Audio-based approaches to head orientation estimation in a smart-room

Alberto Abad, Carlos Segura, Climent Nadeu and Javier Hernando

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

Abstract

The head orientation of human speakers in a smart-room affects the quality of the signals recorded by far-field microphones, and consequently influences the performance of the technologies deployed based on those signals. Additionally, knowing the orientation in these environments can be useful for the development of several multimodal advanced services, for instance, in microphone network management. Consequently, head orientation estimation has recently become a growing interesting research topic. In this paper, we propose two different approaches to head orientation estimation on the basis of multi-microphone recordings: first, an approach based on the generalization of the well-known SRP-PHAT speaker localization algorithm, and second a new approach based on measurements of the ratio between the high and the low band speech energies. Promising results are obtained in both cases, with a generalized better performance of the algorithms based on speaker localization methods.

Index Terms: head orientation estimation, microphone arrays, speaker tracking

1. Introduction

The measurements reported in [1] show that human talkers do not radiate voice sound uniformly in all directions; more energy is radiated in talker's forward direction than towards the side or the rear direction. Additionally, the radiation pattern is frequency dependent; behind the talker, the low speech frequencies are attenuated less than the high speech frequencies. As a consequence, the quality of the speech captured by a far-field microphone in an indoor environment, in addition to be degraded by acoustic noise and room reverberation, it is also influenced by the relative orientation of the speaker with respect the recording microphone. Thus, speech applications based on distant-talking microphones are also affected by head orientation. For instance, the influence of head orientation on the performance of source localization algorithms was studied by the authors in [2]. It was shown that the incorporation of head orientation information permits improving speaker localization.

In general, it can be expected that knowledge about the orientation of human speakers would permit improving speech technologies that are commonly deployed in smart-rooms. For instance, an enhanced microphone network management strategy for microphone selection can be developed based on both speaker position and orientation cues. Additionally, in applications that require human-computer interaction the orientation cue allows a better understanding of what users do or what they refer to. This information can also be exploited by several other multimodal applications, such as automatic camera steering in video conferences. Hence, in the context of the development of smart-room advanced services, the estimation of the head orientation has recently emerged as an interesting field of research.

Actually, the problem of head orientation estimation has been mostly tackled based only on visual cues. However, since humans rely on both acoustic and visual information to perform orientation estimation, one can think that speech should also be used to infer information about source orientation. In fact, the interest on this problem based on multi-channel speech observations is so recent and challenging that very few works can be found in the speech related literature. Most of these works can be coarsely classified into two different classes.

On the one hand, most recent proposals are related to the development of robust speaker localization methods that try to incorporate head orientation as a new search parameter. The aim of these methods is first to achieve more reliable source position estimation performance, since the possible degrading effect of head orientation is accounted for, and second to obtain also an estimation of the orientation. This is the case of [3], that based on the steered response power with PHAT transform (SRP-PHAT) [4] localization algorithm, extends the search with the orientation variable by weighting the contribution of each microphone pair for different possible orientations. A similar approach also based on the SRP-PHAT algorithm can be found in [5], named the Oriented Global Coherence Field (OGCF) method.

On the other hand, some alternative approaches are based on radiation and propagation characteristics of the speech signal [6]. These methods rely on the measurement of the acoustic energy received by several microphones, which is used to infer some information about the speaker orientation. Usually, these methods need to know the speaker position beforehand.

In this paper, two alternative head orientation methods are proposed and evaluated. First, an estimator based on the SRP-PHAT algorithm (similar to the ones of [3] and [5]). Second, a new estimator previously introduced as part of a multi-modal system in [7] based on talker directivity considerations and a measure of the ratio between the high and the low band speech energy that we have named the HLBR measure. In both cases, encouraging results for future research efforts are obtained. Particularly, SRP-PHAT based methods show a generalized better performance, while HLBR methods stand out as simple alternative solutions when the source position is known beforehand.

2. SRP-PHAT head orientation estimation

In this section we describe our algorithm for head orientation estimation based on the well-known SRP-PHAT [4] algorithm for

This work has been partially supported by the EU-funded project CHIL (IP506909) and the Spanish Government funded project ACESCA (TIN-2005-08852). A. Abad is currently working at INESC ID's Spoken Language Systems Laboratory (L^2F).

speaker localization. Indeed, it is based on the robust speaker localization algorithm that we described in [8].

