Evaluation of different microphone arrays and localization algorithms in the context of ambient assisted living

Christian Bartsch*, Andreas Volgenandt†, Thomas Rohdenburg† and Joerg Bitzer*†
* Institute for Hearing Technology and Audiology (IHA), Jade University Of Applied Sciences,
Ofener Str. 16/19, 26121 Oldenburg, Germany, www.hoertechnik-audiologie.de
† Fraunhofer Institute for Digital Media Technology (IDMT) Department Hearing, Speech and Audio Technology (HSA),
Marie-Curie-Str. 2, 26129 Oldenburg, Germany

Abstract—**This paper presents the evaluation of full 3D sound source localization systems for real world living environments. We tested several well-established algorithms. The Generalized Cross Correlation Phase Transform (GCC-PHAT) and Adaptive Eigenvalue Decomposition Phase Transform (AED-PHAT) algorithms were implemented as Time Delay of Arrival (TDOA) estimators. For the localization itself a three dimensional acoustic map was computed using the Global Coherence Field (GCF) as well as the modified Least Squares (LS) algorithm called Least Median of Squares (LMS). The combinations of these techniques are applied to different microphone array configurations composed of two fundamental microphone arrays. These fundamental arrays are a set of ceiling-mounted sensors with large distances and a small spherical array. Finally, a Voice Activity Detector (VAD) was applied in order to avoid false localization estimations during speech pauses. For evaluation we recorded a database of speech signals in a natural living environment. The results show that the combination of ambient microphone arrays with modern localization algorithms are able to locate people in a room in all three dimensions. However, the localization is not perfectly accurate and an error up to 0.4 m has to be tolerated.**

I. Introduction

Knowing the location of a sound source gives great benefit to many applications, e.g. for (multi talker) video conference systems. In this case the knowledge of the current speaker's position offers the possibility to use further speech enhancement techniques like beamforming [1].

In our application we are interested in sound source localization in an Ambient Assisted Living (AAL) apartment. Therefore, the mounted microphones should be as invisible as possible. While the ceiling attached sensors could be concealed completely, hiding spherical arrays is a little more difficult. Our considered solution for that problem is to build a spherical array within a (designer) lamp. The advantages of the combination are that the lamp hides the installed microphones and a very nice indirect light would be created.

The apartment in which the implemented technologies were tested and used is a part of the *"Lower Saxony Research Network Design of Environments for Ageing"* [2], [3]. There are different applications considered within this project which are based on an accurate estimation of the position of the speaker. Some of these are controlling

- a beamformer,
- a multichannel playback system for spatial presentation of sounds,
- other devices, e.g. video control systems or lights.

II. Used Algorithms

For our evaluation we only used well-known and established techniques. The algorithms are based on estimating either cross-correlation (CC) or the Time Delay of Arrival (TDOA) $\widehat{\tau}$ between two microphones. This can be done by using the Generalized Cross Correlation (GCC, [4]) and the Adaptive Eigenvalue Decomposition (AED, [5]–[7]) algorithms. The AED is used to estimate the eigenvector corresponding to the lowest eigenvalue of the combined signals of a microphone pair. It has been shown in [5] that this eigenvector contains a rough estimate of the impulse responses from the source to the two microphones. With an appropriate initialization of the AED [8] and a minimum search, the delay between the direct paths of the microphone signals can be estimated. Generally this is done in the frequency domain. For both TDOA algorithms, GCC and AED, a spectral weigthing of the cross-spectral density between the inputs $S_{x_0x_1}[k]$ is applied, according to the well-known Phase-Transform (PHAT) weighting.

The position of the speaker was estimated by employing an acoustic map. Acoustic maps are functions, defined over a sampled space of potential solutions that represent the plausibility that a source is present at a given point p [9]. In this contribution we compare two localization algorithms based on acoustic maps, the Least Median of Squares (LMS) which is a more robust extension of the well-known Least Squares (LS) approach [9], [10] and the Global Coherence Field GCF [9], [11]. Both algorithms sample the room into a discrete grid with a spacing of 0.1 m in each direction x, y and z. While LMS uses the estimated TDOA $\widehat{\tau}$ only, the GCF takes advantage of the whole cross correlation function which means more computed information will be used.

