published in: *European Conference on Signal Processing (EUSIPCO) 2014*

GEOMETRY CALIBRATION OF DISTRIBUTED MICROPHONE ARRAYS EXPLOITING AUDIO-VISUAL CORRESPONDENCES

Axel Plinge and Gernot A. Fink

Department of Computer Science, TU Dortmund University, Dortmund, Germany

ABSTRACT

Smart rooms are used for a growing number of practical applications. They are often equipped with microphones and cameras allowing acoustic and visual tracking of persons. For that, the geometry of the sensors has to be calibrated. In this paper, a method is introduced that calibrates the microphone arrays by using the visual localization of a speaker at a small number of fixed positions. By matching the positions to the direction of arrival (DoA) estimates of the microphone arrays, their absolute position and orientation are derived. Data from a reverberant smart room is used to show that the proposed method can estimate the absolute geometry with about 0.1 m and 2° precision. The calibration is good enough for acoustic and multi modal tracking applications and eliminates the need for dedicated calibration measures by using the tracking data itself.

Index Terms— microphone array, distributed sensor network; geometry calibration; speaker tracking

1. INTRODUCTION

Smart rooms are increasingly becoming part of our world. Their applications include online lectures and video conferencing [1]. For automated camera control, acoustic [2], visual [3] or multi-modal person detection and tracking [4] is required. For meetings, automated annotation, context information, and speaker diarization [5] can be provided. With the numerous applications of this new technologies, the quest for methods to make them reliable and easy to use is becoming more pressing. Steps along the way include the robust detection and localization of speakers [2] and the reliable classification of acoustic events [6]. Another issue is the geometrical calibration of the sensors. While cameras in a smart room are often installed at fixed known locations, microphone arrays can be set up in an ad hoc fashion.

Several existing geometry calibration methods used absolute time of arrival (ToA) measurements produced by playing calibration sounds at each sensor node. The distance can then be computed assuming the speed of sound. With given distances of all microphones to a number of base points, the geometry can be calculated using multidimensional scaling [7]. This was used with consumer devices like laptops [8] and smartphones [9] that all have at least one speaker and at least one microphone in known relative distance.

When the acoustic sensor nodes are not equipped with speakers, the geometry has to be inferred from unknown source positions not aligned with the sensors. If there is strict time synchronization between the sensors, the time difference of arrival (TDoA) between pairs of sensors can be measured allowing to compute the relative distance to the source. From all these estimates, the geometry can be inferred. Dedicated signals such as white noise or sweep chirps played with a speaker allow good TDoA estimation. To avoid the necessity of an artificial speaker, hand claps can be employed [10]. Assuming the direct path is the shortest to all microphones, the TDoA can be estimated using onset detection [11].

In passive estimation, only speech events may be used. Here the estimation of TDoAs between the sensor nodes may not be reliable if the reverberation in the room is strong. When the sensor nodes are equipped with small microphone arrays, the direction of arrival (DoA) at each node can be computed. Using speech of a single moving person the relative geometry of the nodes can be computed using the random sampling consensus (RANSAC) method [12]. The scaling may not be reliably estimated with only DoAs, but can be estimated in a second step using TDoA estimates [13].

All the above methods require strict time synchronization between the sensor nodes. This imposes an additional requirement that is not needed for acoustic speaker tracking [2]. They also provide only means of relative geometry estimation. For integration with video data and multi modal tracking, additional matching of the two modalities is required. In this paper, a method is introduced that overcomes these shortcomings. The DoA of the speaker at each acoustic sensor node can be estimated by a robust method [14]. Assuming the positions of the cameras are known, visual localization can produce Euclidean tracking of a speaker. By mapping the DoAs to the localizations, the absolute position and orientation of the microphone arrays is derived. If time periods can be identified where the speaker is static or moving slowly, no strict time synchronization is required.

This work was supported by the German Research Foundation (DFG) under contract number Fi 799/5-1.

