Interspeech 2024
1-5 September 2024, Kos, Greece

LibriheavyMix: A 20,000-Hour Dataset for Single-Channel Reverberant Multi-Talker Speech Separation, ASR and Speaker Diarization

Zengrui Jin[1,3]*, Yifan Yang*[1]*, Mohan Shi*[2]*, Wei Kang*[1]*, Xiaoyu Yang*[1]*, Zengwei Yao*[1]*, Fangjun Kuang*[1]*, Liyong Guo*[1]*, Lingwei Meng*[3]*, Long Lin*[1]*, Yong Xu*[2]*, Shi-Xiong Zhang*[2]*, Daniel Povey*[1]

[1]Xiaomi Corp., Beijing, China; [2]Tencent AI Lab, Bellevue, USA;
[3]The Chinese University of Hong Kong, Hong Kong SAR, China

zengrui.jin@link.cuhk.edu.hk; yifanyeung@sjtu.edu.cn; shimohan@g.ucla.edu

Abstract

The evolving speech processing landscape is increasingly focused on complex scenarios like meetings or cocktail parties with multiple simultaneous speakers and far-field conditions. Existing methodologies for addressing these challenges fall into two categories: multi-channel and single-channel solutions. Single-channel approaches, notable for their generality and convenience, do not require specific information about microphone arrays.

This paper presents a large-scale far-field overlapping speech dataset, crafted to advance research in speech separation, recognition, and speaker diarization. This dataset is a critical resource for decoding "Who said What and When" in multi-talker, reverberant environments, a daunting challenge in the field. Additionally, we introduce a pipeline system encompassing speech separation, recognition, and diarization as a foundational benchmark. Evaluations on the WHAMR! dataset validate the broad applicability of the proposed data.

Index Terms: Multi-Talker, Speech Recognition, Speech Separation, Speaker Diarization, Cocktail Party Problem

1. Introduction

Despite the rapid progress of automatic speech recognition (ASR) technologies targeting single-talker, near-field speech [1, 2, 3], these regular methods and datasets [4, 5, 6] cannot handle the scenario where multiple speakers are presented simultaneously.

Existing works on speech separation [7, 8, 9, 10] and multi-talker ASR [11, 12, 13, 14, 15, 16, 17] have been conducted on simulated multi-talker overlapping speech datasets [7, 18, 19, 20]. However, most of these datasets neither take reverberation in the far-field condition into consideration, nor deliver sufficient amount of data for the model to be generalized to other datasets [21, 22]. In addition, most of these datasets are simple cases with only 1 speaker turns, which does not match the real-world conversational scenarios where multiple speaker turns are common. Recently, some real-world recorded multi-talker overlapping speech datasets [23, 24] are proposed with far-field reverberation and multiple speaker turns presented. However, the amount of data delivered by these datasets is still not large enough due to the very high recording cost. Moreover, it is difficult to obtain clean separation targets from real-world recorded data, limiting their capability as training data for speech separation models.

In this work, we propose a 20,000-hour multi-talker overlapping speech dataset LibriheavyMix based on Libriheavy [6], which is a large-scale ASR corpus with richer information including punctuation casing and text context. We conduct preliminary experiments on speech separation and multi-talker ASR on the proposed dataset and present the corresponding baseline results. Compared with previous work, LibriheavyMix presents the following advantages: **(1)** The amount of data is much larger than the others, with 10,000 hours. **(2)** Reverberation is introduced to simulate real-world far-field scenarios. **(3)** Multiple speaker turns, which is consistent with the real-world conversational scenarios, can be further used for speaker diarization [25, 26, 27, 28] and speaker-attributed ASR [29, 30, 31, 32]. **(4)** Punctuation, casing and text context are inherent in transcripts, which can be further combined with the research of punctuation and semantic information [33, 34].

The rest of this paper is organized as follows. Section 2 presents the methods of data simulation. Section 3 shows the baseline systems of speech separation and multi-talker ASR. Section 4 shows experiments and results of baseline systems. Finally, Section 5 concludes this work.

