HYBRID ALGORITHM FOR ROBUST, REAL-TIME SOURCE LOCALIZATION IN REVERBERANT ENVIRONMENTS

J. Michael Peterson and Chris Kyriakakis

Immersive Audio Laboratory
Integrated Media Systems Center
USC Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089

ABSTRACT

The location of an acoustical source can be found robustly using the Steered Response Pattern - Phase Transform (SRP-PHAT) algorithm. However SRP-PHAT can be computationally expensive, requiring a search of a large number of candidate locations. The required spacing between these locations is dependent on sampling rate, microphone array geometry, and source location. In this work, a novel method will be presented that calculates a smaller number of test points using an efficient closed-form localization algorithm. This method significantly reduces the number of calculations, while still remaining robust in acoustical environments.

1. INTRODUCTION

Beamforming is often used for removing noise and reverberation from speech signals by taking advantage of spatial information. The array response is steered to concentrate on the signal at a given location and attenuate noise and interference from other directions. The location is usually not known and must be estimated from the data. Beamformers do not perform well in the presence of steering errors, requiring accurate location estimates [1]. In addition to beamformers, the location could be used in a joint camera-microphone teleconferencing system [2] or for speaker segmentation [3]. So source localization is an integral part of microphone array processing.

Several methods have been developed for estimating an acoustical source location. Algorithms, like SRP-PHAT, have good robustness in the presence of room effects [4]. SRP-PHAT can be quite complex requiring the calculation of a large number of test points in the region of possible source locations. The location is chosen to be the point that results in the highest energy or likelihood. The proper distance between points is determined by the mapping of the Nyquist rate from the time domain to space. As such, the number of candidate locations is dependent on the sampling frequency as well as aperture size and the range of the source. The number of points can be reduced if the source is constrained to a plane or the far field.

Alternatively, the problem can be implemented as a two-step process. First the generalized cross-correlation (GCC) is used to find the time delays. Then those time delays are used to estimate a three dimensional location [2, 5]. Frequently errors, sometimes called anomalies, occur in the time delay estimates [6]. These anomalies are caused by strong reflections of the sound source, which are sometimes greater in energy than the direct signal. The direct path can be obstructed or attenuated because of source and microphone directivity. Anomalous time delay estimates create large errors in estimation. So while these algorithms are quite fast, they lack robustness.

Instead of blindly testing many candidate locations, a novel algorithm, called Hybrid Localization, will be presented. This algorithm is well suited to locating a source in the near field with a large aperture microphone array. Using multiple time delay estimates from each microphone pair, it uses a two-step algorithm to generate a set of candidate locations. These locations become the candidate locations for SRP-PHAT. Although the calculation of these candidate points requires some computation, it reduces the total computational cost compared to SRP-PHAT. This is accomplished without a decrease in the robustness of the location estimates.

2. MODEL

In the following discussion, the received sound signals will be modeled as

$$x_i(n) = \sum_j h_{ij} * s_j(n) + \psi_i(n) \tag{1}$$

where s_j is the signal from the j^{th} sound source, h_{ij} is the filter's impulse response between the j^{th} sound source and the i^{th} microphone. The number of sources is represented by N_{src} and M is the number of microphones. The noise for each channel is represented as $\psi_i(n)$ and it is independent of the noise in other channels. It is also assumed that all sources are independent from each other and from the noise.

GCC is computed in the frequency domain, by converting a frame of time data using an FFT.

$$X_i(\omega) = \sum_j H_{ij} S_j(\omega) + \Psi_i(\omega) \tag{2}$$

The Phase Transform (PHAT) is a GCC defined as

$$R_{ik}(\omega) = \frac{1}{|E[X_i(\omega) X_k^*(\omega)]|} E[X_i(\omega) X_k^*(\omega)] \tag{3}$$

This equation is then transformed back to the time domain. The results are used as the energy function for SRP-PHAT and to estimate time delays. In order for (2) to be valid, the frame must be longer than the length of the impulse response.

