The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement

Simon Leglaive, Léonie Borne, Efthymios Tzinis, Mostafa Sadeghi, Matthieu Fraticelli, Scott Wisdom, Manuel Pariente, Daniel Pressnitzer, John R. Hershey

The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement

Simon Leglaive[1], *Léonie Borne*[2], *Efthymios Tzinis*[3*], *Mostafa Sadeghi*[4], *Matthieu Fraticelli*[5], *Scott Wisdom*[6], *Manuel Pariente*[2], *Daniel Pressnitzer*[5], *John R. Hershey*[6]

[1]CentraleSupélec, IETR UMR CNRS 6164, France [2]Pulse Audition, France [3]University of Illinois at Urbana-Champaign, USA [4]Inria, France [5]École Normale Supérieure, PSL University, CNRS, France [6]Google, USA

Abstract

Supervised speech enhancement models are trained using artificially generated mixtures of clean speech and noise signals, which may not match real-world recording conditions at test time. This mismatch can lead to poor performance if the test domain significantly differs from the synthetic training domain. This paper introduces the unsupervised domain adaptation for conversational speech enhancement (UDASE) task of the 7th CHiME challenge. This task aims to leverage real-world noisy speech recordings from the target domain for unsupervised domain adaptation of speech enhancement models. The target domain corresponds to the multi-speaker reverberant conversational speech recordings of the CHiME-5 dataset, for which the ground-truth clean speech reference is unavailable. Given a CHiME-5 recording, the task is to estimate the clean, potentially multi-speaker, reverberant speech, removing the additive background noise. We discuss the motivation for the CHiME-7 UDASE task and describe the data, the task, and the baseline system.

Index Terms: CHiME challenge, multi-speaker conversational speech, speech enhancement, unsupervised domain adaptation.

1. Introduction

Modern speech technologies enable us to connect with each other much beyond standard telephony, for instance through video sharing on social media, remote conferencing, or assistive hearing. For these technologies to be truly effective, they must rely on dependable speech processing algorithms that can work optimally in diverse and uncontrolled acoustic environments. Unfortunately, recordings of speech in real-life situations are inevitably contaminated by unwanted background noise, necessitating the use of speech enhancement algorithms to improve the quality and intelligibility of speech.

The speech enhancement task is to estimate a clean speech signal from a noisy recording [1]. In recent years, there has been great progress in speech enhancement thanks to the use of deep learning algorithms. Most speech enhancement models today rely on deep neural networks that are trained in a supervised manner [2]. Supervised speech enhancement requires a training dataset consisting of noisy speech signals and their corresponding clean reference signals. We say that the noisy signals are labeled with the clean speech signals, which are used as the training targets for the speech enhancement model. Given the impossibility of acquiring such labeled data in real conditions, datasets are generated artificially by creating synthetic mixtures of isolated speech and noise signals. However, it is difficult if not impossible to generate a synthetic training dataset that matches arbitrary acoustic conditions at test time, in terms of noise type and level, recording equipment, speaker-to-microphone distance, reverberation, etc. Artificially generated training data are thus inevitably mismatched with real-world noisy speech recordings, which can result in poor speech enhancement performance in case of severe mismatch. Supervised speech enhancement models are therefore domain-specific; if the test domain deviates from the synthetic training domain, it will be necessary to rebuild a training dataset and retrain the model. These limitations of supervised deep learning methods for speech enhancement contrast with the impressive adaptability of the human auditory system when it comes to perceiving speech in unknown adversary acoustic conditions [3, 4, 5]. They also contrast with early approaches to supervised speech enhancement, e.g., based on hidden Markov models [6] or nonnegative matrix factorization [7], in which adaptation to unseen acoustic conditions was part of the proposed methodologies.

Previous data challenges for single-channel speech enhancement have focused on such supervised setups, where labeled training data are provided that match the evaluation domain. Such challenges include the series of deep noise suppression (DNS) Challenges [8, 9, 10, 11], which also developed the non-intrusive DNSMOS [12] model for automatic quality assessment. The DNS Challenge training sets have consisted of a large amount of multi-condition data intended to cover diverse conditions. Such approaches can be effective, so long as test-time conditions are covered by the training data. In contrast, the proposed challenge is intended to study a different situation where we are targeting single-channel speech enhancement in a specific domain for which no well-matched labeled data are available for training.

