

# A Person Tracking System for CHIL meetings

Alessio Brutti\*  
brutti@itc.it

FBK-irst,  
Via Sommarive 18,  
38050 Povo di Trento,  
Italy

**Abstract.** This paper presents the audio based tracking system designed at FBK-irst laboratories for the CLEAR 2007 evaluation campaign. The tracker relies on the Global Coherence Field theory that has proved to efficiently deal with the foreseen scenarios. Particular emphasis is given to the post-processing of localization hypotheses which guarantees smooth speaker trajectories and is crucial for the overall performance of the system. The system is also equipped with a speech activity detector based on HMM. The performance delivered by the proposed tracker presents a considerable gain with respect to the previous evaluation.

## 1 Introduction

The goal of an audio based person tracker is to estimate the position of an active speaker from a set of acoustic measurements. The CLEAR 2007 evaluation is focused on meetings that are held in smart rooms equipped with a Distributed Microphone Network (DMN) as envisioned in the CHIL project [1]. In particular, the DMN implementation adopted in the evaluation consists in a set of T-shaped microphone arrays and the 64 channels NIST MarkIII linear array. The meeting chunks taken in consideration for the evaluation involved 4 or 5 participants and showed considerably different levels of interaction. In some cases silence and noise sources were predominant while in other cases speech was present most of the time. As in the previous evaluation, a reference file, including speaker coordinates and speech/silence information, was available for each meeting with 1 s time resolution. The performance is measured in terms of localization precision (MOTP) and localization accuracy (A-MOTA), the latter accounting for gross errors, misses and false positives. Refer to the evaluation plan [2] for further details about the adopted metrics.

The audio based tracker designed at FBK-irst laboratories for the CLEAR 2007 evaluation campaign is a direct evolution of the system adopted in the previous evaluation [11]. In particular the system was modified in order to fit the meeting scenario that was adopted in this evaluation. Some efforts were also devoted to devise a more efficient implementation in an attempt to reduce the overall computational load.

---

\* This work was partially supported by the EU under the IP CHIL and STREP DICIT

## 2 Person Tracking based on Global Coherence Field

Almost all the localization algorithms presented in the literature rely on an analysis of the time difference of arrivals (TDOAs) at two or more microphones which, in far field conditions, is directly related to the direction of arrivals (DOAs) of the acoustic waves. The most common approach exploits the phase of the cross-spectrum of the received signals through the well-known Generalized Cross Correlation PHASE Transform (GCC-PHAT) [15, 16]. Let us assume that a sensor pair  $i$ , consisting of the microphones  $m_{i1}$  and  $m_{i2}$ , is available and let us denote as  $x_{i1}(n)$  and  $x_{i2}(n)$  the digitalized sequences captured by the sensors, where  $n = 0, \dots, L - 1$  and  $L$  is the analysis window length. The GCC-PHAT at microphone  $i$  and time instant  $t$  is computed as follows:

$$C_i(t, l) = FFT^{-1} \left\{ \frac{FFT(x_{i1}(n)) \cdot FFT^*(x_{i2}(n))}{|FFT(x_{i1}(n))| \cdot |FFT(x_{i2}(n))|} \right\} \quad (1)$$

where  $l$  is the time lag. GCC-PHAT delivers a measure of the similarity, or coherence measure, between signals realigned according to the time delay  $l$  and, in ideal conditions, presents a prominent peak in correspondence to the actual TDOA [22]. Although several alternative solutions to the TDOA estimation problem have been investigated over the years [20, 5, 13, 8], GCC-PHAT remains the technique of choice among the sound source localization community. The reasons of its success are mainly its rather simple and direct implementation and the proved robustness to highly adverse and noisy environments. It must be taken into account anyway that as soon as the level of reverberation increases and reflections get stronger, the performance of GCC-PHAT decreases [6].

