Speaker Localization based on Oriented Global Coherence Field

Alessio Brutti, Maurizio Omologo, Piergiorgio Svaizer

Istituto Trentino di Cultura (ITC)-irst
Via Sommarive 18, 38050 Povo, Trento, Italy
{brutti|omologo|svaizer}@itc.it

Abstract

This paper proposes a new speaker localization method that is based on a preliminary estimation of the head orientation. The basic information on which the estimation is accomplished is called Oriented Global Coherence Field (OGCF).

The new algorithm is shown to be significantly more robust than the traditional ones so far explored. Its robustness is also due to an effective speech activity detection, implicitly performed by a thresholding technique applied to OGCF information. To show the performance of the proposed system, experiments were conducted on the NIST RT-05 Spring Evaluation source localization task, which is based on real recordings of lectures in noisy and reverberant environments.

Index Terms: speaker localization, head orientation, microphone arrays, global coherence field.

1. Introduction

Since 1990, several Speaker LOCalization (SLOC) techniques have been proposed as reported in [1, 2]. Most of the traditional SLOC techniques are based on the estimation of time differences of wavefront arrival at each sensor and on a consequent application of geometrical information to infer the acoustic source positions. One of the most common techniques for Time Delay Estimation (TDE) is based on Generalized Cross-Correlation Phase Transform (GCC-PHAT) [3, 4]. Other effective SLOC techniques are based on a preliminary computation of an acoustic map, as for instance the Global Coherence Field (GCF) [5] representation, from which the most likely source position is derived through maximization in space.

This paper aims at describing a new SLOC method that was conceived starting from the effectiveness of the Oriented Global Coherence Field (OGCF), introduced in [6], which allows to characterize the orientation of an active speaker's head with a satisfactory accuracy (in terms of angle error) even under reverberant conditions. By exploiting OGCF information, one can also derive more robust speaker position estimates, since they are mostly related to the propagation of a direct wavefront from a given point. On the other hand, previous SLOC techniques did not deal with the way the sound is being radiated from a hypothesized position in space.

The proposed method requires to use a distributed microphone network similar to those available in the laboratories involved in the EC CHIL[1] project. Interestingly, the method does not require a fine tuning to room acoustics, to changes in the room geometries and to the microphone array distribution in space. These facts will be made evident by the experiments described in the remainder of the paper.

Note that in 2005 international benchmarking activities have been started (coordinated by NIST and mostly related to research conducted under AMI and CHIL EC projects) to better characterize the behaviour of SLOC systems in real-world situations. The primary goal of a SLOC system is, in fact, its accuracy in a real noisy and reverberant environment. The experiments and results reported in the following refer to the task of speaker localization in a lecture room, which was investigated in the NIST RT-05 Spring Evaluation [7, 8].

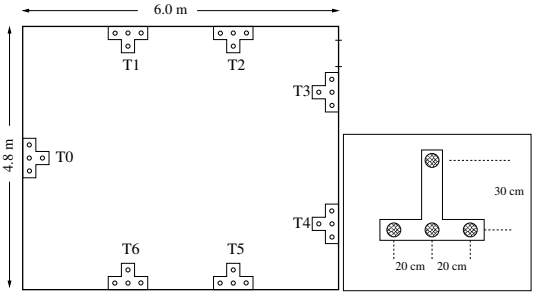

Figure 1: The CHIL room available at ITC-irst and the geometry of a T-shaped microphone array.

2. Distributed Microphone Networks

A goal of the CHIL project is to realize a smart room equipped with acoustic sensors and able to track the position of each active speaker in a noisy and reverberant environment, even under a multi-speaker context. The set-up chosen to face these issues is a distributed microphone network: this corresponds to have a set of microphones distributed all around the room.

