Spatial Diarization for Meeting Transcription with Ad-Hoc Acoustic Sensor Networks

Tobias Gburrek, Joerg Schmalenstroeer and Reinhold Haeb-Umbach
Department of Communications Engineering
Paderborn University, Germany
{gburrek, schmalen, haeb}@nt.uni-paderborn.de

***Abstract*—We propose a diarization system, that estimates "who spoke when" based on spatial information, to be used as a front-end of a meeting transcription system running on the signals gathered from an acoustic sensor network (ASN). Although the spatial distribution of the microphones is advantageous, exploiting the spatial diversity for diarization and signal enhancement is challenging, because the microphones' positions are typically unknown, and the recorded signals are initially unsynchronized in general. Here, we approach these issues by first blindly synchronizing the signals and then estimating time differences of arrival (TDOAs). The TDOA information is exploited to estimate the speakers' activity, even in the presence of multiple speakers being simultaneously active. This speaker activity information serves as a guide for a spatial mixture model, on which basis the individual speaker's signals are extracted via beamforming. Finally, the extracted signals are forwarded to a speech recognizer. Additionally, a novel initialization scheme for spatial mixture models based on the TDOA estimates is proposed. Experiments conducted on real recordings from the LibriWASN data set have shown that our proposed system is advantageous compared to a system using a spatial mixture model, which does not make use of external diarization information.**

***Index Terms*—Diarization, time difference of arrival, ad-hoc acoustic sensor network, meeting transcription**

I. Introduction

When transcribing a meeting, often not only the information of what has been said is of interest but also the information "who spoke when", i.e., diarization information. Additionally, diarization information can also be helpful for speech enhancement, e.g., using the guided source separation (GSS) [1] framework. However, gathering diarization information is a challenging task due to the highly dynamic nature of spontaneous conversations with alternating silence and speech regions, as well as overlapping speech from multiple speakers.

In particular, the segments with overlapping speech are challenging for diarization. For example, the performance of methods, that rely on spectro-temporal information, often tends to degrade with an increasing amount of overlapping speech. This especially holds for early diarization systems [2]. Although nowadays diarization systems, like TS-VAD [3], are able to cope much better with overlap, their performance is often still negatively affected by overlap [4].

In a typical meeting scenario with multiple speakers sitting around a table at spatially well separated, (quasi-)fixed positions the information "when and at which position" a speaker is active also reveals the diarization information. In such a scenario spatial information can be a promising alternative to cope with speech overlap. Typically, direction of arrival (DOA) information, which is gathered using a compact microphone array, is employed as source of spatial information [5]–[9].

Discriminating between two speakers based on DOA information might be challenging, if the distance between the speakers and the microphone array is large and the speakers sit close to each other. The spatial diversity of an ASN comes in handy in such situations by offering TDOA information, which allows for a better distinction between those speakers. However, ASNs are typically formed ad-hoc, e.g., by smartphones. Hence, the microphone positions are generally unknown and the recorded signals are typically asynchronous, which makes it difficult to infer the speakers' position from the TDOA estimates. In [10] we approached these issues by using geometry calibration [11] and a complex synchronization method [12], which maintains the information about the microphones' and speakers' positions, as preprocessing steps before diarization.

Here, a much simpler approach to synchronization is employed, which however distorts the information about the microphones' and speakers' positions by constant TDOA offsets. Although, these distortions make it difficult to map the TDOAs to the coordinates of the speakers' positions anymore, the TDOAs still uniquely represent the speakers' positions. Hence, we propose to derive diarization information by clustering estimates stemming from a multi-speaker TDOA estimator, which delivers estimates at frame rate.

The resulting diarization information is used as a guide for a spatial mixture model in the GSS framework, to force the posterior probability to be zero when a speaker is inactive. In experiments on the LibriWASN [13] data set we show that the guided spatial mixture model is able to outperform a blind spatial mixture model, which does not employ external diarization information. Additionally, a time-frequency bin wise initialization scheme for a spatial mixture-model based on the TDOA estimates is proposed to speed up the convergence.

In the following we describe the considered meeting scenario in Section II and give an overview of the meeting transcription pipeline in Section III. Afterwards, the proposed TDOA-based diarization system is introduced in Section IV, followed by a description, how the diarization information and the TDOA estimates can be employed to support source extraction, in Section V. Experimental results are reported in Section VI. Finally, conclusions are drawn in Section VII.

