Spatio-spectral diarization of meetings by combining TDOA-based segmentation and speaker embedding-based clustering

Tobias Cord-Landwehr, Tobias Gburrek*, Marc Deegen, Reinhold Haeb-Umbach*

Paderborn University, Communications Engineering Department, Germany

{cord,gburrek,deegen,haeb}@nt.upb.de

Abstract

We propose a spatio-spectral, combined model-based and data-driven diarization pipeline consisting of TDOA-based segmentation followed by embedding-based clustering. The proposed system requires neither access to multi-channel training data nor prior knowledge about the number or placement of microphones. It works for both a compact microphone array and distributed microphones, with minor adjustments. Due to its superior handling of overlapping speech during segmentation, the proposed pipeline significantly outperforms the single-channel pyannote approach, both in a scenario with a compact microphone array and in a setup with distributed microphones. Additionally, we show that, unlike fully spatial diarization pipelines, the proposed system can correctly track speakers when they change positions.

Index Terms: diarization, meeting data, spatial, spectral, spatio-spectral

1. Introduction

Diarization systems assign regions of speech activity to the individual participants of a conversation, thus answering the question "Who spoke when?". Essentially, they solve two tasks, segmentation and speaker assignment. The first is on identifying regions (segments) of constant speaker activity, while the second assigns speaker labels to each segment. There exists a large variety of methods for how these tasks are solved [1–5].

Here, we categorize diarization systems according to whether they use spectral or spatial cues, or both. Early diarization systems using spectral information employed statistical models [6], while recent systems rely on speaker embeddings, e.g., x-vectors or d-vectors [2, 7, 8], extracted from audio segments, which are then clustered. Alternatively, they are used to directly predict the speech activity of all participants in a conversation on a frame-by-frame basis as in End-to-End Neural Diarization (EEND) systems [3].

If multi-channel input is available, spatial cues have been shown to deliver strong diarization results [1, 9–11]. In particular, they excel over spectral systems in regions of overlapping speech [12, 13]. However, one should be aware of the fact that segments of speech activity are assigned to positions or directions in space, rather than to speakers, with the consequence that speaker movements or speaker position changes can confuse the system. Additionally, strong reflections can result in so-called phantom positions, indicating activity from a direction, where actually no speaker is present. It is also known that the quality of spatial cues depends on the inter-microphone distance, with reduced informativeness if this distance is small [14].

There are only few examples of systems that use both spectral and spatial cues. Multi-channel information is used as auxiliary input of an otherwise spectral diarization system in [15–17], leading to improved performance. However, this approach requires a dedicated training phase with multi-channel data. That this is a significant impediment became clear in the recently concluded NOTSOFAR-1 challenge, where the lack of in-domain training data was cited as the main reason why only few systems made explicit use of spatial information for diarization [18]. An example of a deeper integration of spectral and spatial information is the integrated model of [19].

In this work, we introduce an alternative spatio-spectral diarization system. It shares similarities with the well-known pyannote diarization system, which is a purely spectral system that consists of (temporally) local segmentation followed by embedding-based global clustering [5]. We propose to do segmentation with spatial features instead, using a model-based approach. With the local segmentation being strictly decoupled from the single-channel, embedding-based clustering stage, the proposed system does not require in-domain training data.

The spatial segmentation model is based on [13]. It employs Time Difference Of Arrivals (TDOAs) estimates to detect segments of speech activity for all active sources. Then, beamforming is applied to all segments with speech activity to enhance the target speaker, and suppress crosstalk in regions of overlapping speech. Next, a speaker embedding extractor is applied to the enhanced speech segments, and global clustering of embedding vectors is carried out to obtain the speaker assignments for all speech segments in the meeting. This spectral clustering stage diminishes the impact of phantom positions, because embedding vectors computed from a segment representing a strong reflection will exhibit strong similarity with the segment containing the direct path signal of that speaker, such that they will be merged during clustering. In this way, the advantages of both spectral and spatial processing are exploited, while mitigating their drawbacks: The spatial processing addresses noise and overlapping speech, while the spectral processing can cope with possible position changes of a speaker.

Unlike [13], which requires globally constant speaker positions, the proposed system requires a speech source to be not moving only for a single segment of speech. In contrast to [15], the system does not require multi-channel training data, and it is independent of the number of microphones and its geometric arrangement. In the experiments, we show that it delivers good results both for a compact microphone array and for distributed microphones, with minimal adjustment of parameters.

