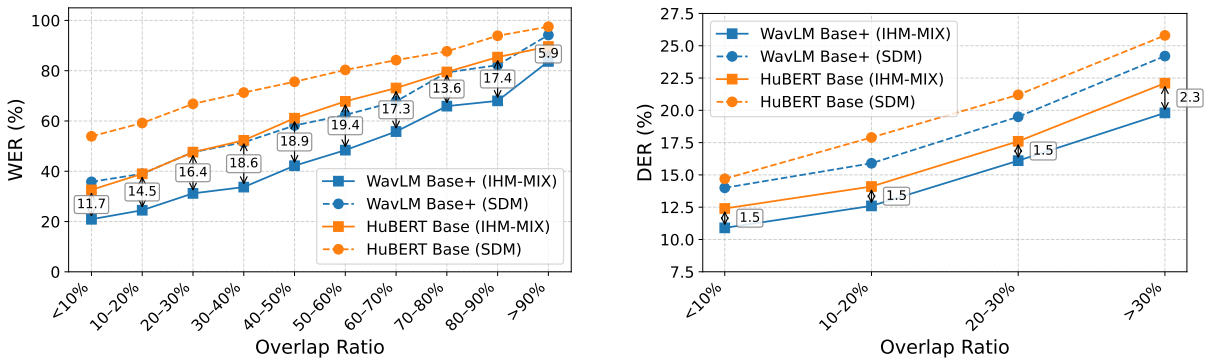
6.2 Impact of Real-World Acoustic Conditions on SSL Model Performance

In this section, we use the HuBERT Base and WavLM Base+ as two SSL model examples to analyze the impact of real-world acoustic conditions on the SSL model performance.

6.2.1 Impact of Recording Conditions

In this section, we leverage the diverse recording setups in the AMI dataset to investigate how acoustic conditions affect the performance of SSL speech models. A summary of these setups is provided in Table 5. The Independent Headset Microphone (IHM) setup uses close-talking condenser microphones worn by each participant. While generally providing clean, speaker-specific recordings, these microphones can occasionally capture side speech from nearby speakers. The IHM-mix configuration sums all IHM streams from a meeting into a single-channel signal, simulating a multi-talker scenario with near-field speech. The Multiple Distant Microphone (MDM) setup uses an 8-channel array of omnidirectional electret microphones placed in the meeting room to capture far-field speech from all participants, often including reverberation and background noise. Finally, the Single Distant Microphone (SDM) condition extracts audio from the first channel of the MDM array and is used as a representative single-channel distant recording for evaluating SSL models.

As shown in Table 6, recording conditions have a substantial effect on the performance of SSL models, including those designed for robustness such as WavLM. The far-field SDM recordings, which are more affected by noise and reverberation, consistently underperform the near-field IHM-mix condition across all downstream tasks. The performance gap between IHM and IHM-mix can be attributed to speaker overlap, which we analyze further in Section 6.2.2. The MDM-BF setup, which leverages spatial cues from multiple microphones through beamforming, consistently outperforms the SDM setup. Among SSL models, WavLM Base+ demonstrates greater robustness compared to HuBERT Base, exhibiting smaller performance degradation under more challenging acoustic conditions.



(a) WER vs. overlap ratio

(b) DER vs. overlap ratio

Figure 2: ASR and SD performance vs. overlap ratio for HuBERT Base and WavLM Base+ under IHM-MIX and SDM conditions. Metrics are computed at the segment level for ASR and the session level for SD.

Condition	Noise	Reverb	Overlap	Far-field
IHM	✓	✗	✗	✗
IHM-MIX	✓	✗	✓	✗
SDM	✓	✓	✓	✓
MDM	✓	✓	✓	✓

Table 5: Summary of recording conditions in the AMI dataset. The table indicates the presence (✓) or absence (✗) of specific acoustic challenges—noise, reverberation, overlapping speech, and far-field effects—across different recording setups.