We would like to thank the anonymous reviewers for their helpful suggestions.

2. METHOD

In this work we consider a smart room setting with several cameras at known positions and microphone arrays with unknown positions and orientations. A recording of a speaker talking at static positions in the room is used to estimate the geometry of the microphone arrays in the following way: Suitable time periods for a number of positions are identified. The Euclidean positions of a speaker are estimated by visual detection and triangulation. The DoA of the utterances at each microphone array and position are estimated. Both Euclidean position and DoA are computed for the projection to the ground floor. Using sets of matched visual 2D localizations and acoustic DoAs, an estimate of the absolute position and orientation of the microphone arrays is computed. By computing a consensus over several such estimates, a reliable estimation is derived.

2.1. Acoustic Speaker Localization

There are various methods of speaker localization that provide DoA estimates for small microphone arrays. The robust bio-inspired speaker localization described in [15] is used since it is robust against reverberation and provides an implicit speech/non-speech decision. To increase robustness, event classification [6] can be employed to discard non-speech events. The DoA azimuth is computed by EM clustering of a spatial likelihood derived from TDoA estimates. Time periods where a speaker is static and robustly localized are identified as periods with a large number of similar estimates. For each person position with index i and microphone array with index m, the median DoA $\Theta_{i,m}$ with respect to the ground plane is computed.

2.2. Visual Person Localization

For the proposed calibration method, any visual localization can be used that estimates positions with respect to the ground plane. Given a conference setting where the person may be sitting, upper body detection can be computed from the camera images with gradient histograms [16]. To remove false alarms due to visual clutter, background subtraction was used [17]. The detections in the camera images can be back-projected into the room since their position and orientation is known. If a person is seen by more than one camera, their Euclidean position can be computed by triangulation. For each position with index i, absolute 2D localizations $\boldsymbol{s}_i \in \mathbb{R}^2$ with respect to the ground plane are estimated [18].

2.3. Geometry Estimation

In order to calibrate the geometry for each circular microphone array, both the 2D position $\boldsymbol{r}_m \in \mathbb{R}^2$ and the orientation o_m have to be estimated. The possible range for the position is given by the camera geometry and the room size,

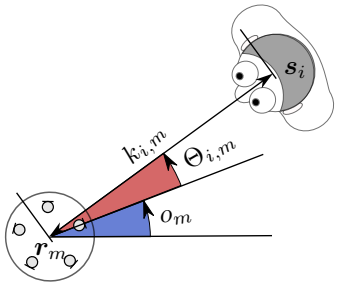

Fig. 1. Geometric relations of a single microphone array at $\boldsymbol{r}_m$ with orientation o_m and one speaker at $\boldsymbol{s}_i$ localized with DoA $\Theta_{i,m}$ in distance $k_{i,m}$.

o_m is in the range $[-\pi, \pi]$. For each person position and microphone array, the vector from the source $\boldsymbol{s}_i$ to the receiver $\boldsymbol{r}_m$ can be expressed by the DoA and the distance $k_{i,m} \in \mathbb{R}^+$, cp. Fig. 1:

$$\boldsymbol{s}_i - \boldsymbol{r}_m = k_{i,m} \begin{pmatrix} \cos\left(o_m + \Theta_{i,m}\right) \\ \sin\left(o_m + \Theta_{i,m}\right) \end{pmatrix} . \tag{1}$$

Solving this directly assumes no error in the localization data. This equations can be rewritten to describe the Euclidean error of the estimate:

$$\boldsymbol{e}_i = \boldsymbol{s}_i - \boldsymbol{r}_m - k_{i,m} \begin{pmatrix} \cos\left(o_m + \Theta_{i,m}\right) \\ \sin\left(o_m + \Theta_{i,m}\right) \end{pmatrix} . \tag{2}$$

