Accuracy Study of a Real-Time Hybrid Sound Source Localization Algorithm

Fernando A. Escobar Juzga[1], Xin Chang[1], Christian Ibala[2], and Carlos Valderrama[1]

[1] Université de Mons, Belgium
{fernando.escobarjuzga,xin.chang,carlos.valderrama}@umons.ac.be
[2] University of Limerick, Ireland
sibala@acm.org

Abstract. Sound source localization in real time can be employed in numerous applications such as filtering, beamforming, security system integration, etc. Algorithms employed in this field require not only fast processing speed but also enough accuracy to properly cope with the application requirements. This work presents accuracy benchmarks of a hybrid approach previously proposed, which is based on the Generalized Cross Correlation (GCC), and the Delay and Sum beamforming (DSB). Tests were performed considering a linear microphone array simulated in MATLAB. Analysis through variations in array size, number of microphones, spacing and other characteristics, were included. Results obtained show that the proposed algorithm is as good as the DSB under some conditions that can be easily met.

Keywords: Accuracy, Sound localization, Generalized Cross Correlation, Beamforming, Computational Complexity, Real Time.

1 Introduction

Numerous applications can be encountered when dealing with sound source localization such as filtering or speech recognition [10], [19], [20], with the use of microphone arrays; by applying beamforming techniques [17], [1], [12], noise, reverberation effects and interference sounds can be filtered by focusing the beam towards the selected sound source location.

One of the well known algorithms to locate sound sources is the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [9], that can provide an angle ϕ which is the sound source direction of arrival (DOA); to compute it, the GCC uses the temporal shift estimation between a pair of microphones i and j that leads to the maximum cross-correlation between them. On the other hand, the Delay and Sum beamforming (DSB) algorithm [12], [3], can be used to build an acoustic energy map (Steered Response Power - SRP) of a predefined Field of View (FoV); under certain conditions it yields the exact position of the sound source. Finally, the Minimum Variance Distortion-less Response (MVDR) is often used to listen to large frequency band signals [2], but implies a big computational cost.

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 146–155, 2013.

In order to optimize and speed up the localization process, we proposed in previous work [7] and [8], a hybrid algorithm that combines the GCC-PHAT with the DSB-SRP to reduce the search area and decrease computational complexity. Our theoretical analysis showed a computational advantage of the hybrid algorithm over the DSB-SRP, thanks to the reduction of evaluated points where the energy response is computed.

Because we are interested in obtaining the exact position of a sound source, the GCC-PHAT solution resulted insufficient as it only provides the source's (DOA); on the other hand, since the SRP computation requires the output from the GCC-PHAT algorithm, it was straightforward to combine them both and optimize its execution. The basic idea of this approach is to create a reduced detection zone by drawing two lines: one above and the other below the GCC detected angle, with an inclination ε chosen by the user; then, the SRP can be computed on the constrained area as shown in Figure 1.

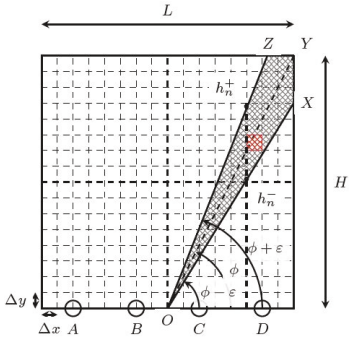


Fig. 1. 2-dimensional, 3x3 m, Field of View (FoV). The constrained area is enclosed by the upper and lower lines and the borders of the FoV.

The main contribution of this work is presented in Section 4 where we show the algorithm's response in terms of localization accuracy and number of detected maxima when changing microphone spacing, the epsilon value and inducing an artificial error to the angle obtained with the GCC algorithm. We employ the toolbox provided by the University of Kentucky [13], which generates synthetic scenarios to analyze with microphone arrays; several simulations were executed to derive accuracy measurements among them all.

The rest of the paper is organized as follows: Section 2 will summarize the latest research on sound localization using microphone arrays. In Section 3, we describe the proposed hydrid algorithm; Section 4 presents our simulated accuracy analysis and finally Section 5 lists our conclusions and suggested further work.

