Effect of Head Orientation on the Speaker Localization Performance in Smart-room Environment

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

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
{alberto,dusan,csegura,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, we investigate the effect of talkers head orientation on the accuracy of acoustical source localization techniques and its relation with the talker directivity pattern and room reverberation. Two different representative speaker localization techniques are assessed, steered response power and a crossing lines based method, in both cases on the basis of the estimated delays between pairs of microphones with the GCC-PHAT algorithm. A small database has been collected at the UPC's smart room for evaluation. The results show how the localization error heavily depends on the head orientation, and also the fact that the space exploration based technique is much more robust to head orientation changes than the crossing lines technique, due to the way the contributions from the various microphones are combined.

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. To locate a speaker from unobtrusive sensors, either video or audio sources can be used, though eventually the most accurate and robust techniques will likely be based on multimodal information [1].

Actually, 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. In this work, developed in the framework of the acoustic processing research that is being carried out in the EU-funded CHIL project [2], we try to get an insight into the effect of talkers head orientation on the accuracy of typical acoustic source localization techniques.

Figure 1 illustrates a typical situation where the various microphones collect not only the sound wave directly coming from the sound source (labeled by 0 in Figure 1) but also indirect sound waves – the ones originated at the same source and reflected from the walls or other surfaces in the room (waves 1-3 in Figure 1), in addition to other possible interferent signals and diffuse background noise. The accuracy of the localization measure depends on the ratios between the energy of the direct sound wave and the energies of the indirect waves. Those ratios are affected by a) the position and orientation of talkers head with respect to the sensors employed for localization and b) the reverberation properties of the room.

On the one hand, the measurements reported in [3] show that human talkers do not radiate voice sound uniformly in all directions; more energy is radiated in talkers forward direction than towards the side or the rear direction. On the other hand, the indirect sound waves can reach the localization sensor with the same or higher energy than the direct sound waves, representing thus a serious competition to the direct wave within the localization algorithm.

In this work, a small database was collected in the UPCs smart room, using 3 T-shaped arrays of 4 microphones each one, in several speaker positions and head orientations. Two different speaker localization techniques were tested and assessed with the labeled database: a steered response power (SRP) [4] technique based on spatial exploration and a crossing lines (CL) based technique [5]. For both techniques, signal delays were estimated with a generalized cross-correlation method with a phase transform weighting function [6]. With that experimental set-up, we have investigated the effect of head orientation on the performance of the selected acoustic source localization techniques in our smart room environment, in order to posteriorly search for ways of alleviating it. A heavy dependence of the localization error on the head orientation is observed. And the SRP-PHAT technique appears as much more robust to head orientation changes than the CL-PHAT technique, due to the way the contributions from the various T-shaped arrays are combined.

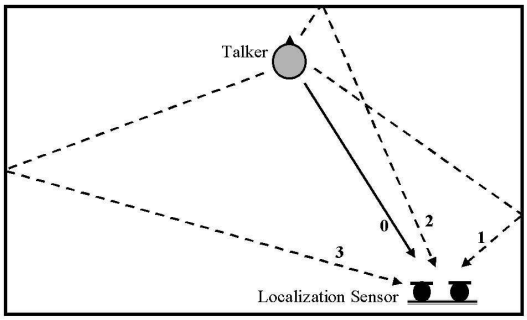

Figure 1: *Multi-path example where direct sound wave is 0 and reflective waves are 1, 2 and 3. In this situation, indirect sound waves could reach sensors with higher energy.*

2. Talker directivity and reverberation: The effect of orientation

As mentioned above, human talkers do not radiate speech uniformly in all directions. Figure 2 shows the A-weighted radiation pattern of human talker in horizontal plane passing through the talkers mouth [3]. This radiation pattern shows about -2dB attenuation on the side of the talker (90 ° or 270 °) and the attenuation behind the talker (180 °) is stronger than -6dB. Similarly, the vertical radiation patter of human talker is not uniform; e.g. there is about -3dB attenuation above the talkers head. In addition, the radiation pattern is frequency dependent [3]; behind the talker, the low speech frequencies are attenuated less than the high speech frequencies.

