Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays

Alessio Brutti, Maurizio Omologo, Piergiorgio Svaizer

Istituto Trentino di Cultura(ITC)-irst
Via Sommarive, 18, Povo-Trento, Italy
brutti/omologo/svaizer@itc.it

Abstract

This paper proposes a new method for estimating the talker's head orientation in a smart room equipped with microphone arrays. The acoustic processing is based on the use of a coherence measure derived from the Cross-power spectrum phase analysis, commonly used for speaker localization and tracking purposes. An Oriented Global Coherence Field function is then introduced to assign to a given point in space different scores according to the possible orientation of the acoustic source.

Preliminary experiments were conducted in the CHIL smart room available at ITC-irst laboratories. A small database was collected, given one speaker uttering a given sentence in different directions. Although the database is rather small, results show that the proposed method is promising and ensures a good estimation of the head orientation.

1. Introduction

An attractive future scenario consists in the development of new workspaces where the so-called "ambient intelligence" is realized through a wide usage of sensors (cameras, microphones, etc.) connected to computers that fade in the background, largely invisible and significantly less intrusive to humans.

Towards this direction of ubiquitous computing, removing any constraint on the distribution of the microphones in space represents an important potential in terms of flexibility of the application, namely speaker tracking or distant-talking Automatic Speech Recognition (ASR).

Speech signals recorded by microphones placed far from a talker are severely degraded by both background noise and reverberation, which depends on spatial relationships between the microphones and the talker him/herself. Although noticeable advances have been made during the last decade in speaker tracking and distant-talking ASR [1], the existing prototypes are often based on the use of a microphone array, located in a predetermined position and characterized by a specific geometry. Moreover, even if the user is not encumbered anymore by hand-held or head-mounted microphones in most of the cases she/he can talk up to a limited distance from the microphones (a few meters), depending on the complexity of the environmental acoustics, and with constraints on the head orientation and on the speaking style.

In the CHIL project [2] [1], a multi-channel scenario with an arbitrary microphone distribution is addressed, as an alternative to a traditional microphone array. Rather than focusing primarily on improving the quality of a spatial filtering process for enhancement purposes, we intend to analyze the given acoustic scenario through a multi-channel processing aimed at extracting basic information to track talkers, classify acoustic events, and eventually recognize what has been uttered. The reference scenarios of CHIL are lectures and meetings, but an extension of the resulting technologies to other contexts, as smart homes, videoconferencing, etc., is envisaged.

In the given context, an important issue is the estimation of the talker's head orientation. From a rather accurate estimation of the head pose, one can obtain an effective combination of audio and visual sensor processing for person tracking purposes as well as a selection of a subset of microphones from which to derive a more significant acoustic representation at front-end processing level for distant-talking ASR purposes.

The automatic estimation of head orientation from acoustic signals is a rather new and challenging topic. A preliminary work was described in [3], which was based on the availability of a huge microphone array consisting of hundreds of microphones linearly distributed along the walls of a given room, this way forming a surrounding aperture in a plane parallel to the floor. The method introduced in [3] was based on an acoustic energy information derived from the related array processing.

This work proposes to use the coherence between microphone pair signals as fundamental information on which the head orientation is estimated. This coherence is combined to provide the Global Coherence Field [4], that represents the plausibility of an acoustic source to be active at a given point in space. Then, a so called Oriented Global Coherence Field (OGCF) is here introduced, which allows to obtain a score for an active source located at a given point in space and oriented in different possible directions.

It is worth noting that the approach introduced here does not require a large-aperture linear microphone arrays. Instead, the experimental activity here described was conducted just using a small set of microphone clusters distributed in the CHIL room available at ITC-irst (see Figure 1). Each cluster is based on a T-shaped array as depicted in Figure 2.

Preliminary results show that the proposed method for the estimation of head orientation is promising both in terms of performance and for what regards required computational load as well as sensor set-up complexity and cost.

2. Head radiation

The concept of head radiation is necessary in the study of speech propagation in an enclosure, since the active speaker can not be approximated as an omnidirectional point source. In general, any sound source (e.g. a loudspeaker) has propagation directivity characteristics that lead to a non-spherical radiation, mainly determined by the size and the shape of the source and the frequency being analyzed.

