# EXPERIMENTS IN ACOUSTIC SOURCE LOCALIZATION USING SPARSE ARRAYS IN ADVERSE INDOORS ENVIRONMENTS

*Antigoni Tsiami*[1,3]*, Athanasios Katsamanis*[1,3]*, Petros Maragos*[1,3] *and Gerasimos Potamianos*[2,3]

[1] School of Electr. and Computer Eng., National Technical University of Athens, 15773 Athens, Greece
[2] Department of Electr. and Computer Eng., University of Thessaly, 38221 Volos, Greece
[3] Athena Research and Innovation Center, 15125 Maroussi, Greece

{antsiami,nkatsam,maragos}@cs.ntua.gr, gpotam@ieee.org

## ABSTRACT

In this paper we experiment with 2-D source localization in smart homes under adverse conditions using sparse distributed microphone arrays. We propose some improvements to deal with problems due to high reverberation, noise and use of a limited number of microphones. These consist of a pre-filtering stage for dereverberation and an iterative procedure that aims to increase accuracy. Experiments carried out in relatively large databases with both simulated and real recordings of sources in various positions indicate that the proposed method exhibits a better performance compared to others under challenging conditions while also being computationally efficient. It is demonstrated that although reverberation degrades localization performance, this degradation can be compensated by identifying the reliable microphone pairs and disposing of the outliers.

***Index Terms***— source localization, reverberation, outlier elimination, sparse arrays

## 1. INTRODUCTION

Smart home environments have recently gained significant attention due to the opportunities they offer in terms of ambient assisted living and control via smart interfaces. Equipment in such environments consists of a wide range of sensors placed in the background enabling a more flexible and less intrusive communication. Among several activities in this area lies the DIRHA European funded project [1], which aims to achieve distant speech interaction for the control of home automation employing distributed microphone arrays.

Of importance in this context is the speaker's location, which can be used either as a front end to an automatic speech recognition/speech enhancement system, or to identify the room of activity in order for the system to respond to a command with the proper action. Microphone arrays distributed across the rooms can be exploited to extract a speaker's location. Although much research has been carried out towards this direction, when environmental conditions are extremely adverse, namely characterized by very high reverberation times ($0.6s < T_{60}$ [1]$< 2s$) and extreme noise levels (negative SNRs), source localization becomes challenging and has not yet been successfully addressed.

Existing source localization algorithms can be divided into three main categories: methods based on a) Steered Response Power (SRP), b) High Resolution Spectral Estimation (HRSE) and c) Time Difference of Arrival (TDOA) estimation. An overview can be found in [3]. In [4] the performance of TDOA estimation is investigated in relation to room reverberation and it is demonstrated that reverberation leads to severe degradations, even when at low levels. In [5] reverberation is modelled in terms of source localization, but this analysis is not applicable to environments where microphones are placed on the walls. Also, reported experiments and results on source localization usually consider very small databases with a limited range of positions and moderate reverberation and do not usually allow a clear understanding of the problems arising at more adverse conditions.

In this work we cope with extremely reverberant and noisy conditions and aim to increase robustness towards reverberation effects without having knowledge of the room impulse response or geometry. We propose a dereverberation step to improve the quality of the data and an iterative outlier elimination stage to improve the final source estimation. We experiment with both simulated and real data including a wide number of different sources and a quite small number of microphone pairs and we end up with several interesting observations.

## 2. PROPOSED SYSTEM

Our source localization system belongs in the TDOA estimation methods category and is based on the popular Generalized Cross Correlation - PHAse Transform (GCC-PHAT) [6], due to our need for computational efficiency. It has been shown [7] that PHAT transform is optimal among other time

This research was supported by the EU project DIRHA with grant FP7-ICT-2011-7-288121.

