Are Identical Sounds Present in Distributed Recordings to Serve as Spatio-Temporal Anchors? A Case Study Using the SINS Database

Takao Kawamura* and Nobutaka Ono*
* Tokyo Metropolitan University, Japan
E-mail: kawamura-takao@ed.tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—In this study, we explore whether distributed recordings, captured using devices equipped with internally synchronized microphones, contain identical sounds that could provide a reference for spatial and temporal alignment. We define anchor sounds as those originating from the same location with the same waveform, such as a TV power-on sound or a door-closing sound. To examine their presence and potential usability, we test a two-stage detection approach that incorporates widely used signal processing techniques, including cross-correlation, to evaluate waveform similarity and time difference of arrival (TDOA) consistency. Our experiments on the SINS database confirm that such identical sounds are present and can be identified using conventional signal processing techniques. While our findings are specific to the SINS database, they suggest similar anchor sounds may also exist in other real-world datasets, potentially enabling applications such as microphone self-localization and synchronization.**

I. Introduction

In smart home environments, monitoring human activities as part of daily routines has become increasingly important. Several studies have focused on acoustic scene analysis [1], [2] and sound source localization [3], [4]. Generally, utilizing microphone arrays is desirable for obtaining spatial information [5]–[9]. The use of distributed microphone arrays is particularly advantageous for capturing large spatial areas and has been studied [7], [8], [10]–[14]. Recently, the number of smartphones, smart speakers, and IoT devices equipped with multiple microphones has rapidly increased. Consequently, research has focused on distributed microphone arrays composed of multiple internally synchronized subarrays (e.g., small microphone arrays within each device) [14].

In distributed microphone arrays, utilizing spatial information requires knowledge of microphone positions and synchronization of microphone recordings. Conventional methods address this through self-calibration [15]–[17] and synchronization techniques [18]–[23]. In these approaches, dedicated calibration signals are often effective [15], [24], [25]. Such signals provide a high signal-to-noise ratio (SNR) and clear time alignment, but playing them in daily-life environments is often impractical or intrusive.

To address this issue in a more practical manner, several studies have explored the use of reliable signal segments directly from recordings, without requiring dedicated reference signals. Among these, signal segments spoken by a single speaker have been effectively utilized in tasks such as direction of arrival (DOA) estimation [26]. In the context of synchronization, Sakanashi *et al.* [18] estimated the sampling rate offset (SRO) between devices by using manually selected segments spoken by a single speaker. Several methods have also been proposed for synchronization based on automatically detected signal segments with consistent acoustic characteristics across recordings [27], [28].

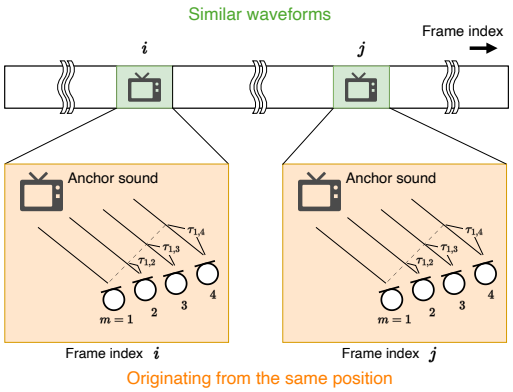

Fig. 1. Definition of anchor sound. The anchor sound consistently originates from the same position with an same waveform.

Inspired by these approaches, we pose a fundamental question from a different perspective: *Are identical sounds present in distributed recordings to serve as spatio-temporal anchors?* In everyday environments, certain naturally occurring sounds, such as a TV power-on sound or a door-closing sound, are expected to originate from the same location and exhibit similar waveforms each time they occur. We refer to such sounds as "anchors," defined as segments that consistently originate from the same position and share similar waveforms across time. This concept is illustrated in Fig. 1. These anchors provide stable time differences of arrival (TDOAs) and DOAs, which can be utilized for synchronization and self-localization in distributed microphone arrays.

