LibriWASN: A Data Set for Meeting Separation, Diarization, and Recognition with Asynchronous Recording Devices

Joerg Schmalenstroeer, Tobias Gburrek, Reinhold Haeb-Umbach

Paderborn University, Department of Communications Engineering
Email: {schmalen, gburrek, haeb}@nt.uni-paderborn.de

Abstract

We present LibriWASN, a data set whose design follows closely the LibriCSS meeting recognition data set, with the marked difference that the data is recorded with devices that are randomly positioned on a meeting table and whose sampling clocks are not synchronized. Nine different devices, five smartphones with a single recording channel and four microphone arrays, are used to record a total of 29 channels. Other than that, the data set follows closely the LibriCSS design: the same LibriSpeech sentences are played back from eight loudspeakers arranged around a meeting table and the data is organized in subsets with different percentages of speech overlap. LibriWASN is meant as a test set for clock synchronization algorithms, meeting separation, diarization and transcription systems on ad-hoc wireless acoustic sensor networks. Due to its similarity to LibriCSS, meeting transcription systems developed for the former can readily be tested on LibriWASN. The data set is recorded in two different rooms and is complemented with ground-truth diarization information of who speaks when.

1 Introduction

Continuous speech separation (CSS) refers to the task of mapping a continuous incoming data stream consisting of the speech of an arbitrary number of, possibly concurrently active, speakers to a fixed number of output channels in such a way that there is no longer speech overlap on any of the output channels. It is designed to be a source separating preprocessing step of a meeting recognition system. *LibriCSS* is a well-known publicly available data set for evaluating and comparing meeting transcription systems [1]. It is structured in sets with different percentages of speech overlap, i.e., the amount of time when two speakers are concurrently active. In each set eight loudspeakers at fixed positions play back clean recordings from the *LibriSpeech* corpus [2] while a synchronous seven-channel circular microphone array with 4.25 cm radius records the audio. Since its publication and use during the 2020 Frederick Jelinek Memorial Summer Workshop, *LibriCSS* has become the de facto standard for evaluation and comparison of meeting transcription systems. Those systems typically consist of a separation and enhancement stage, a diarization component and an automatic speech recognition (ASR) system, where the order of these tasks may vary [3].

Since research on meeting scenarios is getting more attention [4–6] and one data set cannot cater all scenarios, several extensions have been proposed. As *LibriCSS* is merely a test set, the Multi-Speaker Mixture Signal Generator (MMS-MSG) can be used to generate training data for meeting recognition systems: it can be parameterized to take utterances from the *LibriSpeech* corpus to artificially generate meeting data with a desired degree of speech overlap [7]. Further, the authors of [8] re-recorded *LibriSpeech* data with a microphone array consisting of 40 synchronized microphones distributed in a room of 110 m^2 to assess the performance of ad-hoc microphone arrays in a rather large setting. In [6] spatially distributed asynchronous microphones are used to record a data set named *AdHoc-LibriCSS*. It consists of mini-sessions with either two speakers or five speakers. Each mini-session was recorded by 5 recording devices, where the placement of the loudspeakers and recording devices were randomly chosen for each session. However, that data set is not publicly available.

We here also focus on an ad-hoc wireless acoustic sensor network (WASN) scenario. Having in mind a usage scenario where meeting participants use their own smartphones to record the meeting, we record the data with a set of smartphones, whose sampling clocks are not synchronized. The data is further recorded with multi-microphone devices, accounting for a situation, where a multi-channel conference communication system is additionally available for signal capture.

If multiple devices are employed, the sampling rates of each device will slightly differ from the target sampling frequency. This effect, named sampling rate offset (SRO), depends on environmental influences [9], e.g., voltage supply or temperature, and is time-varying [10]. As reported in [11] the SROs of smartphones and audio devices can be in the range between -40 parts per million(ppm) and 416 ppm, whereby devices from the same vendors have less variation.