6.2.2 Impact of Speaker Overlap

Conversational speech is characterized by natural speaker overlap, posing additional challenges for SSL models. In this section, we investigate the impact of speaker overlap on SSL model performance. Figure 2 shows ASR and SD performance as a function of the overlap ratio under the IHM-MIX and SDM conditions of the AMI dataset. The overlap ratio is defined as $overlap\_ratio = \frac{T_{spk\geq 2}}{T_{spk\geq 1}}$, where $T_{spk\geq k}$ is the speech duration with no less than k active speakers.

We compute overlap ratios at the standard evaluation granularity of each task: segment-level for ASR and session-level for SD. Consequently, ASR segments—particularly short backchannel responses occurring while other speakers are active (e.g., "yeah", "hmm")—can exhibit high overlap ratios. In contrast, session-level overlap ratios for SD are substantially lower, since overlap ratio is computed over the entire recording.

As illustrated in Figure 2, the performance of both ASR and SD tasks degrades as the overlap ratio increases. WavLM Base+ consistently outperforms HuBERT Base across all overlap conditions. To isolate the impact of speaker overlap from far-field effects, we analyze the IHM-MIX results (solid lines with square markers). In ASR, the WER gap between HuBERT Base and WavLM Base+ increases from 11.7% at overlap ratios below 10% to a maximum of 19.4% in the 50–60% range. In SD, the absolute DER gap is 1.5% when the overlap ratio is below 30% and widens to 2.3% at higher overlap levels. These results demonstrate the superior robustness of WavLM Base+ to overlapping speech.

7 Conclusion

In this paper, we introduced CSPB, a comprehensive benchmark for evaluating self-supervised speech models under realistic conversational conditions. Through systematic comparisons, we showed that SSL models pretrained on large-scale and diverse corpora with augmentation (e.g., WavLM Large) consistently outperform those trained on clean and controlled speech. We further demonstrated that exploiting spatial information—either through beamforming frontends or multi-channel SSL models such as UniX-Enc—yields additional and consistent performance gains. Beyond SSL approaches, we also evaluated supervised models from the Whisper family, which provide strong performance references, particularly for ASR and SD. We hope that CSPB will facilitate more rigorous evaluation and inspire the development of robust and spatially aware speech representation models for real-world conversational applications.

Limitations

While CSPB represents an important step toward robust evaluation of SSL models in realistic conversational settings, several limitations remain. First,

Condition	HuBERT Base			WavLM Base+		
	ASR (WER%)	SD (DER%)	SE/SS (WER%)	ASR (WER%)	SD (DER%)	SE/SS (WER%)
IHM	32.2	/	/	22.4	/	/
IHM-MIX	47.4	15.3	25.5	34.8	13.8	22.9
MDM-BF	56.6	17.9	38.4	43.3	16.9	33.9
SDM	65.4	18.8	44.0	49.3	17.2	37.0

Table 6: ASR, SD, and SE/SS performance on AMI across different recording conditions.

the benchmark currently focuses on a limited set of tasks and datasets; incorporating additional downstream tasks and more diverse conversational corpora could further strengthen its coverage. Second, although we include Whisper as a supervised reference, CSPB is primarily designed for encoder-only architectures and may not fully reflect the capabilities of encoder-decoder models in conversational speech understanding. Finally, due to limited computational and data resources, our implementation of the UniX-Encoder is not scaled to larger model sizes. However, expanding multi-channel SSL with increased model capacity and more extensive pretraining data remains a promising avenue for future research.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023. AudioLM: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Hervé Bredin. 2023. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. Interspeech 2023*, pages 1983–1987.

Xuankai Chang, Takashi Maekaku, Yuya Fujita, and Shinji Watanabe. 2022. End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation. In *Proc. Interspeech 2022*, pages 3819–3823.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PmLR.