Adopting a distributed microphone network, which consists in a set of 4-microphone clusters (see Figure 1), turns out to be very convenient for speaker localization and tracking purposes. In fact, the basic principle is that in most of the cases one or more clusters will receive a direct wavefront coming from any active source. With the given distributed microphone network a good and uniform coverage in space is guaranteed in order to accurately describe any sound event in the room.

Figure 1 shows the map of the ITC-irst CHIL room, where seven T-shaped arrays are placed at the same height (at about 2 meters). The room size is $6m \times 4.8m \times 4.5m$ and is characterized by a reverberation time $T_{60} \simeq 0.7s$. Note that the SLOC

[1]This work was partially supported by the European Commission under the Integrated Project CHIL, contract number 506909. For further details see http://chil.server.de.

experiments described in the following of this paper were conducted on data acquired in the Karlsruhe University CHIL room whose size is $5.9m \times 7.1m \times 3m$ while the reverberation time is $T_{60} \simeq 0.45s$. Figure 1 also shows the geometry of a T-shaped array: this geometry was chosen to determine both azimuth and elevation angles; merging information from different arrays allows to obtain a source localization in terms of (x, y, z) coordinates [8].

3. Global Coherence Field

The GCF is a function defined over all the possible sound source locations inside a given room, and expresses the plausibility that an active sound source is present at specific coordinates. The overall plausibility is obtained by summing partial contributions from a set of microphone pairs distributed in the room. Each contribution is obtained as a measure of the coherence between microphone pair signals realigned according to the time delay that would be observed when a source is really at the considered coordinates. This approach is similar to the steered beamformer locator [1], but the average coherence is here maximized rather than the power of a beamformer. The Coherence Measure (CM) used to calculate the GCF is based on the Crosspower Spectrum Phase (CSP) [4], which corresponds to GCC-PHAT [3].

Denoting with $s_{l_1}(n)$ and $s_{l_2}(n)$ the discrete time signals acquired by microphones l_1 and l_2, the CSP is defined as:

$$C_l(t,d) = DFT^{-1}\left\{\frac{DFT(s_{l_1}(n)) \cdot DFT^*(s_{l_2}(n))}{|DFT(s_{l_1}(n))| \cdot |DFT(s_{l_2}(n))|}\right\} \quad (1)$$

where d denotes the time lag.

In particular, as shown in [9], a CSP-CM function $C_l(t_0, \tau)$, computed for an interval centered at time instant t_0, has a peak at the delay $\tau = \delta_l$ determined by the direction of wavefront arrival, and it has lower values elsewhere.

In the ideal situation a maximum of GCF is obtained as the sum of the CSP peaks only for the actual coordinates of the active source.

3.1. GCF Computation

Let us consider a set of L microphone pairs Ω_l ($l = 0..L-1$) and denote with $\delta_l(S)$ the theoretical delay for the microphone pair Ω_l if the source is at position $S = (x_s, y_s, z_s)$. Once the CM $C_l(t, \delta_l(S))$ has been computed at instant t, for each microphone pair, the GCF is expressed as:

$$GCF(t,S) = \frac{1}{L}\sum_{l=0}^{L-1} C_l(t, \delta_l(S)) \quad (2)$$

Figure 2 shows an example of GCF restricted to a plane (x, y), and represented by means of gray levels, for data acquired in a real room. The brightest spot in the center of the room corresponds to a maximum of GCF identifying the active speaker.

It can be observed that the maximum peak mainly benefits from the CSP contributions of few microphone pairs. Brighter lines (actually hyperbolic curves) depart from them and represent the loci of potential source locations related to the "directional coherence" observed by the microphone pairs. Other less bright curves and areas in Figure 2 account for the effects of reflections and reverberation inside the room.

In general, as a talker is a quite directional source, only a limited number of microphone pairs receive mainly direct wavefronts, whereas for the other ones reflections are prevalent. This observation leads to consider that an analysis of the GCF helps to obtain clues about head orientation. The study of the "shape" of the GCF around a given point brings to introduce the concept of Oriented GCF.