2.1. Description of the SRP-PHAT localization algorithm

The SRP-PHAT technique is a space exploration algorithm aimed to search for the maximum of the contributions of the cross-correlations between microphone pairs. Concretely, the technique solves the Time Delay of Arrival (TDOA) estimation problem based on generalized cross-correlations with phase transform (GCC-PHAT) and the position estimation problem based on steered response power (SRP) search in an integrated and robust way. Actually, it has turned out the most successful state of the art approach to microphone array sound localization.

Consider that the GCC-PHAT of a microphone pair p is $R_p(\tau)$. It can be expressed in the frequency domain in terms of the Fourier transform of the signals of microphone pair ($X_{p_1}(f)$ and $X_{p_2}(f)$) as follows:

$$R_p(\tau) = \int_{-\infty}^{+\infty} \frac{X_{p_1}(f)X^*_{p_2}(f)}{|X_{p_1}(f)||X^*_{p_2}(f)|} e^{-j2\pi f\tau} \qquad (1)$$

Then, the SRP-PHAT algorithm can be mathematically formulated as the maximization of a Spatial Likelihood Function (SLF) $F(\mathbf{x})$ formed by the contributions of each of the P individual cross-correlations as follows:

$$F(\mathbf{x}) = F(\mathbf{T}(\mathbf{x})) = \sum_{p=1}^{P} R_p(\tau_p(\mathbf{x})) \qquad (2)$$

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \sum_{p=1}^{P} R_p(\tau_p(\mathbf{x})) \qquad (3)$$

That is, for each spatial position $\mathbf{x}$, the time delay vector $\mathbf{T}(\mathbf{x}) = \begin{bmatrix}\tau_1(\mathbf{x}) & \tau_2(\mathbf{x}) & \dots & \tau_P(\mathbf{x})\end{bmatrix}$ formed by theoretical delays to each microphone pair is found and the pairwise cross-correlations are computed. The likelihood assigned to each position is equal to the sum over all pairwise cross-correlations, and the position with a maximum likelihood is the estimated source position.

2.2. Head orientation estimation based on SRP-PHAT

Head orientation estimation can be tackled based on the joint maximization of a SLF depending simultaneously on the potential source positions and orientations. Thus, the SRP-PHAT likelihood function of Equation 2 can be extended to incorporate orientation as follows:

$$F(\mathbf{x}; o) = F(\mathbf{T}(\mathbf{x}), \mathbf{O}(\mathbf{x}; o)) = \sum_{p=1}^{P} \Omega_p(\mathbf{x}, o) R_p(\tau_p(\mathbf{x})) \qquad (4)$$

For each spatial position $\mathbf{x}$ and orientation o, in addition to the time delay vector $\mathbf{T}(\mathbf{x}) = \begin{bmatrix}\tau_1(\mathbf{x}) & \tau_2(\mathbf{x}) & \dots & \tau_P(\mathbf{x})\end{bmatrix}$ formed by theoretical delays to each microphone pair, a weight vector $\mathbf{O}(\mathbf{x}; o) = \begin{bmatrix}\Omega_1(\mathbf{x}, o) & \Omega_2(\mathbf{x}, o) & \dots & \Omega_P(\mathbf{x}, o)\end{bmatrix}$ formed by an appropriate weight representing the influence of each cross-correlation, in terms of the relative orientation, is computed.

In order to calculate $\Omega_p(\mathbf{x}, o)$, first the angle difference between the orientation explored and the line that crosses the middle point of the microphone pair and the position explored ($\theta_p(\mathbf{x})$) is computed:

$$\Delta\theta_p(\mathbf{x}; o) = \theta_p(\mathbf{x}) - o \qquad (5)$$

Then, the following function depending on the angle differences that approximates a normalized talker directivity pattern is used for computing the weights.

$$\Omega_p(\mathbf{x}, o) = F(\Delta\theta_p(\mathbf{x}; o)) = \frac{1}{1 + 0.6sin^2(\Delta\theta_p(\mathbf{x}; o)/2)} \qquad (6)$$

Similarly to conventional SRP-PHAT, the likelihood assigned to each position and orientation is equal to the sum over all the pairwise cross-correlations weighted according to equation 6. The joint estimated position and orientation are the ones that maximize the likelihood function of equation 4:

$$\{\hat{\mathbf{x}}, \hat{o}\} = \arg\max_{\mathbf{x}, o} F(\mathbf{T}(\mathbf{x}), \mathbf{O}(\mathbf{x}; o)) \qquad (7)$$

Notice that the contribution of each cross-correlation is weighted according to the degree of reliable information that provides a microphone pair depending on its relative position and orientation with respect to the source. Consequently, coupling the orientation parameter in the search procedure in this way might presumably provide more robust estimations of the source position.

