LOCALIZATION OF MULTIPLE SPEAKERS BASED ON A TWO STEP ACOUSTIC MAP ANALYSIS

Alessio Brutti, Maurizio Omologo, Piergiorgio Svaizer

Fondazione Bruno Kessler (FBK)-irst
Via Sommarive 18, 38050 Povo, Trento, Italy
{brutti|omologo|svaizer}@fbk.eu

ABSTRACT

An interface for distant-talking control of home devices requires the possibility of identifying the positions of multiple users. Acoustic maps, based either on Global Coherence Field (GCF) or Oriented Global Coherence Field (OGCF), have already been exploited successfully to determine position and head orientation of a single speaker. This paper proposes a new method using acoustic maps to deal with the case of two simultaneous speakers. The method is based on a two step analysis of a coherence map: first the dominant speaker is localized; then the map is modified by compensating for the effects due to the first speaker and the position of the second speaker is detected. Simulations were carried out to show how an appropriate analysis of OGCF and GCF maps allows one to localize both speakers. Experiments proved the effectiveness of the proposed solution in a linear microphone array set up.

Index Terms— microphone array, speaker localization, multiple speakers, global coherence field.

1. INTRODUCTION

During the last two decades many efforts were devoted to investigate Speaker LOCalization (SLOC) technologies [1]. Besides early applications in audio-videoconferencing, more recently several new application areas have been addressed, which include the use of microphone networks for "ambient intelligence". The long-term goal is the capability of monitoring humans in a real noisy and reverberant environment, without any constraint on the number or the distribution of microphones in the space or on the number of speakers active at the same time. Ambient intelligence is realized through the widespread use of sensors (e.g., cameras, microphones) connected to computers that are unobtrusive to their human users.

Along this line, the most recent activities conducted on SLOC under the CHIL project (see for further details http://chil.server.de) showed that acoustic maps, for instance derived from the Oriented Global Coherence Field (OGCF) information [2, 3], allow one to estimate both the position and the head orientation of a single active speaker, even in the most critical and challenging situations.

This paper addresses a scenario including multiple simultaneous speakers [4], which is being investigated under DICIT (Distant-talking Interfaces for Control of Interactive TV), a European Project started in October 2006. The focus of the project is a user-friendly interface that allows access to a virtual smart assistant enabling the interaction with TV-related digital devices and infotainment services. In the given scenario, the user can speak in a natural and comfortable way, not encumbered by any hand-held or head-mounted microphone. It was observed that in the given scenario users tend to pronounce very short sentences, corresponding to single commands to the system. Among the most challenging issues involved in this scenario are the effects of distance between users and microphones as well as the need of implementing Acoustic Echo Cancellation (AEC) to compensate the sound produced by loudspeakers. As a final target, multiple speaker localization will have to perform accurately even when AEC is jointly applied to microphone signals. Moreover it has to be very fast in detecting a new position of the user in space.

In this paper the problem of localizing multiple speakers is dealt with by extending the analysis of Global Coherence Field (GCF) and OGCF acoustic maps, which have proved to effectively tackle the single speaker localization problem. The following section details how acoustic maps are calculated. Section 3 describes the proposed two step analysis. In section 4 experiments and results are reported. Finally, conclusions and future work discussion close the paper.

2. GLOBAL COHERENCE MAPS

Acoustic source localization based on Time Difference Of Arrival (TDOA) and triangulation is computationally efficient but often not robust enough in adverse acoustic situations, characterized by reverberation, reflections and occlusions of the direct path between source and microphones. Spatial maps, in the form of GCF [3, 5] (also known as SRP-PHAT [1]) and OGCF [2], are very effective representations for the given target. Both GCF and OGCF are composed by exploiting not only the maximum peak of generalized cross-correlation [6, 7], but the whole GCC-PHAT based coherence measure at any time lag. Although more computationally expensive, it was shown that they provide reliable results. In particular, when even the maximization of the "global coherence" fails, a suitable analysis and classification of the spatial map yields useful information to localize a speaker and determine his/her head orientation [8].