The error reflects both errors in the the position and angular estimates as well as the geometry estimate. By stacking the linear equations for I speaker positions

$$\begin{aligned} \boldsymbol{e}_1 &= \boldsymbol{s}_1 - \boldsymbol{r}_m - k_{1,m} \begin{pmatrix} \cos\left(o_m + \Theta_{1,m}\right) \\ \sin\left(o_m + \Theta_{1,m}\right) \end{pmatrix} \\ &\vdots \\ \boldsymbol{e}_I &= \boldsymbol{s}_I - \boldsymbol{r}_m - k_{I,m} \begin{pmatrix} \cos\left(o_m + \Theta_{I,m}\right) \\ \sin\left(o_m + \Theta_{I,m}\right) \end{pmatrix} \end{aligned} \tag{3}$$

we can derive a minimization problem stating that the sum l2 norm errors

$$e = \sum_{i=1}^{I} ||\boldsymbol{e}_i|| \tag{4}$$

should be minimal for the correct estimates. If the positions $\boldsymbol{s}_i$ and DoAs $\Theta_{i,m}$ are known from acoustic and visual localization, the offsets o_m and positions, $\boldsymbol{r}_m$ can be estimated with distances $k_{i,m}$. We have to find estimates for $3 + I$ unknowns with $2I$ equations, therefore the system of linear equations is determined for $I \geq 3$. Thus a geometry estimate can be derived by searching for a minimum of e for all possible configurations.

After choosing N random sets of $I \geq 3$ positions, for each array m and random set with index n, estimates $\boldsymbol{r}_{m,n}$, $o_{m,n}$ are computed by minimizing (3) by gradient descent. The Broyden-Fletcher-Goldfarb-Shanno algorithm [19] is used which works with bounding constraints and approximates the

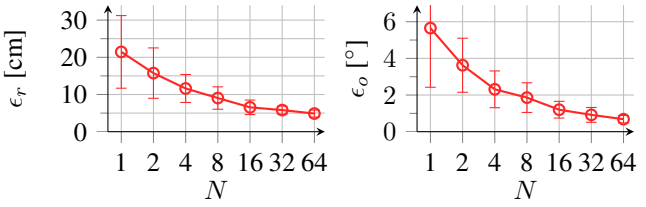

Fig. 2. Estimation error mean and deviation for a different numbers of position sets N. Simulation with a localization error of $\epsilon_v = 20\,\text{cm}$ and $\epsilon_a = 4°$.

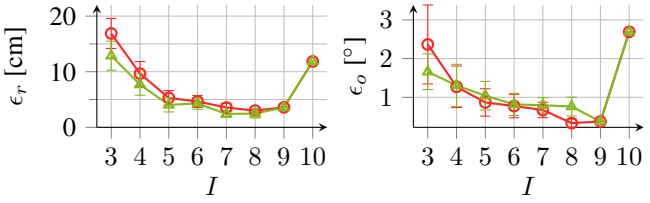

Fig. 3. Calibration error for mean and consensus for different numbers of positions I in simulation ($\epsilon_v = 20\,\text{cm}$, $\epsilon_a = 4°$).

Hessian based on a local neighborhood. A suitable minimum is found regardless of initialization since the problem is close to convex in all paractical cases.

The N estimates may contain outliers due to errors in the localization or the positions being close to co-linear. Over the N geometry estimates, a refined position estimate is computed as the median $\widehat{\boldsymbol{r}_m}$ of all $\boldsymbol{r}_{m,n}$. Then the set of the $N' = N/3$ estimates with the smallest Euclidean distance to the median is used to compute an improved estimate. The average position and orientation of these N' position estimates is the final "consensus" estimate.

3. SIMULATION

In order to test the viability of the method, several simulations were performed. All simulations used the real microphone configuration in the smart room and 15 person positions located around them. The positions were chosen with regard to a conference scenario, either sitting at the table, standing near the table or near the whiteboard.