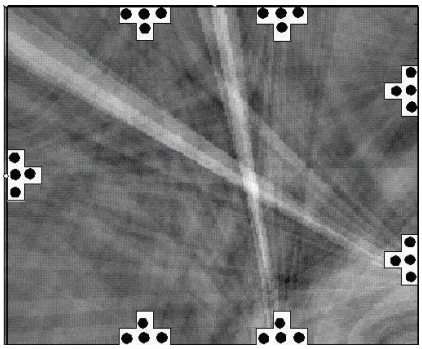

Figure 2: CSP-based 2-dimensional GCF computed in the CHIL room available at ITC-irst. GCF magnitude is represented by gray levels. The brightest spot corresponds to the speaker position.

4. Oriented Global Coherence Field

Assume that the sensor set up consists of L T-shaped arrays distributed in the room and that one microphone pair per array will be used in the following. Then consider the generic potential source location S and orientation O_s chosen from a set of N predefined possible orientations j ($j = 0..N-1$). The Oriented Global Coherence Field is a function of position S and orientation j which represents the plausibility that a talker is at that position and his/her head is aimed according to the considered orientation. The computation of OGCF proceeds from a set of CSP functions as described in the following.

Let us consider a circle C, centered at S and having radius r, and N points P_j on C, which correspond to N possible orientations (see Figure 3). Consider now the intersections Q_l between C and the lines from S to each microphone pair Ω_l.

For a given explored direction j, the set of CSP functions $C_l(t, \delta_l(Q_l))$ is considered, where l identifies each microphone pair Ω_l and point Q_l. OGCF at S is derived as a weighted sum of those values:

$$OGCF_j(t,S) = \sum_{l=0}^{L-1} C_l(t, \delta_l(Q_l)) w(\theta_{lj}) \quad (3)$$

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

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

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 j will emphasize the contributions of CSP in points Q_l closer to P_j (i.e. the direction j) and give less importance to the contributions corresponding to other directions. The orientation j for which $OGCF_j(t,S)$ is maximum is then assumed to indicate the sound source orientation.

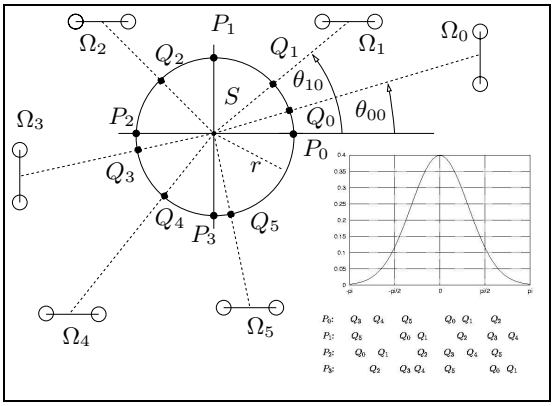

Figure 3: Graphical representation of the orientation estimation scheme described in Sec.4. In this case 6 microphone pairs are available and 4 possible orientations are investigated.

Once defined the number of microphone pairs L, and the number N of possible orientations the described procedure still depends on the values chosen for the parameters r and σ. These may depend on the microphone configuration and on the acoustics of the room and can be optimized empirically. In the SLOC experiments described in the following, we used $r = 5cm$ and $\sigma = 1.7$ and $N = 32$ according to preliminary tuning experiments carried out in the ITC-irst CHIL room [6].

Note that the OGCF can be extended to a 3D spatial domain by considering both elevation and azimuth angles instead of a single orientation angle.

5. OGCF based source localization

Information about talker's head orientation can be advantageously exploited to improve the accuracy of a source localization algorithm. While the GCF based algorithm maximizes the sum of CSP values from all the microphone pairs, uniformly weighted since no information is a priori available about the source orientation, it would be more profitable to give more emphasis to contributions of the microphone pairs receiving direct wavefronts and less emphasis to those collecting mostly reflections.