2.3. Practical problems and alternative solutions

One major problem arises with the proposed algorithm. In practice, the above formulation is equivalent to compute a different SLF for each possible orientation, consequently, the problems of high computational load of the SRP-PHAT algorithm is of major relevance in this case due to the growth of operations and memory requirements. As a result, the exhaustive search of the function of Equation 7 is unfeasible in real time applications. One possibility to partially solve this problem would consist on applying efficient searching strategies, like the proposed two-pass search algorithm described in [8].

Alternatively, it is possible to decouple the problem into two different stages. In the first stage, the source position can be estimated by means of conventional SRP-PHAT algorithm. In a second stage, the likelihood of the various orientations can be computed only at the estimated source position. Then, the orientation that maximizes the likelihood at this position, is the estimated head orientation. Notice that the benefits of introducing orientation information in the source localization problem are lost in this way. Hereinafter, this simplified method will be referred to as the *fast* SRP-PHAT head orientation estimator (SRPPHAT-F), while the exhaustive *joint* SRP-PHAT search estimator will be referred as SRPPHAT-J.

3. HLBR head orientation estimation

Knowledge about the human radiation pattern can be used to estimate the head orientation of an active speaker by computing the energy received at each microphone and searching the angle that best fits the radiation pattern with the energy measures –assuming that a well-distributed network of microphones is available–. This is done in [6] on the basis of a large aperture linear microphone array. However, this approach has several problems since the microphones should be perfectly calibrated and different attenuation at each microphone due to propagation must be accounted for, thus requiring the use of sound propagation models.

In this section alternative solutions for estimating the head orientation from acoustic measurements are presented. In the proposed methods, the computational simplicity is kept by using acoustic energy normalization to solve the aforementioned problems.

3.1. The HLBR measure

The energy at the low frequency band radiated by an active speaker is low directional, while, at the high frequency range the radiation pattern is highly directive [1]. One can make use of this fact to define the High/Low Band Ratio (HLBR) of a radiation pattern. The HLBR of a radiation pattern is defined as the ratio between high and low bands of frequency of the radiation pattern. In Figure 1, a diagram of the speech radiation pattern in the horizontal plane is shown on the left side. On the right, measurements of the HLBR for different orientations are depicted. It can be seen that the measurements of the HLBR of a radiation pattern keeps a similar shape to the one of the speech radiation diagram, thus retaining the dependency with the orientation. In this work, the low band considered is the range of frequencies from 200 Hz to 400 Hz, while the high band is the range from 3500 Hz to 4500 Hz.

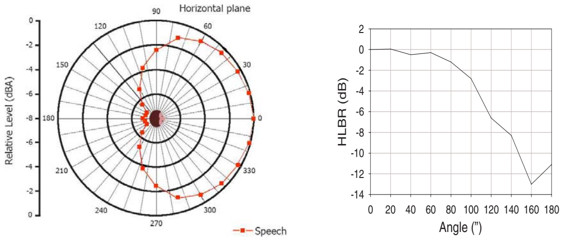

Figure 1: *On the left, talker diagram in the horizontal plane (after [1]). On the right, HLBR of the head radiated pattern.*

3.2. HLBR based head orientation estimators

Alternative methods of head orientation estimation to the computation of the absolute energy received at each microphone can be defined based on the estimation of the HLBR of the acoustic energy of several distributed sensors. The advantage of the HLBR measure is that its value is directly comparable across all microphones since, after the normalization, the effects of bad calibration and propagation losses are cancelled.

On the one hand, these measurements can be used to find the orientation that fits better a model of the HLBR radiation pattern depending on the head pose, as it was done in [7]. On the other hand, in this work we also present an alternative and more simple way of using the HLBR measurements in order to estimate the head orientation.

