

# TUT Acoustic Source Tracking System 2007

Teemu Korhonen, and Pasi Pertilä

Tampere University of Technology,  
Institute of Signal Processing, Audio Research Group  
P.O.Box 553, 33101, Tampere, Finland  
{teemu.korhonen, pasi.pertila}@tut.fi

**Abstract.** This paper is a documentation of the acoustic person tracking system developed by TUT. The system performance was evaluated in the CLEAR 2007 evaluation. The proposed system is designed to track a speaker position in a meeting room domain using only audio data. In the CLEAR 2007 evaluation the audio data consists of recordings from multiple microphone arrays. The meeting rooms are equipped with three to seven arrays.

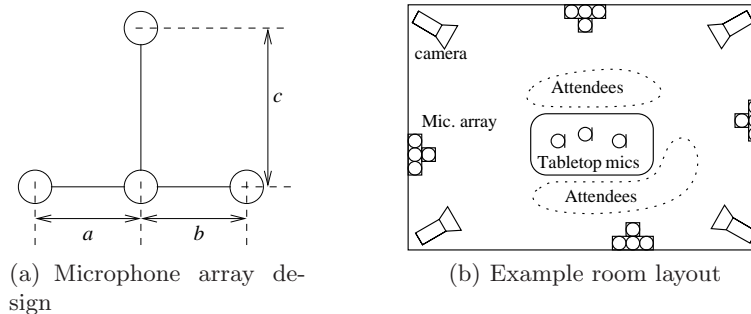
Speaker localization is performed by mapping pairwise cross-correlations of microphone signals into a three dimensional likelihood field. The resulting likelihood is used as source evidence for a particle filtering algorithm. A point estimate for the speaker position for each time frame is derived from the resulting sequential process. Results indicate an 85 % success rate of localization with 15 cm average accuracy.

## 1 Introduction

The CLEAR 2007 evaluation campaign was composed of multiple different tracking and identification tasks. This work proposes a system tackling the 3D person tracking subtask using the provided audio data. The audio recordings contain data sets from different meeting room environments. Each setting comprises of multiple microphone arrays with known coordinates. Performance of the system is evaluated by outside party of TUT using metrics measuring both accuracy and precision.

The TUT source tracking system is based on Bayesian framework where measurement update is done using likelihoods inferred by pairwise cross-correlations of microphone signals. Source tracking algorithm is based on particle filtering where distributions are presented as set of points.

The following section deals with evaluation tasks and metrics associated with them. Section 3 describes the TUT acoustic person tracking system with details of its implementation. The results are given in Sect. 4 with related discussion in Sect. 5. The work is concluded in section 6.



**Fig. 1.** Geometry related to data gathering is illustrated. The geometry of a four microphone T-array is presented in panel 1(a). Dimensions  $a$  and  $b$  are 20 cm and  $c$  is 30 cm for every site except IBM, where the corresponding dimensions are 26 cm and 40 cm. Panel 1(b) illustrates a basic recording room layout for a meeting, equipped with different sensors. Microphone arrays, used by the TUT system, are mounted to the walls.

## 2 Evaluation

### 2.1 Tasks

One of the CLEAR 2007 evaluation tasks is the 3D person tracking, which has been further divided into subtasks using audio or video data, or a combination of both. The TUT system participates in “Acoustic Person Tracking” where the goal is detection and tracking of a speaker using available far-field microphones [1].

### 2.2 Data

A short description about the data used in system development and a performance evaluation is given here. For further details refer to the evaluation plan [1] or description of the CHIL audiovisual corpus [2].

The CHIL corpus is a collection of video and audio recordings from different meeting room settings collected during years 2003–2006. Accompanying the corpus is a manually generated annotation for any notable activities. the 3D person tracking task uses a set of recordings from five different sites listed in Table 1.

The basic setting for each site consists of microphone arrays of different geometries and sizes. The smaller arrays have four microphones in a two dimensional upside-down T-shaped form (Fig. 1). These arrays have been mounted on the walls and Table 1 documents their number per site. Larger arrays housing 64-microphones in a linear setup are not used by the TUT system.

The audio data has been divided into development and testing sets. The development set accompanied by complete annotation is 2.8 hours in length. The testing set contains eight recordings per site, each recording five minutes in

length (with total of 3.3 hours of data). The nature of the recordings is interactive seminar, where the active speaker turn can switch during audience participation.

The audio data was sampled at 44.1 kHz with 24 bit resolution. The reference data for active speaker 3D location was given with time resolution of 1.0 s.

