SOUND MAPPING IN REVERBERANT ROOMS BY A ROBUST DIRECT METHOD

Albenzio Cirillo, Raffaele Parisi, Aurelio Uncini

INFOCOM Dpt., Universita' "Sapienza" di Roma,
via Eudossiana 18, 00184 Roma, Italy

ABSTRACT

Direct methods estimate the position of an acoustic source by sampling the environment through a set of properly placed microphones. SRP-Phat is probably the most popular direct method. It is based on computation of the generalized cross-correlation (GCC) of signals on a grid of preselected points. Anyway, in the presence of reverberation, the functional employed by SRP-Phat can be very irregular from point to point, thus making the source localization a difficult task. In this paper, a new functional is presented that regularizes the SRP-Phat approach and makes it more efficient the use of optimization algorithms to further refine the source position estimation. After a brief introduction, the proposed approach is described and compared to SRP-Phat on simulated and real data at different reverberation levels.

Index Terms— Source localization, reverberation, microphone arrays.

1. INTRODUCTION

Several approaches to estimate a sound source position are available. They can be grouped into two different categories: indirect and direct methods. Indirect methods require that a direction of arrival (DOA) of a sound beam is obtained by microphone recorded signals. Location is estimated in a second step by use of triangulation or different optimization strategies. The time difference of arrival (TDOA) is commonly calculated with a Generalized Cross Correlation (GCC) [1][2] on pairwise signals. In this case, the task of localization becomes more difficult with the increasing of the reverberation time (T_{60})[3] because of the presence of multi-path effect, that rises the difficulty of selecting the correct estimate of DOA on each single microphone pair. To overcome this limit, often a microphone network is adopted, hence redundant information are captured to reinforce the estimator.

In alternative, direct methods (e.g. SRP-Phat, [4]) perform a single step localization. This technique is still based on a distributed microphone network but, this time, each point in the environment is a variable of a proper functional built on the observation of signal cross-correlations. Hence it requires the discretization of the space in an array of points and the consequent weighting of each point according to the corresponding functional. Even if its computational cost requires a careful control, this approach is often adopted. It allows to obtain a visualization of the acoustic field by the representation of a scaled image of the plane containing the estimated position of the acoustic source.

In this paper, after a short description of the background and the SRP-Phat technique, a new direct approach to sound mapping is presented based on a preliminary estimation of time-delays between pairwise microphones. The proposed solution is tested on simulated and real data and compared to the SPR-Phat solution.

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

2. BACKGROUND

2.1. Direction of arrival estimation

The model usually adopted for signals captured by a pair of microphones is

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

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, usually negligible because of the high Signal to Noise Ratios (SNRs) values available. Source localization requires preliminary estimation of DOA. This information is achievable by the Time Difference Of Arrival (TDOA) D between the direct paths from the source to the microphones of all available sensor pairs, based on model (1). D can be obtained by GCC

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

In this equation, $G_{x_1x_2}(f)$ 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

$$\Psi^{PHAT}_g(f) = \frac{1}{|G_{x_1x_2}(f)|}, \tag{3}$$

as it normalizes the cross-spectrum magnitude in order to rely only on phase changes to estimate the cross-correlation. Finally the TDOA D is estimated as

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

The DOA is usually expressed as the angle θ between the line passing through the pair of microphones and the direction of the sound beam

$$\theta = \arccos(\frac{cD}{d}), \quad (5)$$

where c is the speed of sound propagation and d is the distance in between microphones in each pair.

2.2. Linear Intersection

One of the most popular indirect methods, *Linear Intersection* (LI) [5], models the estimated TDOA by a Gaussian probability distribution. In this way each pair of sensors is able to evaluate a probability function for each point of the room. Sensors are arranged in quadruples and the source position is estimated as a weighted sum of points of minimum distance between the bearing lines representing the various DOAs. The point at minimum distance s_{jk} between the $j-th$ and the $k-th$ skew lines is attributed the weight

$$\begin{aligned} w_{jk} &= \sum_{q=1}^{Q} P\left(\tau(\{\mathbf{m}_1^{(q)}, \mathbf{m}_2^{(q)}\}, \mathbf{s}_{jk}), \tau_{12}^{(q)}, \sigma^2\right) \cdot \\ &\cdot P\left(\tau(\{\mathbf{m}_3^{(q)}, \mathbf{m}_4^{(q)}\}, \mathbf{s}_{jk}), \tau_{34}^{(q)}, \sigma^2\right), \end{aligned} \quad (6)$$