[1] $T_{60}$ is defined as the time taken for the reverberant energy to decay by 60 dB once the sound source has been abruptly shut off [2]

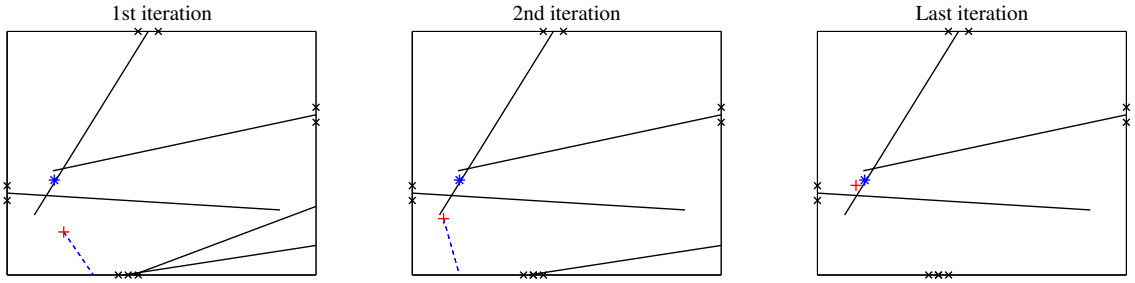


**Fig. 1**: Example of the proposed outlier elimination algorithm (blue "star" is the true source, red "plus" is the estimated source, "X" indicates the microphone positions and dotted line is $max\_dist$, from the estimated position to the possible outlier )

delay estimators when the reverberation is high enough.

### 2.1. TDOA estimation

Given a microphone $i$ we can express its output as:

$$x_i(t) \quad = a_i s(t - \tau_i) + u_i(t) \tag{1}$$

where $x_i$ is the output, $s$ the source signal, $\tau_i$ the time-of-flight (TOF) from the source to the microphone, $u_i$ the noise and $a_i$ the attenuation factor due to the signal propagation delay from the source to the microphone.

The TDOA estimation problem focuses on estimating $\Delta\tau_{ij} = \tau_i - \tau_j$ for a microphone pair $(i, j)$. A lot of methods addressing this issue have been proposed. A brief presentation of some of them can be found in [8]. Our system uses Crosspower Spectrum Phase - Coherence Measure (CSP-CM) [9], based on GCC-PHAT. It is suitable for real-time applications because it is computationally efficient.

We denote by $X_i(f, t)$ the Short-Time Fourier Transform of the signal $x_i(t)$. The CSP-CM method computes

$$C_{ij}(\tau, t) = \int_{-\infty}^{\infty} \frac{X_i(f,t)X_j^*(f,t)}{|X_i(f,t)||X_j(f,t)|} e^{j2\pi f\tau} df \tag{2}$$

and estimates TDOA as $\Delta\tau_{ij} = \arg\max_\tau C_{ij}(\tau, t)$ because $C_{ij}(\tau, t)$ is expected to have a global maximum at $\tau = \Delta\tau_{ij}$.

### 2.2. DOA estimation

After computing a TDOA for each microphone pair, we can extract the source signal's direction-of-arrival (DOA) with respect to that pair. When microphones and source locations lie on the same plane, the root locus of the points that represent possible locations is a half hyperboloid in 2-D space. Assuming a far-field propagation model, we can represent the DOA as a line that connects the source and the middle of the line connecting the two microphones. Then we get [10]:

$$\frac{d\cos\theta}{c} = \Delta\tau_{ij} \Rightarrow \theta = \cos^{-1}\left(\frac{c\Delta\tau_{ij}}{d}\right) \tag{3}$$

where $\theta$ is the angle of DOA, $d$ is the distance between the microphones and $c$ is the sound velocity.

### 2.3. Source location estimation using Least Squares

If the TDOA estimations were ideally correct, all DOA lines should be intersected at a common point. Of course, in practice this does not happen, thus, we have to combine them in order to estimate the final source location. Our approach is based on finding this point that minimizes the sum of the squared distances from DOA lines. Intuitively, this is the point located closest to all DOA lines. We denote as $\mathbf{r}_k$ the unitary vector parallel to the $k$-DOA line (computed from $\theta$) and $\mathbf{p}_{0k}$ the point in the middle of the line connecting the microphone pair $k$. For a random point $\mathbf{a}$ and $\mathbf{a}_{proj_k}$ the projection of $\mathbf{a}$ on the $k$-DOA line, we compute the distance $D_k^2(\mathbf{a}) = ||\mathbf{a} - \mathbf{a}_{proj_k}||^2$, for all $k$. If $\mathbf{A}_k = \mathbf{I} - \mathbf{r}_k\mathbf{r}_k^T$:

$$D_k^2(\mathbf{a}) = (\mathbf{a} - \mathbf{p}_{0k})^T \mathbf{A}_k (\mathbf{a} - \mathbf{p}_{0k}) \tag{4}$$