Zhuo Chen, Naoyuki Kanda, Jian Wu, Yu Wu, Xiaofei Wang, Takuya Yoshioka, Jinyu Li, Sunit Sivasankaran, and Sefik Emre Eskimez. 2023. Speech separation with large-scale self-supervised learning. In *ICASSP*, pages 1–5. IEEE.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Antoni Dimitriadis, Siqi Pan, Vidhyasaharan Sethu, and Beena Ahmed. 2023. Spatial HuBERT: Self-supervised spatial speech representation learning for a single talker from multi-channel audio. *arXiv preprint arXiv:2310.10922*.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, and 1 others. 2021. LeBenchmark: A reproducible framework for assessing self-supervised representation learning from speech. In *Proc. Interspeech 2021*, pages 1439–1443.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2021. Exploring wav2vec 2.0 on speaker verification and language identification. In *Proc. Interspeech 2021*, pages 1509–1513.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, and 1 others. 2021. AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. In *Proc. Interspeech 2021*, pages 3665–3669.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019. End-to-end neural speaker diarization with self-attention. In *IEEE ASRU*, pages 296–303. IEEE.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukáš Burget. 2025. Leveraging self-supervised learning for speaker diarization. In *ICASSP*, pages 1–5. IEEE.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*, pages 15979–15988. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021a. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and 1 others. 2021b. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Proc. Interspeech 2021*, pages 721–725.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, and 1 others. 2024a. Dynamic-SUPERB phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *arXiv preprint arXiv:2411.05361*.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024b. Dynamic-SUPERB: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP*, pages 12136–12140. IEEE.

Kuan Po Huang, Yu-Kuan Fu, Yu Zhang, and Hung-yi Lee. 2022. Improving distortion robustness of self-supervised speech processing tasks with domain adaptation. In *Proc. Interspeech 2022*, pages 2193–2197.

Zili Huang, Yiwen Shao, Shi-Xiong Zhang, and Dong Yu. 2024c. UniX-Encoder: A universal x-channel speech encoder for ad-hoc microphone array speech processing. In *ICASSP*, pages 11991–11995. IEEE.

Kuo-Hsuan Hung, Szu-wei Fu, Huan-Hsin Tseng, Hsin-Tien Chiang, Yu Tsao, and Chii-Wann Lin. 2022. Boosting self-supervised embeddings for speech enhancement. In *Proc. Interspeech 2022*, pages 186–190.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and 1 others. 2003. The ICSI meeting corpus. In *ICASSP*, volume 1, pages I–I. IEEE.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. 2023. IndicSUPERB: A speech processing universal performance benchmark for indian languages. In *AAAI*, pages 12942–12950.

Naoyuki Kanda, Guoli Ye, Yu Wu, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2021. Large-scale pre-training of end-to-end multi-talker asr for meeting transcription with single distant microphone. In *Proc. Interspeech 2021*, pages 3430–3434.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *ICASSP*, pages 4835–4839. IEEE.

Keisuke Kinoshita, Marc Delcroix, and Naohiro Tawara. 2021. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. In *ICASSP*, pages 7198–7202. IEEE.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The AMI meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*.

Yoshiki Masuyama, Xuankai Chang, Samuele Cornell, Shinji Watanabe, and Nobutaka Ono. 2023. End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation. In *SLT*, pages 260–265. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, pages 5206–5210. IEEE.

Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, and 1 others. 2024. LeBenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, page 101622.

Junyi Peng, Takanori Ashihara, Marc Delcroix, Tsubasa Ochiai, Oldrich Plchot, Shoko Araki, and Jan Černockỳ. 2025. TS-SUPERB: A target speech processing benchmark for speech self-supervised learning models. In *ICASSP*, pages 1–5. IEEE.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*, pages 2227–2237.

Chandan KA Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. 2021. INTERSPEECH 2021 Deep Noise Suppression Challenge. In *Proc. Interspeech 2021*, pages 2796–2800.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. MMAU: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *ICASSP*, pages 351–355. IEEE.

Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas. 2020. Dipco—dinner party corpus. In *Proc. Interspeech 2020*, pages 434–436.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, and 1 others. 2023. ML-SUPERB: Multilingual speech universal performance benchmark. In *Proc. Interspeech 2023*, pages 884–888.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. In *ACL*, pages 8906–8937.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. SLUE: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP*, pages 7927–7931. IEEE.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *ICLR*.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, and 1 others. 2022. SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *ACL*, pages 8479–8492.

Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, and 1 others. 2022. HEAR: Holistic Evaluation of Audio Representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR.

Alon Vinnikov, Amir Ivry Mark, Aviv Hurvitz, Igor Abramovski, Sharon Koubi, Ilya Gurvich, Shai Peer, Xiong Xiao, Benjamin Elizalde, Naoyuki Kanda, and 1 others. 2024. NOTSOFAR-1 challenge: New datasets, baseline, and tasks for distant meeting transcription. In *Proc. CHiME 2024*.

Yiming Wang, Jinyu Li, Heming Wang, Yao Qian, Chengyi Wang, and Yu Wu. 2022. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. In *ICASSP*, pages 7097–7101. IEEE.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, and 1 others. 2020. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *Proc. CHiME 2020*, pages 1–7.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, and 1 others. 2024. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *ACL*, pages 1979–1998.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Yufeng Yang, Desh Raj, Ju Lin, Niko Moritz, Junteng Jia, Gil Keren, Egor Lakomkin, Yiteng Huang, Jacob Donley, Jay Mahadeokar, and 1 others. 2025. M-best-rq: A multi-channel speech foundation model for smart glasses. In *ICASSP*, pages 1–5. IEEE.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, and 1 others. 2022. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In *ICASSP*, pages 6167–6171. IEEE.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, and 1 others. 2023. MARBLE: Music Audio Representation Benchmark for Universal Evaluation. *Advances in Neural Information Processing Systems*, 36:39626–39647.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.

Qiu-Shi Zhu, Long Zhou, Jie Zhang, Shu-Jie Liu, Yu-Chen Hu, and Li-Rong Dai. 2023. Robust data2vec: Noise-robust speech representation learning for asr by combining regression and improved contrastive learning. In *ICASSP*, pages 1–5. IEEE.

Kateřina Žmolíková, Simone Merello, Kaustubh Kalgaonkar, Ju Lin, Niko Moritz, Pingchuan Ma, Ming Sun, Honglie Chen, Antoine Saliou, Stavros Petridis, and 1 others. 2024. The CHiME-8 MMCSG Challenge: Multi-modal conversations in smart glasses. In *Proc. CHiME 2024*, pages 7–12.

A Additional Results

A.1 Full ASR Results

A.2 Full Speaker Diarization Results

A.3 Full Speech Enhancement / Separation Results

B Upstream Speech Models for Evaluation

C Multi-Channel Multi-Speaker Mixture Preparation

Algorithm 1 Simulation of Multi-Channel Multi-Speaker Mixture

Require: Clean speech dataset from IHM $\mathcal{D}_{\text{IHM}}$, noise dataset $\mathcal{D}_{\text{noise}}$

1: Sample two clean utterances $s_1, s_2 \in \mathcal{D}_{\text{IHM}}$, and a noise $n \in \mathcal{D}_{\text{noise}}$

2: Randomly sample the room dimensions $\mathbf{r} \sim \mathcal{U}((3,3,2.5),(9,9,4))$ and the reverberation time $\text{RT}_{60} \sim \mathcal{U}(0.05, 0.6)$

3: Randomly sample speaker positions $\mathbf{p}_{s_1}$, $\mathbf{p}_{s_2}$ and $\mathbf{p}_c$ for microphone array center; use the array topology of the target dataset