Recording unlabeled noisy speech signals in the target domain is much easier than engineering synthetic clean speech and noise mixtures that match this domain. However, leveraging such unlabeled data to develop a speech enhancement model is a challenging problem, which is the focus of the unsupervised domain adaptation for conversational speech enhancement (UDASE) task of the CHiME-7 challenge. The problem we propose to address in this task consists of using unlabeled data in the target domain to adapt supervised speech enhancement models trained on synthetic labeled data in a mismatched source domain. This corresponds to an unsupervised domain adaptation task, but the general problem of improving the generalization capability of speech enhancement models to real-world conditions for which labeled data are not available can also be addressed using fully unsupervised or semi-supervised learning algorithms.

*Now with Google.

In the CHiME-7 UDASE task, the target domain corresponds to the real conversational speech recordings of the CHiME-5 dataset [13]. These recordings were made during dinner parties, so they include multiple speakers having a natural conversation in noisy and reverberant environments. For supervised learning in a mismatched source domain, we rely on the artificially generated LibriMix dataset [14], which is derived from LibriSpeech clean utterances [15] and WHAM! noises [16]. The CHiME-7 UDASE task consists of denoising CHiME-5 conversational speech recordings using the LibriMix out-of-domain (OOD) labeled data and the CHiME-5 in-domain unlabeled data. Given a mixture of one or more reverberant speakers and additive noise, the goal is to predict the clean audio signal of the reverberant speaker(s), removing the additive noise. This task is motivated by the assistive listening use case, in which a speech enhancement algorithm can help any individual to better engage in a conversation, by improving the overall multi-speaker speech intelligibility and quality within the ambient noise. For development and evaluation only, we release the reverberant LibriCHiME-5 dataset, which consists of synthetic mixtures generated to be close to the target domain. Systems submitted to the CHiME-7 UDASE task will be first evaluated using objective performance metrics, then the best-performing systems will be evaluated through a subjective listening test following the ITU-T recommendation P.835 [17].

The tools and resources provided to the participants of the CHiME-7 UDASE task are available in the repositories of our GitHub organization[1] and at the CHiME challenge website[2].

The paper is structured as follows. The datasets and the task to be solved are described in Sections 2 and 3. The baseline is presented in Section 4. We conclude in Section 5.

2. Data

The CHiME-7 UDASE task builds upon the following datasets that will be presented in the next subsections: the CHiME-5 in-domain unlabeled data for training, development and evaluation [13]; the LibriMix OOD labeled data for training and development [14]; the reverberant LibriCHiME-5 close-to-in-domain labeled data for development and evaluation.

2.1. CHiME-5 in-domain unlabeled data

The CHiME-5 data consists of twenty 4-people dinner parties or sessions, of between two and three hours, each recorded in a different home with three recording locations per home (kitchen, dining room, living room) [13]. The audio recordings include natural conversations between multiple speakers in reverberant and noisy environments, and these are fully transcribed. Using the CHiME-5 transcription files, we estimated that 22% of the audio recordings contain only noise, 51% contain one single active speaker, and 20%, 5% and 2% contain two, three and four overlapping speakers, respectively (numbers obtained without considering any constraint on the overlap duration).

For the CHiME-7 UDASE task, we only use the right channel of the binaural recordings because the left channel is less reliable. We also discard portions of the data where the participant wearing the microphone speaks, to simulate an assistive listening situation where all voices but the listener's should be enhanced. The extracted audio segments therefore contain up to three simultaneously-active reverberant speakers and background noise. The noisy speech signals are not labeled with the clean speech reference signals. The main objective of the UDASE task is to develop new approaches that can leverage this in-domain unlabeled dataset for speech enhancement.