3.1. Number of position sets N

Assuming a zero-mean distributed error of the localizations, it is clear that the average estimated error for multiple position sets will decrease with the number of sets used. A number of simulations were done with Gaussian distributed localization errors with an RMS of $\epsilon_v = 20\,\text{cm}$ for speaker positions and $\epsilon_a = 4°$ for DoAs. Fig. 2 shows the results for estimations with $I = 5$ and different set sizes N. The estimation error decreases with the number of position sets used. It falls below the localization error for eight or more sets.

3.2. Number of Positions I

The number of positions used in the set was varied from 3 to 10 in another experiment with the same simulated errors. $N = 40$ sets were chosen, and the $N' = 16$ estimates closest to the median were used as consensus to compute an improved estimate. The choice of N' was not found to be critical. Figure 3 shows the resulting errors for both the mean and the consensus. The highest number of positions does not

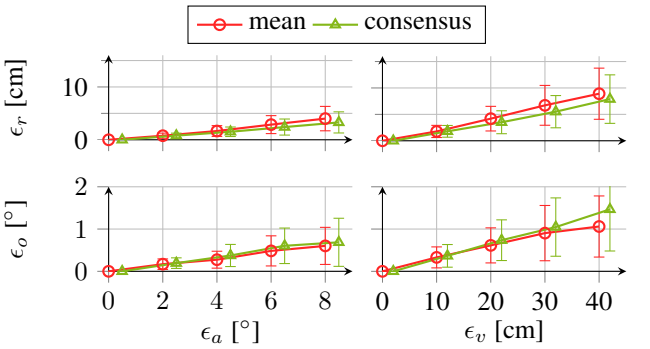

Fig. 4. Estimation error of orientation ϵ_o and absolute position ϵ_r and its standard deviation for different audio and video localization errors ϵ_a, ϵ_v.

yield the smallest error, the minimum error is situated around $I = 7$. The consensus improves the estimate.

3.3. Localization Error

In order to investigate the influence of the localization errors, different audio localization errors $\epsilon_a = 0, 2, \ldots, 8°$ and video localization errors $\epsilon_v = 0, 10, \ldots 40\,\text{cm}$ were simulated. $N = 40$ sets of positions and a $N' = 16$ consensus were used with $I = 5$. The resulting geometry estimation errors are shown in Fig. 2. For all simulations, the orientation error ϵ_o is lower than 2°, which is beneficial for localization target applications, since the triangulation quality decreases rapidly with angular errors. The consensus shows a smaller position estimation error and a similar orientation error for audio localization errors.

4. SMART ROOM RECORDINGS

In order to test the proposed method in a real scenario, a recording was made in the highly reverberant $3.7 \times 6.8 \times 2.6\,\text{m}^3$ conference room of a smart house installation at TU Dortmund university. Signals from three circular microphone arrays with 5 microphones in a 5 cm radius embedded in a table were recorded at 48 kHz. Each array was captured by a separate sound card. Recordings of coherent white noise showed a jitter of 22 μs between the sound cards. A reverberation time of $670 \pm 89\,\text{ms}$ over the microphone signals was

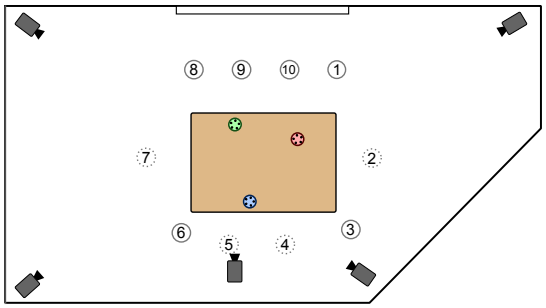

Fig. 5. Smart room scenario for recording. A person takes 11 positions sitting (dotted) and standing while speaking a sentence in the direction of the table. The three microphone arrays are situated on the table (colored cicrles). Five cameras mounted near the walls.