Our data set, called *LibriWASN*, was recorded in two acoustically different rooms, one with a reverberation time of about 200 ms and the other with a reverberation time of about 800 ms. It offers 20 h of transcribed audio from 29 microphones from 9 different devices. This includes five smartphones, one soundcard and three self-developed smart devices based on Raspberry Pis. It has the same directory structure as *LibriCSS* to ease its adoption for those already working with *LibriCSS*. Since playing back the *LibriSpeech* utterances which were also used to create the 60 *LibriCSS* subsets the same ASR tools and models can be used to evaluate the data. We complement the data set with ground-truth information about speaker activity for performance analysis purposes.

The data set is intended to be used to conduct research on multi-channel source separation and meeting transcription tasks using asynchronous audio streams. Furthermore, the combination of multi-channel synchronous smart devices with additional asynchronous single-channel audio streams from smartphones can be studied. Here, approaches which either first estimate and compensate for the SROs [12–14] and then separate the speaker signals as well as approaches which handle the task in an integrated fashion [6] are conceivable.

The paper is structured as follows: In Sec. 2 the recording setup and details on the used hardware devices are presented. Then information about the recorded signals and preprocessing steps are summarized in Sec. 3. After explaining our reference system in Sec. 4 and showing results on the data set in Sec. 5 we conclude in Sec. 6 with a short summary.

2 Recording Setups

Fig. 1 displays the acoustic laboratory room with the recording setup for *LibriWASN*^{200}, i.e., the *LibriWASN* subset for a room reverberation time of $T_{60} \approx 200$ ms. This room has ceiling-high windows on one side of the room with radiators in front of the windows. The ceiling is suspended with mineral fiber boards and the floor is covered with low-pile carpet. A window to the neighboring room, a door and two tables are sound-reflecting elements. The remaining walls are partially covered with a sound-absorbing surface.

The second room, used for recording *LibriWASN*^{800}, is a laboratory room with lightweight walls, a linoleum floor, many window panes and a furnishing consisting of a glass cabinet and tables. This all together leads to an increased reverberation time of $T_{60} \approx 800$ ms. Furthermore, the room contained several computers as noise sources.

A sketch of the device placement for *LibriWASN*^{200} is shown in Fig. 2. In the center of the table a circular microphone array

Figure 1: Recording setup of *LibriWASN*^{200}: Eight loudspeakers surrounding a table with multiple recording devices.

consisting of eight *AKG C 400 bl* microphones with hypercardioid characteristic is placed. It is connected to the soundcard that plays the signals on the indicated loudspeaker positions. We used *JBL Control One* loudspeakers that were facing the table. Further, the smartphones and Raspberry Pis are distributed on the table as shown.

For *LibriWASN*^{800} the soundcard array remained in the center of the table, but the positions of the other devices on the table were changed. Position information for all devices and loudspeakers as well as pictures of the setups and rooms can be found in [15].

2.1 Hardware Devices

The following devices are used to record the data:

- Soundcard: Focusrite Scarlett 18i20 (3rd Gen), 8 channels, circular microphone array (diameter: 20 cm)
- Raspberry Pi 4 Model B
  - *asnupb2* & *asnupb4*: AudioInjector Octo, 6 channel analog frontend, circular microphone array (diameter: 5 cm)
  - *asnupb7*: Soundcard with 4 channels, adjustable sampling rate at part per billion (ppb) precision, quadratic microphone array (edge length: 5 cm)
- Android Smartphones
  - Xiaomi MI A2
  - LG Group Nexus 4
  - Google Pixel 6a (2×)
  - Google Pixel 7

A total of 29 microphone channels are recorded in a parallel process. The audio playback is handled by the soundcard. See [16] and [17] for more details on the 6 channel microphone analog frontend. To distinguish between the two *Pixel 6a* type smartphones we call one "Pixel 6a" and the other "Pixel 6b".

