Audio Person Tracking in a Smart-Room Environment

Alberto Abad, Carlos Segura, Dušan Macho, Javier Hernando and Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
{alberto,csegura,dusan,javier,climent}@talp.upc.edu

Abstract

Reliable measures of speaker positions are needed for computational perception of human activities taking place in a smart-room environment. In this work, it is described the development process and the experiments conducted in the design and implementation of an Audio Person Tracking system for smart-room environments. The proposed system is based on the SRP-PHAT algorithm, as it is known to perform robustly in most environmental conditions. Novelties proposed are aimed to enhance the accuracy of the system independently on the application scenario and to reduce the computational complexity.
Index Terms: sound localization, source tracking, microphone arrays.

1. Introduction

Speaker localization is a basic functionality for computational perception of human activities in a smart-room environment. Additionally, a reliable measure of the talker position is needed for technologies that are often deployed in that environment and use different modalities, like microphone array beamforming or steering of pan-tilt-zoom cameras towards the active speaker.

Conventional acoustic person localization and tracking systems can be generally split into three basic stages. In the first stage, estimations of such information as Time Difference of Arrival [1, 2] or Direction of Arrival [3] is usually obtained from the combination of the different microphones available. In general, in the second stage the set of relative delays or directions of arrival estimations are used to derive the source position that is in the best accordance with them and with the given geometry [4, 5]. In the third optional stage, a tracking of the possible movements of the sources according to a motion model can be employed [6].

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 our previous work [7], we have compared two representative systems to investigate their performances in smart-room environments and, particularly, to study the effect of talkers head orientation. The conclusion of that work is that techniques that join the estimated cross-correlations in a collaborative way, such as SRP-PHAT [8], are in general a convenient choice for developing a robust system almost independently on noise, reverberation or talker head orientation conditions, specially if the microphone network is distributed appropriately in the room.

The objective of the present work, developed in the framework of the acoustic processing research that is being carried out in the EU-funded CHIL project [9], is to show the development process and the novelties introduced in the design of a robust Audio Person Tracking system. Based on the SRP-PHAT algorithm, we first propose an adaptive smoothing factor for cross-power spectrum estimation derived from the estimated velocities of sources. Furthermore, motivated by the need for reducing the computational complexity and by experimental observations of the frequency properties of the cross-correlation, a two-pass search procedure that enhances the accuracy results and reduces the computational cost is proposed. Additionally, other development aspects are discussed.

To assure the usefulness of the proposed techniques, a representative and comparable extract of the data collected by the CHIL consortium, including interactive and non-interactive seminars, is used for evaluation purposes. Results show the convenience of the proposed Audio Person Tracking system and, particularly, of the two novelties introduced in this work.

2. Audio Person Tracking System

Many different approaches exist to tackle acoustic source localization in a smart room environment. Most remarkable differences between them are related with the way in which two basic problems are faced: a) how to infer information from the position of the sources on the basis of the microphone captures and b) how to use these information to obtain a reliable 3D position in the room space.

On the one hand, Time Difference of Arrival (TDOA) 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 source-to-microphone impulse response estimation based techniques [2], among others. On the other hand, the availability of multiple TDOA/DOA estimations lead to a minimization of an over-determined and non linear error function to obtain a unique estimation of the position that can be faced in many different ways, for instance, by means of iterative search algorithms [4], closed-form estimators that approximate to a sub-optimal solution [5] or space exploration based techniques, also known as Steered Response Power (SRP) techniques.

The SRP-PHAT [8] algorithm tackles and integrates these two basic problems of localization in a robust and smart way. In general, the goal of localization techniques based on SRP is to maximize 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. Thus, a simple localization strategy is to search for the energy peak through all the possible positions in 3D space. Concretely, SRP-PHAT algorithm searches for the maximum of the contribution of the cross-correlations between all the microphone pairs across the space. The main strength of this technique consists on the combination of the simplicity of the steered beamformer approach with the robustness offered by the generalized cross-correlation with PHAT weighting (GCC-PHAT) also known in the literature as crosspower-spectrum phase [1].