When a DMN is available as in the addressed scenario, the contributions of each microphone pair are combined to derive a single estimation of the source position. The combination can be performed at the TDOAs level using one among several approaches: a maximum likelihood or least square framework, triangulation, spherical interpolation [21], spherical intersection [19], linear interpolation [3] and so on. Conversely, “direct approaches” derive the source position estimation performing a search, in a beamformer-like [4] fashion, over a grid  $\Sigma$  of potential source positions  $\mathbf{p}$  and maximizing an objective function based either on coherence or energy. A very efficient approach is the Global Coherence Field (GCF) introduced in [18]. In the given multi-microphone scenario, where sensors are distributed all around the room, GCF has proved to efficiently handle and manage the information provided by the sensor network providing satisfactory performance in the single speaker case [11, 17]. Let us consider a set of  $M$  microphone pairs and the corresponding set of theoretical time delays  $\tau_i(\mathbf{p})$  ( $i = 0, \dots, M - 1$ ) measured at microphone pair  $i$  if the source is in  $\mathbf{p}$ . GCF is a function defined over  $\Sigma$  and computed as follows:

$$GCF(t, \mathbf{p}) = \frac{1}{M} \sum_{i=0}^{M-1} C_i(t, \tau_i(\mathbf{p})) \quad (2)$$

GCF represents the plausibility that an active sound source is present in  $\mathbf{p}$  and its maximum is the source position estimation  $\hat{\mathbf{p}}$ :

$$\hat{\mathbf{p}} = \arg \max_{\mathbf{p} \in \Sigma} \text{GCF}(t, \mathbf{p}) \quad (3)$$

The SRP-PHAT [7] and the realizable delay vectors theory [12] are based on very similar concepts but commonly applied to compact microphone clusters instead of DMNs.

Given a traditional GCF framework, the main novelties of the proposed system, with respect to the solution adopted in the last evaluation, are a speech activity detection (SAD) based on HMM and an adaptive smoothing filter [9] that post-processes localization hypotheses in order to remove outliers and deliver regular trajectories. Figure 1 shows the overall system block diagram and highlights the role of the SAD and of the smoothing filter.



**Fig. 1.** Block diagram of the proposed tracking system.

As an alternative to the proposed solution, audio source tracking is commonly tackled using either Kalman Filtering [14] or Particle Filtering [10].

### 3 Tracking System Description

#### 3.1 GCF computation

As already mentioned, the core of the presented tracking system is the computation of GCF sound maps whose peaks identify the source position estimations. In order to find out a trade-off between accuracy and tracking capabilities the analysis window for both GCC-PHAT and GCF computation was set to  $2^{14}$  samples with an overlap factor 4. This is equivalent to a  $0.09 s$  time step, corresponding to about 10 sound map computations per second. Since GCC-PHAT delivers a coherence measure only at integer time delays, an oversampling of factor 2 was introduced to increase the overall accuracy. The GCF function was computed frame by frame independently, on a plane parallel to the floor at  $130 cm$  height. The spatial resolution of  $\Sigma$  was set to  $2 cm$  on both coordinates. The GCF map was computed exploiting each horizontal microphone pair available in the T-shaped arrays. 8 channels of the 64 channels NIST MarkIII linear array, corresponding to 4 further microphone pairs, were also used to process seminars recorded in ITC-irst and UPC CHIL rooms.

### 3.2 Speech Activity Detection

The SAD module is derived from an Acoustic Event Detector (AED) based on Hidden Markov Models (HMM) whose features are the traditional 38 coefficients adopted in speech recognition applications. The proposed AED [23] was adopted to participate to the corresponding CLEAR 2007 task and is able to detect, i.e. to recognize time boundaries, and classifies 13 different acoustic events including speech. It is worth mentioning that this implementation of the AED is site-dependent since a set of event models is trained for each smart room. According to localization experiments conducted on the development data set, the AED was forced to recognize presence of speech when there are not enough clues to derive a safe classification.

### 3.3 The Smoothing Filter

The final smoothing process was specifically designed to handle multiple slowly moving sources that, as observed in the training set, characterize the addressed scenario. With this purpose, the smoothing filter is very static but ready to move to a new area of the room as soon as there are cues that a new source is currently active. In order to filter out potential outliers due to non-human sound sources, the system considers the spatial distance  $d(n, n - 1)$  between the current GCF peak  $\mathbf{u}(n)$  and the last confirmed localization hypothesis  $\tilde{\mathbf{u}}(n - 1)$ . If  $d(n, n - 1)$  is higher than the threshold  $T_d$ , the current GCF peak is skipped. Otherwise, the new localization estimate  $\tilde{\mathbf{u}}(n)$  is computed as follows:

$$\tilde{\mathbf{u}}(n) = \alpha \mathbf{u}(n) + (1 - \alpha) \tilde{\mathbf{u}}(n - 1) \quad (4)$$