The training set consists of raw single-channel audio segments extracted from the binaural recordings. Developing and evaluating a speech enhancement model requires computing objective performance metrics. This is a difficult problem as the CHiME-5 dataset contains noisy multi-speaker speech recordings that are not labeled with clean speech reference signals. For development and evaluation, we therefore used the transcription of the CHiME-5 recordings to extract short audio segments of duration at least 3 seconds labeled with the maximum number of simultaneously-active speakers (0, 1, 2 or 3),[3] which will allow us to compute objective performance metrics. This procedure simulates the reasonable scenario where we can afford to manually annotate a small amount of data with speaker count labels for development and evaluation, but this procedure cannot be easily done for a large training set. The audio segment extracted for the development and evaluation sets is done as follows: (i) we extract all segments where no speaker is active (i.e., noise-only segments); (ii) we extract all segments that were not extracted previously and without overlapping speakers (i.e., single-speaker segments); (iii) we extract all segments that were not extracted previously and with at most two overlapping speakers; (iv) we extract all segments that were not extracted previously and with at most three overlapping speakers. As a post-processing, Brouhaha [18] was used on the 0- and 1-speaker segments to verify the absence and presence of speech, respectively. Misclassified segments were reviewed and removed when appropriate. Noise-only segments are used to create the reverberant LibriCHiME-5 dataset (see Section 2.3), allowing for computing objective performance metrics such as the scale-invariant signal-to-distortion ratio (SI-SDR) [19] on close-to-in-domain data. Single-speaker segments will be used to compute DNSMOS P.835 (simply referred to as DNSMOS hereinafter) metrics [12]. There is no official guideline on how to use the 2- and 3-speaker segments.

We also provide an evaluation subset that will be used for the listening test. It consists of audio samples that were extracted by looking for segments of 4 to 5 seconds with at least 3 seconds of speech and 0.25 second without speech at the beginning and at the end. Additional constraints were taken into account to ensure a balanced subset in terms of speaker gender, recording location, and session.

2.2. LibriMix out-of-domain labeled data

For supervised learning on OOD data, we chose the LibriMix dataset [14] because it is a standard open-source dataset in the community. LibriMix was originally developed for speech separation in noisy environments, it is derived from LibriSpeech clean utterances [15] and WHAM! noises [16]. The Libri2Mix and Libri3Mix versions of the dataset contain noisy speech mixtures with 2 and 3 overlapping speakers, respectively. A single-speaker version of LibriMix (Libri1Mix) can be obtained by simply discarding one of the two speakers in Libri2Mix mixtures. For a complete description of LibriMix, the interested reader is referred to [14].

[1] https://github.com/UDASE-CHiME2023

[2] https://www.chimechallenge.org/

[3] This only corresponds to a maximum value, i.e., through the duration of a segment the number of simultaneously-active speakers can vary between 0 and the maximum value. Moreover, a segment might contain more speakers than the labeled maximum number of simultaneously active speakers. For instance, a segment labeled as single-speaker might contain two active speakers who do not speak simultaneously.

2.3. Reverberant LibriCHiME-5 close-to-in-domain labeled data

In real-world conditions, in particular for the CHiME-5 recordings, it is impossible to have access to the ground-truth clean speech reference signals associated with the noisy recordings due to cross-talk between microphones. Yet, when developing and evaluating a speech enhancement algorithm, it is necessary to compute objective performance metrics. For this purpose, we created the reverberant LibriCHiME-5 dataset for development and evaluation only. This dataset consists of synthetic mixtures of reverberant speech and noise, with up to three simultaneously active speakers, labeled with the clean reference speech signals. Noise signals were extracted from the CHiME-5 recordings using the ground-truth transcriptions, and clean speech utterances were taken from the LibriSpeech dataset [15] and were convolved with room impulse responses (RIRs) from the VoiceHome corpus [20]. These RIRs were recorded in 12 different rooms of 3 real homes, with 4 rooms per home: living room (room 1), kitchen (room 2), bedroom (room 3), and bathroom (room 4). Bathrooms were excluded for the reverberant LibriCHiME-5 dataset. In each room, RIRs were recorded for 2 different positions and geometries of an 8-channel microphone array and 7 to 9 different positions of the loudspeaker.