The proposed system in this work for Audio Person Tracking is based on the SRP-PHAT algorithm with some additional robust modifications that are described next. The system design is fundamentally aimed to develop a robust system with independency on the acoustic and room conditions, such as the number of sources, their maneuvering modes or the number of microphones. Additionally, computational cost is reduced.

2.1. The SRP-PHAT baseline algorithm

As already mentioned above, the SRP-PHAT algorithm searches for the maximum of the contribution of the cross-correlations between all the microphone pairs across the space. The process can be summarized into four basic steps:

- *Step 1* The exploration space is firstly split into small regions (typically of 5-10 cm). Then, theoretical delays from each possible exploration region to each microphone pair is pre-computed and stored.
- *Step 2* Cross-correlations of each microphone pair are estimated for each analysis frame. Concretely, the Generalized Cross Correlation with PHAT weighting is considered. It can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density ($\widehat{G}_{x_1x_2}(f)$) as follows,

$$\widehat{R}_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \frac{\widehat{G}_{x_1x_2}(f)}{|\widehat{G}_{x_1x_2}(f)|} e^{j2\pi f\tau} df \qquad (1)$$

- *Step 3* The contribution of the cross-correlations is accumulated for each exploration region using the delays precomputed in *Step 1*. In this way, it is obtained a kind of *Sound Map* as the one shown in Figure 1.
- *Step 4* Finally, the position with the maximum score is selected as the estimated position.

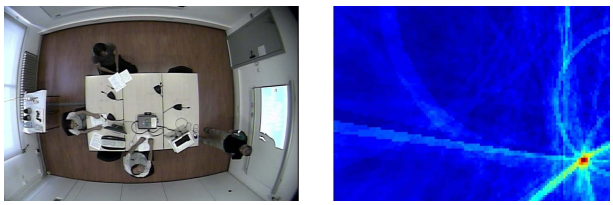

Figure 1: On the left, zenithal camera snapshot. On the right, example of the *Sound Map* obtained with the SRP-PHAT process.

2.2. The proposed Audio Person Tracking system

On the basis of the conventional SRP-PHAT a robust system for Audio Person Tracking is developed. Main novelties and some aspects related to other implementation details are introduced next.

2.2.1. *Adaptive Smoothing Factor for the Cross-Power Spectrum (CPS) Estimations*

Smoothing over time of the GCC-PHAT estimations is a simple and efficient way of adding robustness to the source localization system. This smoothing process can be done in the time domain (GCC-PHAT) or in the frequency domain (CPS). Considering the smoothed cross-power spectrum $\widehat{G}_{x_1x_2}(k,f)$ in time instant k and the instantaneous estimation $G_{x_1x_2}(k,f)$ our system performs the smoothing in the frequency domain as follows,

$$\widehat{G}_{x_1x_2}(k,f) = \beta\widehat{G}_{x_1x_2}(k-1,f) + (1-\beta)G_{x_1x_2}(k,f) \qquad (2)$$

From experimental observation it can be seen that the right selection of this β factor is crucial in the system design. A high smoothing value can greatly enhance the results obtained in an almost static scenario, while it can be dramatically inconvenient in a scenario with various moving speakers.

Hence, an adaptive smoothing factor has been designed. This adaptive factor is obtained based on the velocity estimation provided by a Kalman filter as it is depicted in Figure 2.

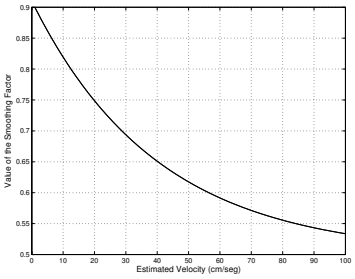

Figure 2: Value of the Adaptive Smoothing Factor depending on the estimated velocity.

2.2.2. *Frequency Masking discussion*

In [10], weighting each frequency bin of the GCC-PHAT according to average speech spectrum is proposed. For this purpose, a fixed template mask is estimated in several mel-scaled critical bands from a development database. It is obvious that as a consequence of this processing lower frequencies will be more emphasized than higher ones.

