Multispeaker Localization and Tracking in Intelligent Environments*

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

Technical University of Catalonia, Barcelona, Spain
{csegura,alberto,javier,climent}@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 SRP-PHAT algorithm, which is known to perform robustly in most scenarios. Data association techniques based on trajectory prediction and spatizal 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. Besides the technology description, experimental results obtained on the CLEAR'07 CHIL database are also reported.

1 Introduction

The automatic analysis of meetings 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.

Many approaches to the task of acoustic source localization in smart environments have been proposed in the literature. The main differences between them lie in the way they gather spatial clues from the acoustic signals, and how this information is processed to obtain a reliable 3D position in the room space. Spatial features, like the Time Difference of Arrival (TDOA) [8] between a pair of microphones or the Direction of Arrival (DOA) to a microphone array, can be obtained on the basis of cross-correlation techniques [1], High Resolution Spectral Estimation techniques [3] or by source-to-microphone impulse response estimation [2].

Conventional acoustic localization systems include a tracking algorithm that smoothes the raw positional measurements to increase precision. Furthermore,

* This work has been partially sponsored by the EC-funded project CHIL (IST-2002-506909) and by the Spanish Government-funded project ACESCA (TIN2005-08852).

R. Stiefelhagen et al. (Eds.): CLEAR 2007 and RT 2007, LNCS 4625, pp. 82–90, 2008.

the localization of multiple speakers simultaneously becomes severely complicated due to speech overlap of participants, since the localization techniques based on the cross-correlation like TDOA estimation assume one impinging wavefront. The task becomes specially difficult in the case of multiple moving speakers. Prior research on speaker tracking usually deals with a single speaker [10], however recently, multispeaker tracking [11] using Kalman [4] and particle [13] filtering techniques has gained interest in the context of smart meeting rooms.

The UPC acoustic localization system proposed in this work is based on the SRP-PHAT [5] localization method. The SRP-PHAT, algorithm although being very robust in reverberant environments, is not very well suited for the case of multiple concurrent speakers. The PHAT weighting introduces a masking effect of dominant acoustic sources over other sources of sound. This is desirable for increasing the robustness of the localization system by masking multipath acoustic propagation and reverberation, but it also hinders the localization of multiple acoustic sources. However, in the case of using a short analysis window ($\sim$23ms), we have observed that the positional estimates produced by the SRP-PHAT jump from one speaker to another at a very high rate due to the non-stationarity of the voice.

In our work we use a multiperson tracker based on the Kalman filter, which models a simple Newtonian motion of the source. The tracker carries out the tasks of detecting potential acoustic sources using spatial clustering and also assigning the raw location estimates to their corresponding speaker tracks using data association techniques. Then the measures assigned to each individual track are spatially smoothed by means of the corresponding Kalman filter [12], acording with the measure error variance estimation method defined in the next section.

2 Acoustic Source Localization

The SRP-PHAT algorithm [5] tackles the task of acoustic localization in a robust and efficient way. In general, the basic operation of localization techniques based on SRP is to search the room space for a maximum in the power of the received sound source signal using a delay-and-sum or a filter-and-sum beamformer. In the simplest case, the output of the delay-and-sum beamformer is the sum of the signals of each microphone with the adequate steering delays for the position that is explored. Concretely, the SRP-PHAT algorithms consists in exploring the 3D space, searching for the maximum of the contribution of the PHAT-weighted cross-correlations between all the microphone pairs. The SRP-PHAT algorithm performs very robustly due the the PHAT weighting, keeping the simplicity of the steered beamformer approach.

Consider a smart-room provided with a set of N microphones from we choose M microphone pairs. Let $\mathbf{x}$ denote a $\mathbf{R}^3$ position in space. Then the time delay

of arrival $TDOA_{i,j}$ of an hypothetic acoustic source located at $\mathbf{x}$ between two microphones i, j with position $\mathbf{m}_i$ and $\mathbf{m}_j$ is:

$$TDOA_{i,j} = \frac{\| \mathbf{x} - \mathbf{m}_i \| - \| \mathbf{x} - \mathbf{m}_j \|}{s}, \tag{1}$$

where s is the speed of sound.

The 3D room space is then quantized into a set of positions with typical separation of 5-10cm. The theoretical TDOA $\tau_{\mathbf{x},i,j}$ from each exploration position to each microphone pair are precalculated and stored.

