Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays

Shanzheng Guan, Shupei Liu, Junqi Chen, Wenbo Zhu, Shengqiang Li, Xu Tan, Ziye Yang, Menglong Xu, Yijiang Chen, Chengdong Liang, Jianyu Wang and Xiao-Lei Zhang
CIAIC, School of Marine Science and Technology, Northwestern Polytechnical University, China
E-mail: gshanzheng@mail.nwpu.edu.cn, xiaolei.zhang@nwpu.edu.cn

***Abstract*—Recently, there is a research trend on ad-hoc microphone arrays. However, most research was conducted on simulated data. Although some datasets were collected with a small number of distributed devices, they were not synchronized which hinders the fundamental theoretical research on ad-hoc microphone arrays. To address this issue, this paper presents a synchronized speech corpus, named *Libri-adhoc40*, which collects the replayed Librispeech data from loudspeakers by ad-hoc microphone arrays of 40 strongly synchronized distributed nodes in a real office environment. Besides, to provide the evaluation target for speech frontend processing and other applications, we also recorded the replayed speech in an anechoic chamber. We trained several multi-device speech recognition systems on both the Libri-adhoc40 dataset and a simulated dataset. Experimental results demonstrate the validity of the proposed corpus which can be used as a benchmark to reflect the trend and difference of the models with different ad-hoc microphone arrays. The dataset is online available at https://github.com/ISmallFish/Libri-adhoc40.**

I. Introduction

Deep learning based speech processing has made significant progress. However, the progress was mostly made with single-channel front-ends or multichannel front-ends on single devices [1]. As we know, the performance of speech processing degrades significantly when the distance between the speech source and the microphone array receiver increases, which is known as the far-field speech processing problem. Fortunately, ad-hoc microphone array can significantly reduce the occurrence probability of far-field pickup scenes [2]. It is a set of distributed microphones collaborating with each other [3]. Conventional methods try to organize the microphones in a blind way, which faces many challenges.

Recently, deep learning has been introduced to the study of ad-hoc microphone arrays [2], [4]–[9], which provides a promising solution to the challenges. In [2], [4], a supervised channel selection strategy based on deep learning was proposed to group the distributed microphones with high signal-to-noise ratios (SNR) into a local microphone array. However, the aforementioned studies were conducted on simulated data only.

Some work was conducted on real-world data. In [6], the authors first conducted single-channel speech separation on a selected reference microphone, and then estimated a beamforming filter for all remaining microphones based on the output of the reference microphone. In [7], a novel neural network architecture was proposed to capture both the inter-channel and temporal correlations from the multi-channel input of ad-hoc microphone arrays. [8] designed a speech recognition system which first makes all channels share the same encoder and then fuses all channels via stream attention. [9] proposed a speaker recognition system based on ad-hoc microphone arrays. It first trains a single-channel speaker recognition system, then applies it to each channel, and finally fuses the outputs of the channels for the final decision. However, their experimental data was recorded with few devices only.

There are already some corpora collected with ad-hoc microphone arrays [9]–[14]. However, most of them were collected with a small number of distributed devices too. In [9], a speaker verification dataset called HI-MIA was collected with 7 recording devices for both training and test. In [10], the CHiME-5 dataset employed 6 Kinect microphone arrays and 4 binaural microphone pairs to record natural conversational speech. To our knowledge, the ad-hoc microphone array in the Massive Distributed Microphone Array dataset [13], which consists of 4 wearable arrays and 12 tabletop arrays, is the largest array that has ever been used for recording publicly available data. As we known, when the ad-hoc nodes are too few, it is difficult to fully explore the potential of ad-hoc microphone arrays.

Moreover, none of the above corpora were collected with synchronized devices. Because the hardware and software processing pipelines between devices are different, the collected data may have significant variations [15], [16]. In [11], the CHiME-6 dataset synchronized the ad-hoc recordings of CHiME-5 via frame-dropping and clock-drift compensation. However, the synchronization technique misses the signal propagation delay information between devices. Although the SINS dataset [14], which adopted 13 sensor nodes to collect data, recorded the timestamps of each node, the timestamps can only provide rough synchronization between the sensor nodes. Post-processing algorithms are still needed if rigorous synchronization is required.