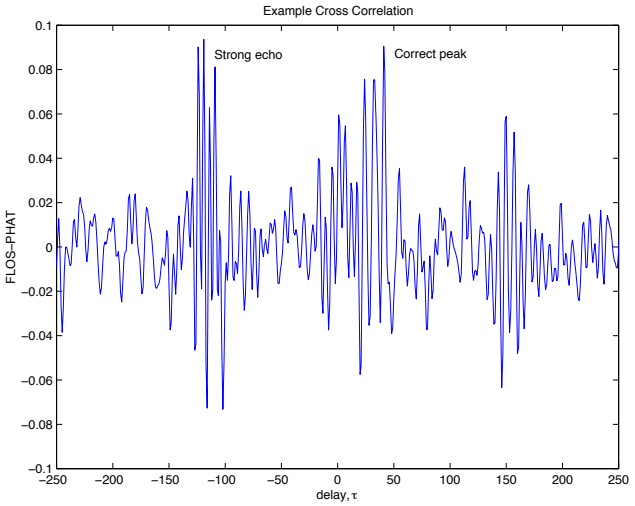

Fig. 1. An example frame using the PHAT method. The GCC consists of multiple delays embedded in noise. Since the direct path is less energetic than one of the reflections, this frame results in an anomaly.

The impulse response can be characterized as having three distinct parts. First is the direct path. This is followed by several discrete reflections and then diffuse reverberation. The length of the impulse response is designated by reverberation time, T_{60}. This is the time that the signal energy decays by 60 dB. Typical rooms have a reverberation time of 300 to 700 ms [7], which results in a very long impulse response.

However, the impulse response can be approximated as

$$h_{ij} \approx \alpha_{ij}^{(0)}\delta(n-\tau_{ij}^{(0)}) + \sum_{l=1}^{N_r} \alpha_{ij}^{(l)}\delta(n-\tau_{ij}^{(l)}) \tag{4}$$

Delay elements, $\delta(n-\tau_{ij}^{(l)})$, represent the direct path and the N_r strongest early reflections. The direct path is designated as $l=0$. The attenuation of the sound energy is represented as $\alpha_{ij}^{(l)}$. The reverberation is included with the noise, $\psi_i(n)$. So the noise includes reverberations of the sources and diffuse noise. Shorter frame sizes can now be used, since the reverberation is no longer considered part of the impulse response.

The resulting GCC will consist of several peaks embedded in a noise floor. The peaks correspond to the delays in the direct path and the early reflections. The noise floor is caused by the reverberation and noise. This results in the following model for the cross-correlation.

$$r_{ik}(n) = \sum_l \acute{\alpha}_{ik}^{(l)}\delta(n-\acute{\tau}_{ik}^{(l)}) + \eta_{ik}(n) \tag{5}$$

where $\eta_{ik}(n)$ is the noise floor with a variance of σ_{ik}^2 and $\acute{\alpha}_{ik}^{(l)}\delta(n-\acute{\tau}_{ik}^{(l)})$ represents the peaks, corresponding to the delay elements in (4). An example frame can be seen in fig. 1. If each channel has N_l reflections, the resulting cross-correlation could have as many as $(N_l+1)^2$ peaks. In practice, the number of significant reflections is usually fairly low.

3. HYBRID LOCALIZATION

The time difference of arrival of the direct path is non-linearly related to location.

$$\tau_{ij} = |\mathbf{r}_s - \mathbf{m}_i| - |\mathbf{r}_s - \mathbf{m}_j| \tag{6}$$

where $\mathbf{r}_s$ is the source location and $\mathbf{m}_i$ is the i^{th} microphone. Spherical intersection (SX) [8] and spherical interpolation (SI) [2] use the time delays corresponding to the maxima in GCC to calculate the location. Starting with (6), a set of linearlized equations can be created and collected into a matrix equation. For example, SX is stated as

$$\hat{\mathbf{r}}_s = (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t(\mathbf{b} - \mathbf{\Delta}\hat{R}_s) \tag{7}$$

