A NEW CONSISTENCY MEASURE FOR LOCALIZATION OF SOUND SOURCES IN THE PRESENCE OF REVERBERATION

Albenzio Cirillo, Raffaele Parisi and Aurelio Uncini

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

ABSTRACT

Sound source localization is an important task in many practical applications. Among existing methods, the Linear Intersection algorithm is a popular technique and performs well in noisy conditions without requiring high computational costs. Unfortunately the performance of Linear Intersection is seriously hampered by the presence of reverberation, which is quite common in closed environments. In order to overcome this limitation, the Optimal Line Selection technique was proposed in the past. In this paper some important modifications of this technique are presented and described. Results on simulated and real data confirm the quality of the proposed solution.

Index Terms— Time delay estimation, Acoustic source localization, Microphone arrays.

1. INTRODUCTION

Localization of acoustic events is an important topic in audio processing, because of the importance of position information in complex systems like speaker diarization, camera steering or audio focusing. Many applications are set up in closed environments, which means that reverberation (usually represented by the reverberation time T_{60}) [1] should be taken into account during processing. Source localization can be performed by beamforming [2] or by methods requiring a prior estimate of the Time Difference Of Arrival (TDOA) between signals acquired by microphone pairs. TDOA can be retrieved either by use of the popular Generalized Cross Correlation (GCC) algorithm [3] or by adaptive algorithms like the Adaptive Eigenvalue Decomposition Algorithm (AEDA) [4]. All these methods are based on a system model that is valid only in very low reverberant conditions and get worsening performance when T_{60} increases.

Linear Intersection (LI) [5] is a popular technique for source localization based on TDOA estimation. Its behaviour is very similar to the maximum likelihood estimator (MLE) [5] in case of noisy conditions. The time delay related to the main peak of the GCC is usually adopted as TDOA estimate. However, in reverberant conditions, strong reflections can originate peaks of correlation that may be stronger than the one due to the direct path, thus leading to an error in the estimate. Nevertheless the GCC preserves information on direct sound origin thanks to secondary peaks of the correlation signal between microphone pairs. This fact can be exploited to design new effective localization techniques.

The Optimal Line Selection (OLS) algorithm [6] is an extension of LI and it takes into account the secondary peaks of cross-correlation. Despite its higher computational complexity, OLS improves the process performance in terms of precision of the estimate in highly reverberating environments.

In this paper a proper modification of the OLS algorithm is introduced and described. Localization of the acoustic event is performed only when information gathered by all the microphone pairs is consistent. Results on simulated and real data are presented to demonstrate the validity of the proposed solution.

2. BACKGROUND

2.1. System 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 gaussian noise. TDOA is the time difference between the direct paths in the two impulse responses. If the sound source is far enough from the microphone pair, the sound wave can be considered as a plane wave. In this case it is possible to get the Direction of Arrival (DOA)[1] from the TDOA τ as

$$\vartheta = \arccos\left(\frac{c \cdot \tau}{d}\right), \tag{2}$$

where c is the speed of sound and d is the distance between the two microphones.

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

[1] The Direction of Arrival is the angle between the direction line and the line connecting the two sensors.

978-1-4244-3298-1/09/$25.00 ©2009 IEEE

2.2. Generalized cross correlation

GCC is defined as [3]

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \Psi(f) G_{x_1x_2}(f) e^{j2\pi f\tau} df, \tag{3}$$

where $G_{x_1x_2}(f)$ is the cross power spectrum of $x_1(t)$ and $x_2(t)$ and $\Psi(f)$ is a proper weighting function, used to mitigate the effects of reverberation. In particular the Phase Transform function (PHAT) [3] is often used. PHAT is based on preservation of the phase of the two signals through the weighting function

$$\Psi(f) = \frac{1}{|G_{x_1x_2}(f)|}, \tag{4}$$

The time delay estimate (TDE) between the two signals is finally

$$\hat{\tau} = arg\,\underset{\tau}{max}\, R_{x_1x_2}(\tau). \tag{5}$$

Unfortunately the GCC technique is not robust for moderate to high reverberation times ($T_{60} > 0.3s$). This behavior has been investigated in [7], showing that reflections due to reverberant environments may originate correlation peaks greater than the one corresponding to the true TDOA.

2.3. Linear Intersection

The LI algorithm is based on the cone approximation of the locus of points that generates a given time delay for a specified couple of sensors. The theoretical hyperboloid is approximated with a cone in case of sufficient distance between the source and the microphone pair [8]. This cone has a semiaperture angle equal to the DOA of the sound wave, its axis is coincident with the pair axis and its vertex is the pair mid point. As a consequence, if two pairs are placed orthogonally with the same mid point (so forming a *quadruple*), two cones with the same vertex will figure out the locus of points where the source lies. As a matter of fact, the intersection of these cones is a line whose cosine directions with respect to the three-dimensional coordinate system can be easily retrieved according to

$$\cos(\alpha)^2 + \cos(\beta)^2 + \cos(\gamma)^2 = 1. \tag{6}$$

where α and β coincide with the measured semiaperture angles of the two cones. This line is aiming at the sound source. If more quadruples are used, all lines should ideally intersect in the source position.