For the TDOA estimation and the localization algorithms the following four combinations were tested: GCC-PHAT/LMS, GCC-PHAT/GCF, AED-PHAT/LMS and AED-PHAT/GCF.

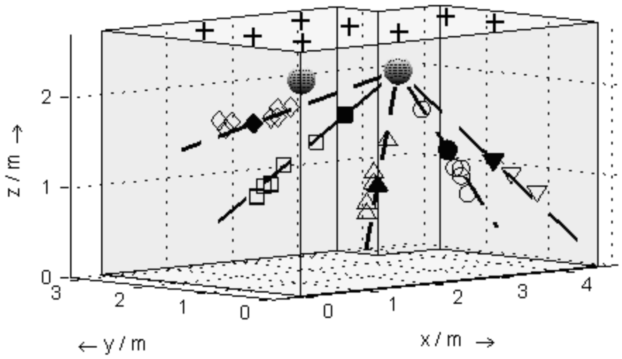

Fig. 1. Schematic view showing the arrangement of the ceiling-mounted microphones (pluses) and spherical arrays (shaded spheres, see also Fig. 2) in the AAL-laboratory at Offis Institute of Information Technology [3]. The diamonds, squares and circles as well as the upward- and downward-pointing triangles represent five different speaker positions (filled marker) and their estimations (unfilled marker) by using the right spherical array. The direction from this array to each true speaker-position is plotted by dashed lines. We tested three sitting-positions (circles, upward- and downward-pointing triangles) as well as two standing-positions (diamonds and squares).

TABLE I
USED ARRAY COMBINATIONS WITH C = CEILING ARRAY AND S_i = SPHERICAL ARRAY i

Combination Number	Arrays used	Number of possible microphone pairs
1	S_2	28
2	S_1 & S_2	120
3	C	28

Typically speech pauses reduce the robustness of localization algorithms for real-world acoustic scenarios. Therefore, a Voice Activity Detection (VAD) based on [12] was applied.

III. EVALUATION

A. Acoustic scenario and microphone arrays

For this evaluation (compare Fig. 1) we used two different kinds of microphone arrays. The first array consisted of eight capacitor microphones which were mounted at the ceiling of the living room. The minimum distance between a pair of microphones in this array was $d_{min} = 0.78$ m while the maximum distance measured $d_{max} = 3.33$ m. As a second and third array we used two self-build spherical arrays which were made from styrofoam spheres with eight low-priced electret capsule microphones. Since they were placed uniformly distributed on the sphere's surface the microphones would span a cube with equal edge lengths (Fig. 2). The radius of those spheres is $r = 0.075$ m. While the first microphone array was mounted into the room's ceiling (height at $s_z = 2.7$m) both spherical arrays were placed at a height of $s_z = 2.1$ m.

Table I shows the microphone array combinations we tested with the implemented algorithms.

Two facts are noteworthy:

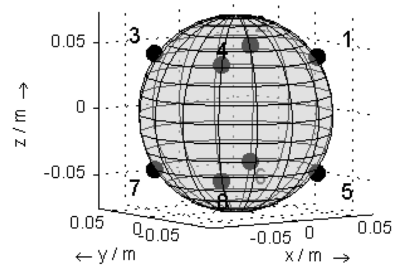

Fig. 2. Schematic view of a spherical array that contains 8 microphones. The microphones are placed uniformly distributed on the sphere's surface which means connecting all neighboring microphones would span a cube with equal edge lengths of 0.087 m.

- We used low-priced microphones that had not been calibrated according to frequency or level in order to have a practical and realistic constraint for AAL applications.
- We did not measured the spatial positions of the microphones and arrays accurate to a millimetre.

Finally, our configuration has 24 microphone channels which are recorded simultaneously at a sampling frequency of $f_s = 48$ kHz and a resolution of 16 bits.

The different speaker positions (three sitting- and two standing-positions), the microphone positions, and the position of the sphere-arrays that we tested are shown in Fig. 1.