2.1. GCF and OGCF

Given a set of L microphone pairs and $\delta_l(\mathbf{s})$ as the theoretical time difference of arrival for microphone pair l when the source is at position $\mathbf{s}$, the GCF at time instant t is expressed as [3]:

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

where $C_l(t,\tau)$ is the GCC-PHAT, a function of the time lag τ (with $|\tau| \leq \tau_{max}$). The maximum valid delay τ_{max} is determined by the

This work was partially supported by the EU under the STREP Project DICIT (FP6 IST-034624). Further details can be found at http://dicit.fbk.eu.

microphone distance, the sampling rate and the propagation speed of sound. If we restrict the analysis to a plane and sample the space of potential source positions defining a grid of points, GCF at a given time instant can be represented as a surface, which exhibits peaks in correspondence of coordinates producing high values of "global coherence" (i.e., high plausibility of source presence). Peaks arise due to the constructive addition of the contributions of multiple microphone pairs receiving direct-path propagation of acoustic wavefronts. Removing the dependence of GCF on time instant t (for a sake of simplicity), the estimation of the speaker position $\hat{\mathbf{s}}$ can hence be derived from maximizing the function over all potential source positions $\mathbf{s}$:

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{\operatorname{argmax}}\, \mathrm{GCF}(\mathbf{s}) \tag{2}$$

Since human are quite directional sources, head orientation affects the shape of the surface around a maximum peak in GCF. The analysis of GCF shape around the peak leads to the concept of Oriented Global Coherence Field (OGCF) [2]. OGCF has proved to effectively address the SLOC problem in a distributed multi-microphone scenario. 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 [2]) according to the formula:

$$\mathrm{OGCF}(t, \mathbf{s}, \varphi_j) = \sum_{l=0}^{L-1} C_l(t, \delta_l(K_l)) \cdot w(\theta_{lj}) \tag{3}$$

where $w(\theta_{lj})$ is a weight computed from a gaussian-like function, whose purpose is to give more emphasis to contributions along directions close to orientation φ_j. Given an OGCF map, the position of an active source is derived by maximizing the map over all possible coordinates and orientations $(\mathbf{s}, \varphi_j)$:

$$(\hat{\mathbf{s}}, \hat{\varphi_{\mathbf{j}}}) = \underset{(\mathbf{s}, \varphi_j)}{\operatorname{argmax}}\, \mathrm{OGCF}(\mathbf{s}, \varphi_j) \tag{4}$$

In principle, OGCF is conceived to operate with a distributed microphone network covering all source orientations. However it can be also applied to a linear array setup, even if its potentialities are not fully exploited because of the lack of a complete angular balance [1].

3. LOCALIZATION OF MULTIPLE SPEAKERS

The mentioned above approaches have been widely adopted to tackle the SLOC problem when limited to a single source. When two, or more, sources are simultaneously active, it was observed that most of the time the coherence map presents two, or more, evident peaks in correspondence of the sources. However, searching for two local maxima may fail in the given context. First of all, some spurious peaks may be generated at points that combine secondary GCC-PHAT peaks related to both sources. As a second aspect, one must take into account that sharpness and magnitude of GCC-PHAT peaks, and consequently of peaks in the coherence maps, are strongly related to the spectral content of the emitted sounds. Wideband sounds, such as fricative, generate sharp peaks, while a source producing vowels generates less defined peaks of GCC-PHAT. As a consequence, in the presence of two speakers, depending on the spectral contents of the involved speech signals, the main peak jumps from one source to the other while the second peak may be considerably lower than the main one and may be overtaken by spurious peaks. Figure 1 shows an example of a map when two sources are active. In this case the two sources, denoted by the circles, are on the left and on the right of a linear microphone array that is placed in the upper part of the picture. It can be observed that most of the coherence concentrates around the speaker on the right, while the peak on the left is quite smooth.