II. Scenario Description

In the following a meeting-like conversation of I speakers is considered, which should be transcribed. It is assumed that the speakers sit at spatially well separated, fixed but unknown positions around a table. During the conversation, there are periods in time without speech activity, periods in time with a single speaker being active and a significant amount of periods in time with two speakers being active at the same time. On the table, $M \geq 4$ microphones, forming an ad-hoc ASN, are distributed, which are used to record the meeting. The microphones are located at fixed but unknown positions.

Since the devices in an ad-hoc ASN are generally independent of each other, the microphone signals are sampled with slightly different sampling frequencies even though the devices have the same nominal sampling rate. This introduces a sampling rate offset (SRO) between the microphone signals. Furthermore, the devices usually start their recordings at different points in time, which causes a sampling time offset (STO) between the microphone signals.

III. Meeting Transcription System

The meeting transcription system, which will be considered in the following, is depicted in Fig. 1. Firstly, the microphone signals are synchronized w.r.t. a reference channel. To do so, first the STOs are compensated for by a correlation-based coarse synchronization [12], [14], which forces the TDOAs between the signals to be close to zero at the beginning of the recordings. Afterwards, the SROs are compensated for via resampling [12]. The diarization information, which is gathered from TDOA information, as well as the estimated TDOAs are used to support the extraction of the single speakers' signals from the noisy and reverberant speech mixtures. Finally, the extracted signals are transcribed.

IV. TDOA-based Diarization

We here propose to cluster frame-wise TDOA estimates as representation of the active speakers' positions in order to gather diarization information. Therefore, a TDOA estimator, that is able to cope with overlapping speech, is introduced.

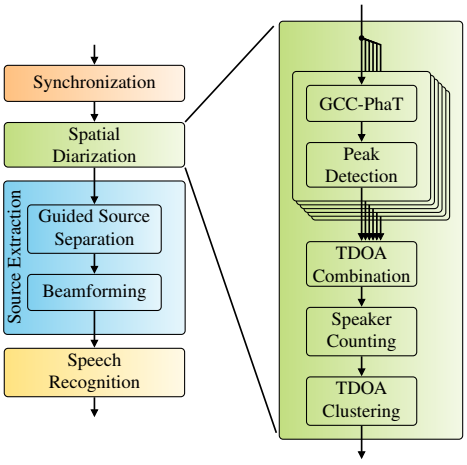

Fig. 1. Meeting transcription pipeline

A. Effect of Asynchronous Recordings

In [12] it was shown that the TDOA $\tau_{i,mm'}[\ell]$ between the m-th and the m'-th channel corresponds to a superposition of the time difference of flight (TDOF) of the i-th speaker's signal between the m-th and the m'-th channel, a constant offset due to the STO and a time-varying SRO-induced delay. Here, ℓ denotes the time frame index. The TDOF is a characteristic of the i-th speaker's position relative to the microphones and, thus, contains spatial information.

The coarse synchronization compensates not only for an STO but rather for a combination of STO, SRO-induced delay and TDOF. Due to this fact the TDOAs cannot be mapped to the coordinates of the speakers' positions anymore. However, the coarse synchronization affects the TDOAs in form of a constant value, which solely depends on the microphone pair. Thus, each source position still can be uniquely represented by a vector of all pairwise TDOAs $\boldsymbol{\tau}_i = [\tau_{i,12},\, \tau_{i,13},\, \ldots,\, \tau_{i,M-1M}]^T$ after synchronization.

B. Multi-Speaker TDOA Estimation

As a basis for diarization TDOA vectors are estimated in each time frame (see right half of Fig. 1). To this end, the generalized cross-correlation with phase transform (GCC-PhaT) [15] $g_{mm'}(\ell, \lambda)$, with λ being the time lag, is firstly estimated for all microphone pairs. In order to get more robust TDOA estimates, the GCC-PhaT $g_{mm'}(\ell, \lambda)$ is averaged across L consecutive time frames. Moreover, the GCC-PhaT is only calculated on the basis of the frequency range from 125 Hz to 3.5 kHz, i.e., the frequency range for which speech has significant power.