Table 1: *Statistics of simulated speech separation datasets. Note that the # Hours listed for the training sets of the LibriheavyMix dataset is determined by summing the durations of all mixtures involving 1-4 speakers in total.*

Dataset	wsj0-mix [7]	WHAM! [18]	Libri2Mix [19]	Libri3Mix [19]	WHAMR! [20]	**LibriheavyMix** (Ours)
Reverberant	-	-	-	-	✓	✓
Multi-Turn	-	-	-	-	-	✓
Split	train (30h) dev (8h) test (5h)	train (30h) dev (8h) test (5h)	train-360 (212h) train-100 (58h) dev (11h) test (11h)	train-360 (146h) train-100 (40h) dev (11h) test (11h)	train (30h) dev (8h) test (5h)	train-*small* (240h) train-*medium* (2,000h) train-*large* (**18,000h**)

2. Data Simulation

Simulation of Overlapped Speech: As described in Algorithm 1, $D_{=\text{spk}}$, $D_{\neq\text{spk}}$ and D_{ovlp} stand for the distribution of "duration of pause between the same speaker", "duration of pause between two different speakers" and "duration of overlapping" respectively. The statistics of these distributions are derived from the target sessions provided, and the duration sampled from the distribution is utilized to blend the source utterances. Such a strategy is adopted as it has been successfully applied to improve end-to-end neural diarization [35].

Given the distribution of the target session on "pause between the same speaker" ($D_{=\text{spk}}$), "pause between different speakers" ($D_{\neq\text{spk}}$), "duration of overlapping" (D_{ovlp}) and "probability of overlapping" (P_{ovlp}), a mixture $\mathcal{M}$ of k speakers with a maximum duration of T is simulated by first sampling source utterances from the provided samples $\mathcal{U} = \{\mathcal{U}_{s_1}, \ldots, \mathcal{U}_{s_k}\}$ containing utterances from $\mathcal{S}$ speakers, k denotes the index for distinct speakers. The starting time of each of the selected utterances is sampled based on the speaker

* Equal contribution was made between the three authors.

10.21437/Interspeech.2024-90

Algorithm 1: Simulation of a session of K speakers.

Data: $\mathcal{U}, D_{=\text{spk}}, D_{\neq\text{spk}}, D_{\text{ovlp}}, P_{\text{ovlp}}$
Result: $\mathcal{M}$

$\mathcal{M} \leftarrow \varnothing$
$\mathcal{U} \leftarrow \text{shuffle}(\mathcal{U}_{s_1}, \ldots, \mathcal{U}_{s_k})$
offset $\leftarrow 0$, num_spk $\leftarrow 0$
for $i \leftarrow 1$ **to** range($|\mathcal{U}|$) **do**
  if $\mathcal{U}_i$.spk == $\mathcal{U}_{i-1}$.spk **then**
    ot $\leftarrow$ sample($D_{=\text{spk}}$)
  else
    num_spk+ $= 1$
    if *Bernoulli*($P_{\text{ovlp}} > 0.5$) **then**
      ot $\leftarrow -$sample(D_{ovlp});
    else ot $\leftarrow$ sample($D_{\neq\text{spk}}$);
  end
  offset $\leftarrow$ offset + ot
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{U_i, \text{offset}\}$
  if num_spk $== K$ **then** break;
end

and the provided distributions $D_{=\text{spk}}$, $D_{\neq\text{spk}}$ and D_{ovlp} as described in Algorithm 1. An SNR value randomly selected within $[-5, 5]$ is assigned to utterances of each speaker before the segments are zero-padded and overlapped to form the *anechoic* single channel training samples.

Simulation of Reverberation: Reverberation was also introduced using FAST-RIR [36], which provides GPU accelerated GAN-based model to generate room impulse responses and convolved with dry clean source utterances to extend the simulated data to a more challenging reverberant scenario.

Given a session $\mathcal{U} = \{\mathcal{U}_1, \ldots, \mathcal{U}_k\}$ composed of source segments from k speakers, the reverberation time (T_{60}), room dimension (RD) and listener position (LP) are identical for all sources in the same session to form a consistent acoustic environment. Meanwhile, the source position (SP) for each speaker is slightly perturbed within the range of previously set RD to model the variation of positions of each source speaker. D_{RD} and $D_{T_{60}}$ indicate the distribution of room dimension and reverberation time. As described in Algorithm 2, each source is convolved with the room impulse response of an identical acoustic environment, but with various locations generated by FAST-RIR. This results in a reverberant session $\mathcal{U}'$. The *reverberant* mixture $\mathcal{M}'$ can be derived from $\mathcal{U}'$ and offsets of the original $\mathcal{M}$.

