Comparison of Different Algorithms for Acoustic Source Localization

Benedikt Loesch, Parisa Ebrahim and Bin Yang

Chair of System Theory and Signal Processing, University of Stuttgart
Email: {benedikt.loesch, bin.yang}@LSS.uni-stuttgart.de

Abstract

Recently, position estimation for acoustic signals has been studied intensively and many different algorithms have been proposed. The different methods can be classified into indirect (estimation of time-difference of arrival (TDOA) and then position) and direct (direct estimation of position) methods. Furthermore, they can be classified according to whether they use a model for the mixing channel (model-based) or not. In this paper, we compare three different algorithms for source localization: one non-model based indirect method (DATEMM), one non-model based direct method (SRP-PHAT) and one model-based direct method (ICA-SCT). We compare them with respect to the underlying concept and the localization performance using simulations and real room recordings. We evaluate the influence of number of microphones, number of sources, microphone arrangement and reverberation time.

1 Introduction

The task of acoustic source localization is to estimate the position of one or multiple sound sources by using an array of microphones. There are indirect localization methods over the explicit estimation of the time-difference of arrival (TDOA) such as DATEMM [1] and direct spatial scanning methods. Among the latter ones, some methods like ADP [2] and SCT [3] are model-based since they explicitely model and estimate the acoustic channel by using independent component analysis (ICA). Other algorithms like SRP-PHAT [4] and DATEMM are not model-based as they perform the localization without a channel model. Since many different approaches have been proposed, this paper aims at comparing three different algorithms: SRP-PHAT, DATEMM and ICA-SCT. The comparison is done with respect to the underlying concept and the localization performance using simulations and real room recordings.

We assume a convolutive mixing model

$$x_m[i] = \sum_{n=1}^{N} h_{mn}[i] * s_n[i], \quad m = 1 \ldots M, \tag{1}$$

where $x_m[i]$ are the microphone signals, $s_n[i]$ are the source signals and $h_{mn}[i]$ are the impulse responses. The convolutive model in the time-domain can be transformed into an instantaneous mixing model in the time-frequency domain:

$$\mathbf{X}[k,l] \approx \mathbf{H}[k]\mathbf{S}[k,l] \tag{2}$$

where $1 \leq k \leq K$ is the frequency bin index and l is the time frame index.

One common approach for localization is a TDOA estimation by the generalized cross-correlation with phase transform (GCC-PHAT), followed by a position estimation. However, it suffers from several ambiguities:

- Multipath: GCC-PHAT shows multiple maxima and the direct path maximum is not necessarily the strongest one.
- Multiple sources: When multiple sources are present, GCC-PHAT shows multiple maxima and it is difficult to group the TDOAs originating from the same source.

These two ambiguities are shown in Fig. 1 which plots the GCC-PHAT for a two-source scenario in a medium-reverberant room. The two direct path TDOAs are indicated by dashed lines.

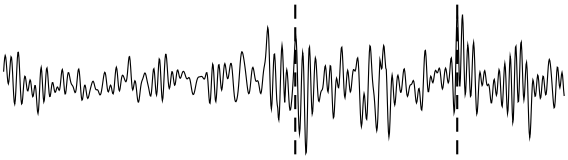

Figure 1: GCC-PHAT for two sources in a real room

2 Localization Methods

- **SRP-PHAT** combines the GCC-PHAT for all $|\mathscr{I}|$ microphone pairs with $\mathscr{I} \subseteq \{(a,b) | 1 \leq a < b \leq M\}$. The combination is done by summing the individual GCC-PHAT functions $c_{ab}[\cdot]$ evaluated at the expected TDOA $\tau_{ab}(\mathbf{p})$ for a potential source located at $\mathbf{p}$.

  $$\mathscr{H}_{\text{SRP-PHAT}}(\mathbf{p}) = \sum_{(a,b)\in\mathscr{I}} |c_{ab}[\tau_{ab}(\mathbf{p})]| \tag{3}$$

  SRP-PHAT aims to resolve the ambiguities of GCC-PHAT by combining multiple GCC-PHAT functions.