[1] This work was partially supported by the EU under the Integrated Project CHIL (IP 506909).

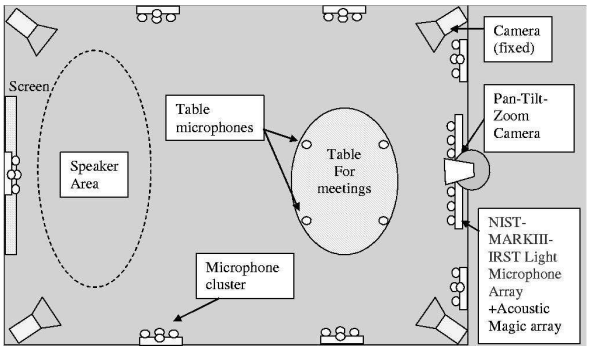

Figure 1: Map of the CHIL room at ITC-irst. The set-up was conceived to collect data from seminars and meetings.

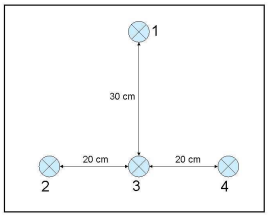

Figure 2: Geometry of the T-shaped array of microphones used in the CHIL smart rooms.

For what regards speech, given that the head diameter is slightly less than 20 cm and that the sound is mainly diffused from the mouth of the speaker, one can expect a more directional radiation for frequencies above 500 Hz, with different properties in the horizontal and in the vertical planes [5].

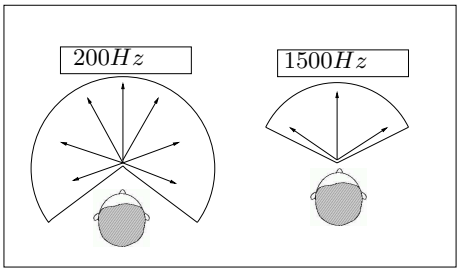

Figure 3: Shape of the head radiation at different frequencies.

As a first consequence of these facts, for higher frequencies the human head influences the timbre and causes a radiation between 5 and 15 dB lower behind the head than in front of it. Moreover, one can observe that the direct wavefront propagates just in the frontal hemisphere, and in a way that also depends on the vertical orientation of the head. For instance, when a person is sitting at a table and is speaking with the mouth oriented toward the hard surface of the table, there will be a dominance of reflected waves in the speech propagation in the environment.

Here, the basic idea is to identify the head orientation on the basis of the coherence between the signals recorded by a subset of microphone pairs distributed in the given environment.

3. CSP-based Coherence Measure

Given a set of microphones, the information on mutual delay between microphone pair signals can be associated to a Coherence Measure (CM) function $C_{ik}(t,\tau)$ that expresses, for a hypothesized delay τ, the similarity between segments (centered at time t) extracted from two generic signals s_i and s_k. This measure can be derived from the CSP analysis [6], as described in the following.

Denoting with $s_i(n)$ and $s_k(n)$ the discrete time sequences in the given interval, which were obtained by sampling the signals acquired by microphones i and k, the CSP is defined as:

$$C_{ik}(t,l) = DFT^{-1}\left\{\frac{DFT(s_i(n))\cdot DFT^{*}(s_k(n))}{|DFT(s_i(n))|\cdot|DFT(s_k(n))|}\right\} \quad (1)$$

where l denotes the time lag.

CSP analysis, also known as Generalized Cross-Correlation PHAse Transform (GCC-PHAT), is commonly used for time delay estimation in speaker localization and tracking systems, thanks to its independence from spectral characteristics of input signals. More details can be found in [7].

In particular, as shown in [4],[6], the CSP-CM function $C_{ik}(t_0,\tau)$, computed for a time interval centered at the time instant t_0, has a prominent peak at a delay $\tau = \delta_{ik}$ corresponding to the direction of wavefront arrival. Note that, for a given microphone pair, the resulting direction of arrival identifies one of the two sheets of a hyperboloid in the most general case, or a cone when one assumes a far-field propagation modeling.