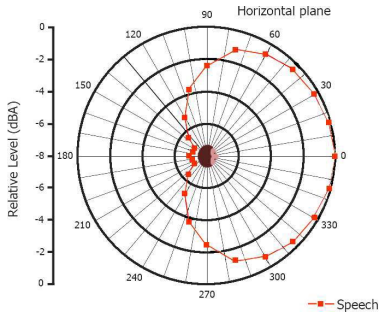

Figure 2: *Talker diagram in the horizontal plane (after [3]).*

Room reverberation produces a set of indirect sound waves reflected from the walls and other surfaces present in the room. The extent of reverberation depends on the dimension of the room and the materials used for the surfaces. The sound absorption capacity of each material is given by its absorption coefficient which indicates what portion of the incoming sound energy is absorbed by the given surface (e.g. the absorption coefficient equals to 0.02 for ceramic tiles and 0.63 for curtains at 1 kHz).

Two kinds of reflections can be considered: early reflections and late reflections. The early reflections cause a distortion called coloration, however it is mostly the late reflections that smear the speech spectra and reduce the intelligibility and quality of speech signals. In speaker localization, however, we are not concerned about the speech intelligibility; actually, the early reflections are usually of high energy and they can arrive to the localization sensor from positions that are very different from the talker position. Figure 1 illustrates a few early reflections. In that talker-localization sensor configuration (talker is facing the closest wall), the energy of the direct wave is attenuated due to the talkers radiation pattern. On the other hand, the wave number 2 before it is reflected is stronger than the direct wave number 0 and it can reach the localization sensor with a higher energy than the direct wave, depending on the absorption coefficient of the wall.

The example mentioned above illustrates that given the room properties and the positioning of the localization sensor(s), the reflected speech sound waves can represent a serious competition to the direct speech sound wave within a talker localization algorithm and they can degrade its performance. It is obvious that the degradation depends on the talkers orientation with respect to the localization sensor.

3. Talker localization

Acoustic source localization and tracking can be split into three basic stages. In the first stage, estimations of such information as Time Difference of Arrival or Direction of Arrival is usually obtained from the combination of the different microphones. 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. In the third optional stage, a tracking of the possible movements of the sources according to a motion model can be employed, but it in this work it has been discarded since is preferable for studing the effect of talker orientation.

3.1. TDOA estimation

There are basically two observable characteristics of the signals arriving to microphones and related to the source position that can be estimated: the Time Difference of Arrival (TDOA) between a pair of microphones and the Direction of Arrival (DOA) to a microphone array.

In this work we will focus on Time Delay Estimation techniques. Conretely, the most common and popular approaches in speech applications are based on the computation of Cross-Correlation. In order to increase robustness in noisy and reverberant conditions, the Cross-Correlation function is usually weighted attending to different optimality criterions, in what is named Generalized Cross-Correlation (GCC) [6]. GCC can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density ($\widehat{G}_{x_1x_2}(f)$) and a weighting function ($\psi(f)$) as follows,

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

A commonly used weighting function is the Phase Transform (PHAT) that is usually considered useful in reverberant conditions.

$$\psi_{PHAT}(f) = \frac{1}{|\widehat{G}_{x_1x_2}(f)|} \qquad (2)$$

3.2. Position estimation

The availability of multiple TDOA estimations lead to a minimization of an over-determined and non linear error function to obtain a unique estimation of the position. The objective of this error function is to find, by means of minimization, the position from which the theoretical delays are in best accordance with the ones estimated. Direct approaches to solve this minimization problem can be based on iterative search algorithms.

Alternatively to the iterative methods, some closed-form estimators that approximate to a sub-optimal solution at an inexpensive cost have been developed to solve the minimization problem. In this work we will use the simple method for 2-D position estimation proposed in [5]. It basically consists on the weighted contribution of the crossing points of the various bearing lines derived from the angle estimation of each microphone pair. Table 1 (left side) summarizes the basic steps of the algorithm, henceforth referred as Crossing Lines (CL).