For each mixture in the reverberant LibriCHiME-5 dataset, we randomly choose the maximum number $n \in \{1, 2, 3\}$ of simultaneously-active speakers in the mixture, with $p(n = i) = 0.60, 0.35, 0.05$ for $i = 1, 2, 3$, respectively, which is consistent with the distribution of the segmented CHiME-5 dataset. In the VoiceHome corpus, we randomly and successively sample a home, a room, an array position/geometry, n speaker positions without replacement, and a channel of the microphone array, which gives the RIRs for the current mixture. LibriSpeech utterances are convolved with the selected RIRs to obtain the reverberant speech utterances. These are mixed following speech activity patterns extracted from the CHiME-5 transcription files to simulate a natural conversation between multiple speakers. Multi-speaker reverberant speech and noise mixtures are created such that the per-speaker signal-to-noise ratio (SNR) is distributed as a Gaussian with a mean of 5 dB and a standard deviation of 7 dB, to match the SNR distribution of the CHiME-5 dataset as estimated by Brouhaha [18]. This is achieved by first sampling a global per-mixture SNR $x \sim \mathcal{N}(5, \sigma_1^2)$ and then sampling a local per-speaker SNR $y \sim \mathcal{N}(x, \sigma_2^2)$, with $\sigma_1 = 6.7082$ and $\sigma_2 = 2$ ($\sqrt{\sigma_1^2 + \sigma_2^2} \approx 7$ dB). The value of σ_2 is chosen such that the loudness difference between multiple speakers is moderate, this is again to simulate a conversation.

Despite the effort to generate a synthetic dataset that matches the distribution of the target domain as much as possible, there still exists a mismatch between the reverberant LibriCHiME-5 dataset and the CHiME-5 dataset, e.g., read speech for the latter and spontaneous speech for the former; the reverberation times might also differ. It is indeed impossible to create synthetic labeled data that perfectly match real-world unlabeled recordings, hence the CHiME-7 UDASE task. Nevertheless, as already mentioned, it is required for development and evaluation to be able to compute objective performance metrics, complementary to listening tests. DNSMOS [12] provides a way to evaluate the performance on single-speaker segments of the CHiME-5 data without having access to the clean speech reference signals, but this is not sufficient as a non-negligible proportion of the CHiME-5 data contains simultaneously-speaking people. We believe it is reasonable to expect systems that successfully managed to leverage the unlabeled CHiME-5 data to have better results on the reverberant LibriCHiME-5 dataset than fully supervised systems only trained on the labeled LibriMix dataset. Indeed, in the reverberant LibriCHiME-5, the speech utterances were convolved with real RIRs measured in domestic environments, the noise signals were extracted from the CHiME-5 recordings, the per-speaker SNR was chosen to approximately match that of the CHiME-5 data, and the speech utterances were mixed to simulate a conversation using the CHiME-5 transcription. We can thus hope that the performance computed on the reverberant LibriCHiME-5 dataset corresponds to an imperfect estimate of the performance on the CHiME-5 dataset.

Subset	# samples	Sample length (s) Mean	Sample length (s) STD	Total duration (HH:MM:SS)
`train`	27 517	10.91	14.10	83:22:29
`dev/0`	912	6.50	4.10	1:38:49
`dev/1`	5 719	5.89	3.49	9:21:53
`dev/2`	3 835	5.23	2.43	5:34:33
`dev/3`	667	4.61	1.84	0:51:14
`eval/0`	977	5.73	3.35	1:33:19
`eval/1`	3 013	5.54	2.94	4:35:05
`eval/2`	1 552	4.88	2.04	2:06:07
`eval/3`	233	4.21	1.17	0:16:21
`eval/LT`	241	4.72	0.34	0:18:58

Table 1: *Segmented CHiME-5 dataset. Dev and eval subsets are labeled with the maximum number of simultaneously active speakers (0, 1, 2, 3).* `eval/LT` *corresponds to the evaluation subset for the listening test.*

Subset	# samples	Sample length (s) Mean	Sample length (s) STD	Total duration (HH:MM:SS)
`dev/1`	1 187	7.14	4.67	2:21:09
`dev/2`	565	5.37	2.24	0:50:31
`dev/3`	65	4.81	1.66	0:05:12
`eval/1`	1 394	6.25	3.75	2:25:17
`eval/2`	494	4.44	1.34	0:36:35
`eval/3`	64	4.21	1.07	0:04:29

Table 2: *Reverberant LibriCHiME-5 dataset. The subsets are labeled with the maximum number of simultaneously active speakers (0, 1, 2, 3).*

3. Task

3.1. Training, development, and evaluation sets

In the original CHiME-5 dataset [13], the 20 sessions (or dinner parties) were divided into disjoint training (train), development (dev), and evaluation (eval) sets [13]. For the CHiME-7 UDASE task, we move sessions S07 and S17 from the train set to the dev set to obtain a sufficient amount of noise-only segments for the generation of the reverberant LibriCHiME-5 dataset. There is no overlap between speakers in each set. The segmented CHiME-5 dataset for the CHiME-7 UDASE task is summarized in Table 1.