4. Global Coherence Field

A Global Coherence Field (GCF) is a function, defined over the space of possible sound source locations, which represents the plausibility that an active sound source is present at a given point. It was conceived starting from the Power Field (PF) introduced in [8] and then used for speaker localization and tracking purposes.

The Power Field (PF) represents the power of the signal obtained at the output of a beamformer, as a function of the point of space at which the array is steered. In other words, if the location space is subdivided by a grid Σ of potential source locations $\mathbf{p_s} = (x_s, y_s, z_s)$ and the corresponding sets $\tau_{\mathbf{s}}$ of steering delays are used to "scan" the space by means of the array, the power of the output signal, when the array is steered at a given location, can be used to derive a degree of plausibility that the source is located at that point.

In a similar manner, GCF is defined by considering the average coherence between signals realigned by the beamformer, instead of the power of its output. An example is GCF that derives from the CSP-based Coherence Measure introduced in the previous section.

Let us consider a set Ω of Q microphone pairs and denote with $\delta_{ik}(x,y,z)$ the theoretical delay for the microphone pair (i,k) if the source is at position (x,y,z). Once the Coherence Measure $C_{ik}(t,\delta_{ik}(x,y,z))$ has been computed at instant t, for each microphone pair (i,k) belonging to Ω, the GCF is expressed as:

$$GCF_{\Omega}(t,x,y,z) = \frac{1}{Q}\sum_{(i,k)\in\Omega} C_{ik}(t,\delta_{ik}(x,y,z)). \quad (2)$$

In this work, GCF function was computed exploiting each T-shaped array, and considering the three pairs (1,3),(2,3),(3,4) of each array (i.e. Q=3). Figure 4 shows an example of GCF restricted to a plane (x,y).

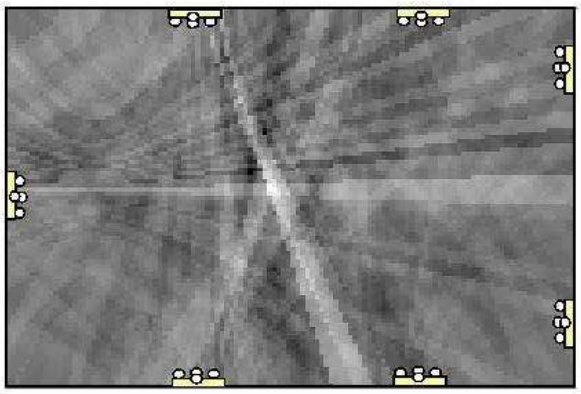

Figure 4: CSP-based 2-dimensional GCF computed using the 5 leftmost microphone clusters. GCF magnitude is represented by the whiteness of the plotted points. The whitest spot in the center of the room corresponds to the active speaker.

5. Oriented GCF

Let us now assume that the sensor set up consists of L T-shaped arrays Ω_l distributed in the room (see Figure 5) and that the sound source is located in $S = (x_s, y_s, z_s)$. The source location can be accomplished by maximizing GCF over a given grid of points in the room, or using a generic speaker location method as described in [1]. The following section introduces a method to derive an oriented GCF function that depends on the radiation pattern generated by the sound source.

5.1. Orientation Estimation

Let us consider a circle C, centered at a generic point S and having radius r, and N equispaced points P_j on C, which correspond to N possible orientations (see Figure 5). Consider now the intersections Q_l between the lines from S to each microphone cluster Ω_l.

A score for every orientation o_j, with j from 0 to $N-1$, is computed through the GCFs evaluated at the points Q_l. In other words, every $GCF_{\Omega_l}(Q_l)$ is computed assuming that the sound source be located in Q_l. Hence, the Oriented Global Coherence Field at S is computed as:

$$O_j(S) = \sum_{l=0}^{L-1} GCF_{\Omega_l}(Q_l) w(\theta_{lj}) \tag{3}$$

where $w(\theta_{lj})$ is a weigth computed from a gaussian function:

$$w(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \tag{4}$$

and $\theta_{lj} \in [-\pi, \pi]$ is the angle between the line passing through S and P_j and the line from S to Q_l.