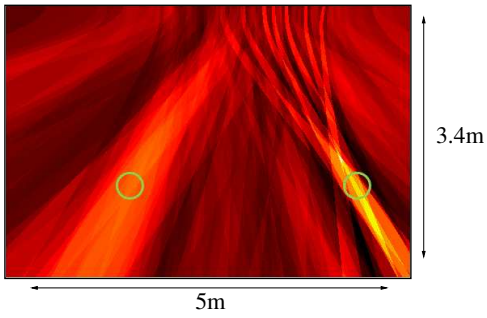

Fig. 1. Example of OGCF map in presence of two sources. Bright colors represent high value, while dark colors identify low values. Notice how the presence of the main peak tends to compress the contrast of the remaining part of the map.

In order to deal with these problems we suggest an approach that attempts to de-emphasize the main peak in order to make the detection of the second one easier. Our proposed method works as follows:

- search for the main peak in the coherence map (e.g., OGCF);
- modify each single GCC-PHAT by lowering those values at time lags that generated the main peak;
- compute a new map and maximize it.

The core of our method is the GCC-PHAT de-emphasis performed in step 2 and described in the following section.

3.1. GCC-PHAT de-emphasis

Let us consider a coherence map $\mathrm{OGCF}(\mathbf{s}, \varphi_j)$ and let us assume that $\mathbf{s}_1$ is the point that maximizes the map. Now let us consider the microphone pair l and its corresponding GCC-PHAT function $C_l(\tau)$. Given the theoretical time delay $\delta_l(\mathbf{s}_1)$, we can compute a modified version of GCC-PHAT, named $C'_l(\tau)$, that de-emphasizes those time lags that gave rise to the peak in $\mathbf{s}_1$:

$$C'_l(\tau) = \phi(\tau, \delta_l(\mathbf{s}_1)) \cdot C_l(\tau) \tag{5}$$

where $\phi(\cdot)$ is a de-emphasis function defined as follows:

$$\phi(\tau, \mu) = \frac{1}{2\alpha}\left[\frac{1}{b} - \frac{1}{b} e^{-\frac{|\tau-\mu|}{b}}\right] \tag{6}$$

Note that b is a parameter that determines the sharpness of the notch, while α is a normalization factor that guarantees that:

$$\sum_{\tau=-\tau_{max}}^{\tau_{max}} C'_l(\tau) = \sum_{\tau=-\tau_{max}}^{\tau_{max}} \phi(\tau, \delta_l(\mathbf{s}_1)) \cdot C_l(\tau) = 1 \tag{7}$$

As shown in Figure 2, in practice a null is set around the lag that corresponds to the direction of the dominant speaker. Small values

[1] For further details on the potential of OGCF for real-time speaker localization purposes one can find a video clip at http://shine.fbk.eu.

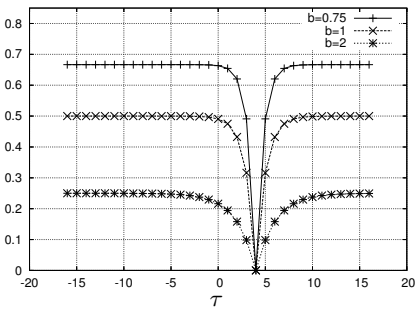

Fig. 2. Example of function ϕ for three different values of b when $\mu = 4$ and $\alpha = 1$.

of b generate very selective de-emphasis functions. Given the new set of de-emphasized coherence measures a new map is computed and maximized. Figure 3 shows the same situation as in Figure 1 after the GCC-PHAT de-emphasis process. It is worth noting that the de-emphasis process not only highlights the second source but also increases the relative level of background amplitude in the map due to reverberation. However, in general the method is very effective in enhancing the peak related to the second sound source position.

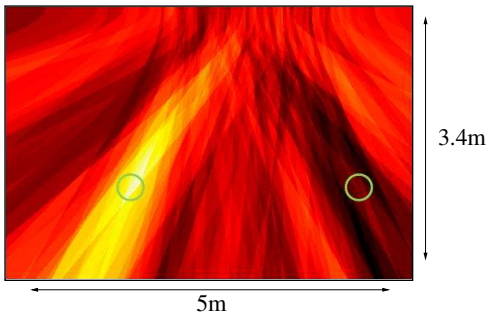

Fig. 3. The picture shows the effect of de-emphasis on the OGCF map of Figure 1. Notice how the peak corresponding to the second source has gained evidence and assumed the clear role of primary source.