The parameter  $\alpha \in [0.12, 1]$  is adapted according to the following equation:

$$\alpha = \begin{cases} \alpha/1.5 & \text{if } d'(n, n - 1) < T_\alpha \\ 1.05\alpha & \text{otherwise} \end{cases} \quad (5)$$

where  $d'(n, n - 1)$  is the euclidean distance between  $\tilde{\mathbf{u}}(n)$  and  $\tilde{\mathbf{u}}(n - 1)$ , while the threshold  $T_\alpha$  is commonly set to  $0.4T_d$ . Notice that the filter does not take into account the actual temporal distance between  $\tilde{\mathbf{u}}(n - 1)$  and  $\mathbf{u}(n)$  in case several GCF peaks are skipped or long pauses occur. In order to enable the tracker to switch to different areas when a new speaker takes turn, the whole process is reset and moved when a given number ( $N_w$ ) of GCF peaks gather in the same area with a range equal to  $T_d$ .

Since each CHIL room presents peculiar acoustic characteristics and different DMN implementations, the tracker loads different parameter configurations depending on the room it is dealing with. Parameters are tuned on the basis of experiments conducted on the development set. Typical values are  $T_d = 30 - 50 \text{ cm}$  and  $N_w = 7 - 10$ .

### 3.4 Computational Load

As for the computational load of the tracker, the overall real time factor is 0.4, as measured on a Pentium 2.4 GHz machine. In particular, GCC-PHAT computation of one single microphone pair is responsible for 0.3, while the remaining localization process, including both GCF search and post-processing, contributes with 0.1. Notice that GCC-PHAT computations can be parallelized in order to save computational load. In case this is not possible, a measure of the computational load due to the GCC-PHAT computation is not feasible since it depends on the number of exploited pairs. Finally, the algorithm is single-pass and fully automatic.

## 4 Results and Conclusions

The evaluation was conducted on 40 segments, each one 5 minutes long, extracted from a set of meetings recorded in the CHIL rooms. The evaluation data set covers different phase of meetings. i.e. beginning, single speaker presentations, discussions, pauses and so on, resulting in a wide variety of acoustic activities.

Table 1 reports on the results obtained in the CLEAR 2007 evaluation campaign by the presented audio based tracking system. In order to have a direct

| MOTP   | A-MOTA | Miss Rate | False Pos | Err>50cm |
|--------|--------|-----------|-----------|----------|
| 20.8cm | 45.18% | 29.47%    | 25.35%    | 18.27%   |

**Table 1.** Performance of the audio tracker in terms of 2007 metrics.

understanding of the improvements obtained from the last evaluation, table 2 compares performance with the results obtained by the ITC-irst system in the CLEAR 2006 evaluation. In this case the evaluation criteria are the same as last year and silences are neglected. Notice the considerable gain in terms of A-MOTA and the consistent reduction of “Miss Rate”. On the other hand the precision, represented by the metrics MOTP, shows only a little and not significant improvement. Since the GCF computation is more or less the same and there is no significant gain in terms of precision, it is reasonable to assume that the improvement is mainly to ascribe to the new adaptive post-processing. Moreover the AED system seems to provide satisfactory performance even if it is responsible of at least a 7% drop in terms of A-MOTA. However this is only a rough estimation as the whole localization system is influenced by the behaviour of the SAD module.

It is anyway worth mentioning that, even if performance is satisfactory and encouraging for further works in this direction, there seem to be high margins for more improvements. Moreover, an analysis on single segments highlights that performance varies hugely depending on the conditions the systems is facing. In particular, those meetings where silences or non-speech sounds are predominant,

| System | MOTP   | A-MOTA | Miss Rate | False Pos |
|--------|--------|--------|-----------|-----------|
| 2006   | 21.8cm | 15.6%  | 65%       | 19%       |
| 2007   | 20.8cm | 52.26% | 29.47%    | 18.27%    |

**Table 2.** Comparison between the performance obtained in 2006 and 2007 CLEAR evaluations in the multi-person tracking task. Silences are neglected as foreseen in the 2006 metrics.

in contrast with the development set where speech was almost always present, are characterized by very bad performance that affects the overall scoring. From this point of view, further investigations aimed at increasing the overall reliability and robustness of the tracker must be conducted.

## 5 Multimodal Person Tracking

The FBK-irst multimodal person tracker combines independently computed outputs of the visual and acoustic tracking systems, which are described in detail in the previous single-modality sections. Acoustic events trigger the dumping of the most likely visual track segment that may have generated it. Once speech activity is detected and located by the audio based system, the spatially closest visual track is linked to it, and dumped as multimodal output. Such link is continuously updated until speech activity terminates. Once terminated, the link remains active until a new acoustic activity is detected. The multimodal system exploits the same data as the acoustic and visual trackers. Computational overhead for the above described fusion strategy is negligible with respect to the single modality computational loads.