Libriheavy and LibriheavyMix: To simulate the LibriheavyMix dataset with a realistic distribution, corresponding statistics are obtained from the AliMeeting [24] dataset, which is a publicly available conference scenario dataset with human-annotated segmentation. Source utterances are from the Libriheavy [6] dataset, which is an ASR corpus for large-scale supervised training consisting of 50,000 hours of data. The Libriheavy dataset provides richer information for system construction such as punctuation, casing, and text context, which are also provided along with the speaker identity and corresponding timestamps for further investigation. During simulation, mixtures are generated by randomly selecting utterances for different speakers, each speaker is assigned with no longer than 15 seconds of the source utterances. Utterances with a duration longer than 15 seconds are first aligned using wav2vec2.0 [37] to obtain boundaries of sub-utterances for mixture simulation. Unlike the wsj0-mix [7], each utterance from the Libriheavy

Algorithm 2: Simulation of the reverberant session.

Data: $\mathcal{U} = \{\mathcal{U}_1, \ldots, \mathcal{U}_k\}$, D_{RD}, $D_{T_{60}}$
Result: $\mathcal{U}' = \{\mathcal{U}'_1, \ldots, \mathcal{U}'_k\}$

$RD_X, RD_Y, RD_Z \leftarrow \text{sample}(D_{RD})$
$LP_X, LP_Y, LP_Z \leftarrow \text{sample}(D_{RD})$
$T_{60} \leftarrow \text{sample}(D_{T_{60}})$
$\mathcal{X}' \leftarrow \varnothing$
for $i \leftarrow 1$ **to** k **do**
  $SP_X, SP_Y, SP_Z \leftarrow \text{sample}(D_{RD_{X,Y,Z}})$
  $\mathcal{U}'_i \leftarrow$ FAST-RIR($\mathcal{U}_i; RD_{X,Y,Z}; LP_{X,Y,Z}; SP_{X,Y,Z}; T_{60}$)
  $\mathcal{U}' \leftarrow \mathcal{U}' \cup \{\mathcal{U}'_i\}$
end

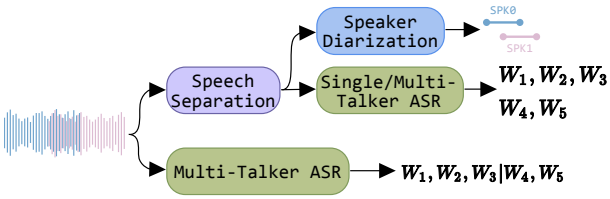

Figure 1: *An illustration of the crafted pipeline system.*

training set is used only once during the process of simulation, creating enough diversity in the simulated training data. The simulated dataset is provided with a *max* and *min* version, shorter sources in the *max* version are padded to the longest one, while mixtures in the *min* version were truncated to align with the source with shortest duration. This results in approximately 100 hours, 900 hours, and 9,000 hours of data in the *max* version of the small, medium, and large training set, against 45 hours of the wsj-mix dataset and 456 hours of the LibriMix dataset. Each training set of LibriheavyMix uniformly includes conversations involving 1-4 speakers, noted as *small*{*1-4*}*spk*, *medium*{*1-4*}*spk* and *large*{*1-4*}*spk*, respectively. For the dev and test sets, mixtures containing 2 to 4 speakers are derived from the dev, test-clean and test-other sets of the original Libriheavy corpus, noted as *dev*{*2-4*}*spk*, *test-clean*{*2-4*}*spk* and *test-other*{*2-4*}*spk*, respectively. The variety of speakers is much wider in LibriheavyMix's training set with around 6,000 distinct speakers in the *large* training set against 1,000 speakers in LibriMix and 100 speakers in wsj0-mix.

3. Baseline Systems

The pipeline system constructed involves a multi-talker ASR system, speech separation model and a diarization system as illustrated in Fig. 1.

