CLASSIFICATION OF ACOUSTIC MAPS TO DETERMINE SPEAKER POSITION AND ORIENTATION FROM A DISTRIBUTED MICROPHONE NETWORK

Alessio Brutti, Maurizio Omologo, Piergiorgio Svaizer, Christian Zieger

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

ABSTRACT

Acoustic maps created on the basis of the signals acquired by distributed networks of microphones allow to identify position and orientation of an active talker in an enclosure. In adverse situations of high background noise, high reverberation or unavailability of direct paths to the microphones, localization may fail. This paper proposes a novel approach to talker localization and estimation of head orientation based on the classification of Global Coherence Field (GCF) or Oriented GCF maps. Preliminary experiments with data obtained by simulated propagation as well as with data acquired in a real room show that the match with precalculated map models provides a robust behavior in adverse conditions.

Index Terms— Speaker localization, head orientation, microphone arrays, room acoustics, distributed microphone networks.

1. INTRODUCTION

Traditional methods for acoustic source localizations are based either on geometrical derivation of the optimal source position based on the arrival directions estimated by sets of microphones (DOA based methods) or on the maximization of a quantity obtained by steering a microphone array to all the potential source positions (steered-response based techniques). A further possible approach consists in learning the acoustic response of the environment by examples and in trying to classify the observed data according to their similarity to a predefined set of models.

The advantage of this solution is that it does not require an accurate modeling of the acoustic environment as it can robustly deal with adverse phenomena like acoustic reflections and reverberations. The disadvantage is that a training phase is required to create the models. This type of approach was used in [1] where a time delay classification based on histograms is proposed, and more recently in [2], where magnitude and phase of the cross-spectrum calculated from microphone pairs are used as discriminating features.

This work was partially supported by the EU under the Integrated Project CHIL (IP506909) and the STREP Project DICIT (FP6 IST-034624).

In this paper we propose to use more articulated features to model the interaction between the emitting source (active talker) and the surrounding environment, as perceived by a distributed network of microphones (e.g. a set of microphone arrays disseminated on the walls of a room as in the CHIL and DICIT projects [3]). Information about directionality of the acoustic field produced by the speaker at various microphone pairs can be condensed in a "global coherence" map, as the Global Coherence Field (GCF) [4] or SRP-PHAT [5] used to recover the location of the emitting source. Acoustic waves generated in an enclosure by active sources reach the microphones along both direct paths and as reflected and diffracted wavefronts, all contributing to the shape of the resulting GCF map. The peak of the map can generally be directly associated to the source position, but the whole map in its entirety provides additional information on the generated acoustic field. In fact, it can be exploited to derive further clues about the location (and orientation) of the speaker with respect to the microphones as well as about room acoustics.

If the non-omnidirectional directivity of the talker is also accounted for, by properly weighting the information contained in the map of global coherence, a more informative map can be obtained, called Oriented Global Coherence Field (OGCF). Use of OGCF has been shown [6, 7] to provide improved performance with respect to GCF in the task of talker localization. This paper proposes to integrate localization obtained as maximum peak of GCF or OGCF with a classification step considering the whole GCF or OGCF maps. This seems reasonable as even when the peak-based localization fails due to reflections, reverberation, or unavailability of direct paths, the patterns in the maps of global coherence may be associated to specific talker positions and orientations.

2. GLOBAL COHERENCE FIELD (GCF) AND ORIENTED GCF (OGCF)

A Global Coherence Field (GCF) is a function, defined over the space of possible sound source locations, which expresses the plausibility that an active sound source is present at a given point $\mathbf{s}$. The GCF is obtained by summing partial

plausibility contributions from a set of microphone pairs distributed in the room. For each pair the related contribution represents the degree of coherence of the two signals at a time lag corresponding to the interchannel delay observed when a source is in $\mathbf{s}$.

Here the coherence between the discrete time signals $x_{l_1}(n)$ and $x_{l_2}(n)$ acquired by microphone pair l is calculated on intervals $\mathbf{x_{l_1}}$ and $\mathbf{x_{l_2}}$ centered around time instant t by means of the GCC-PHAT [8, 9] as follows:

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

where d denotes the time lag.