- **DATEMM** is based on the observation of two TDOA constraints implying information redundancy. By applying these constraints to TDOA estimates derived from GCC-PHAT, the ambiguity of TDOA estimation can be significantly reduced. The first constraint is the relationship between the extremum positions in the cross-correlation and autocorrelation of the microphone signals, called *raster condition*. The second important observation is: For each subset of microphones and the same number of corresponding direct or echo paths, a *zero cyclic sum condition* always holds for TDOAs originating from the same source. Using the raster condition we obtain sets of direct path TDOA estimates which are then synthesized into consistent TDOA graphs using the zero cyclic sum condition. Each consistent graph contains the TDOAs of a single source which are then used to estimate its position by an appropriate method.

- **ICA-SCT** is based on the "self-steering" capability of frequency-domain ICA. The ideal ICA solution is

  $$\mathbf{W}[k] \approx \mathbf{\Pi}[k]\mathbf{D}[k]\mathbf{H}^{-1}[k]. \tag{4}$$

  $\mathbf{D}[k]$ is a diagonal complex-valued scaling matrix and $\mathbf{\Pi}[k]$ is a permutation matrix. The ICA-SCT approach uses the inverse of the estimated demixing matrix $\mathbf{W}^{-1}[k] \sim \mathbf{H}[k]$ to compare an estimated propagation model with an assumed propagation model by using the so-called state coherence transform (SCT) [3, 5]. The main idea of the state coherence transform is to compare the "state" $e^{-j\omega\tau}$ from the propagation model $\mathbf{H}[k]$ against its estimate from the result of ICA. The position estimation is achieved by spatial scanning as in SRP-PHAT. For frequency-domain ICA, we use the approach from [6].

Table 1 summarizes key properties of the different algorithms. SRP-PHAT uses the simplest TDOA disambiguation by summing over microphone pairs, while DATEMM uses the raster condition and the concept of consistent TDOA graphs. ICA-SCT relies on system identification by ICA to yield state vectors containing the TDOAs of a single source. Another important aspect is the computational complexity: DATEMM uses a TDOA estimation for each microphone pair and then forms consistent graphs to directly estimate the location of the sources. A high resolution spatial scanning is not necessary. Hence, it is fast and can be implemented in real-time [1]. SRP-PHAT is a direct spatial scanning method and hence requires the computation of the cost function over a two- or three-dimensional grid. The search grid resolution directly influences the localization accuracy. Although there are methods like multi-resolution approaches to overcome the high computation cost, SRP-PHAT is still considerably more computationally intensive than DATEMM. ICA-SCT has the highest computational complexity among the three methods, since it first uses an ICA algorithm and then performs spatial scanning as in SRP-PHAT.

method	direct	model	TDOA disambiguation
SRP-PHAT	yes	no	over microphone pairs
DATEMM	no	no	raster condition, consistent graphs
ICA-SCT	yes	yes	system identification by ICA, state vector contains TDOAs of a single source

Table 1: Comparison of the different methods

3 Experiments

We compare the performance of the different methods using simulations with the MATLAB RoomSim toolbox [7] and real room recordings. Although we use stationary source positions, we evaluate the different methods under the assumption that the sources could slowly move. This means that we assume a maximum usable data length of $\approx 500\,\text{ms}$. We study the influence of the microphone set, number of microphones, number of sources and reverberation time on each method. Fig. 2 shows the microphone and source positions for the real room. We use $M = 4$ or $M = 8$ microphones and for each case we evaluate two different microphone sets, shown in Table 2. For $M = 8$, the microphones are located at the left and bottom side of the sources in set 1, while the sources are surrounded by microphones in set 2. In the simulation, we use similar microphone and source positions.