Section 2 describes the proposed spatio-spectral diarization pipeline, which is evaluated in Section 3 both in a distributed and compact microphone setup in terms of Diarization Error Rate (DER) and Word Error Rate (WER). Finally, Section 4

*These authors contributed equally.

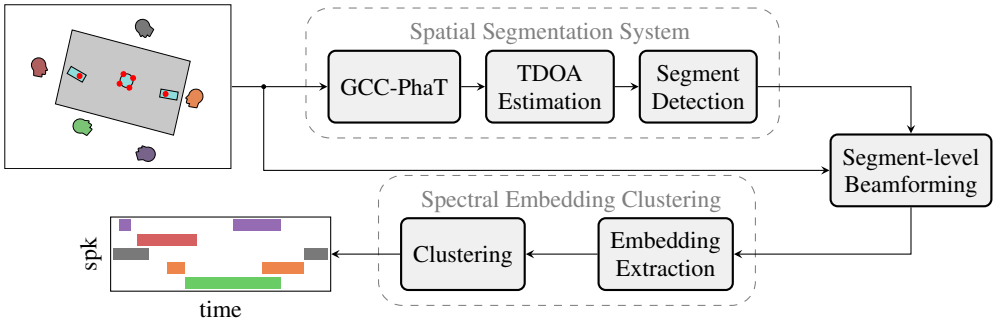

Figure 1: *Illustration of the proposed spatio-spectral diarization pipeline.*

offers some conclusions and an outlook on future work.

2. Spatio-spectral diarization pipeline

A multi-channel recording of a meeting with K speakers can be modeled in the time-frequency domain as the summation

$$y_c(t,f) = \sum_{l(k)} s_{l(k)}(t,f) h_{c,l(k)}(t,f) \tag{1}$$

of delayed, clean speech signals $s_{l(k)}(t,f)$ which are padded to match the length of the conversation. Here, $h_{c,l(k)}(t,f)$ denotes the acoustic transfer function between speaker k and microphone c for each speech segment $l(k)$ in time frame t and frequency bin f.

The proposed system cascades a TDOA-based local segment detection stage with a global spectral embedding-based clustering stage. Fig. 1 illustrates this pipeline.

2.1. Multi-source TDOA estimation

The TDOA $\tau_{i,j}$ of a signal between two microphones i and j for a single active sound source can be found as the position of the maximum of the Generalized Cross Correlation with Phase-Transform (GCC-PhaT) [20]

$$\tau_{i,j} = \arg\max \left(\mathrm{IFFT} \left(\frac{y_i(t,f) y_j(t,f)^*}{|y_i(t,f) y_j(t,f)^*|} \right) \right). \tag{2}$$

All pairwise TDOA estimates are gathered in a TDOA vector

$$\boldsymbol{\tau} = [\tau_{0,1}, \tau_{0,2}, \tau_{1,2}, \ldots, \tau_{C-1,C}]^{\mathrm{T}}, \tag{3}$$

where C is the total number of microphones. GCC-PhaT exhibits one local maximum per source. Since multiple sources can be active at the same time, P local maxima are chosen as possible delays τ_p, $1 \leq p \leq P$ [13]. Thus, $P^{\frac{(C-1)C}{2}}$ different TDOA vectors can be constructed from the estimated delays by combining all individual TDOAs, while only up to K vectors are physically grounded. To address this, a TDOA vector is kept only if the sum of the delays over each closed loop of microphones is close to zero, e.g.,

$$\tau_{0,1} + \tau_{1,2} + \tau_{2,0} < \tau_{\mathrm{th}},$$

where the threshold τ_{th} is set to a small value to account for numerical errors [21, 13]. By performing this TDOA estimation for each time frame, a set of TDOA vectors and corresponding frame indices of speech activity is estimated, which are next grouped into speech segments.

2.2. Temporally constrained segment detection

The speech segment detection is performed as in [13] by a temporally constrained leader-follower clustering. Here, the individual segments are determined by the pairwise Euclidean distance between all TDOA vectors. Two TDOA vectors $\boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_j$ can only belong to the same cluster if they do not exceed a maximal Euclidean distance $\Delta\tau_{\max}$, and if the temporal distance between frames t_i and t_j is smaller than 1 s. This allows individual segments to contain short regions of either silence or where no TDOA could be detected and prevents the formation of too large segments. Additionally, it prevents two consecutive segments from being merged since small segments are favored. After clustering, the detected segments $\hat{l}$ are specified by their start and end frames $t_{\mathrm{on},\hat{l}}$ and $t_{\mathrm{off},\hat{l}}$ as well as their median TDOA vector $\bar{\boldsymbol{\tau}}_{\hat{l}}$, while their respective speaker labels are yet unknown.