The zeroth microphone is located at the origin. The matrix $\mathbf{A}$ is composed of the remaining microphone locations and $\mathbf{A}^t$ is its transpose. $\hat{R}_s$ is the distance of the source to the origin and is equal to the norm of $\mathbf{r}_s$. The vector $\mathbf{\Delta}$ is composed of delay elements $d_{i0} = c/F_s\tau_{i0}$, where c is the speed of sound and F_s is the sampling rate. The i^{th} row of vector $\mathbf{b}$ is

$$b_i = \frac{1}{2}(||\mathbf{m}_i||^2 - d_{i0}^2) \tag{8}$$

Since $\hat{R}_s$ is unknown, it must be estimated. This is accomplished by squaring (7), substituting $||\hat{\mathbf{r}}_s||^2$ with $\hat{R}_s^2$ and solving the resulting quadratic equation. The location is estimated with the resulting $\hat{R}_s$.

Unfortunately, the time delay estimates of individual cross-correlations are quite noisy, introducing significant localization errors. The underlying assumption of SX and SI is that the dominant peak is at the correct time delay for the source. However, frequently the estimated time delay corresponds to a strong reflection or an interfering source near the microphone pair. Reflections can be stronger than the direct path because of an obstruction of the direct path or the directivity of the sound source. For example, a human speaker is not an omni-directional sound source and the sound propagating in front of a person is more energetic than the sound propagating behind [9].

Although SX is not robust, it is quite fast, so it can be used to generate a set of candidate locations quickly. Several maxima from each channel pair are used to find several possible time delays. The number of time delays is denoted by N_p. The individual time delay estimates can be combined in N_p^M different combinations, where M represents the number of microphone pairs. This results in an unrealistically large number of combinations. So in practice, it is best to use a subset of the channel pairs to estimate an initial set of locations. When using SX, only three pairs are needed to create a set of points, which results in N_p^3 candidate locations.

The next step matches the time delays from the other microphone pairs to the initial locations.

$$d_{i0}^{(c)} = \arg\min_k ||\mathbf{m}_i^t\hat{\mathbf{r}}_c - (b_i^{(k)} - \hat{R}_c d_{i0}^{(k)})||^2 \tag{9}$$

for all i not in the original set of microphone pairs. The time delays for the remaining microphone pairs are chosen to minimize the error, which is derived from the SX equations.

After the best time delay estimates are found for each candidate location, the locations are re-estimated using all the channel

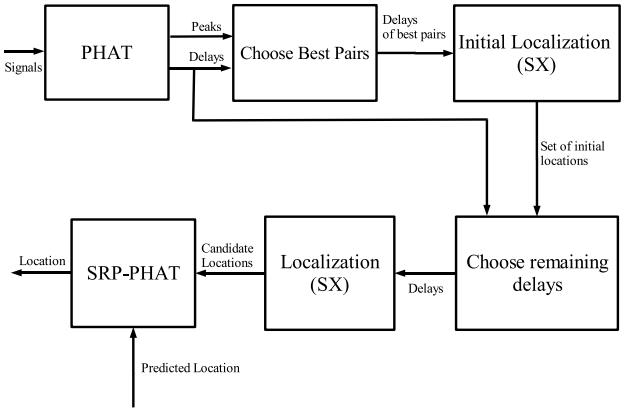

Fig. 2. A block diagram of the Hybrid Algorithm

pairs for SX. These locations are used as the test locations, $\mathbf{q}$, in SRP-PHAT.

$$\hat{\mathbf{r}}_s = \arg\max_{q} \sum_{i=1}^{M} \sum_{j=1}^{M} r_{ij}(\tau_{ij}) \tag{10}$$

where τ_{ij} is related to the test location $\mathbf{q}$ and r_{ij} is the PHAT GCC. Traditionally, SRP-PHAT must sample a large set of points. Many of these points are very unlikely to be the source location. However the candidate locations generated by Hybrid Localization are all very likely to be the location of the source. As long as N_p is small, there are relatively few candidate locations.

Frequently, it is known that a source is located in a certain region. So all candidates outside of this region can be pruned. In addition, past location estimates can be used to predict the current location. This is called tracking and the results of tracking are included in the set of candidates. These modifications increase localization accuracy.

