

# MULTISPEAKER LOCALIZATION AND TRACKING IN INTELLIGENT ENVIRONMENTS

C. Segura, A. Abad, C. Nadeu, and J. Hernando

Technical University of Catalonia, Barcelona, Spain,  
{csegura,alberto,climent,javier}@gps.tsc.upc.edu

**Abstract.** Automatic speaker localization is an important task in several applications such as acoustic scene analysis, hands-free videoconferencing or speech enhancement. Tracking speakers in multiparty conversations constitutes a fundamental task for automatic meeting analysis. In this work, we present the acoustic Person Tracking system developed at the UPC for the CLEAR'07 evaluation campaign. The designed system is able to track the estimated position of multiple speakers in a smart-room environment. Preliminary speaker locations are provided by the well known SRP-PHAT algorithm. Data association techniques based on trajectory prediction and spatial clustering are used to match the raw positional estimates with potential speakers. These positional measurements are then finally spatially smoothed by means of Kalman filtering. Experimental results obtained on the CLEAR'07 CHIL database are also reported.

## 1 Introduction

The automatic analysis of meetings recorded in multisensor rooms is an emerging research field. In this domain, localizing and tracking people and their speaking activity play fundamental roles in several applications, like scene analysis, hands-free videoconferencing or speech enhancement techniques.

The degree of reliable information provided by speaker localization systems on the basis of the audio signals collected in a smart-room environment with a distributed microphone network, depends on a number of factors such as environmental noise, room reverberation, talker movement and head orientation. These factors, among others, demand an effort on the development of new robust issues capable of dealing with independence on the environmental conditions.

In the present work, we get an insight on the development and design of a robust Acoustic Person Tracking system in the framework of the CHIL [9] research activities conducted at UPC.

---

DRAFT-only system description for the CLEAR'07 workshop. Paper will be updated for final inclusion in the workshop proceedings.

## 2 Audio Person Tracking System

### 3 Evaluation

Audio Person Tracking evaluation is run on an extract of the data collected by the CHIL consortium for the CLEAR 07 evaluation. These data are audiovisual recordings of seminars given at each partner site involving a smaller group of attendees listen to a presentation, ask questions, maybe take turns, etc. A complete description of the data and the evaluation can be found in [11].

#### 3.1 Summary of the Experimental Set-Up

**Data description** Room set-ups of the contributing sites present two basic common groups of devices: the *audio* and the *video* sensors.

Audio sensors set-up is composed by 1 (or more) NIST Mark III 64-channel microphone array, 3 (or more) T-shaped 4-channel microphone cluster and various table-top and close-talk microphones.

Video sensors set-up is basically composed by 4 (or more) fixed cameras. In addition to the fixed cameras, some sites are equipped with 1 (or more) PTZ camera.

**Evaluation metrics** Two metrics are considered for evaluation and comparison purposes:

*Multiple Object Tracking Precision (MOTP) [mm]* This is the precision of the tracker when it comes to determining the exact position of a tracked person in the room. It is the total Euclidian distance error for matched *ground truth-hypothesis* pairs over all frames, averaged by the total number of matches made. It shows the ability of the tracker to find correct positions, and is independent of its errors in keeping tracks over time, estimating the numbers of persons, etc.

*Multiple Object Tracking Accuracy (A-MOTA) [%]* This is the accuracy of the tracker when it comes to keeping correct correspondences over time, estimating the number of people, recovering tracks, etc. It is the sum of all errors made by the tracker, false positives, misses, over all frames, divided by the total number of ground truth points. This metric is like the *video* MOTA in which all mismatch errors are ignored and it is used to measure tracker performance only for the active speaker at each point in time for better comparison with the acoustic person tracking results (where identity mismatches are not evaluated).

#### 3.2 Audio Person Tracking Results

We have decided to use all the *T-clusters* available in the different seminars and only to use the *MarkIII* data for the sites (ITC, UKA and UPC)

In general, only microphone pairs of the same *T-cluster* or *MarkIII* array are considered by the algorithm.