2.3. Segment-level beamforming

According to the W-disjoint orthogonality property of speech [22], each time-frequency bin (tf-bin) can be modeled to be either populated by a single source or by noise. This assumption underlying mask-based beamforming [23] is used to estimate binary masks to perform segment-wise beamforming.

First, for each processed segment the tf-bins containing only noise are estimated. This is done according to [24] via the eigenvalue gap of the tf-wise Spatial Covariance Matrix (SCM) estimates of the observation vector $\mathbf{y}(t,f) = (y_1(t,f), \ldots, y_C(t,f))^T$, which are gathered by averaging the outer product of the observation vector over a small local context. Since these tf-wise SCM estimates depict a dominant eigenvalue only for speech regions, bins are assigned to the noise mask if the eigenvalue gap between the first and second largest eigenvalue is below a threshold.

All remaining tf-bins are assigned to the mask of the processed segment or the interfering segments as follows. First, "prototype" SCMs are computed as the outer product of the steering vectors $\mathbf{a}_{\hat{l}}$ corresponding to each $\boldsymbol{\tau}_{\hat{l}}$

$$\boldsymbol{\Phi}_{\hat{l}} = \mathbf{a}_{\hat{l}} \mathbf{a}_{\hat{l}}^{\mathrm{H}}. \tag{4}$$

These prototypes are compared against the instantaneous matrix of pairwise phase terms of the observation

$$\boldsymbol{\Psi} = \frac{\mathbf{y}(t,f)\mathbf{y}(t,f)^{\mathrm{H}}}{|\mathbf{y}(t,f)\mathbf{y}(t,f)^{\mathrm{H}}|} \tag{5}$$

using the spatial covariance distance measure from [25]. The binary mask of each speech segment is now formed by those tf-bins, whose SCM is closest to the same prototype SCM.

The segment-forming process can also result in superfluous segments that are caused by phantom positions. Before beamforming, these segments need to be identified and removed to prevent using a speaker's own reflection as an interferer during beamforming. Reflections are characterized by the fact that its SCM shows a stronger deviation from the prototype SCM at high frequencies, which is caused by phase errors so that the corresponding binary masks are sparsely populated for higher frequencies. Therefore, the average mask activity between 150 Hz to 3500 Hz is compared against a threshold to determine whether the activity matches that of a speech signal. Segments containing too little activity are declared as caused by reflections and discarded.

Finally, the binary masks of the remaining segments need to be refined to fill up missing tf-bins that were assigned to discarded segments. This can either be done by repeating the mask estimation on the reduced set of segments, or by using a complex Angular Central Gaussian Mixture Model (cACGMM) for mask refinement as proposed in [13, 16]. In the latter approach, a statistical mixture model is fitted to the data, which is initialized with the binary masks. This refinement step is of low computational complexity because only few iterations are needed and because the model is applied to a single segment and not the whole meeting.

2.4. Embedding extraction & clustering

For each beamformed segment, the ResNet34-based d-vector model from [26] is employed to extract a speaker embedding. Then, HDBSCAN [27], a hierarchical, density-based clustering approach is applied to the embeddings. Here, similar speaker embeddings are grouped by the pairwise cosine distance between them. Additionally, HDBSCAN marks outliers. These outlier segments are merged into the most similar cluster. If two intersecting segments are assigned to the same cluster, the activity of both segments is merged. This allows merging phantom positions caused by reflections that were not detected in the beamforming stage. Since this step employs spectral information only, embeddings of a speaker changing their position can still be merged into the same cluster.

3. Experiments

3.1. Experimental Setup

For evaluation, the proposed pipeline is applied to the LibriCSS [28] and LibriWASN [14] data sets. LibriCSS consists of re-recordings of simulated LibriSpeech 8-speaker meetings ranging from 0 % to 40 % overlapping speech with a duration of 10 min. LibriWASN is an additional re-recording of the same synthetic meetings as LibriCSS, albeit in a distributed setup with multiple recording devices in two different rooms, exhibiting a T60 time of 200 ms (LibriWASN_{200}) and 800 ms (LibriWASN_{800}).