4: Simulate room impulse responses (RIRs) $\mathbf{R}_1, \mathbf{R}_2 \in \mathbb{R}^{C \times T_{\text{rir}}}$ for speaker 1 and 2

5: Convolve clean speech with RIRs:

$$s_1^{\text{reverb}} = \mathbf{R}_1 * s_1, \quad s_2^{\text{reverb}} = \mathbf{R}_2 * s_2$$

6: Randomly shift s_1^{reverb} or s_2^{reverb} to create partial overlap

7: Scale s_2^{reverb} to target SIR $\in [-5, 5]$ dB, and noise n to target SNR $\in [5, 20]$ dB:

$$\tilde{s}_2^{\text{reverb}} = \alpha \cdot s_2^{\text{reverb}}, \quad \tilde{n} = \beta \cdot n$$

where α, β are scaling factors computed to achieve target SIR/SNR

8: Mix signals to create multi-channel multi-speaker mixture:

$$\mathbf{y} = s_1^{\text{reverb}} + \tilde{s}_2^{\text{reverb}} + \tilde{n}$$

D UniX-Enc Implementation

Our implementation of the multi-channel SSL model UniX-Encoder generally follows the design proposed by Huang et al. (2024c), but introduces several key differences. (1) Instead of using raw waveforms as input, we concatenate STFT magnitude and cosine STFT phase as the input features. (2) Rather than pretraining exclusively on synthetic mixtures derived from LibriSpeech (Panayotov et al., 2015), we leverage a combination of real and simulated data. The real datasets include AISHELL-4 (Fu et al., 2021), AliMeeting (Yu et al., 2022), AMI (Kraaij et al., 2005), CHiME-6 (Watanabe et al., 2020), ICSI (Janin et al., 2003), and NOTSOFAR-1 (Vinnikov et al., 2024), while the simulated datasets are generated using LibriSpeech (Panayotov et al., 2015) and DNS noise (Reddy et al., 2021). (3) While the original UniX-Encoder employs 6 cross-channel and 6 cross-frame Transformer layers, our implementation uses a shallower configuration of 3 cross-channel Transformers followed by 9 cross-frame Transformers.

E Microphone channel configurations for multi-channel track

Table 11: Microphone channel configurations used for multi-channel evaluation across datasets. For each dataset, we define 2-chan, 4-chan, and all-chan settings based on the original microphone array topology.

Dataset	2-chan	4-chan	All-chan
AMI	0, 4	0, 2, 4, 6	0–7
AliMeeting	0, 4	0, 2, 4, 6	0–7
MMCSG	0, 2	0, 2, 3, 4	0–6
DiPCo	0, 3	0, 2, 4, 6	0–6