In this study, we investigate whether such sounds are present in real-world multi-channel recordings. To detect these sounds, we focus on internally synchronized subarrays and employ a two-stage detection approach based on waveform similarity and TDOA consistency within each subarray. We apply this approach to the SINS database [13] and confirm that

979-8-3315-7206-8/25/$31.00

such anchor sounds, including TV power-on and door-closing sounds, are present in real-world recordings. We also discuss the potential usefulness of these anchor sounds for microphone synchronization by analyzing the stability of time differences between subarrays.

II. PROBLEM SETTING

We consider the problem of detecting an anchor sound using signals recorded from a subarray with M microphones. We assume that the microphones within the subarray are closely located and synchronized. We divide the recorded signal into short signals as frames. In this study, we assume each frame contains at most one anchor sound and formulate the problem of detecting frames that contain anchor sound. The following equation represents the recorded signal of the mth microphone at frame i:

$$x_{m,i}[t] = s[t - \xi_{m,i}] * h_m[t] + n_{m,i}[t], \tag{1}$$

where $\xi_{m,i}$, $h_m[t]$, and $n_{m,i}[t]$ indicate a time position of the anchor sound $s[t]$ in the frame i, an impulse response from sound source to microphone, and noise signal, respectively.

III. BLIND DETECTION OF ANCHOR SOUNDS

The detection method consists of two stages. In the first stage, we identify the temporal frames where the peak of the cross-correlation between different frames exceeds a threshold, thereby detecting signals with similar waveforms (see Sec. III-A). In the second stage, we calculate TDOAs between microphones within a subarray and detect frames where these TDOAs coincide, indicating sounds likely originating from the same position (see Sec. III-B). The detected frames are grouped according to the pair information (see Sec. III-C).

A. Detection of Frame Pairs Based on Waveform Similarity

In the first stage, we detect pairs of frames with similar waveforms. Since microphones within each subarray are closely located, we select a representative microphone m from the subarray to perform the detection. If the anchor sounds are contained in both frames, the peak of the normalized cross-correlation between the frames is expected to be high. The normalized cross-correlation can be calculated as

$$\phi_{i,j}(\tau) = \mathcal{F}^{-1}_{f\to\tau}\left(\frac{X_{m,i}[f]X^*_{m,j}[f]}{\sqrt{\sum_f |X_{m,i}[f]|^2}\sqrt{\sum_f |X_{m,j}[f]|^2}}\right), \tag{2}$$

where $\mathcal{F}^{-1}_{f\to\tau}(\cdot)$ is the inverse Fourier transform from f to τ, $X_{m,i}[f]$ is the Fourier transform of $x_{m,i}[t]$ in Eq. (1), and $\{\cdot\}^*$ denotes the complex conjugate. Here, the denominator works to adjust the scale of the cross-correlation. The pairs of frames are detected as candidates containing anchor sounds when the peak value of the cross-correlation function $\phi_{i,j}(\tau)$ exceeds a threshold θ. The set of the detected pairs are obtained as

$$P_1 = \{(i,j) | \max_\tau \phi_{i,j}(\tau) > \theta\}. \tag{3}$$

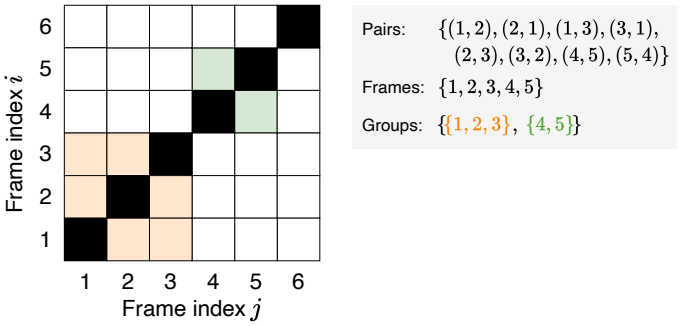

Fig. 2. Example of detected results. Colored off-diagonal components indicate detected pairs (i, j). Here, the numbers of pairs, frames, and groups are 8, 5, and 2, respectively.

