SOURCE LOCALIZATION IN REVERBERANT ENVIRONMENTS BY CONSISTENT PEAK SELECTION

Raffaele Parisi, Albenzio Cirillo, Massimo Panella and Aurelio Uncini

INFOCOM Dpt., University of Rome "La Sapienza"
via Eudossiana 18, 00184 Rome, Italy

ABSTRACT

Acoustic source localization in the presence of reverberation is a difficult task. Conventional approaches, based on time delay estimation performed by generalized cross correlation (GCC) on a set of microphone pairs, followed by geometric triangulation, are often unsatisfactory. Prefiltering is usually adopted to reduce the spurious peaks due to reflections.

In this work an alternative strategy is proposed, based on the concept that secondary peaks of the GCCs can be crucial in order to correctly locate the source. More specifically, an iterative weighting procedure is introduced, based on the rationale that peaks corresponding to the actual source position should be *consistently* weighted. The position estimate is then refined by use of an effective and fast clustering technique. Experimental results on simulated data demonstrate the effectiveness of the proposed solution.

Index Terms— Source localization, reverberation, microphone arrays.

1. INTRODUCTION

Localization of acoustic sources by microphone arrays is an important task in many applications of practical interest. Typical examples can be found in videoconferencing, multimedia, surveillance, hand-free talking systems. A popular class of localization algorithms is based on a two-step strategy. In the first step a set of relative time differences of arrival (TDOA) between pairs of microphone signals are estimated. Usually the well-known generalized crosscorrelation (GCC) is employed [1]. In the second step the source position is obtained by exploiting the estimated TDOAs according to some specified strategy (e.g. by geometrical triangulation).

Unfortunately, in the presence of even moderate reverberation levels, traditional GCC-based approaches are seriously hampered, due to the presence of spurious peaks [2]. In this case, in order to mitigate the effects of spurious peaks, some approaches have been proposed in literature (e.g. [3, 4]). In this paper a different approach is followed. The Linear Intersection (LI) method [5] is applied also to selected secondary peaks of the GCCs in order to locate potential source positions. More specifically, an iterative weighting procedure is introduced, based on the idea that peaks corresponding to the actual source position should be weighted in a consistent way.

In the following, after a brief summary of the background, the proposed approach and some preliminary results on simulated data are described.

This work was partially funded by the Italian "Ministero dell'Istruzione, dell'Università e della Ricerca".

2. BACKGROUND

2.1. Signal model

Signals received by a pair of microphones are modelled as

$$\begin{aligned} x_1(t) &= s(t) * h_1(t) + n_1(t) \\ x_2(t) &= s(t) * h_2(t) + n_2(t), \end{aligned} \tag{1}$$

where $s(t)$ is the source signal, $h_i(t)$ $(i = 1, 2)$ is the room impulse response between the source and the i-th microphone and $n_i(t)$ is uncorrelated noise, neglected in this work. Source localization requires preliminary estimation of the TDOA between the direct paths from the source to the microphones of all available microphone pairs, based on model (1).

2.2. TDOA estimation

The TDOA is usually estimated by the generalized crosscorrelation function (GCC) [1]

$$R^{(g)}_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \Psi_g(f) G_{x_1x_2}(f) e^{j2\pi f\tau} df, \tag{2}$$

where $G_{x_1x_2}(f)$ is the cross power spectrum of $x_1(t)$ and $x_2(t)$ and $\Psi_g(f)$ is a proper weighting function used to mitigate the effects of reverberation. The Phase Transform function (PHAT) [1] is very popular

$$\Psi_g^{PHAT}(f) = \frac{1}{|G_{x_1x_2}(f)|}. \tag{3}$$

The TDOA D is then estimated as

$$\hat{D} = arg \underset{\tau}{max}\, R^{(g)}_{x_1x_2}(\tau). \tag{4}$$

2.3. Source localization

For each received signal frame, use of (4) for each microphone pair makes it available a set of TDOAs, which can be employed to estimate the source position. Several approaches are possible and a thorough review can be found in [6].

In this work the Linear Intersection method [5] was adopted. LI is a closed-form localization method that demonstrated particularly robust and accurate. It requires the microphones to be arranged in quadruples, where sensors of each quadruple lie on the midpoints of a rectangle. Two TDOAs per quadruple are estimated by considering the main peaks of the GCCs of the two pairs of received signals. Since each TDOA estimate approximately determines a cone in 3D (under the assumption of far-field source), the intersection of the two cones can be computed in closed form yielding a single *bearing line* for each quadruple. The final source position is estimated by properly weighting the points of "closest intersection" between all pairs of lines [5].

