Passive online geometry calibration of acoustic sensor networks

Axel Plinge, *Member, IEEE,* Gernot A. Fink, *Senior Member, IEEE*,
and Sharon Gannot *Senior Member, IEEE*

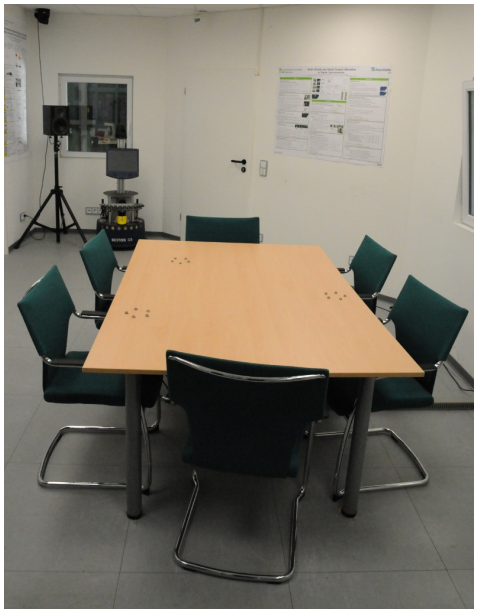

Figure 1. Smart conference room

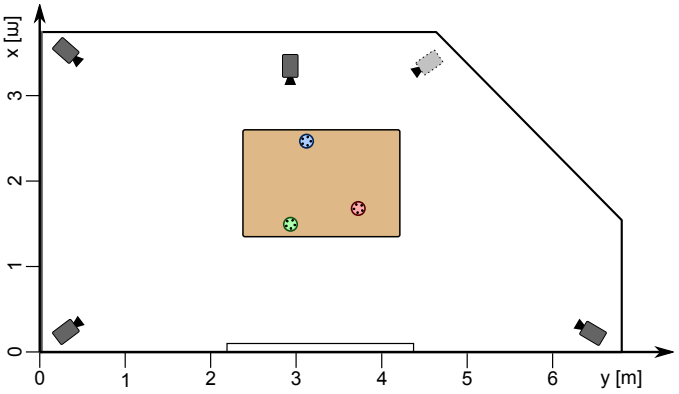

Figure 2. Camera and microphone array positions in the smart room

APPENDIX

In this appendix, the modes of distance estimation are investigated in further detail. The accuracy and influence of using the PoAP sparse spike representation will be evaluated. By comparison with a calibration sequence where the source is in the same plane as the microphones, the 2D assumption will be tested.

Additionally, the method will be compared to the previous one and two others from the literature. A multimodal method and another passive speech based technique that is fast to compute.

A. Smart room recordings

Several sequences with a single source directed towards the table were recorded. The setup is the same as in [1], Fig. 2 shows the positioning of the microphone arrays and cameras in the smart room. For all recordings, the speaker positions were known from floor markings.

A. Plinge and G. A. Fink are with the Department of Computer Science, TU Dortmund University, Dortmund, Germany. S. Gannot is with the Faculty of Engineering at Bar Ilan Univesity, Ramat Gan, Israel.
We would like to thank Shmulik Markovich-Golan for helpful discussions.
This work was supported by a fellowship within the FITweltweit program of the German Academic Exchange Service (DAAD).

In sequence #1 a smartphone was used to play white noise at table height, in the other sequences #2, #3, and #4 a human speaker was uttering a sentence from positions sitting in a chair or standing up in the room. In sequence #1 and sequence #2, the same ten positions were used. In #3 and #4 15 and 19 positions were used, respectively. The additional positions were situated at the whiteboard further away from the table.

In the recordings #2 and #4, the five cameras were used to track the speaker visually by upper body detector using histograms of oriented gradientss (HoGs) and triangulation to allow for multimodal calibration [2].

B. TDoA estimation

In order to asses the influence of the method used to compute the time difference of arrival (TDoA), all four sequences were used. By using differential evolution optimization on all positions for each sequence, the achievable optimum dependent on the measurement was computed. Table I summarizes the accuracy of the measurements and the resulting geometry calibration error.

The direction of arrival (DoA) estimation was used to automatically determine the time segments corresponding to the different positions. The angular RMS error slightly higher error for the latter sequences, probably caused by the speaker being further away and the signals having higher reverberation.

The error in TDoA measurements is higher in the human speaker case, due to his elevation. When computing the TDoA error with respect to the three-dimensional ground truth positions, the RMS error is lowered by 4-6 cm.