Since multiple speakers can be active within a time frame, the C time lags λ_c, belonging to the C highest local maxima of the GCC-PhaT $g_{mm'}(\ell, \lambda)$, are considered as possible TDOA candidates. Due to the fact that the direct path signal corresponds to a delayed and attenuated version of the source signal, only time lags λ_c, belonging to positive local maxima, are considered as TDOA candidates [16]. Furthermore, the local maximum has to be larger than twice the standard deviation of the GCC-PhaT, which is calculated w.r.t. the time lag λ for the ℓ-th time frame.

Afterwards, the pairwise TDOA candidates have to be combined to form consistent TDOA vectors. All elements of a consistent TDOA vector have to fulfill the cyclic consistency condition, i.e., in case of three microphones m, n and o

$$\tau_{mn} - \tau_{mo} + \tau_{on} \leq \tau_{\text{th}}, \tag{1}$$

has to be fulfilled, where τ_{th} is a small value of a few samples. Since we do not check for exact equality to zero in (1), additional valid TDOA vectors, e.g., stemming from multi-speaker ambiguities or echos, are possible. Here, we tackle this issue by utilizing the fact that speaker positions of equal TDOA lie on a hyperboloid and the speakers' positions are associated with the point of intersection of the hyperboloids belonging to the different microphone pairs. Moving along the hyperboloid of equal TDOA of one microphone pair, changes the points of intersection so that the TDOAs of all other microphone

pairs have to change. Hence, at maximum one element is allowed to be equal for two TDOA vectors. In case of multiple TDOA vectors having more than one common element, only the TDOA vector with the largest steered-response power with phase transform (SRP-PhaT) is kept. Thereby, the SRP-PhaT is efficiently computed from the previously calculated pairwise GCC-PhaTs $g_{mm'}(\ell, \lambda)$.

Finally, the number of speakers being active within a time frame is determined. To decide whether there is speech, an energy-based voice activity detection (VAD) is utilized. In case of speech activity the TDOA vector with the largest SRP-PhaT is considered to belong to an active speaker. In addition to that, the SRP-PhaT is used to decide whether multiple speakers are active. Additional TDOA vectors and, thus, additional speakers for a time frame are considered if the corresponding SRP-PhaT is larger than the mean of the largest SRP-PhaT value per frame minus twice their standard deviation.

C. TDOA Clustering

Diarization information is gathered by clustering the estimated frame-wise TDOA vectors. First, temporally local clusters, corresponding to speaker activity information approximately at utterance-level, are formed. These temporally local clusters are determined via a leader-follower clustering [17]. Thereby, the TDOA vector of the most recent frame within a cluster becomes its new leader. The temporal locality of the clusters is forced by considering only TDOA vectors which do not lie more than $1\,\mathrm{s}$ in the past as possible leaders. We use the maximum of the element-wise absolute difference between two TDOA vectors as clustering metric.

Subsequently, a single-linkage clustering [18] is employed to obtain the global diarization information from the temporally local clusters. To this end, the temporally local clusters are represented by the element-wise median of the TDOA vectors of their cluster members and the mean-squared deviation (MSD) between the TDOA vectors is used as clustering metric. The clustering is aborted when the MSD is larger than a certain threshold to address outlier TDOA vectors.

The final clustering result often contains more clusters than there are speakers. These clusters mostly belong to TDOA vectors which correspond to a combination of direct path TDOAs and TDOAs of early reflections or a combination of direct path TDOAs of multiple speakers. To mitigate these influences, we first sort the estimated speakers' activities by the amount of frames with activity. If a cluster with a smaller amount of activity intersects more than $50\,\%$ with a cluster with a larger amount of activity and more than one element of the TDOA vectors of both clusters match each other (see hyperboloild property of TDOA vectors described above), the cluster with the smaller amount of activity is discarded. After all, a dilation and an erosion filter are applied to the estimated activities to smooth the activity estimates [19].

V. Source Extraction

As shown in Fig. 1 mask-based beamforming is utilized to extract the single speakers' signals. The masks, which are used to calculate the beamformer coefficients, are estimated via a spatial mixture model using the TDOA-based diarization information as guide.

A. Guided Source Separation

A time-frequency mask for each speaker and an additional mask for noise are estimated using GSS. In the GSS framework, the TDOA-based diarization is employed to force the class posterior probability of a spatial mixture model, i.e., the time-frequency masks, to be zero when the corresponding speaker is not active. In contrast to the original GSS method from [1] we here use a complex Angular Central Gaussian Mixture Model (cACGMM) [20] with time-dependent instead of frequency-dependent mixture weights [21] as spatial mixture model. Since the segmentation needed for GSS, which is given by the TDOA-based diarization, may also contain segments whose length is underestimated, a context of $\pm 5\,\mathrm{s}$ and additional non-guided Expectation Maximization (EM) iterations, that follow the guided EM iterations, are utilized.