The diarization pipeline is applied to 4 microphone channels, which is the smallest possible number for TDOA-based source localization. In the compact setup, the 4-element microphone array *asnupb7* is used for LibriWASN, and 4 of the non-center microphones in LibriCSS. For a distributed setup, four smartphones of LibriWASN, the two *Pixel6*, one *Pixel7* and a *Xiaomi* device, are used, and all channels are assumed to be synchronized both in terms of Sampling Rate Offset (SRO) and Sampling Time Offset (STO).

The delay thresholds τ_{th} are set to 1 and 2 for the compact and distributed microphone setup, respectively, and the maximum delays during segment detection $\Delta\tau_{\max}$ are set to 1 and 0.75 samples, to accommodate for the very different inter-microphone distances. All remaining parameters are chosen independently of the scenarios, which encompass three different rooms and five different microphone setups.*

In addition to the DER as a performance measure, the transcription performance of the downstream ASR system from [29] is evaluated in terms of concatenated minimum-permutation Word Error Rate (cpWER). To this end, Guided Source Separation (GSS) [30] is applied as in [13] to extract the speech sources. For the ASR experiment, all 7 microphone channels of LibriCSS are used to be comparable with the literature, while only 4 channels are used for diarization. For calculating the DER according to [31], no forgiveness collar is used, and the cpWER is obtained using the *meeteval* toolkit [32].

3.2. Diarization performance

First, the proposed pipeline is evaluated w.r.t. its capability to be employed both in a distributed and a compact microphone setup. Table 1 shows that, in the distributed setup of LibriWASN_{200}, the proposed system can perform diarization equally well in overlap and in single-speaker regions, achieving an average DER of 3.78 % and of 4.19 % when only evaluating regions of overlapping speech. For the LibriWASN_{800} database, the system still can achieve similar average (DER_{avg}) and overlap-DERs (DER_{OV}) of 3.92 % and 5.28 %, respectively. Compared to other systems like [33] and [34] that try to address overlapping speech on a fully spectral level and achieve a DER_{OV} of 25 % to 30 % for single-channel processing, this underlines the advantage of dedicated multi-channel processing to handle overlapping speech in diarization.

When switching to a compact scenario, the total performance decreases by 1 % to 2 % absolute in terms of DER and WER, which is to be expected since the spatial cues used for TDOA estimation become less informative and speakers are harder to separate. Still, the system is able to consistently obtain similar DERs in single-speaker and overlap regions.

3.3. Comparison to other systems

Table 2 compares the spatio-spectral pipeline against other systems in the compact microphone setup, which is the more common application for multi-channel meeting processing. We compared with the embedding-based, overlap-aware diarization system from [33, 35] and the state-of-the-art, hybrid diarization and enhancement system SSND [15] on LibriCSS.

To have a comparison also for LibriWASN, we implemented the following spatial and spectral systems as references: a spatial-only pipeline directly clustering the median TDOA vectors of the detected segments using single-linkage agglomerative clustering with outlier rejection, and the single-channel pyannote 3.1 pipeline without any further modifications.

It can be seen that the proposed spatio-spectral system outperforms both a purely spectral and spatial approach. This shows that the proposed system effectively combines both systems' advantages. Here, the spectral system shows stable, but lower performance due to solely using single-channel information, while the spatial model shows higher errors due to a coarser resolution and reflections in the environment.

On LibriWASN, the proposed system even proves slightly better when omitting the cACGMM, demonstrating good performance even without additional segment refinement before

*github.com/fgnt/spatiospectral_diarization

Table 1: *DER and WER Performance of the proposed pipeline (with cACGMM refinement) for a distributed and compact setup.*

Setup	Database	0S	0L	OV10	OV20	OV30	OV40	DER_{avg}	DER_{OV}	cpWER
Distributed	$LibriWASN_{200}$	3.46	3.90	3.30	3.92	3.98	4.10	3.79	4.19	3.36
	$LibriWASN_{800}$	2.70	3.71	3.11	3.90	5.17	4.53	3.92	5.28	3.60
Compact	$LibriWASN_{200}$	3.52	3.81	5.34	4.93	5.57	6.89	5.16	7.00	5.13
	$LibriWASN_{800}$	3.08	4.59	4.24	4.99	6.38	6.09	5.00	7.08	5.50
	LibriCSS	5.87	5.90	5.90	7.46	8.16	8.79	7.17	9.97	6.53