3.1. Baseline for Multi-Talker ASR

For the multi-talker ASR baseline system, serialized output training (SOT) [14] based Conformer [2] Attention Encoder Decoder (AED) model is employed. Given input $\boldsymbol{X} = \{x_1, \cdots, x_T\}$, a single-speaker AED model produces the output sequence $\boldsymbol{Y} = \{y_1, \cdots, y_N\}$ as follows. Firstly, the encoder converts the input sequence $\boldsymbol{X}$ into a sequence embeddings by

$$\boldsymbol{H}^{enc} = \{h_1^{enc}, \cdots, h_T^{enc}\} = \text{Encoder}(\boldsymbol{X}) \quad (1)$$

Then, given the previous output $y_{[1:n-1]}$ and the encoder embeddings $\boldsymbol{H}^{enc}$, the output y_n is estimated by the autoregres-

Table 2: *Word error rate (%) of the AED model pre-trained on FAST-RIR augmented LibriSpeech 100-hour data on LibriheavyMix test sets of 2 to 4 speakers.*

Sys.	Training Set	Test Set # spkr	dry clean			reverb clean			reverb mixture		
			dev	test-clean	test-other	dev	test-clean	test-other	dev	test-clean	test-other
1	LS100 w. RIR	2	35.6	34.3	39.3	35.2	34.0	38.2	106.2	114.0	106.7
		3	35.3	31.1	38.7	33.8	29.6	37.8	121.6	122.3	121.6
		4	34.3	29.3	39.0	33.8	27.4	37.5	130.9	139.0	137.7

sive Transformer-based decoder.

$$y_n = \text{Decoder}\left(y_{[1:n-1]}, \boldsymbol{H}^{enc}\right) \tag{2}$$

To incorporate the SOT paradigm, a special symbol $\langle sc \rangle$ is inserted in the concatenation of multiple references to represent "speaker change" between each turn. Given a two-speaker conversation with speaker A and B, the reference word sequence will be given as $\boldsymbol{W} = \{\boldsymbol{w}_A^1, \cdots, \boldsymbol{w}_A^n, \langle sc \rangle, \boldsymbol{w}_B^1, \cdots, \boldsymbol{w}_B^m, \langle sc \rangle, \boldsymbol{w}_A^1, \cdots, \boldsymbol{w}_A^o\}$, where n, m and o represent number of tokens in the transcript of each utterance. Reference labels in the resulting concatenation $\boldsymbol{W}$ are sorted by their start times in a first-in, first-out fashion. In this way, the AED model learns to identify the turning point in a given utterance of multiple speakers marked by the special symbol $\langle sc \rangle$, thereby separating the transcript.

3.2. Baseline for Speech Separation

The Conv-TasNet [9] is selected as the baseline system for speech separation task. The model is a fully convolutional model specifically designed to separate individual speakers from a given mixed time-domain segment $\mathbf{x} \in \mathbb{R}^{1\times L}$, where L represents the number of samples of the given mixture. The network involves three stages: encoder, separation, and decoder. Encoder maps the segment to a high-dimensional representation $\boldsymbol{H}^{enc}$ using

$$\boldsymbol{H}^{enc} = \mathcal{H}(\mathbf{x}\mathbf{U}) \tag{3}$$

where $\mathbf{U} \in \mathbb{R}^{N\times L}$ represents 1-D convolution operations with N kernels, each of length L. This operation can be represented as a matrix multiplication. $\mathcal{H}(\cdot)$ is a rectified linear unit (ReLU) to ensure the non-negativity of $\boldsymbol{H}^{enc}$. The separation stage involves a series of 1-D convolution blocks of different dilation factors to estimate the masks for each of the target sources based on the encoder output. Estimated masks are multiplied with the encoded high-dimensional representation, a decoder further reconstructs the masked feature to waveforms using a 1-D transposed convolution operation as

$$\tilde{\mathbf{x}} = \tilde{\boldsymbol{H}}^{enc}\,\mathbf{V} \tag{4}$$

where $\tilde{\boldsymbol{H}}^{enc}$ and $\tilde{\mathbf{x}}$ stand for the masked feature and reconstructed waveforms, $\mathbf{V} \in \mathbb{R}^{N\times L}$ represents N kernels of the convolution operation, each with a dimension of L.

The model is trained end-to-end by minimizing the negative scale-invariant source-to-noise ratio (SI-SNR) loss $\mathcal{L}_{\text{SI-SNR}}$ given by

$$\mathcal{L}_{\text{SI-SNR}} = -10\log_{10}\frac{||\mathbf{s}_{target}||^2}{||\mathbf{e}_{noise}||^2} \tag{5}$$

in which $\mathbf{s}_{target}$, $\mathbf{e}_{noise}$ are obtained through $\mathbf{s}_{target} = \frac{\langle \hat{\mathbf{s}}, \mathbf{s}\rangle \mathbf{s}}{\langle \mathbf{s}, \mathbf{s}\rangle}$ and $\mathbf{e}_{noise} = \hat{\mathbf{s}} - \mathbf{s}_{target}$ given estimated sources $\hat{\mathbf{s}} \in \mathbb{R}^{1\times T}$ and original clean sources $\mathbf{s} \in \mathbb{R}^{1\times T}$. $\hat{\mathbf{s}}$ and $\mathbf{s}$ are normalized to zero-mean before loss calculation. To address the source permutation problem, utterance-level permutation invariant training (uPIT) [8] is incorporated during training.