One way to employ the TDOA-based diarization information for initialization of the spatial mixture model is to broadcast the speakers' activities over all frequencies as in the original implementation of GSS. We here propose to utilize the estimated TDOA vectors to derive an initial time-frequency mask for each source. Therefore, a steering vector based minimum variance distortionless response (MVDR) beamformer [22] per speaker is derived from the TDOA vectors, assuming anechoic signal propagation. The spatial covariance matrices (SCMs) of the interference are calculated as sum of the outer products of the steering vectors of all possibly interfering speakers. Afterwards the MVDR beamformers are applied in the short-time Fourier transform (STFT) domain. Assuming W-disjoint orthogonality of speech [23] each time-frequency bin is assigned to the mask of the active speaker whose beamformer has the largest output power.

Finally, the method from [24] is used to identify the time-frequency bins which are dominated by a single speaker: The SCM of the microphone signals is estimated for each time-frequency bin based on a short temporal and frequency context. Afterwards, the ratio of the largest and the second-largest eigenvalue of the SCMs is compared to a certain threshold. If the largest eigenvalue is significantly larger than the second-largest eigenvalue, the time-frequency bin is assumed to be dominated by a single speaker. All time-frequency bins which are not dominated by a single speaker are assigned to the initial noise mask.

B. Beamforming

We utilize an MVDR beamformer in the formulation of [25], [26] to extract the signals of the single speakers. Therefore, we first re-segment the segments used for GSS based on the target speakers' activities, which are calculated from the estimated prior probabilities of the spatial mixture model as described in [19]. The beamforming coefficients are calculated for each resulting segment, defined by continuous activity of the target

speaker, whose signal should be extracted. The SCM of the target speaker is calculated via

$$\mathbf{\Phi}_i(k) = \frac{1}{|\mathcal{T}_i|} \sum_{\ell \in \mathcal{T}_i} \gamma_i^2(\ell, k) \cdot \boldsymbol{Y}(\ell, k) \cdot \boldsymbol{Y}^H(\ell, k), \quad (2)$$

with $\mathcal{T}_i$ corresponding to the set of time frames, which belong to the segment, $\gamma_i(\ell, k)$ being the time-frequency mask of the target speaker and $\boldsymbol{Y}(\ell, k)$ denoting the vector of stacked STFTs of all microphone signals. The frequency bin index is denoted by k.

Since the set of active interfering speakers typically varies over time during a segment, we divide the segment into subsegments, whose boundaries are given by the change points of the interfering speakers' activities. For each subsegment new beamformer coefficients are calculated based on the interference SCM $\bar{\mathbf{\Phi}}_{i,b}(k)$, which is estimated via

$$\bar{\mathbf{\Phi}}_{i,b}(k) = \frac{1}{|\mathcal{T}_{i,b}|} \sum_{\ell \in \mathcal{T}_{i,b}} (1 - \gamma_i(\ell, k))^2 \cdot \boldsymbol{Y}(\ell, k) \cdot \boldsymbol{Y}^H(\ell, k). \quad (3)$$

Here, b denotes the index of the subsegment and $\mathcal{T}_{i,b}$ the set of time frames, which belong to the b-th subsegment. The reference channel for beamforming is chosen such that the expected signal-to-distortion ratio (SDR) of the sub-segment, which exhibits the lowest expected SDR, is maximized [26].

VI. Experiments

For the experiments we utilize the LibriWASN data set. The LibriWASN data set consists of recordings of replayed meetings with various overlap conditions, including no overlap (0L & 0S) as well as $10\,\%$ (OV10) to $40\,\%$ (OV40) of speech overlap. Moreover, the data set offers recordings from two different rooms resulting in the subsets LibriWASN^{200} (reverberation time $T_{60} \approx 200\,\mathrm{ms}$) and LibriWASN^{800} ($T_{60} \approx 800\,\mathrm{ms}$ and computer fan noise in background). The meetings were recorded by an ASN consisting of multiple smartphones and Raspberry Pis, which were equipped with soundcards.