2.2 Adjustable Sampling Rate Offset

SRO manipulations in software have high computational costs [18] and, furthermore, may attenuate the upper frequencies bands due to the required low-pass filter [19]. Hence, we developed an audio front-end for Raspberry Pis whose sampling frequency can be adjusted in hardware at ppb granularity (see Fig. 3 right). It uses a field programmable gate array (FPGA) for handling the interface

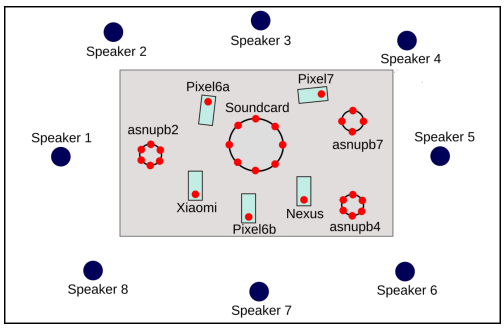

Figure 2: *LibriWASN*^{200} recording and playback devices: Eight loudspeakers surrounding a table with 5 smartphones, one 8-channel microphone array (soundcard) and three Raspberry Pi devices. Red dots indicate microphones on devices.

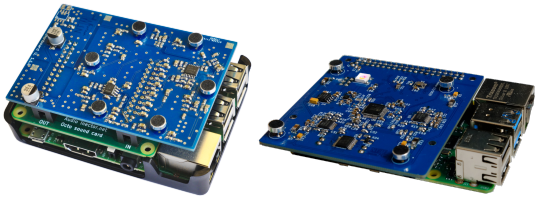

Figure 3: Raspberry Pis with mounted soundcards. Left: AudioInjector Octo with 6 channel microphone analog frontend (*asnupb2* & *asnupb4*). Right: FPGA-based 4 channel soundcard (*asnupb7*).

communication between hardware and kernel space drivers. Furthermore, it has the ability to compensate for SROs without additional computational requirements with smooth frequency changes in the range of ±1000 ppm during recording and a lossless recording of all upper frequencies w.r.t. the Nyquist theorem. To this end, on the frontend an Si514 chip (any-frequency inter-integrated circuit (I^2C) programmable crystal oscillator (xo)) is deployed to generate the sampling signal clock of the analog-digital converter (ADC).

While the hardware is originally intended to compensate for SROs, we here misuse it to artificially increase the SROs. Hardware bought from the market comes with a random SRO and, depending on the individual devices, a certain SRO spread can be observed. As reported in [11] typical SRO values range between −40 ppm and 416 ppm whereby values exceeding ±100 ppm are rarely observed. Hence, we set the sampling frequency of *asnupb7* to 16 kHz with an SRO of −100 ppm.

Researchers working on the data set can select a subset of devices to meet their requirements w.r.t. the SRO-range. In Sec. 5 some SRO measurements are presented.

2.3 Smartphones & Streaming

The Android operating system offers multiple different audio sources for recording. Some of them preprocess the microphone signals, e.g., type *voice recognition* suppresses environmental noise to a certain extent. To prevent signal loss or artifacts on all smartphones the least processed type was selected.

The recorded audio data was streamed wirelessly from each smartphone to a central data acquisition server. Similarly, the Raspberry Pis send their data via a wired connection. All connections are transmission control protocol (TCP)-based and have been continuously supervised to work losslessly.

3 Signals & Preprocessing

The recordings were performed during the regular operating of our laboratory at the university. As a result they may contain minimal interference from road traffic or impulsive interference from doors.

3.1 Data Quantization & Normalization

The Android smartphones sample the data at 16 kHz with a quantization of signed 16 Bit. The Raspberry Pis share the same sampling rate with a quantization of signed 32 Bit and the soundcard has a sampling rate of 48 kHz with 32 Bit resolution. The gain of the analog circuits was adjusted so that no clipping occurred during the recordings and all audio stream samples are converted to 16 Bit little endian values.

3.2 Downsampling

High-quality multi-channel soundcards with native support of a sampling rate of 16 kHz are currently not available on the market. Hence, we used a *Focusrite Scarlett 18i20 (3rd Gen)* universal serial bus (USB) soundcard to playback and record the *LibriSpeech* data at 48 kHz. To this end, we implemented an up/down-sampling method employing a linear phase, optimal equiripple finite impulse response (FIR) filter with a stop band attenuation of 50 dB and a filter length of 256 taps.