B. Methodology

All tested algorithms were implemented in a MATLAB block processing framework. The size of the blocks was 12.5 ms which results in a length of 1024 samples at the given sampling frequency of $f_s = 48$ kHz. Before calculating anything all data were filtered through a third order Butterworth high-pass filter with a cut-off frequency of $f_c = 100$ Hz. This filter was supposed to avoid estimation errors caused by foot fall sounds which may temporarily occur.

Originally the test sounds were used to train a speech recognition system. Therefore, speech pauses where introduced deliberately, which are responsible for a high amount of estimation errors. These are caused by a computer rack outside the room, which was localized in the absence of speech. To prevent this a combination of a voice activity detector (VAD) and a moving root mean square (MRMS) threshold decision method was implemented. As VAD we used the technique proposed by [12] which performs quite well in noisy environments. However, the VAD can produce false alarms in quiet situations, so we used the MRMS decision method as a second step. Only if the MRMS is greater than a predefined threshold and if the VAD signalizes *"speech"* on all microphone channels, acoustical activity will be identified and the current data block will be used for localization.

The calculation of the GCC/AED algorithm starts instantly after acoustical activity is detected. However, the position estimation through GCF/LMS starts with a little time-shift to GCC/AED. The time shift is $t_s = 0.1$ s which equates to the smoothing time of the cross spectrum calculation for the GCC. This approach takes advantages of the fact that the TDOA estimation needs a few blocks of processing to adapt. The estimated positions were not smoothed over time because speaker tracking is not a topic in this article. We also tried to improve the time delay estimation for the spherical arrays by incorporating head models [13]. However, no significant advantages justifying the higher computational complexity could be found.

C. Determining the hit rate

If we estimate a position $\widehat{p}(n) = [\widehat{x}(n), \widehat{y}(n), \widehat{z}(n)]$ in the three-dimensional space at a time step n, the error $e(n)$ of this estimation is given as

$$e(n) = \|\widehat{p}(n) - p\| \text{ , with the true source position } p. \quad (1)$$

That means the error $e(n)$ stands for the spatial distance between true position and estimation. Further we defined a percentage-correct measure depending on the radius r_{corr} of a lock-in range sphere around the true position. If $\widehat{p}(n)$ lies inside this sphere the estimation is identified as *"correct"*:

$$P_{corr}(n) = \begin{cases} 1, & \text{if } e(n) < r_{corr} \\ 0, & \text{else} \end{cases} \quad (2)$$

Finally, the hit rate of a full trial can be calculated as

$$P_{corr,trial} = \sum_{n=1}^{N} P_{corr}(n) \text{ , with number of time steps } N. \quad (3)$$

We evaluated the arrays and algorithms with different lock-in range radiuses in the range of $r_{corr} = 0 \ldots 1$ m.

IV. Results

An overview of all results by using a lock-in range radius of $r_{corr} = 0.4$ m is given in Fig. 4. The worst hit rates were produced by using only one spherical array. This could be explained by the fact that the mapped area of a sound source is not circular but long drawn-out in the acoustical map [11]. Moreover, it holds that a smaller distance between the microphone pairs leads to extended mapped areas. Thus, finding the correct position is more difficult if the mapped area is very long drawn-out. However, if you look at Fig. 1, it is clear that only the estimation of the exact position fails when using only one spherical array whereas the estimation of the direction is quite good.

GCC-PHAT/LMS seems to be the worst algorithm combination with the largest variances. A reason for that could be that the LMS only uses the estimated TDOA $\widehat{\tau}$ between a pair of microphones, but not the whole cross-correlation data like the GCF does. Furthermore, if the spatial distance between the used pair of microphones is large, which is the case when using both spheres or the ceiling array, AED-PHAT might produce better TDOA estimations than GCC-PHAT. This explains why the degradation of GCC-PHAT/LMS