Both techniques based on HLBR first need the position of the source $\mathbf{x}$ to be known beforehand or estimated by means of any source localization method. Then, the vectors $\mathbf{v}_q$ from the speaker to each microphone $\mathbf{m}_q$ with module $|\mathbf{v}_q|$ (equal to the HLBR measure of the microphone) and angle θ_q are computed.

In [7], the estimated speaker orientation is computed by searching the angle that maximizes the correlation between a mathematical model of the HLBR of a radiated pattern $\mathbf{G}(\theta)$ and the HLBR of the acoustic energy measured at each microphone:

$$\hat{o} = \arg\max_{\theta} \sum_{q=1}^{Q} |\mathbf{v}_q| \mathbf{G}(\theta - \theta_q) \qquad (8)$$

In order to model $\mathbf{G}(\theta)$, an appropriate Gaussian function or the function defined by Equation 6 have been experimentally tested to provide good results. In the following experiments, the function of Equation 6 is the one considered. Hereinafter, this estimator will be referred to as the *basic* HLBR head orientation estimation method (HLBR-B).

A new efficient solution to the head orientation estimation problem also based on HLBR measures, named the *vectorial* HLBR head orientation estimation method (HLBR-V), is also proposed. The advantage of the proposed technique with respect to the HLBR-B is that it does not depend on a model of the HLBR radiated pattern. Additionally, it does not need to be evaluated for a constrained number of potential orientations. Concretely, the algorithm simply consists on the computation of the sum vector of all the HLBR vectors of each microphone. The angle of the resulting sum vector is considered to be the estimated head orientation:

$$\mathbf{v}_{sum} = \sum_{q=1}^{Q} \mathbf{v}_q \qquad \hat{o} = \angle \mathbf{v}_{sum} \qquad (9)$$

4. Experimental evaluation

The two proposed head orientation estimation methods described above are tested and compared in this section: the SRP-PHAT based method in its two versions – SRPPHAT-J and SRPPHAT-F – and the HLBR method in its two versions also – HLBR-B and HLBR-V–.

Regarding implementation details, SRP-PHAT based methods are based on the system proposed in [8]. In both cases, the number of possible orientations is fixed to 8. With respect to the HLBR based techniques, the source position is previously estimated also with the system described in [8]. Additionally, a first order filter is used to smooth the estimations of the high and low frequency band with a forgetting factor equal to 0.7.

Notice that a Kalman filter could be used to obtain enhanced estimations in both SRP-PHAT and HLBR methods [7]. In this work, it has not been considered in order to better evaluate the potentials of each method.

4.1. Data description and evaluation metrics

Head orientation estimation is evaluated with the CLEAR head pose database [9]. It consists on an extract of 3 seminars from the data collected by the CHIL –Computers In the Human Interaction Loop– consortium for the CLEAR 2006 evaluation that was labelled for particular head pose evaluation purposes. The seminars were recorded in a non-interactive indoor scenario where a person was giving a talk, for a total of approximately 15 min. All results described in this work were obtained using a set of four T-shaped 4-channel microphone clusters.

Metrics and scoring of the systems has been done following the common agreement of the CHIL consortium for head pose evaluation. Three basic metrics are defined:

Pan Mean Average Error (PMAE) [degrees] This is the precision of the head orientation angle estimation.

Pan Correct Classification (PCC) [%] This is the ability of the system to correctly classify the head position within 8 classes spanning 45° each.

Pan Correct Classification within a Range (PCCR) [%] This is the ability of the system to correctly classify the head position within 8 classes spanning 45° each, allowing a classification error of ±1 adjacent class.

4.2. Experimental results

Table 1 summarizes the results obtained by the methods under study. First, it is clear that methods based on SRP-PHAT show a generalized better performance than methods based on HLBR. On the one hand, the SRPPHAT-J and SRPPHAT-F show in practice almost identical head orientation estimation performances. Since SRPPHAT-F is much less computationally demanding, it can be considered the most convenient option for head pose orientation estimation. On the other hand, it seems that the HLBR-V method is slightly superior to the HLBR-B. Consequently it can be considered the most convenient option to estimate head orientation if the source position is known beforehand or obtained with a different source localization estimator than the SRP-PHAT algorithm.