Frontend	Upstream	AMI WER%↓	ALM CER%↓	MMCSG WER%↓	DiPCo WER%↓
1-channel					
—	FBANK	89.3	60.2	86.1	97.0
—	wav2vec 2.0 Base	69.0	52.1	57.7	93.1
—	wav2vec 2.0 Large	67.1	54.4	46.4	92.0
—	HuBERT Base	65.4	51.3	51.4	91.6
—	HuBERT Base Robust MGR	62.7	51.2	50.6	91.0
—	HuBERT Large	56.3	49.2	40.8	85.2
—	WAVLM Base	62.8	49.5	47.1	90.5
—	WAVLM Base+	49.3	46.7	37.1	82.6
—	WAVLM Large	**35.5**	42.6	**24.3**	**63.6**
—	UniX-Enc	59.2	45.9	51.0	85.0
—	Whisper Base	55.8	44.4	41.5	80.0
—	Whisper Small	48.6	39.8	34.7	70.7
—	Whisper Medium	45.4	**37.4**	32.0	66.7
2-channel					
BeamformIt	FBANK	87.6	59.3	85.6	96.6
BeamformIt	HuBERT Base	61.0	49.1	47.9	90.6
BeamformIt	HuBERT Large	52.5	47.0	39.7	84.6
BeamformIt	WAVLM Base	57.9	46.8	44.4	88.3
BeamformIt	WAVLM Base+	46.0	44.0	35.3	79.4
BeamformIt	WAVLM Large	**34.0**	40.2	**23.8**	**63.6**
—	UniX-Enc (CPE-F)	56.9	42.5	49.8	83.3
—	UniX-Enc (CPE-T)	54.9	41.4	50.6	81.6
—	Whisper Small	46.3	38.0	34.0	70.2
—	Whisper Medium	43.7	**35.5**	31.3	65.7
4-channel					
BeamformIt	FBANK	86.9	57.4	85.5	96.7
BeamformIt	HuBERT Base	58.3	47.6	48.1	90.1
BeamformIt	HuBERT Large	50.5	45.6	39.6	84.4
BeamformIt	WAVLM Base	54.9	45.5	45.0	87.6
BeamformIt	WAVLM Base+	44.8	42.9	35.3	79.5
BeamformIt	WAVLM Large	**33.1**	39.0	**24.0**	**63.9**
—	UniX-Enc (CPE-F)	55.9	41.8	49.2	82.6
—	UniX-Enc (CPE-T)	54.1	40.2	50.1	81.9
—	Whisper Small	45.2	37.3	34.3	70.5
—	Whisper Medium	42.4	**34.8**	31.5	65.7
All-channel					
BeamformIt	FBANK	86.3	57.4	85.2	96.9
BeamformIt	HuBERT Base	56.6	47.1	47.8	89.7
BeamformIt	HuBERT Large	49.3	45.1	38.8	83.7
BeamformIt	WAVLM Base	53.7	45.0	44.0	88.1
BeamformIt	WAVLM Base+	43.3	42.0	34.8	79.0
BeamformIt	WAVLM Large	**32.7**	38.5	**23.3**	**63.4**
—	UniX-Enc (CPE-F)	55.6	41.5	49.2	83.3
—	UniX-Enc (CPE-T)	53.4	40.2	48.9	81.0
—	Whisper Small	44.2	36.7	33.5	70.1
—	Whisper Medium	41.4	**34.5**	31.6	64.8

Table 7: ASR performance on AMI, ALM, MMCSG, and DiPCo using various upstream representations under different microphone configurations. Results are reported as WER or CER, with lower scores indicating better performance. Rows highlighted in gray correspond to single-channel self-supervised speech models (optionally combined with a BeamformIt front-end). Rows highlighted in yellow correspond to UniX-Enc, a multi-channel SSL model. Fully supervised Whisper models are highlighted in blue to provide a strong performance reference.