The LI approach was shown to perform well in the presence of moderate reverberation. However, when reverberation increases, the main peaks of the GCC may likely not correspond to the TDOAs of the direct paths, so leading to gross localization errors. In order to overcome this limitation, a new approach is proposed in this paper.

3. PROPOSED APPROACH

In the presence of reverberation the performance of GCC-based localization approaches rapidly degrades with the reverberation time [2], due to the presence of significant spurious peaks in the GCC, i.e. peaks not corresponding to the delay between the direct paths. Several solutions to reduce the spurious peaks have been devised in the past, including cepstral and common pole prefiltering [4]. In this paper a different strategy is pursued, based on a new criterion for the selection of *consistent* peaks of the GCC. Specifically, the LI paradigm is adopted, but selection of less significant peaks is also allowed, according to the following strategy.

3.1. Multi-Peak Linear Intersection

The first step of the proposed approach is the selection of most significant peaks of each GCC. Assume that Q quadruples are available. For each quadruple, two GCCs are computed and for each GCC at most K peaks are retained according to the following two constraints:

- only peaks exceeding a prespecified threshold are selected. The threshold is determined by considering the mean value of the GCC in an interval around the peak.
- Peaks corresponding to non-admissible delays are discarded. Admissible delays must lie in the interval $[-t_D^{max}, t_D^{max}]$, where $t_D^{max} = d/c$, being d the inter-microphone distance and c the speed of sound.

After peak selection, LI is applied to all possible combinations of selected peaks at each quadruple, so that *a set* of at most K^2 bearing lines (or rays) is generated, having the center of the quadruple as a common origin. Fig. 1 illustrates a typical case, where points of closest intersection are also shown; only interior points are considered. For each set of rays, at most one ray steers at the actual source position, while the others are generated by spurious peaks due to reflections. A criterion is needed in order to select the set of lines (one per quadruple) that *most consistently* steer at the actual source position.

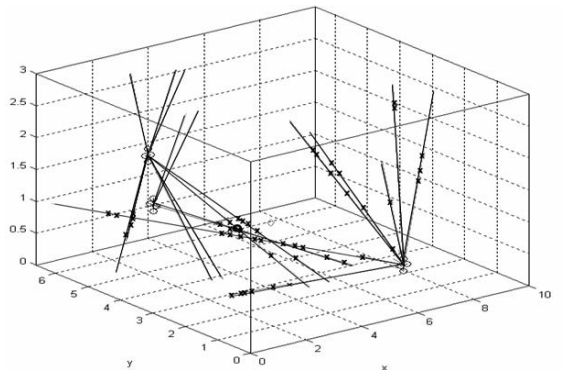

Fig. 1. Application of LI to multiple peaks.

3.2. Optimal line selection (OLS)

Consider a single set of lines (one line per quadruple). The corresponding points of closest intersection (or potential source locations) can be used to estimate a possible source position. In particular, similarly to [5], the point $\mathbf{s}_{jk}$ of closest intersection between the j-th and k-th lines is given the following weight

$$w_{jk} = \sum_{q=1}^{Q} P\left(\tau(\{\mathbf{m}_1^{(q)}, \mathbf{m}_2^{(q)}\}, \mathbf{s}_{jk}), \tau_{12}^{(q)}, \sigma^2\right) \cdot \cdot P\left(\tau(\{\mathbf{m}_3^{(q)}, \mathbf{m}_4^{(q)}\}, \mathbf{s}_{jk}), \tau_{34}^{(q)}, \sigma^2\right). \quad (5)$$

In (5) Q is the number of quadruples, $P(x, m, \sigma^2)$ is a normal distribution of mean m and variance σ^2, evaluated at x, $\mathbf{m}_i^{(q)}$ $(i = 1, \ldots, 4)$ is the position of the i-th microphone of the q-th quadruple, τ is the time delay corresponding to the potential location $\mathbf{s}_{jk}$ and $\tau_{lm}^{(q)}$ ($lm = 12$ or $lm = 34$) is the estimated time delay, corresponding to the selected peak of the GCC. Figure 2 visualizes the application of (5) to a generic quadruple. The closer a point is to all lines, the higher its weight. The selected set of lines is given a weight equal to the sum of the weights of all corresponding potential sources. This weight is a measure of the consistency of the selected lines in terms of minimum distance among the potential source solutions.