The dev (resp. eval) set of the reverberant LibriCHiME-5 dataset is created from the 'dev-clean' (resp. 'test-clean') subset of LibriSpeech, noise-only segments from the dev (resp. eval) set of CHiME-5, and a subset of VoiceHome RIRs. RIRs from home 2 (rooms 1, 2, 3) and home 3 (rooms 1, 3) are used for the

dev set, and RIRs from home 3 (room 2) and home 4 (rooms 1, 2, 3) are used for the eval set. The reverberant LibriCHiME-5 dataset is summarized in Table 2.

The original split of the LibriMix dataset [14] into train, dev, and eval (test) subsets is kept for the CHiME-7 UDASE task.

3.2. Rules

The train and dev sets of the LibriSpeech and WHAM! datasets (from which LibriMix is generated) can be used individually (e.g., to train isolated speech and noise models) or they can be used to create synthetic mixtures similar to the original LibriMix dataset. Specifically, participants are allowed to create synthetic mixtures using noise-only segments that would be extracted from the binaural recordings of the CHiME-5 training set, only if this extraction does not rely on the CHiME-5 ground-truth transcription. Participants are also allowed to use RIRs to create reverberant utterances from LibriSpeech, as long as the RIRs are synthetic. Using any other datasets of clean speech signals, noise signals, or measured RIRs is not allowed. Finally, the Kinect recordings of the CHiME-5 dataset cannot be used. Although a synthetic labeled dataset better matching with the real CHiME-5 data could be created with more engineering effort and knowledge about the target domain, the goal of the CHiME-7 UDASE task is to simulate more realistic conditions where knowledge about in-domain data is scarce. The motivation for the above rules is to encourage participants to use a relatively identical synthetic dataset and show that models trained with OOD labeled data can be adapted using unsupervised, self- or semi-supervised learning from in-domain unlabeled data.

All speech enhancement system parameters should be tuned on the training set or development set of the LibriMix, CHiME-5 and reverberant LibriCHiME-5 datasets as described in Section 3.1, or variations that comply with the above rules. During evaluation or inference, the submitted systems must use as input only noisy speech waveforms and process them independently of one another. Participants can use external pre-trained and frozen models for voice activity detection, diarization, speaker counting, or signal-to-noise ratio estimation.

3.3. Evaluation

The submitted systems will follow a two-step evaluation process. They will first be evaluated in terms of SI-SDR [19] on the complete eval set of the reverberant LibriCHiME-5 dataset and in terms of DNSMOS scores on the `eval/1` subset of the CHiME-5 dataset. DNSMOS is a non-intrusive objective metric that provides performance scores for the speech signal quality (SIG), the background intrusiveness (BAK), and the overall quality (OVRL) [12]. The four best-performing systems in terms of SI-SDR or OVRL score will then be evaluated by a listening test using audio samples from the `eval/LT` subset of the CHiME-5 dataset. In case a team submits multiple entries, only the one that obtains the best performance during the first evaluation stage will be qualified for the listening test.

The evaluation data will be released two weeks before the submission deadline. Participants are asked to evaluate their system using the provided evaluation scripts, and they are asked to return the performance scores for each audio file of the `eval/1` subset of the CHiME-5 dataset and of the `eval/{1,2,3}` subsets of the reverberant LibriCHiME-5 dataset. They are also asked to submit the output signals of their system to allow their scores to be verified, and a technical report describing their system. Participants are asked to normalize the output signals at a loudness of -30 LUFS (Loudness Unit Full Scale) before computing the DNSMOS performance scores, using the Python package `pyloudnorm` [21]. The motivation for this normalization is that DNSMOS scores (especially the SIG and BAK scores) are very sensitive to a change in the input signal loudness. This sensitivity would make it difficult to compare different systems without a common normalization procedure. The same normalization is considered for the listening test material.