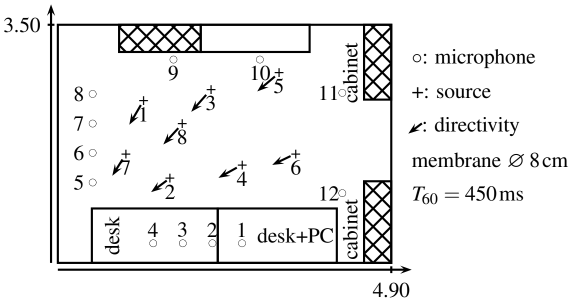

Figure 2: Room layout and experimental setup

	microphone set 1	microphone set 2
$M = 4$	$1, 2, 7, 8$	$1, 4, 5, 8$
$M = 8$	$1, \cdots, 8$	$1, 4, 5, 8, 9, 10, 11, 12$

Table 2: Microphone sets

For SRP-PHAT and ICA-SCT we perform spatial scanning with a grid resolution of 1 cm and locate the sources by finding the N highest peaks (with minimum distance of 20 cm) of the cost function in each block. For DATEMM, we accept all estimated positions (possibly $> N$) regardless of the their quality measure. Unfortunately, different localization algorithms might require different block sizes to work optimally. SRP-PHAT and DATEMM use non-overlapping frames of length 170.67 ms and a sampling frequency of 96 kHz. ICA-SCT uses a block size of 512 ms and a sampling frequency of 16 kHz. To make a fair comparison, we cluster the positions estimated by SRP-PHAT and DATEMM in a time interval of 512 ms by agglomerative hierarchical clustering using the nearest neighbour distance metric D between clusters. The clustering algorithm starts with each estimated position as an individual cluster. It then repeatedly merges the two clusters with minimum D if $D < 20\,\text{cm}$. Otherwise the clustering procedure stops. Note, that the number of resulting clusters can be larger than N. The final estimated positions in each time interval of 512 ms are given by the cluster centroids.

For the evaluation, we associate each estimated position $\hat{\mathbf{p}}_i$ with the true source position $\mathbf{p}_n$ if $d_i = \min_{\tilde{n}} \|\hat{\mathbf{p}}_i - \mathbf{p}_{\tilde{n}}\| = \|\hat{\mathbf{p}}_i - \mathbf{p}_n\|$ is smaller than a threshold t (e.g. 20 cm). In each time interval of 512 ms, every $\mathbf{p}_n$ can only be associated with the closest $\hat{\mathbf{p}}_i$. Non-associated $\hat{\mathbf{p}}_i$ with $d_i > t$ are considered as false positives. For the complete recording of 24 s and each $\mathbf{p}_n$, we count the number of associated $\hat{\mathbf{p}}_i$ and calculate from these the mean localization error $\bar{d}_n$ of each source n. Important performance criteria are the mean localization error, the number of correct source detections (true positives TP_n) for source n, the total number of false positives (FP) and the total number of non-associated $\hat{\mathbf{p}}_i$ with $d_i < t$ (NA). Ideally, an algorithm should yield a low mean localization error with a large TP_n and small FP and NA. In the result tables, we normalize TP_n, FP and NA by the number of blocks and give the resulting value in percent. Note that FP and NA can be larger than 100%. For the ICA-SCT algorithm, the sum of TP_n $(1 \leq n \leq N)$, FP and NA should be N. For DATEMM and SRP-PHAT, this sum could be larger or smaller than N, because these two algorithms do not return exactly N source position estimates per 512 ms block. In Table 3, for example, TP_1=91 means that SRP-PHAT detects source 1 in 91% of the 512 ms blocks. FP=26 and NA=45 mean that there are on average 0.26 false positives and 0.45 non-associated estimates per block.

In the simulations, we have used source 1 and 2 for $N = 2$ and source 1 to source 4 for $N = 4$. In general, all algorithms allow an accurate localization with a mean localization error of less than 7 cm, even for a reverberation time of $T_{60} = 300\,\text{ms}$ (see Table 3). However, the localization error of DATEMM is in most cases higher than that of SRP-PHAT or ICA-SCT.

		microphone set 1		microphone set 2	
	T_{60}	100 ms	300 ms	100 ms	300 ms
DATEMM	TP_n	100 96	93 100	98 91	96 98
	$\bar{d}_n[cm]$	5.6 4.4	6.1 6.3	2.7 1.1	3.5 1.6
	FP/NA	67/74	489/178	20/41	230/161
SRP-PHAT	TP_n	93 91	93 91	93 96	93 91
	$\bar{d}_n[cm]$	0.8 1.7	1.9 3.2	0.0 0.2	0.1 0.4
	FP/NA	26/46	37/59	30/61	43/59
ICA-SCT	TP_n	70 89	7 74	89 93	76 91
	$\bar{d}_n[cm]$	1.5 1.3	1.1 4.4	0.0 0.1	0.3 1.7
	FP/NA	41/0	93/26	17/0	33/0