When using the peak over average position (PoAP) spike representation in order to compute the TDoA measurements,

Table I
ERROR OF THE MEASUREMENTS USED FOR ACOUSTIC GEOMETRY ESTIMATION WITH CALIBRATION SEQUENCES AND THE RESULTING ERROR OF THE CALIBRATION WITH DIFFERENTIAL EVOLUTION OPTIMIZATION ON ALL POSITIONS.

sequence	T	measurement RMS			calibration error			
		ϵ_a	ϵ_τ		ϵ_o		ϵ_r	
		PoAP	PHAT	PoAP	PHAT	PoAP	PHAT	PoAP
#1 noise	10	4.44°	3.34 cm	5.27 cm	2.40°	2.01°	2.84 cm	2.73 cm
#2 speech	10	3.93°	8.92 cm	6.47 cm	2.79°	1.97°	5.94 cm	3.64 cm
#3 speech	15	7.28°	12.89 cm	12.96 cm	1.43°	1.48°	9.06 cm	8.90 cm
#4 speech	19	5.45°	12.06 cm	10.17 cm	1.35°	0.69°	8.92 cm	7.06 cm

the RMS error with respect to the two-dimensional ground truth positions is similar or better that the steered response power with phase transform (SRP-PHAT) based estimates. For the white noise it is slightly worse with 5.3 cm. This can be understood since the spike representation is not tuned for broadband noise signals.

C. Comparison with other approaches

In order to asses the quality of the proposed method in comparison to other approaches from the literature, the two sequences with video recordings were selected. Thus our previously proposed multimodal approach [2] could be used as well. For the acoustic approach, the PoAP spike correlation was used in conjunction with the differential evolution optimization. Both the off-line [1] and online version were run with position subsets of size $t_0 = 6$.

For comparison, the approach by Pertilä et al. was reimplemented [3]. This method capable of working in real time. Their approach is based on applying multidimensional scaling (MDS) in order to estimate the geometry from all pairwise distances [4]. MDS can be related to the principal component analysis (PCA). In both methods the eigenvectors corresponding to large eigenvalues yield the desired representation. Here a matrix consisting of the squared distances of the sensors is used. The eigenvectors of the decomposed inner product matrix are an estimate of the relative sensor coordinates. For microphone distances in the order of 1 m, the pairwise distances cannot be estimated using diffuse noise as in [5]. Instead, Pertilä et al. compute the TDoA of speech. Assuming at some point the speaker will be in the endfire position with regard to each microphone pair, the distance can be estimated by taking the maximum TDoA value.

Fig. 3 shows the results for the different methods. All proposed acoustic methods calibrate with an error of 7 cm and around 1°. The position error is partly due to the speakers elevation. In the short dedicated calibration sequence #2, the position error is lower with about 4 cm. This is likely due to the speakers higher proximity to the nodes.

The mTDoA method achieves a comparable position accuracy. However, the orientation cannot be estimated reliably from the microphone positions. The orientation error is beyond 5° when using minimum angular difference for alignment of the known geometry, as is shown. Using singular value decomposition (SVD) as in [6] did not improve the results.

As the multi-modal approach optimizes the arrays positions independently, the error in measurement translates more directly. Additionally, the visual localization had an higher error than the correlation based distance measurements. Thus, a higher position error 15 cm and 10 cm is observed.

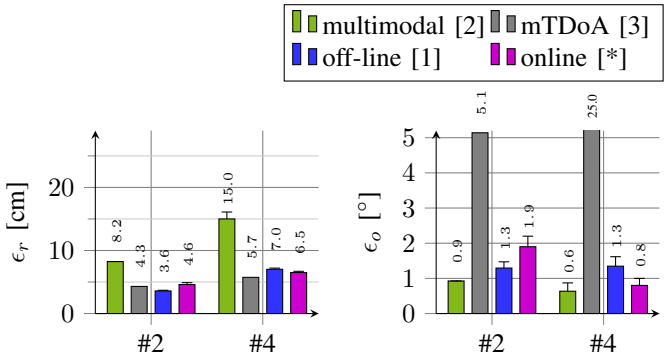

Figure 3. Comparison of different methods for array configuration calibration with the proposed online method [*]. Position (left) and orientation error (right) for two calibration sequences with 10 and 19 speaker positions.

REFERENCES

[1] A. Plinge and G. A. Fink, "Geometry calibration of multiple microphone arrays in highly reverberant environments," in *Int. Works. on Acoustic Signal Enh.*, Antibes – Juan les Pins, France, Sept. 2014.

[2] A. Plinge and G. A. Fink, "Geometry calibration of distributed microphone arrays exploiting audio-visual correspondences," in *European Signal Process. Conf.*, Lisbon, Portugal, Sept. 2014.

[3] P. Pertilä, M. Mieskolainen, and M. S. Hämäläinen, "Passive self-localization of microphones using ambient sounds," in *European Signal Process. Conf.*, Bucharest, Romania, Aug. 2012, pp. 1314–1318.

[4] S. T. Birchfield, "Geometric microphone array calibration by multidimensional scaling," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Process.*, Hong Kong, Apr. 2003.

[5] I. McCowan and M. Lincoln, "Microphone array shape calibration in diffuse noise fields," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 16, no. 3, pp. 666–670, 2008.

[6] P. Pertilä, M. Mieskolainen, and M. S. Hämäläinen, "Closed-form self-localization of asynchronous microphone arrays," in *J. Works. on Hands-Free Speech Commun. and Microphone Arrays*, Edinburgh, UK, May 2011, pp. 139–144.