where $P(x, m, \sigma^2)$ is a normal distribution of mean m and variance σ^2, evaluated at x, Q is the number of quadruples of microphones, each one composed by microphones positioned at $m_i^{(q)}$ $(i = 1, \ldots, 4)$, while τ_{12} and τ_{34} are the TDOAs estimated by the two pairs considered in each quadruple. The source position is obtained as

$$\hat{s} = \frac{\sum_{j=1}^{Q}\sum_{k=1,k\neq j}^{Q} w_{jk}\mathbf{s}_{jk}}{\sum_{j=1}^{Q}\sum_{k=1,k\neq j}^{Q} w_{jk}}. \quad (7)$$

This technique is important because it shows that it is actually possible to build a likelihood function based on the main peak of a GCC-Phat. This concept can be extended to a multichannel case obtaining good results in terms of estimation error. However the good performance of the algorithm is guaranteed only in low reverberation conditions.

2.3. SRP-Phat

Among the direct methods, the Steered Response Power (SRP) is very popular [4]. The point with the highest likelihood value is chosen as an estimate of the position. This function $F_{SRP}(s)$, being s a generic position in the room, is obtained by computing the GCC in each point of the space:

$$F_{SRP}(s) \equiv F_{SRP}(\tau_s) = \int_{-\infty}^{\infty} \Psi_g(f) G_{x_1x_2}(f) e^{j2\pi f\tau_s} df, \quad (8)$$

where τ_s is the geometrically calculated TDOA related to point s and to a single pair of microphones. Equation (8) can be easily extended to the case of M pairs of sensors

$$F_{SRP}(s) = \sum_{i=1}^{M} \int_{-\infty}^{\infty} \Psi_g(f) G_{x_1x_2}(f) e^{j2\pi f\tau_s^{(i)}} df, \quad (9)$$

where $\tau_s^{(i)}$ is the TDOA related to point s and to the $i-th$ pair. SRP has become a reference in this class of algorithms and will be used in the following as a term of comparison.

3. SMOOTHED LIKELIHOOD FUNCTION (SLF)

GCC-based algorithms are limited by the effects of high reverberation, hence the presence of spurious peaks in the GCC [6]. Optimal Line Selection (OLS) algorithm [7] is based on a criterion that selects secondary peaks in order to improve the estimation of the position. According to TDOA estimation methods and to LI, a Gaussian function can be centered on each main peak of the GCC in the time domain. This operation can be repeated for the k most significant peaks of the GCC of a pair, scaling each Gaussian curve with the peak value, thus creating a smoothed likelihood function (SLF) that describes the envelope of the correspondent k Gaussian distribution:

$$F(s) = arg \underset{j}{max}(V_j \cdot P(\tau_s, \tau_j, \sigma)) \quad (10)$$

where $j = 1, \ldots, k$, is the index related to the k peaks of GCC, $P(x, \mu, \sigma)$ is a Normal distribution and V_j is the value of the j-th peak of the GCC.

This function can be easily extended to a multiple pair system,

$$F(s) = \prod_{i=1}^{N_p} arg \underset{j}{max}(V_j^{(i)} \cdot P(\tau_s^{(i)}, \tau_j^{(i)}, \sigma)), \quad (11)$$

where N_p is the number of used pairs.

The standard deviation σ depends on the accuracy of the used TDOA estimator, but it is reasonable to assume a value not bigger than the maximum allowable TDOA $t_{max} = d/c$,. Since main peaks should not hide secondary peaks, a good rule is to set σ as a fraction of t_{max}. This is a logical choice due to the fact that the probability of estimate the correct position does not rely only on a single peak but on multiple peaks, each one contributing with a smaller variance.

It is possible to estimate (with a quantization error) the position of the acoustic source as the point with the highest weight:

$$\hat{s} = arg \underset{s_q}{max} F(s_q), \quad (12)$$

where s_q belongs to the set of points of the grid.