### 2.3 Metrics

The CLEAR 2007 Evaluation uses a set of metrics for analyzation of the system performance. Proposed metrics measure performance in terms of precision and accuracy. Further details and discussion can be read from [1].

Acoustic person tracking accuracy is evaluated with the metrics similar to video person tracking task. However, identity mismatches are removed from the calculations since audio tracking is not expected to distinguish between different speakers. Following notations are used for different metrics and related types of error:

- MOTP [mm]: Multiple Object Tracking Precision
- MISS [%]: number of misses out of ground truth points
- FALSEPOS [%]: number of false positives out of ground truth points
- A-MOTA [%]: (Audio-) Multiple Object Tracking Accuracy

The MOTP metric is defined as an average Euclidean distance for non-missed positions. Errors divided into misses and false positives are classified using a threshold set to 500 mm. The ground truth points without a tracker hypothesis within the threshold are classified as misses. False positives are, on the other hand, hypothesis points without ground truth support. The combination of both of the errors is reflected in A-MOTA metric.

**Table 1.** Information about the data used in the evaluation is presented. The smart rooms used in the evaluation are: Society in Information Technologies at Athens Information Technology, Athens, Greece (AIT); the IBM T.J. Watson Research Center, Yorktown Heights, USA (IBM); the Centro per la ricerca scientifica e tecnologica at the Instituto Trentino di Cultura, Trento, Italy (ITC-irst); the Interactive Systems Labs of the Universitat Karlsruhe, Germany (UKA); and the Universitat Politecnica de Catalunya, Barcelona, Spain (UPC). [2]

		Data length [minutes]		Room size [m]			
Site	T-arrays	Dev set	Eval set	$x$	$y$	$z$	Recording type
UKA	4	44	8 * 5	5.9	7.1	3.0	Meeting
ITC	7	31	8 * 5	4.7	5.9	4.5	Meeting
IBM	4	37	8 * 5	7.2	5.9	2.7	Meeting
UPC	3	23	8 * 5	4.0	5.2	4.0	Meeting
AIT	3	31	8 * 5	5.0	3.7	2.6	Meeting

### 3 System description

The proposed acoustic source tracking system is based on Bayesian framework utilizing particle filtering. Audio signals from microphones are processed pairwise within each array using a derivative cross-correlation method. Combination of the resulting likelihoods follows in a room-encompassing spatial likelihood field which the particle filtering algorithm processes as a source evidence. A point estimate can be extracted from the filtering process for each frame of interest.

#### 3.1 Sound source likelihood estimation

It is assumed that when an active speaker is present the received signals between spatially separate microphones inside an array differ mainly in the reception time. The microphone data is processed in frames of length  $L$ . The signal from microphone  $i$  belonging to array  $a \in [1, N]$  is denoted  $\mathbf{x}_i^a(t) = [x_i^a(t-L+1), \dots, x_i^a(t)]^T$ . The similarity of two signals from spatially separated microphones  $i, j$  from array  $a$  is estimated using the generalized cross-correlation (GCC) with the PHAT weighting algorithm [3]

$$r_{i,j}^a(\tau) = \mathcal{F}^{-1} \left\{ \frac{X_i^a(\omega_u)X_j^{a*}(\omega_u)}{|X_i^a(\omega_u)X_j^{a*}(\omega_u)|} \right\}, \omega_u = 0, \dots, L-1, \quad (1)$$

where  $\mathcal{F}^{-1}\{\cdot\}$  denotes inverse discrete Fourier transform (IDFT) and  $X_i^a(\omega_u)$  is the DFT of microphone signal  $\mathbf{x}_i^a(t)$ ,  $\omega_u$  is the  $u$ th frequency sample,  $\{\cdot\}^*$ ,  $\{\cdot\}^T$  denote complex conjugate transpose and transpose respectively, and  $t, \tau$  are discrete time indices. The window length was set to 44100 samples and an overlap of 22050 samples was used. The system therefore produces location estimates at the rate of 0.5 second.

From any hypothetical point  $\mathbf{h}$  a discrete time difference of arrival (TDOA) value  $\Delta\tau_{\mathbf{h},i,j}^a$  can be calculated between a microphone pair  $i, j$  in array  $a$

$$\Delta\tau_{\mathbf{h},i,j}^a = Q \left[ \frac{f_s \cdot (\|\mathbf{h} - \mathbf{m}_i^a\| - \|\mathbf{h} - \mathbf{m}_j^a\|)}{c} \right], \quad (2)$$

where  $c$  is the speed of sound,  $f_s$  is the sampling rate,  $\mathbf{m}_i^a$  denotes the microphone  $i$  position and  $Q[\cdot]$  is a quantization operator. Here  $c$  was set to 343 m/s.