Finally the set of lines with the highest weight is selected to locate the source. In particular, the location of the acoustic source is estimated as a weighted sum of the corresponding points according to

$$\hat{s} = \frac{\sum_{j=1}^{Q} \sum_{k=1,k\neq j}^{Q} w_{jk} s_{jk}}{\sum_{j=1}^{Q} \sum_{k=1,k\neq j}^{Q} w_{jk}}. \quad (6)$$

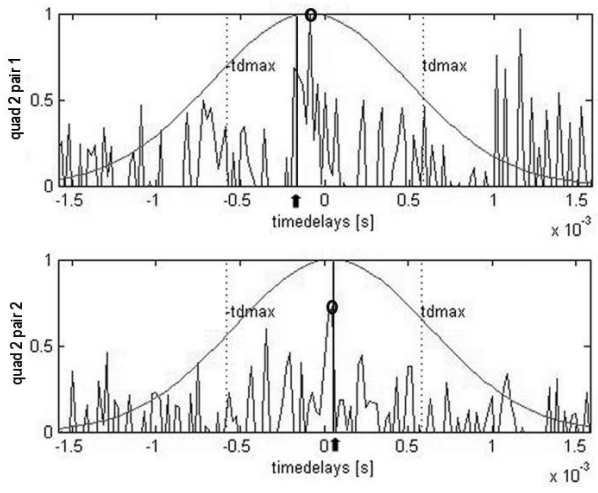

Fig. 2. GCCs of a generic quadruple. A normal distribution is placed on each selected peak. Arrows indicate the time delays corresponding to the potential source position.

3.3. Outlier elimination by clustering

In the presence of high reverberation levels some of the selected lines may still significantly deviate from the others, thus behaving like *outliers* of the location estimate. The source position estimate can be improved by removing in (6) contributions of all points lying on the outlier line. These points can be efficiently determined by a proper clustering technique.

Many clustering techniques are available in literature. The most convenient solution should be chosen based on different aspects, such as accuracy of cluster geometric models, real-time and on-line implementation requirements, computational complexity, and so on. As a good compromise for the problem at hand the *Unsupervised Splitting Hierarchical Expectation-Maximization* (USHEM) clustering algorithm [7] was adopted in this work. The USHEM algorithm is particularly suited for applications where a high degree of automation is required.

More specifically, in USHEM a mixture of Gaussian models is determined via the well-known EM algorithm. Each model is associated with a cluster and the number of clusters is automatically determined by a constructive approach. The centroid of each cluster is computed as the weighted mean of its points, using as weights the probabilities that each point be generated by the corresponding Gaussian distribution. Points are assigned to the cluster scoring the highest probability and their weights determine the final weight of the cluster. Finally, the heaviest cluster is selected and its centroid is chosen as the source position estimate.

4. EXPERIMENTAL RESULTS

The proposed algorithm was tested on synthetic data generated by the image method [8]. A room of size $10 \times 6.6 \times 3$ m was modelled, at different reverberation levels. An acoustic source, radiating white noise, was placed in $(3, 3, 1)$. Sampling frequency was 44100 Hz. Three quadruples of microphones were placed in the room, their centers being set in $(5.5, 0, 0.8)$, $(4, 10, 1.3)$ and $(0, 3, 1.7)$ respectively. The inter-microphone distance was $d = 20$ cm. 1024-point FFTs were applied to compute the GCC for each pair of microphones. Up to $K = 4$ peaks from each normalized GCC were selected according to a specified threshold, empirically determined. Parabolic interpolation was used to get intra sample resolution.

LI was applied to all combinations of TDOAs of the two microphone pairs of each quadruple. Then all points at minimum distance between pairs of skew lines were weighted according to (5) for all possible combination of lines (one per quadruple), to determine the best set of lines according to the OLS strategy. Of course the complexity of this step was limited for the specific values of K considered. Higher values of K would require a proper suboptimal solution, currently under investigation. The final estimate of the source location was obtained by (6) and successively refined by clustering.

The three approaches (LI, OLS and OLS with clustering) were compared for different reverberation times T_R (from 0 to 0.65 s), in terms of mean error and standard deviation of the location estimate (average on 100 realizations). Figures 3,4,5 and 6 show the results for $K = 3$ and $K = 4$ on a logarithmic scale. The improvement w.r.t. the straight LI method is clear for $T_R > 0.4s$.