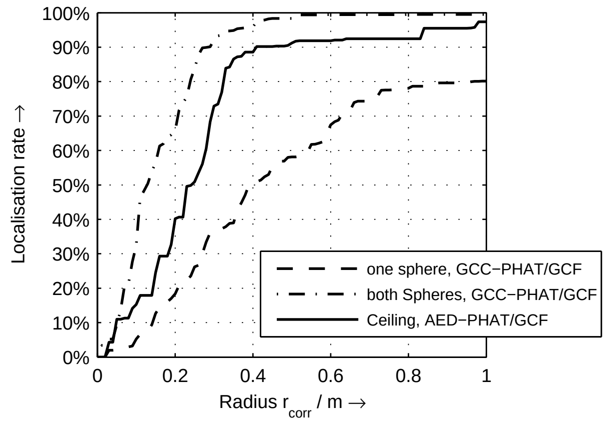

Fig. 3. Localization rates in percent depending on the lock-in range radius r_{corr} and plotted for the three best array-algorithm-combinations which are identified in Fig. 4, position 6. Curves are calculated by using the data from all speakers and all speaker positions.

is significant compared to the other algorithm combinations when using both spheres or the ceiling array, but not when using only one spherical array.

Using both spherical arrays causes the best of all hit rates for all the algorithm combinations, which is not surprising because in this case 16 microphones instead of 8 microphones were used. But also the ceiling-mounted microphones alone can produce quite good estimation results by using AED-PHAT/GCF.

Furthermore, Fig. 4 shows, that LMS does not enhance the hit rates compared to GCF. Since LMS needs more processing power than GCF, we have selected GCF as the better localization technique for our microphone array setups at the AAL-laboratory.

Moreover, we haven chosen the three best array-algorithm combinations and plotted their localization rates depending on the lock-in range radiuses r_{corr} (Fig. 3). This figure demonstrates that an application of both spherical arrays can estimate the speaker position within a radius of $r_{corr} = 0.3$ m with a hit rate of more than 90 %. The ceiling-mounted array needs approximately $r_{corr} = 0.4$ m to reach the same hit rate of 90 %. As explained before, one spherical array is insufficient to estimate the accurate position.

When using both spherical arrays as well as the ceiling-mounted microphones there is a strong increase of the localization rate in the range of $r_{corr} \approx 0 \ldots 0.3$ m and $r_{corr} \approx 0 \ldots 0.4$ m, respectively. For larger r_{corr} the gradient decreases which means choosing r_{corr} larger than these ranges does not result in significantly higher localization rates. In case of using one sphere only, the point where the localization rate gradient turns smaller lies at $r_{corr} \approx 0.7 \ldots 0.8$ m.

V. Conclusions

In this paper we evaluated well-known algorithms for 3D acoustical localization with constraints on the array design. In the context of ambient assisted living the array should be as