To fascinate the fundamental research of ad-hoc microphone arrays on how well it can improve the performance ideally, we need to collect synchronized data from large ad-hoc microphone arrays, leaving the device synchronization problem as a separate topic at the current research stage. To address this issue, in this paper, we create a dataset, named *Libri-adhoc40*, which collects the replayed Librispeech data

978-988-14768-9-0/21/$31.00 ©2021 APSIPA

[17] from loudspeakers by ad-hoc microphone arrays of 40 synchronized distributed microphones, where the 'train-clean-100', 'dev-clean' and 'test-clean' subsets of Librispeech were used as the speech source. To provide the evaluation target for speech frontend processing and other applications, we also recorded the replayed speech in an anechoic chamber. Eventually, Libri-adhoc40 contains 4510 hours data in total with 110 hours data per microphone. We conducted a speech recognition evaluation on the test set of Libri-adhoc40, where both the simulated data and the training set of Libri-adhoc40 were used for model training. Experimental results demonstrate the validity of Libri-adhoc40.

The rest of this paper is organized as follows. We first provide an overview of the dataset and its recording method in Sections 2 and 3 respectively, then conduct a baseline evaluation in Section 4, and finally conclude in Section 5.

II. Description of Libri-adhoc40

The Librispeech corpus [17] is derived from audiobooks that are part of the LibriVox project. It contains 1000 hours of clean speech with a sampling rate of 16 kHz. The gender and per-speaker duration are reasonably balanced. The Libri-adhoc40 dataset takes the 'train-clean-100', 'dev-clean', and 'test-clean' subsets of Librispeech as the clean speech source, which contains about 110 hours of US English speech from 331 speakers.

A. Recording environment

We replayed the subsets of Librispeech in an office room and an anechoic chamber individually which are described as follows:

- *Office room*: The plane structure of the office room is shown in Figure 1. The height of the room is 4.2 m. Because the room size is large, and because the floor is laid with smooth tiles, the room is highly reverberant with the T_{60} around 900 ms. Because the room is far from noisy environments, the recorded speech has little additive noise. A directional loudspeaker and 40 omnidirectional microphones of the same type were placed in the room. The sampling rate is 16 kHz.
- *Anechoic chamber*: The size of the net space of the anechoic chamber is $11.8\times4.2\times3.8$ m after the installation of sound-absorbing materials. The same loudspeaker and a handy recorder were placed in the anechoic chamber. The speech was recorded at 48 kHz, and further downsampled to 16 kHz.

B. Training data

As shown in Figure 1(a), the loudspeaker was placed at 9 positions with 10 orientations, where the loudspeaker at 'pos 9' has 2 opposite orientations. The distances between the loudspeaker and the microphones are ranged from 0.8 m to 7.4 m. The speech source is the 'train-clean-100' corpus of Librispeech, which contains 251 speakers. We replayed the corpus with about 20 to 40 speakers per position. A detailed configuration, including the coordinates of the loudspeaker and microphones, as well as the relationship between the speaker identities and the positions, are described at https://github.com/ISmallFish/Libri-adhoc40.

TABLE I: Recording equipment.

Device	Product model	Quantity
Microphone	Superlux ECM 999	40
Preamplifier	Focusrite Scarlett Octopre 8	4
Sound card	RME Fireface UFX II	2
Handy recorder	Zoom H1N	1
Loudspeaker	JBL One Series 104	2

C. Development and test data

As shown in Figure 1(b), the loudspeaker was placed at 8 positions. The distance between the loudspeaker and the microphones ranges from 0.8 m to 7.4 m as well. The positions of the loudspeaker and 40 microphones for preparing the development and test data are different from those for preparing the training data, which is designed for evaluating the generalization ability of speech processing algorithms on different array patterns. The speech sources for development and test are the 'dev-clean' and 'test-clean' corpora of Librispeech respectively, each of which contains 40 speakers. We replayed the corpus with 10 speakers per position.