2 Related Work

Several works have been reported in the field of sound source localization. Valin et al. use an 8 microphone array to estimate the time delay of arrival (TDOA) using the GCC-PHAT technique on a moving robot [18] and report an average of 3 degrees accuracy in computing the direction of arrival (DOA); the exact position of the detected target is left as future work though. On a second work [19], the authors use a particle filtering technique for tracking moving sound sources with 2, 8-microphone array configurations. They report high accuracy detecting both elevation and azimuth angles.

Another common technique to locate sound mainly in human shaped robots, is called Head Related Transfer Function (HRTF) which estimates the difference in level intensity between the two ears (microphones); however, whilst human ears are shaped in a special way to enhance localization, the algorithm is rather complex for real time implementation. Some related work can be found in [11], [14] and [6].

There are some researches working on three dimensional sound localization such as [16], [5] and [15]. Reports in [16] describe the first working, scalable and cost-effective array that offers high-precision localization of conversational speech in large semi-structured spaces; it achieves high throughput for real-time updates of tens of active sources. Yoko et. al [15], proposed a spherical microphone array design for spatial sound localization. This structure has 64 microphones arranged in a 350-mm-diameter sphere. It is designed to be mounted on a mobile robot with omni-directional directivity in both azimuth and elevation angles. In order to achieve better accuracy, the number of microphones is substantially raised in this kinds of designs. Since the computation load increased, high performance processors are required. Adittionally, bigger areas are occupied.

3 Proposed Hybrid Algorithm

Since our algorithm has already been presented in previous articles [4], [7], [8], only a brief explanation of its operation will be provided; readers are encouraged to revise the cited references for more detailed information.

The Generalized Cross Correlation between two signals provides the estimation of the temporal shift between two microphones i and j that leads to the maximum cross-correlation between them as in Equation 1:

$$\Delta_{ij} = arg_k \ \ max \ R(k) \tag{1}$$

The cross correlation between two microphones is computed by taking the inverse Fourier transform of the product of the first microphone FFT (Fast Fourier Transform) and the conjugated FFT of the second one. To correct the effect of phase, and improve robustness against noise and other undesired effects, there is a correction called PHAT, i.e. phase transform that can be applied yielding Equation 2:

$$R(k) = \quad IFFT \quad \left(\frac{FFT(f(t)).FFT^*(g(t))}{\mid FFT(f(t)).FFT^*(g(t))\mid^{\beta}} \right) \tag{2}$$

Equation 2 is defined as the GCC-PHAT; β is a coefficient factor in the interval $(0, 1)$. The IFFT is performed to go back to the time domain and extract the corresponding value of index k. The value of k can be computed by taking the index of the maximum value from the GCC-PHAT output. Using the far field approximation, the cosine of the angle of arrival, measured by microphones i and j, can be computed as in Equation 3:

$$cos(\phi)_{ij} = \quad \frac{kv_s}{f_s d_{ij}} \tag{3}$$

Where f_s is the sampling frequency, d_{ij} is the distance between microphone i and j, and v_s is the sound speed.

The output of the GCC algorithm can be used to compute the energy response of a predefined FoV. By assuming the sound source to be located at a certain point in space, it is possible to establish the theoretical delay of the signal between every pair of microphones; using such delays, we can extract and add up the energy contribution of every pair from the GCC output. Under certain microphone array configurations and, assuming a single sound source, there will only be one point in the space where, all delays will match the maximum energy possible, that is, the real source location.

As shown in Figure 1, we can restrain the search region by focusing on the relevant part of the FoV, using the computed angle. Since the hybrid approach only restricts the search area, the output is expected to be as accurate as the regular algorithm. In terms of computational cost, great reduction is obtained by considering less points but other computations are required to establish the restricted area boundaries. More specifically, it's necessary to perform some tangents and cotangents whose amount, depend on the size of the small squares. As shown in Figure 1, the number of small squares to be evaluated in each column, depend on the heights h_1^- and h_1^+ which are defined as follows:

$$h_n^- = \quad n \cdot \Delta_x \cdot \tan(\phi - \varepsilon) \tag{4}$$

$$h_n^+ = \quad n \cdot \Delta_x \cdot \tan(\phi + \varepsilon) \tag{5}$$

In summary, the total number of tangents for a specific FoV of length X and resolution R is:

$$NumTangents = \quad \left(\frac{2 \cdot X}{R} \right) \tag{6}$$

The following subsection will present our accuracy estimations for the proposed algorithm.