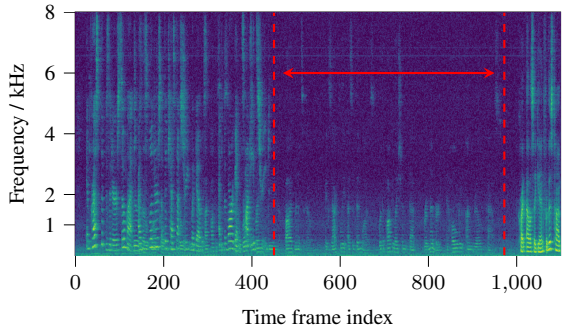

Figure 2: *Spectrogram of a LibriCSS segment with a single speech region of the defective loudspeaker (highlighted in red).*

Table 2: *Comparison of the proposed pipeline to other systems in a compact microphone setup.*

System	$LibriWASN_{200}$		$LibriWASN_{800}$		LibriCSS	
	DER	WER	DER	WER	DER	WER
SC + OV [35]	–	–	–	–	11.3	12.1
SSND [15]	–	–	–	–	4.7	5.1
pyannote[†]	12.8	8.8	12.6	11.6	14.1	10.2
Spatial	11.5	4.6	12.3	6.5	15.8	9.8
Proposed	4.5	4.1	5.2	5.1	9.9	8.9
+ cACGMM	5.2	5.1	5.0	5.5	7.2	6.5

beamforming. However, for LibriCSS, which is comparable in difficulty and acoustic properties to $LibriWASN_{200}$, unexpectedly high error rates occur without the cACGMM refinement.

After a closer analysis, these errors could be traced back to a single loudspeaker used during the recordings exhibiting a low-pass characteristic. This loudspeaker significantly attenuates frequencies above 1.5 kHz, as can be seen in Fig. 2. Therefore, the filtering stage aimed at identifying reflections through the energy distribution of a speech signal inadvertently removes this loudspeaker's signal before embedding extraction. The cACGMM-based mask refinement mitigates this effect and fills in gaps in the estimated speech activity, but cannot compensate for completely missed segments. With refinement, the proposed system approaches the results of the state-of-the-art SSND system [15] on LibriCSS. While SSND also employs a spatio-spectral approach, it employs a fully data-driven model to cascade diarization and separation. Due to the transformer-based multi-channel EEND network used for diarization, it requires matching training data on the corresponding microphone array geometry. Compared to it, the proposed model only requires training on VoxCeleb [36] for the embedding extractor and is agnostic to microphone placement and geometry.

[†]Applied to the first microphone channel using pyannote 3.1 [5]

Table 3: *Performance of the proposed system in the "semi-static" setup. "pm" denotes the cpWER on the individual meetings, "chg" on the concatenated meetings.*

	System	$LibriWASN_{200}$		$LibriWASN_{800}$		LibriCSS	
		pm	chg	pm	chg	pm	chg
Dist.	Spatial	3.1	24.7	4.5	27.6	–	–
	Proposed	3.4	3.3	3.6	4.0	–	–
Compact	Spatial	4.6	74.2	6.5	73.7	9.8	29.8
	pyannote [†]	8.8	10.2	11.6	11.1	10.2	12.4
	Proposed	5.1	5.3	5.5	5.6	6.5	6.5

3.4. Spatio-spectral diarization for changing positions

So far, all evaluations consider a static scenario with constant speaker positions. To verify the proposed, spatio-spectral pipeline's robustness against position changes, a "semi-static" LibriCSS and LibriWASN scenario is created as follows. Two meetings are concatenated so that each overlap subset (0S – OV40) consists of 20 min long meetings, where the speaker positions change after the first half of the meeting. Here, on average five out of the eight speakers are replaced by new speakers, while three speakers change location, resulting in meetings with 12-15 different speakers but only 8 speaker positions. This allows checking the system's performance both in case of different sources from the same location and speakers who change their position during the meeting.

Table 3 shows that the proposed system, similar to the fully spectral pyannote pipeline, only marginally degrades when switching from a per-meeting evaluation (*pm*) to the semi-static scenario (*chg*). Compared to this, the spatial-only system, as expected, cannot handle this scenario, since neither speakers changing their position nor multiple speakers from the same position can be accurately detected solely through their TDOAs.