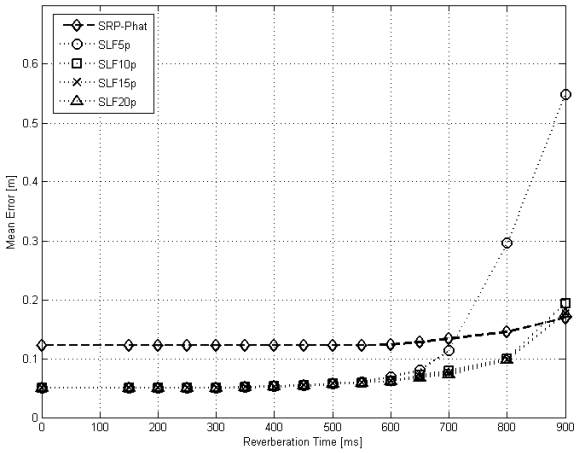
4. EXPERIMENTAL RESULTS

4.1. Synthetic Data

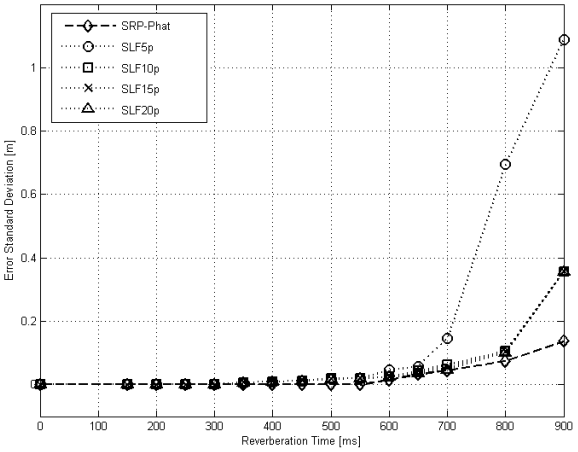
The image method [8] was used to calculate Room Impulse Responses in a synthetic large hall room, with the following dimensions ($x \times y \times z$): $10m \times 6.6m \times 3m$. 6 couples have been used to capture the signal.

To test the theoretical aspect of the likelihood function, a source has been placed in $P \equiv (3m, 3m, 1m)$, emitting a random omni-directional white noise signal sampled at $f_S = 44100Hz$. For each microphone, 100 frames of 1024 samples each have been considered separately to locate the source adopting the proposed algorithm. Peaks of GCC have been selected according to a threshold equal to the 60% of the power of the whole GCC sequence, in order to avoid noisy peaks. For low reverberation it is common that less than k peaks are actually selected. The value of $\sigma = t_{max}/100$ has been chosen empirically, observing that larger values would have hidden the information brought by secondary peaks.

A number of 100 trials has been produced for each of the investigated reverberation times. The comparison of mean errors (figure 1(a)) shows that there is a general improvement on the precision of the estimator with respect to the SRP-Phat approach. What it is expected is that for low reverberation times, the consideration of few peaks ($k = 5$) produces more accurate results, while performance rapidly degradate with T_{60} increasing. It is expected that higher values of k become more effective at higher reverberation levels ($> 1s$). Actually higher number of peaks improve the performance in this case, obtaining results very similar to SRP-Phat. Considering the first $k = 5$ peaks, the precision of the estimator is better until $0.7s$ and then it rapidly decreases. Increasing the number of peaks, there is a substantial improvement. The tests were stopped when, adopting $k = 20$, no further advantages were obtained respect to the previous case of $k = 15$.



(a)



(b)

Fig. 1. (a) mean error and (b) error standard deviation versus reverberation time in the case of noise signal.