However, attending to radiation and room acoustics considerations the importance of each frequency band can be interpreted in an alternative way. It is known that talker's radiation pattern is more directive for high frequencies. Furthermore, most of the construction materials of typical rooms show higher absorption of sound for higher frequencies. Hence, if we can consider a source localization system relatively independent on orientation (like SRP-PHAT), enhancing high frequencies (that are assumed

to be less reverberated) will improve the GCC estimation and the performance of the overall system.

Obviously, this last idea is in strong contradiction to the previous one and suggests a dilemma. It becomes necessary to investigate the role of each frequency bin in the estimation of the cross-correlations. From experimental observations we were not able to draw a strong conclusion about this point, however, we could generally observe that most of the information for a rough localization seems to be concentrated in the low-frequency bins of the GCC-PHAT, while high frequency bins seem to be useful in order to obtain a finest estimation given a first coarse estimation.

2.2.3. Two-pass SRP Search

A typical problem of the SRP based techniques is the excess of computational load due to the exploration process, specially if a reduced space resolution is desired.

In this way, a two step search procedure is proposed with a double intentionality. Firstly, it is intended to reduce the computational complexity of the exploration. Secondly, and most importantly, it is aimed to enhance the accuracy of the system incorporating ideas derived from the discussion about frequency masking above. The proposed two-pass SRP search consists on the following two steps:

Coarse Search This search procedure is performed only in the *x*-*y* axis (*z* is assumed to be 1.5 m), with a searching cell dimension of 20 cm and only using the low frequency information of the cross-correlations ($f < 9kHz$). A first coarse estimation is obtained from this search, say $(x_1, y_1, 150)$ cm.

Fine Search A new limited search area around the obtained *coarse* estimation is defined $(x_1 - 25 : x_1 + 25, y_1 - 25 : y_1 + 25, 110 : 190)$ cm. In this new fine search, dimension of the cell search is fixed to 4 cm for the *x*-*y* axis and to 8 cm for the *z*-axis. In the *fine search* all the frequency information of the cross-correlations is used and a more accurate estimation is obtained.

2.2.4. Other considerations of the proposed system

The SRP-PHAT algorithm selects the position with the maximum value obtained from the accumulated contributions of all the correlations (*Step 4*). This value 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.

Finally, it is worth noting that although a Kalman filter is used for the estimation of the adaptive CPS smoothing factor, it is not considered for tracking purposes. The reason is that the Kalman filter design and the data association strategies adopted showed a different impact depending on the scenario. In other words, it showed to be too much dependent on the number and the velocities of sources to perform correctly.

3. Evaluation

Person Tracking evaluation is run on an extract of the data collected by the CHIL consortium for the CLEAR 06 evaluation [11]. These data are audiovisual recordings of seminars given at each partner site.

Two types of seminars were recorded: Non-interactive seminars, where mostly one presenter is speaking in front of a larger audience, and highly interactive seminars, where a smaller group of attendees listen to a presentation, ask questions, maybe take turns, etc.

3.1. Experimental Set-Up

3.1.1. Data description

The intention of the experiments is to evaluate the proposed system in two different environments. For this purpose, only data collected by IBM (interactive) and University of Karlsruhe, UKA (non-interactive) is considered. Each room is equipped with 4 T-shaped 4-channel microphone clusters appropriately distributed. In general, only microphone pairs of the same *T-cluster* array are used by the algorithms.

3.1.2. Evaluation metrics

Metrics and scoring of the systems has been done following the common agreement of the CHIL consortium for audio person tracking evaluation. The evaluation is run comparing the systems to 3D references at a rate of 1 label per second. Only time periods with one active speaker are considered. Two basic metrics are defined:

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 (i.e. Euclidean distance less than 500 mm) over all frames, averaged by the total number of matches made.

Acoustic Multiple Object Tracking Accuracy (A-MOTA) [%] This is the accuracy of the tracker evaluated only for the active speaker at each time instant. It is calculated as one minus the ratio of the sum of errors over all frames and the total number of frames; the errors can be misses (i.e. Euclidian distance higher than 500mm) or false positives.