We find the source location $\mathbf{a}$ by minimizing:

$$E(\mathbf{a}) = \sum_{k=1}^{M} D_k^2(\mathbf{a}) \tag{5}$$

where $M$ is the number of DOA lines. Essentially, we end up with a closed-form solution using Least Squares [11].

### 2.4. Improving Robustness

In practice, the accuracy of the TDOA estimation highly depends on the reverberation time, the noise level and the orientation of the speaker. Thus, we propose a dereverberation process in order to improve the quality of the data, described in Sec. 2.4.1. As it will be demonstrated in Sec. 4, although this step indeed improves the performance compared to the baseline system, the presence of outlier DOA lines continues to degrade final estimation. For this purpose, we also consider an outlier elimination step described in Sec. 2.4.2.

#### *2.4.1. Pre-filtering*

In order to cope with extreme reverberation times, a dereverberation step that precedes that of TDOA estimation was introduced in the system. Reverberation effects can be expressed as follows:

$$x_i(t) = s(t) * h_i(t) + u_i(t) \tag{6}$$

where $h_i(t)$ is the impulse response between channel $i$ and source position. The dereverberation process is mainly based on cepstral prefiltering [12]. Assuming static sources and linear channels that vary slowly with time we can switch to the cepstral domain, transforming the convolutive component into an additive one, with which we can deal via linear filtering. The complex cepstum is preferred instead of the real one as it retains the phase information which is necessary for the signal reconstruction. If we denote by $\hat{h}_i[k]$ the cepstrum of the impulse response, it can be shown that $\hat{h}_i[k] = \hat{h}_{i,ap}[k] + \hat{h}_{i,min}[k]$, where $\hat{h}_{i,ap}[k]$ is the all-pass component and $\hat{h}_{i,min}[k]$ the minimum phase component (MPC). It is claimed [12] that dereverberation can be achieved by subtracting the MPC component of the channel cepstrum, assuming that the additive noise is negligible compared to the convolutive one, i.e. the reverberation and that the MPC of the source signal cepstrum is zero-mean.

#### *2.4.2. Eliminating outliers*

As stated before, reverberation and noise effects may result in erroneous TDOA estimation for some microphone pairs. The pre-filtering technique indeed reduces the reverberation but not totally. As it will be demonstrated in Sec. 4, among the available microphone pairs in most cases there are some that can accurately estimate a source location. Thus, our effort concentrated towards detecting these pairs and disposing of the rest, which are considered as outliers. In search of an objective metric for this purpose, we implemented three methods: i) an SNR-based pair selection assuming that the most reliable microphone pairs should have the highest SNR (based on [10]), ii) a cross-correlation peak value pair selection assuming that the larger peak ensures a better estimation and iii) a TDOA variance-based pair selection so as to dispose of pairs that give significantly different estimations for consecutive frames. None of these hypotheses were valid for the full range of data, because of the severe degradation and false correlation peaks imposed by reverberation. In [13], the authors proposed an iterative method to solve the system of DOA equations, as those for the minimization of (5), which allows the disposition of outliers, namely the projection method. Expressing the system of DOA lines as $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{x}$ is the unknown location, the proposed algorithm iterates over all DOA line equations by projecting the solution on the hyperplanes represented by each individual equation. At step $i+1$, the projected solution is:

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \frac{e_i}{|\mathbf{a}_p|^2}\mathbf{a}_p^T \tag{7}$$

where $\mathbf{a}_p$ is the $p^{th}$ row of matrix $\mathbf{A}$. At the $i^{th}$ iteration, the $p^{th}$ row is utilized and $e_i = b_p - \mathbf{a}_p\mathbf{x}^i$ denotes the error, where $b_p$ is the $p^{th}$ element of vector $\mathbf{b}$. The equation with the maximum distance to the projection point is the possible outlier and will be removed if the error is over a threshold. This algorithm eliminates one equation at a time and is terminated when all the errors are under the threshold or the number of iterations exceeds a pre-set number. This method is efficient when the number of available equations is large as it asymptotically converges to the Least Squares solution. In our case, due to the small number of microphone pairs, this method is not effective. Thus, we propose an alternative iterative method which is described in Alg. 1. First, a source location is estimated using all available DOA lines, as explained in Sec. 2.3. Then, we compute the distances between every DOA line and the estimated source as stated in (4). We choose the maximum among the latter which we compare to a threshold. If it exceeds that threshold, the corresponding DOA line is removed and a new location is estimated using the remaining lines. This procedure is continued until either no DOA line distance exceeds the defined threshold or we are left with only two DOA lines. An example is depicted in Fig. 1.

**Algorithm 1** Proposed Outlier Elimination

```
N ← number_of_DOA_lines
while N ≥ 2 or max_dist > threshold do
    compute source location (sloc) via LS using N lines
    D(k) ← dist(sloc, line(k))^2 for each k-line
    max_dist ← max D(k)
    if max_dist > threshold then
        N ← N − {k}
    end if
end while
```

## 3. DATABASES

### 3.1. DIRHA simulated and real corpora

For the source localization experiments we used two sets of data [14], simulated and real, provided by the DIRHA project, based on a smart home (apartment) located in Trento, Italy. This is equipped with forty microphones, distributed into twelve 2- or 3-element arrays located at the apartment walls, and two 6-microphone arrays located at the ceilings of the two rooms considered of interest, namely the living room and the kitchen (see also Fig. 2). It should be noted that the apartment exhibits significant reverberation.

In the case of simulated data, speech is first recorded in a clean environment in four languages (Austrian German, Greek, Italian, and European Portuguese). The data are then convolved with estimated impulse responses of the DIRHA smart home for a wide range of static source locations, while pre-recorded acoustic events and background noise at high SNR levels are superimposed, giving rise to multi-microphone, noisy, far-field speech data. A total of 40 1-min multi-channel simulated sequences are available (10 for each language). These sequences contain more than one speech segments in different source locations and/or orientations. The whole database consists of 159 different speech segments. It should be noted that only few wall microphone pairs are available for the source localization task, because the arrays are sparsely distributed inside the rooms.

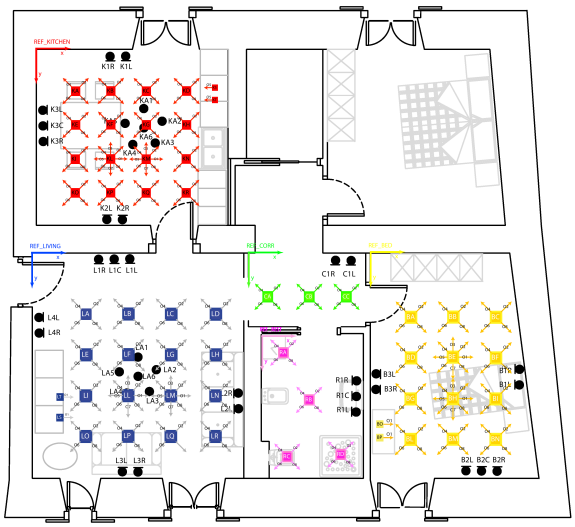

**Fig. 2**: Floorplan of the DIRHA apartment, with all source positions and orientations as well as the 40 microphone positions depicted (from [15]).