The similarity function (Eq. 1) is indexed with the TDOA value (Eq. 2) to get a likelihood of source location  $\mathbf{h}$  from a single pairwise similarity measure

$$L(\mathbf{h}(t)|\mathbf{x}_i^a(t), \mathbf{x}_j^a(t)) = r_{i,j}^a(\Delta\tau_{\mathbf{h},i,j}^a). \quad (3)$$

The likelihoods from all pairwise similarities for point  $\mathbf{h}$  are combined via multiplication. The multiplication of normalized likelihoods can be interpreted as a logical “and” operation. Only points in which all pairwise similarities are significant the source likelihood can be significant. This differs from the approach of summing up non-negative likelihood values [4][5] which can be interpreted

as a logical “or” operation. This approach has not been used previously elsewhere according to the knowledge of the authors. The pairwise similarity estimation (Eq. 1) is performed for every microphone pair within each microphone array. Each array consists of four microphones. A total of six microphone pairs are used per array. The pairwise similarity values for point  $\mathbf{h}$  are first multiplied together between all pairs inside each microphone array. Then the resulting array likelihoods are combined. The likelihood of a sound source existing in point  $\mathbf{h}(t)$  at discrete time  $t$  can now be written as

$$L(\mathbf{h}(t)|\mathbf{X}^1(t), \dots, \mathbf{X}^N(t)) = \prod_{a=1}^N \prod_{\substack{i=3, j=4 \\ i=1, \\ j=i+1}} r_{i,j}^a(\Delta\tau_{\mathbf{h},i,j}), \quad (4)$$

where  $r_{i,j}^a$  denotes pairwise similarity of microphone signals  $i, j$  from array  $a = 1, \dots, N$  and the data from array  $a$  is written as

$$\mathbf{X}^a(t) = [\mathbf{x}_1^a(t), \dots, \mathbf{x}_4^a(t)].$$

### 3.2 Source tracking

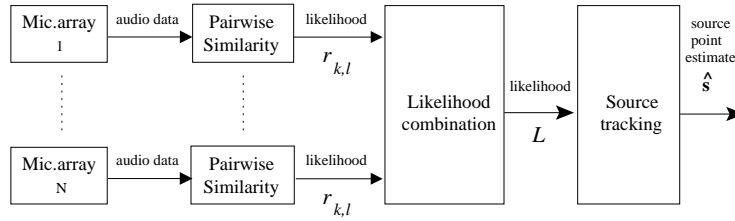
The source tracking was performed using sequential Monte Carlo method known as particle filtering. Specifically, the Sampling Importance Resampling (SIR) algorithm is used [6] where transition prior is used as importance function.

Particle filtering approximates a probability density function (pdf) with a set of  $M$  weighted random samples  $\mathcal{X}_t = \{\mathbf{s}_t^{(n)}, w_t^{(n)}\}_{n=1}^M$  for each time instant  $t$ . The samples known as particles are propagated over time and resampled according to their fit on the evidence from measurements. A point approximation from the particle set can be evaluated with many different methods. Here, a median of particle positions  $\mathbf{s}_t$  is used due to its robustness. Median is calculated separately for each of the three dimensions.

The initial set of particles  $\mathcal{X}_0$  is sampled from an uniform distribution constrained within room coordinates and at the height of 1.2 to 1.8 m. The number of particles is constant for any frame and set to  $M = 50,000$ . During each iteration particles are sampled from importance function modeled as multi-Gaussian prior (with deviations of 50 mm and 200 mm). Particle weights indicating fitness of each particle are calculated from the spatial likelihood field given in (4) and the particles are resampled according to the weights gained.

### 3.3 System output

The system outputs a speaker location hypothesis every 0.5 second interval. The results from TUT source tracking system were evaluated against annotation from 3D labels for acoustic person tracking subtask.



**Fig. 2.** A block diagram showing flow of information within presented system. Microphone signals are processed pairwise to produce source likelihoods. The combination of those likelihoods is then used as measurement evidence for source tracking system.

## 4 Results

Evaluation scores were calculated using metrics specified in 2.3 and results are presented in Table 2 with specifics in Table 4.