Method	PMAE	PCC	PCCR
SRPPHAT-J	44.68°	37.32%	73.38%
SRPPHAT-F	44.23°	37.71%	73.89%
HLBR-B	52.92°	29.85%	67.99%
HLBR-V	50.98°	32.61%	68.94%

Table 1: *Head pose orientation results of the four methods evaluated.*

Both SRPPHAT-J and SRPPHAT-F algorithms obtain similar results in the head pose orientation estimation task. However, it should be confirmed that source localization performance provided by the SRPPHAT-F method is equivalent to the SRPPHAT-J technique, despite it does not incorporate the orientation parameter in the search process. Table 2 shows the source tracking results of the two algorithms evaluated with the seminars of the CLEAR head pose database. Speaker localization is assessed in terms of the Multiple Object Tracking Precision (MOTP) [mm] –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– and the Acoustic Multiple Object Tracking Accuracy (A-MOTA) [%] –it is one minus the ratio of the sum of misses and false positives over all frames and the total number of frames–.

System	MOTP	Miss	FalsePos	A-MOTA
SRPPHAT-J	157mm	18.28%	10.18%	71.53%
SRPPHAT-F	160mm	17.42%	11.90%	70.67%

Table 2: *Audio person tracking results of the SRPPHAT-J and SRPPHAT-F algorithms evaluated with the CLEAR head pose database.*

From these results it can be stated that the incorporation of orientation information in the search of the SRP-PHAT algorithm provides an enhanced source localization performance as it was expected. However, the improvement is achieved in exchange of an high increase in the computational load of the algorithm. Since the enhancement obtained is not very remarkable and the orientation estimation performance of both approaches is equivalent, the SRPPHAT-F method can be still considered as the most convenient approach to both speaker localization and head orientation estimation.

5. Conclusions

In this paper, it has been shown that head orientation estimation can be achieved by means of both SRP-PHAT based approaches and more simple techniques based on speech radiation considerations. On the one hand, coupling the source localization and orientation estimation problem in a joint search permits obtaining enhanced source localization estimations and reasonably good results in head pose estimation, but in exchange of high memory and computational expense requirements. In order to partially solve these problems, a faster head orientation estimation consisting in the evaluation of the likelihoods of the various possible orientations at the estimated source position can be applied, without a remarkable drop of performance in both source localization and head pose estimation tasks. On the other hand, inexpensive head orientation estimation methods based on the ratio between the energies of the high and low band frequencies (HLBR) measure have been also presented. Although these approaches show in general a worse performance compared to the more sophisticated SRP-PHAT based algorithms, they can become an interesting alternative for instance when source position is known beforehand, when the source localization algorithm is not based on SRP-PHAT, or when memory and computational cost are constrained.

6. References

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

[2] Abad, A., Macho, D., Segura, C., Hernando, J. and Nadeu, C., "Effect of Head Orientation on the Speaker Localization Performance in Smart-room Environment", in Proceedings of Interspeech, pp. 145–148, 2005.

[3] Mungamuru, B. and Aarabi, P., "Enhanced Sound Localization", IEEE Transactions on Systems, Man and Cybernetics, Vol. 34(3), pp. 1526–1540, 2004.

[4] DiBiase, J., Silverman, H. and Brandstein, M., "Robust Localization in Reverberant Rooms", in Microphone Arrays: Signal Processing Techniques and Applications, Chapter 8, pp. 157–180, Springer-Verlag, 2001.

[5] Brutti, A., Omologo, M. and Svaizer, P., "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays", in Proceedings of Interspeech, pp. 2337–2340, 2005.

[6] Sachar, J. M. and Silverman, H. F., "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array", in Proceedings of ICASSP, Vol. 4, pp. 65–68, 2004.

[7] Segura, C., Cantón, C., Abad, A., Casas, J.R. and Hernando, J., "Multimodal head orientation towards attention tracking in smart rooms", in Proceedings of ICASSP, 2007.

[8] Abad, A., Segura, C., Macho, D., Hernando, J. and Nadeu, C., "Audio Person Tracking in a Smart-Room Environment", in Proceedings of Interspeech, pp. 2590–2593, 2006.

[9] Stiefelhagen, R. and Garofolo, J. (Eds.), "Multimodal Technologies for Perception of Humans. First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006", Lecture Notes in Computer Science, Vol. 4122, Springer-Verlag, 2007.