B. Detection of Frame Pairs Based on TDOA Consistency

In the second stage, from the detected pairs at the first stage, we select those with matching TDOAs. The TDOA of the ith frame between the microphone m and n within the same subarray can be calculated as

$$\hat{\tau}_{m,n,i} = \operatorname*{argmax}_\tau \quad \psi_{m,n,i}(\tau), \tag{4}$$

$$\psi_{m,n,i}(\tau) = \mathcal{F}^{-1}_{f\to\tau}\left(\frac{X_{m,i}[f]X^*_{n,i}[f]}{\sqrt{\sum_f |X_{m,i}[f]|^2}\sqrt{\sum_f |X_{n,i}[f]|^2}}\right). \tag{5}$$

Without loss of generality, we use the first microphone as the reference microphone and calculate the TDOA vectors for the ith frame as

$$\tau_i = (\hat{\tau}_{1,2,i}, \cdots, \hat{\tau}_{1,M,i}). \tag{6}$$

The pairs of frames are detected when the TDOA vectors τ_i and τ_j are coincided. The set of the detected pairs are obtained as

$$P_2 = \{(i,j) |\ \tau_i = \tau_j, (i,j) \in P_1\}. \tag{7}$$

Note that we assume $\hat{\tau}_{m,n,i}$ is obtained in discrete time and takes an integer value. Therefore, $\tau_i = \tau_j$ indicates that the TDOAs coincide with sample-level precision.

Furthermore, because the method uses only TDOA information, it remains effective regardless of the specific array geometry or the number of microphones, provided that the microphone configuration is fixed.

C. Grouping the Same Kind of Anchor Sounds

To simply detect the frames containing anchor sounds, it is sufficient to extract all frame indices included in P_2. However, in this study, we also perform grouping of frames to classify their types of anchor sounds roughly. If two frames are detected as a pair in P_2, they are similar and arriving from the same direction, thus considered as the same type of anchor sound. To group these sounds, we consider a graph where frames are nodes and pair relationships are edges. We then detect disconnected subgraphs to form the groups. Figure 2 shows an example of detected results by the detection method.

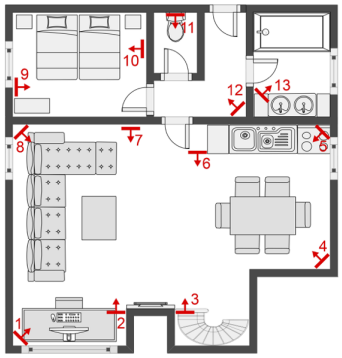

Fig. 3. Arrangement of subarrays in SINS database (figure from [13]). Red arrows indicate directions of subarrays containing four microphones, and the number indicates the subarray index.

TABLE I
NUMPBER OF DETECTED SOUNDS. "# PAIR," "# FRAME," "# GROUP" INDICATE THE NUMBER OF DETECTED PAIRS, FRAMES, AND GROUPS, RESPECTIVELY.

	Subarray index						
	1	2	3	4	6	7	8
# Pair	697	1178	1156	993	1023	1312	697
# Frame	193	193	277	282	238	319	162
# Group	33	25	31	31	54	42	29

IV. EXPERIMENTS

In this experiment, we applied the detection method to one of the real-world datasets (see Sec. IV-B). Then, we observed the examples of detected sounds (see Sec. IV-C) and discussed the potential usefulness of these sounds (see Sec. IV-D).

A. Experimental Setup

We used the SINS database [13], which is real-world recordings collected by 13 subarrays that consist of four microphones (see Fig. 3). In the SINS database, along with each audio recording, internal counter (timer) values from the control board are recorded every second. The subarrays are synchronized using these values. We used the recorded signal labeled as "Other," which contains a transition between activities and is expected to contain various sounds. The total time of the recorded signal was 145 minutes. The sampling rate was $16,000$ Hz. The recorded signals were segmented into 2-second frames without overlap, resulting in a total of $4,360$ frames. The number of pairs considered using brute force is $4,360^2$. We applied the detection methods to seven subarrays in the living room (1 through 8, excluding subarray 5, which was unavailable). The peak threshold for the detection process in Sec. III-A was set to a value that detected 3000 pairs in each subarray.