Frontend	Upstream	AMI DER%↓	ALM DER%↓	MMCSG DER%↓	DiPCo DER%↓
1-channel					
—	FBANK	22.3	19.8	10.2	31.4
—	wav2vec 2.0 Base	18.6	16.6	8.5	29.3
—	wav2vec 2.0 Large	18.5	16.1	8.2	29.1
—	HuBERT Base	18.8	16.6	8.5	29.4
—	HuBERT Base Robust MGR	17.8	16.7	8.5	28.6
—	HuBERT Large	18.3	15.9	8.3	29.9
—	WAVLM Base	17.6	15.7	8.2	28.8
—	WAVLM Base+	17.2	15.0	7.7	28.1
—	WAVLM Large	15.7	13.6	7.9	26.8
—	UniX-Enc	17.0	14.5	8.1	26.2
—	Whisper Base	16.8	15.0	7.7	27.4
—	Whisper Small	15.5	13.5	7.6	24.8
—	Whisper Medium	**15.3**	**12.9**	**6.9**	**23.6**
2-channel					
BeamformIt	FBANK	21.8	19.8	9.9	31.8
BeamformIt	HuBERT Base	18.7	16.6	8.2	29.5
BeamformIt	HuBERT Large	17.4	16.0	8.4	29.2
BeamformIt	WAVLM Base	17.9	15.7	7.8	28.5
BeamformIt	WAVLM Base+	17.1	15.0	7.6	27.7
BeamformIt	WAVLM Large	15.5	13.4	7.9	27.0
—	UniX-Enc (CPE-F)	17.0	14.1	7.9	26.2
—	UniX-Enc (CPE-T)	16.3	13.6	8.1	26.1
—	Whisper Small	15.3	12.9	7.3	24.7
—	Whisper Medium	**14.6**	**12.5**	**7.1**	**23.7**
4-channel					
BeamformIt	FBANK	22.4	19.8	10.2	32.7
BeamformIt	HuBERT Base	18.2	16.9	8.2	29.4
BeamformIt	HuBERT Large	18.0	15.7	8.0	28.8
BeamformIt	WAVLM Base	17.6	15.5	8.2	28.8
BeamformIt	WAVLM Base+	16.7	14.6	7.6	27.9
BeamformIt	WAVLM Large	15.2	13.6	7.9	26.8
—	UniX-Enc (CPE-F)	17.0	14.3	7.8	26.1
—	UniX-Enc (CPE-T)	16.3	13.4	8.0	26.0
—	Whisper Small	15.9	13.3	7.2	25.5
—	Whisper Medium	**15.1**	**12.9**	**6.9**	**23.5**
All-channel					
BeamformIt	FBANK	21.5	19.3	10.4	31.5
BeamformIt	HuBERT Base	17.9	16.3	8.5	29.5
BeamformIt	HuBERT Large	17.5	15.6	8.4	29.1
BeamformIt	WAVLM Base	17.5	15.5	7.7	28.8
BeamformIt	WAVLM Base+	16.9	14.5	7.9	28.8
BeamformIt	WAVLM Large	16.0	13.8	7.7	27.1
—	UniX-Enc (CPE-F)	17.0	13.8	7.6	26.2
—	UniX-Enc (CPE-T)	16.1	13.8	7.5	25.8
—	Whisper Small	15.6	13.4	7.2	25.4
—	Whisper Medium	**15.4**	**12.5**	**7.0**	**23.5**

Table 8: Diarization performance on AMI, ALM, MMCSG, and DiPCo using various upstream representations under different microphone configurations. Results are reported as DER, with lower scores indicating better performance. Rows highlighted in gray correspond to single-channel self-supervised speech models (optionally combined with a BeamformIt front-end). Rows highlighted in yellow correspond to UniX-Enc, a multi-channel SSL model. Fully supervised Whisper models are highlighted in blue to provide a strong performance reference.