			mean		consensus	
I	N	N'	ϵ_r[cm]	ϵ_o[°]	ϵ_r[cm]	ϵ_o[°]
3	35	11	9.6	3.25	8.8	3.24
4	35	11	8.5	3.07	7.2	1.80
5	21	7	7.5	2.80	7.4	1.62
6	7	4	6.9	2.34	6.6	1.89
7	1	1	6.6	2.01	–	–

Table 1. Calibration results for smart room recordings.

calculated using a blind estimation algorithm [20]. Five cameras mounted at the ceiling captured the scene at 10 fps and 384×288 pixel resolution. They have a field of view (FoV) of $48° \times 36°$. A person took the ten positions shown in Fig. 5 in the room and spoke a sentence at each position. The recording provided a total of seven visual localization (position nr. 2-5,7,8,10) with an position error of around $\epsilon_v \approx 20\,\text{cm}$, For the audio localizations, the median angle at each position was chosen, the error was around $\epsilon_a \approx 5°$.

4.1. Calibration

The calibration was done with the proposed method using all available position sets and an $N' = N/3$ consensus. The results are shown in Tab. 1. The absolute position error is around $\epsilon_r = 7\,\text{cm}$ and the orientation error around $2°$. This is already achieved using $I = 4$ and the consensus method for outlier removal. In Fig. 6 the absolute position estimates their consensus estimate are plotted. The mean estimate improves when using more positions. The consensus estimate is consistently better.

4.2. Speaker Tracking

A multi-array speaker tracking method based on [2] was applied using both the measured ground truth and the estimated geometry on the calibration sequence itself and a second similar sequence with 18 speaker positions. The tracking performance using both calibrations is shown in table 2. The Euclidean RMS ϵ_l decreases slightly when the automated calibration is used. The angular localization error ϵ_a is similar or better. A distance of $0.5\,\text{m}$ is used as margin for the allowed Euclidean error to reflect what error may be tolerable for practical applications. With respect to that, the precision (P) and recall (R) are similar.

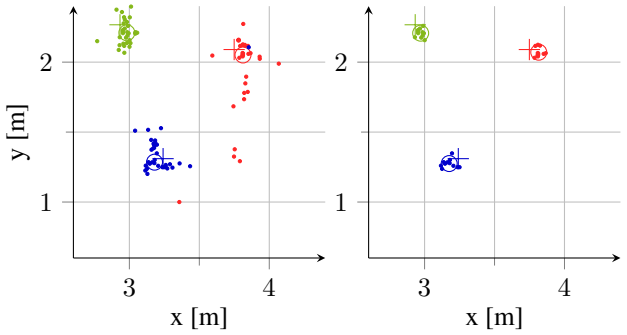

Fig. 6. Ground truth positions (+), individual estimates (·) and average estimate (o) for the three microphone arrays using $N = 35$ position sets (left) and the $N' = 11$ consensus (right) for $I = 4$ using the smart room recording with all speaker positions. Note that only the section of the total search space containing the estimates is shown.

	measured				proposed			
	ϵ_a[°]	ϵ_l[cm]	P [%]	R [%]	ϵ_a[°]	ϵ_l[cm]	P [%]	R [%]
#1	4.9	18.7	100	89	4.9	17.3	100	89
#2	5.1	26.1	99	90	4.7	23.2	97	90

Table 2. Acoustic speaker tracking results for manual measurement and calibration using the proposed method for the calibration sequence (#1) and a second one (#2).

5. CONCLUSION

A method for absolute calibration of distributed microphone arrays in a smart room was introduced. It requires no additional calibration step or strict time synchronization, only a recording of a single speaker talking at a small number of fixed positions, where absolute Euclidean localization is done using the cameras. The DoA at the individual microphone arrays is estimated by a robust acoustic method. By matching the positions to the DoAs, the absolute position and orientation of each microphone array is computed. Application in a reverberant smart room achieved around 0.1 m and $2°$ accuracy. When applying the calibration result and performing acoustic speaker tracking, the localization RMS decreases slightly over manual calibration, the precision and recall are almost identical with respect to practical applications. The proposed method is an alternative to manual measurements for an absolute calibration.