In the experiments where the *MarkIII* is used, 16 microphone channels are selected for GCC-PHAT computation

The pairs selected out of the *MarkIII* are 42 in total, spanning an inter-microphone separation of 16cm, 24cm, and 32cm. The number of microphone pairs used in *MarkIII* is greater than those used of the *T-Clusters*, thus a corrective weight is given to the *MarkIII* contribution to the SRP-PHAT algorithm in order to have approximately the same importance as one *T-Cluster*

In Table 1 individual results for each data set and average results Acoustic Person Tasks are shown. Notice that the results are not directly the mean of the individual results, since the scores are recomputed jointly.

**Table 1.** Results for acoustic person tracking.

Task	MOTP	Misses	False Positives	A-MOTA
AIT data	201mm	48.15%	8.17%	43.68%
IBM data	206mm	35.01%	18.09%	46.91%
ITC data	157mm	38.31%	38.97%	22.72%
UKA data	175mm	41.55%	22.56%	35.89%
UPC data	117mm	30.35%	13.69%	55.96%
Total Average	168mm	37.86%	20.97%	41.17%

## 4 Conclusions

In this paper we have presented the audio Person Tracking system developed by UPC for the CLEAR evaluation campaign. Novelty proposed in the systems have been specially designed to add robustness to scenario and environment variabilities. Results show that the use of the *MarkIII* data yields to a better precision but more false positives may arise due to acoustic noise sources. Improvement of the Kalman filtering and association rules and the introduction of a SAD algorithm are also expected to enhance the tracking system.

## References

- [1] Omologo, M, Svaizer, P., “Use of the crosspower-spectrum phase in acoustic event location”, IEEE Trans. on Speech and Audio Processing, 1997.
- [2] Chen, J., Huang, Y.A., Benesty, J. “An adaptive blind SIMO identification approach to joint multichannel time delay estimation”, in Proceedings of IEEE ICASSP, Montreal, May 2004.
- [3] Potamitis, I., Tremoulis, G., Fakotakis, N., “Multi-speaker doa tracking using interactive multiple models and probabilistic data association”, in Proceedings of Eurospeech 2003, Geneva, Sep 2003.

- [4] Yu, Y., Silverman, H.F., “An improved TDOA-based location estimation algorithm for large aperture microphone arrays”, in Proceedings of IEEE ICASSP 2004, Montreal, May 2004.
- [5] Brandstein, M.S., Adcock, J.E., Silverman, H.F., “A closed-form location estimator for use with room environment microphone arrays”, IEEE Trans. on Speech and Audio Processing, 1997.
- [6] Sturim, D.E., Brandstein, M.S., Silverman, H.F., “Tracking multiple talkers using microphone-array measurements”, in Proceedings of IEEE ICASSP 1997, Munich, April 1997.
- [7] Abad, A., Macho, D., Segura, C., Hernando, J., Nadeu, C., “Effect of Head Orientation on the Speaker Localization Performance in Smart-room Environment”, in Proceedings of Interspeech 2005, Lisboa, Sep 2005.
- [8] DiBiase, J., Silverman, H., Brandstein, M., “Microphone Arrays. Robust Localization in Reverberant Rooms”, Chapter 8, Springer, Jan 2001.
- [9] CHIL Computers In the Human Interaction Loop. Integrated Project of the 6th European Framework Programme (506909). <http://chil.server.de/>, 2004- 2007.
- [10] Denda, Y., Nishiura, T., Yamashita, Y., “A study of weighted CSP analysis with average speech spectrum for noise robust talker localizatio”, in Proceedings of Interspeech 2005, Lisboa, Sep 2005.
- [11] The Spring 2007 CLEAR Evaluation and Workshop. <http://www.clear-evaluation.org/>.