Optionally, participants are also invited to submit SI-SDR scores for the LibriMix dataset ('max' version), using the `test/mix_single` and `test/mix_both` subsets of Libri2Mix (containing 3000 single-speaker and 2-speaker examples, respectively) and the `test/mix_both` subset of Libri3Mix (containing 3000 3-speaker examples). SI-SDR results on LibriMix will not be used to rank systems because it would not be consistent with the purpose of the CHiME-7 UDASE task. They will only be used to compare the performance on the (close to) in-domain and OOD datasets.

The listening test will follow the ITU-T Recommendation P.835 [17]. It will be conducted in person in a listening booth at the University of Sheffield. Participants will listen over headphones to short speech samples (45 seconds). Each trial will consist of three presentations of the same sample, to collect three different subjective reports. In the different presentations participants will be instructed to either focus on the speech signal and rate how natural it sounds, focus on the background noise and rate how noticeable or intrusive this background is, or attend to both the speech and the background noise and rate the overall quality of the sample, quality being defined in the perspective of everyday speech communication. The order of presentations will be counterbalanced across participants. The ratings will be reported on 5-point Likert scales and mean opinion scores (MOS) will be computed.

We target a total number of 32 subjects, separated into 4 panels of 8 listeners. Each panel will be associated with a distinct set of 32 audio samples taken from the `eval/LT` subset of the CHiME-5 dataset, resulting in a total of $32 \times 4 = 128$ audio samples for the entire listening test. For each audio sample we will have 5 different experimental conditions (4 systems and the noisy input condition). Each listener will evaluate all experimental conditions for each audio sample associated with his/her panel, according to the three rating scales. For each pair of audio sample and experimental condition, a MOS will be computed out of 8 votes, leading to an overall $8 \times 128 = 1024$ votes for each experimental condition. The final ranking of the systems will be based on statistical analysis of the MOS results.

4. Baseline

4.1. Baseline system

The CHiME-7 UDASE baseline system is based on RemixIT [22, 23], a self-supervised learning approach for unsupervised domain adaptation of a speech enhancement model pre-trained (in a supervised or a self-supervised manner) on OOD noisy speech data. In semi-supervised RemixIT, a teacher model is trained on OOD noisy speech signals alongside the corresponding clean speech and noise reference waveforms. In the second step, RemixIT performs inference on a batch of noisy in-domain speech recordings to obtain pseudo-labels that are used to train a student model for speech enhancement in the target domain without the need for in-domain reference signals.

4.1.1. OOD Sudo rm -rf teacher model

The teacher is based on a Sudo rm -rf [24, 25] sound separation model, which is an end-to-end framework with three main blocks: (i) an encoder network processing the raw waveform of the input audio mixture; (ii) a separator network that operates on the encoder output to provide separation masks; (iii) a decoder network to estimate the audio source signals from the encoder output and the estimated masks. The encoder and decoder architectures consist of a one-dimensional convolution and transpose convolution, respectively, with 512 filters of 41 taps and a hop size of 20 samples. The backbone structure of the separator network consists of 8 U-Conv blocks where each block extracts and aggregates information from multiple resolutions.

The Sudo rm -rf teacher model is trained fully supervised on the LibriMix OOD labeled dataset. We use the 3-speaker version of the dataset (Libri3Mix) and discard one (two) speakers in the original Libri3Mix mixtures to obtain 2- (single-) speaker data. The proportion of single-speaker, 2-speaker, and 3-speaker mixtures are 0.50, 0.25, and 0.25, respectively. The model is trained by minimizing the negative SI-SDR loss for both the speech and the noise components with equal weights.

4.1.2. Self-supervised RemixIT student model

The student model follows the same architecture as the teacher model, and it is initialized using the model parameters resulting from the supervised training on LibriMix. The RemixIT learning framework consists of feeding noisy speech signals from the training set of the unlabeled CHiME-5 dataset in the frozen teacher model to get estimates of the isolated speech and noise, which will serve as pseudo-labels. In a second step, new bootstrapped mixtures are synthesized by permuting the aforementioned noise estimates and remixing them with the speech estimates. These bootstrapped mixtures and the corresponding pseudo-labels are finally used to train the student model, again by minimizing the SI-SDR loss between the student's speech and noise estimates and the corresponding teachers' pseudo-targets. The teacher model is also continuously updated using the parameters of the student model, following an exponential moving average update. The final student model is chosen as the one obtaining the highest mean overall MOS (OVRL) as computed by DNSMOS on the single-speaker subset of the CHiME-5 dev set (`dev/1` subset).