The following section describes an experimental activity conducted by applying the resulting algorithm.

4. EXPERIMENTS AND RESULTS

The application scenario, as envisioned in the DICIT project, is outlined in figure 4: up to four persons are sitting in front of a linear array of microphones and producing utterances.

The entire sensor setup employed in the DICIT project includes 13 microphones arranged in a harmonic fashion (with an overall distance between the first and the last one of 192 cm) plus two microphones placed 20 cm above the two extremities in order to derive clues for 3D speaker localization. In the following experiments we exploit only a subset composed of 7 sensors with a uniform distance of 32 cm (see Figure 5).

A simulated data collection is generated to evaluate the proposed approach in the DICIT setup. The image method [9], modified to account for source directivity, is employed to generate the impulse responses at each microphone when a speaker is sitting at one of

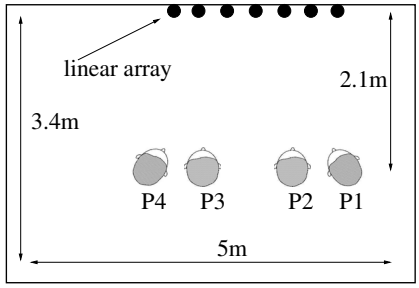

Fig. 4. Layout of the experimental setup. 4 positions were investigated at 2.1 m distance from a linear microphone array.

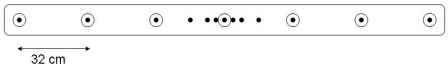

Fig. 5. Configuration of the harmonic linear array. The circled sensors form a uniform array of 7 elements.

the 4 locations, at 2.1 m from the microphones as reported in Figure 4. An orientation of $\pm 45^{\circ}$ is assumed for sources in positions P1 and P4. A cardioid polar pattern, which roughly models the human radiation pattern, is assumed as the directivity of the talker in a horizontal plane [10]. Four phonetically rich sentences, with lengths of about 6 s, are filtered with the impulse responses in order to simulate the wave propagation through the room. Different sentences are exploited to investigate the effects of different spectral contents. The simulated room has a size of 5 m$\times$3.4 m$\times$3 m and the reverberation time is 0.7 s. The sampling rate is set to 16 kHz and the spatial resolution is 1 cm. In the experiments we assume that speakers are always simultaneous in pairs. The search for the peaks is performed over all the room and the knowledge about the four predefined positions is not exploited.

Given the linear microphone layout under investigation, it makes sense to analyze the localization performance in terms of azimuth estimation error with respect to the center of the array. Figure 6 shows an example of localization outcome.

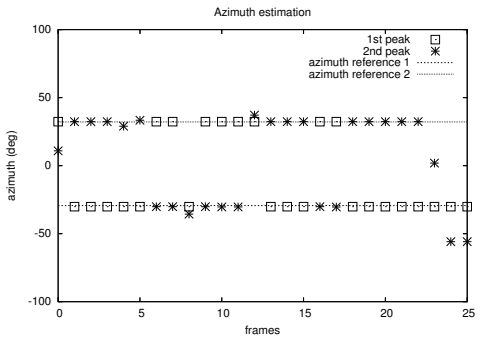

Fig. 6. Example of azimuth estimations. Squares refer to the main peak, asterisks identify the estimation obtained after the de-emphasis. Notice that the main peak moves from one source to the other and the errors are always due to the second peak.

Besides the reliability of the estimation, it is worth noting how the main peak jumps from one source to the other. A more detailed analysis shows that it depends on the particular spectral contents of the involved signals. However, the target is to have two accurate localizations, no matter how the speakers are being associated.

The proposed algorithm was compared to a traditional search for the two highest peaks, both in the GCF and in the OGCF case. The search is only constrained to guarantee that the two peaks are at least 50 cm apart.

In order to evaluate the proposed approach, we define two different metrics:

- localization rate: percentage of estimations whose distance from the actual source position is lower than 50 cm;
- azimuth rms error: root mean square error of the azimuth estimation with respect to the center of the array.