B. Results of Anchor Sound Detection

We counted the number of detected anchor sounds. Table I shows the number of detected pairs, frames, and groups using the two-stage detection method. The number of pairs, frames, and groups are $|P_2|$, unique frame indices included in P_2, and the number of subgraphs. The detection method detected several hundred frames and approximately 30 groups. We confirmed that in subarrays 1 and 8, which were located in the corners of the living room, the number of detected pairs was lower than in other subarrays.

TABLE II
SUBJECTIVE LABELED RESULTS AT SUBARRAY 7.

Detected sounds	# Group	# Frame
Stationary noise	28	190
Door-closing sounds	8	104
Notification sounds	2	12
TV power-on sounds	2	7
Harmonic noise	1	2
Beating-something sounds	1	4

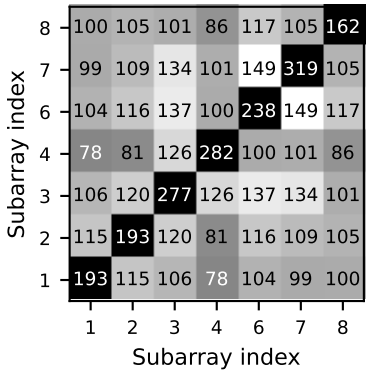

Fig. 4. Co-occurrence of detection on the subarrays. The diagonal elements represent the number of frames detected in each subarray. The total number of frames detected in all subarrays was 57 (labeled "Door-closing sound").

To observe the detected sounds, they were subjectively labeled by one person. The subjective labeling was performed on the detected sounds in subarray 7 (42 groups). Table II shows the subjectively labeled results. "Stationary noise" was background noise, primarily consisting of low-power sounds. "Harmonic noise" was noise with a harmonic structure. "Door-closing sounds" and "Beating-something sounds" were impulsive sounds. "TV power-on sounds" were short music that contained harmonic components. "Notification sounds" were short sounds that contained harmonic components. The detection method detected various sounds, including impulsive and harmonic sounds. These results suggest the possibility of detecting additional anchor sounds when exploring other scene labels (e.g., "Cooking" and "Working") or other real-world datasets [10]–[12].

We also investigated the co-occurrence of detections across subarrays. Figure 4 illustrates the co-occurrence of detections across the subarrays. Here, co-occurrence denotes the count of frame indices that overlap among detections in each subarray. The results indicate that anchor sounds were detected simultaneously across all subarrays, with the number of pairwise co-occurrences exceeding approximately 80. A recent method estimates the TDOA or sampling rate offsets (SROs) using a pairwise approach and computes their consensus [23], [29]. These results suggest that these methods are applicable.

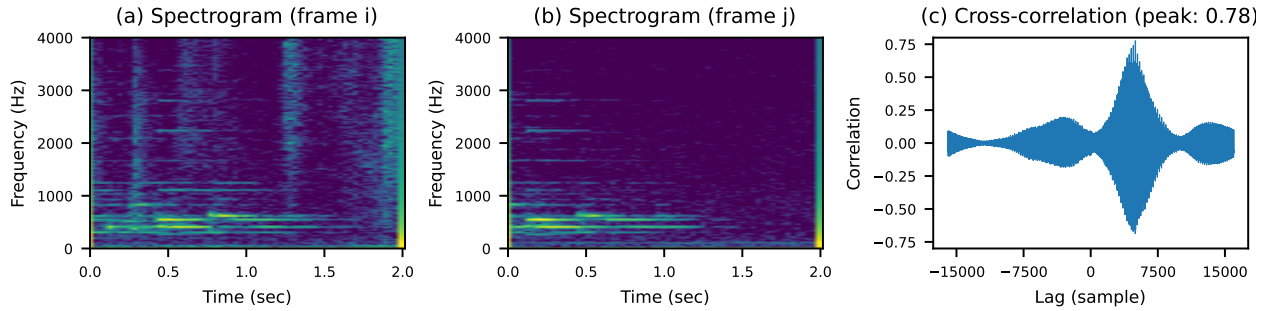

Fig. 5. Example of "TV power-on sounds" at Table II (subarray 7). (a) and (b) show spectrograms ranging from 0 Hz to 4000 Hz of frame i and j, respectively. (c) shows the normalized cross-correlation function $\phi_{i,j}(\tau)$. The time difference vector was $[0, 0, 0]$.