3.3 High-Pass Filtering

The recordings of *LibriWASN*^{200} showed low-frequency interference from the water pipes of the heating and an air conditioner from an adjacent lab room, which we removed by high-pass filtering (3 dB cutoff frequency at 75 Hz) from all recordings. After removing the low frequency noise the signals have a signal-to-noise ratio (SNR) of approximately 30 dB depending on their position and the active source. For example, the Raspberry Pi *asnupb4* has a measured SNR of 15.25 dB (in set *OV10, session 3, first speaker*) and after high-pass filtering an SNR of 33.21 dB. The recordings of *LibriWASN*^{800} contain more noise from the computers in the background and no noise from the air conditioning system (e.g., *asnupb4* has an approximate SNR of 20 dB, a high-pass filter increases it by 3 dB). Thus, we have taken them unprocessed and leave any noise suppression to future users.

3.4 Sampling Time Offset Reduction

The asynchronous devices in the network start recording at unknown points in time. The resulting initial time offset, called sampling time offset (STO), is not only influenced by quantities that can be controlled, e.g., the selected packet size when interacting with the soundcard and the size of the soundcard data buffer, but also by random quantities that cannot be controlled, e.g., the latency in the network, the packet size of the network and the scheduling of the kernel.

We reduce the STO to a technically possible minimum by first starting the recording on all devices and merging the data streams at a central node. After all devices have delivered data, we discard already received data in the queue and start signal recording. This reduces the STO to a size of a few packets. The remaining offset is determined by correlating the signals and then reduced to an approximate range of ± 40 samples during the first 10 s of each recording. Note that the described STO minimization is done per device so that the intra-device time differences of arrival (TDOAs) are maintained.

The inter-device TDOAs correspond to a superposition of the STO and the time difference of flight (TDOF) [10], which is caused by the different distances of the sources to the microphones. Thus, the above time offset removal, which forces these TDOAs to be close to zero at the beginning of the recordings, does not only remove the STO but also manipulates the TDOF information. Since the latter is carrying the source position information source localization based on inter-device TDOAs, estimated from the processed signals, cannot be performed.

4 Reference System

In addition to the data set we also provide a meeting transcription pipeline as reference/baseline system for future works. In the experimental section we provide results for this pipeline using single device and multi device setups. To enable a coherent processing of signals gathered by different devices, we firstly compensate for SROs. For this we use the dynamic weighted average coherence drift (DWACD) method [10] to estimate the SROs and afterwards compensate for them using the short-time Fourier transform (STFT)-resampling from [18].

In order to extract the single speakers' signals, mask-based beamforming is utilized. A complex Angular Central Gaussian Mixture Model (cACGMM) [20] with time-dependent instead of frequency-dependent mixture weights [21] is used to estimate a mask for each of the speakers and an additional mask for noise. The initialization of the cACGMM is based on the idea to divide the meeting into segments consisting of multiple frames, which are clustered afterwards [22]. For each segment a spatial covariance matrix (SCM) is estimated. To avoid ambiguities due to speech pauses or overlapping speech, a rank-1 approximation of the SCMs is conducted. In addition to that, the ratio of the largest and the second largest eigenvalue of the SCMs is used as indicator if one speaker is dominant within a segment. If this ratio is below a certain threshold, indicating either a speech pause or overlapping speech, the segment is assigned to the noise class. The segments are clustered based on the similarity of their SCMs, which is measured by the the correlation matrix distance from [23]. By quantizing the similarity measure to zero (not the same speaker) or one (the same speaker) for each segment based on a certain threshold, results in an activity pattern that indicates in which segments the same speaker is active. Subsequently, a leader-follower clustering is used to group the segments whose activity patterns intersect most. To avoid that the cACGMM diverges too much from the initialization, the latter is used as guide for the first expectation maximization (EM) iterations.

The speakers' signals are extracted using a minimum variance distortionless response (MVDR) beamformer in the formulation of [24]. Therefore, the meeting is segmented using the priors of the cACGMM as described in [22]. In each segment the SCM of the target speaker and the SCM of the interference are estimated using the masks obtained from the cACGMM. The mask used to estimate the interference SCM is obtained as sum of all masks except the one of the target speaker. Finally, a pretrained ASR system [25] for *LibriSpeech* from the ESPnet framework [26] is used to transcribe the separated signals. This ASR system has a transformer architecture and is trained on *LibriSpeech*. On the clean test set of *LibriSpeech* it achieves a word error rate (WER) of 2.7 %.