### 4.1 Computation time

The system was implemented and run in Matlab using only internal libraries. The computer used was a 3.20 GHz Intel Pentium 4 with 2 GB of RAM. Running environment was Linux based. The total processing time is given in Table 3.

A realtime factor (RT) of 5.08 was achieved with given setup using the eval data set. If all processed data is modeled as a single concatenated mono channel (equaling 206296 s), the RT factor would be 0.30.

## 5 Discussion

The task evaluated was 3D tracking of multiple persons using acoustic recordings from a set of microphones. Applied system performed well regardless the lack of

**Table 2.** TUT system’s evaluation scores for the both data sets. The tasks are defined in Section 2.1 and the metrics are defined in Section 2.3. The scores for the CLEAR 2006 evaluation are presented for comparison: acoustic multiple person tracking (MPT-A) and single person tracking (SPT-A).

MOT Scores	CLEAR 2007		CLEAR 2006	
	eval	dev	MTP-A	SPT-A
MOTP [mm]	<b>152</b>	165	334	245
MISS [%]	<b>14.96</b>	7.91	83.32	27.93
FALSEPOS [%]	<b>37.00</b>	24.23	83.22	27.86
A-MOTA [%]	<b>48.04</b>	67.85	-66.53	44.21

**Table 3.** The total processing times of the TUT system. The IO overhead is the time spent by the system while loading audio data and is accounted for in the RT factor estimation.

+ GCC-PHAT	25992 s
– IO overhead	5492 s
+ Particle filter	40365 s
=	60865 s

speech activity detection (SAD). This naturally reflects directly as a degradation of the false positives metric when every frame is tagged as source. Large number of false positives also affect the A-MOTA metric directly.

Detailed evaluation results with recording specific metrics are in Table 4.

## 6 Summary

A system for acoustic person tracking was presented complying the task presented in CLEAR 2007 evaluation. The acoustic tracking system uses data provided by microphone arrays with known coordinates. Measurements are processed pairwise within each microphone array with cross-correlation producing a similarity value for each time delay. The similarity values are processed as

**Table 4.** The CLEAR 2007 evaluation results for accuracy and precision separated with sites and recordings for the proposed system is presented. Anomalous results with values over twice the average have been bolded for clarity.

Site	Metric	Recording							
		1	2	3	4	5	6	7	8
AIT	MISS [%]	12.7	8.8	13.1	16.3	13.3	<b>60.5</b>	24.4	15.8
	MOTP [mm]	178	166	207	156	164	263	243	213
IBM	MISS [%]	19.2	22.1	11.5	7.5	16.1	4.6	3.2	4.5
	MOTP [mm]	198	184	171	154	227	112	213	202
ITC	MISS [%]	6.2	6.7	4.1	<b>40.6</b>	13.3	47	16.6	9.9
	MOTP [mm]	105	138	113	106	123	94	135	134
UKA	MISS [%]	7.8	9.1	7.1	8.8	<b>48.4</b>	7.8	<b>30.2</b>	15.6
	MOTP [mm]	128	107	88	123	97	101	145	123
UPC	MISS [%]	8.3	8.2	10.2	8.5	15.6	<b>44.3</b>	23.7	14.8
	MOTP [mm]	142	128	128	147	169	282	160	143

spatial likelihoods and combined over all microphone pairs. Resulting likelihood is used in particle filtering algorithm as speaker location evidence and a point estimate for speaker location is derived from the sequential process. The system is accurate to 15 cm for roughly 85 % of the time.

## 7 Acknowledgments

The authors wish to thank Mikko Parviainen, Tuomo Pirinen, Sakari Tervo and Ari Visa for their efforts in the development of the presented acoustic person tracking system.

## References

1. Bernardin, K.: Clear 2007 evaluation plan v.1.0. <http://isl.ira.uka.de/clear07/downloads/?download=CLEAR07-3DPT-2007-03-09.pdf> (2007)
2. Mostefa, D. *et al.*: The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. *Journal of Language Resources and Evaluation* (2006) (Submitted to).
3. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics, Speech, and Signal Processing* **4** (1976) 320–327
4. Aarabi, P.: The Fusion of Distributed Microphone Arrays for Sound Localization. *EURASIP Journal on Applied Signal Processing* **4** (2003) 338–347
5. DiBiase, J., Silverman, H., Brandstein, M.: 8. In: *Microphone Arrays*. Springer-Verlag (2001)
6. Gordon, N., Salmond, D., Smith, A.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F* **140** (1993) 107–113