D. Ground-truth clean speech

The loudspeaker may introduce an unwanted mismatch between the original recordings and the output of the loudspeaker, so we replayed the clean speech of Librispeech in the anechoic chamber to provide the ground-truth clean speech of Libri-adhoc40. The distance between the loudspeaker and the recording device was 40 cm. The sound volume of the loudspeaker was set the same as that in the office room.

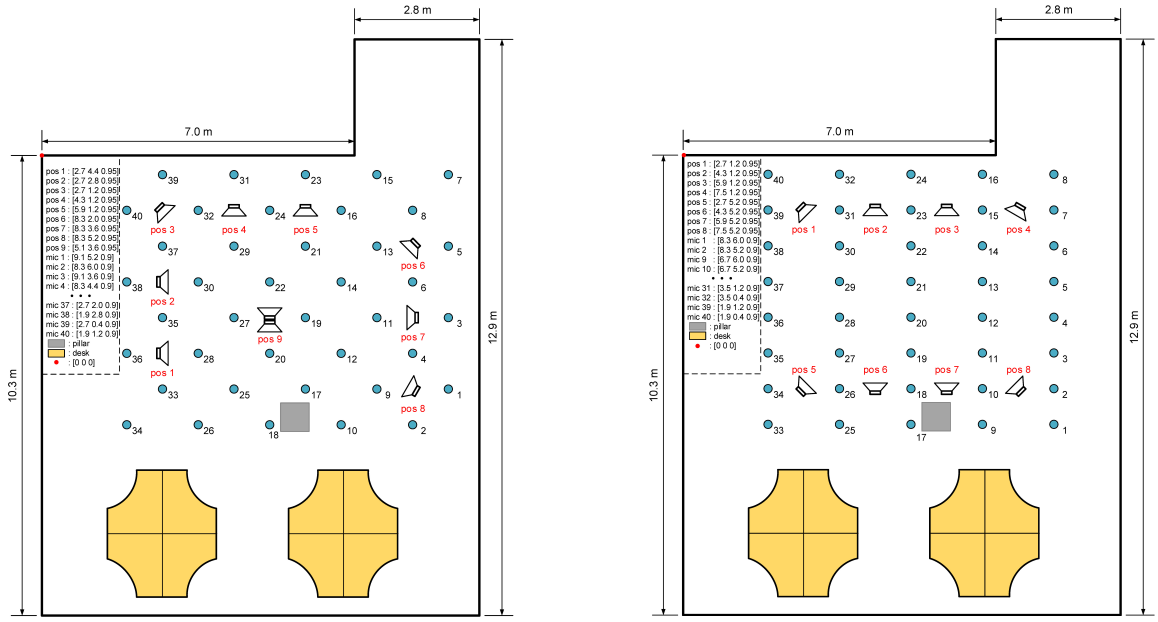
III. Methodology

A. Recording equipment

The equipment for recording the data is listed in Table I. 'JBL One Series 104' was selected as the loudspeaker. In this loudspeaker, a high-frequency driver aligned with a precisely contoured woofer cone is used to deliver accurate response. Because its treble and bass units are tightly arranged together, this design makes the loudspeaker behave like a point source.

To reduce the difference of the distortion between devices and eliminate device asynchronization, 40 'Superlux ECM 999' microphones and two 'RME Fireface UFX II' sound cards were pre-synchronized.

Specifically, each sound card was connected with 20 microphones separately, where 4 microphones were connected to the sound card directly, and the other 16 microphones were connected indirectly through 2 'Focusrite Scarlett Octopre' microphone preamplifiers, each of which connects 8 microphones. After a careful evaluation, we found that the time delay caused by the preamplifiers could be neglected. To reduce the difference in gains between the two sets of channels, the gains of the sound cards and preamplifiers were



(a) Training set (b) Development and test sets

Fig. 1: Recording environment and setting of Libri-adhoc40. The red dot indicates the origin of the reference axes. The blue dots indicate the positions of the microphones, whose coordinates are listed in the upper-left corner. The positions and orientations of the loudspeaker are marked by loudspeaker icons. The terms 'pos' is short for position. The term 'mic' is short for microphone.

adjusted in advance. A 'Zoom H1N' handy recorder was employed to record anechoic signals at a sampling rate of 48 kHz. The recordings were further downsampled to 16 kHz manually.