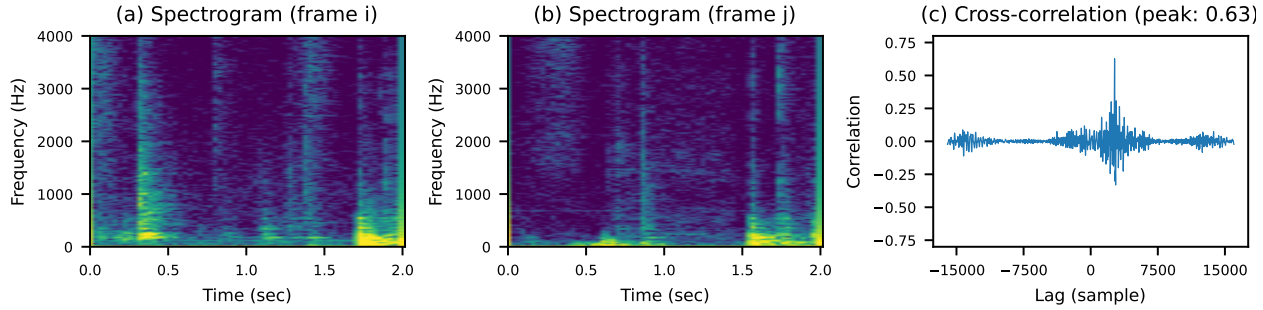

Fig. 6. Example of "Door-closing sounds" at Table II (subarray 7). (a) and (b) show spectrograms ranging from 0 Hz to 4000 Hz of frame i and j, respectively. (c) shows the normalized cross-correlation function $\phi_{i,j}(\tau)$. The time difference vector was $[-1, -2, -3]$.

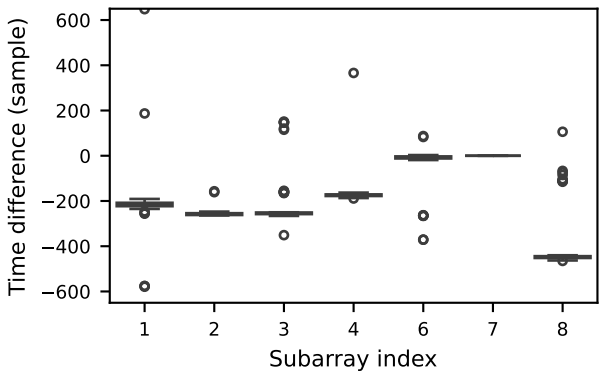

Fig. 7. Time difference between subarray 7 and all subarrays (subarray 1, 2, ..., 8) for the label "Door-closing sounds" (57 frames).

C. Examples of Anchor Sounds

Figure 5 and 6 show examples of anchor sounds detected by the detection method. These frames were selected from the categories "TV power-on sounds" and "Door-closing sounds" as listed in Table II. Figure 5 confirms the harmonic components associated with TV power-on sounds. Figure 6 displays the impulse component of a door-closing sound. In both examples, peaks appeared in the cross-correlation function, with peak values approximately around 0.7. We also observed TDOA vectors. The TDOA vectors were $[0, 0, 0]$ and $[-1, -2, -3]$ for the "TV power-on sound" and "Door-closing sound," respectively. This suggests that the TV may be in front of subarray 7, and the door may be located adjacent to subarray 7 (see Fig. 3). These results suggest that the TDOA vectors of anchor sounds may provide spatial information.