4 Accuracy Analysis

The hybrid approach has two error sources: the one introduced by the GCC algorithm when obtaining the angle of arrival, and the one inherent to the DSB-SRP

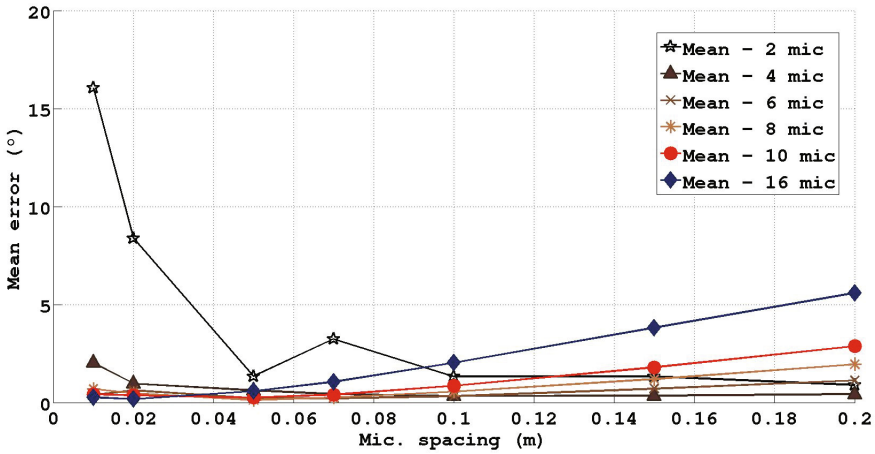


Fig. 2. GCC mean error for different spacings and number of microphones. The algorithm provides a good accuracy with more than 2 microphones; increasing their number does not provide better results. The error obtained is less than 5 degrees for the majority of cases.

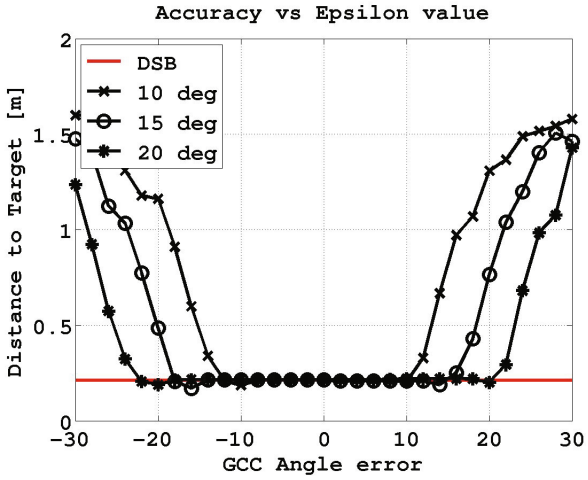
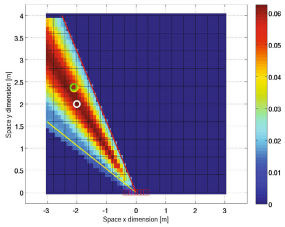


Fig. 3. Accuracy for different values of epsilon vs GCC error. The hybrid algorithm performs as good as the DSB when the value of epsilon is greater or equal than the error of the GCC, if any.

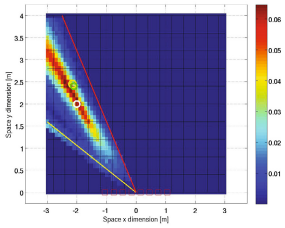
itself. Because both algorithms can yield errors at the same time we present the results considering them both. For all the following simulations we varied the target position for a better generalization and present the mean value as the final result.

We conducted a first study related to GCC precision using the aforementioned toolbox [13]; after considering different scenarios for linear arrays, we obtained an average error of $2-4°$ for arrays of at least 4 cms long with more than 3 microphones. Arrays that did not respect such conditions yielded up to 16-degrees errors. This results are shown in Figure 2; when the number of microphones increases, so does the error. Since every pair of microphones provide an estimated angle, the error increase can be caused by the averaging of all measured values.