This is exactly what the OGCF based localization does, given the hypothesized source position for all the possible orientations. Instead of performing two separated steps for estimating source *position* (based on GCF) and *orientation* (based on OGCF), the idea here is to perform directly an analysis of the OGCF extended to all the possible source positions, in order to find the joint optimal solution for both the quantities.

Hence, the proposed localization algorithm consists of two steps:

1. the source position is estimated on a plane (x, y) by maximizing the OGCF function for all possible angles:

$$(\hat{x}, \hat{y}) = \arg \max_{(x,y,j)} = OGCF_j(x, y) \qquad (5)$$

2. given the resulting 2D localization, the z coordinate is derived with a TDE approach using the vertical pair that provides the highest CSP value.

6. Experiments and Results

Localization experiments were carried out to compare the new proposed OGCF based localization algorithm with two reference algorithms previously adopted for evaluation campaigns on speaker localization and tracking. For both the former algorithms some parameters (e.g. analysis window size) had been optimized and the following results refer to the best performance that were obtained.

The first method simply exploits TDE between the signals of two orthogonal microphone pairs and derives the source position by means of triangulation in two steps: first on a (x, y) plane using horizontal microphone pairs and then for the vertical coordinate z by means of vertical pairs. The second method is based on the maximization of the GCF function computed on a (x, y) plane. Then, as for the OGCF approach, the z coordinate is determined in a separate step by means of the most reliable delay estimate. Further details on the two reference methods can be found in [8].

The basic metric to evaluate SLOC methods is the *localization error* that is the euclidean distance between the coordinates delivered by the localization system and the reference coordinates. An error is classified as *fine* if it is lower than $50cm$; otherwise it is classified as *gross*. Given this metric, the comparison between the SLOC algorithms is carried out in terms of [8, 10]:

- Output rate: average number of localizations produced per second;
- False Alarm (FA) rate: percentage of frames for which the algorithm produces a localization output even if nobody is speaking;
- Deletion rate: percentage of frames for which the system does not produce any localization hypothesis even if there is an active speaker;
- Localization rate: percentage of fine errors with respect to all the localization outputs;
- RMSE: overall root mean square error;
- fine RMSE: root mean square error computed only on fine errors;
- Bias: single coordinate average error.

For all the systems a postprocessing was applied in order to select the most reliable frames, based on the amplitude of the peaks of either the CSP, or the GCF or the OGCF functions exceeding predefined thresholds. This step acts as a sort of Speaker Activity Detection (SAD) and has the purpose of properly balancing performance with a trade off between FA rate and Deletion rate.

The real data of the NIST RT-05 Spring Evaluation database for speaker localization was adopted to test the given methods. It includes excerpts from 13 real lectures held at the CHIL room of Karlsruhe University, acquired by means of 4 T-shaped arrays, and manually annotated to obtain speech boundaries as well as reference speaker coordinates.

Table 1 reports on results obtained with the three given algorithms, considering different SAD thresholds for GCF and OGCF.

As first comment, one can notice that the chosen SAD threshold values have a direct effect on performance reported on Table 1. Consider that the two thresholds for GCF or OGCF can not be compared one each other due to different ranges assumed by the two functions.

In practice, when the SAD threshold increases the output rate and the FA rate decrease, which leads to a less reactive but quite robust system. For high values of the SAD threshold, localization

Technique (SAD threshold)	Output Rate [1/s]	FA Rate [%]	Del. Rate [%]	Loc. Rate [%]	RMSE [mm]	fine RMSE [mm]	Bias [mm]
TDE	2.25	42	41	95	309	203	(59,-78,-41)
GCF(0)	6.21	81	7	87	479	226	(43,-64,-77)
GCF(0.38)	1.94	39	48	92	327	198	(40,-47,-51)
GCF(0.75)	0.07	03	96	91	238	159	(80,-22,-57)
OGCF(0.15)	5.09	68	13	95	298	193	(-1,-7,-55)
OGCF(0.20)	3.91	55	23	95	272	193	(-12,-10,-47)
OGCF(0.25)	2.84	44	36	95	266	192	(-23,-10,-41)
OGCF(0.30)	2.01	33	50	95	249	191	(-37,-14,-33)