As the line direction depends on TDOA estimates, usually lines are skew because of quantization errors and measurement errors of sensor positions. Hence a pair of points at minimum distance s_{ij} and s_{ji} are calculated for each couple of lines (i, j) [5]. The final set of points is weighted according to

$$w_{ij} \;=\; \prod_{q=1}^{Q} P\left(T(\{\mathbf{m}_1^{(q)}, \mathbf{m}_2^{(q)}\}, \mathbf{s}_{ij}), \hat{\tau}_q, \sigma^2\right), \tag{7}$$

where Q is the number of sensor pairs, $P(x, m, \sigma^2)$ is a normal distribution of mean m and variance σ^2, evaluated at x, $\mathbf{m}_i^{(q)}$ $(i = 1, 2)$ is the position of the i-th microphone of the q-th pair, T is the time delay corresponding to the potential location $\mathbf{s}_{ij}$ and $\hat{\tau}_q$ is the TDE with respect to the q-th pair. The final source position is estimated by the weighted sum

$$\hat{s} = \frac{\sum_{i,j} w_{ij} s_{ij}}{\sum_{i,j} w_{ij}}. \tag{8}$$

LI has become very popular as a localization strategy. Its main advantages consist of light computation and robustness to noise [5].

3. OLS ALGORITHM AND CONSISTENCY CHECK

Ideally lines generated by each microphone quadruple should point at the same region of the space. This behaviour inspired the OLS algorithm [6], which considers all possible combinations among the most k significant peaks of the GCC. In this way a set of lines belonging to a single quadruple are generated, including the line that correctly aims at the source. The combination of k peaks for each orthogonal pair of a quadruple can originate up to k^2 lines. The algorithm originally was set to find the combination of lines, one per quadruple, whose points at minimum distance would produce the heaviest weight according to formula (7). Results obtained in this way [6] were good with respect to LI but still suffered from a performance degradation with increasing T_{60}. To tackle this problem, some modifications to the original algorithm have been introduced.

As a first consideration, it is useless to consider a couple of points s_{ij} and s_{ji} at minimum distance between the i-th and j-th line if

$$\|s_{ij} - s_{ji}\| > Th, \tag{9}$$

where Th is a proper threshold. In fact, if (9) is verified, lines i and j do not point at the same area of the room. This is a very helpful control because not all the quadruples often workout correctly (for example quadruples standing in the back of a directive source) and it is adviceable not to consider them when generating the set of lines likely aiming at the source. In this way bad quadruples are simply discarded.

As a consequence of this first check, there could remain a set composed by points produced by only two lines. Even if this could be a good set for the estimation, relying on an estimate consistent only for two quadruples is not a robust choice. So sets of points composed by less than three points should be discarded to ensure that at least three lines compose the set. This control can be changed in a tougher condition if the number of minimum points rises, so as to constraint the simultaneous consistency of a higher number of quadruples.

A second consideration is related to the measure of compactness of each set of lines. Differently from [6], compactness C

is now evaluated by referring to the variance of the distance of the set L of points with respect to the mid point b_L of the set

$$C_L = \frac{1}{N_L} \sum_{(i,j)\in L} (\|s_{ij} - b_L\|)^2, \qquad (10)$$

In this equation N_L is the total number of points belonging to the L set and (i,j) are the combination of the lines who passed the check in (9). Even sets of lines whose C_L is bigger than an imposed threshold should be discarded from the estimation process as this could mean that lines composing those sets are surely diverging. The most compact set of points among all the ones who passed the checks (namely the one that minimizes the C_L value) is chosen as the set to be used for the final estimate. Position estimate is then obtained by adopting formula (7) and (8) applied to the chosen set.
Imposition of constraints in the search for the correct estimate is the main difference with OLS. In the next part of the paper results demonstrating the validity of this consistency check will be shown. It should be noted that if a frame does not satisfy the consistency check, it is discarded and the localization is missed. Considering this aspect, it is very useful to consider the probability to miss localization versus reverberation time, while varying the number k of peaks considered. At meantime, the probability should be related to the corresponding error measurement that will be presented in terms of Mean Square Error (MSE).

4. EXPERIMENTAL RESULTS

The localization algorithm was tested both on simulated and on real data. Main results are presented in the following.