We provide two versions of the student model. The first one is trained on the raw audio segments of the CHiME-5 `train` set, which may result in sub-optimal performance because these audio segments do not always contain speech. Remixing possibly perturbed noise waveforms with almost zero teacher's speech estimates will not always lead to valid new bootstrapped mixtures for training the student model. Therefore, we provide a second student model that was trained on audio segments of the CHiME-5 `train` set which have been automatically labeled as containing speech by Brouhaha's voice activity detector (VAD) [18].

4.2. Results and discussion

Results are shown in Table 3. In this table, OOD teacher corresponds to the fully supervised Sudo rm -rf teacher model, while RemixIT and RemixIT-VAD correspond to the student models, the latter being trained on the data preprocessed by Brouhaha's VAD. SIG, BAK, and OVRL correspond to the DNSMOS scores (in between 1 and 5, the higher the better) and are averaged over the 1-speaker subset of the CHiME-5 dev or eval sets. For the reverberant LibriCHiME-5 dataset, the SI-SDR scores (in dB, the higher the better) are averaged over the entire dev or eval sets (including all 1-, 2-, and 3-speaker mixtures). This is the same for the LibriMix dataset, where the `eval` row contains the results averaged over the eval subsets indicated in Section 3.3, and the `dev` row contains the results averaged over the equivalent subsets of the LibriMix dev set (i.e., `dev/{mix_single, mix_both}` for Libri2Mix and `dev/mix_both` for Libri3Mix).

Subset	Metric	Input	OOD teacher	RemixIT	RemixIT-VAD
		LibriMix dataset			
`dev`	SI-SDR	5.2	**13.2**	11.9	12.3
`eval`		4.9	**13.2**	11.5	12.2
		CHiME-5 dataset			
`dev/1`	SIG	**3.64**	3.48	3.44	3.46
	BAK	3.04	3.79	**3.85**	**3.85**
	OVRL	3.03	3.08	3.07	**3.09**
`eval/1`	SIG	**3.48**	3.33	3.26	3.28
	BAK	2.92	3.59	**3.64**	3.62
	OVRL	2.84	**2.88**	2.82	2.84
		Reverberant LibriCHiME-5 dataset			
`dev`	SI-SDR	6.6	8.3	9.5	**9.9**
`eval`		6.6	7.8	9.4	**10.1**

Table 3: *Results computed from the unprocessed noisy speech signals (Input) and from the output signals of the three baseline systems (OOD teacher, RemixIT, and RemixIT-VAD).*

Using RemixIT training, we expect the performance of the speech enhancement student model to improve on the target domain (corresponding to the CHiME-5 data) and to deteriorate on the synthetic training domain (corresponding to the LibriMix data), compared to the OOD teacher model. This is globally what we observe in Table 3. In terms of SI-SDR on the LibriMix dataset, the self-supervised student models, namely, RemixIT and RemixIT-VAD perform between 0.9 and 1.7 dB worse than the fully-supervised teacher model as they are getting more fine-tuned towards the target domain. However, as will be discussed in the next paragraphs, the RemixIT student models globally outperform the fully-supervised teacher model on the CHiME-5 and reverberant LibriCHiME-5 datasets.

On the CHiME-5 dataset, the student models obtain better BAK scores compared to the OOD teacher on both the dev and eval set (between +0.03 and +0.06 points), indicating a reduction of the background noise intrusiveness in the output signal. Regarding, the SIG metric, all models obtain lower performance than the unprocessed noisy speech signals, which is expected because these are contaminated by noise but not distorted. Unfortunately, as indicated by the SIG scores the distortion introduced by the speech enhancement process is more important for the student models than for the teacher, which is particularly true for the `eval/1` subset. Regarding the OVRL metric on the `dev/1` set, the performance of all models is close. RemixIT-VAD outperforms RemixIT and obtains a marginal improvement of 0.01 point compared to the OOD teacher, which we nevertheless considered as sufficient for the baseline of a new challenge. However, the OOD teacher outperforms the student models by a margin between 0.04 and 0.06 points in terms of OVRL score on the CHiME-5 `eval/1` set, which we assume is mainly due to the increased distortion for the student models.