3.1.3. Evaluated systems

Four systems are evaluated to show independently the usefulness of the new proposals: a baseline SRP-PHAT system (*Baseline*), the same system with adaptive smoothing factor for the CPS estimations (*Adaptive*), a SRP-PHAT system with the two-pass proposed search algorithm (*Two-pass*) and the complete proposed system including adaptive smoothing factor and two-pass algorithm (*Proposed*).

Due to excess of computational load, the double search processing is applied also in the case of the *Baseline* and *Adaptive* systems, however the *coarse* step search is done on the basis of the cross-correlation without frequency masking, that is, all the frequency band is considered.

The non-daptive β factor of the *Baseline* and the *Two-pass* systems is fixed to 0.7.

3.2. Experimental results

Table 1 shows comparative results of the four studied systems in interactive environments. Comparing *Baseline* with *Two-pass* and *Adaptive* with *Proposed* results, a clear improvement thanks to the two-pass search algorithm in both precision and accuracy scores

can be confirmed. The adaptive smoothing factor technique (*Baseline* vs. *Adaptive* and *Two-pass* vs. *Proposed*) does not show such an important influence in the results, with a slight improvement in MOTP and a slight decrease in A-MOTA score.

Table 1: Audio person tracking results of interactive (IBM) seminars.

System	MOTP	Misses	False Positives	A-MOTA
Baseline	189,5mm	31,08%	21,90%	47,05%
Adaptive	185,8mm	31,48%	22,78%	45,78%
Two-pass	184,5mm	19,50%	11,10%	69,40%
Proposed	180,3mm	19,08%	12,13%	68,80%

Person tracking results in non-interactive environments are presented in Table 2. Similar improvements are obtained with the two-pass algorithm to those observed in the interactive environment, while in this case a generalized improvement is observed also with the adaptive smoothing technique.

Table 2: Audio person tracking results of non-interactive (UKA) seminars.

System	MOTP	Misses	False Positives	A-MOTA
Baseline	161,2mm	28,25%	20,95%	50,81%
Adaptive	155,0mm	26,79%	19,29%	53,93%
Two-pass	147,0mm	19,32%	12,64%	68,04%
Proposed	142,4mm	17,98%	11,04%	70,97%

Attending to results in both scenarios, the proposed adaptive smoothing factor technique compared to the application of a fixed value shows equivalent results in the interactive scenario, while a slight improvement in the non-interactive scenario is observed. Indeed, this difference justifies the need for the adaptive technique. For instance, we could likely find (experimentally) a fixed smoothing value to improve the interactive results in exchange of probably degrading the performance in other scenarios. However, with the adaptive factor, we can develop a system working reasonably well independently on the number of speakers and their dynamics. In other words, the objective is not to improve the best possible results, the objective is to obtain good results in most environments. Anyway, the adaptive smoothing factor for the CPS estimations in the SRP-PHAT algorithm must be further investigated, for example, the relation between the velocity and the assigned value (see Figure 2) or the possible impact of less accurate Kalman estimations due to data association rules applied to assign estimated positions to tracks in multiple-speaker environments.

The proposed two-pass SRP algorithm based on frequency masking shows a great performance compared to the use of the complete frequency band of the cross-correlations. It is clearly convenient for tracking applications independently on the environment. Furthermore, it is a practical solution for reducing the computational cost of the SRP search procedure.

4. Conclusions

In this paper we have presented a robust Audio Person Tracking system for smart-room environments based on the well-known SRP-PHAT algorithm. Two simple and efficient novelties have been described and evaluated in interactive and non-interactive seminars collected by the CHIL consortium. Firstly, using an adaptive smoothing factor for the cross-power spectral estimations has shown to be convenient independently on the dynamics of the speakers. Secondly, the Two-Pass Search algorithm based on frequency masking, significantly improves the precision and accuracy of the tracker, besides reducing the computational cost of the search procedure.

5. Acknowledgements

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). Alberto Abad is currently supported by a Catalan Government grant.

6. 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 2006 CLEAR Evaluation and Workshop. http://www.clear-evaluation.org/.