B. Recording process

We played back Librispeech in a streaming fashion, where the sentences from the same speaker were concatenated into a sequence and played back continuously. A picture of the real recording environment is shown in Figure 2.

C. Postprocessing

The two independent sound cards introduce a device asynchronization problem into the two sets of the microphones. It was mainly caused by (i) the asynchronization of the recording start time and (ii) the random drop of the sample points. To compensate the start time difference, we conducted a time delay estimation by playing white noise before the recording, which makes us possible in inferring the time delay difference.

Although the sample drop happened occasionally, the accumulation of the negative effect cannot be neglected if we played a long sequence continuously. To compensate the sample drop caused by the two independent sound cards, we first carefully selected one microphone per sound card and then calculated the time difference of arrival between the two microphones for each position of the loudspeaker, before the data recording. Finally, if the time delay difference of the recorded data changed at some point-in-time, we compensated the detected sample drop at the time.

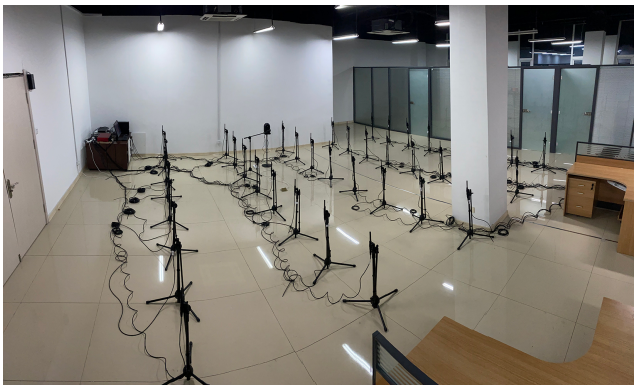


Fig. 2: A picture of the recording environment for replaying the 'dev-clean' and 'test-clean' corpora. The loudspeaker was placed at *pos 2*. The microphones 1 to 20 were connected to one sound card, while the microphones 21 to 40 were connected to the other sound card.

At last, we partitioned the recorded continuous speech according to the original segmentation lengths of the Librispeech utterances. Each partitioned segment was saved with the same name of its corresponding Librispeech utterance in a subdirectory that has the same name with Librispeech.

This strict synchronization setting makes the dataset gener-

alizable for simulating device asynchronized situations by, e.g., performing bandpass filtering, waveform amplitude clipping, and delay perturbation operations [18] to the data.

IV. Experiments

In this section, we evaluate the validity of Libri-adhoc40 in an automatic speech recognition (ASR) task with ad-hoc microphone arrays.

A. Datasets

To evaluate the performance of the ASR with ad-hoc microphone arrays, we simulated a similar dataset with Libri-adhoc40, named *Libri-adhoc40-simu*. Because all ASR systems in evaluation were tested on the test set of Libri-adhoc40, Libri-adhoc40-simu consists of only a training set and a development set, which were generated from the 'train-clean-100' and 'dev-clean' corpora of Librispeech respectively. The simulation environment is described as follows.

To roughly match the recording environments of Libri-adhoc40, we simulated a room with a size of $10\times10\times4$ m. Forty simulated microphones for both training and development were placed at the same locations as Libri-adhoc40. For each utterance, a simulated loudspeaker for playing back the utterance was placed *randomly* in the room with its position located in the covering range of the ad-hoc microphone arrays and at least 0.6 meter away from the microphones; the room impulse response was generated by an image source model [19], where the T_{60} was sampled from a Gaussian distribution with a mean value of 0.7 second, a standard deviation of 0.1 second, a lower bound of 0.5 second, and an upper bound of 1.2 second.

We constructed three test scenarios. The first two scenarios randomly select 10 and 25 channels respectively for each test utterance. The third scenario uses all 40 channels for evaluation.