D. Evaluation of Inter-Array Time Differences for Potential Passive Synchronization

We examined the stability of inter-array time differences for anchor sounds to assess their potential for synchronization. In our investigations, the only detected anchor sounds across all subarrays were door-closing sounds, totaling 57 frames. We calculated the time differences between subarray 7 and all other subarrays (subarray 1, 2, ..., 8) for these 57 frames to analyze their stability. Figure 7 presents a boxplot where the horizontal and vertical axes represent the subarray index and inter-array time difference, respectively. Since the SINS database ensures subarray synchronization, we expect the inter-array time differences to remain constant if the detected sounds originate from the same location. Although a few outliers were observed, possibly resulting from incorrect detection of a spurious peak in the normalized cross-correlation function, the results confirmed that the observed time differences were highly stable, supporting the reliability of anchor sounds in synchronized environments. This finding suggests that if the subarrays were not synchronized, identifying such anchor sounds and adjusting inter-array delays to maintain their consistency could provide a means of passive synchronization. In fact, Sakanashi *et al.* [18] demonstrated that if signal segments from the same sound source are observed twice, SRO between devices can be estimated from the ratio of the time intervals between the two segments on each device. Their method was based on the assumption that such segments were manually specified by the user. In contrast, our approach enables blind detection of anchor sounds with similar properties, making it effective for SRO estimation. This allows their SRO estimation technique to be automatically applied in practical systems without manual intervention.

V. CONCLUSION

In this study, we explored anchor sounds that provide spatial and temporal alignment cues in distributed recordings. We defined anchor sounds as those originating from the same position with the same waveform. A two-stage detection process was applied to short frames. Firstly, we identified temporal frame pairs where the peak of cross-correlation between different frames exceeded a threshold, thereby detecting signals with similar waveforms. In the second stage, we calculated time differences between microphones within a subarray and detected frame pairs where these time differences coincided, indicating sounds likely originating from the same position. The detected pair information was then used to assemble the relevant groups of anchor sounds. In our experiments, we applied the detection method to the SINS database and confirmed that six types of sounds, including TV power-on sounds, impulsive sounds such as door-closing sounds, and noise, were present in the real recordings. We also analyzed the stability of inter-array time differences of these sounds to explore their potential usefulness for microphone synchronization. The observed stability suggests that such anchor sounds could contribute to passive synchronization methods. While our findings are specific to the SINS database, these results suggested the possibility of detecting other anchor sounds when exploring other scene labels or other real-world datasets.

Future work will focus on refining detection performance and expanding the analysis to additional datasets to further evaluate the role of anchor sounds in distributed microphone arrays.

ACKNOWLEDGMENT

This work was supported by JST SICORP Grant Number JPMJSC2306 and JSPS KAKENHI Grant Number JP24KJ1866.

REFERENCES

[1] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, vol. 11, pp. 106 620–106 649, 2023. DOI: 10.1109/ACCESS.2023.3318015.

[2] B. Ding, T. Zhang, C. Wang, *et al.*, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121 902, 2024. DOI: 10.1016/j.eswa.2023.121902.

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Berlin Heidelberg, 2001, pp. 157–180. DOI: 10.1007/978-3-662-04619-7_8.

[4] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–24, 2017. DOI: 10.1155/2017/3956282.

[5] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019, pp. 30–34.

[6] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "On the effectiveness of spatial and multi-channel features for multi-channel polyphonic sound event detection," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020, pp. 115–119.

[7] P. Giannoulis, G. Potamianos, and P. Maragos, "Room-localized speech activity detection in multi-microphone smart homes," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, p. 15, 1 2019. DOI: 10.1186/s13636-019-0158-8.

[8] T. Kawamura, Y. Kinoshita, N. Ono, and R. Scheibler, "Effectiveness of inter- and intra-subarray spatial features for acoustic scene classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096935.

[9] X. Jiang, C. Han, Y. A. Li, and N. Mesgarani, "Exploring self-supervised contrastive learning of spatial sound event representation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1281–1285. DOI: 10.1109/ICASSP48485.2024.10447391.