Both metrics are applied, on a frame by frame basis, to the first and the second resulting peaks. The step size characterizing the analysis rate is set to 200 ms. When calculating the distance of an estimate from the true speaker position, each estimate is always associated with the closest active speaker.

Experiments are conducted by investigating all the possible cases derived using 4 sentences and 6 pairs of speakers chosen from the 4 predefined positions. As above mentioned, the techniques based on the proposed analysis (denoted as OGCF-d, GCF-d) are compared to a traditional search for the two highest maxima in the coherence map (OGCF, GCF).

	GCF	OGCF	GCF-d	OGCF-d
loc rate 1st peak	96.5%	84.3%	98.7%	99%
loc rate 2nd peak	19.3%	30.9%	73%	73.7%
rms 1st peak	8.8°	17.9°	4.5°	3.8°
rms 2nd peak	35.5°	31.4°	23.6°	18.4°

Table 1. Overall results for all possible combinations of 4 different speech sequences and 4 different positions.

As reported in Table 1 a traditional search either on GCF or on OGCF does not provide satisfactory results. In particular note that the higher sensitivity of OGCF is due to the fact that the second speaker often interfered through spurious peaks with the accurate and selective angle-dependent analysis provided by OGCF. As main achievement of this work, Table1 shows that the de-emphasis process allows one to considerably increase the localization rate of the second peak, for instance from 19.3% to 73% in the GCF case. Moreover, OGCF-d produces more accurate results than the GCF-d as evidenced by both the rms measures. As final remark, notice that the main peak can almost always ($\sim 99\%$) correctly identify one of the sources with a quite small azimuth rms error (3.8°), while the second source is generally localized with a less accurate but anyway satisfactory performance ($\sim 73\%$ loc. rate, 18.4° azimuth rms error). The gain in the 1st peak performance is due to the association process that in some cases assigns both estimates to the same source. When the de-emphasis is not applied, it is more frequent that the second peak is close to the first one resulting in reduced performance.

5. CONCLUSIONS AND FUTURE WORK

An acoustic map based approach to address the problem of localizing multiple speakers has been presented in this paper. The proposed method operates in two steps and attempts to highlight the weaker source by masking the main peak in the initial acoustic map. The algorithm has proved to provide satisfactory results in the given scenario. Experiments are restricted to the case of two simultaneous speakers. In case the number of speakers is not known, the amplitude of the secondary peak may be exploited to check whether a second source is possibly active or not. Potentially, the same solution can be extended to the case of three or more simultaneous speakers, even if the discriminative power of acoustic maps decreases as the number of sources increases. Preliminary experiments on data collected with human speakers in a real scenario confirm the capability of the proposed approach to deal with two simultaneous speakers.

Future work will address the adaptation of the parameter b in (6) according to the selectivity required to spatially discriminate different positions. In a real-time application, this selectivity can eventually be adapted according to previously estimated speaker positions.

Furthermore a robust criterion to detect overlapping speakers that need joint localization will be investigated by analyzing the time persistence of primary and secondary peaks in the acoustic maps.

Finally, it is worth noting that a real-time localization system based on the described maps is memoryless. In other words, it can provide frame by frame reliable speaker positions that do not depend on previous position estimates. This is a very important feature, necessary in a scenario as that addressed under DICIT, which is characterized by multiple simultaneous speakers, often uttering short commands or background speech irrelevant to the application.

6. REFERENCES

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

[2] 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.

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

[4] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. on Speech and audio processing*, vol. 12, no. 5, pp. 520–528, September 2004.

[5] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event localization," Tech. Rep. 9303-13, ITC-irst Istituto per la Ricerca Scientifica e Tecnologica, 1993.

[6] 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.

[7] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on SAP*, vol. 5, no. 3, pp. 288–292, May 1997.

[8] A. Brutti, M. Omologo, and P. Svaizer, "Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network," in *Proc. of IEEE ICASSP*, Honolulu, Hawaii, USA, 2007.

[9] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustic Society of America*, vol. 65, no. 4, pp. 943–950, April 1979.

[10] H. Kuttruff, *Room Acoustics*, Elsevier Applied Science, 1991.