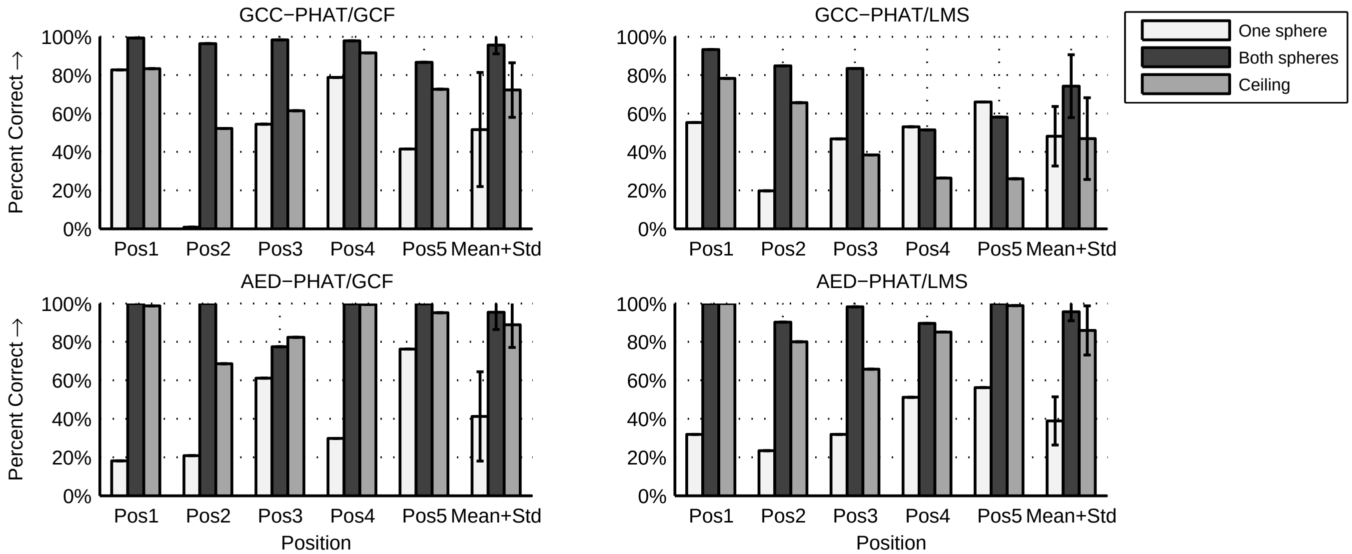

Fig. 4. Percent-correct hit rates calculated by using the estimated data from all five speakers. The x-axis represents the different speaker positions 1-5 and an average for all positions. The first and second lines correspond to the GCC-PHAT and AED-PHAT TDOE estimators, respectively. The first and the second columns correspond to the GCF and LMS algorithm, respectively. The lock-in range radius was set to $r_{corr} = 0.4$ m.

invisible as possible and the overall costs should be moderate. Therefore, we built small spherical arrays which could be hidden in lamps and an array of microphones hidden in the ceiling. For both designs the microphones and the corresponding amplifier and conversion chips were low-cost devices. The results clearly indicate that localization is possible with these arrays, if an error of 0.4 m can be tolerated. However, only one sphere is not enough for exact localization. Only the direction is estimated very well, which would be enough for the beamformer application. One interesting result was the effectiveness of the ceiling array with the AED algorithm, even though the problem of spatial aliasing is present at all relevant frequencies. At the moment we cannot present a reason for this behavior and it is a question of ongoing research. Other open issues to enhance the overall performance are

- smoothing the estimated positions over time to prevent outlier.
- multi speaker tracking to prevent jumping back and forth between the estimated position.
- tracking of temporary as well as constant noise sources.

ACKNOWLEDGEMENT

This research was (partly) funded by grant *VWZN2420* (*"Lower Saxony Research Network Design of Environments for Ageing"*) from the *Ministry for Science and Culture of lower saxony*. The views and conclusions contained in this document, however, are those of the authors.

REFERENCES

[1] J. Bitzer and K. U. Simmer, "Superdirective Microphone Arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Berlin, Heidelberg, New York: Springer, May 2001, ch. 2, pp. 19–37.

[2] Lower Saxony Research Network Design of Environments for Ageing. (2010, May). [Online]. Available: http://altersgerechte-lebenswelten.de/index.php?id=21&L=1

[3] Offis e.V. (2010, May) Institute for Information Technology. [Online]. Available: http://www.offis.de/en/start.html

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

[5] Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," *IEEE*, no. 0-7803-5041-3/99, pp. 937–940, 1999.

[6] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source lcalization," *Journal of Acoustical Society of America*, vol. 107, no. 1, Jan. 2000.

[7] J. Chen, J. Benesty, and Y. A. Huang, "Time Delay Estimation using Spatial Correlation Techniques," *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 207–210, Sep. 2003.

[8] G. Doblinger, "Localization and Tracking of Acoustical Sources," in *Topics in Acoustic Echo and Noise Control*, E. Haensler and G. Schmidt, Eds. Berlin: Springer Verlag, 2006, pp. 91–124.

[9] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," *IEEE HSCMA*, no. 978-1-4244-2338-5/08/, pp. 69–72, 2008.

[10] P. J. Rousseeuw, "Least Median of Squares Regression," *Journal of the American Statistical Association*, vol. 79, no. 388, Dec. 1984.

[11] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," *IEEE ICASSP*, no. 1-4244-1484-9/08/, pp. 4349–4352, 2008.

[12] M. Marzinzik and B. Kollmeier, "Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 2, pp. 109–118, Feb 2002.

[13] T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier, "Combined source tracking and noise reduction for application in hearing aids," in *8. ITG-Fachtagung Sprachkommunikation*, no. 8. VDE VERLAG GMBH, Oct. 2008.