Finally, it can be seen that RemixIT's unsupervised adaptation of the OOD Sudo rm -rf teacher model on the in-domain CHiME-5 dataset improved the speech enhancement performance on the close-to-in-domain reverberant LibriCHiME-5

dataset. The student models obtain better performance compared to the fully supervised teacher model, and RemixIT-VAD is the best-performing system with an improvement between 1.6 and 2.3 dB in terms of SI-SDR. Consequently, an unsupervised domain adaptation method that leverages the CHiME-5 data could obtain better separation performance on the reverberant LibriCHiME-5 dataset than a fully-supervised model only trained on the OOD LibriMix data.

5. Conclusion

In this paper, we presented the CHiME-7 UDASE task, which aims to foster new methods toward more ecologically valid and robust speech enhancement models. Ecological validity describes the extent to which an experimental setting and task correspond to real-life conditions [26]. Fully-supervised speech enhancement models trained (and most of the time also evaluated) on synthetic data cannot always capture the distribution of real-world acoustic recordings, which has important implications in terms of generalization capability. Evaluating unsupervised domain adaptation methods for speech enhancement is by definition a challenging task because one cannot have access to the ground-truth clean speech signals in the target domain. Hopefully, the design of the CHiME-7 UDASE task will enable the further development and evaluation of such methods in the future. We will evaluate the systems submitted to the CHiME-7 UDASE task and the results will be announced during the CHiME-2023 workshop.

6. Acknowledgments

The authors thank the CHiME Steering Group (Jon Barker, Emmanuel Vincent, Shinji Watanabe, Michael Mandel, Marc Delcroix, Leibny Paola Garcia Perera) for their support in the organization of the CHiME-7 UDASE task and for their suggestion of using the binaural microphone recordings of the CHiME-5 dataset to create the task material.

7. References

[1] P. C. Loizou, *Speech enhancement: Theory and practice*. CRC press, 2013.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] T. Bent, A. Buchwald, and D. B. Pisoni, "Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech," *J. Acoust. Soc. Am.*, vol. 126, no. 5, 2009.

[4] E. Brandewie and P. Zahorik, "Prior listening in rooms improves speech intelligibility," *J. Acoust. Soc. Am.*, vol. 128, no. 1, 2010.

[5] M. Cooke, O. Scharenborg, and B. T. Meyer, "The time course of adaptation to distorted speech," *J. Acoust. Soc. Am.*, vol. 151, no. 4, 2022.

[6] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 6, no. 5, 1998.

[7] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, 2013.

[8] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. INTERSPEECH*, 2020.

[9] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 6623–6627.

[10] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *Proc. INTERSPEECH*, 2021.

[11] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper *et al.*, "ICASSP 2022 deep noise suppression challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 9271–9275.

[12] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 886–890.

[13] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. INTERSPEECH*, 2018, pp. 1561–1565.

[14] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 5206–5210.

[16] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in *Proc. INTERSPEECH*, 2019, pp. 1368–1372.

[17] I. Recommendation, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T recommendation*, p. 835, 2003.

[18] M. Lavechin, M. Mtais, H. Titeux, A. Boissonnet, J. Copet, M. Rivire, E. Bergelson, A. Cristia, E. Dupoux, and H. Bredin, "Brouhaha: Multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation," *arXiv preprint arXiv: Arxiv-2210.13248*, 2022.

[19] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 626–630.

[20] N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, É. Lamandé, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom *et al.*, "A French corpus for distant-microphone speech processing in real homes," in *Proc. INTERSPEECH*, 2016, pp. 2781–2785.

[21] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES Convention*, 2021.

[22] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, and A. Kumar, "Continual self-training with bootstrapped remixing for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022, pp. 6947–6951.

[23] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1329–1341, 2022.

[24] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2020, pp. 1–6.

[25] E. Tzinis, Z. Wang, X. Jiang, and P. Smaragdis, "Compute and memory efficient universal sound source separation," *J. Signal Process. Syst.*, vol. 94, no. 2, pp. 245–259, 2022.

[26] H. Hung, E. Gedik, and L. Cabrera Quiros, "Complex conversational scene analysis using wearable sensors," in *Multimodal Behavior Analysis in the Wild*, ser. Computer Vision and Pattern Recognition, X. Alameda-Pineda, E. Ricci, and N. Sebe, Eds. Academic Press, 2019, pp. 225–245.