REFERENCES

[1] Damien Kelly, Anil Kokaram, and Frank Boland, "Voxel-Based Viterbi Active Speaker Tracking (V-VAST) with Best View Selection for Video Lecture Post-Production," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011, pp. 2296–2299.

[2] Axel Plinge and Gernot Fink, "Multi-Speaker Tracking using Multiple Distributed Microphone Arrays," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014.

[3] Keni Bernardin, Tobias Gehrig, and Rainer Stiefelhagen, "Multi- and Single View Multiperson Tracking for Smart Room Environments," in *Workshop on Classification of Events, Actions and Relations (CLEAR)*, 2006.

[4] Shankar T. Shivappa, Mohan Manubhai Trivedi, and Bhaskar D. Rao, "Audiovisual Information Fusion in Human-Computer Interfaces and Intelligent Environments: A Survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct. 2010.

[5] Deepu Vijayasenan, Fabio Valente, and H. Bourlard, "An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 431–438, 2011.

[6] Axel Plinge, Rene Grzeszick, and Gernot Fink, "A Bag-of-Features Approach to Acoustic Event Detection," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014.

[7] Stanley T. Birchfield, "Geometric Microphone Array Calibration by Multidimensional Scaling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[8] Vikas C. Raykar, Igor V. Kozintsev, and Rainer Lienhart, "Position Calibration of Microphones and Loudspeakers in Distributed Computing Platforms," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, 2005.

[9] Marius H. Hennecke and Gernot A. Fink, "Towards Acoustic Self-Localization of Ad Hoc Smartphone Arrays," in *Proc. Workshop on Hands-Free Speech Communication and Microphone Arrays*, Edinburgh, UK, 2011, pp. 127–132.

[10] Nikolay Gaubitch, Willem Bastiaan Kleijn, and Richard Heusdens, "Auto-Localization in Ad-Hoc Microphone Arrays," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancover, Canada, 2013.

[11] Yubin Kuang and Kalle Aström, "Stratified Sensor Network Self-Calibration From TDOA Measurements," in *European Signal Processing Conference*, Marrakesh, Morocco, 2013.

[12] Florian Jacob, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "DoA-based Microphone Array Postion Self-Calibration using Circular Statistisc," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancover, Canada, 2013.

[13] Joerg Schmalenstroeer, Florian Jacob, Reinhold Haeb-Umbach, Marius H. Hennecke, and Gernot A. Fink, "Unsupervised Geometry Calibration of Acoustic Sensor Networks using Source Correspondences," in *Interspeech*, 2011.

[14] Axel Plinge, Marius H. Hennecke, and Gernot A. Fink, "Reverberation-Robust Online Multi-Speaker Tracking by using a Microphone Array and CASA Processing," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, 2012.

[15] Axel Plinge and Gernot A. Fink, "Online Multi-Speaker Tracking Using Multiple Microphone Arrays Informed by Auditory Scene Analysis," in *European Signal Processing Conference*, Marrakesh, Morocco, 2013.

[16] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 886–893.

[17] Pakorn Kaewtrakulpong and Richard Bowden, "An improved Adaptive Background Mixture Model for Real-Time Tracking with Shadow Detection," in *European Workshop on Advanced Video-Based Surveillance Systems*, 2001.

[18] Steve Brischke, "Multikamera Personelokalisierung in Inteligenten Umgebungen (Multicamera Person Localization in Intelligent Environments)," Diploma thesis, TU Dortmund University, Dec. 2013.

[19] Richard H. Byrd, Peihuang Lu, and Jorge Nocedal, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific and Statistical Computing*, vol. 16, pp. 1190–1208, 1995.

[20] Heinrich W. Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary, "An Improved Algorithm for Blind Reverberation Time Estimation," in *International Workshop on Acoustic Echo and Noise control*, Tel Aviv, Israel, 2010.