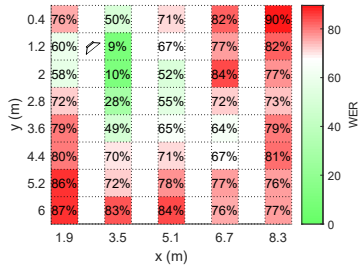
B. ASR systems

We used a single-channel conformer based automatic speech recognition (ASR) system [20] and a multichannel ASR based on the Scaling Sparsemax stream attention [21] as the ASR systems, which are described as follows:

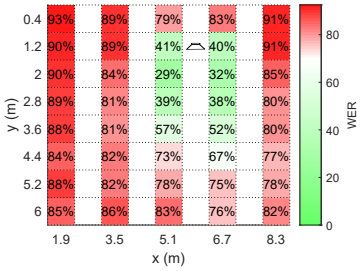
Single-channel conformer (oracle one-best): We trained the single-channel ASR with the clean speech of the original Librispeech corpora directly. In the test stage, we picked the channel that was physically closest to and also faced by the loudspeaker as the input of the single channel ASR system.

Scaling Sparsemax stream attention based multichannel ASR (Scaling Sparsemax): We first used the single-channel ASR trained with the clean speech of Librispeech as the initialization of the multichannel ASR. Then, we trained the stream attention module of the multichannel ASR with 20 randomly selected channels per utterance. If the training utterances were from Libri-adhoc40-simu, the training condition is denoted as *simu train*. If the training utterances were from Libri-adhoc40, the training condition is denoted as *real train*.

We conducted the evaluation on the test data of Libri-adhoc40 in terms of word error rate (WER).



(a) Loudspeaker placed at pos 1



(b) Loudspeaker placed at pos 3

Fig. 3: Visualization of the WER (%) results of the single-channel conformer-based ASR system on the test data of Libri-adhoc40.

C. Results

Figure 3 shows the average WER results of the single-channel conformer-based ASR system on each channel of the test set at two positions. From the figure, we see that (i) the average WER at the closest channel is 9% in Figure 3(a) and 29% in Figure 3(b); (ii) the WERs of the channels that the loudspeaker faces to in Figure 3(a) is significantly lower than those in Figure 3(b). The phenomena indicate that the performance was not only affected by the distance between the speaker and the microphone, but also affected by the orientation of the speaker.

Table II lists the comparison results when the ad-hoc microphone array contains 10, 25, and 40 channels respectively. From the table, we can see that the models, no matter trained on simulated data or semi-real data, can be used on the semi-real test data of the proposed Libri-adhoc40. The systems in the *real train* condition perform better than those in the *simu train* condition. When there is no microphone in the orientation of the loudspeaker, such as 'pos2' and 'pos3', all methods behave poorly.

Besides the general phenomena, the results can also reflect the trend and difference of the models with different ad-hoc microphone arrays. Specifically, (i) as the number of channels increases, the performance of the ASR systems is gradually improved. For example, the WER of the Scaling Sparsemax in the *real train* condition is reduced by 46.6% relatively when the channel number is increased from 10 to 40, which demonstrates the importance of increasing the number of the channels. (ii) When the channel number is 40, the Scaling Sparsemax achieves a relative WER reduction

TABLE II: Comparison results (in WER (%)) on the test set of Libri-adhoc40. The term 'Pos#' means that the test data is a subset of Libri-adhoc40 test data where the loudspeaker was placed at *pos #* described in Figure 1(b).

Method	Training condition	Pos1	Pos2	Pos3	Pos4	AVG
10 channels						
Oracle	Librispeech	32.5	46.2	43.6	39.4	40.4
Scaling	*simu train*	28.6	43.5	36.3	35.8	36.1
Sparsemax	*real train*	**25.7**	**38.5**	**33**	**31.7**	**32.2**
25 channels						
Oracle	Librispeech	12.4	33	32.4	16.9	23.6
Scaling	*simu train*	12.5	32.9	27.4	17.4	22.5
Sparsemax	*real train*	**12.3**	**27.9**	**24.4**	**16.5**	**20.3**
40 channels						
Oracle	Librispeech	9.1	31.7	32.2	12.6	21.4
Scaling	*simu train*	**8.7**	29.1	24.9	12.9	18.9
Sparsemax	*real train*	9.4	**25.3**	**21.8**	**12.3**	**17.2**