As a result, the weights $w(\theta_{lj})$ related to the orientation o_j will emphasize the contributions of GCFs in points Q_l closer to P_j (i.e. the direction o_j) and deemphasize the contributions corresponding to points in the opposite directions. The orientation index j for which $O_j(S)$ is maximum can then be assumed to indicate the sound source orientation.

Note that the proposed method depends on several choices, namely: the radius r of the circle, the weighting function $w(\cdot)$, the number of directions N to analyze for a given possible sound source position, and the number of points on the circle C which in this case corresponds to the number of microphone clusters in the room.

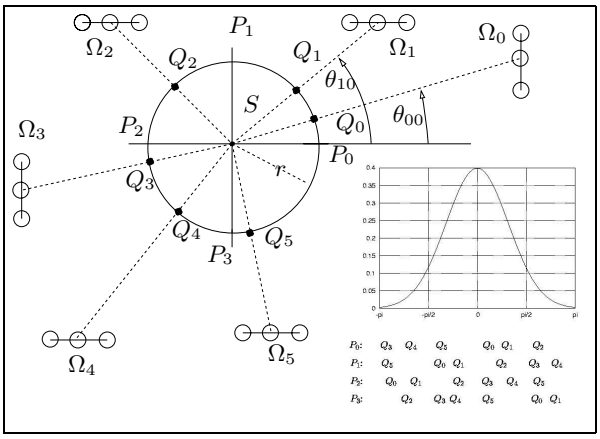

Figure 5: Graphical representation of the orientation estimation scheme described in Sec. 5. In this case 6 microphone clusters are available and 4 possible orientations are investigated.

6. Experiments and Results

6.1. Experimental set-up

For a preliminary experimental activity, aimed at verifying the effectiveness of the proposed method, the CHIL sensor set up available at ITC-irst was adopted. In particular, given a central position of the real speaker, the 5 leftmost T-shaped arrays of Figure 6 were exploited.

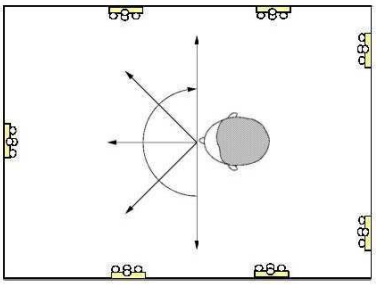

Figure 6: Outline of the speaker position and the 5 orientations adopted in the experiment. The two arrays at the right side of the room were not used in the given experimental framework.

The room was characterized by a reverberation time of more than 600 ms, which makes both speaker location and head orientation estimation problems quite difficult.

Collected data consisted of a sequence of sentences uttered by one male speaker. The speaker repeated the same sentence (of about 7 seconds length) five times, standing at the same position but every time with a different orientation, rotating by a step of about 45 degrees in order to sweep the range between 0 and 180 degrees. The distance between the speaker and the microphone clusters was about 3 meters.

Signals were sampled at 44.1 kHz. The total sequence length was about 45 seconds.

The procedure to compute the OGFC consisted in the three following steps:

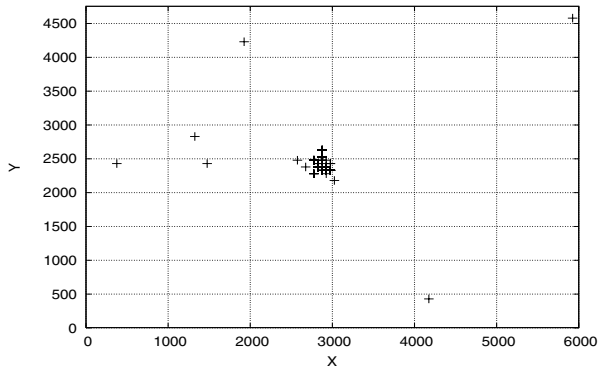

Figure 7: Sound source locations estimated for the given test set.

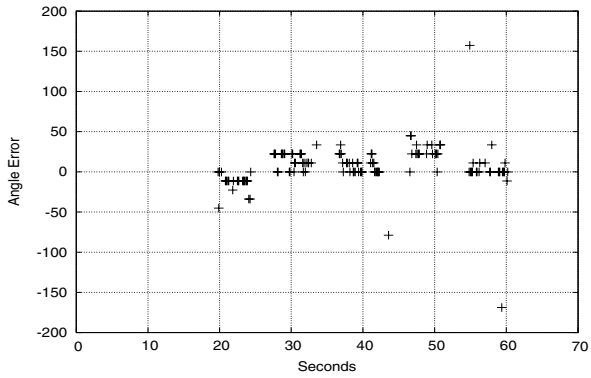

Figure 8: Orientation error, expressed in degrees, as a function of time.