Table 3: *Performance of the Serialized Output Training (SOT) [14] models on dev/test sets of 2 to 4 speakers. Systems are initialized using the single channel ASR model in Table 2 (Sys. 1) and trained on the small, medium and large in the "Training Set" column stand for the small{1-4}spk, medium{1-4}spk and large{1-4}spk training sets. cpWER represents the concatenated minimum-permutation word error rate [40].*

Sys.	Training Set	Test Set # spkr	cpWER (%) ↓			Spkr. Counting Acc. (%) ↑		
			dev	test-clean	test-other	dev	test-clean	test-other
1	*small*	2	57.8	59.5	61.0	45.19	42.56	46.84
		3	68.3	70.5	73.4	33.51	25.97	28.21
		4	76.0	76.3	79.5	25.22	25.04	26.85
2	*medium*	2	27.2	25.7	27.5	54.12	52.89	56.55
		3	35.8	35.8	40.4	41.04	33.20	38.16
		4	52.3	48.3	54.5	22.02	20.66	21.39
3	*large*	2	21.0	19.0	21.7	55.48	54.18	56.25
		3	28.8	27.7	31.7	41.48	37.84	41.01
		4	40.4	38.9	43.3	22.63	20.59	21.62

3.3. Baseline for Speaker Diarization

Pre-trained pyannote.audio 3.1 system [1] [38, 39] is applied as the baseline system for speaker diarization experiments. The system first utilizes a neural speaker segmentation model, incorporating a sliding window to obtain local speaker segmentation. Local speaker embeddings are then extracted from each window, and classical agglomerative hierarchical clustering with centroid linkage is then applied to the extracted embeddings. The final aggregating step produces the actual speaker diarization results on top of the clustered local speaker segmentation.

4. Experiments and Results

4.1. Performance of Multi-Talker ASR Baseline

The recipe for serialized output training (SOT) [14] is modified from the LibriMix recipe of the ESPnet [41] toolkit. Trained Conformer models are of 12 encoder blocks and 6 Transformer decoder blocks, with a total of 43 million parameters. The dimension of feed forward layers in both encoder and decoder blocks is set to 2048 with 4 attention heads, each attention head has a dimension of 256, kernel size of all convolutional layers is set to 31[2]. To help convergence, systems trained were initialized using a Conformer model with an identical setup pre-trained on LibriSpeech [4] 100-hour training set augmented using FAST-RIR. Performance of the pre-trained system is presented in Tab. 2, Sys. 1. The training data includes all available mixtures, the transcript of which is obtained by concatenating the transcript of each source according to its starting time in a "first-in, first-out" fashion. SpecAugment [42] is incorporated for all systems. Speed perturbation [43] is further applied except for the ones with *large{1-4}spk* training set involved. The evaluation metric for all results obtained from SOT systems is the concatenated minimum-permutation word error rate (cpWER) [40]. This metric is calculated by first concatenating all utterances of each speaker for both the reference and hypothesis files. Then, the permutation of speakers that yields the lowest word error rate when compared to the reference is picked.

Sys. 1-3 (Tab. 3) were trained on the small, medium, and large training sets of the proposed LibriheavyMix corpus re-

[1] `https://huggingface.co/pyannote/speaker-diarization-3.1/`

[2] More details can be found at `egs2/librimix/sot_asr1/conf/tuning/train_sot_asr_conformer.yaml` of the ESPnet toolkit [41].