4. Conclusion

In this work, we presented an approach that combines spatial segmentation with a spectral, embedding-based clustering model for the diarization of meetings without requiring in-domain training data. The proposed model can be deployed in compact and distributed microphone setups without large performance differences and with only minimal parameter changes. It was shown to robustly handle regions of overlapping speech and speaker position changes. Because the spatial subsystem is model-based instead of data-driven, it does not require in-domain training data and prior knowledge about the microphone configuration.

In future work, we will focus on completely integrating the speech enhancement stage into the segment-level beamforming of the spatio-spectral pipeline and extend the TDOA segmentation to handle continuous speaker movements.

5. Acknowledgements

Computational Resources were provided by BMBF/NHR/PC2.

6. References

[1] 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 (HSCMA)*, 2008, pp. 29–32.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5329–5333.

[3] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022.

[4] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. ISCA Interspeech*, 2020, pp. 274–278.

[5] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. ISCA Interspeech*, 2023.

[6] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[7] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang *et al.*, "Deep speaker: An end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[8] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. ISCA Interspeech*, 2020, pp. 3830–3834.

[9] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[10] 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, 2011.

[11] 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, pp. 1–5.

[12] J. Wang, Y. Liu, B. Wang, Y. Zhi, S. Li *et al.*, "Spatial-aware speaker diarization for multi-channel multi-party meeting," in *Proc. ISCA Interspeech*, 2022.

[13] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "Spatial diarization for meeting transcription with ad-hoc acoustic sensor networks," in *57th Asilomar Conference on Signals, Systems, and Computers*, 2023, pp. 1399–1403.

[14] 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*, Sep 2023.

[15] H. Taherian and D. Wang, "Multi-channel conversational speaker separation via neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[16] R. Wang, S. Niu, G. Yang, J. Du, S. Qian *et al.*, "Incorporating spatial cues in modular speaker diarization for multi-channel multi-party meetings," *arXiv preprint arXiv:2409.16803*, 2024.

[17] N. Zheng, N. Li, J. Yu, C. Weng, D. Su *et al.*, "Multi-channel speaker diarization using spatial features for meetings," in *Proc. IEEE ICASSP*, 2022, pp. 7337–7341.

[18] I. Abramovski, A. Vinnikov, S. Shaer, N. Kanda, X. Wang *et al.*, "Summary of the NOTSOFAR-1 challenge: Highlights and learnings," *arXiv preprint arXiv:2501.17304*, 2025.

[19] T. Cord-Landwehr, C. Boeddeker, and R. Haeb-Umbach, "Simultaneous diarization and separation of meetings through the integration of statistical mixture models," *arXiv preprint arXiv:2410.21455*, 2024.

[20] C. Knapp and G. 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.

[21] 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.

[22] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE ICASSP*, vol. 1, 2002, pp. I–529–I–532.

[23] 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.

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

[26] C. Boeddeker, T. Cord-Landwehr, and R. Haeb-Umbach, "Once more diarization: Improving meeting transcription systems through segment-level speaker reassignment," in *Proc. ISCA Interspeech*, 2024, pp. 1615–1619.

[27] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Trans. Knowl. Discov. Data*, 2015.

[28] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng *et al.*, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE ICASSP*, 2020, pp. 7284–7288.

[29] S. Watanabe, "ESPnet2 pretrained model, Shinji Watanabe/librispe ech_asr_train_asr_transformer_e18_raw_bpe_sp_valid .acc.best, fs=16k, lang=en," Jul. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3966501

[30] C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann *et al.*, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. 5th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2018, pp. 35–40.

[31] O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "NIST 2021 speaker recognition evaluation plan," 2021-07-12 04:07:00 2021.

[32] T. von Neumann, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "MeetEval: A toolkit for computation of word error rates for meeting transcription systems," in *Proc. 7th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2023, pp. 27–32.

[33] D. Raj, Z. Huang, and S. Khudanpur, "Multi-class spectral clustering with overlaps for speaker diarization," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 582–589.

[34] T. Cord-Landwehr, C. Boeddeker, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, "Geodesic interpolation of frame-wise speaker embeddings for the diarization of meeting scenarios," in *Proc. IEEE ICASSP*, 2024, pp. 11 886–11 890.

[35] D. Raj, D. Povey, and S. Khudanpur, "GPU-accelerated guided source separation for meeting transcription," in *Proc. ISCA Interspeech*, 2023, pp. 3507–3511.

[36] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.