The system proposed in this contribution is completed by the synchronization and automatic speech recognition (ASR) building blocks from the reference system provided with the LibriWASN data set in [13]. In order to measure the meeting transcription performance, we employ the concatenated minimum-permutation word error rate (cpWER) [27].

TABLE I

Comparison of the time frame wise initialization by broadcasting the diarization information along all frequencies (T-Init) and the proposed time-frequency bin wise initialization (TF-Init) for different amounts of guided EM iterations followed by one additional non-guided EM iteration. The signals of the smartphones (Pixel6a, Pixel6b, Pixel7, Xiaomi) from the LibriWASN^{800} data set are used.

Guided Iter.	Init.	cpWER / %						
		0L	0S	OV10	OV20	OV30	OV40	Avg.
1	T-Init	3.33	3.30	3.58	4.00	5.15	5.03	4.17
	TF-Init	3.13	2.98	3.11	3.36	4.00	3.90	3.46
2	T-Init	3.36	3.10	3.37	3.56	4.41	4.28	3.74
	TF-Init	3.11	2.93	3.18	3.41	3.88	3.76	3.42
5	T-Init	3.25	2.93	3.22	3.31	3.83	3.83	3.43
	TF-Init	3.11	2.97	3.20	3.34	3.84	3.67	3.39
20	T-Init	3.11	2.96	3.20	3.33	3.91	3.66	3.40
	TF-Init	3.10	2.94	3.18	3.35	3.88	3.62	3.38

TABLE II

Comparison of the blind spatial mixture model from [13] and the TDOA-based GSS system. Clean denotes transcribing the original LibriSpeech utterances, which were replayed to record the LibriWASN data set.

Data set	Devices	System	cpWER / %						
			0L	0S	OV10	OV20	OV30	OV40	Avg.
		Clean	2.92	2.61	2.60	2.48	2.61	2.43	2.59
LibriWASN^{200}	*Phones*	Blind	2.97	2.75	2.86	2.85	3.46	2.93	2.98
		Guided	2.91	2.77	2.76	2.69	3.11	2.76	2.83
	All	Blind	2.86	2.74	2.84	3.73	3.23	3.07	3.10
		Guided	2.93	2.75	2.76	2.77	2.91	2.74	2.80
LibriWASN^{800}	*Phones*	Blind	3.04	3.09	3.75	4.76	7.71	6.08	4.96
		Guided	3.11	2.93	3.18	3.23	3.64	3.54	3.30
	All	Blind	3.09	2.93	3.25	4.46	3.69	3.12	3.45
		Guided	3.00	2.88	2.96	2.82	3.08	2.85	2.93

A. Time-Frequency Mask vs. Time Mask Initialization

The influence of the different initialization strategies for the spatial mixture model, i.e., time frame wise initialization by broadcasting the diarization information along all frequencies (T-Init) and the proposed time-frequency bin wise initialization (TF-Init), is shown in Table I. GSS with up to 20 guided EM iterations followed by an additional non-guided EM iteration is considered.

It can be seen that the proposed time-frequency bin wise initialization is able to outperform the time frame wise initialization. This especially holds for the subsets with more speech overlap and when only a few EM iterations are used. Moreover, it becomes obvious that the cpWER already converges after a few EM iterations for the time-frequency bin wise initialization. Significantly more EM iterations are needed for the time frame wise initialization to end up at a similar performance as for the time-frequency bin wise initialization.

B. Guided Source Separation vs. Blind Source Separation

Table II compares the meeting transcription performance which can be reached with the proposed TDOA-based GSS system to the performance which can be achieved with the blind spatial mixture model of the LibriWASN reference system from [13], which does not utilize external diarization information. Furthermore, the initialization of the blind spatial mixture model is not able to cope with overlapping speech and the entire meeting is used at once to estimate the parameters of the blind spatial mixture model. For a fair comparison, the source extraction proposed in this contribution is adopted to the baseline system. GSS uses five guided EM iterations followed by five non-guided EM iterations. In order to investigate the influence of the amount of available channels, we consider the set *Phones* with four channels, stemming from smartphones

(Pixel6a, Pixel6b, Pixel7, Xiaomi) and the set *all* with seven channels stemming from the smartphones and three additional Raspberry Pis with soundcards (asnupb2, asnupb4, asnupb7).