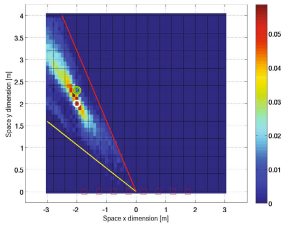
The hybrid algorithm was analysed in terms of the GCC error; a few parameters can be changed to tune up its output but according to our results, only some of them are relevant. Initially, we varied parameter ε and measured the



(a) 10 cm



(b) 30 cm

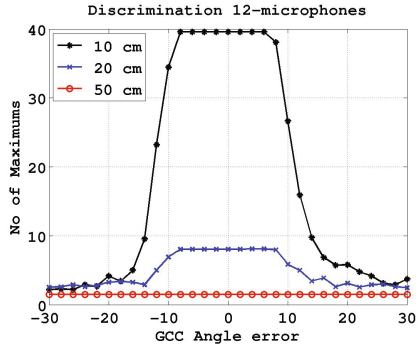


(c) 50 cm

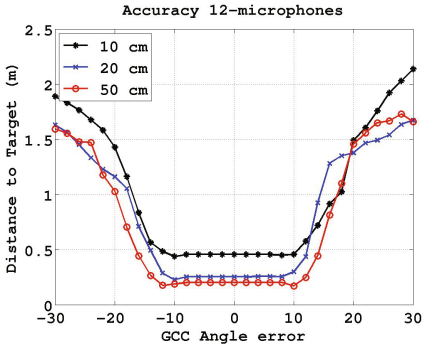
Fig. 4. SRP response of an array of 8 microphones at different spacings using the hybrid GCC-DSB algorithm

distance of the detected point to the real source location. The test was performed under a configuration of 8 microphones, 50 cm apart, and assuming a GCC error range of $[-30°, 30°]$; we obtained Figure 3. The horizontal red line represents the output when using the DSB-SRP, while the black lines show the behavior of the GCC-DSB for different epsilon values; when the GCC angle error is greater than the value of epsilon, accuracy tends to be rapidly lost. On the other hand, if the GCC error lies whithin the range comprised by epsilon, the accuracy is the same as with the DSB-SRP.

Through the simulations performed, we noticed that one of the most important parameters that affected localization accuracy was the microphone spacing; for instance, consider Figure 4, where the same scenario is presented for three different spacings; in Figure 4a the microphones are only 10 cm apart and a big red fringe of points are detected as possible source location. On the contrary, Figures 4b and 4c show that, when increasing their separation, the energy plot is much clearer, thus enabling better accuracy. Since more than one point can be detected as the maximum, we decided to select the mean point between all of them as the source location. In Figure 5 the algorithm is tested with 12 microphones and a 10° epsilon, changing their spacing. In Figure 5a we can see



(a) Number of maximums detected



(b) Mean distance to target obtained

Fig. 5. Performance on a 12-microphone array for different spacings between each one

the amount of points that are detected for three different cases; although the number of maximums between 20 and 50 cms does not change much, the difference in array size greatly does. In accordance with this result, yet with smaller differences, Figure 5b shows that the accuracy improves for longer spaced arrays. Once more, the validity of the result holds as long as the GCC error is less than the value of epsilon.

A final study was carried on to understand the behavior of the algorithms in relatively small arrays; experimental results using the aforementioned database showed that for arrays of 10 cms long or less, even with no GCC angle error, is not possible to obtain a good precision. According to our estimations, sufficiently accurate results can be obtained with arrays of at least 20 cms long; although, as shown in Figure 4, a considerable amount of maximums will be detected, on average, good estimations can still be obtained with this configuration. Our results show that precision obtained tends to slightly improve with more microphones but is practically the same in all cases.

5 Conclusions

Through the analysis performed in this work, we could verify that the proposed hybrid algorithm can perform as accurately as the traditional DSB-SRP with linear microphone arrays. An important factor that can drastically change the output from the algorithm is the error induced by the GCC-PHAT algorithm; when the angle error is greater than the value of parameter epsilon, the algorithm loses its accuracy.

Linear microphone arrays whose microphone spacing is less than $40 - 50$ cms present difficulties to establish the exact coordinate of the sound source; when the array length is fixed, irrespective of the number of microphones, the localization accuracy is very similar. Since accuracy changed with target position and angle, in general terms we consider the best results were obtained when using between 4 to 10 microphones, for both the GCC and the GCC-DSB.