4.1. Simulated data

OLS was tested on two different simulated data set, namely on white noise and on speech signal. Different reverberant conditions were considered by computing the impulse responses with the Image Method [9] and adopting model (1). The reverberation time was varied in between 0 and 2 seconds. Synthetic data were produced at $f_S = 48000Hz$ in a simulated environment with dimensions $10 \times 6.6 \times 3$ [m], placing six quadruples on the four peripheral walls. Constraints chosen for the OLS algorithm were set at $Th = 0.25m$, a minimum number of 3 points for each set and a maximum variance $C_{MAX} = 0.0625m^2$. A number of 100 frames were examined in both cases. The values of the MSE versus reverberation time were represented in figures 1 and 2. LI performances were compared with OLS by adopting maximum number of GCC peaks $k = 1, 2, 3$. Figures 3 and 4 show the probability to obtain a missed localization with OLS in case of noise and speech signal. In figure 1 the MSE is not evaluated for $k = 1$ at high T_{60} ($T_{60} > 1s$) because it is not possible to achieve consistency by considering only the most significant peak. By observing the results, it can be claimed that even for high T_{60} localization obtained with OLS keeps maintaining a low MSE value despite of the increasing probability to have frames not suitable for localization. However, increasing the k parameter, this probability can be reduced. Comparing the different figures, it is clear the advantage brought by using OLS with multiple peaks, even if the algorithm complexity rises.

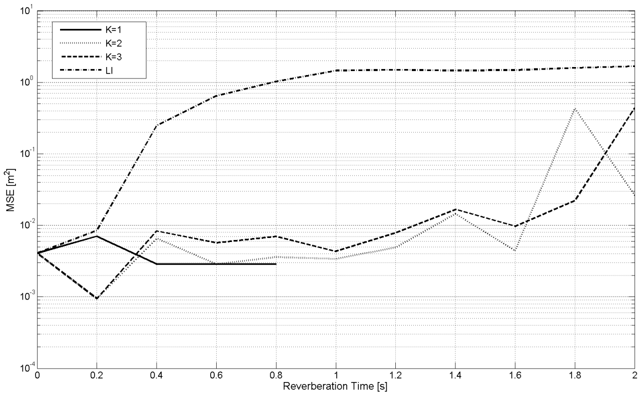

Fig. 1. MSE vs T_{60} for white noise signal.

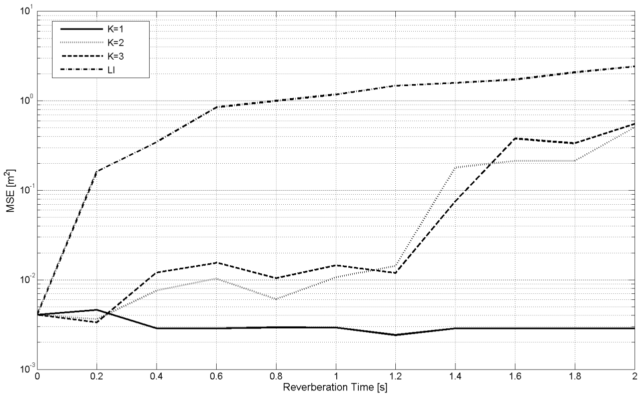

Fig. 2. MSE vs T_{60} for speech signal.

4.2. Real Data

Experiments were carried out also in the ISPAC lab at the INFOCOM dept., which is a room with size $6.31 \times 4.66 \times 2.9$ [m], with two windows and with a T_{60} of about 0.4s. Data were recorded on four quadruples each one placed on a different wall of the room. A loudspeaker was placed in four different points in four different temporal intervals lasting 17s. In each location the back of the loudspeaker was turned toward the closest quadruple. White noise signal was reproduced and recorded at $f_S = 48000Hz$. Real data, composed by 200 frames of 4096 samples for each position, were elaborated with both the LI algorithm and the OLS algorithm with the settings $k = 2$ and $Th = 0.3m$. At least 6 points must be considered to form a set, which means at least three lines.

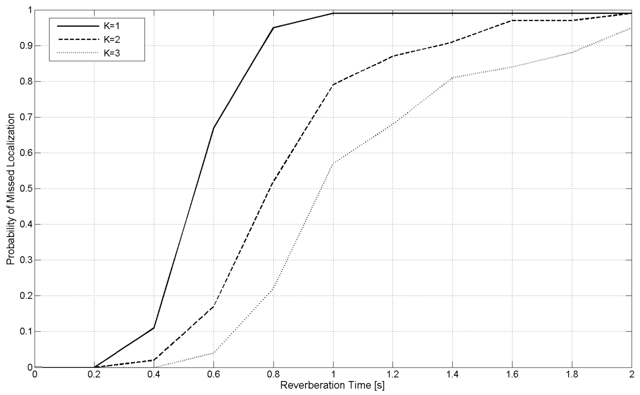

Fig. 3. Probability of missed localization vs T_{60} for white noise signal.