Table 3: Simulation results for $N = 2$ and $M = 8$

		microphone set 1								microphone set 2							
	T_{60}	100 ms				300 ms				100 ms				300 ms			
DATEMM	TP$_n$	91	91	48	96	93	100	0	91	98	100	76	59	91	89	50	33
	$\bar{d}_n[cm]$	4.1	4.8	2.6	2.5	5.0	6.9	–	4.4	2.8	0.8	1.4	1.8	4.4	1.7	1.5	3.4
	FP/NA	78/148				283/209				22/43				291/191			
SRP-PHAT	TP$_n$	93	89	50	91	93	85	46	87	96	93	80	80	93	87	63	65
	$\bar{d}_n[cm]$	1.9	3.8	3.1	0.1	3.3	4.4	4.1	0.4	0.1	0.1	0.3	0.4	0.1	0.4	0.3	0.6
	FP/NA	74/143				137/211				126/228				243/243			
ICA-SCT	TP$_n$	85	83	11	93	2	30	0	85	93	91	83	83	72	85	35	48
	$\bar{d}_n[cm]$	1.2	1.0	9.4	0.1	19.1	3.2	–	1.9	0.0	0.2	0.2	0.4	0.2	1.2	0.7	4.5
	FP/NA	111/17				248/30				37/13				130/30			

Table 4: Simulation results for $N = 4$ and $M = 8$

		microphone set 1								microphone set 2							
	source	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
DATEMM	TP$_n$	85	80	76	89	54	87	74	85	15	20	65	50	59	70	0	57
	$\bar{d}_n[cm]$	4.9	3.9	5.3	7.6	11.2	9.0	5.5	5.7	12.9	10.5	7.0	10.7	11.3	10.6	–	6.5
	FP/NA	24/35		50/57		172/72		20/35		20/2		2/22		4/20		4/0	
SRP-PHAT	TP$_n$	96	89	87	93	83	96	89	89	91	96	89	93	91	96	83	91
	$\bar{d}_n[cm]$	3.5	2.4	3.9	7.2	10.4	8.5	5.9	3.7	5.6	6.2	6.3	10.2	8.8	7.9	8.7	8.0
	FP/NA	30/57		37/67		39/96		35/65		61/85		48/76		65/78		122/109	
ICA-SCT	TP$_n$	87	72	80	83	70	91	74	72	76	67	74	78	74	72	26	72
	$\bar{d}_n[cm]$	2.6	2.1	2.1	4.3	8.8	6.9	5.2	2.1	3.3	2.3	2.2	6.3	8.2	7.0	6.1	3.1
	FP/NA	39/2		35/2		37/2		52/2		57/0		48/0		54/0		98/4	

Table 5: Real room results for $N = 2$ and $M = 8$

		$M = 8$								$M = 4$							
		microphone set 1				microphone set 2				microphone set 1				microphone set 2			
DATEMM	TP$_n$	74	76	24	26	4	7	4	0	9	4	0	2	4	13	2	2
	$\bar{d}_n[cm]$	4.4	3.9	6.7	8.4	9.2	10.1	5.2	–	14.0	10.8	–	12.2	16.6	12.5	11.3	4.7
	FP/NA	20/46				7/0				46/0				15/0			
SRP-PHAT	TP$_n$	91	93	72	61	93	87	65	76	91	35	76	26	89	87	35	43
	$\bar{d}_n[cm]$	5.3	2.6	3.5	6.6	4.9	6.0	6.0	9.4	3.3	2.2	7.3	5.2	3.3	2.5	5.1	6.6
	FP/NA	117/211				257/230				350/167				374/209			
ICA-SCT	TP$_n$	85	72	37	30	74	52	22	9	80	9	39	11	65	54	26	35
	$\bar{d}_n[cm]$	2.7	2.0	2.5	5.5	2.7	2.4	4.6	4.4	5.0	6.0	8.1	7.9	3.6	2.7	2.4	6.4
	FP/NA	163/13				220/24				220/41				185/35			