If we consider a set of L microphone pairs and indicate with $\delta_l(\mathbf{s})$ the theoretical delay for microphone pair l when the source is at position $\mathbf{s} = (x_s, y_s, z_s)$, the GCF is expressed as:

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

If restricted to a plane (x, y) the GCF at a given instant can be represented as a map or grey-level image, with bright pixels in correspondence of coordinates producing high values of "global coherence" (i.e. high plausibility of source presence). The contribution of each single microphone pair can generally be easily identified as one or more brighter lines (actually hyperbolic curves) departing from the pair and passing through the source and/or the points of reflection of the generated wavefronts. Constructive interference of contributions gives rise, in favorable situations, to a single emerging maximum peak in correspondence of the source (see an example in left part of Figure 1).

In general, as a talker is a quite directional source, only a limited number of microphone pairs receive a prevalence of direct wavefronts, whereas for the other ones energy of reflections is predominant. Besides, direct wavefronts produce higher coherence levels than reflected/reverberated components. These facts can be exploited to obtain clues about the directivity of the source (e.g. head orientation) from a study of the "shape" of the GCF around a given point, leading to the concept of Oriented Global Coherence Field (OGCF). Given L microphone pairs the OGCF maps can be derived for a set of predefined possible orientations φ_j $(j = 0..N-1)$ considering the coherence contributions on L points K_l on a circle around the given point $\mathbf{s}$ (see [6, 7, 10]) according to the formula:

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

where $w(\theta_{lj})$ is a weight computed from a gaussian function, whose purpose is to give more emphasis to contributions along directions close to orientation φ_j.

3. THE PROPOSED APPROACH

During test of a real-time talker localization system based on GCF and implemented in the CHIL room [3] at ITC-irst it was observed that a satisfactory localization performance is obtained, except when the talker is closely facing the walls, the corners, or is speaking curved toward a table or other reflecting surfaces. In these unfavorable cases, the GCF cannot be easily "decoded" by simply detecting the maximum peak and associating it with the source coordinates (see right part of Figure 1). Nevertheless, the GCF map still contains information "encoded" in its particular shape, useful to detect and classify the particular cases. Even if the direct wavefronts do not provide sufficient clues for source localization, the particular patterns of reflected wavefronts may still be enough to uniquely characterize the position of the emitting source. As a direct modeling of the complex patterns of reflections in a real environment is not easily obtained, the classification based on examples seems to be a valid alternative to extract reliable information about talker position and orientation.

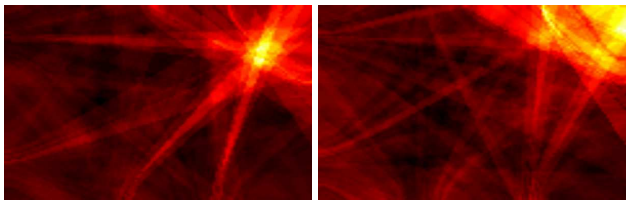

Fig. 1. Two examples of GCF maps with an active speaker on the right upper corner of the room. On the left map the speaker is oriented toward the center of the room and is easily localizable. On the right map the speaker is in the same position but is facing the corner: reflections are predominant.

A "pattern classification" approach is the first and straightforward attempt to exploit the additional information about patterns of direct paths and reflections. To this purpose it is necessary to create a set of models corresponding to various talker positions and orientations. It is then necessary to define a *distance* between the observed data and the stored models.

For the sake of simplicity we will now suppose to restrict the localization task to a 2-dimensional space, i.e. $\mathbf{s} = (x_s, y_s)$, and to drop from notation the dependency from instant t. Let us consider a room in which the set of possible talker positions is identified by $\mathbf{s} \in A$, where A is a spatial sampling of the room (e.g. with resolution $5cm \times 5cm$). Source position (and orientation) estimates can be obtained as

$$(\hat{x},\hat{y}) = \arg\max_{(\mathbf{s}\in A)} GCF(\mathbf{s}) \quad (4)$$

or

$$(\hat{x},\hat{y},\hat{\varphi_j}) = \arg\max_{(\mathbf{s}\in A,j)} OGCF_j(\mathbf{s}) \quad (5)$$