The aforementioned analysis were done at the theoretical level using an artificial database, which generated an impulse response, noised signal, for each microphone however, we consider that a similar study, with real microphone and signals is necessary to verify the robustness of the simulator engine, and to confirm the validity of the conclusions reached in this paper.

The severity of the localization error of the algorithm can vary depending on the application domain; if employed for detecting small objects, an error of 10 cm can be unacceptable, but the contrary might occur for locating speakers, since a person's personal space can be at least $30cm^2$. Improvements can be also obtained by changing the decision function when different maxima have been found or by applying other methodologies to compute the SRP after the angle has been measured. This is however out of the scope of the presented work.

Acknowledgements. This work was funded by the Wallonia Region DG06 Belgium under grant 917005-CTEUC 2009 Eureka ITEA DiYSE.

References

1. Benesty, J., Chen, J., Huang, Y., Dmochowski, J.: On microphone-array beamforming from a mimo acoustic signal processing perspective. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1053–1065 (2007)
2. Cox, H., Zeskind, R., Owen, M.: Robust adaptive beamforming. IEEE Transactions on Acoustics, Speech and Signal Processing 35(10), 1365–1376 (1987)
3. Dmochowski, J.P., Benesty, J., Affes, S.: A generalized steered response power method for computationally viable source localization. IEEE Transactions on Audio, Speech, and Language Processing 15(8), 2510–2526 (2007)
4. Escobar, F.A., Ibala, C., Chang, X., Valderrama, C.: Fast accurate hybrid algorithm for sound source localization in real time. International Journal of Sensors and Related Networks 1(1), 1–7 (2013)
5. Fréchette, M., Letourneau, D., Valin, J.-M., Michaud, F.: Integration of sound source localization and separation to improve dialogue management on a robot. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2358–2363. IEEE (2012)
6. Hörnstein, J., Lopes, M., Santos-Victor, J.: Sound localization for humanoid robots building audio-motor maps based on the hrtf (2006)
7. Ibala, C., Vachaudez, J., Fourtounis, G., Possa, P., Valderrama, C.: Combining sound source tracking algorithms based on microphone array to improve real-time localization. In: 2012 Proceedings of the 19th International Conference on Mixed Design of Integrated Circuits and Systems (MIXDES), pp. 478–483 (May 2012)
8. Ibala, C., Escobar, F.A., Chang, X., Valderrama, C.: Hybrid algorithm computation methodology to accelerate sound source localization. International Journal of Microelectronics and Computer Science 3(3), 99–110 (2012)
9. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing 24(4), 320–327 (1976)
10. Li, Q., Zhu, M., Li, W.: A portable usb-based microphone array device for robust speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pp. 1301–1304 (April 2009)
11. MacDonald, J.A.: An algorithm for the accurate localization of sounds (2005)
12. McCowan, I.: Robust speech recognition using microphone arrays (2001)
13. University of Kentucky. Performance analysis of srcp image based sound source detection algorithms (2010)
14. Rothbuncher, M., Kronmuller, D., Durkovic, M., Habigt, T., Diepold, K.: Hrtf sound localization (2011)
15. Sasaki, Y., Kabasawa, M., Thompson, S., Kagami, S., Oro, K.: Spherical microphone array for spatial sound localization for a mobile robot. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 713–718. IEEE (2012)
16. Sun, D., Canny, J.: A high accuracy, low-latency, scalable microphone-array system for conversation analysis (2012)
17. Tashev, I.J.: Sound Capture and Processing: Practical Approaches. Wiley Publishing (2009)

18. Valin, J.-M., Michaud, F., Rouat, J., Letourneau, D.: Robust sound source localization using a microphone array on a mobile robot. In: Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), vol. 2, pp. 1228–1233 (October 2003)
19. Valin, J.-M., Michaud, F., Rouat, J.: Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Robotics and Autonomous Systems 55(3), 216–228 (2007)
20. Zwyssig, E., Lincoln, M., Renals, S.: A digital microphone array for distant speech recognition. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5106–5109 (March 2010)