### 5.1 Results and Conclusions

Table 3 reports on the results obtained by the multimodal system. Since the task

| MOTP   | MOTA   | Miss   | False Pos | Mismatches |
|--------|--------|--------|-----------|------------|
| 13.5mm | 34.30% | 33.81% | 29.71%    | 209(2.18%) |

**Table 3.** Performance obtained by the multimodal tracker.

considerably differs from the previous one, a direct comparison of the results is not feasible. It is anyway clear that this kind of approach suffers the weaknesses of both systems. A real multimodal system, that treats the audio contribution as an extra source of likelihood, as presented in [11] would be more suitable to the given scenario.

## References

1. <http://server.chil.de/>.
2. <http://www.clear-evaluation.org/>.
3. M. Brandstein, J. Adcock, and H. Silverman. A closed-form location estimator for use with room environment microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 5(1):45–50, January 1997.
4. M. Brandstein and D. Ward, editors. *Microphone Arrays*. Springer Verlag, 2001.
5. J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceeding of IEEE*, 57(8):1408–1418, August 1969.
6. B. Champagne, S. Bedard, and A. Stephenne. Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, 4(2):148–152, March 1996.
7. J. DiBiase. *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. PhD thesis, Brown University, May 2000.
8. S. Doclo and M. Moonen. Robust time-delay estimation in highly adverse acoustic environments. In *Proceeding of IEEE WASPAA*, pages 59–62, New Platz, NY, USA, October 21-24 2001.
9. A. Abad et al. Audio person tracking in a smart-room environment. In *Proceedings of Interspeech*, pages 2590–2593, Pittsburgh, PA, USA, September 17-21 2006.
10. F. Antonacci et al. Tracking multiple acoustic sources using particle filtering. In *Proceedings of the European Signal Processing Conference*, Florence, Italy, September 4-8 2006.
11. R. Brunelli et al. A generative approach to audio-visual person tracking. In Rainer Stiefelhagen and John Garofolo, editors, *Multimodal Technologies for Perception of Humans*, pages 55–68. Springer LNCS 4122, 2006. First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006.
12. S. Griebel and M. Brandstein. Microphone array source localization using realizable delay vectors. In *IEEE WASPAA*, pages 71–74, New Platz, NY, USA, October 21-24 2001.
13. Y. Huang, J. Benesty, and G. Elko. Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system. In *Proceedings of IEEE ICASSP*, volume 2, pages 937–940, Phoenix, AZ, USA, March 15-19 1999.
14. U. Klee, T. Gehrig, and J. McDonough. Kalman filters for time delay of arrival-based source localization. In *Proceedings of Interspeech*, pages 2289–2292, Lisbon, Portugal, September 4-8 2005.
15. C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 24(4):320–327, August 1976.
16. M. Omologo and P. Svaizer. Use of Crosspower-Spectrum Phase in acoustic event location. *IEEE Transactions on Speech and Audio Processing*, 5(3):288–292, May 1997.
17. M. Omologo, P. Svaizer, A. Brutti, and L. Cristoforetti. Speaker localization in CHIL lectures: Evaluation criteria and results. In Steve Renals and Springer Berlin/Heidelberg Samy Bengio, editors, *MLMI 2005: Revised and selected papers*, pages 476–487, Edinburgh, UK, July 11-13 2005.
18. M. Omologo, P. Svaizer, and R. DeMori. *Spoken Dialogue with Computers*. Academic Press, 1998. Chapter 2: *Acoustic Transduction*.

19. H. Schau and A. Robinson. Passive source localization employing intersecting spherical surfaces from time-of-arrival differences. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 35(12):1661–1669, December 1987.
20. R. Schmidt. *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*. PhD thesis, Stanford University, 1981.
21. J. Smith and J. Abel. Closed-form least-square source location estimation from range-difference measurements. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 35(12):1661–1669, December 1987.
22. P. Svaizer, M. Matassoni, and M. Omologo. Acoustic source location in a three-dimensional space using crosspower spectrum phase. In *Proceedings of IEEE ICASSP*, volume 1, pages 231–234, Munich, Germany, April 21-24 1997.
23. C. Zieger. An HMM based system for acoustic event detection. Second International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2007.