Smartphone	SRO
Pixel6a	16.62 ppm
Pixel6b	13.58 ppm
Pixel7	13.73 ppm
Nexus	−0.11 ppm
Xiaomi	−1.44 ppm

Device	SRO
Soundcard	0.0 ppm (Ref.)
asnupb2	−10.35 ppm
asnupb4	−23.18 ppm
asnupb7	−109.03 ppm

Table 1: Estimated average SROs of WASN devices w.r.t. soundcard as reference.

5 Experiments

In the following two aspects of the data set are investigated. First, the SROs between the different recording devices are analyzed. Afterwards, the proposed reference system is investigated for varying recording setups.

5.1 Sampling Rate Offsets

We estimated the average SRO of the hardware device from the audio recordings by employing the DWACD method. As shown in Tab. 1 the hardware dependent SROs of the devices are in the range between ± 20 ppm w.r.t. the sampling rate of the soundcard.

Since [11] reported SROs of more than 100 ppm, we decided to enrich the data set with a device that has an artificially higher SRO. To this end, we intentionally set the changeable sampling frequency of *asnupb7* to 15 998.4 kHz relative to its build-in oscillator corresponding to an additional SRO of -100 ppm. This results in an average SRO of -109.03 ppm relative to the soundcard. Note that during the separate recording sessions the device temperatures changed and, thus, the SROs are slightly time-varying.

In Fig. 4 the estimated delays between the first channel of the soundcard and the first channel of the other devices are exemplary depicted over time. While the measures described in the last section lead to a delay of nearly zero at the start of the recording, the delays drift over time due to the different SROs. One can also clearly see the abrupt changes in the delay values, that are caused by changing TDOFs due to speaker changes.

5.2 Source Separation

Table 2 shows a comparison of the *LibriCSS* data set and the two subsets of the *LibriWASN* data set w.r.t. the performance of the proposed reference meeting transcription system. As measure for the meeting transcription performance the concatenated minimum-permutation word error rate (cpWER) [27] is utilized.

	System information [Device]	Sync.	Devices	Channels	cpWER / % 0L	0S	OV10	OV20	OV30	OV40	Avg.
	Clean	-	-	-	2.90	2.58	2.58	2.43	2.51	2.31	2.52
LibriCSS	Sys-1: Segmentation (Oracle act.)	-	1	1	4.36	4.49	10.61	18.72	27.21	35.98	18.56
	Sys-2: cACGMM & MVDR	-	1 / 1	7 / 7	3.57	3.44	3.76	4.38	5.40	5.38	4.43
LibriWASN^{200} $T_{60} \approx 200$ ms	Sys-1: Segmentation (Oracle act.) [Pixel7]	-	1	1	3.69	3.47	12.21	21.84	30.73	39.95	20.59
	Sys-2: cACGMM [asnupb4] & MVDR [asnupb4]	-	1 / 1	6 / 6	3.11	3.2	4.68	5.35	5.00	4.60	4.43
	Sys-3: cACGMM [asnupb4] & MVDR [all]	✓	1 / 9	6 / 9	3.22	3.01	4.38	5.48	3.67	4.50	4.11
	Sys-4: cACGMM [all] & MVDR [all]	✓	9 / 9	9 / 9	3.20	3.56	5.36	10.37	3.43	5.54	5.38
LibriWASN^{800} $T_{60} \approx 800$ ms	Sys-1: Segmentation (Oracle act.) [Pixel7]	-	1	1	4.71	4.70	13.40	22.97	32.03	41.71	21.89
	Sys-2: cACGMM [asnupb4] & MVDR [asnupb4]	-	1 / 1	6 / 6	3.86	3.92	5.26	7.09	7.15	6.91	5.90
	Sys-3: cACGMM [asnupb4] & MVDR [all]	✓	1 / 9	6 / 9	3.87	3.55	4.14	5.33	4.94	4.59	4.47
	Sys-4: cACGMM [all] & MVDR [all]	✓	9 / 9	9 / 9	3.93	3.62	3.90	3.99	4.46	10.01	5.20