In case the peak of GCF or OGCF map is not clearly identified, we can also compare the obtained maps with pre-

calculated model maps $\mu_{p,q}$, one for each predefined potential position $p \in \{P_0, P_1, ..., P_{M-1}\}$ and orientation $q \in \{Q_0, Q_1, ..., Q_{R-1}\}$ of the speaker. All the models $\mu_{p,q}$ are normalized by mean value subtraction and scaling to unitary energy. The decision can be based on minimizing a similarity measure expressing the difference between the calculated map, after a normalization step, and the stored models. The simplest comparison can be accomplished by means of the L_1 norm $d(p,q)$ taken "pixel by pixel" between the normalized map $\Lambda(\mathbf{s})$ and the models $\mu_{p,q}(\mathbf{s})$:

$$d(p,q) = \sum_{\mathbf{s} \in A} |\Lambda(\mathbf{s}) - \mu_{p,q}(\mathbf{s})| \tag{6}$$

$$(\hat{p}, \hat{q}) = \arg\min_{p,q} d(p,q) \tag{7}$$

As an alternative to the L_1 norm a correlation-based measure $d'(p,q)$ between the map $\Lambda(\mathbf{s})$ and the models can be calculated [11] (and in this case maximized):

$$d'(p,q) = \sum_{\mathbf{s} \in A} [\Lambda(\mathbf{s}) \cdot \mu_{p,q}(\mathbf{s})] \tag{8}$$

or more sophisticated morphological distances can be adopted.

4. EXPERIMENTS AND RESULTS

In order to validate the effectiveness of the proposed approach a set of experiments was carried out using the distributed microphone network available at the ITC-irst CHIL room. The sensor network, consisting of 7 T-shaped microphone clusters, and a map of the room are depicted in Figure 2. It is worth mentioning that the room is characterized by a reverberation time $T_{60} = 0.7s$ which makes the localization task quite hard.

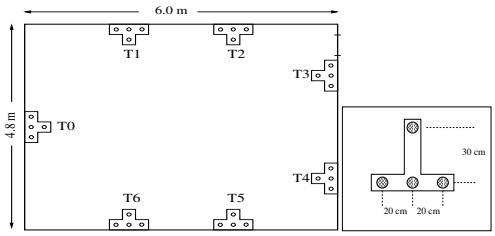

Fig. 2. Outline of microphone placement in the CHIL room available at ITC-irst (left), and geometry of a T-shaped microphone array (right). Each T-shaped array is placed vertically on the walls.

For a preliminary evaluation a database was generated by simulating speech production and propagation in the room. Impulse responses were generated using a modified version of the image method [12] which allows to account for source directivity. A cardioid directivity pattern, roughly resembling

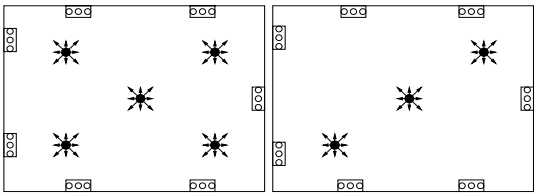

Fig. 3. Positions and orientations of the source in the simulated data (left) and in the real database (right).

the characteristics of a talker, was adopted for the experiments. Close-talk speech segments were then filtered with the precalculated impulse responses in order to produce two different sets of signals: the first one exploited to generate GCF and OGCF based models and the second one to evaluate algorithm performance. Real background noise acquired at each sensor in the room was added to the signals in order to make data more realistic. The overall simulated database includes 4 SNR levels (30, 10, 5 and 0 dB), 5 positions and 8 different orientations for each positions, as depicted in left part of Figure 3.

A second database of real data was also acquired to test the proposed approach in a real scenario. In particular GCF and OGCF based models were created on the basis of data produced using a loudspeaker as source, placed and oriented as shown in right part of Figure 3. Test data were instead produced with a real human talker as source, with an average resulting SNR about 20 dB. It is important to note that position and orientation of the real talker are only nominally the same as those of the loudpeaker used to produce the corresponding models. Exact location of talker's head is not easily determinable, however this fact contributes to assess the feasibility and robustness of the proposed method.

Given a position p and orientation q of the source, the model $\mu_{p,q}$ was computed by processing the whole signals acquired by the sensor network in order to produce a single global coherence map. In the evaluation process, instead, an analysis step of $100ms$ was adopted; localization and orientation estimates were produced only during intervals of source activity (a speech activity detector is used to detect frames containing speech). The analysis was restricted to the (x, y) space and only horizontal microphone pairs were involved in the computations of the maps. Unreliable estimates were detected by comparing the distance between observed data and the models, with a threshold depending on the adopted metrics. If the distances from all the models exceeded the threshold, the corresponding data frame was assigned to the rejection class Π.

Two methods to compute the global coherence maps (and the corresponding models) were compared:

- **GCF**: in this case $\Lambda(\mathbf{s}) = GCF(\mathbf{s})$ and the maps are calculated on the test data according to eq. 2, and compared with the $M \cdot R$ GCF models $\mu^{GCF}_{p,q}(\mathbf{s})$.