of 19.6% lower than the oracle one-best in the *real train* condition, which demonstrates the importance of channel selection. (iii) Although the orientation of the loudspeaker affects the performance significantly, Scaling Sparsemax reduces the negative effect. For example, when the channel number is 40, the average WER of the oracle one-best at 'pos2' and 'pos3' is increased by about 66% over that at 'pos1' and 'pos4', while the relative WER increase of Scaling Sparsemax in the *real train* condition is only 54%, which demonstrates the merit of ad-hoc microphone arrays.

To summarize, the above phenomena demonstrate the effectiveness of Libri-adhoc40 as an evaluation benchmark.

V. Conclusions and discussion

This paper presents a semi-real dataset recorded by synchronized ad-hoc microphone arrays, named Libri-adhoc40. Its validity has been evaluated in the speech recognition task. It facilitates the study and development of speech processing algorithms based on ad-hoc microphone arrays.

The dataset can be used as a benchmark corpus of many speech processing tasks beyond speech recognition, including speech enhancement, dereverberation, given that the anechoic recordings are provided in Libri-adhoc40. It can also be used for speech separation by mixing the speech signals at different positions of the loudspeakers, since the loudspeakers at different positions replay different speakers. As for speaker recognition, we may take the entire set of 331 speakers for evaluation. As for the research on the synchronization techniques of devices, we may also construct asychronized test environments by adding various interruptions to the channels.

References

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

[2] X.-L. Zhang, "Deep ad-hoc beamforming," *Computer Speech and Language*, vol. 68, p. 101201, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000085

[3] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, 2004.

[4] Z. Yang, S. Guan, and X.-L. Zhang, "Deep ad-hoc beamforming based on speaker extraction for target-dependent speech separation," *arXiv preprint arXiv:2012.00403*, 2020.

[5] R. Li, G. Sell, X. Wang, S. Watanabe, and H. Hermansky, "A practical two-stage training strategy for multi-stream end-to-end speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7014–7018.

[6] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 260–267.

[7] D. Wang, Z. Chen, and T. Yoshioka, "Neural speech separation using spatially distributed microphones," *arXiv preprint arXiv:2004.13670*, 2020.

[8] R. Li, X. Wang, S. H. Mallidi, S. Watanabe, T. Hori, and H. Hermansky, "Multi-stream end-to-end speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 646–655, 2020.

[9] X. Qin, H. Bu, and M. Li, "Hi-mia: A far-field text-dependent speaker verification database and the baselines," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7609–7613.

[10] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth'chime'speech separation and recognition challenge: dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[11] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[12] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The dirha-english corpus and related tasks for distant-speech recognition in domestic environments," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 275–282.

[13] R. Corey, M. Skarha, and A. Singer, "Massive distributed microphone array dataset," 2019.

[14] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. Van den Bergh, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The sins database for detection of daily activities in a home environment using an acoustic sensor network," *Detection and Classification of Acoustic Scenes and Events 2017*, 2017.

[15] A. Mathur, F. Kawsar, N. Berthouze, and N. D. Lane, "Libri-adapt: a new speech dataset for unsupervised domain adaptation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7439–7443.

[16] A. Mathur, A. Isopoussu, F. Kawsar, N. Berthouze, and N. D. Lane, "Mic2mic: Using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems," in *2019 18th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, 2019, pp. 169–180.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] D. Wang, T. Yoshioka, Z. Chen, X. Wang, T. Zhou, and Z. Meng, "Continuous speech separation with ad hoc microphone arrays," *arXiv preprint arXiv:2103.02378*, 2021.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[21] J. Chen and X.-L. Zhang, "Scaling sparsemax based channel selection for speech recognition with ad-hoc microphone arrays," *arXiv preprint arXiv:2103.15305*, 2021.