Table 4: *Performance of the Conv-TasNet [9] models on the LibriheavyMix and WHAMR! dataset. "tt" stands for the test set of WHAMR!. "dev", "test-clean" and "test-other" denote the dev2spk, test-clean2spk and test-other2spk of LibriheavyMix.*

Sys.	Training Set				SI-SDR ↑				Δ SI-SDR ↑			
	LibriheavyMix (*2spk*) small	medium	large	WHAMR!	tt	dev	test-clean	test-other	tt	dev	test-clean	test-other
1	-			✓	9.36	1.36	1.99	1.51	9.36	1.95	2.04	1.65
2	✓			-	5.11	6.41	7.83	6.14	5.11	7.08	7.88	6.28
3		✓		-	8.19	9.27	11.33	10.11	8.20	9.86	11.37	10.25
4			✓	-	9.23	10.70	12.94	11.54	9.23	11.55	12.90	11.52
5	✓			✓	10.02	7.24	9.19	7.53	10.02	7.83	9.24	7.67
6		✓		✓	10.49	9.75	12.06	10.58	10.49	10.33	12.11	10.72
7			✓	✓	10.33	10.66	12.81	11.35	10.34	11.24	12.87	11.49

Table 5: *Performance of the SOT [14] models on the WHAMR! dataset. Other naming conventions follow the one in Table 4. Note that only performance on* **2-speaker** *test sets are presented in this table for simplicity.*

Sys.	Training Set				cpWER (%) ↓				Spkr. Counting Acc. (%) ↑			
	LibriheavyMix small	medium	large	WHAMR!	tt	dev	test-clean	test-other	tt	dev	test-clean	test-other
1	-			✓	61.4	85.8	86.5	87.4	99.20	45.46	44.05	46.49
2	✓			-	76.1	57.8	59.5	61.0	70.60	45.19	42.56	46.84
3		✓		-	43.9	27.2	25.7	27.5	54.80	54.12	52.89	56.55
4			✓	-	28.1	21.0	19.0	21.7	77.30	55.48	54.18	56.25
5	✓			✓	23.9	45.5	45.5	49.4	99.30	50.80	51.12	53.94
6		✓		✓	15.1	23.6	22.9	24.5	99.70	55.41	55.25	58.31
7			✓	✓	13.6	21.0	19.6	21.4	99.40	59.73	58.14	59.79

spectively. All training data containing 1 to 4 speakers were involved to evaluate the capability of SOT systems on generalizing to mixtures of various numbers of speakers. Results suggest that scaling up the amount of training data demonstrates a significant reduction on cpWER and speaker counting accuracy. Sys. 2 consistently outperforms Sys. 1 on all test sets, while a similar trend can also be observed between Sys. 3 and 2 except for a slight performance degradation on speaker counting accuracy is obtained on the most challenging 4-speaker scenario.

4.2. Performance of Speech Separation Baseline

The Conv-TasNet [9] model trained has 8.98 million parameters. The encoder contains 512 filters, the length of each filter is set to 40, bottleneck dimension is set to 256. The repeat number is set to 4, each repeat contains 8 convolutional blocks with kernel size set to 3 and number of channels set to 512. Global layer normalization and ReLU are adopted for normalization and non-linearity respectively. Training is done by minimizing the negative permutation-invariant, SI-SNR loss on 4-second segments. All systems were trained with identical parameters. Since the SI-SDR is not defined for silent sources, all results reported were trained on the 8kHz *min* version of the training sets and evaluated on the *max* version of test sets.

The performance of the Conv-TasNet model on LibriheavyMix dataset is presented in Tab. 4, Sys. 2-4. All models were trained and evaluated on the 2-speaker sets of LibriheavyMix. Results suggest that scaling up the training data demonstrates significant performance improvements on all test sets, as Sys. 4 trained on approx. 7,000-hour *large2spk* consistently outperforms Sys. 3 trained on approx. 500-hour *medium2spk* set. A similar trend is also obtained on Sys. 3 and Sys. 2 trained on the approx. 70-hour *small2spk* set.

4.3. Generalization on the WHAMR! dataset

The WHAMR! [20] dataset is a public benchmark built upon wsj0-2mix [7] and WHAM! [18] for the task of overlapped speech separation and recognition under reverberant and noisy conditions. It serves as one of the publicly available benchmarks for speech separation and recognition under reverberant and overlapping conditions. Further experiments were conducted to evaluate the generalizability of the proposed LibriheavyMix dataset. Note that all experiments involving WHAMR! use the *clean_reverb* data to match the acoustic environment of LibriheavyMix.