5. CONCLUSION

A new approach to source localization in the presence of reverberation, based on a novel consistency measure of selected peaks of the GCCs, was presented. Improvement w.r.t. conventional solutions was demonstrated on simulated white noise data. Further studies will investigate the application to speech data in real environments.

6. REFERENCES

[1] C. H. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE*

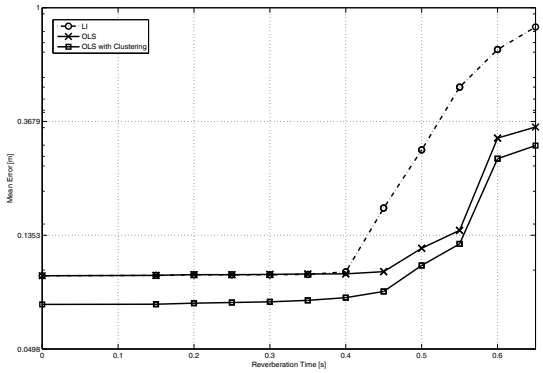

Fig. 3. Mean error vs. T_R, $K = 3$ peaks. LI (circles), OLS (crosses), OLS with clustering (squares)

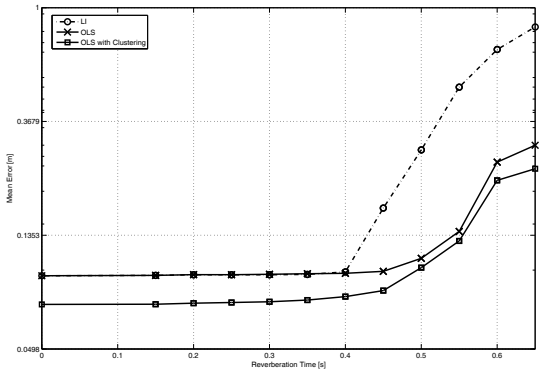

Fig. 5. Mean error vs. T_R, $K = 4$ peaks. LI (circles), OLS (crosses), OLS with clustering (squares)

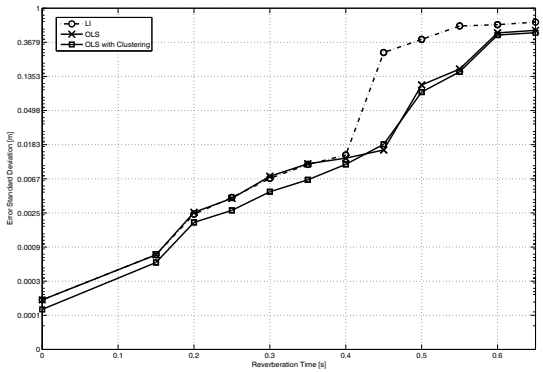

Fig. 4. Error standard deviation vs. T_R, $K = 3$ peaks. LI (circles), OLS (crosses), OLS with clustering (squares)

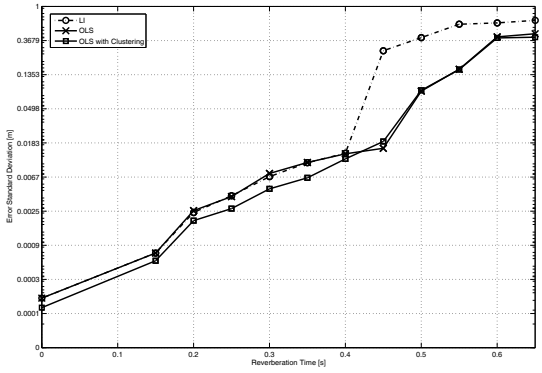

Fig. 6. Error standard deviation vs. T_R, $K = 4$ peaks. LI (circles), OLS (crosses), OLS with clustering (squares)

Trans. on Acoust., Speech and Signal Processing, vol. 24, no. 4, Aug. 1976.

[2] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, Mar. 1996.

[3] A. Stephenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, pp. 253–266, 1997.

[4] R. Parisi, R. Gazzetta, and E. D. Di Claudio, "Prefiltering approaches for time delay estimation in reverberant environments," in *IEEE 2002 International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[5] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form estimator for use with room environment microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, January 1997.

[6] M.S Brandstein and H.F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, no. 2, pp. 91–126, April 1997.

[7] M. Panella, A. Rizzi, F.M. Frattale Mascioli, and G. Martinelli, "A constructive EM approach to density estimation for learning," in *Proc. of the IEEE Int. Joint Conf. on Neural Networks (IJCNN'2001)*, 2001, vol. 4, pp. 2608–2613.

[8] J. B. Allen and A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.