Frontend	Upstream	AMI WER		ALM CER	
		1spk%↓	2spk%↓	1spk%↓	2spk%↓
1-channel					
—	FBANK	45.1	62.6	32.5	64.9
—	HuBERT Base	37.8	52.6	25.0	53.3
—	HuBERT Base Robust MGR	35.2	51.1	24.8	53.2
—	HuBERT Large	36.2	48.6	24.4	46.5
—	wav2vec 2.0 Base	37.2	52.1	26.2	56.0
—	wav2vec 2.0 Large	38.2	52.3	27.7	55.5
—	WAVLM Base	34.6	48.8	24.7	48.2
—	WAVLM Base+	31.4	45.0	21.3	41.5
—	WAVLM Large	**28.6**	**41.4**	**18.6**	**35.0**
—	UniX-Enc	31.6	47.6	19.2	43.0
—	Whisper Base	36.2	48.6	22.4	49.4
—	Whisper Small	32.8	44.8	20.1	41.2
—	Whisper Medium	31.5	45.2	19.7	39.8
2-channel					
BeamformIt	FBANK	38.1	59.8	25.2	53.4
BeamformIt	HuBERT Base	32.5	51.0	20.6	48.2
BeamformIt	HuBERT Large	31.3	47.4	19.3	41.0
BeamformIt	WAVLM Base	30.9	47.3	18.9	42.5
BeamformIt	WAVLM Base+	30.2	43.3	17.4	39.4
BeamformIt	WAVLM Large	**25.7**	**40.2**	16.3	**31.7**
—	UniX-Enc (CPE-F)	30.4	46.1	**16.1**	40.7
—	UniX-Enc (CPE-T)	30.4	45.9	16.8	37.3
—	Whisper Small	31.2	45.1	16.8	35.2
—	Whisper Medium	31.6	44.6	17.1	35.0
4-channel					
BeamformIt	FBANK	37.6	59.5	24.6	52.6
BeamformIt	HuBERT Base	30.9	49.5	19.3	45.8
BeamformIt	HuBERT Large	30.4	46.8	20.3	41.0
BeamformIt	WAVLM Base	29.3	47.0	18.0	39.6
BeamformIt	WAVLM Base+	28.7	42.9	17.2	39.6
BeamformIt	WAVLM Large	**25.3**	**38.7**	16.0	**32.1**
—	UniX-Enc (CPE-F)	30.1	46.3	15.9	38.6
—	UniX-Enc (CPE-T)	30.1	45.0	**15.8**	34.0
—	Whisper Small	30.5	44.7	17.3	39.2
—	Whisper Medium	31.7	45.4	17.0	34.4
All-channel					
BeamformIt	FBANK	35.2	57.5	24.4	51.9
BeamformIt	HuBERT Base	31.3	48.4	19.8	45.3
BeamformIt	HuBERT Large	30.9	45.9	19.5	42.6
BeamformIt	WAVLM Base	28.6	44.8	17.8	39.5
BeamformIt	WAVLM Base+	28.0	42.2	17.3	36.4
BeamformIt	WAVLM Large	**24.7**	**38.8**	**15.4**	**31.3**
—	UniX-Enc (CPE-F)	28.9	45.5	15.8	36.6
—	UniX-Enc (CPE-T)	30.2	44.7	16.0	37.3
—	Whisper Small	29.2	41.9	15.5	35.7
—	Whisper Medium	29.6	42.8	16.3	32.8

Table 9: Speech enhancement/separation performance on AMI and ALM using various upstream representations under different microphone configurations. We report WER or CER on utterance groups containing 1 or 2 speakers, where lower values indicate better performance. Rows highlighted in gray correspond to single-channel self-supervised speech models (optionally combined with a BeamformIt front-end). Rows highlighted in yellow correspond to UniX-Enc, a multi-channel SSL model. Fully supervised Whisper models are highlighted in blue to provide a strong performance reference.

Method	Network	#Params	Stride	Input	Corpus size
FBANK	-	0	10ms	waveform	-
wav2vec 2.0 Base	7-Conv, 12-Trans	95.0M	20ms	waveform	1khr
wav2vec 2.0 Large	7-Conv, 24-Trans	317.4M	20ms	waveform	60khr
HuBERT Base	7-Conv, 12-Trans	94.7M	20ms	waveform	1khr
HuBERT Large	7-Conv, 24-Trans	316.6M	20ms	waveform	60khr
WavLM Base	7-Conv, 12-Trans	94.4M	20ms	waveform	1khr
WavLM Base+	7-Conv, 12-Trans	94.4M	20ms	waveform	94khr
WavLM Large	7-Conv, 24-Trans	315.5M	20ms	waveform	94khr
UniX-Enc	2-Conv, 12-Trans	90.7M	20ms	STFT	0.7khr
Whisper Base	2-Conv, 6-Trans	19.8M	20ms	log-Mel spec	680khr
Whisper Small	2-Conv, 12-Trans	87.0M	20ms	log-Mel spec	680khr
Whisper Medium	2-Conv, 24-Trans	305.7M	20ms	log-Mel spec	680khr

Table 10: Summary of upstream speech models for evaluation.