Another possible solution for the positioning problem based on space exploration are the Steered Response Power (SRP) techniques. The goal is to maximize the power of the received sound source signal using a steerable beamformer. According to this idea an hybrid approach named SRP-PHAT [4] that searches the maximum of the contribution of the cross-correlations between all the microphone pairs across the space is proposed as a robust solution. Table 1 (right side) shows a summary of the SRP-PHAT algorithm.

CL algorithm	SRP algorithm
1. Angle estimation of each selected microphone pair for each analysis frame. 2. Compute the lines with the angle estimations that crosses the midpoints of the microphone pairs. 3. Compute candidate points from the crossing lines and compute the error associated o each candidate. 4. Final estimation (and error) obtained with the average of the half of the points with less error.	1. Pre-compute theoretical delays from each possible exploration position to each microphone pair. 2. For each analysis frame compute the cross-correlations of each microphone pair. 3. For each position accumulate the contribution of cross-correlations (using delays pre-computed in 1). 4. Select the position with the maximum score.

Table 1: *Summary of CL algorithm and SRP algorithm.*

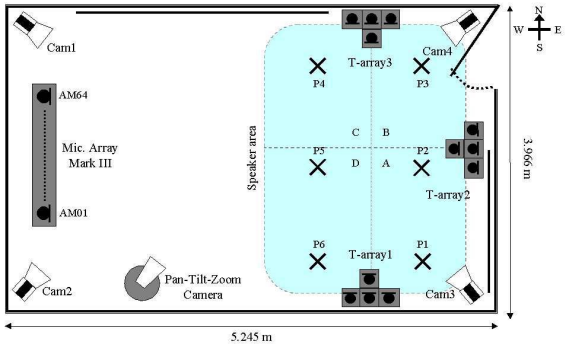

Figure 3: *Sensors set-up in the UPC smart-room.*

4. Database for talker orientation experiments

4.1. UPC Smart-Room

The testing database was collected in UPC smart-room. It is a meeting room equipped with several multimodal sensors such as microphone arrays, table-top microphones, and fixed or pan-tilt-zoom video cameras. The room dimensions are 3966 x 5245 x 4000 mm, which correspond to x, y, z coordinates respectively, and its measured reverberation time is approximately 400 msec.

For speaker localization, the important sensors are the three T-shaped microphone arrays and presumably, the 64 element linear microphone array (from NIST). In this work only the T-shaped microphone arrays were used (from now on referred as T-arrays). The T-arrays are placed on three different walls of the room at the height of 2300 mm (see Figure 3). Each T-array has 4 microphones (Shure Microflex), three of them form a horizontal line with an inter-sensor separation of 200 mm and the last one is placed vertically 300 mm above the central microphone of the horizontal line.

4.2. Database recording

To collect the database, a male subject was moving through a predefined set of six points depicted at Figure 3 (from P1 to P6). At each point, the subject stopped and uttered a few Spanish sentences using four different orientations in such a way that he headed towards each of the four different walls of the room for 10-15 sec. The orientations are denoted as North (N), West (W), South (S) and East (E).

For the recording, the three T-arrays and a close-talk microphone were used. The signals coming from microphones were acquired by Hammerfall audio card at 44.1 kHz sampling rate. The position of subject was marked manually by listening to the audio signal and using the coordinates of the predefined six points (the precision of a reference obtained by using a recording from a calibrated Zenithal camera was not satisfactory). The z coordinate was added based on subject's height. The collected data is freely available at <http://gps-tsc.upc.es/veu/personal/alberto/db/db.tgz>.

5. Experimental results and discussion

The two different techniques summarized in Table 1 called CL-PHAT and SRP-PHAT have been used to carry out the experiments of the influence of talker orientation on the two dimensional talker localization task. The step size in SRP-PHAT exploration is 40 mm and only cross-correlations of the microphones belonging to the same T-array are used. In the case of CL-PHAT 3 bearing lines are estimated for each T-array. In both cases, the delay estimation is achieved by means of the GCC-PHAT of Hanning windowed frames of 4096 samples (approx. 93 msec) with a 50% overlap. Furthermore, cross-power spectral density estimations are smoothed over time with a forgetting factor equal to 0.3.