Table 2: Comparison of cpWERs on *LibriCSS* and *LibriWASN*. The cpWER is calculated for the different overlap ratios / silence conditions defined by the *LibriCSS* data set: no overlap with long silence (0L), no overlap with short silence (0S) and between 10 % (OV10) and 40 % (OV40) overlap. *Clean* denotes transcribing the original LibriSpeech utterances used to record *LibriCSS* and *LibriWASN*. *Segmentation* means that oracle activity information is used to cut the unprocessed signal of one channel into blocks, which correspond to single utterances. The first number in the *Devices*/*Channels*-column defines the number of devices/channels used for mask estimation and the second number the number of devices/channels used for beamforming. If signals of multiple devices are jointly aggregated only the first channel of each device is utilized.

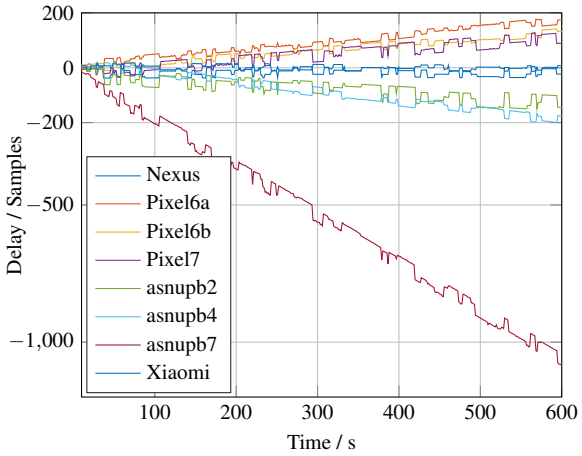

Figure 4: Estimated delay between soundcard and first channel of each device (*LibriWASN*^{200}, OV40, Ses. 9).

In order to get an impression of the acoustic conditions of *LibriWASN* w.r.t. the ASR performance, we segmented the unprocessed signal of one channel using oracle activity information for each speaker (Sys-1). It becomes obvious that the acoustic conditions (see sessions without overlap: 0L and 0S) of *LibriWASN*^{200} seem to be less challenging w.r.t. ASR than the acoustic conditions of *LibriCSS*. In contrast, the acoustic conditions of *LibriWASN*^{800} seem to be a little bit more challenging than the acoustic conditions of *LibriCSS* due to greater reverberation and more noise.

Using a single microphone array for mask estimation and beamforming (Sys-2), it can be seen that similar results can be achieved for *LibriCSS* and *LibriWASN*^{200}. Again, the performance for *LibriWASN*^{800} is slightly worse due to the more challenging acoustic conditions. If the masks are still estimated using a single array but all devices are used for beamforming (Sys-3), the results for both subsets of *LibriWASN* can be improved. This especially holds for *LibriWASN*^{800}. Thus, the achieved synchronization based on the DWACD method seems to be good enough to enable a coherent signal processing and, therefore, to benefit from the spatial diversity of the distributed recording setup.

Utilizing all devices for mask estimation and beamforming (Sys-4), the overall performance of the reference system degrades slightly. This is caused by a few sessions with very large cpWERs. We hypothesize that these outlier results are caused by errors made during the initialization of the cACGMM, e.g., by mixing the activity of two speakers. However, the good results for some overlap ratios indicate that mask estimation using spatial information can also benefit from spatial diversity of the recording setup.

Although the results achieved on the *LibriWASN* data set are already decent there is still room for further improvement when comparing the results to the cpWER which can be achieved by transcribing the clean utterances from *LibriSpeech* (first result line of Table 2). Further, it is to be mentioned that the proposed reference system is based on spatial information, which is a very powerful source of information in a static setup with fixed source positions. How to achieve comparable performance with a single-channel system or with moving speakers, remains an open problem.

6 Summary

We have presented a data set which consists of re-recordings of the *LibriCSS* data set in two different acoustic environments with 9 devices and a total of 29 channels. Its intended use is for diarization and meeting transcription research in ad-hoc wireless acoustic sensor networks, with a focus on synchronization and multi-channel signal processing.

Download & Licence

The data set is available under Creative Commons Attribution 4.0 International License (CC BY 4.0) from Zenodo.