PHAT-weighted cross-correlations [1] of each microphone pair are estimated for each analysis frame. It can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density ($G_{m_1 m_2}(f)$) as follows,

$$R_{m_i m_j}(\tau) = \int_{-\infty}^{\infty} \frac{G_{m_i m_j}(f)}{|G_{m_i m_j}(f)|} e^{j2\pi f \tau} df, \tag{2}$$

The estimated acoustic source location is the position of the quantized space that maximizes the contribution of the cross-correlation of all microphone pairs:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \sum_{i,j \in \mathbb{S}} R_{m_i m_j}(\tau_{\mathbf{x},i,j}), \tag{3}$$

where $\mathbb{S}$ is the set of microphone pairs. The sum of the contributions of each microphone pair cross-correlation is assumed to be well-correlated with the likelihood of the estimation given. Hence, this value is compared to a fixed threshold (depending on the number of microphone pairs used) to reject/accept the estimation. The threshold has been experimentally fixed to 0.5 for each 6 microphone pairs. It is important to note that in the case of concurrent speakers or acoustic events, this technique will only provide an estimation for the dominant acoustic source at each iteration.

3 Multiple Speaker Tracking

One of the major problems faced by acoustic tracking systems is the lack of a continuous stream of features provided by the localization module. Moreover, in the case of spontaneous speech, we have to deal with acoustic events that are sporadic and others that are concurrent.

The proposed method makes use of spatial segmentation to detect tracks and associate incoming raw estimates to them. Each tracked acoustic source has an associated acceptance region and a Kalman filter. When a raw estimate falls within a region of a track, it is asigned to that track and then used by the Kalman filter. If no measurement falls within this acceptance region, then the predicted position is used as the measurement for the Kalman filter.

We have no constraint on the number of acoustic sources that the algorithm is able to track. The method dynamically estimates the number of sources based

on a birth/death system. The track detection uses a spatial segmentation algorithm to group locations that are close to each other in space and time. When a minimum number of locations N_b are found in a space region over a defined time window T_b, the tracking system decides it is a new track. Similarly, if a track does not have any measurements that fall within its acceptance region for a given amount of time T_d, then the track is dropped. The ratio between T_b and N_b used in the track detection module is a design parameter. It must be high enough to filter out noises and outliers, but also not too high in order to be able to detect sporadic acoustic events. In our experiments N_b is set to 4, T_b is $460ms$ and T_d is also $460ms$.

3.1 Kalman Filter

The Kalman filter [12] has been widely used in tracking applications.

The motion of a specified talker is modelled by a simple Newtonian model defined by the state difference equation:

$$\mathbf{s}_{k+1} = \phi_k \mathbf{s}_k + \mathbf{\Gamma}_k \mathbf{w}_k, \tag{4}$$

where $\mathbf{s}_k$ is the system state, ϕ_k is the transition matrix that propagates the state, $\mathbf{w}_k$ is the process noise vector and $\mathbf{\Gamma}_k$ is the gain matrix.

In this work we have chosen the state, as a 6-component vector consisting in the 2-dimensional source position, velocity and acceleration:

$$\mathbf{s}_k = [x_k \; y_k \; \dot{x}_k \; \dot{y}_k \; \ddot{x}_k \; \ddot{y}_k]^T. \tag{5}$$

The process noise vector, $\mathbf{w}_k = [w_{x,k} \; w_{y,k}]^T$, whose components are uncorrelated, zero-mean Gaussian variables with equal variance σ_w, is used to model variations in the acceleration of the source motion. The transition matrix ϕ and the gain matrix $\mathbf{\Gamma}$ are defined by:

$$\phi = \left(\begin{array}{c|c|c} \mathbf{I}_2 & \Delta t \mathbf{I}_2 & \frac{\Delta t^2}{2}\mathbf{I}_2 \\ \hline \mathbf{0}_2 & \mathbf{I}_2 & \Delta t \mathbf{I}_2 \\ \hline \mathbf{0}_2 & \mathbf{0}_2 & \mathbf{I}_2 \end{array}\right), \tag{6}$$

$$\mathbf{\Gamma} = \left(\frac{\Delta t^3}{6}\mathbf{I}_2 \middle| \frac{\Delta t^2}{2}\mathbf{I}_2 \middle| \Delta t \mathbf{I}_2\right)^T, \tag{7}$$

where Δt is the time period between positional measures provided by the localization system, $\mathbf{I}_2$ is the identity matrix and $\mathbf{0}_2$ is a zero matrix.