Figure 4 shows the distance error of the SRP-PHAT position estimation at the point P4 for the all four considered orientations (N, W, S, E). The three different error curves correspond to using a) only the T-array from the S wall (upper curve), b) the two T-arrays from the S and E walls (middle), and c) the all three T-arrays (bottom). The dashed line in each graph denotes the threshold when the localization is considered as correct (300 mm).

A high dependence of the localization error on the talker orientation is obvious from the results when using only one T-array. In this case, when the subject faces S, the direct sound wave is stronger than the reflected waves and the obtained localization errors are near the correct localization threshold. For the other orientations (N, W, E), the effect of reverberation together with the loss of directivity causes the reflected sound waves are stronger than the direct one and the localization error increases. Particularly, the combined effect of the talker directivity and the room reverberation can be observed when the subject faces N and W. Attending only to the talker directivity pattern, we would expected a better performance when the subject faces W than when he faces N. However, better average results are obtained when facing N – Root Mean Square Error (RMSE) of 803mm vs. 1205mm for W – basically due to the fact that in

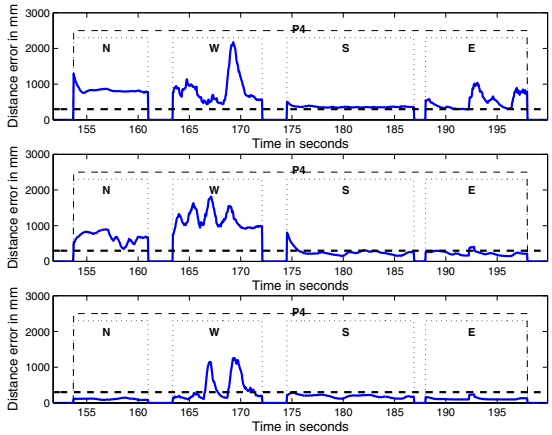

Figure 4: *Error in mm for various orientations in point P4 using T-array 1 (up), T-arrays 1&2 (middle) and the 3 T-arrays (bottom).*

this case the main contribution of the indirect reflections comes from a position closer to the actual subject position.

When using the two T-arrays from the S and E walls, there is a clear improvement when the subject faces E in comparison to using only the S wall T-array. A slight improvement can also be observed when he faces S, but the performance for the other two orientations remains poor. As expected, the performance improves further when using the all three T-arrays; the errors are below the correct localization threshold for the cases when the subject faces N, S, and E. When he faces W (the wall without T-array), the performance improved significantly compared to the one or two T-array case, but it remains still poor.

From these observations some nice properties of the SRP-PHAT technique in relation to the orientation effect can be extracted. This technique combines the contributions of the different GCC estimations in an additive and collaborative way, which means that the contribution of worse microphone pairs (in the sense of their relative position and orientation with respect to the talker) does not seriously affect the performance of the technique if the best microphones are included in the selected set. In other words, with an appropriate distribution of microphones in the room the effect of talker orientation on the localization performance can be importantly diminished using this technique.

A comparison between SRP-PHAT and CL-PHAT is shown in Table 2 for the two following cases: a) using the all three T-arrays and b) using the best possible selection of T-arrays according to the experimental results with all the combinations assuming known position and orientation. Notice that the case b) can be considered as an upper-bound in a hypothetical case of an ideal selection of T-arrays. In Table 2, SRP-PHAT is more robust and it obtains better results than CL-PHAT when using the three T-arrays. By using the ideal selection of T-arrays, a much higher improvement in relative terms is achieved for CL-PHAT (37.13%) than for SRP-PHAT (19.37%). This result indicates that the influence of orientation, and as a consequence the influence of the correct selection of microphones, on the localization performance is stronger in CL-PHAT than in SRP-PHAT.