The performance of Conv-TasNet models is presented in Tab. 4. All models were trained on *min* version of the WHAMR! and LibriheavyMix and evaluated on *max* test sets of the corresponding corpora. Sys. 1 suggests that the model trained on WHAMR! performs not as well on LibriheavyMix as it demonstrates on WHAMR!. This observation aligns with the previous study [21] indicating that the Conv-TasNet model trained on wsj0-2mix demonstrates poor generalization on other datasets. Sys. 2-4 suggest that by introducing more diversity into the training data and scaling up the amount of training data, the Conv-TasNet model achieved a significant improvement in generalization even on the unseen WHAMR! data. The performance on both LibriheavyMix and WHAMR! can be further boosted when incorporating training data from WHAMR! dataset, as Sys. 5-7 consistently outperform Sys. 2-4 on all test sets involved.

The performance of the SOT models is presented in Tab. 5. Results suggest that scaling the amount of training data demonstrates a significant WER reduction especially on the most complicated 4-speaker scenario. Although a similar trend is also observed in terms of the speaker counting accuracy, it is still challenging especially when multi turn conversations are presented in test sets.

Table 6: *Performance of the pyannote.audio diarization system and cascaded systems on LibriheavyMix test sets.*

Sys.	Test Set # spkr	Diarization Error Rate (%) ↓ dev	test-clean	test-other
1	2	30.90	31.72	28.20
	3	42.13	41.96	40.17
	4	50.27	48.49	47.42
2 (Sys. 4, Tab. 4 → Sys. 1, Tab. 6)	2	19.68	21.20	19.47
3 (Sys. 7, Tab. 4 → Sys. 1, Tab. 6)	2	19.39	21.03	19.40
Sys.	**Test Set # spkr**	**Word Error Rate (%) ↓ dev**	**test-clean**	**test-other**
4 (Sys. 4, Tab. 4 → Sys. 4, Tab. 5)	2	44.9	42.5	47.1
5 (Sys. 7, Tab. 4 → Sys. 7, Tab. 5)	2	43.4	41.0	45.9

4.4. Performance of Speaker Diarization Baseline

For simplicity, speaker diarization was directly performed on the dev and test sets of LibriheavyMix using pre-trained pyannote.audio 3.1 system. Performance of the pyannote.audio system is presented in Tab. 6. The speech separation module demonstrates its effectiveness by delivering a remarkable absolute DER reduction of up to 11.51% to the diarization system.

5. Conclusions

This work releases a large-scale (20,000 hours) synthesized corpus [3]. for overlapped speech separation and recognition under reverberant conditions. A series of baseline systems are constructed to evaluate the performance of the proposed dataset. Further evaluation using a public benchmark for far-field overlapped speech separation and recognition validates the effectiveness and generalizability of the proposed dataset.

[3] `https://huggingface.co/datasets/zrjin/LibriheavyMix-{dev,test,small,medium,large}`

6. References

[1] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao *et al.*, "Transformer-Based Acoustic Modeling for Hybrid Speech Recognition," *IEEE ICASSP*, 2020.

[2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *INTERSPEECH*, 2020.

[3] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang *et al.*, "Zipformer: A Faster and Better Encoder for Automatic Speech Recognition," *ICLR*, 2024.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," in *IEEE ICASSP*, 2015.

[5] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AIShell-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline," in *Oriental COCOSDA*, 2017.

[6] W. Kang, X. Yang, Z. Yao, F. Kuang, Y. Yang *et al.*, "Libriheavy: A 50,000 Hours ASR Corpus with Punctuation Casing and Context," in *IEEE ICASSP*, 2024.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *IEEE ICASSP*, 2016.

[8] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation," in *IEEE ICASSP*, 2017.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM TASLP*, 2019.

[10] S. Zhao, Y. Ma, C. Ni, C. Zhang, H. Wang *et al.*, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," *CoRR*, vol. abs/2312.11825, 2023.

[11] D. Yu, X. Chang, and Y. Qian, "Recognizing Multi-Talker Speech with Permutation Invariant Training," in *INTERSPEECH*, 2017.

[12] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "MIMO-Speech: End-to-End Multi-Channel Multi-Speaker Speech Recognition," in *IEEE ASRU*, 2019.

[13] W. Zhang, X. Chang, Y. Qian, and S. Watanabe, "Improving End-to-End Single-Channel Multi-Talker Speech Recognition," *IEEE/ACM TASLP*, 2020.

[14] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized Output Training for End-to-End Overlapped Speech Recognition," in *INTERSPEECH*, 2020.