Table 1: Results obtained applying different localization systems to the NIST RT-05 Spring Evaluation test set.

rate and RMSE are also improved. However, it is worth noting that an RMSE of about $24cm$ is achieved by the GCF method only when a non realistic $0.07/s$ output rate (i.e. one localization every 14 seconds) is obtained by the given SAD threshold of 0.75. As a result the OGCF based method turns out to be the most interesting and best performing one: with the highest SAD threshold, it ensures a RMSE of $25cm$ with an output rate of more than 2 localizations per second; with the lowest SAD threshold (i.e. 0.15), it ensures a RMSE of less than $30cm$ with an output rate of more than 5 localizations per second (i.e. very good real time tracking capabilities).

Finally, one can note that the fine RMSE is close to $19cm$. This is an important result, since it expresses the error observed when gross errors are discarded (sometimes caused by cross-talk effects generated in the audience and not annotated by manual labelers).

7. Conclusions and Future Works

This paper introduced a new SLOC method based on the OGCF information extracted from coherence at microphone pairs of a distributed microphone network. Experimental results show the relevant improvement in terms of accuracy and robustness provided by the given algorithm, when compared to other ones previously investigated.

Note that the system operates in a completely unsupervised manner. It was tuned in a given room and then tested in another room with less microphone pairs and different acoustic characteristics. This fact shows the robustness and portability of the proposed solution.

The two-step algorithm here adopted represents a suboptimal approach. In fact, a direct optimization of GCF in the (x, y, z) space, and of OGCF in the (x, y, z, j) space, although possible, was not adopted because of the high computational requirements, exceeding the limits of potential real-time implementation required by benchmark tests [7]. On the other hand, the current solution is also implemented in real time (see http://shine.itc.it).

Next work is planned to explore different possible directions for improvement. A first step will regard a statistically based optimization of the weighting function, currently defined in a rather intuitive but empirical way.

Secondly, a detailed analysis will be done on the relationship between head orientation and contribution to OGCF provided by each microphone pair. In fact, when primary reflections are dominant over most of the direct wavefronts one can try to exploit them to adapt the system according to the environmental acoustics (instead of focusing only on direct wavefronts).

Finally, the proposed method seems to be very promising to cope with multiple active speaker contexts. Next work will address this issue.

8. References

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

[2] Y. Huang, J. Benesty, "Audio Signal Processing for Next-Generation Multimedia Communication Systems", Kluwer Academic Publishers, Boston, 2004.

[3] C.H. Knapp, C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", IEEE Trans. On ASSP, vol.24, pp. 320-327, 1976.

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

[5] R. DeMori, "Spoken Dialogue with Computers", Chapter 2, Academic Press, 1998.

[6] A. Brutti, M. Omologo, P. Svaizer, "Oriented Global Coherence Field for the Estimation of the Head Orientation in Smart Rooms equipped with distributed microphone Arrays", Proc. Interspeech, Lisboa, 4-8 September 2005.

[7] URL: http://www.nist.gov/speech/tests/rt/rt2005/spring/

[8] M. Omologo, A. Brutti, P. Svaizer, L. Cristoforetti, "Speaker Localization in CHIL lectures: Evaluation Criteria and Results", "MLMI 2005: Revised selected papers", edited by Steve Renals and Samy Bengio, Springer Berlin/Heidelberg, pp. 476-487, 2006.

[9] 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.

[10] T. Nishiura, T. Yamada, S. Nakamura, K. Shikano, "Localization of multiple sound sources based on a CSP analysis with a microphone array", Proc. IEEE ICASSP, Istanbul 2000, vol. 2, pp. 1053-1056.