In the other hand, the source positional observation at the k^{th} iteration, $\mathbf{z}_k$ is modelled in the conventional as the true 2D source position corrupted by the measurement noise $\mathbf{v}_k$.

$$\mathbf{z}_k = \mathbf{H}\mathbf{s}_k + \mathbf{v}_k. \tag{8}$$

In this work, the measurement matrix is given by:

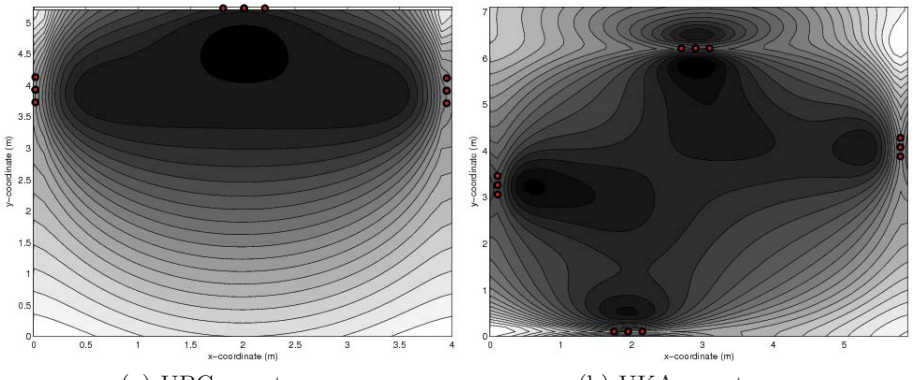
$$\mathbf{H} = [\mathbf{I}_2 | \mathbf{0}_2 | \mathbf{0}_2]. \tag{9}$$

The covariance matrix of the measurement noise $\mathbf{R}_k = E\left[\mathbf{v}_k \mathbf{v}_k^T\right]$ is calculated as a function of the estimated source location, sensor position and environmental conditions as proposed in [14], where the error covariance of the localization estimation is computed as a function of the variances of the TDOAs estimation and a pure geometrical weight matrix:

$$\mathbf{R}_k = \left(\mathbf{M}^T \cdot \mathbf{V} \cdot \mathbf{M}\right)^{-1}, \tag{10}$$

$$\mathbf{V} = \begin{pmatrix} \frac{1}{\sigma^2_{\tau_1}} & & & \\ & \frac{1}{\sigma^2_{\tau_2}} & & \\ & & \vdots & \\ & & & \frac{1}{\sigma^2_{\tau_N}} \end{pmatrix}, \tag{11}$$

where the weight matrix $\mathbf{M}$ [14] models the sensitivity of the microphone array at the estimated position of the speaker and $\mathbf{V}$ is the diagonal matrix consisting of the inverse of the TDOA variances $\sigma^2_{\tau_i}$ at the microphone pair i. The figure 1 shows a simulation of the error variance for the rooms at UKA and UPC.



(a) UPC smartroom (b) UKA smartroom

Fig. 1. Simulation of the localization error variance at heigh= $1.7m$ for UPC and UKA CHIL-Rooms. The brightness in the figure is related to the predicted error at a given position. Brighter zones are more prone to localization errors.

The SRP-PHAT algorithm does not provide an estimation of the variance of the time differences of arrival, because the TDOAs $\hat{\tau}_{\hat{\mathbf{x}},i,j}$ are estimated indirectly calculating the distance differences from the detected location of the acoustic source to each microphone. The only measure available is the value of the cross-correlation $\rho = R_{m_i m_j}(\hat{\tau}_{\hat{\mathbf{x}},i,j})$ at each microphone pair. In principle, lower values of the cross-correlation function should correspond with high variance TDOA estimation. Preliminary experimental results have led us to propose an exponential function to model the relationship between ρ and σ^2_τ :

$$\sigma^2_\tau = e^{-\frac{\rho}{\delta}} \cdot \beta. \tag{12}$$

The parameter δ must be set according with the microphone array configuration, since microphones that are closer exhibit a higher cross-correlation. In our work we have chosen $\delta = 0.05$ and $\beta = 5 \cdot 10^{-4}$.