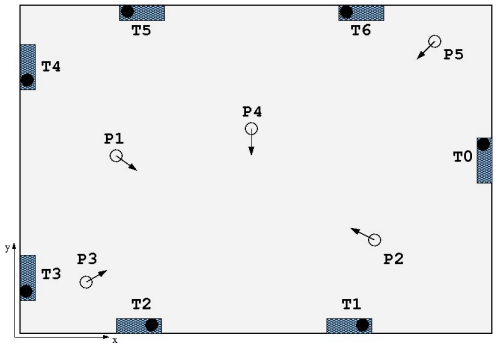


**Fig. 3**: DMN floorplan

The real data contain 10 sessions of recorded wizard-of-Oz like interaction between users and a speech-enabled home-automation system in Italian. In all cases, the user is located in the living room or kitchen, may be moving, and no acoustic events are present (see also [15]). The number of speech segments in this case is 79.

### 3.2. DMN database

For further validation we performed experiments in one more database, namely the Distributed Microphone Network database (DMN), provided by Fondazione Bruno Kessler (FBK). This database was collected in a smart room with 21 microphones distributed in 7 triads on the four walls (see Fig. 3). It consists of five single speaker recordings in five different positions. It is a small database with relatively high reverberation time and noise levels. In contrast to DIRHA databases, this one has a quite large number of available microphone pairs, since all 21 microphones are located within a single room. Source localization results for DMN have also been reported in [16].

| | |
|---|---|
| Oracle | Upper bound without pre-filtering |
| Oracle-D | Upper bound with pre-filtering |
| CSP | CSP-CM |
| CSP-O | CSP-CM with outlier elimination |
| CSP-D | CSP-CM with pre-filtering |
| CSP-D-O | CSP-CM with pre-filtering and outlier elimination |
| SRP | SRP-PHAT |
| SRP-D | SRP-PHAT with pre-filtering |

**Table 1**: Various source localization approaches and acronyms.

| | Pcor | RMSE (in cm) |
|---|---|---|
| CSP | 100% | 34 |
| SRP | 100% | 9 |

**Table 2**: Results on DMN database.

## 4. EXPERIMENTAL EVALUATION

For source localization, fine and gross estimation errors are distinguished, the former corresponding to cases when the distance between the reference and hypothesized sources is less than 50 cm. The percentage of such errors over all speech segments is referred to as the "Pcor" metric. For the computation, the full speech events are considered and one position per speech event was computed. Also, the root mean square error is calculated, separately for fine errors (RMSEf) and for all errors (RMSE). Table 1 summarizes the various implemented methods. For all methods, the window length was $50ms$ and the overlap $30ms$.

First, the results for the DMN database are presented in Table 2. These results concern two baseline systems, the CSP-CM as described in Sec. 2 without the additional steps and the SRP-PHAT system [17, 18]. Both systems achieve 100% correct source estimation, while SRP-PHAT seems to yield more accurate estimations, with RMSE just 9cm.

Next, we experimented with DIRHA simulated and real corpora. In order to evaluate our method, we first obtained the best estimations our algorithm could achieve (the upper bound) if we could estimate the most reliable microphone pairs for each position. The motivation behind this lies in the observation that although CSP-CM fails in most cases to produce satisfying estimations, among all microphone pairs there are several that indeed yield correct ones. Thus, knowing the ground truth source positions, we experimented with all possible microphone pair combinations and obtained an "oracle" result both with and without the pre-filtering step. In terms of comparison, we also experimented with SRP-PHAT both with and without a pre-filtering stage.

Table 3 summarizes the results for all implemented methods on DIRHA databases. As it can be noticed, the problem is very challenging. All baseline methods degrade and fail to produce correct estimations for the whole database. However, the "Oracle-D" result indicates that if we knew or could estimate the most reliable microphone pairs, we could achieve a more accurate source localization result. The two proposed steps, the pre-filtering and the outlier elimination increase the robustness of the baseline system, not achieving however

| | Simulated Data | | | Real Data | | |
|---|---|---|---|---|---|---|
| | Pcor | RMSEf (in cm) | RMSE (in cm) | Pcor | RMSEf (in cm) | RMSE (in cm) |
| Oracle | 77.3% | 16 | 37 | 89.8% | 20 | 27 |
| Oracle-D | 84.3% | 15 | 28 | 82.3% | 17 | 28 |
| CSP | 30.0% | 27 | 109 | 49.1% | 32 | 63 |
| CSP-O | 40.3% | 19 | 210 | 46.8% | 29 | 69 |
| CSP-D | 44.7% | 27 | 75 | 49.1% | 27 | 98 |
| CSP-D-O | 51.0% | 18 | 78 | 49.1% | 27 | 75 |
| SRP | 16.9% | 19 | 178 | 11.0% | 23 | 166 |
| SRP-D | 48.0% | 21 | 105 | 18.7% | 21 | 159 |