[10] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans*. Springer Berlin Heidelberg, 2007, pp. 311–322. DOI: 10.1007/978-3-540-69568-4_29.

[11] V. Libal, B. Ramabhadran, N. Mana, *et al.*, "Multimodal classification of activities of daily living inside smart homes," in *Proc. International Work-Conference on Artificial Neural Networks (IWANN)*, 2009, pp. 687–694. DOI: 10.1007/978-3-642-02481-8_103.

[12] K. Imoto and N. Ono, "RU multichannel domestic acoustic scenes 2019: A multichannel dataset recorded by distributed microphones with various properties," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019, pp. 104–108.

[13] G. Dekkers, S. Lauwereins, B. Thoen, *et al.*, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2017, pp. 32–36.

[14] K. Imoto, "Graph cepstrum: Spatial feature extracted from partially connected microphones," *IEICE Transactions on Information and Systems*, vol. E103.D, pp. 631–638, 3 2020. DOI: 10.1587/transinf.2019EDP7162.

[15] V. Raykar, I. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Transactions on Speech*

and Audio Processing, vol. 13, pp. 70–83, 1 2005. DOI: 10.1109/TSA.2004.838540.

[16] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 106–110. DOI: 10.1109/ICASSP.2013.6637618.

[17] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of ad-hoc arrays using time difference of arrivals," *IEEE Transactions on Signal Processing*, vol. 64, pp. 1018–1033, 4 2016. DOI: 10.1109/TSP.2015.2498130.

[18] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Speech enhancement with ad-hoc microphone array using single source activity," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2013, pp. 1–6. DOI: 10.1109/APSIPA.2013.6694323.

[19] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Processing*, vol. 107, pp. 185–196, 2015. DOI: 10.1016/j.sigpro.2014.09.015.

[20] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 651–661, 3 2017. DOI: 10.1109/TASLP.2017.2655259.

[21] S. Wozniak and K. Kowalczyk, "Passive joint localization and synchronization of distributed microphone arrays," *IEEE Signal Processing Letters*, vol. 26, pp. 292–296, 2 2019. DOI: 10.1109/LSP.2018.2889438.

[22] A. Chinaev, P. Thuene, and G. Enzner, "Double-cross-correlation processing for blind sampling-rate and time-offset estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1881–1896, 2021. DOI: 10.1109/TASLP.2021.3071967.

[23] D. Hu, H. Zhang, F. Bao, and R. Wang, "Distributed sampling rate offset estimation over acoustic sensor networks based on asynchronous network newton optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 301–312, 2023. DOI: 10.1109/TASLP.2022.3224256.

[24] N. Ono, K. Shibata, and H. Kameoka, "Self-localization and channel synchronization of smartphone arrays using sound emissions," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2016, pp. 1–5. DOI: 10.1109/APSIPA.2016.7820778.

[25] A. Kovalyov, K. Patel, and I. Panahi, "Joint calibration and synchronization of two arrays of microphones and loudspeakers using particle swarm optimization," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 535–544, 2021. DOI: 10.1109/OJSP.2021.3118574.

[26] A. Schwartz, O. Schwartz, S. E. Chazan, and S. Gannot, "Multi-microphone simultaneous speakers detection and localization of multi-sources for separation and noise reduction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 50, 2024. DOI: 10.1186/s13636-024-00365-3.

[27] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Estimation of sampling frequency mismatch between distributed asynchronous microphones under existence of source movements with stationary time periods detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 785–789. DOI: 10.1109/ICASSP.2019.8683192.

[28] A. Chinaev, N. Knaepper, and G. Enzner, "Long-term synchronization of wireless acoustic sensor networks with nonpersistent acoustic activity using coherence state," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095792.

[29] K. Yamaoka, T. Nakashima, Y. Wakabayashi, and N. Ono, "Minimum-spanning-tree-based time delay estimation robust to outliers," *IEEE Access*, vol. 11, pp. 121 284–121 294, 2023. DOI: 10.1109/ACCESS.2023.3327011.