3.2 Data Association

In situations dealing with multiple, possibly moving, concurrent speakers, the purpose of the data association method is to assign raw location measures to a specific acoustic source and also to filter out outliers that appear due to noise and reverberation. This is done through the use of acceptance regions [15]. The acceptance region is a segment of the space around the position predicted by the corresponding track. The region size is set dynamically according to the measure noise variance and state estimation uncertainty:

$$\left(\mathbf{z} - \mathbf{z}_k^-\right)^T \cdot \mathbf{S}_k^- 1 \cdot \left(\mathbf{z} - \mathbf{z}_k^-\right) \leq \gamma. \tag{13}$$

The variable $\mathbf{z}$ defines the acceptance region in space, γ is a fixed bound value , $\mathbf{z}_k^-$ is the source position predicted by the Kalman filter and $\mathbf{S}_k$ is the covariance matrix of the positional observations, that can be formulated recursively as follows:

$$\mathbf{S}_k = \mathbf{H} \cdot \mathbf{P}_k^- \cdot \mathbf{H}^T + \mathbf{R}_k, \tag{14}$$

where $\mathbf{P}_k^-$ is a matrix provided by the Kalman filter, that predicts the error covariance of the estimated state. A high value of the measure noise covariance matrix $\mathbf{R}_k$ or a high uncertainty in the estimation of the state, for instance due motion of the source, yields to a bigger acceptance region.

4 Evaluation

Audio Person Tracking evaluation is run on an extract of the data collected by the CHIL consortium for the CLEAR 07 evaluation. The data consists of meetings recorded at each partner site involving presentations and discussions. A complete description of the data and the evaluation can be found in [7].

4.1 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.

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 one minus 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).

4.2 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 eather the same *T-cluster* or within the *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 microphones 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 for the Acoustic Person Tracking tasks are shown. Notice that the average results are not directly the mean of the individual results, since the scores are recomputed jointly.

Table 1. Results for acoustic person tracking

Site	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%

5 Conclusions

In this paper we have presented the audio Person Tracking system developed by UPC for the CLEAR evaluation campaign. A method for estimating the localization error covariance matrix of the SRP-PHAT algorithm has been presented, that can be used in conjuction with a Kalman tracking filter to add robustness to scenario and environment variables. Results show that the use of the *MarkIII* data yields a better precision but more false positives, which may be attributable to non-speech acoustic sources. Improvement of the Kalman filtering and association rules and the introduction of a SAD algorithm, are 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: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (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 (September 2003)
[4] Sturim, D.E., Brandstein, M.S., Silverman, H.F.: Tracking multiple talkers using microphone-array measurements. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Munich (April 1997)
[5] DiBiase, J., Silverman, H., Brandstein, M.: Microphone Arrays, ch. 8. In: Robust Localization in Reverberant Rooms, Springer, Heidelberg (2001)
[6] CHIL Computers In the Human Interaction Loop. Integrated Project of the 6th European Framework Programme (506909) (2004-2007), `http://chil.server.de/`
[7] The Spring 2007 CLEAR Evaluation and Workshop, `http://www.clear-evaluation.org/`
[8] Brandstein, M.S.: A Framework for Speech Source Localization Using Sensor Arrays. Ph.D. Thesis, Brown University (1995)
[9] Bernardin, K., Gehring, T., Stiefelhagen, R.: Multi- and Single View Multiperson Tracking for Smart Room Environments. In: Stiefelhagen, R., Garofolo, J.S. (eds.) CLEAR 2006. LNCS, vol. 4122, Springer, Heidelberg (2007)
[10] Vermaak, J., Blake, A.: Nonlinear filtering for speaker tracking in noisy and reverberant environments. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP) (2001)
[11] Claudio, E., Parisi, R.: Multi-source localization strategies. In: Brandstein, M.S., Ward, D.B. (eds.) Microphone Arrays: Signal Processing Techniques and Applications, ch. 9, pp. 181–201. Springer, Heidelberg (2001)
[12] Welch, G., Bishop, G.: An introduction to the Kalman filter. TR 95-041, Dept. of Computer Sc., Uni. of NC at Chapel Hill (2004)

[13] Checka, N., Wilson, K., Siracusa, M., Darrell, T.: Multiple person and speaker activity tracking with a particle filter. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Montreal (May 2004)
[14] Brandstein, M.S., Adcock, J.E., Silverman, H.F.: Microphone array localization error estimation with application to optimal sensor placement. J. Acoust. Soc. Am. 99(6), 3807–3816 (1996)
[15] Bar-Shalom, Y., Fortman, T.E.: Tracking and Data association. Academic Press, London (1988)