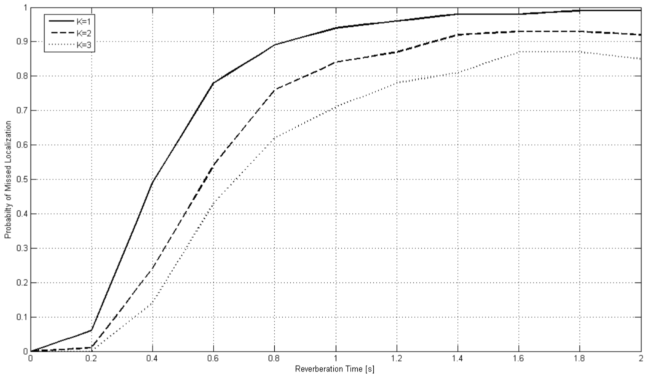

Fig. 4. Probability of missed localization vs T_{60} for speech signal.

$C_{MAX} = 0.16m^2$ was finally chosen. A scattering plot of the estimates is shown in figure 5 for LI and figure 6 for OLS. In the OLS figure the percentage of frames that passed the consistency check is indicated for each position. As a final result it is clear that OLS produced a more robust localization with respect to LI. For each position the variance of the distance of all the estimates from their midpoint was evaluated. LI and OLS are compared in table 1.

var [m^2]	Position 1	Position 2	Position 3	Position 4
LI	0.0015413	0.27518	0.65018	0.0039985
OLS	0.031342	0.05188	0.33886	0.031174

Table 1. Variance of the estimated distance for each loudspeaker position.

5. CONCLUSIONS AND FUTURE WORKS

In this paper it has been demonstrated that OLS brings a clear improvement in terms of accuracy and robustness in sound localization with respect to LI, despite of the possibility to miss localization in some frames. Future work will be devoted to the problem of multi-source localization in the presence of reverberation.

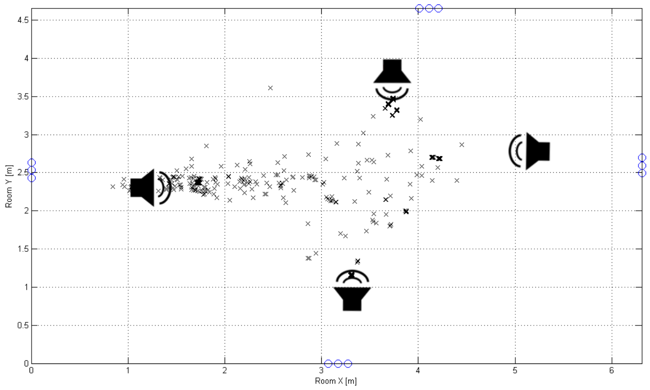

Fig. 5. Scattering plot of the LI estimates, circles represent microphone positions.

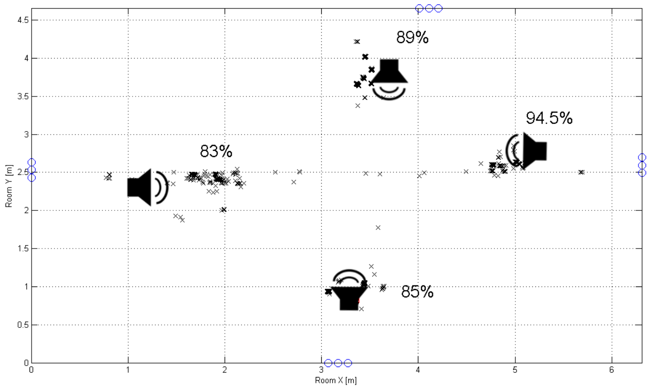

Fig. 6. Scattering plot of the OLS estimates, circles represent microphone positions.

6. REFERENCES

[1] Heinrich Kuttruff, *Room Acoustics*, Taylor & Francis, 4th edition, 2000.

[2] S. Haykin, *Array Signal Processing*, Prentice-Hall, Inc.,, 1984.

[3] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, Aug 1976.

[4] Jacob Benesty Yiteng (Arden) Huang, Ed., *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Kluwer Academic, 2004.

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

[6] R. Parisi, A. Cirillo, M. Panella, and A. Uncini, “Source localization in reverberant environments by consistent peak selection,” in *Proc. of the IEEE ICASSP 2007, Honolulu, Hawaii*, 2007.

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

[8] M. S. Brandstein, *A Framework for Speech Source Localization Using Sensor Arrays*, Ph.D. thesis, Providence, RI: Brown Univ., May 1995.

[9] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of the Acoustical Society of America*, vol. vol. 65, pp. pp. 943–950, 1979.