This difference between SRP-PHAT and CL-PHAT in the orientation effect is mainly due to the multiplicative way in which the microphone pairs are combined in the CL-PHAT technique, i.e., the candidate points are obtained from the crossings of bearing lines and furthermore these lines are estimated based on the maximum of the cross-correlation, which is a hard decision. Hence, it can be deduced that in contrast to SRP-PHAT, the contributions of different T-arrays do not combine in a collaborative way in the CL-PHAT technique. As a result, using extra microphones in addition to the "right" ones not only does not help, but it can harm the final performance, unless a reliable best array indication is available.

Some basic experiments were carried out to investigate how the effect of the talker orientation on the localization performance can be alleviated in a more realistic scenario where it is assumed that the absolute orientation is known. For this purpose, in the case of the SRP technique we propose to use a different combination of T-arrays for the computation of the score at each position depending on the zone being explored. We divided the talker area in 4 sub-areas (see A, B, C and D in Figure 3; the boundaries of the sub-areas were determined by the T-arrays positioning) where a different set of T-arrays is used depending on the orientation. In these experiments, only a 3% relative improvement was obtained compared to the usage of the all three T-arrays, in part due to the already mentioned robustness of the technique to the effect of orientation.

	SRP	CL
3 T-arrays	228.45	533.29
Best T-array selection	184.2	335.16
Relative difference	19.37%	37.15%

Table 2: *RMSE in mm of SRP and CL techniques (mean of all the points and orientations).*

In the case of CL-PHAT, as a first attempt, we computed simultaneously the solution for the given orientation and for the four speaker areas A, B, C and D with different T-clusters set, in order to select the position with the lowest estimated error given by the error function employed in CL-PHAT. The results obtained in this case were not satisfactory. Then we used a table indicating the best selection of T-arrays depending on the orientation and independent on the zone. In this case, a 7% relative improvement was achieved in comparison to the usage of the all three T-arrays.

6. Conclusions

In this work we have shown that talker orientation strongly affects the performance of acoustic localization in smart-rooms due to the combinative effects of talker directivity pattern and room reverberation. However, techniques that join the estimated cross-correlations in a collaborative way, such as SRP-PHAT, have shown to be able to perform nearly independently on the talker orientation if the microphones are distributed appropriately in the room. In the case of coarse techniques based on crossing lines, it has been shown that the multiplicative nature of the fusion of angle estimations yields in a highly dependent orientation method that needs a reliable best array indication to perform robustly. In addition, we proposed basic ideas for introducing the orientation cue into the selected talker localization techniques obtaining slight improvements.

7. Acknowledgements

The authors would like to thank Borja Pachan, Pere Pujol, Dr. Josep Ramon Casas and Dr. Maurizio Omologo for their help. The work has been partially supported by the EU-funded project CHIL (IP506909). Additional support comes from Catalan Government (A. Abad) and Spanish Government (D. Macho).

8. References

[1] Strobel, N., Spors, S. and Rabenstein, R., "Joint Audio-Video Object Localization and Tracking", IEEE Signal Processing Magazine, Jan 2001.

[2] Macho, D. et al., "Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus", ICME05, July 2005.

[3] Chu, W.T. and Warnock, A.C.C., "Detailed directivity of sound fields around human talkers", Tech. Rep. RR-104, National Research Council Canada, Dec 2002.

[4] DiBiase, J. and Silverman, H. and Brandstein, M., "Microphone Arrays. Robust Localization in Reverberant Rooms, chapter 8", Springer, Jan 2001.

[5] Brandstein, M.S. and Adcock, J.E. and Silverman, H.F., "A Practical Time-Delay Estimator for Localizing Speech Sources with a Microphone Array", Computer Speech and Language, Apr 1995.

[6] Knapp, C.H. and Carter, G.C., "The Generalized Correlation Method for Estimation of Time Delay", IEEE Trans. on Acoustics, Speech, and Signal Processing, Aug 1976.