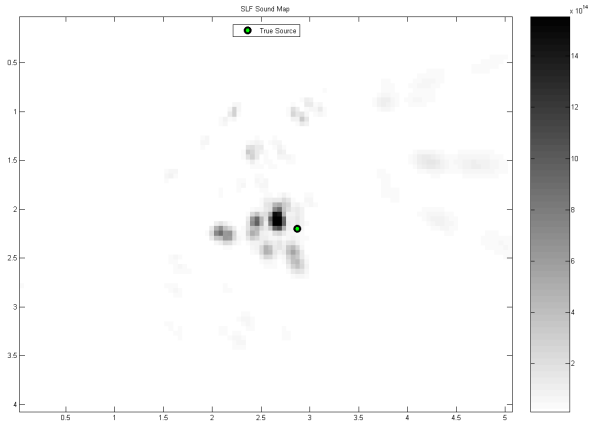
4.2. Real Data

The algorithm was also tested in a real environment. Speech was recorded in the *ISPAC Lab* at INFOCOM Dpt, that is a laboratory room with several noise sources due to computer machines and external environment and large reflecting windows. Hence it is quite a stressing test for the algorithm because of the quantity of disturbing factors. The T_{60} was estimated to be about $0.3s$. Two quadruple of microphones were adopted for the recording and they were located only on two walls of the room. The talker's mouth was established as origin of the sound. SRP-Phat and SLF were applied on 100 frames of the same length (1024 samples for a signal sampled at $44.1Khz$) by considering 15 peaks. Frames were completely different from each other, so to obtain a statistical description of the two estimators. An histogram of the error values in both cases were created (figure 3). It is evident that SLF is lightly unbiased respect to SRP-Phat, even if mean value and a variance are similar in both cases, thus confirming the result obtained with synthetic data.

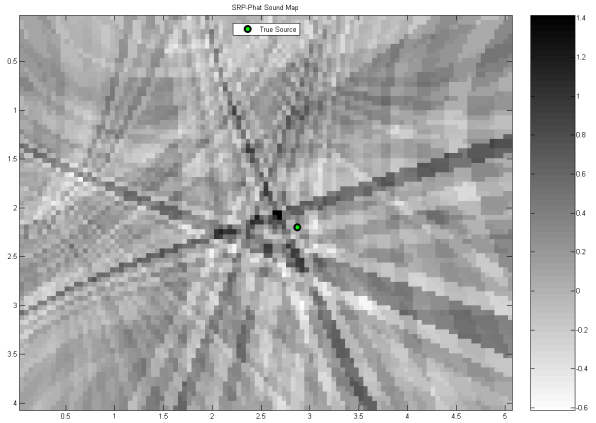
Figure 2(a) and 2(b) shows the difference among the sound maps obtained with the two approaches: SLF shows a contour of the sound that can be helpful to discriminate its nature (talker instead of noise). Of course the acoustic image could be improved using a higher number of microphones.

5. CONCLUSION

A new weighting function for sound source localization has been introduced. The proposed function can be exploited to construct robust sound maps with a clearer definition on lo-



(a)



(b)

Fig. 2. 2D sound field map of ISPAC Lab obtained with (a) SLF and (b) SRP-Phat.

cal maxima with respect to previous approaches. Moreover it is clear that only essential information from GCC can be selected and the number of peak to be considered is a fundamental parameter related to the reverberation time taken in consideration. It is evident that the sound map created with SLF is more significative and can support additive information like sound directivity. For example, in figure 2(a) a bottom-left direction of the speaker can be supposed and actually corresponds to the true propagation of voice.

6. REFERENCES

[1] 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.

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

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

[4] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, vol. 34, no. 3, pp. 1526–1540, June 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] 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.

[7] 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.

[8] 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.

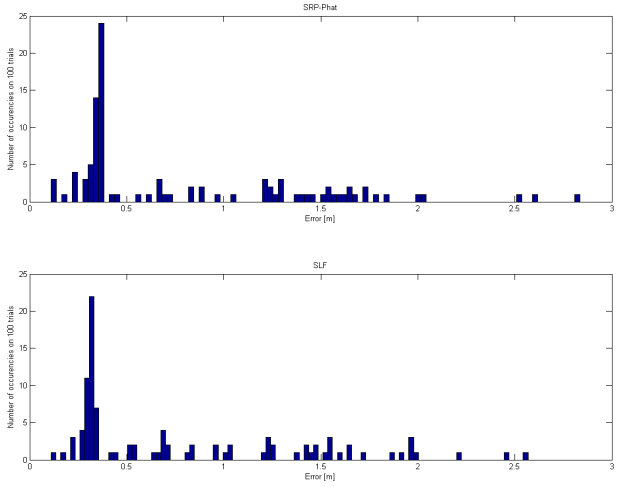


Fig. 3. Histogram of the occurencies of localization error[m] on 100 trials for SRP-Phat and SLF