**Table 3**: Results on DIRHA corpora.

a high "Pcor" rate. In case of simulated data, the best performance is achieved by the proposed system,"CSP-D-O", yielding $51\%$ fine errors, while "SRP-D" achieves a rate of $48\%$ fine errors. In case of real data, where the speaker moves slowly with time, the best result comes from "CSP","CSP-D" and "CSP-D-O". Here it seems that the proposed additions do not offer much in terms of increasing the system's robustness. This can be explained in two aspects: first, considering dereverberation, it should be noted that the approach followed in Sec. 2.4.1 makes the assumption of a static speaker which means that the subsequent hypothesis of a slowly varying channel impulse response is not accurate. Secondly, in contrast to the simulated corpora where the "speaker" is actually a loudspeaker (meaning a more directive source), here because of real human voice which is less directive, the elimination of outliers does not seem to add much. In some cases it even seems to compromise some not so bad estimations.

Concerning the SRP-PHAT algorithm, its low performance for the DIRHA corpora in comparison to the one for DMN database can be attributed to the very small number of available microphone pairs (5 pairs for Living room and 4 pairs for Kitchen) and the high noise of the former. It seems that although it can give accurate estimations, it needs quite a large number of microphone pairs in order to compensate for the reverberation and noise effects. Also, for real data, this hypothesis of a static speaker is false.

Lastly, it should be pointed out that the CSP approaches yield a source location estimation in much less time than the signal's duration, while SRP-PHAT as implemented in [18] is quite slow and not appropriate for real-time applications.

## 5. CONCLUSION

Results reported in this paper indicate that when the conditions are extremely adverse and the number of available microphone pairs too small, the source localization task becomes quite challenging. We have provided two algorithmic improvements that increase robustness, one based on an efficient way to eliminate outliers and another on pre-filtering to reduce reverberation. We have also demonstrated that even under these conditions, a satisfying accuracy can be achieved if outliers are properly detected. Further study is needed towards modelling reverberation and noise in smart rooms and successfully eliminating their effects, as well as detecting the outliers.

## Acknowledgments

The authors would like to thank M. Omologo and the SHINE group at Fondazione Bruno Kessler (FBK) for having provided us with the DIRHA and DMN corpora.

## REFERENCES

[1] "The DIRHA (Distance-speech Interaction for Robust Home Applications) EU project," [Online] Availble at: `http://dirha.fbk.eu/`.

[2] P.A. Naylor and N.D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[3] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

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

[5] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 791–803, 2003.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[7] C. Zhang, D. Florêncio, and Z. Zhang, "Why does phat work well in low noise, reverberative environments?," in *Proc. ICASSP*, 2008.

[8] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1, Springer, 2008.

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

[10] M. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech and Language*, vol. 9, no. 2, pp. 153–169, 1995.

[11] I. Rodomagoulakis, P. Giannoulis, Z.-I. Skordilis, P. Maragos, and G. Potamianos, "Experiments on far-field multichannel speech processing in smart homes," in *Proc. DSP*, 2013.

[12] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[13] E. Jan and J. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," in *Proc. ICSLP*, 1996.

[14] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. LREC*, 2014.

[15] "Speech detection and speaker localization in domestic environments," [Online] Available at: `http://dirha.fbk.eu/hscma`.

[16] A. Brutti, M. Omologo, P. Svaizer, and C. Zieger, "Classification of acoustic maps to determine speaker position and orientation from a distributed microphone network," in *Proc. ICASSP*, 2007.

[17] J. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, 2000.

[18] H. Do, H.F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. ICASSP*, 2007.