Table 6: Real room results $N = 4$ (source 1 to 4)

		sources 1,2,7,8												sources 3,4,5,6											
	method	short block (sb)				sb + clustering				long block				short block (sb)				sb + clustering				long block			
DATEMM	TP$_n$	29	20	11	18	65	41	30	39	72	51	19	19	18	27	30	44	46	48	46	74	38	57	53	68
	$\bar{d}_n[cm]$	4.6	2.4	5.6	4.1	4.4	4.7	5.6	5.7	4.5	3.0	5.5	3.8	5.8	6.3	9.6	8.8	6.3	8.2	10.2	9.1	6.9	6.0	10.0	8.1
	FP/NA	39/57				35/28				34/109				216/101				141/111				383/219			
SRP-PHAT	TP$_n$	78	73	50	51	91	91	87	83	87	76	60	47	41	50	62	74	72	74	85	96	51	47	64	87
	$\bar{d}_n[cm]$	3.1	2.6	5.6	3.4	4.1	2.7	5.7	3.4	3.1	2.3	5.5	2.6	3.5	7.0	10.2	8.5	3.4	7.3	9.8	8.6	3.2	6.5	10.0	8.0
	FP/NA	114/32				115/196				118/13				120/53				146/233				84/67			
ICA-SCT	TP$_n$	68	52	38	40	87	89	76	76	80	65	43	63	36	46	55	69	76	87	89	93	52	59	63	89
	$\bar{d}_n[cm]$	3.0	2.1	5.6	2.5	3.1	2.3	5.7	2.8	2.6	1.9	5.3	2.0	2.8	5.9	8.8	7.1	2.9	6.3	8.8	7.7	2.2	5.0	8.7	7.0
	FP/NA	176/26				327/231				130/17				156/38				276/229				100/37			

Table 7: Real room results, $N = 4$, $M = 8$, microphone set 1

From the *simulation results*, we find:

- *Influence of microphone set:* As shown in Table 3 and 4, for $M = 8$ microphones, all methods yield lower localization error with set 2 than with set 1. This is as expected, since for omnidirectional sources and microphones, a microphone set surrounding the sources is optimal [8].
- *Influence of number of sources:* As the number of sources increases from $N = 2$ (Table 3) to $N = 4$ (Table 4), it becomes more difficult to estimate all sources in all blocks. Hence, TP decreases for all algorithms. Furthermore, we can observe that the mean localization error $\bar{d}_n$ for $N = 4$ is comparable to $N = 2$. Only for microphone set 1, large reverberation time, and large source numbers ($N = 4$ or $N = 6$ (not shown)), DATEMM and ICA-SCT fail to detect all sources. However, TP for the not detected sources is already low at $T_{60} = 100$ ms for all algorithms.
- *Influence of reverberation time:* An increase in the reverberation time from $T_{60} = 100$ ms to $T_{60} = 300$ ms clearly leads to a larger localization error and a larger FP for all algorithms. The degree of the performance loss differs considerably for different microphone sets. Hence it is difficult to draw general conclusions.

Table 5 to 7 summarize the results of real room experiments. In most cases, all three methods allow a pretty accurate localization, though DATEMM somteimes fails to detect all sources. In the real room experiments, we have tried different source combinations as shown in the first or second row of Table 5 and 7.

From the *real room results*, we conclude:

- *Influence of source position:* The localization error increases when the source to microphone distance increases, as shown in Table 5 for microphone set 1 and $M = 8$. The sources 1 and 2 are quite close to the microphone array and hence the localization error is small for all methods. In comparison, the sources 5 and 6 are far away and hence the localization error is higher. Among the three methods, ICA-SCT shows the smallest localization error.
- *Influence of microphone geometry:* For $M = 8$, all three methods show a larger localization error with microphone set 2 as shown in Table 5 and 6. This is in contrast to the results with room simulations in Table 3 and 4. Furthermore, SRP-PHAT and ICA-SCT show more false positives with microphone set 2. A possible explanation is the directivity of the sources which influences the direct-to-reverberant ratio at each microphone. In the real room experiment, the loudspeakers were approximately facing the lower left corner of the room (see Fig. 2). Hence, the loudspeakers were facing away from the microphones 9 to 12 and the direct-to-reverberant ratio was low at these microphones. As a consequence, set 2 performs worse than set 1.
- *Influence of number of microphones:* DATEMM sometimes fails to estimate all source positions, in particular if the number of microphones is small. This has been verified in several experiments not shown here. Furthermore, TP for DATEMM is rather small. A possible explanation is that with a small number of microphones it is more difficult to find consistent graphs with a sufficient number of TDOAs to properly locate the sources.
- *Effect of block size:* Table 7 compares results for different block sizes and post-processing (clustering). The short block size is 170.67 ms for SRP-PHAT and DATEMM, and 128 ms for ICA-SCT. The long block has a length of 512 ms for all three methods. SRP-PHAT works acceptably well for small and large block sizes and also for the clustering. Compared to the large block size, the use of a short block size with clustering leads to an increase in the TP and only to a small increase in localization error. For DATEMM, the use of a small block size combined with clustering gives a good TP, small FP and acceptable localization error. ICA-SCT performs best with a large block size because the ICA algorithm needs sufficient data to converge to a demixing matrix which is approximately the inverse of the mixing matrix. With a small block size and no clustering, the TP are rather low and FP is quite high, whereas clustering leads to a higher TP but dramatically increased FP.

4 Conclusion

In this paper, we have compared three different algorithms for acoustic source localization:

- DATEMM, which is based on TDOA estimation and consistent graphs, has the lowest computational complexity. It allows quite accurate localization but has problems when the number of microphones is small. It works best with a small block size and clustering.
- SRP-PHAT, which is based on a sum of individual GCC-PHAT functions for each microphone pair and spatial scanning, has a much higher computational complexity. It performs very well and allows accurate localization even with a small number of microphones. Its performance is not very sensitive to the block size.
- ICA-SCT, which is based on a mixing channel identification through ICA and spatial scanning, has the highest computational complexity. It shows lower localization error than SRP-PHAT, especially when the sources are far away from the microphones. However, compared to SRP-PHAT with small block size and clustering, the detection rate is lower. Furthermore, ICA-SCT is more sensitive to a short block size because ICA algorithms require a sufficient amount of data to work properly. Another aspect is that ICA suffers from poor convergence for distributed microphone arrays. However, [9] alleviates this problem.

When comparing simulation results with real room results, we notice that the directivity of both sources and microphones do have a great impact on the localization performance. Most simulations up to now assumed omnidirectional sources and microphones and hence their results are not realistic.

References

[1] J. Scheuing and B. Yang, “Disambiguation of TDOA estimation for multiple sources in reverberant environments,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1479–1489, Nov. 2008.

[2] A. Lombard, T. Rosenkranz, H. Buchner, and W. Kellermann, “Multidimensional localization of multiple sound sources using averaged directivity patterns of blind source separation systems,” *Proc. ICASSP*, 2009.

[3] B. Loesch, S. Uhlich, and B. Yang, “Multidimensional localization of multiple sound sources using frequency domain ICA and an extended state coherence transform,” *Proc. IEEE Workshop on Statistical Signal Processing (SSP)*, 2009.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein, Ed. Springer Verlag, 2001.

[5] F. Nesta, M. Omologo, and P. Svaizer, “A novel robust solution to the permutation problem based on a joint multiple TDOA estimation,” *Proc. International Workshop for Acoustic Echo and Noise Control (IWAENC)*, 2008.

[6] F. Nesta, P. Svaizer, and M. Omologo, “Separating short signals in highly reverberant environment by a recursive frequency domain BSS,” *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2008.

[7] E. Lehmann, “Image-source method for room impulse response simulation (room acoustics),” `http://www.watri.org.au/~ericl/ism_code.html`, 2008.

[8] B. Yang and J. Scheuing, “A theoretical analysis of 2D sensor arrays for TDOA based localization,” *Proc. ICASSP*, 2006.

[9] F. Nesta and M. Omologo, “Cooperative Wiener-ICA for source localization and separation by distributed microphone arrays,” *Proc. ICASSP*, 2010.