4. A METRIC FOR BEST CHANNEL PAIRS

If all channel pairs were equally good, then it wouldn't matter which pairs were used. Unfortunately, this is not the case. So a metric should be developed in order to choose which channel pairs best estimate the set of initial locations. The metric used in this paper can be developed using (5). Due to the PHAT weighting, the energy of the cross-correlation over all time delays is unity, and σ_{ik}^2 can be estimated by subtracting the energy of the peaks from the total energy.

$$\sigma_{ik}^2 = \sum_{n} r_{ik}^2(n) - \sum_{l} (\acute{\alpha}_{ik}^{(l)})^2 = 1 - \sum_{l} (\acute{\alpha}_{ik}^{(l)})^2 \tag{11}$$

Intuitively, the best channel pairs to use are those with the lowest energy noise floor or alternatively those pairs with the highest peak energy.

It turns out that (11) is related to the early energy to total energy ratio of the impulse response.

$$D = \frac{\int_0^{50ms} [h(t)]^2 dt}{\int_0^{\infty} [h(t)]^2 dt} \tag{12}$$

This measure is often used to determine intelligibility of sound when designing acoustic spaces [7]. Intuitively, the best microphone pairs are those with the highest early energy to late energy ratio.

To recap, a block diagram of Hybrid Localization can be seen in fig. 2. First, PHAT is used to find N_p time-delay estimates for each microphone pair. The metric (11) determines which three microphone pairs SX uses to estimate the set of initial candidate locations. Using the remaining microphone pairs, the time delays that correspond to the candidate locations are determined using (9). Finally, the candidate locations are re-estimated and SRP-PHAT is used to test these locations for the one with maximum energy.

5. SIMULATIONS

The hybrid algorithm was tested in both simulated and real scenarios. The resulting estimates were compared with those obtained using SRP-PHAT [4] and SI [2]. The SRP-PHAT candidate locations are chosen in a non-linear optimal fashion based on the Nyquist rate. This method requires fewer candidate locations then a linear spacing. Hybrid source localization was performed using increasing values of N_p from 2 to 6. The resulting candidate locations were pruned to match the region of interest used in SRP-PHAT and the previous location estimate was added to the candidates. The error is defined as

$$E = \frac{1}{L} ||\mathbf{r}_s - \hat{\mathbf{r}}_s||^2 \tag{13}$$

where L is the number of frames. While the hybrid algorithm could be used to find multiple sources, only the single source case was tested in this paper.

The simulated room had dimensions of 4 m by 5 m by 3 m. The omni-directional microphones are placed in the middle of each wall at the elevation of 1 m and 0.1 m below the ceiling at a distance of 0.1 m from the wall. The region of interest is defined as a box with dimensions of 2 m by 2.5 m and 0.4 m. In order to adequately test the space, SRP-PHAT requires about 20,000 points.

The source locations were placed in random locations inside the region of interest. The sources consist of human speech by both males and females in English and French. Human speech is not omni-directional, so the directivity data from [9] was used. The impulse response was created using the image method. The ceiling had an absorption coefficient of 0.3 and the floor had a coefficient of 0.7. The walls had an absorption coefficient ranging from 0.05 to 0.3, which results in a reverberation time of 430 ms to 270 ms calculated using Sabine's formula. Low frequency noise was added to the signal at a SNR of 40 dB.

The resulting estimation errors for the various algorithms can be seen in fig. 3. The hybrid algorithm used $N_p = 3$ time delays. It can be seen that the hybrid algorithm has approximately the same error as SRP-PHAT, while both methods vastly outperform SI.

6. EXPERIMENTS IN AN ACTUAL ROOM

The hybrid algorithm was also tested in an acoustically untreated room. The test data included human speakers standing in marked locations, so that their location could be easily determined. Their voice was recorded by eight microphones spread out on one wall and the ceiling. The region of interest was defined as a box with