In general, the TDOA-based GSS system is able to outperform the blind spatial mixture model. Thereby, the gap in performance is larger under the more challenging conditions of the LibriWASN^{800} data set with more reverberation and noise. This especially holds for the sub sets with a larger amount of speech overlap. In addition to that, it becomes clear that the TDOA-based GSS-system profits from more microphones although decent results can already be achieved with the recordings of four smartphones.

VII. Conclusions

In this contribution we have shown that spatial information in form of TDOA information is a powerful source for diarization information when using an ad-hoc ASN in a quite static scenario with spatially well separated speakers like a typical meeting. Thereby, the benefits of the spatial distribution predominate the challenges arising from the ad-hoc nature of the ASN, e.g., unknown microphone positions and asynchronous recordings. For gathering diarization information, we proposed to cluster TDOA estimates from a mutli-speaker TDOA estimator.

Experiments on real recordings have shown that source extraction via mask-based beamforming benefits from the derived diarization information and the TDOA estimates from which the diarization information is derived. On the one hand, a spatial mixture model, which utilizes the TDOA-based diarization as guide, outperforms a blind spatial mixture model with state-of-the-art initialization. On the other hand, a time-frequency bin wise initialization based on the TDOA estimates leads to a faster convergence of the spatial mixture model compared to a conventional time frame wise initialization scheme.

Acknowledgment

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

References

[1] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-End Processing for the CHiME-5 Dinner Party Scenario," in *CHiME5 Workshop*, 2018.

[2] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *Interspeech*, 2019, pp. 978–982.

[3] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *Interspeech*, 2020, pp. 274–278.

[4] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe, "Target-Speaker Voice Activity Detection with Improved i-Vector Estimation for Unknown Number of Speaker," in *Interspeech*, 2021, pp. 3555–3559.

[5] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 29–32.

[6] K. Ishiguro, T. Yamada, S. Araki, T. Nakatani, and H. Sawada, "Probabilistic speaker diarization with bag-of-words representations of speaker angle information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 447–460, 2012.

[7] M. Fakhry, N. Ito, S. Araki, and T. Nakatani, "Modeling audio directional statistics using a probabilistic spatial dictionary for speaker diarization in real meetings," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[8] N. Zheng, N. Li, J. Yu, C. Weng, D. Su, X. Liu, and H. Meng, "Multi-channel speaker diarization using spatial features for meetings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7337–7341.

[9] H. Taherian and D. Wang, "Multi-channel conversational speaker separation via neural diarization," *arXiv preprint arxiv.2311.08630*, 2023.

[10] T. Gburrek, C. Boeddeker, T. von Neumann, T. Cord-Landwehr, J. Schmalenstroeer, and R. Haeb-Umbach, "A meeting transcription system for an ad-hoc acoustic sensor network," *arXiv preprint arxiv.2205.00944*, 2022.

[11] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "Geometry calibration in wireless acoustic sensor networks utilizing DoA and distance information," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[12] ——, "On synchronization of wireless acoustic sensor networks in the presence of time-varying sampling rate offsets and speaker changes," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[13] J. Schmalenstroeer, T. Gburrek, and R. Haeb-Umbach, "Libriwasn: A data set for meeting separation, diarization, and recognition with asynchronous recording devices," in *ITG conference on Speech Communication (ITG)*, 2023.

[14] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[15] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[16] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1479–1489, 2008.

[17] P. Vijaya, M. Narasimha Murty, and D. Subramanian, "Leaders–subleaders: An efficient hierarchical clustering algorithm for large data sets," *Pattern Recognition Letters*, vol. 25, no. 4, pp. 505–513, 2004.

[18] J. C. Gower and G. J. S. Ross, "Minimum spanning trees and single linkage cluster analysis," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 18, no. 1, pp. 54–64, 1969.

[19] C. Boeddeker, T. Cord-Landwehr, T. von Neumann, and R. Haeb-Umbach, "An Initialization Scheme for Meeting Separation with Spatial Mixture Models," in *Interspeech 2022*, 2022, pp. 271–275.

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

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

[22] S. Haykin, *Adaptive Filter Theory - Fourth Edition*. Prentice-Hall, Information and system science series, 2002.

[23] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[24] B. Yang, H. Liu, C. Pang, and X. Li, "Multiple sound source counting and localization based on TF-wise spatial spectrum clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1241–1255, 2019.

[25] M. Souden, J. Benesty, and S. 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.

[26] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

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