- CSP based Global Coherence Field computation;
- Speaker localization based on maximizing GCF over a given grid of points;
- Orientation estimation for the located speaker position.

6.2. Speaker localization

For speaker localization purposes GCF functions were evaluated on a 50×50 mm grid. The analysis frame on which CSP was computed had a length of about 180 ms.

A confidence threshold (empirically set to 0.4) was then defined for GCF, in order to make a decision of "located active speaker". When GCF is below the threshold, a deletion error occurs. For what concerns the 134 frames classified as "located active speaker", 6 of them corresponded to errors larger than 50 cm from the true speaker position (at x=2.9 m, y=2.4 m), while the other 128 can be considered as fine localization errors, as shown in Figure 7.

6.3. Head orientation

For the head orientation estimation task, every frame classified as "located active speaker" was then analyzed by using the OGCF-based method proposed above.

Based on a preliminary analysis, r was fixed at 3 cm, while the number of microphone clusters and of directions N to investigate were 5 and 32, respectively. In practice, a small radius is necessary to have consistency among the coherence measures extracted from the given signals: although we observed a reasonable behaviour of the system also with larger values of r, the limited size and characteristics of the test data set do not suggest to draw any further conclusion at this level.

Figure 8 represents the error in the orientation estimate with respect to the true head orientation. From this result, one can observe that except for two frames the proposed method always ensures an orientation error lower than 45 degrees, with an average error of about 5 degrees and a related RMSE of 17 degrees.

7. Conclusions and Future work

This paper introduced a new method to estimate the speaker's head orientation, given a distributed microphone network. Although the work has to be considered as a only preliminary study, the experimental results obtained in a real noisy and reverberant environment are promising. With the proposed method, one can derive a sound source orientation that has a good correlation with the true speaker head orientation adopted during the data collection.

Currently, a new database is being collected through the use of a loudspeaker and with a very accurate calibration of the orientation angle, in order to extend this activity towards a better established evaluation process.

Moreover, a detailed analysis of the influence of some parameters as, for instance, the radius r, the number of possible orientations, and the type of weighting function is planned.

The resulting system for the estimation of head orientation will then be integrated with the speech activity detection and the speaker localization and tracking system being developed in CHIL and, eventually, with the systems for person tracking and head pose estimation based on visual sensor processing.

8. References

[1] M. Brandstein, D. Ward, "Microphone Arrays", *Springer Verlag*, 2001.

[2] D. Macho et al., "Automatic Speech Activity Detection, Source Localization, and Speech Recognition on the CHIL Seminar Corpus", to appear in *Proc. ICME*, 2005.

[3] J.M. Sachar, H.F. Silverman, "A Baseline algorithm for estimating Talker Orientation using Acoustical Data from a Large-aperture Microphone Array", *Proc. IEEE ICASSP*, Montreal 2004, vol. 4, pp. 65–68.

[4] R. De Mori, "Spoken Dialogues with Computers", Chapter 2, *Academic Press*, 1998.

[5] H. Kuttruff, "Room Acoustics", *Elsevier Applied Science*, Amsterdam, 1991.

[6] M. Omologo, P. Svaizer, "Acoustic Event Localization using a Crosspower-Spectrum Phase based Techniques", *Proc. IEEE ICASSP*, Adelaide 1994, vol. 2, pp. 273–276.

[7] M. Omologo, P. Svaizer, "Use of the Crosspower-Spectrum Phase in Acoustic Event Location", *IEEE Trans. on SAP*, vol. 5, N. 3, pp. 288–292, May 1997.

[8] V. Alvarado, "Talker Localization and Optimal Placement of Microphones for a Linear Microphone Array using Stochastic Region Contraction", PhD Thesis, *Technical Report LEMS-69*, Brown University, 1990.