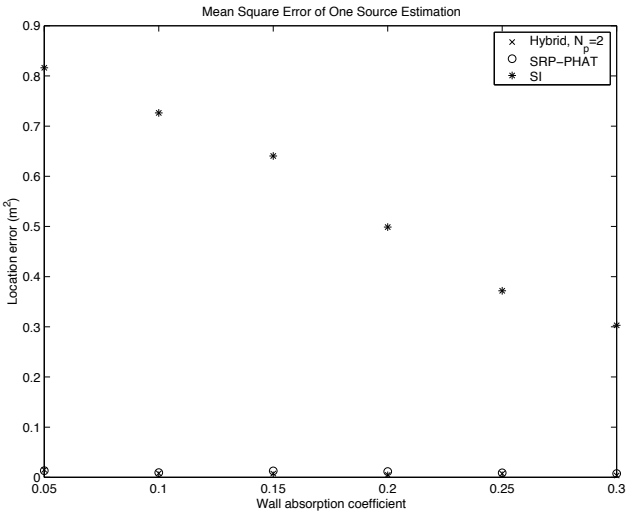

Fig. 3. Results of single source simulation for several absorption values, α.

Method	MSE of Location (m^2)
SI	3.3858
SRP-PHAT	0.3346
Hybrid, $N_p = 2$	0.3709
Hybrid, $N_p = 3$	0.3125
Hybrid, $N_p = 4$	0.3020
Hybrid, $N_p = 5$	0.4004
Hybrid, $N_p = 6$	0.3780

Fig. 4. Table of real room results using SI, SRP-PHAT, and the new Hybrid algorithm for several values of N_p.

a volume of 9 m^3. For this experiment, SRP-PHAT required only 5000 non-linearly spaced points.

As can be seen in fig. 4, even with $N_p = 2$, the resulting error of Hybrid Localization is statistically insignificant when compared to the computationally more expensive SRP-PHAT. It vastly outperforms the SI algorithm. So Hybrid Localization retains the robustness of SRP-PHAT.

One concern for real-time localization is the speed of computation. By counting the number of operations required to estimate locations, it can be shown that Hybrid Localization is much faster than SRP-PHAT. To increase the speed of SRP-PHAT, a table look-up method is used to find the delays, which are calculated beforehand. Fig. 5 shows the number of points that can be in the region of interest for SRP-PHAT to have an equivalent computational cost compared to Hybrid Localization. In this case M represents the number of channel pairs. With eight microphones, $M = 28$; so Hybrid Localization, with $N_p = 2$, is equivalent in cost to searching 290 points. The room experiment required 5000 points for the look-up table so Hybrid Localization requires a tenth of the computation cost.

7. CONCLUSION

Hybrid Localization is a good compromise between robustness and ease of computation. It uses SX to create a set of candidate locations to be used in SRP-PHAT. This algorithm combines the best aspects of SX and SRP-PHAT. It greatly reduces the computation cost of SRP-PHAT, while still retaining the robustness. The new hybrid algorithm is an effective solution for robust real-time source localization.

M	$N_p = 2$	$N_p = 3$	$N_p = 4$	$N_p = 5$
7	450	1640	4130	8570
15	330	1280	3400	7330
28	290	1140	3100	6830
31	280	1120	3060	6770

Fig. 5. This table shows how many points can be sampled for SRP-PHAT to be equivalent to Hybrid Localization for a given number of channel pairs and number of peaks. If more points are required than Hybrid Localization is more efficient.

8. ACKNOWLEDGMENTS

Research presented this paper was funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and in part by the Department of the Army under contract number DAAD 19-99-D-0046. Any Opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation and the Department of the Army.

9. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*. Springer, 2001, ch. Robust Adaptive Beamforming: Signal Processing Techniques and Applications.

[2] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proceedings of ICASSP*, 2000.

[3] G. Lathoud and I. A. McCowan, "Location based speaker segmentation," in *Proceedings of ICME*, 2003.

[4] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001, ch. Robust Localization in Reverberant Rooms.

[5] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 8, December 1987.

[6] M. Jian, A. Kot, and M. Er, "Performance study of time delay estimation in a room environment," in *Proceedings of ISCAS*, 1998.

[7] H. Kuttruff, *Room Acoustics*. Elsevier Applied Science, 1991.

[8] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 8, August 1987.

[9] W. T. Chu and A. C. Warnock, "Detailed directivity of sound fields around human talkers," Institute for Research in Construction, National Research Council Canada, Tech. Rep., December 2002.