[15] N. Kanda, J. Wu, Y. Wu, X. Xiao, Z. Meng *et al.*, "Streaming Multi-Talker ASR with Token-Level Serialized Output Training," in *INTERSPEECH*, 2022.

[16] L. Meng, J. Kang, M. Cui, Y. Wang, X. Wu *et al.*, "A Sidecar Separator Can Convert A Single-Talker Speech Recognition System to A Multi-Talker One," in *IEEE ICASSP*, 2023.

[17] L. Meng, J. Kang, M. Cui, H. Wu, X. Wu *et al.*, "Unified Modeling of Multi-Talker Overlapped Speech Recognition and Diarization with a Sidecar Separator," in *INTERSPEECH*, 2023.

[18] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn *et al.*, "WHAM!: Extending Speech Separation to Noisy Environments," in *INTERSPEECH*, 2019.

[19] 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.

[20] M. Maciejewski, G. Wichern, and J. Le Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," in *IEEE ICASSP*, 2020.

[21] B. Kadıoğlu, M. Horgan, X. Liu, J. Pons, D. Darcy *et al.*, "An Empirical Study of Conv-TasNet," in *IEEE ICASSP*, 2020.

[22] N. Kanda, G. Ye, Y. Wu, Y. Gaur, X. Wang *et al.*, "Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone," in *INTERSPEECH*, 2021.

[23] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot *et al.*, "The AMI Meeting Corpus: A Pre-announcement," in *MLMI*, ser. Lecture Notes in Computer Science, 2005.

[24] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng *et al.*, "M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge," in *IEEE ICASSP*, 2022.

[25] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe *et al.*, "A Review of Speaker Diarization: Recent Advances with Deep Learning," *Computer Speech & Language*, 2022.

[26] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-End Neural Speaker Diarization with Permutation-Free Objectives," in *INTERSPEECH*, 2019.

[27] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Y. Khokhlov, M. Korenevskaya *et al.*, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *INTERSPEECH*, 2020.

[28] M. He, D. Raj, Z. Huang, J. Du, Z. Chen *et al.*, "Target-Speaker Voice Activity Detection with Improved i-Vector Estimation for Unknown Number of Speaker," in *INTERSPEECH*, 2021.

[29] N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng *et al.*, "End-to-End Speaker-Attributed ASR with Transformer," in *INTERSPEECH*, 2021.

[30] F. Yu, Z. Du, S. Zhang, Y. Lin, and L. Xie, "A Comparative Study on Speaker-attributed Automatic Speech Recognition in Multi-party Meetings," in *INTERSPEECH*, 2022.

[31] M. Shi, Z. Du, Q. Chen, F. Yu, Y. Li *et al.*, "CASA-ASR: Context-Aware Speaker-Attributed ASR," in *INTERSPEECH*, 2023.

[32] M. Shi, J. Zhang, Z. Du, F. Yu, Q. Chen *et al.*, "A Comparative Study on Multichannel Speaker-Attributed Automatic Speech Recognition in Multi-party Meetings," in *IEEE APSIPA ASC*, 2023.

[33] S. Bijwadia, S. Chang, B. Li, T. N. Sainath, C. Zhang *et al.*, "Unified End-to-End Speech Recognition and Endpointing for Fast and Efficient Speech Systems," in *IEEE SLT*, 2022.

[34] M. Shi, Y. Shu, L. Zuo, Q. Chen, S. Zhang *et al.*, "Semantic VAD: Low-Latency Voice Activity Detection for Speech Interaction," in *INTERSPEECH*, 2023.

[35] F. Landini, A. Lozano-Diez, M. Diez, and L. Burget, "From Simulated Mixtures to Simulated Conversations as Training Data for End-to-End Neural Diarization," in *INTERSPEECH*, 2022.

[36] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha *et al.*, "FAST-RIR: Fast Neural Diffuse Room Impulse Response Generator," in *IEEE ICASSP*, 2022.

[37] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[38] H. Bredin, "Pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe," in *INTERSPEECH*, 2023.

[39] A. Plaquet and H. Bredin, "Powerset Multi-Class Cross Entropy Loss for Neural Speaker Diarization," in *INTERSPEECH*, 2023.

[40] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora *et al.*, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[41] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *INTERSPEECH*, 2018.

[42] D. S. Park, W. Chan, Y. Zhang *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *INTERSPEECH*, 2019.

[43] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *INTERSPEECH*, 2015.