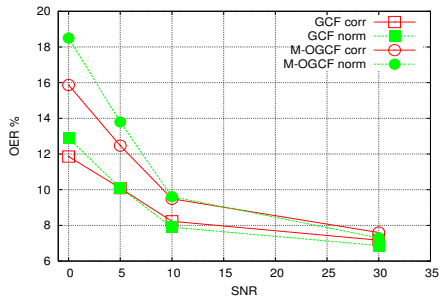

Fig. 4. OER obtained with simulated speakers as a function of the SNR. GCF and M-OGCF methods are compared using both L_1 norm (*norm.*) and correlation (*corr.*).

- **M-OGCF**: in this case $\Lambda(\mathbf{s}) = \max_j[OGCF_j(\mathbf{s})]$, where $OGCF_j(\mathbf{s})$ is calculated according to eq. 3, and the maps are compared with a set of models $\mu_{p,q}^{OGCF}(\mathbf{s})$

For the two methods both the L_1 norm and the correlation between observed maps and models (see equations 6 and 8) were applied.

Performance was measured in terms of Localization Error Rate (LER) and Orientation Error Rate (OER) as percentages of errors over the number of estimates.

With simulated data the thresholding process was not performed and LER ranged from a minimum of 4.3% at 30 dB to a maximum of 7.6% at 0 dB, with no significant difference between the two methods and the two distance measures. Performance in terms of OER is reported in Figure 4. It is worth mentioning that most of the orientation errors reported in the figure are within the contiguous angles ($\pm 45°$), the worst case being at 0 dB SNR with 90% of estimates within this tolerance.

Table 1 summarizes results obtained with real-talker data. It can be noted that even in the case of models acquired with a different source (loudspeaker), and a limited accuracy in the position/orientation of the real talker, performance is still satisfactory in terms of LER. A limited drop in OER is mainly due to classification into the directions adjacent to the correct one. These results were obtained with thresholds on the distance measures empirically set to values such to guarantee that a maximum of 25% of speech frames were rejected by the classification procedure.

Method	LER	OER	E$\leq 45°$
GCF norm.	0.37%	9.0%	99.5%
GCF corr.	0%	9.3%	100%
M-OGCF norm.	0.16%	11.8%	99.5%
M-OGCF corr.	0%	11.9%	99.8%

Table 1. Results obtained on the real-talker data.

A further experiment on real data and models obtained from simulated impulse responses did not provide encouraging results as the image method would require a more accurate geometrical model of the environment.

5. CONCLUSIONS

Results of the preliminary experiments reported in this paper show that the proposed classification method offers reliable information that could be advantageously integrated in GCF- and, in particular, in OGCF-based localization systems to resolve critical cases of unfavorable position of the talker with respect to the microphones.

Further study is needed to achieve more discriminant distance measures, to determine the number of models necessary in realistic scenarios and to assess the robustness of the models in case of deviations with respect to predefined positions and orientations.

6. REFERENCES

[1] N. Strobel and R. Rabenstein, "Classification of time delay estimates for robust speaker localization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, 1999, vol. 6.

[2] P. Smaragdis and P. Boufounos, "Learning source trajectories using wrapped-phase hidden markov models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2005.

[3] *http://chil.server.de; http://dicit.itc.it*.

[4] R. DeMori, *Spoken Dialogue with Computers*, Academic Press, London, 1998, chapter 2.

[5] M. Brandstein and D. Ward, *Microphone Arrays*, Springer Verlag, 2001.

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

[7] A. Brutti, M. Omologo, and P. Svaizer, "Speaker Localization based on Oriented Global Coherence Field," in *Interspeech*, Pittsburgh, PA, USA, September 2006, pp. 2606–2609.

[8] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 24, no. 4, 1976.

[9] M. Omologo and P. Svaizer, "Acoustic event localization using a cross-power spectrum based techniques," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, 1994, vol. 2, pp. 273–276.

[10] A. Brutti, M. Omologo, and P. Svaizer, "Estimation of talker's head orientation based on Oriented Global Coherence Field," in *120th Convention Audio Engineering Society*, Paris, France, May 2006.

[11] R. Brunelli and S. Messelodi, "Robust estimation of correlation with applications to computer vision," in *Pattern Recognition*, 1995, vol. 28, pp. 833–841.

[12] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, pp. 943–950, April 1979.