- Zenodo data set link:
  `https://zenodo.org/record/7960972`
- Scripts for data set handling and code of the reference system:
  `https://github.com/fgnt/libriwasn`

Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project 282835863.

References

[1] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li, "Continuous speech separation: Dataset and analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[2] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[3] Desh Raj, Pavel Denisov, and Zhuo Chen et al., "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 897–904.

[4] Takuya Yoshioka, Dimitrios Dimitriadis, Andreas Stolcke, William Hinthorn, Zhuo Chen, Michael Zeng, and Xuedong Huang, "Meeting transcription using asynchronous distant microphones," in *Proc. Interspeech*, September 2019.

[5] Shota Horiguchi, Yusuke Fujita, and Kenji Nagamatsu, "Utterance-wise meeting transcription system using asynchronous distributed microphones," in *Interspeech*, 2020.

[6] Dongmei Wang, Takuya Yoshioka, Zhuo Chen, Xiaofei Wang, Tianyan Zhou, and Zhong Meng, "Continuous speech separation with ad hoc microphone arrays," in *European Signal Processing Conference (EUSIPCO)*, 2021.

[7] Tobias Cord-Landwehr, Thilo von Neumann, Christoph Boeddeker, and Reinhold Haeb-Umbach, "MMS-MSG: A multi-purpose multi-speaker mixture signal generator," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

[8] Shanzheng Guan, Shupei Liu, and Junqi Chen et al., "Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1116–1120.

[9] Fred L. Walls and Jean-Jacques Gagnepain, "Environmental sensitivities of quartz oscillators," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 39, pp. 241–9, 02 1992.

[10] Tobias Gburrek, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "On synchronization of wireless acoustic sensor networks in the presence of time-varying sampling rate offsets and speaker changes," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[11] Mario Guggenberger, Mathias Lux, and Laszlo Böszörmenyi, "An analysis of time drift in hand-held recording devices," in *MultiMedia Modeling*. 2015, pp. 203–213, Springer International Publishing.

[12] Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2012.

[13] Shigeki Miyabe, Nobutaka Ono, and Shoji Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in stft domain," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 674–678.

[14] Mohamad Hasan Bahari, Alexander Bertrand, and Marc Moonen, "Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 674–686, 2017.

[15] Joerg Schmalenstroeer, Tobias Gburrek, and Reinhold Haeb-Umbach, "Zenodo: LibriWASN data set (open access)," https://zenodo.org/record/7960972, 2023.

[16] DFG FOR 2457, "Homepage acoustic sensor networks project - open hardware," https://upb.de/asn/hardware, 2023.

[17] Haitam Afifi, Joerg Schmalenstroeer, Joerg Ullmann, Reinhold Haeb-Umbach, and Holger Karl, "MARVELO - a framework for signal processing in wireless acoustic sensor networks," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[18] Joerg Schmalenstroeer and Reinhold Haeb-Umbach, "Efficient sampling rate offset compensation - an overlap-save based approach," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018.

[19] Aleksej Chinaev, Gerald Enzner, and Joerg Schmalenstroeer, "Fast and accurate audio resampling for acoustic sensor networks by polyphase-Farrow filters with FFT realization," in *Proc. of ITG Fachtagung Sprachkommunikation (Speech Communications)*, Oct. 2018.

[20] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2016.

[21] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[22] Christoph Boeddeker, Tobias Cord-Landwehr, Thilo von Neumann, and Reinhold Haeb-Umbach, "An initialization scheme for meeting separation with spatial mixture models," in *Proc. INTERSPEECH*, 2022.

[23] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, "Correlation matrix distance, a meaningful measure for evaluation of non-stationary MIMO channels," in *2005 IEEE 61st Vehicular Technology Conference*, 2005, vol. 1, pp. 136–140 Vol. 1.

[24] Mehrez Souden, Jacob Benesty, and Sofiène Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2010.

[25] Shinji Watanabe, "ESPnet2 pretrained automatic speech recognition model, https://doi.org/10.5281/zenodo.3966501," July 2020.

[26] Shinji Watanabe, Takaaki Hori, and Shigeki Karita et al., "ESPnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018.

[27] Shinji Watanabe, Michael Mandel, and Jon Barker et al., "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.